

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 19, 2014

J. Dong
M. Zhang
Huawei Technologies
B. Zhang
The University of Arizona
M. Boucadair
France Telecom
October 16, 2013

Requirements for Power Aware Network
draft-dong-panet-requirement-02

Abstract

Energy consumption of networks is rising fast, which results in the increase of network operational costs. There are emerging demands from operators for power-aware networking (PANET) which could adaptively reduce the network energy consumption when possible. This document presents the requirements which should be considered in building a power aware network.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 19, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Requirements on Network Elements	3
3. Requirements on the Whole Network	3
4. Requirements on Network Control Plane	5
5. Requirements on Management Plane	5
6. IANA Considerations	6
7. Security Considerations	6
8. Acknowledgements	6
9. References	6
9.1. Normative References	6
9.2. Informative References	6
Authors' Addresses	7

1. Introduction

With the increase of network services and exponential growth of traffic volume, the network operators are expanding their infrastructures with more high-capacity, full-featured network devices, which also leads to the increase of network energy consumption. Besides, today's service provider networks are mostly designed for high performance and reliability, without much consideration of energy efficiency. These networks usually have redundant routers and links, over-provisioned link capacity, and multiple paths for load-balancing and protection, which make the networks far from energy efficient. As energy price continues to rise, the increasing network energy consumption becomes a significant portion of the network operational costs. The energy consumption problem in service provider networks is detailed in [I-D.zhang-panet-problem-statement]. Some use cases of reducing network energy consumption are described in [I-D.zhang-panet-use-cases].

While energy consumption has become an important issue, network operators are very cautious about energy conservation solutions due to the concerns about the potential impacts on the network performance and resiliency.

This document presents a set of requirements for building a Power Aware NETWORK (PANET) while meeting operators' requirements on performance and resiliency.

2. Requirements on Network Elements

Today's network elements are mostly designed for high throughput and availability. With the increase of throughput capacity, energy consumption of network element is also rising accordingly. Since most of time the network elements in the network would not work in the full loaded state, if the energy consumption of network elements could be proportional to the carried traffic load, energy conservation could be achieved. Typically after a network element is turned on, the base energy consumption is relatively high, and the energy consumption of the device does not vary a lot from zero load state to full loaded state. While there has been a lot of efforts aiming at making the energy consumption of network device proportional to the load it carries, it is not quite easy for the network elements getting to this stage in the near term.

Thus for near term energy saving, In practical the network elements should meet the following requirements:

- o Network elements should support a set of energy saving modes (e.g. sleeping mode, etc. as defined in IETF EMAN working group). The energy consumption under energy saving modes should be much lower than that under the normal mode.
- o Network elements should support the report of energy consumption and state information.
- o The transition between different energy modes SHOULD not cost a lot of energy, otherwise there will not be no much benefit of transiting between different energy modes.
- o Network elements should support the transition between different energy modes within acceptable time period, e.g. subsecond.
- o Network elements should support some approach of reducing the packet loss during the transition of energy modes.

3. Requirements on the Whole Network

While energy awareness and conservation of individual network element is fundamental, currently there are many limits in reducing the energy consumption at network element level. Besides, different from terminal devices like PC and mobile phones, network elements usually cannot be shut down arbitrarily as this may affect the services carried in the network. Thus mechanisms which could reduce the energy consumption from the whole network point of view should also be considered.

Most of the existing networks are over-provisioned for better service performance and redundancy, which means they are not energy efficient by default. In order to save energy, the entire network should become power aware, then it can make appropriate decisions to save energy when possible. Since in most time the network does not carry the peak traffic volume, which means there is chance for the network to coordinate network elements and create opportunity for some of the network elements to enter energy saving modes. Meanwhile, reducing energy consumption of the network should not undermine the performance of services carried by the network.

For energy conservation of the whole network, the network should meet the following requirements:

- o The network should try to keep all the active network elements with a reasonable utilization rate, network elements with low utilization should be informed to enter energy saving modes. For example, the network elements with utilization lower than specific threshold may be put into low rate mode to reduce energy consumption, or the traffic carried by these network elements may be migrated to other paths such that these network elements could be put into sleeping mode.
- o With energy conservation, the network should retain enough network availability and resiliency against node and link failures. In other words, the redundancy of the network should be kept at a reasonable level, e.g. 2-connected.
- o Energy saving of the network should not induce increase of latency nor induce traffic loss which exceed the tolerance of the services in the network. QoS metrics such as end-to-end delay, loss and jitter should be kept at a desired level.
- o The network should reserve enough spare capacity or be able to react quickly to absorb traffic spikes in order to minimize packet loss due to congestions.
- o The network stability should be preserved. Particularly, traffic oscillation should be avoided.

- o Energy saving should not conflict with other policies (e.g. performance at the highest priority) in the network.

4. Requirements on Network Control Plane

Most of the existing network control protocols do not take energy awareness or efficiency into consideration, and some protocols may not work properly when some of the network elements in the network are in energy saving modes. For example, when a network link is put into sleeping mode, the protocols run on this link may be impacted.

For energy saving of the whole network, control plane should meet the following requirements:

- o Control plane should be able to work properly when some of the network elements are in energy saving mode.
- o Control plane should support the advertisement of energy related information (e.g. current energy saving mode) of network elements in the network.
- o Control plane should be able to coordinate the energy saving operations of network elements to achieve the overall network energy saving.
- o Control plane should be able to maximize the opportunity for network elements to enter the energy saving modes.
- o Control plane should be aware of the network elements in energy saving modes, and should be able to calculate available paths (e.g. which do not traverse the network elements in sleeping mode).
- o Control plane should be able to calculate the path set for all services carried by the network in a way that energy conservation of the whole network is achieved.

Some considerations on control plane when using energy saving mechanism are also specified in [I-D.retana-rtgwg-eacp].

5. Requirements on Management Plane

Management plane would also be necessary for building a power aware network. IETF EMAN working group is working on the requirements [I-D.ietf-eman-requirements] and mechanisms for energy management. Such management requirements include identification of energy-managed devices and their components, monitoring of a series of power states and power properties. It may further includes controlling of the power supply and power states of the managed devices.

6. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

7. Security Considerations

TBD

8. Acknowledgements

TBD

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

9.2. Informative References

[I-D.ietf-eman-requirements]
Quittek, J., Chandramouli, M., Winter, R., Dietz, T., and B. Claise, "Requirements for Energy Management", draft-ietf-eman-requirements-14 (work in progress), May 2013.

[I-D.retana-rtgwg-eacp]
Retana, A., White, R., and M. Paul, "A Framework and Requirements for Energy Aware Control Planes", draft-retana-rtgwg-eacp-01 (work in progress), February 2013.

[I-D.zhang-panet-problem-statement]
Zhang, B., Shi, J., Dong, J., Zhang, M., and M. Boucadair, "Power-Aware Networks (PANET): Problem Statement", draft-zhang-panet-problem-statement-03 (work in progress), October 2013.

[I-D.zhang-panet-use-cases]

Zhang, M., Dong, J., Zhang, B., and B. Khargharia, "Use Cases for Power-Aware Networks", draft-zhang-panet-use-cases-03 (work in progress), October 2013.

Authors' Addresses

Jie Dong
Huawei Technologies
Beijing 100095
China

Email: jie.dong@huawei.com

Mingui Zhang
Huawei Technologies
Beijing 100095
China

Email: zhangmingui@huawei.com

Beichuan Zhang
The University of Arizona
USA

Email: bzhang@cs.arizona.edu

Mohamed Boucadair
France Telecom
France

Email: mohamed.boucadair@orange.com

Routing Area Working Group
Internet-Draft
Intended status: Informational
Expires: April 24, 2014

G. Enyedi, Ed.
A. Csaszar
Ericsson
A. Atlas, Ed.
C. Bowers
Juniper Networks
A. Gopalan
University of Arizona
October 21, 2013

Algorithms for computing Maximally Redundant Trees for IP/LDP Fast-
Reroute
draft-enyedi-rtgwg-mrt-frr-algorithm-04

Abstract

A complete solution for IP and LDP Fast-Reroute using Maximally Redundant Trees is presented in [I-D.ietf-rtgwg-mrt-frr-architecture]. This document defines the associated MRT Lowpoint algorithm that is used in the default MRT profile to compute both the necessary Maximally Redundant Trees with their associated next-hops and the alternates to select for MRT-FRR.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 24, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents

(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Terminology and Definitions	4
3. Algorithm Key Concepts	6
3.1. Partial Ordering for Disjoint Paths	6
3.2. Finding an Ear and the Correct Direction	8
3.3. Low-Point Values and Their Uses	11
3.4. Blocks in a Graph	13
3.5. Determining Local-Root and Assigning Block-ID	15
4. Algorithm Sections	16
4.1. MRT Island Identification	17
4.2. Root Selection	18
4.3. Initialization	18
4.4. MRT Lowpoint Algorithm: Computing GADAG using lowpoint inheritance	19
4.5. Augmenting the GADAG by directing all links	21
4.6. Compute MRT next-hops	23
4.6.1. MRT next-hops to all nodes partially ordered with respect to the computing node	24
4.6.2. MRT next-hops to all nodes not partially ordered with respect to the computing node	24
4.6.3. Computing Redundant Tree next-hops in a 2-connected Graph	25
4.6.4. Generalizing for graph that isn't 2-connected	27
4.6.5. Complete Algorithm to Compute MRT Next-Hops	28
4.7. Identify MRT alternates	30
4.8. Finding FRR Next-Hops for Proxy-Nodes	34
5. MRT Lowpoint Algorithm: Complete Specification	36
6. Algorithm Alternatives and Evaluation	37
6.1. Algorithm Evaluation	37
7. Algorithm Work to Be Done	47
8. IANA Considerations	47
9. Security Considerations	47
10. References	47
10.1. Normative References	47
10.2. Informative References	47
Appendix A. Option 2: Computing GADAG using SPF's	49
Appendix B. Option 3: Computing GADAG using a hybrid method	53
Authors' Addresses	55

1. Introduction

MRT Fast-Reroute requires that packets can be forwarded not only on the shortest-path tree, but also on two Maximally Redundant Trees (MRTs), referred to as the MRT-Blue and the MRT-Red. A router which experiences a local failure must also have pre-determined which alternate to use. This document defines how to compute these three things for use in MRT-FRR and describes the algorithm design decisions and rationale. The algorithm is based on those presented in [MRTLinear] and expanded in [EnyediThesis].

Just as packets routed on a hop-by-hop basis require that each router compute a shortest-path tree which is consistent, it is necessary for each router to compute the MRT-Blue next-hops and MRT-Red next-hops in a consistent fashion. This document defines the MRT Lowpoint algorithm to be used as a standard in the default MRT profile for MRT-FRR.

As now, a router's FIB will contain primary next-hops for the current shortest-path tree for forwarding traffic. In addition, a router's FIB will contain primary next-hops for the MRT-Blue for forwarding received traffic on the MRT-Blue and primary next-hops for the MRT-Red for forwarding received traffic on the MRT-Red.

What alternate next-hops a point-of-local-repair (PLR) selects need not be consistent - but loops must be prevented. To reduce congestion, it is possible for multiple alternate next-hops to be selected; in the context of MRT alternates, each of those alternate next-hops would be equal-cost paths.

This document defines an algorithm for selecting an appropriate MRT alternate for consideration. Other alternates, e.g. LFAs that are downstream paths, may be preferred when available and that policy-based alternate selection process[I-D.ietf-rtgwg-lfa-manageability] is not captured in this document.

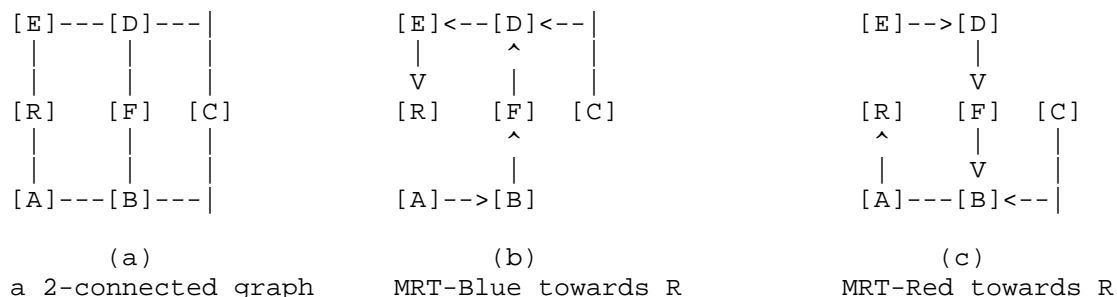
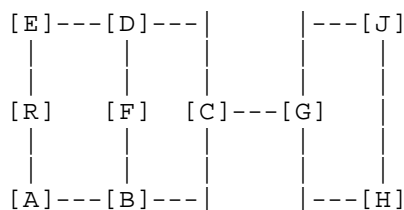
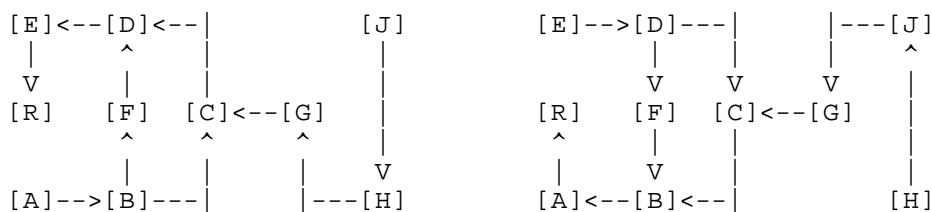


Figure 1

Algorithms for computing MRTs can handle arbitrary network topologies where the whole network graph is not 2-connected, as in Figure 2, as well as the easier case where the network graph is 2-connected (Figure 1). Each MRT is a spanning tree. The pair of MRTs provide two paths from every node X to the root of the MRTs. Those paths share the minimum number of nodes and the minimum number of links. Each such shared node is a cut-vertex. Any shared links are cut-links.



(a) a graph that isn't 2-connected



(b) MRT-Blue towards R

(c) MRT-Red towards R

Figure 2

2. Terminology and Definitions

network graph: A graph that reflects the network topology where all links connect exactly two nodes and broadcast links have been transformed into the standard pseudo-node representation.

Redundant Trees (RT): A pair of trees where the path from any node X to the root R on the first tree is node-disjoint with the path from the same node X to the root along the second tree. These can be computed in 2-connected graphs.

Maximally Redundant Trees (MRT): A pair of trees where the path from any node X to the root R along the first tree and the path from the same node X to the root along the second tree share the minimum number of nodes and the minimum number of links. Each such shared node is a cut-vertex. Any shared links are cut-links. Any RT is an MRT but many MRTs are not RTs.

MRT-Red: MRT-Red is used to describe one of the two MRTs; it is used to describe the associated forwarding topology and MT-ID. Specifically, MRT-Red is the decreasing MRT where links in the GADAG are taken in the direction from a higher topologically ordered node to a lower one.

MRT-Blue: MRT-Blue is used to describe one of the two MRTs; it is used to describe the associated forwarding topology and MT-ID. Specifically, MRT-Blue is the increasing MRT where links in the GADAG are taken in the direction from a lower topologically ordered node to a higher one.

cut-vertex: A vertex whose removal partitions the network.

cut-link: A link whose removal partitions the network. A cut-link by definition must be connected between two cut-vertices. If there are multiple parallel links, then they are referred to as cut-links in this document if removing the set of parallel links would partition the network.

2-connected: A graph that has no cut-vertices. This is a graph that requires two nodes to be removed before the network is partitioned.

spanning tree: A tree containing links that connects all nodes in the network graph.

back-edge: In the context of a spanning tree computed via a depth-first search, a back-edge is a link that connects a descendant of a node *x* with an ancestor of *x*.

2-connected cluster: A maximal set of nodes that are 2-connected. In a network graph with at least one cut-vertex, there will be multiple 2-connected clusters.

block: Either a 2-connected cluster, a cut-edge, or an isolated vertex.

DAG: Directed Acyclic Graph - a digraph containing no directed cycle.

ADAG: Almost Directed Acyclic Graph - a digraph that can be transformed into a DAG with removing a single node (the root node).

GADAG: Generalized ADAG - a digraph, which has only ADAGs as all of its blocks. The root of such a block is the node closest to the global root (e.g. with uniform link costs).

DFS: Depth-First Search

DFS ancestor: A node n is a DFS ancestor of x if n is on the DFS-tree path from the DFS root to x .

DFS descendant: A node n is a DFS descendant of x if x is on the DFS-tree path from the DFS root to n .

ear: A path along not-yet-included-in-the-GADAG nodes that starts at a node that is already-included-in-the-GADAG and that ends at a node that is already-included-in-the-GADAG. The starting and ending nodes may be the same node if it is a cut-vertex.

$X \gg Y$ or $Y \ll X$: Indicates the relationship between X and Y in a partial order, such as found in a GADAG. $X \gg Y$ means that X is higher in the partial order than Y . $Y \ll X$ means that Y is lower in the partial order than X .

$X > Y$ or $Y < X$: Indicates the relationship between X and Y in the total order, such as found via a topological sort. $X > Y$ means that X is higher in the total order than Y . $Y < X$ means that Y is lower in the total order than X .

proxy-node: A node added to the network graph to represent a multi-homed prefix or routers outside the local MRT-fast-reroute-supporting island of routers. The key property of proxy-nodes is that traffic cannot transit them.

3. Algorithm Key Concepts

There are five key concepts that are critical for understanding the MRT Lowpoint algorithm and other algorithms for computing MRTs. The first is the idea of partially ordering the nodes in a network graph with regard to each other and to the GADAG root. The second is the idea of finding an ear of nodes and adding them in the correct direction. The third is the idea of a Low-Point value and how it can be used to identify cut-vertices and to find a second path towards the root. The fourth is the idea that a non-2-connected graph is made up of blocks, where a block is a 2-connected cluster, a cut-edge or an isolated node. The fifth is the idea of a local-root for each node; this is used to compute ADAGs in each block.

3.1. Partial Ordering for Disjoint Paths

Given any two nodes X and Y in a graph, a particular total order means that either $X < Y$ or $X > Y$ in that total order. An example would be a graph where the nodes are ranked based upon their unique IP loopback addresses. In a partial order, there may be some nodes

for which it can't be determined whether $X \ll Y$ or $X \gg Y$. A partial order can be captured in a directed graph, as shown in Figure 3. In a graphical representation, a link directed from X to Y indicates that X is a neighbor of Y in the network graph and $X \ll Y$.

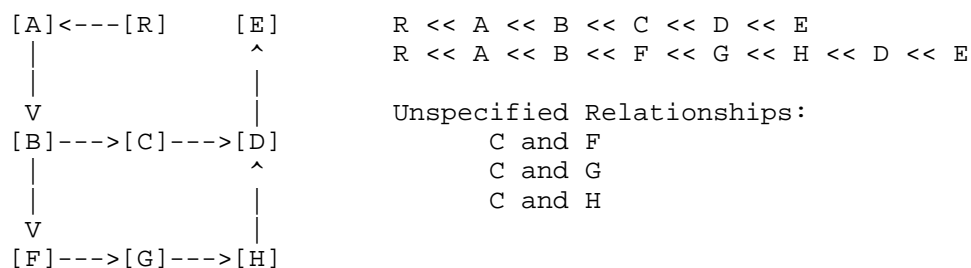


Figure 3: Directed Graph showing a Partial Order

To compute MRTs, the root of the MRTs is at both the very bottom and the very top of the partial ordering. This means that from any node X, one can pick nodes higher in the order until the root is reached. Similarly, from any node X, one can pick nodes lower in the order until the root is reached. For instance, in Figure 4, from G the higher nodes picked can be traced by following the directed links and are H, D, E and R. Similarly, from G the lower nodes picked can be traced by reversing the directed links and are F, B, A, and R. A graph that represents this modified partial order is no longer a DAG; it is termed an Almost DAG (ADAG) because if the links directed to the root were removed, it would be a DAG.

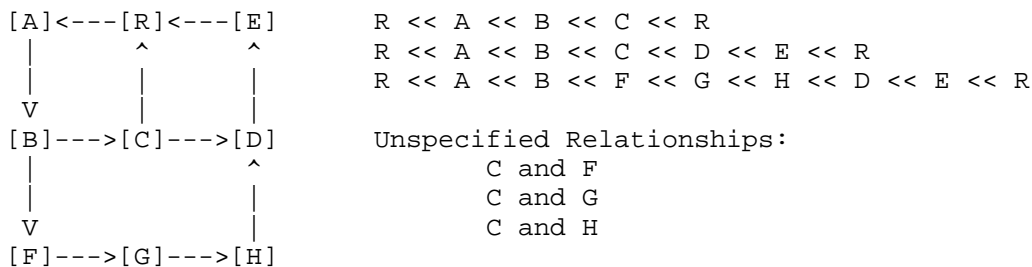


Figure 4: ADAG showing a Partial Order with R lowest and highest

Most importantly, if a node $Y \gg X$, then Y can only appear on the increasing path from X to the root and never on the decreasing path. Similarly, if a node $Z \ll X$, then Z can only appear on the decreasing path from X to the root and never on the increasing path.

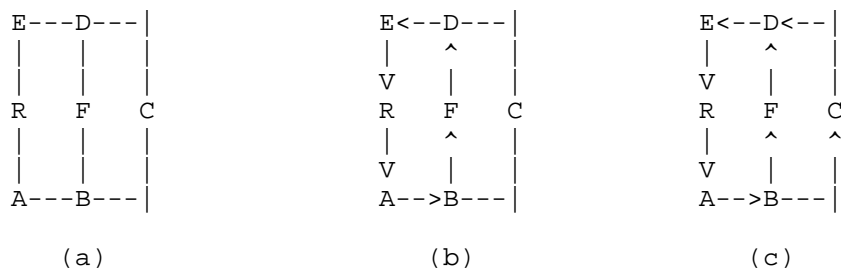
When following the increasing paths, it is possible to pick multiple higher nodes and still have the certainty that those paths will be disjoint from the decreasing paths. E.g. in the previous example node B has multiple possibilities to forward packets along an increasing path: it can either forward packets to C or F .

3.2. Finding an Ear and the Correct Direction

For simplicity, the basic idea of creating a GADAG by adding ears is described assuming that the network graph is a single 2-connected cluster so that an ADAG is sufficient. Generalizing to multiple blocks is done by considering the block-roots instead of the GADAG root - and the actual algorithm is given in Section 4.4.

In order to understand the basic idea of finding an ADAG, first suppose that we have already a partial ADAG, which doesn't contain all the nodes in the block yet, and we want to extend it to cover all the nodes. Suppose that we find a path from a node X to Y such that X and Y are already contained by our partial ADAG, but all the remaining nodes along the path are not added to the ADAG yet. We refer to such a path as an ear.

Recall that our ADAG is closely related to a partial order, more precisely, if we remove root R , the remaining DAG describes a partial order of the nodes. If we suppose that neither X nor Y is the root, we may be able to compare them. If one of them is definitely lesser with respect to our partial order (say $X \ll Y$), we can add the new path to the ADAG in a direction from X to Y . As an example consider Figure 5.



(a) A 2-connected graph
 (b) Partial ADAG (C is not included)
 (c) Resulting ADAG after adding path (or ear) B-C-D

Figure 5

In this partial ADAG, node C is not yet included. However, we can find path B-C-D, where both endpoints are contained by this partial ADAG (we say those nodes are **ready** in the sequel), and the remaining node (node C) is not contained yet. If we remove R, the remaining DAG defines a partial order, and with respect to this partial order we can say that $B \ll D$, so we can add the path to the ADAG in the direction from B to D (arcs B→C and C→D are added). If B were strictly greater than D, we would add the same path in reverse direction.

If in the partial order where an ear's two ends are X and Y, $X \ll Y$, then there must already be a directed path from X to Y already in the ADAG. The ear must be added in a direction such that it doesn't create a cycle; therefore the ear must go from X to Y.

In the case, when X and Y are not ordered with each other, we can select either direction for the ear. We have no restriction since neither of the directions can result in a cycle. In the corner case when one of the endpoints of an ear, say X, is the root (recall that the two endpoints must be different), we could use both directions again for the ear because the root can be considered both as smaller and as greater than Y. However, we strictly pick that direction in which the root is lower than Y. The logic for this decision is explained in Section 4.6

A partial ADAG is started by finding a cycle from the root R back to itself. This can be done by selecting a non-ready neighbor N of R and then finding a path from N to R that doesn't use any links between R and N. The direction of the cycle can be assigned either way since it is starting the ordering.

Once a partial ADAG is already present, we can always add ears to it: just select a non-ready neighbor N of a ready node Q, such that Q is not the root, find a path from N to the root in the graph with Q removed. This path is an ear where the first node of the ear is Q, the next is N, then the path until the first ready node the path reached (that second ready node is the other endpoint of the path). Since the graph is 2-connected, there must be a path from N to R without Q.

It is always possible to select a non-ready neighbor N of a ready node Q so that Q is not the root R . Because the network is 2-connected, N must be connected to two different nodes and only one can be R . Because the initial cycle has already been added to the ADAG, there are ready nodes that are not R . Since the graph is 2-connected, while there are non-ready nodes, there must be a non-ready neighbor N of a ready node that is not R .

```
Generic_Find_Ears_ADAG(root)
  Create an empty ADAG. Add root to the ADAG.
  Mark root as IN_GADAG.
  Select an arbitrary cycle containing root.
  Add the arbitrary cycle to the ADAG.
  Mark cycle's nodes as IN_GADAG.
  Add cycle's non-root nodes to process_list.
  while there exists connected nodes in graph that are not IN_GADAG
    Select a new ear. Let its endpoints be  $X$  and  $Y$ .
    if  $Y$  is root or  $(Y \ll X)$ 
      add the ear towards  $X$  to the ADAG
    else // (a)  $X$  is root or (b)  $X \ll Y$  or (c)  $X, Y$  not ordered
      Add the ear towards  $Y$  to the ADAG
```

Figure 6: Generic Algorithm to find ears and their direction in 2-connected graph

Algorithm Figure 6 merely requires that a cycle or ear be selected without specifying how. Regardless of the way of selecting the path, we will get an ADAG. The method used for finding and selecting the ears is important; shorter ears result in shorter paths along the MRTs. The MRT Lowpoint algorithm's method using Low-Point Inheritance is defined in Section 4.4. Other methods are described in the Appendices (Appendix A and Appendix B).

As an example, consider Figure 5 again. First, we select the shortest cycle containing R , which can be $R-A-B-F-D-E$ (uniform link costs were assumed), so we get to the situation depicted in Figure 5 (b). Finally, we find a node next to a ready node; that must be node C and assume we reached it from ready node B . We search a path from C to R without B in the original graph. The first ready node along this is node D , so the open ear is $B-C-D$. Since $B \ll D$, we add arc $B \rightarrow C$ and $C \rightarrow D$ to the ADAG. Since all the nodes are ready, we stop at this point.

3.3. Low-Point Values and Their Uses

A basic way of computing a spanning tree on a network graph is to run a depth-first-search, such as given in Figure 7. This tree has the important property that if there is a link (x, n) , then either n is a DFS ancestor of x or n is a DFS descendant of x . In other words, either n is on the path from the root to x or x is on the path from the root to n .

```

global_variable: dfs_number

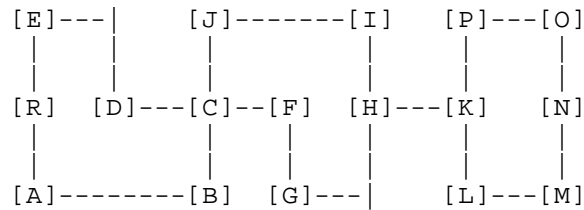
DFS_Visit(node x, node parent)
    D(x) = dfs_number
    dfs_number += 1
    x.dfs_parent = parent
    for each link (x, w)
        if D(w) is not set
            DFS_Visit(w, x)

Run_DFS(node root)
    dfs_number = 0
    DFS_Visit(root, NONE)

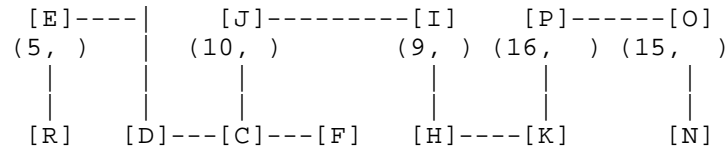
```

Figure 7: Basic Depth-First Search algorithm

Given a node x , one can compute the minimal DFS number of the neighbours of x , i.e. $\min(D(w) \text{ if } (x,w) \text{ is a link})$. This gives the highest attachment point neighbouring x . What is interesting, though, is what is the highest attachment point from x and x 's descendants. This is what is determined by computing the Low-Point value, as given in Algorithm Figure 9 and illustrated on a graph in Figure 8.



(a) a non-2-connected graph



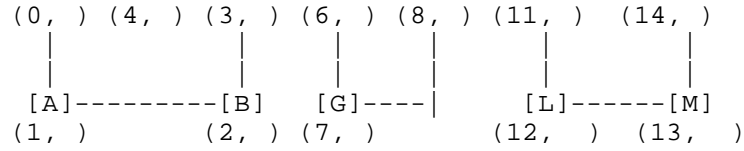
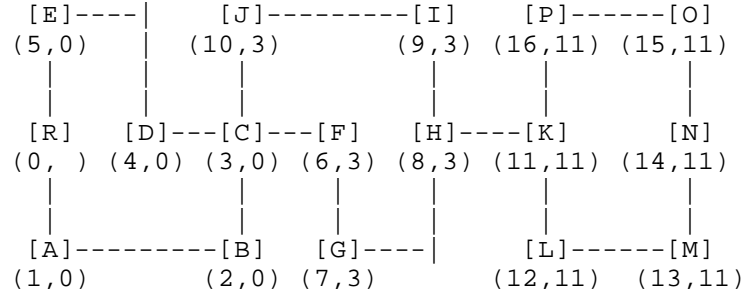
(b) with DFS values assigned $(D(x), L(x))$ (c) with low-point values assigned $(D(x), L(x))$

Figure 8

```
global_variable: dfs_number
```

```
Lowpoint_Visit(node x, node parent, interface p_to_x)
    D(x) = dfs_number
    L(x) = D(x)
    dfs_number += 1
    x.dfs_parent = parent
    x.dfs_parent_intf = p_to_x
    x.lowpoint_parent = NONE
    for each interface intf of x:
        if D(intf.remote_node) is not set
            Lowpoint_Visit(intf.remote_node, x, intf)
        if L(intf.remote_node) < L(x)
            L(x) = L(intf.remote_node)
            x.lowpoint_parent = intf.remote_node
            x.lowpoint_parent_intf = intf
        else if intf.remote_node is not parent
            if D(intf.remote_node) < L(x)
                L(x) = D(intf.remote)
                x.lowpoint_parent = intf.remote_node
                x.lowpoint_parent_intf = intf

Run_Lowpoint(node root)
    dfs_number = 0
```

```
Lowpoint_Visit(root, NONE, NONE)
```

Figure 9: Computing Low-Point value

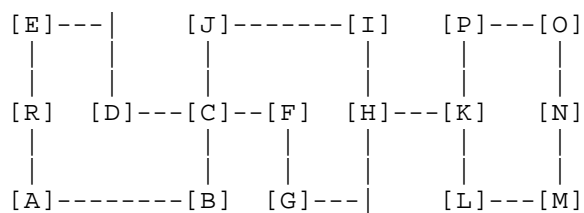
From the low-point value and lowpoint parent, there are two very useful things which motivate our computation.

First, if there is a child c of x such that $L(c) \geq D(x)$, then there are no paths in the network graph that go from c or its descendants to an ancestor of x - and therefore x is a cut-vertex. This is useful because it allows identification of the cut-vertices and thus the blocks. As seen in Figure 8, even if $L(x) < D(x)$, there may be a block that contains both the root and a DFS-child of a node while other DFS-children might be in different blocks. In this example, C 's child D is in the same block as R while F is not.

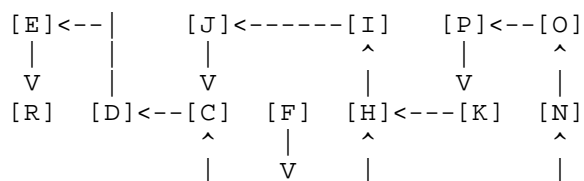
Second, by repeatedly following the path given by `lowpoint_parent`, there is a path from x back to an ancestor of x that does not use the link $[x, x.\text{dfs_parent}]$ in either direction. The full path need not be taken, but this gives a way of finding an initial cycle and then ears.

3.4. Blocks in a Graph

A key idea for an MRT algorithm is that any non-2-connected graph is made up by blocks (e.g. 2-connected clusters, cut-links, and/or isolated nodes). To compute GADAGs and thus MRTs, computation is done in each block to compute ADAGs or Redundant Trees and then those ADAGs or Redundant Trees are combined into a GADAG or MRT.



(a) A graph with four blocks that are:
3 2-connected clusters and a cut-link



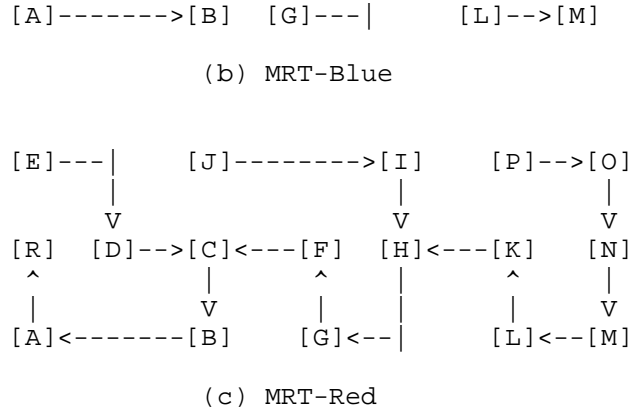


Figure 10

Consider the example depicted in Figure 10 (a). In this figure, a special graph is presented, showing us all the ways 2-connected clusters can be connected. It has four blocks: block 1 contains R, A, B, C, D, E, block 2 contains C, F, G, H, I, J, block 3 contains K, L, M, N, O, P, and block 4 is a cut-edge containing H and K. As can be observed, the first two blocks have one common node (node C) and blocks 2 and 3 do not have any common node, but they are connected through a cut-edge that is block 4. No two blocks can have more than one common node, since two blocks with at least 2 common nodes would qualify as a single 2-connected cluster.

Moreover, observe that if we want to get from one block to another, we must use a cut-vertex (the cut-vertices in this graph are C, H, K), regardless of the path selected, so we can say that all the paths from block 3 along the MRTs rooted at R will cross K first. This observation means that if we want to find a pair of MRTs rooted at R, then we need to build up a pair of RTs in block 3 with K as a root. Similarly, we need to find another one in block 2 with C as a root, and finally, we need the last one in block 1 with R as a root. When all the trees are selected, we can simply combine them; when a block is a cut-edge (as in block 4), that cut-edge is added in the same direction to both of the trees. The resulting trees are depicted in Figure 10 (b) and (c).

Similarly, to create a GADAG it is sufficient to compute ADAGs in each block and connect them.

It is necessary, therefore, to identify the cut-vertices, the blocks and identify the appropriate local-root to use for each block.

3.5. Determining Local-Root and Assigning Block-ID

Each node in a network graph has a local-root, which is the cut-vertex (or root) in the same block that is closest to the root. The local-root is used to determine whether two nodes share a common block.

```

Compute_Localroot(node x, node localroot)
  x.localroot = localroot
  for each DFS child c
    if L(c) < D(x)    //x is not a cut-vertex
      Compute_Localroot(c, x.localroot)
    else
      mark x as cut-vertex
      Compute_Localroot(c, x)

Compute_Localroot(root, root)

```

Figure 11: A method for computing local-roots

There are two different ways of computing the local-root for each node. The stand-alone method is given in Figure 11 and better illustrates the concept; it is used by the MRT algorithms given in the Appendices Appendix A and Appendix B. The method for local-root computation is used in the MRT Lowpoint algorithm for computing a GADAG using Low-Point inheritance and the essence of it is given in Figure 12.

```

Get the current node, s.
Compute an ear(either through lowpoint inheritance
or by following dfs parents) from s to a ready node e.
(Thus, s is not e, if there is such ear.)
if s is e
  for each node x in the ear that is not s
    x.localroot = s
else
  for each node x in the ear that is not s or e
    x.localroot = e.localroot

```

Figure 12: Ear-based method for computing local-roots

Once the local-roots are known, two nodes X and Y are in a common block if and only if one of the following three conditions apply.

- o Y's local-root is X's local-root : They are in the same block and neither is the cut-vertex closest to the root.

- o Y's local-root is X: X is the cut-vertex closest to the root for Y's block
- o Y is X's local-root: Y is the cut-vertex closest to the root for X's block

Once we have computed the local-root for each node in the network graph, we can assign for each node, a block id that represents the block in which the node is present. This computation is shown in Figure 13.

```

global_var: max_block_id

Assign_Block_ID(x, cur_block_id)
  x.block_id = cur_block_id
  foreach DFS child c of x
    if (c.local_root is x)
      max_block_id += 1
      Assign_Block_ID(c, max_block_id)
    else
      Assign_Block_ID(c, cur_block_id)

max_block_id = 0
Assign_Block_ID(root, max_block_id)

```

Figure 13: Assigning block id to identify blocks

4. Algorithm Sections

This algorithm computes one GADAG that is then used by a router to determine its MRT-Blue and MRT-Red next-hops to all destinations. Finally, based upon that information, alternates are selected for each next-hop to each destination. The different parts of this algorithm are described below. These work on a network graph after, for instance, its interfaces are ordered as per Figure 14.

1. Compute the local MRT Island for the particular MRT Profile. [See Section 4.1.]
2. Select the root to use for the GADAG. [See Section 4.2.]
3. Initialize all interfaces to UNDIRECTED. [See Section 4.3.]
4. Compute the DFS value, e.g. $D(x)$, and lowpoint value, $L(x)$. [See Figure 9.]
5. Construct the GADAG. [See Section 4.4]

6. Assign directions to all interfaces that are still `UNDIRECTED`. [See Section 4.5.]
7. From the computing router `x`, compute the next-hops for the MRT-Blue and MRT-Red. [See Section 4.6.]
8. Identify alternates for each next-hop to each destination by determining which one of the blue MRT and the red MRT the computing router `x` should select. [See Section 4.7.]

To ensure consistency in computation, all routers **MUST** order interfaces identically. This is necessary for the DFS, where the selection order of the interfaces to explore results in different trees, and for computing the GADAG, where the selection order of the interfaces to use to form ears can result in different GADAGs. The required ordering between two interfaces from the same router `x` is given in Figure 14.

```
Interface_Compare(interface a, interface b)
  if a.metric < b.metric
    return A_LESS_THAN_B
  if b.metric < a.metric
    return B_LESS_THAN_A
  if a.neighbor.loopback_addr < b.neighbor.loopback_addr
    return A_LESS_THAN_B
  if b.neighbor.loopback_addr < a.neighbor.loopback_addr
    return B_LESS_THAN_A
  // Same metric to same node, so the order doesn't matter anymore.
  // To have a unique, consistent total order,
  // tie-break based on, for example, the link's linkData as
  // distributed in an OSPF Router-LSA
  if a.link_data < b.link_data
    return A_LESS_THAN_B
  return B_LESS_THAN_A
```

Figure 14: Rules for ranking multiple interfaces. Order is from low to high.

4.1. MRT Island Identification

The local MRT Island for a particular MRT profile can be determined by starting from the computing router in the network graph and doing a breadth-first-search (BFS), exploring only links that aren't MRT-ineligible.

```
MRT_Island_Identification(topology, computing_rtr, profile_id)
  for all routers in topology
    rtr.IN_MRT_ISLAND = FALSE
```



```
computing_rtr.IN_MRT_ISLAND = TRUE
explore_list = { computing_rtr }
while (explore_list is not empty)
  next_rtr = remove_head(explore_list)
  for each interface in next_rtr
    if interface is not MRT-ineligible
      if ((interface.remote_node supports profile_id) and
          (interface.remote_node.IN_MRT_ISLAND is FALSE))
        interface.remote_node.IN_MRT_ISLAND = TRUE
        add_to_tail(explore_list, interface.remote_node)
```

Figure 15: MRT Island Identification

4.2. Root Selection

In [I-D.atlas-ospf-mrt], a mechanism is given for routers to advertise the GADAG Root Selection Priority and consistently select a GADAG Root inside the local MRT Island. Before beginning computation, the network graph is reduced to contain only the set of routers that support the specific MRT profile whose MRTs are being computed.

Off-line analysis that considers the centrality of a router may help determine how good a choice a particular router is for the role of GADAG root.

4.3. Initialization

Before running the algorithm, there is the standard type of initialization to be done, such as clearing any computed DFS-values, lowpoint-values, DFS-parents, lowpoint-parents, any MRT-computed next-hops, and flags associated with algorithm.

It is assumed that a regular SPF computation has been run so that the primary next-hops from the computing router to each destination are known. This is required for determining alternates at the last step.

Initially, all interfaces must be initialized to UNDIRECTED. Whether they are OUTGOING, INCOMING or both is determined when the GADAG is constructed and augmented.

It is possible that some links and nodes will be marked as unusable, whether because of configuration, IGP flooding (e.g. MRT-ineligible links in [I-D.atlas-ospf-mrt]), overload, or due to a transient cause such as [RFC3137]. In the algorithm description, it is assumed that such links and nodes will not be explored or used and no more discussion is given of this restriction.

4.4. MRT Lowpoint Algorithm: Computing GADAG using lowpoint inheritance

As discussed in Section 3.2, it is necessary to find ears from a node *x* that is already in the GADAG (known as IN_GADAG). There are two methods to find ears; both are required. The first is by going to a not IN_GADAG DFS-child and then following the chain of low-point parents until an IN_GADAG node is found. The second is by going to a not IN_GADAG neighbor and then following the chain of DFS parents until an IN_GADAG node is found. As an ear is found, the associated interfaces are marked based on the direction taken. The nodes in the ear are marked as IN_GADAG. In the algorithm, first the ears via DFS-children are found and then the ears via DFS-neighbors are found.

By adding both types of ears when an IN_GADAG node is processed, all ears that connect to that node are found. The order in which the IN_GADAG nodes is processed is, of course, key to the algorithm. The order is a stack of ears so the most recent ear is found at the top of the stack. Of course, the stack stores nodes and not ears, so an ordered list of nodes, from the first node in the ear to the last node in the ear, is created as the ear is explored and then that list is pushed onto the stack.

Each ear represents a partial order (see Figure 4) and processing the nodes in order along each ear ensures that all ears connecting to a node are found before a node higher in the partial order has its ears explored. This means that the direction of the links in the ear is always from the node *x* being processed towards the other end of the ear. Additionally, by using a stack of ears, this means that any unprocessed nodes in previous ears can only be ordered higher than nodes in the ears below it on the stack.

In this algorithm that depends upon Low-Point inheritance, it is necessary that every node have a low-point parent that is not itself. If a node is a cut-vertex, that may not yet be the case. Therefore, any nodes without a low-point parent will have their low-point parent set to their DFS parent and their low-point value set to the DFS-value of their parent. This assignment also properly allows an ear between two cut-vertices.

Finally, the algorithm simultaneously computes each node's local-root, as described in Figure 12. This is further elaborated as follows. The local-root can be inherited from the node at the end of the ear unless the end of the ear is *x* itself, in which case the local-root for all the nodes in the ear would be *x*. This is because whenever the first cycle is found in a block, or an ear involving a bridge is computed, the cut-vertex closest to the root would be *x* itself. In all other scenarios, the properties of lowpoint/dfs parents ensure that the end of the ear will be in the same block, and

thus inheriting its local-root would be the correct local-root for all newly added nodes.

The pseudo-code for the GADAG algorithm (assuming that the adjustment of lowpoint for cut-vertices has been made) is shown in Figure 16.

```

Construct_Ear(x, Stack, intf, type)
    ear_list = empty
    cur_node = intf.remote_node
    cur_intf = intf
    not_done = true

    while not_done
        cur_intf.UNDIRECTED = false
        cur_intf.OUTGOING = true
        cur_intf.remote_intf.UNDIRECTED = false
        cur_intf.remote_intf.INCOMING = true

        if cur_node.IN_GADAG is false
            cur_node.IN_GADAG = true
            add_to_list_end(ear_list, cur_node)
            if type is CHILD
                cur_intf = cur_node.lowpoint_parent_intf
                cur_node = cur_node.lowpoint_parent
            else type must be NEIGHBOR
                cur_intf = cur_node.dfs_parent_intf
                cur_node = cur_node.dfs_parent
        else
            not_done = false

    if (type is CHILD) and (cur_node is x)
        //x is a cut-vertex and the local root for
        //the block in which the ear is computed
        localroot = x
    else
        // Inherit local-root from the end of the ear
        localroot = cur_node.localroot
    while ear_list is not empty
        y = remove_end_item_from_list(ear_list)
        y.localroot = localroot
        push(Stack, y)

Construct_GADAG_via_Lowpoint(topology, root)
    root.IN_GADAG = true
    root.localroot = root
    Initialize Stack to empty
    push root onto Stack
    while (Stack is not empty)

```

```

x = pop(Stack)
foreach interface intf of x
  if ((intf.remote_node.IN_GADAG == false) and
      (intf.remote_node.dfs_parent is x))
    Construct_Ear(x, Stack, intf, CHILD)
foreach interface intf of x
  if ((intf.remote_node.IN_GADAG == false) and
      (intf.remote_node.dfs_parent is not x))
    Construct_Ear(x, Stack, intf, NEIGHBOR)

Construct_GADAG_via_Lowpoint(topology, root)

```

Figure 16: Low-point Inheritance GADAG algorithm

4.5. Augmenting the GADAG by directing all links

The GADAG, regardless of the algorithm used to construct it, at this point could be used to find MRTs but the topology does not include all links in the network graph. That has two impacts. First, there might be shorter paths that respect the GADAG partial ordering and so the alternate paths would not be as short as possible. Second, there may be additional paths between a router *x* and the root that are not included in the GADAG. Including those provides potentially more bandwidth to traffic flowing on the alternates and may reduce congestion compared to just using the GADAG as currently constructed.

The goal is thus to assign direction to every remaining link marked as `UNDIRECTED` to improve the paths and number of paths found when the MRTs are computed.

To do this, we need to establish a total order that respects the partial order described by the GADAG. This can be done using Kahn's topological sort [[Kahn_1962_topo_sort](#)] which essentially assigns a number to a node *x* only after all nodes before it (e.g. with a link incoming to *x*) have had their numbers assigned. The only issue with the topological sort is that it works on DAGs and not ADAGs or GADAGs.

To convert a GADAG to a DAG, it is necessary to remove all links that point to a root of block from within that block. That provides the necessary conversion to a DAG and then a topological sort can be done. Finally, all `UNDIRECTED` links are assigned a direction based upon the partial ordering. Any `UNDIRECTED` links that connect to a root of a block from within that block are assigned a direction `INCOMING` to that root. The exact details of this whole process are captured in Figure 17

```

Set_Block_Root_Incoming_Links(topo, root, mark_or_clear)
  foreach node x in topo
    if node x is a cut-vertex or root
      foreach interface i of x
        if (i.remote_node.localroot is x)
          if i.UNDIRECTED
            i.OUTGOING = true
            i.remote_intf.INCOMING = true
            i.UNDIRECTED = false
            i.remote_intf.UNDIRECTED = false
          if i.INCOMING
            if mark_or_clear is mark
              if i.OUTGOING // a cut-edge
                i.STORE_INCOMING = true
                i.INCOMING = false
                i.remote_intf.STORE_OUTGOING = true
                i.remote_intf.OUTGOING = false
                i.TEMP_UNUSABLE = true
                i.remote_intf.TEMP_UNUSABLE = true
              else
                i.TEMP_UNUSABLE = false
                i.remote_intf.TEMP_UNUSABLE = false
            if i.STORE_INCOMING and (mark_or_clear is clear)
              i.INCOMING = true
              i.STORE_INCOMING = false
              i.remote_intf.OUTGOING = true
              i.remote_intf.STORE_OUTGOING = false

Run_Topological_Sort_GADAG(topo, root)
  Set_Block_Root_Incoming_Links(topo, root, MARK)
  foreach node x
    set x.unvisited to the count of x's incoming interfaces
    that aren't marked TEMP_UNUSABLE
  Initialize working_list to empty
  Initialize topo_order_list to empty
  add_to_list_end(working_list, root)
  while working_list is not empty
    y = remove_start_item_from_list(working_list)
    add_to_list_end(topo_order_list, y)
    foreach interface i of y
      if (i.OUTGOING) and (not i.TEMP_UNUSABLE)
        i.remote_node.unvisited -= 1
        if i.remote_node.unvisited is 0
          add_to_list_end(working_list, i.remote_node)
  next_topo_order = 1
  while topo_order_list is not empty
    y = remove_start_item_from_list(topo_order_list)
    y.topo_order = next_topo_order

```

```

        next_topo_order += 1
    Set_Block_Root_Incoming_Links(topo, root, CLEAR)

Add_Undirected_Links(topo, root)
Run_Topological_Sort_GADAG(topo, root)
foreach node x in topo
    foreach interface i of x
        if i.UNDIRECTED
            if x.topo_order < i.remote_node.topo_order
                i.OUTGOING = true
                i.UNDIRECTED = false
                i.remote_intf.INCOMING = true
                i.remote_intf.UNDIRECTED = false
            else
                i.INCOMING = true
                i.UNDIRECTED = false
                i.remote_intf.OUTGOING = true
                i.remote_intf.UNDIRECTED = false

Add_Undirected_Links(topo, root)

```

Figure 17: Assigning direction to UNDIRECTED links

Proxy-nodes do not need to be added to the network graph. They cannot be transited and do not affect the MRTs that are computed. The details of how the MRT-Blue and MRT-Red next-hops are computed and how the appropriate alternate next-hops are selected is given in Section 4.8.

4.6. Compute MRT next-hops

As was discussed in Section 3.1, once a ADAG is found, it is straightforward to find the next-hops from any node X to the ADAG root. However, in this algorithm, we want to reuse the common GADAG and find not only the one pair of MRTs rooted at the GADAG root with it, but find a pair rooted at each node. This is useful since it is significantly faster to compute. It may also provide easier troubleshooting of the MRT-Red and MRT-Blue.

The method for computing differently rooted MRTs from the common GADAG is based on two ideas. First, if two nodes X and Y are ordered with respect to each other in the partial order, then an SPF along OUTGOING links (an increasing-SPF) and an SPF along INCOMING links (a decreasing-SPF) can be used to find the increasing and decreasing paths. Second, if two nodes X and Y aren't ordered with respect to each other in the partial order, then intermediary nodes can be used to create the paths by increasing/decreasing to the intermediary and then decreasing/increasing to reach Y.

As usual, the two basic ideas will be discussed assuming the network is two-connected. The generalization to multiple blocks is discussed in Section 4.6.4. The full algorithm is given in Section 4.6.5.

4.6.1. MRT next-hops to all nodes partially ordered with respect to the computing node

To find two node-disjoint paths from the computing router X to any node Y, depends upon whether $Y \gg X$ or $Y \ll X$. As shown in Figure 18, if $Y \gg X$, then there is an increasing path that goes from X to Y without crossing R; this contains nodes in the interval $[X, Y]$. There is also a decreasing path that decreases towards R and then decreases from R to Y; this contains nodes in the interval $[X, R\text{-small}]$ or $[R\text{-great}, Y]$. The two paths cannot have common nodes other than X and Y.

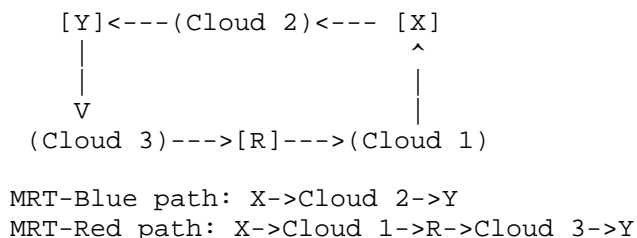


Figure 18: $Y \gg X$

Similar logic applies if $Y \ll X$, as shown in Figure 19. In this case, the increasing path from X increases to R and then increases from R to Y to use nodes in the intervals $[X, R\text{-great}]$ and $[R\text{-small}, Y]$. The decreasing path from X reaches Y without crossing R and uses nodes in the interval $[Y, X]$.

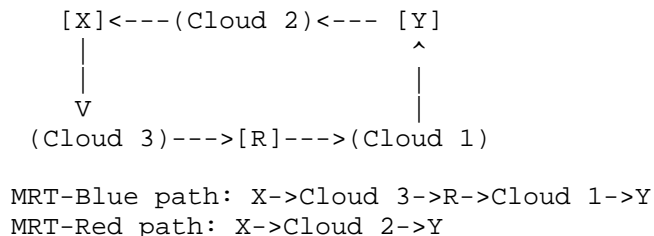
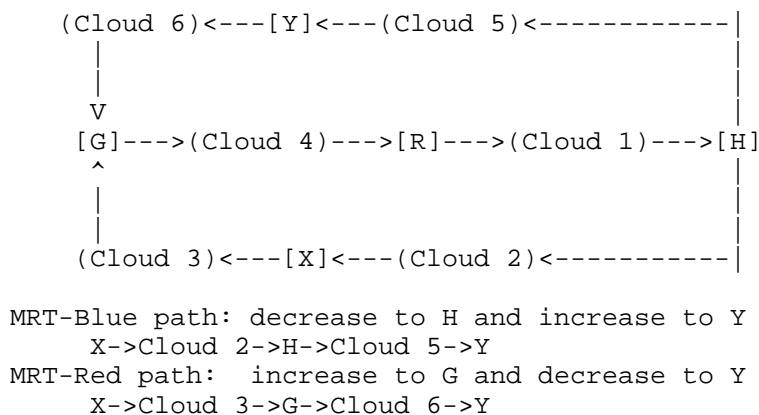


Figure 19: $Y \ll X$

4.6.2. MRT next-hops to all nodes not partially ordered with respect to the computing node

When X and Y are not ordered, the first path should increase until we get to a node G , where $G \gg Y$. At G , we need to decrease to Y . The other path should be just the opposite: we must decrease until we get to a node H , where $H \ll Y$, and then increase. Since R is smaller and greater than Y , such G and H must exist. It is also easy to see that these two paths must be node disjoint: the first path contains nodes in interval $[X, G]$ and $[Y, G]$, while the second path contains nodes in interval $[H, X]$ and $[H, Y]$. This is illustrated in Figure 20. It is necessary to decrease and then increase for the MRT-Blue and increase and then decrease for the MRT-Red; if one simply increased for one and decreased for the other, then both paths would go through the root R .

Figure 20: X and Y unordered

This gives disjoint paths as long as G and H are not the same node. Since $G \gg Y$ and $H \ll Y$, if G and H could be the same node, that would have to be the root R . This is not possible because there is only one incoming interface to the root R which is created when the initial cycle is found. Recall from Figure 6 that whenever an ear was found to have an end that was the root R , the ear was directed from R so that the associated interface on R is outgoing and not incoming. Therefore, there must be exactly one node M which is the largest one before R , so the MRT-Red path will never reach R ; it will turn at M and decrease to Y .

4.6.3. Computing Redundant Tree next-hops in a 2-connected Graph

The basic ideas for computing RT next-hops in a 2-connected graph were given in Section 4.6.1 and Section 4.6.2. Given these two ideas, how can we find the trees?

If some node X only wants to find the next-hops (which is usually the case for IP networks), it is enough to find which nodes are greater and less than X, and which are not ordered; this can be done by running an increasing-SPF and a decreasing-SPF rooted at X and not exploring any links from the ADAG root. (Traversal algorithms other than SPF could safely be used instead where one traversal takes the links in their given directions and the other reverses the links' directions.)

An increasing-SPF rooted at X and not exploring links from the root will find the increasing next-hops to all $Y \gg X$. Those increasing next-hops are X's next-hops on the MRT-Blue to reach Y. An decreasing-SPF rooted at X and not exploring links from the root will find the decreasing next-hops to all $Z \ll X$. Those decreasing next-hops are X's next-hops on the MRT-Red to reach Z. Since the root R is both greater than and less than X, after this increasing-SPF and decreasing-SPF, X's next-hops on the MRT-Blue and on the MRT-Red to reach R are known. For every node $Y \gg X$, X's next-hops on the MRT-Red to reach Y are set to those on the MRT-Red to reach R. For every node $Z \ll X$, X's next-hops on the MRT-Blue to reach Z are set to those on the MRT-Blue to reach R.

For those nodes, which were not reached, we have the next-hops as well. The increasing MRT-Blue next-hop for a node, which is not ordered, is the next-hop along the decreasing MRT-Red towards R and the decreasing MRT-Red next-hop is the next-hop along the increasing MRT-Blue towards R. Naturally, since R is ordered with respect to all the nodes, there will always be an increasing and a decreasing path towards it. This algorithm does not provide the complete specific path taken but just the appropriate next-hops to use. The identity of G and H is not determined.

The final case to considered is when the root R computes its own next-hops. Since the root R is \ll all other nodes, running an increasing-SPF rooted at R will reach all other nodes; the MRT-Blue next-hops are those found with this increasing-SPF. Similarly, since the root R is \gg all other nodes, running a decreasing-SPF rooted at R will reach all other nodes; the MRT-Red next-hops are those found with this decreasing-SPF.



(a) (b)
A 2-connected graph A spanning ADAG rooted at R

Figure 21

As an example consider the situation depicted in Figure 21. There node C runs an increasing-SPF and a decreasing-SPF. The increasing-SPF reaches D, E and R and the decreasing-SPF reaches B, A and R. So towards E the increasing next-hop is D (it was reached through D), and the decreasing next-hop is B (since R was reached through B). Since both D and B, A and R will compute the next hops similarly, the packets will reach E.

We have the next-hops towards F as well: since F is not ordered with respect to C, the MRT-Blue next-hop is the decreasing one towards R (which is B) and the MRT-Red next-hop is the increasing one towards R (which is D). Since B is ordered with F, it will find, for its MRT-Blue, a real increasing next-hop, so packet forwarded to B will get to F on path C-B-F. Similarly, D will have, for its MRT-Red, a real decreasing next-hop, and the packet will use path C-D-F.

4.6.4. Generalizing for graph that isn't 2-connected

If a graph isn't 2-connected, then the basic approach given in Section 4.6.3 needs some extensions to determine the appropriate MRT next-hops to use for destinations outside the computing router X's blocks. In order to find a pair of maximally redundant trees in that graph we need to find a pair of RTs in each of the blocks (the root of these trees will be discussed later), and combine them.

When computing the MRT next-hops from a router X, there are three basic differences:

1. Only nodes in a common block with X should be explored in the increasing-SPF and decreasing-SPF.
2. Instead of using the GADAG root, X's local-root should be used. This has the following implications:
 - a. The links from X's local-root should not be explored.
 - b. If a node is explored in the outgoing SPF so $Y \gg X$, then X's MRT-Red next-hops to reach Y uses X's MRT-Red next-hops to reach X's local-root and if $Z \ll X$, then X's MRT-Blue next-hops to reach Z uses X's MRT-Blue next-hops to reach X's local-root.

- c. If a node W in a common block with X was not reached in the increasing-SPF or decreasing-SPF, then W is unordered with respect to X. X's MRT-Blue next-hops to W are X's decreasing aka MRT-Red next-hops to X's local-root. X's MRT-Red next-hops to W are X's increasing aka Blue MRT next-hops to X's local-root.
- 3. For nodes in different blocks, the next-hops must be inherited via the relevant cut-vertex.

These are all captured in the detailed algorithm given in Section 4.6.5.

4.6.5. Complete Algorithm to Compute MRT Next-Hops

The complete algorithm to compute MRT Next-Hops for a particular router X is given in Figure 22. In addition to computing the MRT-Blue next-hops and MRT-Red next-hops used by X to reach each node Y, the algorithm also stores an "order_proxy", which is the proper cut-vertex to reach Y if it is outside the block, and which is used later in deciding whether the MRT-Blue or the MRT-Red can provide an acceptable alternate for a particular primary next-hop.

```

In_Common_Block(x, y)
    if (((x.localroot is y.localroot) and (x.block_id is y.block_id))
        or (x is y.localroot) or (y is x.localroot))
        return true
    return false

Store_Results(y, direction, spf_root, store_nhs)
    if direction is FORWARD
        y.higher = true
        if store_nhs
            y.blue_next_hops = y.next_hops
    if direction is REVERSE
        y.lower = true
        if store_nhs
            y.red_next_hops = y.next_hops

SPF_No_Traverse_Root(spf_root, block_root, direction, store_nhs)
    Initialize spf_heap to empty
    Initialize nodes' spf_metric to infinity and next_hops to empty
    spf_root.spf_metric = 0
    insert(spf_heap, spf_root)
    while (spf_heap is not empty)
        min_node = remove_lowest(spf_heap)
        Store_Results(min_node, direction, spf_root, store_nhs)
        if ((min_node is spf_root) or (min_node is not block_root))

```

```
    foreach interface intf of min_node
        if (((direction is FORWARD) and intf.OUTGOING) or
            ((direction is REVERSE) and intf.INCOMING) and
            In_Common_Block(spf_root, intf.remote_node))
            path_metric = min_node.spf_metric + intf.metric
            if path_metric < intf.remote_node.spf_metric
                intf.remote_node.spf_metric = path_metric
                if min_node is spf_root
                    intf.remote_node.next_hops = make_list(intf)
                else
                    intf.remote_node.next_hops = min_node.next_hops
                    insert_or_update(spf_heap, intf.remote_node)
            else if path_metric is intf.remote_node.spf_metric
                if min_node is spf_root
                    add_to_list(intf.remote_node.next_hops, intf)
                else
                    add_list_to_list(intf.remote_node.next_hops,
                                    min_node.next_hops)

SetEdge(y)
    if y.blue_next_hops is empty and y.red_next_hops is empty
        if (y.local_root != y) {
            SetEdge(y.localroot)
        }
        y.blue_next_hops = y.localroot.blue_next_hops
        y.red_next_hops = y.localroot.red_next_hops
        y.order_proxy = y.localroot.order_proxy

Compute_MRT_NextHops(x, root)
    foreach node y
        y.higher = y.lower = false
        clear y.red_next_hops and y.blue_next_hops
        y.order_proxy = y
        SPF_No_Traverse_Root(x, x.localroot, FORWARD, TRUE)
        SPF_No_Traverse_Root(x, x.localroot, REVERSE, TRUE)

    // red and blue next-hops are stored to x.localroot as different
    // paths are found via the SPF and reverse-SPF.
    // Similarly any nodes whose local-root is x will have their
    // red_next_hops and blue_next_hops already set.

    // Handle nodes in the same block that aren't the local-root
    foreach node y
        if (y.IN_MRT_ISLAND and (y is not x) and
            (y.localroot is x.localroot) and
            ((y is x.localroot) or (x is y.localroot) or
             (y.block_id is x.block_id)))
            if y.higher
```

```

        y.red_next_hops = x.localroot.red_next_hops
    else if y.lower
        y.blue_next_hops = x.localroot.blue_next_hops
    else
        y.blue_next_hops = x.localroot.red_next_hops
        y.red_next_hops = x.localroot.blue_next_hops

    // Inherit next-hops and order_proxies to other components
    if x is not root
        root.blue_next_hops = x.localroot.blue_next_hops
        root.red_next_hops = x.localroot.red_next_hops
        root.order_proxy = x.localroot
    foreach node y
        if (y is not root) and (y is not x) and y.IN_MRT_ISLAND
            SetEdge(y)

max_block_id = 0
Assign_Block_ID(root, max_block_id)
Compute_MRT_NextHops(x, root)

```

Figure 22

4.7. Identify MRT alternates

At this point, a computing router S knows its MRT-Blue next-hops and MRT-Red next-hops for each destination in the MRT Island. The primary next-hops along the SPT are also known. It remains to determine for each primary next-hop to a destination D, which of the MRTs avoids the primary next-hop node F. This computation depends upon data set in Compute_MRT_NextHops such as each node y's y.blue_next_hops, y.red_next_hops, y.order_proxy, y.higher, y.lower and topo_orders. Recall that any router knows only which are the nodes greater and lesser than itself, but it cannot decide the relation between any two given nodes easily; that is why we need topological ordering.

For each primary next-hop node F to each destination D, S can call Select_Alternates(S, D, F, primary_intf) to determine whether to use the MRT-Blue next-hops as the alternate next-hop(s) for that primary next hop or to use the MRT-Red next-hops. The algorithm is given in Figure 23 and discussed afterwards.

```

Select_Alternates_Internal(S, D, F, primary_intf,
                           D_lower, D_higher, D_topo_order)

    //When D==F, we can do only link protection
    if ((D is F) or (D.order_proxy is F))

```

```
    if an MRT doesn't use primary_intf
        indicate alternate is not node-protecting
        return that MRT color
    else // parallel links are cut-edge
        return AVOID_LINK_ON_BLUE

if (D_lower and D_higher and F_lower and F_higher)
    if F_topo_order < D_topo_order
        return USE_RED
    else
        return USE_BLUE

if (D_lower and D_higher)
    if F_higher
        return USE_RED
    else
        return USE_BLUE

if (F_lower and F_higher)
    if D_lower
        return USE_RED
    else if D_higher
        return USE_BLUE
    else
        if primary_intf.OUTGOING and primary_intf.INCOMING
            return AVOID_LINK_ON_BLUE
        if primary_intf.OUTGOING is true
            return USE_BLUE
        if primary_intf.INCOMING is true
            return USE_RED

if D_higher
    if F_higher
        if F_topo_order < D_topo_order
            return USE_RED
        else
            return USE_BLUE
    else if F_lower
        return USE_BLUE
    else
        // F and S are neighbors so either F << S or F >> S
else if D_lower
    if F_higher
        return USE_RED
    else if F_lower
        if F_topo_order < D_topo_order
            return USE_RED
        else
```

```

        return USE_BLUE
    else
        // F and S are neighbors so either F << S or F >> S
    else // D and S not ordered
        if F.lower
            return USE_RED
        else if F.higher
            return USE_BLUE
        else
            // F and S are neighbors so either F << S or F >> S

Select_Alternates(S, D, F, primary_intf)
    if D.order_proxy is not D
        D_lower = D.order_proxy.lower
        D_higher = D.order_proxy.higher
        D_topo_order = D.order_proxy.topo_order
    else
        D_lower = D.lower
        D_higher = D.higher
        D_topo_order = D.topo_order
    return Select_Alternates_Internal(S, D, F, primary_intf,
                                     D_lower, D_higher, D_topo_order)

```

Figure 23

If either $D \gg S \gg F$ or $D \ll S \ll F$ holds true, the situation is simple: in the first case we should choose the increasing Blue next-hop, in the second case, the decreasing Red next-hop is the right choice.

However, when both D and F are greater than S the situation is not so simple, there can be three possibilities: (i) $F \gg D$ (ii) $F \ll D$ or (iii) F and D are not ordered. In the first case, we should choose the path towards D along the Blue tree. In contrast, in case (ii) the Red path towards the root and then to D would be the solution. Finally, in case (iii) both paths would be acceptable. However, observe that if e.g. $F.topo_order > D.topo_order$, either case (i) or case (iii) holds true, which means that selecting the Blue next-hop is safe. Similarly, if $F.topo_order < D.topo_order$, we should select the Red next-hop. The situation is almost the same if both F and D are less than S.

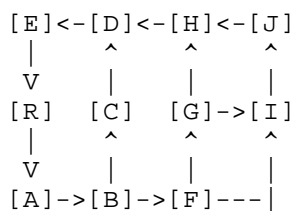
Recall that we have added each link to the GADAG in some direction, so it is impossible that S and F are not ordered. But it is possible that S and D are not ordered, so we need to deal with this case as well. If $F < S$, we can use the Red next-hop, because that path is first increasing until a node definitely greater than D is reached, then decreasing; this path must avoid using F. Similarly, if $F > S$, we should use the Blue next-hop.

Additionally, the cases where either F or D is ordered both higher and lower must be considered; this can happen when one is a block-root or its order_proxy is. If D is both higher and lower than S, then the MRT to use is the one that avoids F so if F is higher, then the MRT-Red should be used and if F is lower, then the MRT-Blue should be used; F and S must be ordered because they are neighbors. If F is both higher and lower, then if D is lower, using the MRT-Red to decrease reaches D and if D is higher, using the Blue MRT to increase reaches D; if D is unordered compared to S, then the situation is a bit more complicated.

In the case where $F < S < F$ and D and S are unordered, the direction of the link in the GADAG between S and F should be examined. If the link is directed $S \rightarrow F$, then use the MRT-Blue (decrease to avoid that link and then increase). If the link is directed $S \leftarrow F$, then use the MRT-Red (increase to avoid that link and then decrease). If the link is $S \leftrightarrow F$, then the link must be a cut-link and there is no node-protecting alternate. If there are multiple links between S and F, then they can protect against each other; of course, in this situation, they are probably already ECMP.

Finally, there is the case where D is also F. In this case, only link protection is possible. The MRT that doesn't use the indicated primary next-hop is used. If both MRTs use the primary next-hop, then the primary next-hop must be a cut-edge so either MRT could be used but the set of MRT next-hops must be pruned to avoid that primary next-hop. To indicate this case, `Select_Alternates` returns `AVOID_LINK_ON_BLUE`.

As an example, consider the ADAG depicted in Figure 24 and first suppose that G is the source, D is the destination and H is the failed next-hop. Since $D > G$, we need to compare `H.topo_order` and `D.topo_order`. Since $D.topo_order > H.topo_order$, D must be not smaller than H, so we should select the decreasing path towards the root. If, however, the destination were instead J, we must find that $H.topo_order > J.topo_order$, so we must choose the increasing Blue next-hop to J, which is I. In the case, when instead the destination is C, we find that we need to first decrease to avoid using H, so the Blue, first decreasing then increasing, path is selected.



(a)

a 2-connected graph

Figure 24

4.8. Finding FRR Next-Hops for Proxy-Nodes

As discussed in Section 10.2 of [I-D.ietf-rtgwg-mrt-frr-architecture], it is necessary to find MRT-Blue and MRT-Red next-hops and MRT-FRR alternates for a named proxy-nodes. An example case is for a router that is not part of that local MRT Island, when there is only partial MRT support in the domain.

A first incorrect and naive approach to handling proxy-nodes, which cannot be transited, is to simply add these proxy-nodes to the graph of the network and connect it to the routers through which the new proxy-node can be reached. Unfortunately, this can introduce some new ordering between the border routers connected to the new node which could result in routing MRT paths through the proxy-node. Thus, this naive approach would need to recompute GADAGs and redo SPTs for each proxy-node.

Instead of adding the proxy-node to the original network graph, each individual proxy-node can be individually added to the GADAG. The proxy-node is connected to at most two nodes in the GADAG. Section 10.2 of [I-D.ietf-rtgwg-mrt-frr-architecture] defines how the proxy-node attachments MUST be determined. The degenerate case where the proxy-node is attached to only one node in the GADAG is trivial as all needed information can be derived from that attachment node; if there are different interfaces, then some can be assigned to MRT-Red and others to MRT-Blue.

Now, consider the proxy-node that is attached to exactly two nodes in the GADAG. Let the `order_proxies` of these nodes be A and B. Let the current node, where next-hop is just being calculated, be S. If one of these two nodes A and B is the local root of S, let `A=S.local_root` and the other one be B. Otherwise, let `A.topo_order < B.topo_order`.

A valid GADAG was constructed. Instead doing an increasing-SPF and a decreasing-SPF to find ordering for the proxy-nodes, the following simple rules, providing the same result, can be used independently for each different proxy-node. For the following rules, let $X=A.local_root$, and if A is the local root, let that be strictly lower than any other node. Always take the first rule that matches.

Rule	Condition	Blue NH	Red NH	Notes
1	$S=X$	Blue to A	Red to B	
2	$S<<A$	Blue to A	Red to R	
3	$S>>B$	Blue to R	Red to B	
4	$A<<S<<B$	Red to A	Blue to B	
5	$A<<S$	Red to A	Blue to R	S not ordered w/ B
6	$S<<B$	Red to R	Blue to B	S not ordered w/ A
7	Otherwise	Red to R	Blue to R	S not ordered w/ A+B

These rules are realized in the following pseudocode where P is the proxy-node, X and Y are the nodes that P is attached to, and S is the computing router:

```

Select_Proxy_Node_NHs(P, S, X, Y)
  if (X.order_proxy.topo_order < Y.order_proxy.topo_order)
    //This fits even if X.order_proxy=S.local_root
    A=X.order_proxy
    B=Y.order_proxy
  else
    A=Y.order_proxy
    B=X.order_proxy

  if (S==A.local_root)
    P.blue_next_hops = A.blue_next_hops
    P.red_next_hops  = B.red_next_hops
    return
  if (A.higher)
    P.blue_next_hops = A.blue_next_hops
    P.red_next_hops  = R.red_next_hops
    return
  if (B.lower)
    P.blue_next_hops = R.blue_next_hops
    P.red_next_hops  = B.red_next_hops
    return
  if (A.lower && B.higher)
    P.blue_next_hops = A.red_next_hops
    P.red_next_hops  = B.blue_next_hops
    return
  if (A.lower)

```

```

        P.blue_next_hops = R.red_next_hops
        P.red_next_hops  = B.blue_next_hops
        return
    if (B.higher)
        P.blue_next_hops = A.red_next_hops
        P.red_next_hops  = R.blue_next_hops
        return
    P.blue_next_hops = R.red_next_hops
    P.red_next_hops  = R.blue_next_hops
    return

```

After finding the the red and the blue next-hops, it is necessary to know which one of these to use in the case of failure. This can be done by `Select_Alternates_Inner()`. In order to use `Select_Alternates_Internal()`, we need to know if P is greater, less or unordered with S, and P.topo_order. P.lower = B.lower, P.higher = A.higher, and any value is OK for P.topo_order, until A.topo_order<=P.topo_order<=B.topo_order and P.topo_order is not equal to the topo_order of the failed node. So for simplicity let P.topo_order=A.topo_order when the next-hop is not A, and P.topo_order=B.topo_order otherwise. This gives the following pseudo-code:

```

Select_Alternates_Proxy_Node(S, P, F, primary_intf)
    if (F is not P.neighbor_A)
        return Select_Alternates_Internal(S, P, F, primary_intf,
                                           P.neighbor_B.lower,
                                           P.neighbor_A.higher,
                                           P.neighbor_A.topo_order)
    else
        return Select_Alternates_Internal(S, P, F, primary_intf,
                                           P.neighbor_B.lower,
                                           P.neighbor_A.higher,
                                           P.neighbor_B.topo_order)

```

Figure 25

5. MRT Lowpoint Algorithm: Complete Specification

This specification defines the MRT Lowpoint Algorithm, which include the construction of a common GADAG and the computation of MRT-Red and MRT-Blue next-hops to each node in the graph. An implementation MAY select any subset of next-hops for MRT-Red and MRT-Blue that respect the available nodes that are described in Section 4.6 for each of the MRT-Red and MRT-Blue and the selected next-hops are further along in the interval of allowed nodes towards the destination.

For example, the MRT-Blue next-hops used when the destination $Y \gg S$, the computing router, MUST be one or more nodes, T , whose `topo_order` is in the interval $[X.topo_order, Y.topo_order]$ and where $Y \gg T$ or Y is T . Similarly, the MRT-Red next-hops MUST be have a `topo_order` in the interval $[R-small.topo_order, X.topo_order]$ or $[Y.topo_order, R-big.topo_order]$.

Implementations SHOULD implement the `Select_Alternates()` function to pick an MRT-FRR alternate.

In a future version, this section will include pseudo-code describing the full code path through the pseudo-code given earlier in the draft.

6. Algorithm Alternatives and Evaluation

This specification defines the MRT Lowpoint Algorithm, which is one option among several possible MRT algorithms. Other alternatives are described in the appendices.

In addition, it is possible to calculate Destination-Rooted GADAG, where for each destination, a GADAG rooted at that destination is computed. Then a router can compute the blue MRT and red MRT next-hops to that destination. Building GADAGs per destination is computationally more expensive, but may give somewhat shorter alternate paths. It may be useful for live-live multicast along MRTs.

6.1. Algorithm Evaluation

This section compares MRT and remote LFA for IP Fast Reroute in 19 service provider network topologies, focusing on coverage and alternate path length. Figure 26 shows the node-protecting coverage provided by local LFA (LLFA), remote LFA (RLFA), and MRT against different failure scenarios in these topologies. The coverage values are calculated as the percentage of source-destination pairs protected by the given IPFRR method relative to those protectable by optimal routing, against the same failure modes. More details on alternate selection policies used for this analysis are described later in this section.

Topology	percentage of failure scenarios covered by IPFRR method		
	NP_LLFA	NP_RLFA	MRT
T201	37	90	100
T202	73	83	100
T203	51	80	100
T204	55	81	100
T205	92	93	100
T206	71	74	100
T207	57	74	100
T208	66	81	100
T209	79	79	100
T210	95	98	100
T211	68	71	100
T212	59	63	100
T213	84	84	100
T214	68	78	100
T215	84	88	100
T216	43	59	100
T217	78	88	100
T218	72	75	100
T219	78	84	100

Figure 26

For the topologies analyzed here, LLFA is able to provide node-protecting coverage ranging from 37% to 95% of the source-destination pairs, as seen in the column labeled NP_LLFA. The use of RLFA in addition to LLFA is generally able to increase the node-protecting coverage. The percentage of node-protecting coverage with RLFA is provided in the column labeled NP_RLFA, ranges from 59% to 98% for these topologies. The node-protecting coverage provided by MRT is 100% since MRT is able to provide protection for any source-destination pair for which a path still exists after the failure.

We would also like to measure the quality of the alternate paths produced by these different IPFRR methods. An obvious approach is to take an average of the alternate path costs over all source-destination pairs and failure modes. However, this presents a problem, which we will illustrate by presenting an example of results for one topology using this approach (Figure 27). In this table, the average relative path length is the alternate path length for the IPFRR method divided by the optimal alternate path length, averaged

over all source-destination pairs and failure modes. The first three columns of data in the table give the path length calculated from the sum of IGP metrics of the links in the path. The results for topology T208 show that the metric-based path lengths for NP_LLFA and NP_RLFA alternates are on average 78 and 66 times longer than the path lengths for optimal alternates. The metric-based path lengths for MRT alternates are on average 14 times longer than for optimal alternates.

Topology	average relative alternate path length					
	IGP metric			hopcount		
	NP_LLFA	NP_RLFA	MRT	NP_LLFA	NP_RLFA	MRT
T208	78.2	66.0	13.6	0.99	1.01	1.32

Figure 27

The network topology represented by T208 uses values of 10, 100, and 1000 as IGP costs, so small deviations from the optimal alternate path can result in large differences in relative path length. LLFA, RLFA, and MRT all allow for at least one hop in the alternate path to be chosen independent of the cost of the link. This can easily result in an alternate using a link with cost 1000, which introduces noise into the path length measurement. In the case of T208, the adverse effects of using metric-based path lengths is obvious. However, we have observed that the metric-based path length introduces noise into alternate path length measurements in several other topologies as well. For this reason, we have opted to measure the alternate path length using hopcount. While IGP metrics may be adjusted by the network operator for a number of reasons (e.g. traffic engineering), the hopcount is a fairly stable measurement of path length. As shown in the last three columns of Figure 27, the hopcount-based alternate path lengths for topology T208 are fairly well-behaved.

Figure 28, Figure 29, Figure 30, and Figure 31 present the hopcount-based path length results for the 19 topologies examined. The topologies in the four tables are grouped based on the size of the topologies, as measured by the number of nodes, with Figure 28 having the smallest topologies and Figure 31 having the largest topologies. Instead of trying to represent the path lengths of a large set of alternates with a single number, we have chosen to present a histogram of the path lengths for each IPFRR method and alternate selection policy studied. The first eight columns of data represent

the percentage of failure scenarios protected by an alternate N hops longer than the primary path, with the first column representing an alternate 0 or 1 hops longer than the primary path, all the way up through the eighth column representing an alternate 14 or 15 hops longer than the primary path. The last column in the table gives the percentage of failure scenarios for which there is no alternate less than 16 hops longer than the primary path. In the case of LLFA and RLFA, this category includes failure scenarios for which no alternate was found.

For each topology, the first row (labeled OPTIMAL) is the distribution of the number of hops in excess of the primary path hopcount for optimally routed alternates. (The optimal routing was done with respect to IGP metrics, as opposed to hopcount.) The second row (labeled NP_LLFA) is the distribution of the extra hops for node-protecting LLFA. The third row (labeled NP_LLFA_THEN_NP_RLFA) is the hopcount distribution when one adds node-protecting RLFA to increase the coverage. The alternate selection policy used here first tries to find a node-protecting LLFA. If that does not exist, then it tries to find an RLFA, and checks if it is node-protecting. Comparing the hopcount distribution for RLFA and LLFA across these topologies, one can see how the coverage is increased at the expense of using longer alternates. It is also worth noting that while superficially LLFA and RLFA appear to have better hopcount distributions than OPTIMAL, the presence of entries in the last column (no alternate < 16) mainly represent failure scenarios that are not protected, for which the hopcount is effectively infinite.

The fourth and fifth rows of each topology show the hopcount distributions for two alternate selection policies using MRT alternates. The policy represented by the label NP_LLFA_THEN_MRT_LOWPOINT will first use a node-protecting LLFA. If a node-protecting LLFA does not exist, then it will use an MRT alternate. The policy represented by the label MRT_LOWPOINT instead will use the MRT alternate even if a node-protecting LLFA exists. One can see from the data that combining node-protecting LLFA with MRT results in a significant shortening of the alternate hopcount distribution.

Topology name and alternate selection policy evaluated	percentage of failure scenarios protected by an alternate N hops longer than the primary path									
	0-1	2-3	4-5	6-7	8-9	10-11	12-13	14-15	no alt	<16
T201(avg primary hops=3.5)										
OPTIMAL	37	37	20	3	3					
NP_LLFA	37									63
NP_LLFA_THEN_NP_RLFA	37	34	19							10
NP_LLFA_THEN_MRT_LOWPOINT	37	33	21	6	3					
MRT_LOWPOINT	33	36	23	6	3					
T202(avg primary hops=4.8)										
OPTIMAL	90	9								
NP_LLFA	71	2								27
NP_LLFA_THEN_NP_RLFA	78	5								17
NP_LLFA_THEN_MRT_LOWPOINT	80	12	5	2	1					
MRT_LOWPOINT_ONLY	48	29	13	7	2	1				
T203(avg primary hops=4.1)										
OPTIMAL	36	37	21	4	2					
NP_LLFA	34	15	3							49
NP_LLFA_THEN_NP_RLFA	35	19	22	4						20
NP_LLFA_THEN_MRT_LOWPOINT	36	35	22	5	2					
MRT_LOWPOINT_ONLY	31	35	26	7	2					
T204(avg primary hops=3.7)										
OPTIMAL	76	20	3	1						
NP_LLFA	54	1								45
NP_LLFA_THEN_NP_RLFA	67	10	4							19
NP_LLFA_THEN_MRT_LOWPOINT	70	18	8	3	1					
MRT_LOWPOINT_ONLY	58	27	11	3	1					
T205(avg primary hops=3.4)										
OPTIMAL	92	8								
NP_LLFA	89	3								8
NP_LLFA_THEN_NP_RLFA	90	4								7
NP_LLFA_THEN_MRT_LOWPOINT	91	9								
MRT_LOWPOINT_ONLY	62	33	5	1						

Figure 28

Topology name and alternate selection policy evaluated	percentage of failure scenarios protected by an alternate N hops longer than the primary path								
	0-1	2-3	4-5	6-7	8-9	10-11	12-13	14-15	no alt <16
T206(avg primary hops=3.7)									
OPTIMAL	63	30	7						
NP_LLFA	60	9	1						29
NP_LLFA_THEN_NP_RLFA	60	13	1						26
NP_LLFA_THEN_MRT_LOWPOINT	64	29	7						
MRT_LOWPOINT	55	32	13						
T207(avg primary hops=3.9)									
OPTIMAL	71	24	5	1					
NP_LLFA	55	2							43
NP_LLFA_THEN_NP_RLFA	63	10							26
NP_LLFA_THEN_MRT_LOWPOINT	70	20	7	2	1				
MRT_LOWPOINT_ONLY	57	29	11	3	1				
T208(avg primary hops=4.6)									
OPTIMAL	58	28	12	2	1				
NP_LLFA	53	11	3						34
NP_LLFA_THEN_NP_RLFA	56	17	7	1					19
NP_LLFA_THEN_MRT_LOWPOINT	58	19	10	7	3	1			
MRT_LOWPOINT_ONLY	34	24	21	13	6	2	1		
T209(avg primary hops=3.6)									
OPTIMAL	85	14	1						
NP_LLFA	79								21
NP_LLFA_THEN_NP_RLFA	79								21
NP_LLFA_THEN_MRT_LOWPOINT	82	15	2						
MRT_LOWPOINT_ONLY	63	29	8						
T210(avg primary hops=2.5)									
OPTIMAL	95	4	1						
NP_LLFA	94	1							5
NP_LLFA_THEN_NP_RLFA	94	3	1						2
NP_LLFA_THEN_MRT_LOWPOINT	95	4	1						
MRT_LOWPOINT_ONLY	91	6	2						

Figure 29

Topology name and alternate selection policy evaluated	percentage of failure scenarios protected by an alternate N hops longer than the primary path									
	0-1	2-3	4-5	6-7	8-9	10-11	12-13	14-15	no alt	<16
T211(avg primary hops=3.3)										
OPTIMAL	88	11								
NP_LLFA	66	1								32
NP_LLFA_THEN_NP_RLFA	68	3								29
NP_LLFA_THEN_MRT_LOWPOINT	88	12								
MRT_LOWPOINT	85	15	1							
T212(avg primary hops=3.5)										
OPTIMAL	76	23	1							
NP_LLFA	59									41
NP_LLFA_THEN_NP_RLFA	61	1	1							37
NP_LLFA_THEN_MRT_LOWPOINT	75	24	1							
MRT_LOWPOINT_ONLY	66	31	3							
T213(avg primary hops=4.3)										
OPTIMAL	91	9								
NP_LLFA	84									16
NP_LLFA_THEN_NP_RLFA	84									16
NP_LLFA_THEN_MRT_LOWPOINT	89	10	1							
MRT_LOWPOINT_ONLY	75	24	1							
T214(avg primary hops=5.8)										
OPTIMAL	71	22	5	2						
NP_LLFA	58	8	1	1						32
NP_LLFA_THEN_NP_RLFA	61	13	3	1						22
NP_LLFA_THEN_MRT_LOWPOINT	66	14	7	5	3	2	1	1	1	1
MRT_LOWPOINT_ONLY	30	20	18	12	8	4	3	2	3	3
T215(avg primary hops=4.8)										
OPTIMAL	73	27								
NP_LLFA	73	11								16
NP_LLFA_THEN_NP_RLFA	73	13	2							12
NP_LLFA_THEN_MRT_LOWPOINT	74	19	3	2	1	1	1			
MRT_LOWPOINT_ONLY	32	31	16	12	4	3	1			

Figure 30

Topology name and alternate selection policy evaluated	percentage of failure scenarios protected by an alternate N hops longer than the primary path								
	0-1	2-3	4-5	6-7	8-9	10-11	12-13	14-15	no alt <16
T216(avg primary hops=5.2)									
OPTIMAL	60	32	7	1					
NP_LLFA	39	4							57
NP_LLFA_THEN_NP_RLFA	46	12	2						41
NP_LLFA_THEN_MRT_LOWPOINT	48	20	12	7	5	4	2	1	1
MRT_LOWPOINT	28	25	18	11	7	6	3	2	1
T217(avg primary hops=8.0)									
OPTIMAL	81	13	5	1					
NP_LLFA	74	3	1						22
NP_LLFA_THEN_NP_RLFA	76	8	3	1					12
NP_LLFA_THEN_MRT_LOWPOINT	77	7	5	4	3	2	1	1	
MRT_LOWPOINT_ONLY	25	18	18	16	12	6	3	1	
T218(avg primary hops=5.5)									
OPTIMAL	85	14	1						
NP_LLFA	68	3							28
NP_LLFA_THEN_NP_RLFA	71	4							25
NP_LLFA_THEN_MRT_LOWPOINT	77	12	7	4	1				
MRT_LOWPOINT_ONLY	37	29	21	10	3	1			
T219(avg primary hops=7.7)									
OPTIMAL	77	15	5	1	1				
NP_LLFA	72	5							22
NP_LLFA_THEN_NP_RLFA	73	8	2						16
NP_LLFA_THEN_MRT_LOWPOINT	74	8	3	3	2	2	2	2	4
MRT_LOWPOINT_ONLY	19	14	15	12	10	8	7	6	10

Figure 31

In the preceding analysis, the following procedure for selecting an RLFA was used. Nodes were ordered with respect to distance from the source and checked for membership in Q and P-space. The first node to satisfy this condition was selected as the RLFA. More sophisticated methods to select node-protecting RLFAs is an area of active research.

The analysis presented above uses the MRT Lowpoint Algorithm defined in this specification with a common GADAG root. The particular choice of a common GADAG root is expected to affect the quality of the MRT alternate paths, with a more central common GADAG root resulting in shorter MRT alternate path lengths. For the analysis above, the GADAG root was chosen for each topology by calculating node centrality as the sum of costs of all shortest paths to and from a given node. The node with the lowest sum was chosen as the common GADAG root. In actual deployments, the common GADAG root would be chosen based on the GADAG Root Selection Priority advertised by each router, the values of which would be determined off-line.

In order to measure how sensitive the MRT alternate path lengths are to the choice of common GADAG root, we performed the same analysis using different choices of GADAG root. All of the nodes in the network were ordered with respect to the node centrality as computed above. Nodes were chosen at the 0th, 25th, and 50th percentile with respect to the centrality ordering, with 0th percentile being the most central node. The distribution of alternate path lengths for those three choices of GADAG root are shown in Figure 32 for a subset of the 19 topologies (chosen arbitrarily). The third row for each topology (labeled MRT_LOWPOINT (0 percentile)) reproduces the results presented above for MRT_LOWPOINT_ONLY. The fourth and fifth rows show the alternate path length distribution for the 25th and 50th percentile choice for GADAG root. One can see some impact on the path length distribution with the less central choice of GADAG root resulting in longer path lengths.

We also looked at the impact of MRT algorithm variant on the alternate path lengths. The first two rows for each topology present results of the same alternate path length distribution analysis for the SPF and Hybrid methods for computing the GADAG. These two methods are described in Appendix A and Appendix B. For three of the topologies in this subset (T201, T206, and T211), the use of SPF or Hybrid methods does not appear to provide a significant advantage over the Lowpoint method with respect to path length. Instead, the choice of GADAG root appears to have more impact on the path length. However, for two of the topologies in this subset (T216 and T219) and for this particular choice of GADAG root, the use of the SPF method results in noticeably shorter alternate path lengths than the use of the Lowpoint or Hybrid methods. It remains to be determined if this effect applies generally across more topologies or is sensitive to choice of GADAG root.

Topology name	percentage of failure scenarios protected by an alternate N hops longer than the primary path									
MRT algorithm variant										
(GADAG root centrality percentile)	0-1	2-3	4-5	6-7	8-9	10-11	12-13	14-15	no alt	<16
T201(avg primary hops=3.5)										
MRT_HYBRID (0 percentile)	33	26	23	6	3					
MRT_SPF (0 percentile)	33	36	23	6	3					
MRT_LOWPOINT (0 percentile)	33	36	23	6	3					
MRT_LOWPOINT (25 percentile)	27	29	23	11	10					
MRT_LOWPOINT (50 percentile)	27	29	23	11	10					
T206(avg primary hops=3.7)										
MRT_HYBRID (0 percentile)	50	35	13	2						
MRT_SPF (0 percentile)	50	35	13	2						
MRT_LOWPOINT (0 percentile)	55	32	13							
MRT_LOWPOINT (25 percentile)	47	25	22	6						
MRT_LOWPOINT (50 percentile)	38	38	14	11						
T211(avg primary hops=3.3)										
MRT_HYBRID (0 percentile)	86	14								
MRT_SPF (0 percentile)	86	14								
MRT_LOWPOINT (0 percentile)	85	15	1							
MRT_LOWPOINT (25 percentile)	70	25	5	1						
MRT_LOWPOINT (50 percentile)	80	18	2							
T216(avg primary hops=5.2)										
MRT_HYBRID (0 percentile)	23	22	18	13	10	7	4	2	2	
MRT_SPF (0 percentile)	35	32	19	9	3	1				
MRT_LOWPOINT (0 percentile)	28	25	18	11	7	6	3	2	1	
MRT_LOWPOINT (25 percentile)	24	20	19	16	10	6	3	1		
MRT_LOWPOINT (50 percentile)	19	14	13	10	8	6	5	5	10	
T219(avg primary hops=7.7)										
MRT_HYBRID (0 percentile)	20	16	13	10	7	5	5	5	3	
MRT_SPF (0 percentile)	31	23	19	12	7	4	2	1		
MRT_LOWPOINT (0 percentile)	19	14	15	12	10	8	7	6	10	
MRT_LOWPOINT (25 percentile)	19	14	15	13	12	10	6	5	7	
MRT_LOWPOINT (50 percentile)	19	14	14	12	11	8	6	6	10	

Figure 32

7. Algorithm Work to Be Done

Broadcast Interfaces: The algorithm assumes that broadcast interfaces are already represented as pseudo-nodes in the network graph. Given maximal redundancy, one of the MRT will try to avoid both the pseudo-node and the next hop. The exact rules need to be fully specified.

8. IANA Considerations

This document includes no request to IANA.

9. Security Considerations

This architecture is not currently believed to introduce new security concerns.

10. References

10.1. Normative References

[I-D.ietf-rtgwg-mrt-frr-architecture]
Atlas, A., Kebler, R., Envedi, G., Csaszar, A., Tantsura, J., Konstantynowicz, M., and R. White, "An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees", draft-ietf-rtgwg-mrt-frr-architecture-03 (work in progress), July 2013.

10.2. Informative References

[EnyediThesis]
Enyedi, G., "Novel Algorithms for IP Fast Reroute", Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics Ph.D. Thesis, February 2011, <http://www.omikk.bme.hu/collections/phd/Villamosmernoki_es_Informatikai_Kar/2011/Enyedi_Gabor/ertekezes.pdf>.

[I-D.atlas-ospf-mrt]
Atlas, A., Hegde, S., Chris, C., and J. Tantsura, "OSPF Extensions to Support Maximally Redundant Trees", draft-atlas-ospf-mrt-00 (work in progress), July 2013.

[I-D.ietf-rtgwg-ipfrr-notvia-addresses]
Bryant, S., Previdi, S., and M. Shand, "A Framework for IP and MPLS Fast Reroute Using Not-via Addresses", draft-ietf-rtgwg-ipfrr-notvia-addresses-11 (work in progress), May 2013.

- [I-D.ietf-rtgwg-lfa-manageability]
Litkowski, S., Decraene, B., Filsfils, C., and K. Raza,
"Operational management of Loop Free Alternates", draft-
ietf-rtgwg-lfa-manageability-00 (work in progress), May
2013.
- [I-D.ietf-rtgwg-remote-lfa]
Bryant, S., Filsfils, C., Previdi, S., Shand, M., and S.
Ning, "Remote LFA FRR", draft-ietf-rtgwg-remote-lfa-02
(work in progress), May 2013.
- [Kahn_1962_topo_sort]
Kahn, A., "Topological sorting of large networks",
Communications of the ACM, Volume 5, Issue 11 , Nov 1962,
<<http://dl.acm.org/citation.cfm?doid=368996.369025>>.
- [LFARevisited]
Retvari, G., Tapolcai, J., Enyedi, G., and A. Csaszar, "IP
Fast ReRoute: Loop Free Alternates Revisited", Proceedings
of IEEE INFOCOM , 2011, <http://opti.tmit.bme.hu/~tapolcai/papers/retvari2011lfa_infocom.pdf>.
- [LightweightNotVia]
Enyedi, G., Retvari, G., Szilagyi, P., and A. Csaszar, "IP
Fast ReRoute: Lightweight Not-Via without Additional
Addresses", Proceedings of IEEE INFOCOM , 2009,
<<http://mycite.omikk.bme.hu/doc/71691.pdf>>.
- [MRTLlinear]
Enyedi, G., Retvari, G., and A. Csaszar, "On Finding
Maximally Redundant Trees in Strictly Linear Time", IEEE
Symposium on Computers and Communications (ISCC) , 2009,
<<http://opti.tmit.bme.hu/~enyedi/ipfrr/distMaxRedTree.pdf>>.
- [RFC3137] Retana, A., Nguyen, L., White, R., Zinin, A., and D.
McPherson, "OSPF Stub Router Advertisement", RFC 3137,
June 2001.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast
Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC
5714, January 2010.

[RFC6571] Filsfils, C., Francois, P., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.

Appendix A. Option 2: Computing GADAG using SPF

The basic idea in this option is to use slightly-modified SPF computations to find ears. In every block, an SPF computation is first done to find a cycle from the local root and then SPF computations in that block find ears until there are no more interfaces to be explored. The used result from the SPF computation is the path of interfaces indicated by following the previous hops from the minimized IN_GADAG node back to the SPF root.

To do this, first all cut-vertices must be identified and local-roots assigned as specified in Figure 12.

The slight modifications to the SPF are as follows. The root of the block is referred to as the block-root; it is either the GADAG root or a cut-vertex.

- a. The SPF is rooted at a neighbor x of an IN_GADAG node y. All links between y and x are marked as TEMP_UNUSABLE. They should not be used during the SPF computation.
- b. If y is not the block-root, then it is marked TEMP_UNUSABLE. It should not be used during the SPF computation. This prevents ears from starting and ending at the same node and avoids cycles; the exception is because cycles to/from the block-root are acceptable and expected.
- c. Do not explore links to nodes whose local-root is not the block-root. This keeps the SPF confined to the particular block.
- d. Terminate when the first IN_GADAG node z is minimized.
- e. Respect the existing directions (e.g. INCOMING, OUTGOING, UNDIRECTED) already specified for each interface.

```
Mod_SPF(spf_root, block_root)
  Initialize spf_heap to empty
  Initialize nodes' spf_metric to infinity
  spf_root.spf_metric = 0
  insert(spf_heap, spf_root)
  found_in_gadag = false
  while (spf_heap is not empty) and (found_in_gadag is false)
```



```

min_node = remove_lowest(spf_heap)
if min_node.IN_GADAG is true
    found_in_gadag = true
else
    foreach interface intf of min_node
        if ((intf.OUTGOING or intf.UNDIRECTED) and
            ((intf.remote_node.localroot is block_root) or
             (intf.remote_node is block_root)) and
            (intf.remote_node is not TEMP_UNUSABLE) and
            (intf is not TEMP_UNUSABLE))
            path_metric = min_node.spf_metric + intf.metric
            if path_metric < intf.remote_node.spf_metric
                intf.remote_node.spf_metric = path_metric
                intf.remote_node.spf_prev_intf = intf
                insert_or_update(spf_heap, intf.remote_node)
    return min_node

SPF_for_Ear(cand_intf.local_node, cand_intf.remote_node, block_root,
            method)
    Mark all interfaces between cand_intf.remote_node
        and cand_intf.local_node as TEMP_UNUSABLE
    if cand_intf.local_node is not block_root
        Mark cand_intf.local_node as TEMP_UNUSABLE
    Initialize ear_list to empty
    end_ear = Mod_SPF(spf_root, block_root)
    y = end_ear.spf_prev_hop
    while y.local_node is not spf_root
        add_to_list_start(ear_list, y)
        y.local_node.IN_GADAG = true
        y = y.local_node.spf_prev_intf
    if(method is not hybrid)
        Set_Ear_Direction(ear_list, cand_intf.local_node,
                           end_ear, block_root)
    Clear TEMP_UNUSABLE from all interfaces between
        cand_intf.remote_node and cand_intf.local_node
    Clear TEMP_UNUSABLE from cand_intf.local_node
    return end_ear

```

Figure 33: Modified SPF for GADAG computation

Assume that an ear is found by going from y to x and then running an SPF that terminates by minimizing z (e.g. $y \leftrightarrow x \dots q \leftrightarrow z$). Now it is necessary to determine the direction of the ear; if $y \ll z$, then the path should be $y \rightarrow x \dots q \rightarrow z$ but if $y \gg z$, then the path should be $y \leftarrow x \dots q \leftarrow z$. In Section 4.4, the same problem was handled by finding

all ears that started at a node before looking at ears starting at nodes higher in the partial order. In this algorithm, using that approach could mean that new ears aren't added in order of their total cost since all ears connected to a node would need to be found before additional nodes could be found.

The alternative is to track the order relationship of each node with respect to every other node. This can be accomplished by maintaining two sets of nodes at each node. The first set, *Higher_Nodes*, contains all nodes that are known to be ordered above the node. The second set, *Lower_Nodes*, contains all nodes that are known to be ordered below the node. This is the approach used in this algorithm.

```

Set_Ear_Direction(ear_list, end_a, end_b, block_root)
// Default of A_TO_B for the following cases:
// (a) end_a and end_b are the same (root)
// or (b) end_a is in end_b's Lower_Nodes
// or (c) end_a and end_b were unordered with respect to each
// other
direction = A_TO_B
if (end_b is block_root) and (end_a is not end_b)
    direction = B_TO_A
else if end_a is in end_b.Higher_Nodes
    direction = B_TO_A
if direction is B_TO_A
    foreach interface i in ear_list
        i.UNDIRECTED = false
        i.INCOMING = true
        i.remote_intf.UNDIRECTED = false
        i.remote_intf.OUTGOING = true
else
    foreach interface i in ear_list
        i.UNDIRECTED = false
        i.OUTGOING = true
        i.remote_intf.UNDIRECTED = false
        i.remote_intf.INCOMING = true
if end_a is end_b
    return
// Next, update all nodes' Lower_Nodes and Higher_Nodes
if (end_a is in end_b.Higher_Nodes)
    foreach node x where x.localroot is block_root
        if end_a is in x.Lower_Nodes
            foreach interface i in ear_list
                add i.remote_node to x.Lower_Nodes
        if end_b is in x.Higher_Nodes
            foreach interface i in ear_list
                add i.local_node to x.Higher_Nodes

```

```

else
  foreach node x where x.localroot is block_root
    if end_b is in x.Lower_Nodes
      foreach interface i in ear_list
        add i.local_node to x.Lower_Nodes
    if end_a is in x.Higher_Nodes
      foreach interface i in ear_list
        add i.remote_node to x.Higher_Nodes

```

Figure 34: Algorithm to assign links of an ear direction

A goal of the algorithm is to find the shortest cycles and ears. An ear is started by going to a neighbor x of an IN_GADAG node y. The path from x to an IN_GADAG node is minimal, since it is computed via SPF. Since a shortest path is made of shortest paths, to find the shortest ears requires reaching from the set of IN_GADAG nodes to the closest node that isn't IN_GADAG. Therefore, an ordered tree is maintained of interfaces that could be explored from the IN_GADAG nodes. The interfaces are ordered by their characteristics of metric, local loopback address, remote loopback address, and ifindex, as in the algorithm previously described in Figure 14.

The algorithm ignores interfaces picked from the ordered tree that belong to the block root if the block in which the interface is present already has an ear that has been computed. This is necessary since we allow at most one incoming interface to a block root in each block. This requirement stems from the way next-hops are computed as will be seen in Section 4.6. After any ear gets computed, we traverse the newly added nodes to the GADAG and insert interfaces whose far end is not yet on the GADAG to the ordered tree for later processing.

Finally, cut-edges are a special case because there is no point in doing an SPF on a block of 2 nodes. The algorithm identifies cut-edges simply as links where both ends of the link are cut-vertices. Cut-edges can simply be added to the GADAG with both OUTGOING and INCOMING specified on their interfaces.

```

add_eligible_interfaces_of_node(ordered_intfs_tree,node)
  for each interface of node
    if intf.remote_node.IN_GADAG is false
      insert(intf,ordered_intfs_tree)

check_if_block_has_ear(x,block_id)
  block_has_ear = false
  for all interfaces of x
    if (intf.remote_node.block_id == block_id) &&
      (intf.remote_node.IN_GADAG is true)

```

```

        block_has_ear = true
    return block_has_ear

Construct_GADAG_via_SPF(topology, root)
    Compute_Localroot (root,root)
    Assign_Block_ID(root,0)
    root.IN_GADAG = true
    add_eligible_interfaces_of_node(ordered_intfs_tree,root)
    while ordered_intfs_tree is not empty
        cand_intf = remove_lowest(ordered_intfs_tree)
        if cand_intf.remote_node.IN_GADAG is false
            if L(cand_intf.remote_node) == D(cand_intf.remote_node)
                // Special case for cut-edges
                cand_intf.UNDIRECTED = false
                cand_intf.remote_intf.UNDIRECTED = false
                cand_intf.OUTGOING = true
                cand_intf.INCOMING = true
                cand_intf.remote_intf.OUTGOING = true
                cand_intf.remote_intf.INCOMING = true
                cand_intf.remote_node.IN_GADAG = true
            add_eligible_interfaces_of_node(
                ordered_intfs_tree,cand_intf.remote_node)
        else
            if (cand_intf.remote_node.local_root ==
                cand_intf.local_node) &&
                check_if_block_has_ear
                    (cand_intf.local_node,
                     cand_intf.remote_node.block_id))
                /* Skip the interface since the block root
                   already has an incoming interface in the
                   block */
            else
                ear_end = SPF_for_Ear(cand_intf.local_node,
                                       cand_intf.remote_node,
                                       cand_intf.remote_node.localroot,
                                       SPF method)
                y = ear_end.spf_prev_hop
                while y.local_node is not cand_intf.local_node
                    add_eligible_interfaces_of_node(
                        ordered_intfs_tree,
                        y.local_node)
                    y = y.local_node.spf_prev_intf

```

Figure 35: SPF-based GADAG algorithm

Appendix B. Option 3: Computing GADAG using a hybrid method

In this option, the idea is to combine the salient features of the above two options. To this end, we process nodes as they get added to the GADAG just like in the lowpoint inheritance by maintaining a stack of nodes. This ensures that we do not need to maintain lower and higher sets at each node to ascertain ear directions since the ears will always be directed from the node being processed towards the end of the ear. To compute the ear however, we resort to an SPF to have the possibility of better ears (path lengths) thus giving more flexibility than the restricted use of lowpoint/dfs parents.

Regarding ears involving a block root, unlike the SPF method which ignored interfaces of the block root after the first ear, in the hybrid method we would have to process all interfaces of the block root before moving on to other nodes in the block since the direction of an ear is pre-determined. Thus, whenever the block already has an ear computed, and we are processing an interface of the block root, we mark the block root as unusable before the SPF run that computes the ear. This ensures that the SPF terminates at some node other than the block-root. This in turn guarantees that the block-root has only one incoming interface in each block, which is necessary for correctly computing the next-hops on the GADAG.

As in the SPF gadag, bridge ears are handled as a special case.

The entire algorithm is shown below in Figure 36

```

find_spf_stack_ear(stack, x, y, xy_intf, block_root)
  if L(y) == D(y)
    // Special case for cut-edges
    xy_intf.UNDIRECTED = false
    xy_intf.remote_intf.UNDIRECTED = false
    xy_intf.OUTGOING = true
    xy_intf.INCOMING = true
    xy_intf.remote_intf.OUTGOING = true
    xy_intf.remote_intf.INCOMING = true
    xy_intf.remote_node.IN_GADAG = true
    push y onto stack
    return
  else
    if (y.local_root == x) &&
      check_if_block_has_ear(x,y.block_id)
      //Avoid the block root during the SPF
      Mark x as TEMP_UNUSABLE
    end_ear = SPF_for_Ear(x,y,block_root,hybrid)
    If x was set as TEMP_UNUSABLE, clear it
    cur = end_ear
    while (cur != y)
      intf = cur.spf_prev_hop

```

```

        prev = intf.local_node
        intf.UNDIRECTED = false
        intf.remote_intf.UNDIRECTED = false
        intf.OUTGOING = true
        intf.remote_intf.INCOMING = true
        push prev onto stack
    cur = prev
    xy_intf.UNDIRECTED = false
    xy_intf.remote_intf.UNDIRECTED = false
    xy_intf.OUTGOING = true
    xy_intf.remote_intf.INCOMING = true
    return

Construct_GADAG_via_hybrid(topology,root)
Compute_Localroot (root,root)
Assign_Block_ID(root,0)
root.IN_GADAG = true
Initialize Stack to empty
push root onto Stack
while (Stack is not empty)
    x = pop(Stack)
    for each interface intf of x
        y = intf.remote_node
        if y.IN_GADAG is false
            find_spf_stack_ear(stack, x, y, intf, y.block_root)

```

Figure 36: Hybrid GADAG algorithm

Authors' Addresses

Gabor Sandor Enyedi (editor)
 Ericsson
 Konyves Kalman krt 11
 Budapest 1097
 Hungary

Email: Gabor.Sandor.Enyedi@ericsson.com

Andras Csaszar
 Ericsson
 Konyves Kalman krt 11
 Budapest 1097
 Hungary

Email: Andras.Csaszar@ericsson.com

Alia Atlas (editor)
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: akatlas@juniper.net

Chris Bowers
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
USA

Email: cbowers@juniper.net

Abishek Gopalan
University of Arizona
1230 E Speedway Blvd.
Tucson, AZ 85721
USA

Email: abishek@ece.arizona.edu

Routing Area Working Group
Internet-Draft
Intended status: Standards Track
Expires: November 16, 2013

H. Gredler, Ed.
Juniper Networks, Inc.
S. Amante
Level 3 Communications, Inc.
T. Scholl
Amazon
L. Jalil
Verizon
May 15, 2013

Advertising MPLS labels in IGPs
draft-gredler-rtgwg-igp-label-advertisement-05

Abstract

Historically MPLS label distribution was driven by session oriented protocols. In order to obtain a particular routers label binding for a given destination FEC one needs to have first an established session with that node.

This document describes a mechanism to distribute FEC/label mappings through flooding protocols. Flooding protocols publish their objects for an unknown set of receivers, therefore one can efficiently scale label distribution for use cases where the receiver of label information is not directly connected.

Application of this technique are found in the field of backup (Bypass, R-LFA) routing, Label switched path stitching, egress protection, explicit routing and egress ASBR link selection.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months

and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on November 16, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	4
2. Motivation and Applicability	4
3. Use cases for IGP label distribution	5
3.1. Increase LFA backup coverage using 'Directed Forwarding'	5
3.2. Egress ASBR Link Selection	6
3.3. Tail end protection of BGP service routes	7
3.4. Explicit Path Routing through Label Stacking	8
3.5. Link and Node Protection LSPs	10
3.6. Stitching MPLS Label Switched Path Segments	12
3.7. T-LDP replacement for infrastructure labels	13
4. Acknowledgements	14
5. IANA Considerations	14
6. Security Considerations	14
7. References	14
7.1. Normative References	14
7.2. Informative References	15
Authors' Addresses	15

1. Introduction

MPLS label allocations are predominantly distributed by using the LDP [RFC5036], RSVP [RFC5151] or labeled BGP [RFC3107] protocol. All of those protocols have in common that they are session oriented, which means that in order to learn the Label Information database of a particular router one needs to have a direct control-plane session using the given protocol.

There are a couple of practical use cases where the consumer of a MPLS label allocation may not be adjacent to the router having allocated the label. Bringing up an explicit session using existing label distribution protocols between the non-adjacent label allocator and the label consumer is the existing remedy for this dilemma.

For LDP protection routing LDP next next hop labels [NNHOP] have been proposed to provide the 2 hop neighborhood labels. While the 2 hop neighborhood provides good backup coverage for the typical network operator topology it is inadequate for some sparse for example ring like topologies.

Depending on the application, retrieval and setup of forwarding state of such >1 hop label allocations may only be transient. As such configuring and un-configuring the explicit session is an operational burden and therefore should be avoided.

The use cases described in this document are equally applicable to IPv4 and IPv6 carried over MPLS. Furthermore the proposed use of distributing MPLS Labels using IGP protocols adheres to the architectural principles laid out in [RFC3031].

2. Motivation and Applicability

It may not be immediate obvious, however introduction of Remote LFA [I-D.ietf-rtgwg-remote-lfa] technology has implied important changes for an IGP implementation. Previously the IGP had a one-way communication path with the LDP module. The IGP supplies tracking routes and LDP selects the best neighbor based upon FEC to tracking routes exact matching results. Remote LFA changes that relationship such that there is a bi-directional communication path between the IGP and LDP. Now the IGP needs to learn about if a label switched path to a given destination prefix has been established and what the ingress label for getting there is. The IGP needs to push that label for the tracking routes of destinations beyond a remote LFA neighbor.

Since the IGP is now aware of label switched paths and it does create forwarding state based on label information it makes sense to

distribute label switched paths by the IGP as well.

3. Use cases for IGP label distribution

This section lists example use cases which illustrate IGP distribution of MPLS label switched paths.

3.1. Increase LFA backup coverage using 'Directed Forwarding'

Deployment of Loop free alternate backup technology [RFC5286] results in backup graphs whose coverage is highly dependent on the underlying Layer-3 topology. Typical network deployments provide backup coverage less than 100 percent (see RFC 6571 Section 4.3 for Results [RFC6571]) for IGP destination prefixes.

By closer examining the coverage gaps from the referenced production network topologies, it becomes obvious that most topologies lacking backup coverage are close to ring shaped topologies (Figure 1).

Remote LFA [I-D.ietf-rtgwg-remote-lfa] has introduced the notion of a "remote" LFA neighbor. This helper router which is both in P and Q space could forward the traffic to the final destination. Router 'H' is in P space, however due to the actual metric allocation router 'H' is not in Q space.

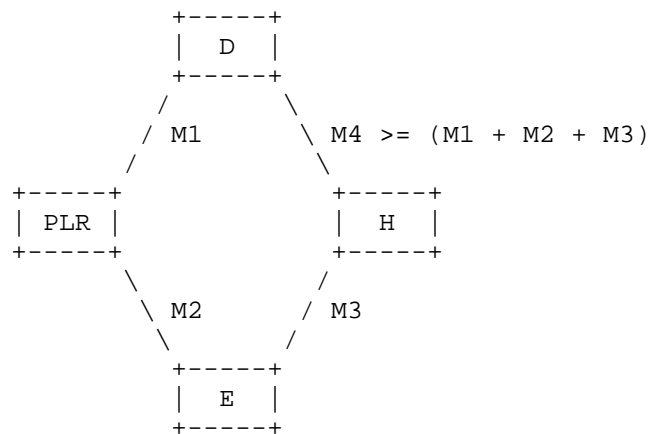


Figure 1: Coverage gap analysis

The protection router (PLR) evaluates for a primary path to destination 'D' if $\{E \rightarrow H \rightarrow D\}$ is a viable backup path. Because the metric $M4 \{H \rightarrow D\}$ is higher than the sum of the original primary path and the path from router 'H' to the PLR, this particular path

would result in a loop and therefore is rejected.

Now consider that router 'H' would advertise a label for FEC 'D', which has the semantics that H will POP the label and forward to the destination node 'D'. This is done irrespective of the underlying IGP metric 'M4' it is a 'strict forwarding' label. The PLR router can now construct a label stack where the outermost label provides transport to router 'H'. The next label on the MPLS stack is the IGP learned 'strict forwarding label' label. Note that the label 'strict forwarding' semantics are similar to a 1-hop ERO (Explicit route object). The Remote 'LFA' calculation would need to get changed, such that even if a node is not in PQ space, but rather in P space, it may get used as a backup neighbor if it advertises a strict forwarding label to the final destination. A recursive version of the algorithm is applicable as well as long a node in P space has some non looping LSP path to the final destination. The PLR router can now program a backup path irrespective of the undesirable underlying layer-3 topology.

Using existing tunnels for backup routing has been previously described in [I-D.bryant-ipfrr-tunnels]. Section 5.2.3 'Directed forwarding' describes an option to insert a single MPLS label between the tunnel and the payload. Traffic may thereby be directed to a particular neighbor. The mechanism described in this document, is an MPLS specific manifestation of 'Directed forwarding'.

3.2. Egress ASBR Link Selection

In the topology described in Figure 2. router 'S' is facing a dilemma. Router S receives a BGP route from all of its 4 upstream routers. Using existing mechanism the provider owning AS1 can control the loading of its direct links *to* its ASBR1 and ASBR2, however it cannot control the load of the links beyond the ASBRs, except manually tweaking the eBGP import policy and filtering out a certain prefix. It would be more desirable to have visibility of all four BGP paths and be able to control the loading of those four paths using Weighted ECMP. Note that the computation of the 'Weight' percentage and the component doing this computation (Router embedded or SDN) is outside the scope of this document.

If all the ASes would be under one common administrative control then the network operator could deploy a forwarding hierarchy by using [RFC3107] to learn about the remote-AS BGP nexthop addresses and associated labels. An ingress router 'S' would then stack the transport label to its local egress ASBR and the remote ASBR supplied label. In reality it is hard to convince a peering AS to deploy another protocol just in order to easier control the egress load on the WAN links for the ingress AS.

calculation would have passed the ERO {S, R1, R4, R2, D} down to RSVP for signaling.

One of the functions that RSVP-TE provides, is that it keeps track of all the reservations over a particular link, enabling support for such traffic engineering features as bandwidth constraints, LSP priorities, and LSP preemption. However, support for these features with RSVP-TE has a cost associated with it, as it does require a node to maintain control and data plane state for all the individual point-to-point LSPs traversing the node (modulo the LSPs that rely on the LSP hierarchy). This is a use case for constructing explicitly routed paths, without the need to maintain per LSP control/data plane state on the nodes traversed by the LSP. This use case assumes that either support for bandwidth constraints, LSP priorities, and LSP preemption is not needed, or that such support is provided by means outside of this document.

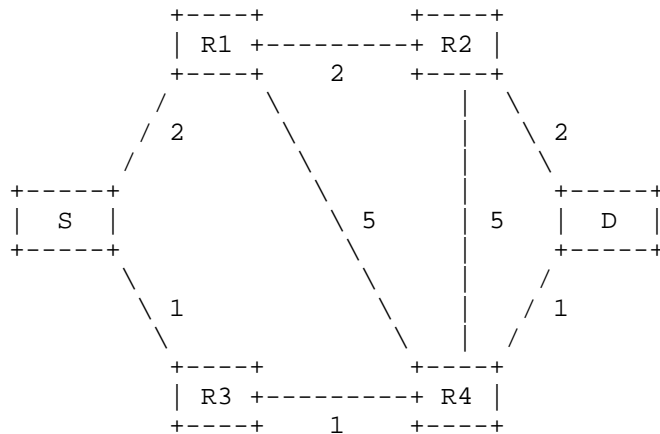


Figure 4: Explicit Routing using Label stacking

Consider now every router along the path does advertise a strict forwarding label for its direct neighbor. Router S could now construct a couple of paths for avoiding the hot links without explicitly signaling them.

- o {S, R1, R2, D}
- o {S, R1, R4, D}
- o {S, R1, R4, R2, D}

Note that not every hop in the ERO needs to be unique label in the label stack. This is undesired as existing forwarding hardware

technology has got upper limits how much labels can get pushed on the label stack. In fact an existing tunnel (for example LDP tunnel {S, R1, R2} can be reused for certain path segments.

3.5. Link and Node Protection LSPs

In a network that is protecting nodes and links using IGP advertised labels, it is critical to perform fast restoration using local-repair, with packet forwarding restoration times comparable to RSVP Fast Re-Route (FRR) [RFC4090] or Loop Free Alternates [RFC5286].

First consider the timing of events assuming control-plane convergence as the sole repair mechanism. In Figure 5 a link failure scenario is illustrated. The best IGP path between {S,D} is {S, R3, R4, D}. When the directly adjacent link between R3 to R4 experiences a failure, (e.g.: fiber cut), the length of time to restore packet forwarding, from S to D, is dependent on several factors:

1. artificial (generation and pacing) delay of link-state updates
2. propagation delay of link-state updates
3. SPF throttling
4. programming forwarding state

The overall length of IGP convergence time, is largely dependent on the slowest router programming changed forwarding state. This is inherent unpredictable due to the CPU load and overall scheduling state in the affected systems, hence control-plane as the sole repair technology is ineffective. In contrast, local-repair technology helps to minimize transient packet loss. In local-repair technology a backup path is programmed ahead of time. Once the link fails a forwarding plane may immediately change forwarding state (= local-repair) to the backup path. This keeps the traffic flowing until the control-plane calculates and installs the new primary path and backup path tuples forwarding state for a given destination in the network.

In the below example, the IGP calculates using C-SPF and pre-establishes a FRR Bypass LSP along {R3, R1, R2, R4} to provide Link Protection of the R3 to R4 link. When that link fails, R3 will local-repair traffic along the {R3, R1, R2, R4} Bypass LSP while simultaneously signaling in the Control Plane to the Head-End LSR, S, that the R3 to R4 link has failed. This allows time for S to run C-SPF to calculate a new, optimal forwarding path around the link failure; signal a new LSP through intermediate LSRs; and, finally, S may perform "make-before-break" to start forwarding traffic on the new LSP.

Note that the algorithmic complexity of a single-destination C-SPF is much less compared to the the all-destination, per-neighbor forward SPF and per-neighbor reverse SPF a router doing Remote LFA [I-D.ietf-rtgwg-remote-lfa] calculations.

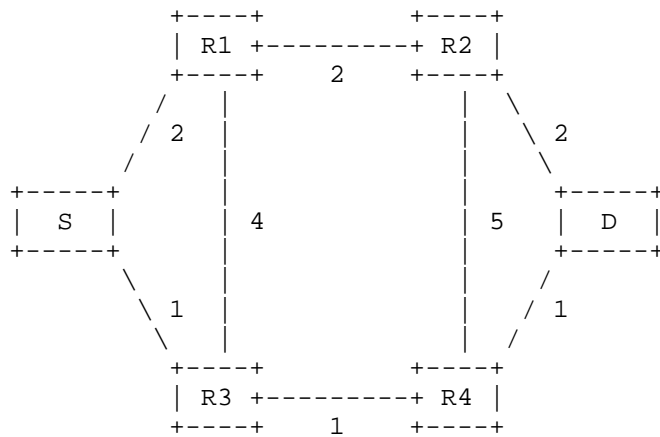


Figure 5: Protection LSPs using Label stacking

For construction of the Bypass LSP a constrained-SPF (C-SPF) calculation is commenced. The C-SPF calculation computes an alternative path to R4, without transiting the {R3, R4} link. Furthermore the backup path MUST not violate any SRLGs with respect to the {R3, R4} link. A possible backup path result for R3 is {R3, R1, R2, R4}.

Next R3 needs to construct the label stack for this particular Bypass LSP. Assume that each router along the Bypass LSP has advertised a label binding for reaching its direct neighbor.

- o R1: to R2, Label 102
- o R2: to R4, Label 204
- o R3: to R1, Label 301

Now R3 can construct the the label-stack fully describing the bypass LSP: For the last hop from R2 to R4, label 204 is pushed on the stack For the penultimate hop from R1 to R2, label 102 is pushed on the stack Since the first hop of the Bypass LSP is a local choice, there is no need to encode an actual label (label 301), but rather program a nexthop forwarding action to R1.

RSVP headends learn about all their bypasses using RESV messages.

When stacking IGP advertised labels, there is no direct comparable concept of a 'single head-end' node. All one-hop LSPs are in fact head-end nodes of their own and since there is no end-to-end signaling there is also no way about learning the bypasses that transit nodes have set up. IGP advertised labels hence mandate that all Bypass LSPs needs to be signaled to the rest of the network, such that the edge routers can have full insight (and control) what links may get utilized during local-repair. This is necessary, such that an edge router who may wants to enforce path policy constraints (e.g. end-to-end delay, hop count, path diversity, SRLG) can prefer or avoid certain paths (and their Bypasses) for path construction.

3.6. Stitching MPLS Label Switched Path Segments

One of the shortcomings of existing traffic-engineering solutions is that existing label switched paths cannot get advertised and shared by many ingress routers in the network. In the example network (Figure 6) a LSP with an ERO of {R4, R2, R6} has been established in order to utilize two unused north / south links. The only way to attract traffic to that LSP is to advertise the LSP as a forwarding adjacency. This causes loss of the original path information which might be interesting for a potential router which might wants to use this LSP for backup purposes. A computing router would need to have all underlying fate-sharing and bandwidth utilization information.

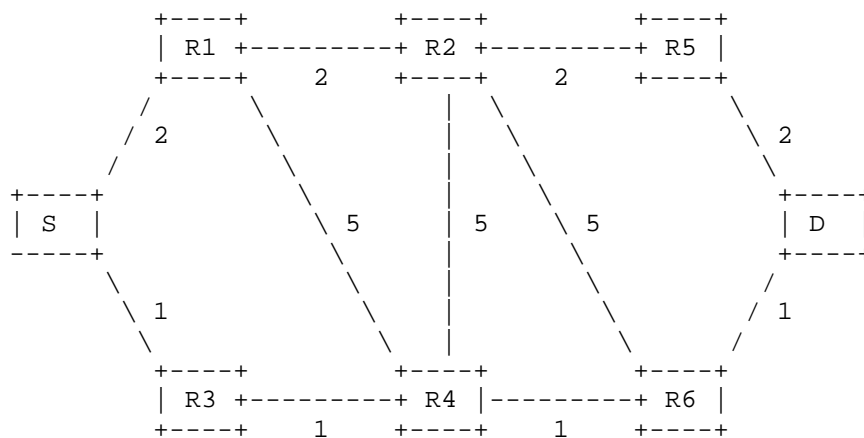


Figure 6: Advertising path segments

The IGP on R4 can now advertise the LSP segment by advertising its ingress label and optionally pass the original ERO, such that any upstream router can do their fate-sharing computations. Potential ingress routers now can use this LSP as a segment of the overall LSP. Furthermore ingress routers can combine label advertisements from

different routers along the path. For example router S could stack its LDP path to R2 {S, R1, R2} plus the IGP learned RSVP LSP {R4, R5, R6} plus a strict forwarding label {R6, D}.

3.7. T-LDP replacement for infrastructure labels

Consider Figure 7. There is a LSP {S, R1, R2, D} which seeks link-protection against failure of the {R1, R2} link using R-LFA.

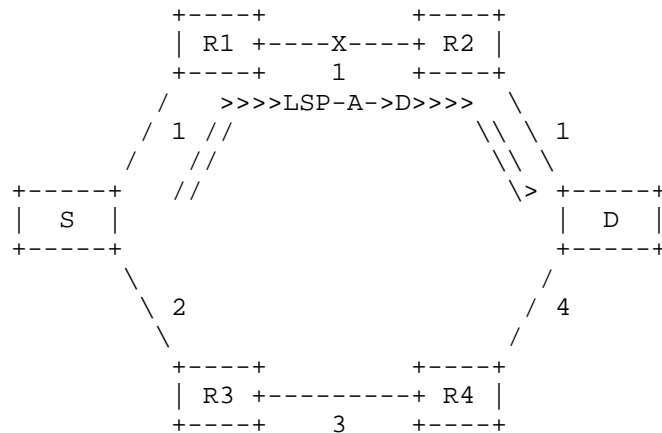


Figure 7: Avoidance of T-LDP for obtaining infrastructure labels

The Remote LFA Calculations results in the following Node sets.

- o Extended P set: {R4}
- o Q set: {R2, D, R4}
- o PQ set: {R4}

The PLR router (R1) needs to obtain the label-bindings from R4 towards the final destination D in order to push the two LSPs {R1, S, R3, R4} and {R4, D}. State of the art is to establish a targeted LDP session between PLR (R1) and the R-LFA Neighbor (R4). It would be desirable to avoid dynamic bringup of T-LDP sessions. Rather the IGP should supply the corresponding Label Bindings. Furthermore it would be desirable to apply some form of message compression, such that (unlike T-LDP) not per-FEC label bindings need to be exchanged. Applying Label Block style encoding [RFC4761] would be a suitable technology to compress the messaging overhead.

4. Acknowledgements

Many thanks to Yakov Rekhter, Ina Minei, Stephane Likowski and Bruno Decraene for their useful comments.

5. IANA Considerations

This memo includes no request to IANA.

6. Security Considerations

This document does not introduce any change in terms of IGP security. It simply proposes to flood existing information gathered from other protocols via the IGP.

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5151] Farrel, A., Ayyangar, A., and JP. Vasseur, "Inter-Domain MPLS and GMPLS Traffic Engineering -- Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 5151, February 2008.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.

- [RFC6571] Filsfils, C., Francois, P., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.

7.2. Informative References

- [I-D.bryant-ipfrr-tunnels]
Bryant, S., Filsfils, C., Previdi, S., and M. Shand, "IP Fast Reroute using tunnels", draft-bryant-ipfrr-tunnels-03 (work in progress), November 2007.
- [I-D.ietf-rtgwg-remote-lfa]
Bryant, S., Filsfils, C., Previdi, S., Shand, M., and S. Ning, "Remote LFA FRR", draft-ietf-rtgwg-remote-lfa-01 (work in progress), December 2012.
- [I-D.minto-2547-egress-node-fast-protection]
Jeganathan, J. and H. Gredler, "2547 egress PE Fast Failure Protection", draft-minto-2547-egress-node-fast-protection-01 (work in progress), October 2012.
- [NNHOP] Chen, E., Shen, N., and A. Tian, "Discovering LDP Next-Next-hop Labels", November 2005, <<http://tools.ietf.org/html/draft-shen-mpls-ldp-nnhop-label-02>>.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

Authors' Addresses

Hannes Gredler (editor)
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: hannes@juniper.net

Shane Amante
Level 3 Communications, Inc.
1025 Eldorado Blvd
Broomfield, CO 80021
US

Email: shane@level3.net

Tom Scholl
Amazon
Seattle, WA
US

Email: tscholl@amazon.com

Luay Jalil
Verizon
1201 E Arapaho Rd.
Richardson, TX 75081
US

Email: luay.jalil@verizon.com

RTGWG
Internet-Draft
Intended status: Informational
Expires: January 16, 2014

S. Ning
Tata Communications
D. McDysan
Verizon
E. Osborne
Cisco
L. Yong
Huawei USA
C. Villamizar
Outer Cape Cod Network
Consulting
July 15, 2013

Advanced Multipath Framework in MPLS
draft-ietf-rtgwg-cl-framework-04

Abstract

This document specifies a framework for support of Advanced Multipath in MPLS networks. As defined in this framework, an Advanced Multipath consists of a group of homogenous or non-homogenous links that have the same forward adjacency (FA) and can be considered as a single TE link or an IP link when advertised into IGP routing.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 16, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal

Provisions Relating to IETF Documents
(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	4
1.1. Background	4
1.2. Architecture Summary	4
1.3. Conventions used in this document	5
1.4. Terminology	5
1.5. Document Issues	5
2. Advanced Multipath Key Characteristics	7
2.1. Flow Identification	7
2.1.1. Flow Identification Granularity	8
2.1.2. Flow Identification Summary	9
2.1.3. Flow Identification Using Entropy Label	9
2.2. Advanced Multipath in Control Plane	10
2.3. Advanced Multipath in Data Plane	13
3. Architecture Tradeoffs	14
3.1. Scalability Motivations	14
3.2. Reducing Routing Information and Exchange	15
3.3. Reducing Signaling Load	15
3.3.1. Reducing Signaling Load using LDP MPTP	16
3.3.2. Reducing Signaling Load using Hierarchy	16
3.3.3. Using Both LDP MPTP and RSVP-TE Hierarchy	17
3.4. Reducing Forwarding State	17
3.5. Avoiding Route Oscillation	17
4. New Challenges	18
4.1. Control Plane Challenges	19
4.1.1. Delay and Jitter Sensitive Routing	19
4.1.2. Local Control of Traffic Distribution	20
4.1.3. Path Symmetry Requirements	20
4.1.4. Requirements for Contained LSP	21
4.1.5. Retaining Backwards Compatibility	21
4.2. Data Plane Challenges	22
4.2.1. Very Large LSP	22
4.2.2. Very Large Microflows	23
4.2.3. Traffic Ordering Constraints	23
4.2.4. Accounting for IP and LDP Traffic	23
4.2.5. IP and LDP Limitations	24
5. Existing Mechanisms	25

5.1.	Link Bundling	25
5.2.	Classic Multipath	26
6.	Mechanisms Proposed in Other Documents	27
6.1.	Loss and Delay Measurement	27
6.2.	Link Bundle Extensions	28
6.3.	Pseudowire Flow and MPLS Entropy Labels	28
6.4.	Multipath Extensions	29
7.	Required Protocol Extensions and Mechanisms	29
7.1.	Brief Review of Requirements	29
7.2.	Proposed Document Coverage	30
7.2.1.	Component Link Grouping	31
7.2.2.	Delay and Jitter Extensions	31
7.2.3.	Path Selection and Admission Control	32
7.2.4.	Dynamic Multipath Balance	32
7.2.5.	Frequency of Load Balance	33
7.2.6.	Inter-Layer Communication	33
7.2.7.	Packet Ordering Requirements	33
7.2.8.	Minimally Disruption Load Balance	34
7.2.9.	Path Symmetry	34
7.2.10.	Performance, Scalability, and Stability	35
7.2.11.	IP and LDP Traffic	35
7.2.12.	LDP Extensions	35
7.2.13.	Pseudowire Extensions	36
7.2.14.	Multi-Domain Advanced Multipath	36
7.3.	Framework Requirement Coverage by Protocol	36
7.3.1.	OSPF-TE and ISIS-TE Protocol Extensions	37
7.3.2.	PW Protocol Extensions	37
7.3.3.	LDP Protocol Extensions	37
7.3.4.	RSVP-TE Protocol Extensions	37
7.3.5.	RSVP-TE Path Selection Changes	37
7.3.6.	RSVP-TE Admission Control and Preemption	37
7.3.7.	Flow Identification and Traffic Balance	37
8.	IANA Considerations	38
9.	Security Considerations	38
10.	Acknowledgments	38
11.	References	39
11.1.	Normative References	39
11.2.	Informative References	39
	Authors' Addresses	42

1. Introduction

Advanced Multipath functional requirements are specified in [I-D.ietf-rtgwg-cl-requirement]. Advanced Multipath use cases are described in [I-D.ietf-rtgwg-cl-use-cases]. This document specifies a framework to meet these requirements.

This document describes an Advanced Multipath framework in the context of MPLS networks using an IGP-TE and RSVP-TE MPLS control plane with GMPLS extensions [RFC3209] [RFC3630] [RFC3945] [RFC5305].

Specific protocol solutions are outside the scope of this document, however a framework for the extension of existing protocols is provided. Backwards compatibility is best achieved by extending existing protocols where practical rather than inventing new protocols. The focus is on examining where existing protocol mechanisms fall short with respect to [I-D.ietf-rtgwg-cl-requirement] and on the types of extensions that will be required to accommodate functionality that is called for in [I-D.ietf-rtgwg-cl-requirement].

1.1. Background

Classic multipath, including Ethernet Link Aggregation has been widely used in today's MPLS networks [RFC4385][RFC4928]. Classic multipath using non-Ethernet links are often advertised using MPLS Link bundling. A link bundle [RFC4201] bundles a group of homogeneous links as a TE link to make IGP-TE information exchange and RSVP-TE signaling more scalable. An Advanced Multipath allows bundling non-homogenous links together as a single logical link.

An Advanced Multipath is a single logical link in MPLS network that contains multiple parallel component links between two MPLS LSR. Unlike a link bundle [RFC4201], the component links in an Advanced Multipath can have different properties such as cost, capacity, delay, or jitter.

1.2. Architecture Summary

Networks aggregate information, both in the control plane and in the data plane, as a means to achieve scalability. A tradeoff exists between the needs of scalability and the needs to identify differing path and link characteristics and differing requirements among flows contained within further aggregated traffic flows. These tradeoffs are discussed in detail in Section 3.

Some aspects of Advanced Multipath requirements present challenges for which multiple solutions may exist. In Section 4 various challenges and potential approaches are discussed.

A subset of the functionality called for in [I-D.ietf-rtgwg-cl-requirement] is available through MPLS Link Bundling [RFC4201]. Link bundling and other existing standards applicable to Advanced Multipath are covered in Section 5.

The most straightforward means of supporting Advanced Multipath requirements is to extend MPLS protocols and protocol semantics and in particular to extend link bundling. Extensions which have already been proposed in other documents which are applicable to Advanced Multipath are discussed in Section 6.

A goal of most new protocol work within IETF is to reuse existing protocol encapsulations and mechanisms where they meet requirements and extend existing mechanisms. This approach minimizes additional complexity while meeting requirements and tends to preserve backwards compatibility to the extent it is practical to do so. These goals are considered in proposing a framework for further protocol extensions and mechanisms in Section 7.

1.3. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

1.4. Terminology

Terminology defined in [I-D.ietf-rtgwg-cl-requirement] is used in this document. The additional terms defined in [I-D.ietf-rtgwg-cl-use-cases] are also used.

The abbreviation IGP-TE is used as a shorthand indicating either OSPF-TE [RFC3630] or ISIS-TE [RFC5305].

1.5. Document Issues

This subsection exists solely for the purpose of focusing the RTGWG meeting and mailing list discussions on areas within this document that need attention in order for the document to achieve the level of quality necessary to advance the document through the IETF process. This subsection will be removed before work group last call.

The following issues need to be resolved.

1. The feasibility of symmetric paths for all flows is questionable. The only case where this is practical is where LSP are smaller than component links and where classic link bundling (not using the all-ones component) is used. Perhaps the emphasis on this

(mis)feature should be reduced in the requirements document. See Section 4.1.3.

2. There is a tradeoff between supporting delay optimized routing and avoiding oscillation. This may be sufficiently covered, but a careful review by others and comments would be beneficial.
3. Any measurement of jitter (delay variation) that is used in route decision is likely to cause oscillation. Trying to optimize a path to reduce jitter may be a fools errand. How do we say this in the draft or does the existing text cover it adequately?
4. RTGWG needs to consider the possibility of using multi-topology IGP extensions in IP and LDP routing where the topologies reflect differing requirements (see Section 4.2.5). This idea is similar to TOS routing, which has been discussed for decades but has never been deployed. One possible outcome of discussion would be to declare TOS routing out of scope in the requirements document.
5. The following referenced drafts have expired:
 - A. [I-D.ospf-cc-stlv]
 - B. [I-D.villamizar-mppls-multipath-extn]

A replacement for [I-D.ospf-cc-stlv] is expected to be submitted. [I-D.villamizar-mppls-multipath-extn] is expected to emerge in a simplified form, removing extensions for which existing workarounds are considered adequate based on feedback at a prior IETF.
6. Clarification of what we intend to do with Multi-Domain Advanced Multipath is needed in Section 7.2.14.
7. The following topics in the requirements document are not addressed. Since they are explicitly mentioned in the requirements document some mention of how they are supported is needed in this document.
 - A. Migration (incremental deployment) may not be adequately covered in Section 4.1.5. It might also be necessary to say more here on performance, scalability, and stability as it related to migration. Comments on this from co-authors or the WG?
 - B. We may need a performance section in this document to specifically address #DR6 (fast convergence), and #DR7 (fast worst case failure convergence). We do already have

scalability discussion and make a recommendation for a separate document. At the very least the performance section would have to say "no worse than before, except were there was no alternative to make it very slightly worse" (in a bit more detail than that). It might also be helpful to better define the nature of the performance criteria implied by #DR6 and #DR7.

The above list has been in this document for the better part of a year with very little discussion (or none) of the above issues on the RTGWG mailing list.

2. Advanced Multipath Key Characteristics

[I-D.ietf-rtgwg-cl-requirement] defines external behavior of Advanced Multipath. The overall framework approach involves extending existing protocols in a backwards compatible manner and reusing ongoing work elsewhere in IETF where applicable, defining new protocols or semantics only where necessary. Given the requirements, and this approach of extending MPLS, Advanced Multipath key characteristics can be described in greater detail than given requirements alone.

2.1. Flow Identification

Traffic mapping to component links is a data plane operation. Control over how the mapping is done may be directly dictated or constrained by the control plane or by the management plane. When unconstrained by the control plane or management plane, distribution of traffic is entirely a local matter. Regardless of constraints or lack of constraints, the traffic distribution is required to keep packets belonging to individual flows in sequence and meet QoS criteria specified per LSP by either signaling or management [RFC2475] [RFC3260].

Key objectives of the traffic distribution are to not overload any component link, and to be able to perform local recovery when a subset of component links fails.

The network operator may have other objectives such as placing a bidirectional flow or LSP on the same component link in both direction, bounding delay and/or jitter, Advanced Multipath energy saving, and etc. These new requirements are described in [I-D.ietf-rtgwg-cl-requirement].

Examples of means to identify a flow may in principle include:

1. an LSP identified by an MPLS label,
2. a pseudowire (PW) [RFC3985] identified by an MPLS PW label,
3. a flow or group of flows within a pseudowire (PW) [RFC6391] identified by an MPLS flow label,
4. a flow or flow group in an LSP [RFC6790] identified by an MPLS entropy label,
5. all traffic between a pair of IP hosts, identified by an IP source and destination pair,
6. a specific connection between a pair of IP hosts, identified by an IP source and destination pair, protocol, and protocol port pair,
7. a layer-2 conversation within a pseudowire (PW), where the identification is PW payload type specific, such as Ethernet MAC addresses and VLAN tags within an Ethernet PW [RFC4448]. This is feasible but not practical (see below).

Although in principle a layer-2 conversation within a pseudowire (PW), may be identified by PW payload type specific information, in practice this is impractical at LSP midpoints when PW are carried. The PW ingress may provide equivalent information in a PW flow label [RFC6391]. Therefore, in practice, item #8 above is covered by [RFC6391] and may be dropped from the list.

2.1.1.1. Flow Identification Granularity

An LSR must at least be capable of identifying flows based on MPLS labels. Most MPLS LSP do not require that traffic carried by the LSP are carried in order. MPLS-TP is a recent exception. If it is assumed that no LSP require strict packet ordering of the LSP itself (only of flows within the LSP), then the entire label stack can be used as flow identification. If some LSP may require strict packet ordering but those LSP cannot be distinguished from others, then only the top label can be used as a flow identifier. If only the top label is used (for example, as specified by [RFC4201] when the "all-ones" component described in [RFC4201] is not used), then there may not be adequate flow granularity to accomplish well balanced traffic distribution and it will not be possible to carry LSP that are larger than any individual component link.

The number of flows can be extremely large. This may be the case when the entire label stack is used and is always the case when IP addresses are used in provider networks carrying Internet traffic.

Current practice for native IP load balancing at the time of writing were documented in [RFC2991] and [RFC2992]. These practices as described, make use of IP addresses.

The common practices described in [RFC2991] and [RFC2992] were extended to include the MPLS label stack and the common practice of looking at IP addresses within the MPLS payload. These extended practices require that pseudowires use a PWE3 Control Word and are described in [RFC4385] and [RFC4928]. Additional detail on current multipath practices can be found in the appendices of [I-D.ietf-rtgwg-cl-use-cases].

Using only the top label supports too coarse a traffic balance. Prior to MPLS Entropy Label [RFC6790] using the full label stack was also too coarse. Using the full label stack and IP addresses as flow identification provides a sufficiently fine traffic balance, but is capable of identifying such a high number of distinct flows, that a technique of grouping flows, such as hashing on the flow identification criteria, becomes essential to reduce the stored state, and is an essential scaling technique. Other means of grouping flows may be possible.

2.1.2. Flow Identification Summary

In summary:

1. Load balancing using only the MPLS label stack provides too coarse a granularity of load balance.
2. Tracking every flow is not scalable due to the extremely large number of flows in provider networks.
3. Existing techniques, IP source and destination hash in particular, have proven in over two decades of experience to be an excellent way of identifying groups of flows.
4. If a better way to identify groups of flows is discovered, then that method can be used.
5. IP address hashing is not required, but use of this technique is strongly encouraged given the technique's long history of successful deployment.

2.1.3. Flow Identification Using Entropy Label

MPLS Entropy Label [RFC6790] provides a means of making use of the entropy from information that would require deeper packet inspection, such as inspection of IP addresses, and putting that entropy in the

form of a hashed value into the label stack. Midpoint LSR that understand the Entropy Label Indicator can make use of only label stack information but still obtain a fine load balance granularity.

2.2. Advanced Multipath in Control Plane

An Advanced Multipath is advertised as a single logical interface between two connected routers, which forms forwarding adjacency (FA) between the routers. The FA is advertised as a TE-link in a link state IGP, using either OSPF-TE or ISIS-TE. The IGP-TE advertised interface parameters for the Advanced Multipath can be preconfigured by the network operator or be derived from its component links. Advanced Multipath advertisement requirements are specified in [I-D.ietf-rtgwg-cl-requirement].

In IGP-TE, an Advanced Multipath is advertised as a single TE link between two connected routers. This is similar to a link bundle [RFC4201]. Link bundle applies to a set of homogenous component links. Advanced Multipath allows homogenous and non-homogenous component links. Due to the similarity, and for backwards compatibility, extending link bundling is viewed as both simple and as the best approach.

In order for a route computation engine to calculate a proper path for a LSP, it is necessary for Advanced Multipath to advertise the summarized available bandwidth as well as the maximum bandwidth that can be made available for single flow (or single LSP where no finer flow identification is available). If an Advanced Multipath contains some non-homogeneous component links, the Advanced Multipath also should advertise the summarized bandwidth and the maximum bandwidth for single flow per each homogeneous component link group.

Both LDP [RFC5036] and RSVP-TE [RFC3209] can be used to signal a LSP over an Advanced Multipath. LDP cannot be extended to support traffic engineering capabilities [RFC3468].

When an LSP is signaled using RSVP-TE, the LSP MUST be placed on the component link that meets the LSP criteria indicated in the signaling message.

When an LSP is signaled using LDP, the LSP MUST be placed on the component link that meets the LSP criteria, if such a component link is available. LDP does not support traffic engineering capabilities, imposing restrictions on LDP use of Advanced Multipath. See Section 4.2.5 for further details.

If the Advanced Multipath solution is based on extensions to IGP-TE and RSVP-TE, then in order to meet requirements defined in

[I-D.ietf-rtgwg-cl-requirement], the following derived requirements MUST be met.

1. An Advanced Multipath MAY contain non-homogeneous component links. The route computing engine MAY select one group of component links for a LSP. The The route computing engine MUST accommodate service objectives for a given LSP when selecting a group of component links for a LSP.
2. The routing protocol MUST make a grouping of component links available in the TE-LSDB, such that within each group all of the component links have similar characteristics (the component links are homogeneous within a group).
3. The route computation used in RSVP-TE MUST be extended to include only the capacity of groups within an Advanced Multipath which meet LSP criteria.
4. The signaling protocol MUST be able to indicate either the criteria, or which groups may be used.
5. An Advanced Multipath MUST place each LSP on a component link or group which meets or exceeds the LSP criteria.

Advanced Multipath capacity is aggregated capacity. LSP capacity MAY be larger than individual component link capacity. Any aggregated LSP can determine a bounds on the largest microflow that could be carried and this constraint can be handled as follows.

1. If no information is available through signaling, management plane, or configuration, the largest microflow is bound by one of the following:
 - A. the largest single LSP if most traffic is RSVP-TE signaled and further aggregated,
 - B. the largest pseudowire if most traffic is carrying pseudowire payloads that are aggregated within RSVP-TE LSP,
 - C. or the largest interface or component link capacity carrying IP or LDP if a large amount of IP or LDP traffic is contained within the aggregate.

If a very large amount of traffic being aggregated is IP or LDP, then the largest microflow is bound by the largest component link on which IP traffic can arrive. For example, if an LSR is acting as an LER and IP and LDP traffic is arriving on 10 Gb/s edge interfaces, then no microflow larger than 10 Gb/s will be present

on the RSVP-TE LSP that aggregate traffic across the core, even if the core interfaces are 100 Gb/s interfaces.

2. The prior conditions provide a bound on the largest microflow when no signaling extensions indicate a bounds. If an LSP is aggregating smaller LSP for which the largest expected microflow carried by the smaller LSP is signaled, then the largest microflow expected in the containing LSP (the aggregate) is the maximum of the largest expected microflow for any contained LSP. For example, RSVP-TE LSP may be large but aggregate traffic for which the source or sink are all 1 Gb/s or smaller interfaces (such as in mobile applications in which cell sites backhauls are no larger than 1 Gb/s). If this information is carried in the LSP originated at the cell sites, then further aggregates across a core may make use of this information.
3. The IGP must provide the bounds on the largest microflow that an Advanced Multipath can accommodate, which is the maximum capacity on a component link that can be made available by moving other traffic. This information is needed by the ingress LER for path determination.
4. A means to signal an LSP whose capacity is larger than individual component link capacity is needed [I-D.ietf-rtgwg-cl-requirement] and also signal the largest microflow expected to be contained in the LSP. If a bounds on the largest microflow is not signaled there is no means to determine if an LSP which is larger than any component link can be subdivided into flows and therefore should be accepted by admission control.

When a bidirectional LSP request is signaled over an Advanced Multipath, if the request indicates that the LSP must be placed on the same component link, the routers of the Advanced Multipath MUST place the LSP traffic in both directions on a same component link. This is particularly challenging for aggregated capacity which makes use of the label stack for traffic distribution. The two requirements are mutually exclusive for any one LSP. No one LSP may be both larger than any individual component link and require symmetrical paths for every flow. Both requirements can be accommodated by the same Advanced Multipath for different LSP, with any one LSP requiring no more than one of these two features.

Individual component link may fail independently. Upon component link failure, an Advanced Multipath MUST support a minimally disruptive local repair, preempting any LSP which can no longer be supported. Available capacity in other component links MUST be used to carry impacted traffic. The available bandwidth after failure MUST be advertised immediately to avoid looped crankback.

When an Advanced Multipath is not able to transport all flows, it preempts some flows based upon holding priority and informs the control plane of these preempted flows. To minimize impact on traffic, the Advanced Multipath MUST support soft preemption [RFC5712]. The network operator SHOULD enable soft preemption. This action ensures the remaining traffic is transported properly. FR#10 requires that the traffic be restored. FR#12 requires that any change be minimally disruptive. These two requirements are interpreted to include preemption among the types of changes that must be minimally disruptive.

2.3. Advanced Multipath in Data Plane

The data plane must identify groups of flows. Flow identification is covered in Section 2.1. Having identified groups of flows the groups must be placed on individual component links. This step following flow group identification is called traffic distribution or traffic placement. The two steps together are known as traffic balancing or load balancing.

Traffic distribution may be determined by or constrained by control plane or management plane. Traffic distribution may be changed due to component link status change, subject to constraints imposed by either the management plane or control plane. The distribution function is local to the routers in which an Advanced Multipath belongs to and its implementation is not specified here.

When performing traffic placement, an Advanced Multipath does not differentiate multicast traffic vs. unicast traffic.

In order to maintain scalability, existing data plane forwarding retains state associated with the top label only. Using UHP (UHP is the absence of the more common PHP), zero or more labels may be POPed and packet and byte counters incremented prior to processing what becomes the top label after the POP operations are completed. Flow group identification may be a parallel step in the forwarding process. Data plane forwarding makes use of the top label to select an Advanced Multipath, or a group of components within an Advanced Multipath or for the case where an LSP is pinned (see [RFC4201]), a specific component link. For those LSP for which the LSP selects only the Advanced Multipath or a group of components within an Advanced Multipath, the load balancing makes use of the set of component links selected based on the top label, and makes use of the flow group identification to select among that group.

The simplest traffic placement techniques uses a modulo operation after computing a hash. This techniques has significant disadvantages. The most common traffic placement techniques uses the

a flow group identification as an index into a table. The table provides an indirection. The number of bits of hash is constrained to keep table size small. While this is not the best technique, it is the most common. Better techniques exist but they are outside the scope of this document and some are considered proprietary.

Requirements to limit frequency of load balancing can be adhered to by keeping track of when a flow group was last moved and imposing a minimum period before that flow group can be moved again. This is straightforward for a table approach. For other approaches it may be less straightforward.

3. Architecture Tradeoffs

Scalability and stability are critical considerations in protocol design where protocols may be used in a large network such as today's service provider networks. Advanced Multipath is applicable to networks which are large enough to require that traffic be split over multiple paths. Scalability is a major consideration for networks that reach a capacity large enough to require Advanced Multipath.

Some of the requirements of Advanced Multipath could potentially have a negative impact on scalability. This section is about architectural tradeoffs, many motivated by the need to maintain scalability and stability, a need which is reflected in [I-D.ietf-rtgwg-cl-requirement], specifically in DR#6 and DR#7.

3.1. Scalability Motivations

In the interest of scalability, information is aggregated in situations where information about a large amount of network capacity or a large amount of network demand provides is adequate to meet requirements. Routing information is aggregated to reduce the amount of information exchange related to routing and to simplify route computation (see Section 3.2).

In an MPLS network large routing changes can occur when a single fault occurs. For example, a single fault may impact a very large number of LSP traversing a given link. As new LSP are signaled to avoid the fault, resources are consumed elsewhere, and routing protocol announcements must flood the resource changes. If protection is in place, there is less urgency to converging quickly. If multiple faults occur that are not covered by shared risk groups (SRG), then some protection may fail, adding urgency to converging quickly even where protection is deployed.

Reducing the amount of information allows the exchange of information

during a large routing change to be accomplished more quickly and simplifies route computation. Simplifying route computation improves convergence time after very significant network faults which cannot be handled by preprovisioned or precomputed protection mechanisms. Aggregating smaller LSP into larger LSP is a means to reduce path computation load and reduce RSVP-TE signaling (see Section 3.3).

Neglecting scaling issues can result in performance issues, such as slow convergence. Neglecting scaling in some cases can result in networks which perform so poorly as to become unstable.

3.2. Reducing Routing Information and Exchange

Link bundling provides a means of aggregating control plane information. Even where the all-ones component link supported by link bundling is not used, the amount of control information is reduced by the number of component links in a bundle.

Fully deaggregating link bundle information would negate this benefit. If there is a need to deaggregate, such as to distinguish between groups of links within specified ranges of delay, then no more deaggregation than is necessary should be done.

For example, in supporting the requirement for heterogeneous component links, it makes little sense to fully deaggregate link bundles when adding support for groups of component links with common attributes within a link bundle can maintain most of the benefit of aggregation while adequately supporting the requirement to support heterogeneous component links.

Routing information exchange is also reduced by making sensible choices regarding the amount of change to link parameters that require link readvertisement. For example, if delay measurements include queuing delay, then a much more coarse granularity of delay measurement would be called for than if the delay does not include queuing and is dominated by geographic delay (speed of light delay).

3.3. Reducing Signaling Load

Aggregating traffic into very large hierarchical LSP in the core very substantially reduces the number of LSP that need to be signaled and the number of path computations any given LSR will be required to perform when a network fault occurs.

In the extreme, applying MPLS to a very large network without hierarchy could exceed the 20 bit label space. For example, in a network with 4,000 nodes, with 2,000 on either side of a cutset, would have 4,000,000 LSP crossing the cutset. Even in a degree four

cutset, an uneven distribution of LSP across the cutset, or the loss of one link would result in a need to exceed the size of the label space. Among provider networks, 4,000 access nodes is not at all large. Hierarchy is an absolute requirement if all access nodes were interconnected in such a network.

In less extreme cases, having each node terminate hundreds of LSP to achieve a full mesh creates a very large computational load. Computational complexity is a function of the number of nodes (N) and links (L) in a topology, and the number of LSP that need to be set up. In the common case where L is proportional to N (relatively constant node degree with growth), the time complexity of one CSPF computation is $\text{order}(N \log N)$. If each node must perform $\text{order}(N)$ computations when a fault occurs, then the computational load increases as $\text{order}(N^2 \log N)$ as the number of nodes increases (where $^$ is the power of operator and N^2 is read "N-squared"). In practice at the time of writing, this imposes a limit of a few hundred nodes in a full mesh of MPLS LSP before the computational load is sufficient to result in unacceptable convergence times.

Two solutions are applied to reduce the amount of RSVP-TE signaling. Both involve subdividing the MPLS domain into a core and a set of regions.

3.3.1. Reducing Signaling Load using LDP MPTP

LDP can be used for edge-to-edge LSP, using RSVP-TE to carry the LDP intra-core traffic and also optionally also using RSVP-TE to carry the LDP intra-region traffic within each region. LDP does not support traffic engineering, but does support multipoint-to-point (MPTP) LSP, which require less signaling than edge-to-edge RSVP-TE point-to-point (PTP) LSP. A drawback of this approach is the inability to use RSVP-TE protection (FRR or GMPLS protection) against failure of the border LSR sitting at a core/region boundary.

3.3.2. Reducing Signaling Load using Hierarchy

When the number of nodes grows too large, the amount of RSVP-TE signaling can be reduced using the MPLS PSC hierarchy [RFC4206]. A core within the hierarchy can divide the topology into M regions of on average N/M nodes. Within a region the computational load is reduced by more than M^2 . Within the core, the computational load generally becomes quite small since M is usually a fairly small number (a few tens of regions) and each region is generally attached to the core in typically only two or three places on average.

Using hierarchy improves scaling but has two consequences. First, hierarchy effectively forces the use of platform label space. When a

containing LSP is rerouted, the labels assigned to the contained LSP cannot be changed but may arrive on a different interface. Second, hierarchy results in much larger LSP. These LSP today are larger than any single component link and therefore force the use of the all-ones component in link bundles.

3.3.3. Using Both LDP MPTP and RSVP-TE Hierarchy

It is also possible to use both LDP and RSVP-TE hierarchy. MPLS networks with a very large number of nodes may benefit from the use of both LDP and RSVP-TE hierarchy. The two techniques are certainly not mutually exclusive.

3.4. Reducing Forwarding State

Both LDP and MPLS hierarchy have the benefit of reducing the amount of forwarding state. Using the example from Section 3.3, and using MPLS hierarchy, the worst case generally occurs at borders with the core.

For example, consider a network with approximately 1,000 nodes divided into 10 regions. At the edges, each node requires 1,000 LSP to other edge nodes. The edge nodes also require 100 intra-region LSP. Within the core, if the core has only 3 attachments to each region the core LSR have less than 100 intra-core LSP. At the border cutset between the core and a given region, in this example there are 100 edge nodes with inter-region LSP crossing that cutset, destined to 900 other edge nodes. That yields forwarding state for on the order of 90,000 LSP at the border cutset. These same routers need only reroute well under 200 LSP when a multiple fault occurs, as long as only links are affected and a border LSR does not go down.

Interior to the core, the forwarding state is greatly reduced. If inter-region LSP have different characteristics, it makes sense to make use of aggregates with different characteristics. Rather than exchange information about every inter-region LSP within the intra-core LSP it makes more sense to use multiple intra-core LSP between pairs of core nodes, each aggregating sets of inter-region LSP with common characteristics or common requirements.

3.5. Avoiding Route Oscillation

Networks can become unstable when a feedback loop exists such that moving traffic to a link causes a metric such as delay to increase, which then causes traffic to move elsewhere. For example, the original ARPANET routing used a delay based cost metric and proved prone to route oscillations [DBP].

Delay may be used as a constraint in routing for high priority traffic, when this high priority traffic makes a minor contribution to total load, such that the movement of the high priority traffic has a small impact on the delay experienced by other high priority traffic. The safest way to measure delay is to make measurements based on traffic which is prioritized such that it is queued ahead of the lower priority traffic which will be affected if high priority traffic is moved. The amount of high priority traffic must be constrained to consume a fraction of link capacities with the remaining capacity available to lower priority traffic.

Any measurement of jitter (delay variation) that is used in route decision is likely to cause oscillation. Jitter that is caused by queuing effects and cannot be measured using a very high priority measurement traffic flow.

It may be possible to find links with constrained queuing delay or jitter using a theoretical maximum or a probability based bound on queuing delay or jitter at a given priority based on the types and amounts of traffic accepted and combining that theoretical limit with a measured delay at very high priority. Using delay or jitter as path metrics without creating oscillations is challenging.

Instability can occur due to poor performance and interaction with protocol timers. In this way a computational scaling problem can become a stability problem when a network becomes sufficiently large.

4. New Challenges

New technical challenges are posed by [I-D.ietf-rtgwg-cl-requirement] in both the control plane and data plane.

Among the more difficult challenges are the following.

1. The requirements related to delay or jitter conflict with requirements for scalability and stability (see Section 4.1.1),
2. The combination of ingress control over LSP placement and retaining an ability to move traffic as demands dictate can pose challenges and such requirements can even be conflicting (see Section 4.1.2),
3. Path symmetry requires extensions and is particularly challenging for very large LSP (see Section 4.1.3),
4. Accommodating a very wide range of requirements among contained LSP can lead to inefficiency if the most stringent requirements

are reflected in aggregates, or reduce scalability if a large number of aggregates are used to provide a too fine a reflection of the requirements in the contained LSP (see Section 4.1.4),

5. Backwards compatibility is somewhat limited due to the need to accommodate legacy multipath interfaces which provide too little information regarding their configured default behavior, and legacy LSP which provide too little information regarding their LSP requirements (see Section 4.1.5),
6. Data plane challenges include those of accommodating very large LSP, large microflows, traffic ordering constraints imposed by a subset of LSP, and accounting for IP and LDP traffic (see Section 4.2).

4.1. Control Plane Challenges

Some of the control plane requirements are particularly challenging. Handling large flows which aggregate smaller flows must be accomplished with minimal impact on scalability. Potentially conflicting are requirements for jitter and requirements for stability. Potentially conflicting are the requirements for ingress control of a large number of parameters, and the requirements for local control needed to achieve traffic balance across an Advanced Multipath. These challenges and potential solutions are discussed in the following sections.

4.1.1. Delay and Jitter Sensitive Routing

Delay and jitter sensitive routing are called for in [I-D.ietf-rtgwg-cl-requirement] in requirements FR#2, FR#7, FR#8, FR#9, FR#15, FR#16, FR#17, FR#18. Requirement FR#17 is particularly problematic, calling for constraints on jitter.

A tradeoff exists between scaling benefits of aggregating information, and potential benefits of using a finer granularity in delay reporting. To maintain the scaling benefit, measured link delay for any given Advanced Multipath SHOULD be aggregated into a small number of delay ranges. IGP-TE extensions MUST be provided which advertise the available capacities for each of the selected ranges.

For path selection of delay sensitive LSP, the ingress SHOULD bias link metrics based on available capacity and select a low cost path which meets LSP total path delay criteria. To communicate the requirements of an LSP, the ERO MUST be extended to indicate the per link constraints. To communicate the type of resource used, the RRO SHOULD be extended to carry an identification of the group that is

used to carry the LSP at each link bundle hop.

4.1.2. Local Control of Traffic Distribution

Many requirements in [I-D.ietf-rtgwg-cl-requirement] suggest that a node immediately adjacent to a component link should have a high degree of control over how traffic is distributed, as long as network performance objectives are met. Particularly relevant are FR#18 and FR#19.

The requirements to allow local control are potentially in conflict with requirement FR#21 which gives full control of component link select to the LSP ingress. While supporting this capability is mandatory, use of this feature is optional per LSP.

A given network deployment will have to consider this set of conflicting requirements and make appropriate use of local control of traffic placement and ingress control of traffic placement to best meet network requirements.

4.1.3. Path Symmetry Requirements

Requirement FR#21 in [I-D.ietf-rtgwg-cl-requirement] includes a provision to bind both directions of a bidirectional LSP to the same component. This is easily achieved if the LSP is directly signaled across an Advanced Multipath. This is not as easily achieved if a set of LSP with this requirement are signaled over a large hierarchical LSP which is in turn carried over an Advanced Multipath. The basis for load distribution in such a case is the label stack. The labels in either direction are completely independent.

This could be accommodated if the ingress, egress, and all midpoints of the hierarchical LSP make use of an entropy label in the distribution, and the ingress use a fixed value per contained LSP in the entropy label. A solution for this problem may add complexity with very little benefit. There is little or no true benefit of using symmetrical paths rather than component links of identical characteristics.

Traffic symmetry and large LSP capacity are a second pair of conflicting requirements. Any given LSP can meet one of these two requirements but not both. A given network deployment will have to make appropriate use of each of these features to best meet network requirements.

4.1.4. Requirements for Contained LSP

[I-D.ietf-rtgwg-cl-requirement] calls for new LSP constraints. These constraints include frequency of load balancing rearrangement, delay and jitter, packet ordering constraints, and path symmetry.

When LSP are contained within hierarchical LSP, there is no signaling available at midpoint LSR which identifies the contained LSP let alone providing the set of requirements unique to each contained LSP. Defining extensions to provide this information would severely impact scalability and defeat the purpose of aggregating control information and forwarding information into hierarchical LSP. For the same scalability reasons, not aggregating at all is not a viable option for large networks where scalability and stability problems may occur as a result.

As pointed out in Section 4.1.3, the benefits of supporting symmetric paths among LSP contained within hierarchical LSP may not be sufficient to justify the complexity of supporting this capability.

A scalable solution which accommodates multiple sets of LSP between given pairs of LSR is to provide multiple hierarchical LSP for each given pair of LSR, each hierarchical LSP aggregating LSP with common requirements and a common pair of endpoints. This is a network design technique available to the network operator rather than a protocol extension. This technique can accommodate multiple sets of delay and jitter parameters, multiple sets of frequency of load balancing parameters, multiple sets of packet ordering constraints, etc.

4.1.5. Retaining Backwards Compatibility

Backwards compatibility and support for incremental deployment requires considering the impact of legacy LSR in the role of LSP ingress, and considering the impact of legacy LSR advertising ordinary links, advertising Ethernet LAG as ordinary links, and advertising link bundles.

Legacy LSR in the role of LSP ingress cannot signal requirements which are not supported by their control plane software. The additional capabilities supported by other LSR has no impact on these LSR. These LSR however, being unaware of extensions, may try to make use of scarce resources which support specific requirements such as low delay. To a limited extent it may be possible for a network operator to avoid this issue using existing mechanisms such as link administrative attributes and attribute affinities [RFC3209].

Legacy LSR advertising ordinary links will not advertise attributes

needed by some LSP. For example, there is no way to determine the delay or jitter characteristics of such a link. Legacy LSR advertising Ethernet LAG pose additional problems. There is no way to determine that packet ordering constraints would be violated for LSP with strict packet ordering constraints, or that frequency of load balancing rearrangement constraints might be violated.

Legacy LSR advertising link bundles have no way to advertise the configured default behavior of the link bundle. Some link bundles may be configured to place each LSP on a single component link and therefore may not be able to accommodate an LSP which requires bandwidth in excess of the size of a component link. Some link bundles may be configured to spread all LSP over the all-ones component. For LSR using the all-ones component link, there is no documented procedure for correctly setting the "Maximum LSP Bandwidth". There is currently no way to indicate the largest microflow that could be supported by a link bundle using the all-ones component link.

Having received the RRO, it is possible for an ingress to look for the all-ones component to identify such link bundles after having signaled at least one LSP. Whether any LSR collects this information on legacy LSR and makes use of it to set defaults, is an implementation choice.

4.2. Data Plane Challenges

Flow identification is briefly discussed in Section 2.1. Traffic distribution is briefly discussed in Section 2.3. This section discusses issues specific to particular requirements specified in [I-D.ietf-rtgwg-cl-requirement].

4.2.1. Very Large LSP

Very large LSP may exceed the capacity of any single component of an Advanced Multipath. In some cases contained LSP may exceed the capacity of any single component. These LSP may make use of the equivalent of the all-ones component of a link bundle, or may use a subset of components which meet the LSP requirements.

Very large LSP can be accommodated as long as they can be subdivided (see Section 4.2.2). A very large LSP cannot have a requirement for symmetric paths unless complex protocol extensions are proposed (see Section 2.2 and Section 4.1.3).

4.2.2. Very Large Microflows

Within a very large LSP there may be very large microflows. A very large microflow is one which cannot be further subdivided and contributes a very large amount of capacity. Flows which cannot be subdivided must be no larger than the capacity of any single component link.

Current signaling provides no way to specify the largest microflow that can be supported on a given link bundle in routing advertisements. Extensions which address this are discussed in Section 6.4. Absent extensions of this type, traffic containing microflows that are too large for a given Advanced Multipath may be present. There is no data plane solution for this problem that would not require reordering traffic at the Advanced Multipath egress.

Some techniques are susceptible to statistical collisions where an algorithm to distribute traffic is unable to disambiguate traffic among two or more very large microflow where their sum is in excess of the capacity of any single component. Hash based algorithms which use too small a hash space are particularly susceptible and require a change in hash seed in the event that this were to occur. A change in hash seed is highly disruptive, causing traffic reordering among all traffic flows over which the hash function is applied.

4.2.3. Traffic Ordering Constraints

Some LSP have strict traffic ordering constraints. Most notable among these are MPLS-TP LSP. In the absence of aggregation into hierarchical LSP, those LSP with strict traffic ordering constraints can be placed on individual component links if there is a means of identifying which LSP have such a constraint. If LSP with strict traffic ordering constraints are aggregated in hierarchical LSP, the hierarchical LSP capacity may exceed the capacity of any single component link. In such a case the load balancing may be constrained through the use of an entropy label [RFC6790]. This and related issues are discussed further in Section 6.4.

4.2.4. Accounting for IP and LDP Traffic

Networks which carry RSVP-TE signaled MPLS traffic generally carry low volumes of native IP traffic, often only carrying control traffic as native IP. There is no architectural guarantee of this, it is just how network operators have made use of the protocols.

[I-D.ietf-rtgwg-cl-requirement] requires that native IP and native LDP be accommodated (DR#2 and DR#3). In some networks, a subset of services may be carried as native IP or carried as native LDP. Today

this may be accommodated by the network operator estimating the contribution of IP and LDP and configuring a lower set of available bandwidth figures on the RSVP-TE advertisements.

The only improvement that Advanced Multipath can offer is that of measuring the IP and LDP traffic levels and automatically reducing the available bandwidth figures on the RSVP-TE advertisements. The measurements would have to be filtered. This is similar to a feature in existing LSR, commonly known as "autobandwidth" with a key difference. In the "autobandwidth" feature, the bandwidth request of an RSVP-TE signaled LSP is adjusted in response to traffic measurements. In this case the IP or LDP traffic measurements are used to reduce the link bandwidth directly, without first encapsulating in an RSVP-TE LSP.

This may be a subtle and perhaps even a meaningless distinction if Advanced Multipath is used to form a Sub-Path Maintenance Element (SPME). A SPME is in practice essentially an unsignaled single hop LSP with PHP enabled [RFC5921]. An Advanced Multipath SPME looks very much like classic multipath, where there is no signaling, only management plane configuration creating the multipath entity (of which Ethernet Link Aggregation is a subset).

4.2.5. IP and LDP Limitations

IP does not offer traffic engineering. LDP cannot be extended to offer traffic engineering [RFC3468]. Therefore there is no traffic engineered fallback to an alternate path for IP and LDP traffic if resources are not adequate for the IP and/or LDP traffic alone on a given link in the primary path. The only option for IP and LDP would be to declare the link down. Declaring a link down due to resource exhaustion would reduce traffic to zero and eliminate the resource exhaustion. This would cause oscillations and is therefore not a viable solution.

Congestion caused by IP or LDP traffic loads is a pathologic case that can occur if IP and/or LDP are carried natively and there is a high volume of IP or LDP traffic. This situation can be avoided by carrying IP and LDP within RSVP-TE LSP.

It is also not possible to route LDP traffic differently for different FEC. LDP traffic engineering is specifically disallowed by [RFC3468]. It may be possible to support multi-topology IGP extensions to accommodate more than one set of criteria. If so, the additional IGP could be bound to the forwarding criteria, and the LDP FEC bound to a specific IGP instance, inheriting the forwarding criteria. Alternately, one IGP instance can be used and the LDP SPF can make use of the constraints, such as delay and jitter, for a

given LDP FEC.

5. Existing Mechanisms

In MPLS the one mechanism which supports explicit signaling of multiple parallel links is Link Bundling [RFC4201]. The set of techniques known as "classis multipath" support no explicit signaling, except in two cases. In Ethernet Link Aggregation the Link Aggregation Control Protocol (LACP) coordinates the addition or removal of members from an Ethernet Link Aggregation Group (LAG). The use of the "all-ones" component of a link bundle indicates use of classis multipath, however the ability to determine if a link bundle makes use of classis multipath is not yet supported.

5.1. Link Bundling

Link bundling supports advertisement of a set of homogenous links as a single route advertisement. Link bundling supports placement of an LSP on any single component link, or supports placement of an LSP on the all-ones component link. Not all link bundling implementations support the all-ones component link. There is no way for an ingress LSR to tell which potential midpoint LSR support this feature and use it by default and which do not. Based on [RFC4201] it is unclear how to advertise a link bundle for which the all-ones component link is available and used by default. Common practice is to violate the specification and set the Maximum LSP Bandwidth to the Available Bandwidth. There is no means to determine the largest microflow that could be supported by a link bundle that is using the all-ones component link.

[RFC6107] extends the procedures for hierarchical LSP but also extends link bundles. An LSP can be explicitly signaled to indicate that it is an LSP to be used as a component of a link bundle. Prior to that the common practice was to simply not advertise the component link LSP into the IGP, since only the ingress and egress of the link bundle needed to be aware of their existence, which they would be aware of due to the RSVP-TE signaling used in setting up the component LSP.

While link bundling can be the basis for Advanced Multipath, a significant number of small extension needs to be added.

1. To support link bundles of heterogeneous links, a means of advertising the capacity available within a group of homogeneous links needs to be provided.

2. Attributes need to be defined to support the following parameters for the link bundle or for a group of homogeneous links.
 - A. delay range
 - B. jitter (delay variation) range
 - C. group metric
 - D. all-ones component capable
 - E. capable of dynamically balancing load
 - F. largest supportable microflow
 - G. support for entropy label
3. For each of the prior extended attributes, the constraint based routing path selection needs to be extended to reflect new constraints based on the extended attributes.
4. For each of the prior extended attributes, LSP admission control needs to be extended to reflect new constraints based on the extended attributes.
5. Dynamic load balance must be provided for flows within a given set of links with common attributes such that Performance Objectives are not violated including frequency of load balance adjustment for any given flow.

5.2. Classic Multipath

Classic multipath is described in [I-D.ietf-rtgwg-cl-use-cases].

Classic multipath refers to the most common current practice in implementation and deployment of multipath. The most common current practice makes use of a hash on the MPLS label stack and if IPv4 or IPv6 are indicated under the label stack, makes use of the IP source and destination addresses [RFC4385] [RFC4928].

Classic multipath provides a highly scalable means of load balancing. Dynamic multipath has proven value in assuring an even loading on component link and an ability to adapt to change in offered load that occurs over periods of hundreds of milliseconds or more. Classic multipath scalability is due to the ability to effectively work with an extremely large number of flows (IP host pairs) using relatively little resources (a data structure accessed using a hash result as a key or using ranges of hash results).

Classic multipath meets a small subset of Advanced Multipath requirements. Due to scalability of the approach, classic multipath seems to be an excellent candidate for extension to meet the full set of Advanced Multipath forwarding requirements.

Additional detail can be found in [I-D.ietf-rtgwg-cl-use-cases].

6. Mechanisms Proposed in Other Documents

A number of documents which at the time of writing are works in progress address parts of the requirements of Advanced Multipath, or assist in making some of the goals achievable.

6.1. Loss and Delay Measurement

Procedures for measuring loss and delay are provided in [RFC6374]. These are OAM based measurements. This work could be the basis of delay measurements and delay variation measurement used for metrics called for in [I-D.ietf-rtgwg-cl-requirement].

Currently there are three documents that address delay and delay variation metrics.

draft-ietf-ospf-te-metric-extensions

[I-D.ietf-ospf-te-metric-extensions] provides a set of OSPF-TE extension to support delay, jitter, and loss. Stability is not adequately addressed and some minor issues remain.

I-D.previdi-isis-te-metric-extensions

[I-D.previdi-isis-te-metric-extensions] provides the set of extensions for ISIS that [I-D.ietf-ospf-te-metric-extensions] provides for OSPF. This draft mirrors [I-D.ietf-ospf-te-metric-extensions] sometimes lagging for a brief period when the OSPF version is updated.

I-D.atlas-mppls-te-express-path

[I-D.atlas-mppls-te-express-path] provides information on the use of OSPF and ISIS extensions defined in [I-D.ietf-ospf-te-metric-extensions] and [I-D.previdi-isis-te-metric-extensions] and a modified CSPF path selection to meet LSP performance criteria such as minimal delay paths or bounded delay paths.

Delay variance, loss, residual bandwidth, and available bandwidth extensions are particular prone to network instability. The question as to whether queuing delay and delay variation should be considered, and if so for which diffserv Per-Hop Service Class (PSC) is not

adequately addressed in the current versions of these drafts. These drafts are actively being discussed and updated and remaining issues are expected to be resolved.

6.2. Link Bundle Extensions

A set of extension are needed to indicate a group of component links in the ERO or RRO, where the group is given an interface identification like the bundle itself. The extensions could also be further extended to support specification of the all-ones component link in the ERO or RRO.

[I-D.ospf-cc-stlv] provides a baseline draft for extending link bundling to advertise components. A new component TLV (C-TLV) is proposed, which must reference an Advanced Multipath Link TLV. [I-D.ospf-cc-stlv] is intended for the OSPF WG and submitted for the "Experimental" track. The 00 version expired in February 2012. A replacement is expected that will be submitted for consideration on the standards track.

6.3. Pseudowire Flow and MPLS Entropy Labels

Two documents provide a means to add entropy for the purpose of improving load balance. MPLS encapsulation can bury information that is needed to identify microflows. These two documents allow a pseudowire ingress and LSP ingress respectively to add a label solely for the purpose of providing a finer granularity of microflow groups.

[RFC6391] allows pseudowires which carry a large volume of traffic, where microflows can be identified to be load balanced across multiple members of an Ethernet LAG or an MPLS link bundle. This is accomplished by adding a flow label below the pseudowire label in the MPLS label stack. For this to be effective the link bundle load balance must make use of the label stack up to and including this flow label.

[RFC6790] provides a means for a LER to put an additional label known as an entropy label on the MPLS label stack. Only the LER can add the entropy label. The LER of a PSC LSP would have to add a entropy label for contained LSPs for which it is a midpoint LSR.

Core LSR acting as LER for aggregated LSP can add entropy labels based on deep packet inspection and place an entropy label indicator (ELI) and entropy label (EL) just below the label being acted on. This would be helpful in situations where the label stack depth to which load distribution can operate is limited by implementation or is limited for other reasons such as carrying both MPLS-TP and MPLS with entropy labels within the same hierarchical LSP.

6.4. Multipath Extensions

The multipath extensions drafts address the issue of accommodating LSP which have strict packet ordering constraints in a network containing multipath. MPLS-TP has become the one important instance of LSP with strict packet ordering constraints and has driven this work.

[I-D.ietf-mpls-multipath-use] proposed to use MPLS Entropy Label [RFC6790] to allow MPLS-TP to be carried within MPLS LSP that make use of multipath. Limitations of this approach in the absence of protocol extensions is discussed.

[I-D.villamizar-mpls-multipath-extn] provides protocol extensions needed to overcome the limitations in the absence of protocol extensions is discussed in [I-D.ietf-mpls-multipath-use].

7. Required Protocol Extensions and Mechanisms

Prior sections have reviewed key characteristics, architecture tradeoffs, new challenges, existing mechanisms, and relevant mechanisms proposed in existing new documents.

This section first summarizes and groups requirements specified in [I-D.ietf-rtgwg-cl-requirement] (see Section 7.1). A set of documents coverage groupings are proposed with existing works-in-progress noted where applicable (see Section 7.2). The set of extensions are then grouped by protocol affected as a convenience to implementors (see (see Section 7.3)).

7.1. Brief Review of Requirements

The following list provides a categorization of requirements specified in [I-D.ietf-rtgwg-cl-requirement] along with a short phrase indication what topic the requirement covers.

routing information aggregation

FR#1 (routing summarization), FR#20 (Advanced Multipath may be a component of another Advanced Multipath)

restoration speed

FR#2 (restoration speed meeting performance objectives), FR#12 (minimally disruptive load rebalance), DR#6 (fast convergence), DR#7 (fast worst case failure convergence)

load distribution, stability, minimal disruption

FR#3 (automatic load distribution), FR#5 (must not oscillate), FR#11 (dynamic placement of flows), FR#12 (minimally disruptive load rebalance), FR#13 (bounded rearrangement frequency), FR#18 (flow placement must satisfy performance objectives), FR#19 (flow identification finer than per top level LSP), MR#6 (operator initiated flow rebalance)

backward compatibility and migration

FR#4 (smooth incremental deployment), FR#6 (management and diagnostics must continue to function), DR#1 (extend existing protocols), DR#2 (extend LDP, no LDP TE)

delay and delay variation

FR#7 (expose lower layer measured delay), FR#8 (precision of latency reporting), FR#9 (limit latency on per LSP basis), FR#15 (minimum delay path), FR#16 (bounded delay path), FR#17 (bounded jitter path)

admission control, preemption, traffic engineering

FR#10 (admission control, preemption), FR#14 (packet ordering), FR#21 (ingress specification of path), FR#22 (path symmetry), DR#3 (IP and LDP traffic), MR#3 (management specification of path)

single vs multiple domain

DR#4 (IGP extensions allowed within single domain), DR#5 (IGP extensions disallowed in multiple domain case)

general network management

MR#1 (polling, configuration, and notification), MR#2 (activation and de-activation)

path determination, connectivity verification

MR#4 (path trace), MR#5 (connectivity verification)

The above list is not intended as a substitute for

[I-D.ietf-rtgwg-cl-requirement], but rather as a concise grouping and reminder or requirements to serve as a means of more easily determining requirements coverage of a set of protocol documents.

7.2. Proposed Document Coverage

The primary areas where additional protocol extensions and mechanisms are required include the topics described in the following subsections.

There are candidate documents for a subset of the topics below. This

grouping of topics does not require that each topic be addressed by a separate document. In some cases, a document may cover multiple topics, or a specific topic may be addressed as applicable in multiple documents.

7.2.1. Component Link Grouping

An extension to link bundling is needed to specify a group of components with common attributes. This can be a TLV defined within the link bundle that carries the same encapsulations as the link bundle. Two interface indices would be needed for each group.

- a. An index is needed that if included in an ERO would indicate the need to place the LSP on any one component within the group.
- b. A second index is needed that if included in an ERO would indicate the need to balance flows within the LSP across all components of the group. This is equivalent to the "all-ones" component for the entire bundle.

[I-D.ospf-cc-stlv] can be extended to include multipath treatment capabilities. An ISIS solution is also needed. An extension of RSVP-TE signaling is needed to indicate multipath treatment preferences.

If a component group is allowed to support all of the parameters of a link bundle, then a group TE metric would be accommodated. This can be supported with the component TLV (C-TLV) defined in [I-D.ospf-cc-stlv].

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "routing information aggregation" set of requirements. The "restoration speed", "backward compatibility and migration", and "general network management" requirements must also be considered.

7.2.2. Delay and Jitter Extensions

A extension is needed in the IGP-TE advertisement to support delay and delay variation for links, link bundles, and forwarding adjacencies. Whatever mechanism is described must take precautions that insure that route oscillations cannot occur. The following set of drafts address this.

1. [I-D.ietf-ospf-te-metric-extensions]
2. [I-D.previdi-isis-te-metric-extensions]

3. [I-D.atlas-mpis-te-express-path]

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "delay and delay variation" set of requirements. The "restoration speed", "backward compatibility and migration", and "general network management" requirements must also be considered.

7.2.3. Path Selection and Admission Control

Path selection and admission control changes must be documented in each document that proposes a protocol extension that advertises a new capability or parameter that must be supported by changes in path selection and admission control.

It would also be helpful to have an informational document which covers path selection and admission control issues in detail and briefly summarizes and references the set of documents which propose extensions. This document could be advanced in parallel with the protocol extensions.

The primary focus of this document, among the sets of requirements listed in Section 7.1 are the "load distribution, stability, minimal disruption" and "admission control, preemption, traffic engineering" sets of requirements. The "restoration speed" and "path determination, connectivity verification" requirements must also be considered. The "backward compatibility and migration", and "general network management" requirements must also be considered.

7.2.4. Dynamic Multipath Balance

FR#11 explicitly calls for dynamic placement of flows. Load balancing similar to existing dynamic multipath would satisfy this requirement. In implementations where flow identification uses a coarse granularity, the adjustments would have to be equally coarse, in the worst case moving entire LSP. The impact of flow identification granularity and potential dynamic multipath approaches may need to be documented in greater detail than provided here.

The primary focus of this document, among the sets of requirements listed in Section 7.1 are the "restoration speed" and the "load distribution, stability, minimal disruption" sets of requirements. The "path determination, connectivity verification" requirements must also be considered. The "backward compatibility and migration", and "general network management" requirements must also be considered.

7.2.5. Frequency of Load Balance

IGP-TE and RSVP-TE extensions are needed to support frequency of load balancing rearrangement called for in FR#13, and FR#15-FR#17. Constraints are not defined in RSVP-TE, but could be modeled after administrative attribute affinities in RFC3209 and elsewhere.

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "load distribution, stability, minimal disruption" set of requirements. The "path determination, connectivity verification" must also be considered. The "backward compatibility and migration" and "general network management" requirements must also be considered.

7.2.6. Inter-Layer Communication

Lower layer to upper layer communication called for in FR#7 and FR#20. Specific parameters, specifically delay and delay variation, need to be addressed. Passing information from a lower non-MPLS layer to an MPLS layer needs to be addressed, though this may largely be generic advice encouraging a coupling of MPLS to lower layer management plane or control plane interfaces. This topic can be addressed in each document proposing a protocol extension, where applicable.

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "restoration speed" set of requirements. The "backward compatibility and migration" and "general network management" requirements must also be considered.

7.2.7. Packet Ordering Requirements

A document is needed to define extensions supporting various packet ordering requirements, ranging from requirements to preserve microflow ordering only, to requirements to preserve full LSP ordering (as in MPLS-TP). This is covered by [I-D.ietf-mpls-multipath-use] and [I-D.villamizar-mpls-multipath-extn].

The primary focus of this document, among the sets of requirements listed in Section 7.1 are the "admission control, preemption, traffic engineering" and the "path determination, connectivity verification" sets of requirements. The "backward compatibility and migration" and "general network management" requirements must also be considered.

7.2.8. Minimally Disruption Load Balance

The behavior of hash methods used in classic multipath needs to be described in terms of FR#12 which calls for minimally disruptive load adjustments. For example, reseeding the hash violates FR#12. Using modulo operations is significantly disruptive if a link comes or goes down, as pointed out in [RFC2992]. In addition, backwards compatibility with older hardware needs to be accommodated.

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "load distribution, stability, minimal disruption" set of requirements.

7.2.9. Path Symmetry

Protocol extensions are needed to support dynamic load balance as called for to meet FR#22 (path symmetry) and to meet FR#11 (dynamic placement of flows).

Currently path symmetry can only be supported in link bundling if the path is pinned. When a flow is moved both ingress and egress must make the move as close to simultaneously as possible to satisfy FR#22 and FR#12 (minimally disruptive load rebalance). There is currently no protocol to coordinate this move.

If a group of flows are identified using a hash, then the hash must be identical on the pair of LSR at the endpoint, using the same hash seed and with one side swapping source and destination. If the label stack is used, then either the entire label stack must be a special case flow identification, since the set of labels in either direction are not correlated, or the two LSR must conspire to use the same flow identifier. For example, using a common entropy label value, and using only the entropy label in the flow identification would satisfy the forwarding requirement. There is no protocol to indicate special treatment of a label stack within a hierarchical LSP. Adding such an extension may add significant complexity and ultimately may prove unscalable.

The primary focus of this document, among the sets of requirements listed in Section 7.1 are the "load distribution, stability, minimal disruption" and the "admission control, preemption, traffic engineering" sets of requirements. The "backward compatibility and migration" and "general network management" requirements must also be considered. Path symmetry simplifies support for the "path determination, connectivity verification" set of requirements, but with significant complexity added elsewhere.

7.2.10. Performance, Scalability, and Stability

A separate document providing analysis of performance, scalability, and stability impacts of changes may be needed. The topic of traffic adjustment oscillation must also be covered. If sufficient coverage is provided in each document covering a protocol extension, a separate document would not be needed.

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "restoration speed" set of requirements. This is not a simple topic and not a topic that is well served by scattering it over multiple documents, therefore it may be best to put this in a separate document and put citations in documents called for in Section 7.2.1, Section 7.2.2, Section 7.2.3, Section 7.2.9, Section 7.2.11, Section 7.2.12, Section 7.2.13, and Section 7.2.14. Citation may also be helpful in Section 7.2.4, and Section 7.2.5.

7.2.11. IP and LDP Traffic

A document is needed to define the use of measurements of native IP and native LDP traffic levels which are then used to reduce link advertised bandwidth amounts.

The primary focus of this document, among the sets of requirements listed in Section 7.1 are the "load distribution, stability, minimal disruption" and the "admission control, preemption, traffic engineering" set of requirements. The "path determination, connectivity verification" must also be considered. The "backward compatibility and migration" and "general network management" requirements must also be considered.

7.2.12. LDP Extensions

Extending LDP is called for in DR#2. LDP can be extended to couple FEC admission control to local resource availability without providing LDP traffic engineering capability. Other LDP extensions such as signaling a bound on microflow size and LDP LSP requirements would provide useful information without providing LDP traffic engineering capability.

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "admission control, preemption, traffic engineering" set of requirements. The "backward compatibility and migration" and "general network management" requirements must also be considered.

7.2.13. Pseudowire Extensions

Pseudowire (PW) extensions such as signaling a bound on microflow size and signaling requirements specific to PW would provide useful information. This information can be carried in the PW LDP signaling [RFC3985] and the the PW requirements could then be used in a containing LSP.

The primary focus of this document, among the sets of requirements listed in Section 7.1 is the "admission control, preemption, traffic engineering" set of requirements. The "backward compatibility and migration" and "general network management" requirements must also be considered.

7.2.14. Multi-Domain Advanced Multipath

DR#5 calls for Advanced Multipath to span multiple network topologies. Component LSP may already span multiple network topologies, though most often in practice these are LDP signaled. Component LSP which are RSVP-TE signaled may also span multiple network topologies using at least three existing methods (per domain [RFC5152], BRPC [RFC5441], PCE [RFC4655]). When such component links are combined in an Advanced Multipath, the Advanced Multipath spans multiple network topologies. It is not clear in which document this needs to be described or whether this description in the framework is sufficient. The authors and/or the WG may need to discuss this. DR#5 mandates that IGP-TE extension cannot be used. This would disallow the use of [RFC5316] or [RFC5392] in conjunction with [RFC5151].

The primary focus of this document, among the sets of requirements listed in Section 7.1 are "single vs multiple domain" and "admission control, preemption, traffic engineering". The "routing information aggregation" and "load distribution, stability, minimal disruption" requirements need attention due to their use of the IGP in single domain Advanced Multipath. Other requirements such as "delay and delay variation", can more easily be accommodated by carrying metrics within BGP. The "path determination, connectivity verification" requirements need attention due to requirements to restrict disclosure of topology information across domains in multi-domain deployments. The "backward compatibility and migration" and "general network management" requirements must also be considered.

7.3. Framework Requirement Coverage by Protocol

As an aid to implementors, this section summarizes requirement coverage listed in Section 7.2 by protocol or LSR functionality affected.

Some documentation may be purely informational, proposing no changes and proposing usage at most. This includes Section 7.2.3, Section 7.2.8, Section 7.2.10, and Section 7.2.14.

Section 7.2.9 may require a new protocol.

7.3.1. OSPF-TE and ISIS-TE Protocol Extensions

Many of the changes listed in Section 7.2 require IGP-TE changes, though most are small extensions to provide additional information. This set includes Section 7.2.1, Section 7.2.2, Section 7.2.5, Section 7.2.6, and Section 7.2.7. An adjustment to existing advertised parameters is suggested in Section 7.2.11.

7.3.2. PW Protocol Extensions

The only suggestion of pseudowire (PW) extensions is in Section 7.2.13.

7.3.3. LDP Protocol Extensions

Potential LDP extensions are described in Section 7.2.12.

7.3.4. RSVP-TE Protocol Extensions

RSVP-TE protocol extensions are called for in Section 7.2.1, Section 7.2.5, Section 7.2.7, and Section 7.2.9.

7.3.5. RSVP-TE Path Selection Changes

Section 7.2.3 calls for path selection to be addressed in individual documents that require change. These changes would include those proposed in Section 7.2.1, Section 7.2.2, Section 7.2.5, and Section 7.2.7.

7.3.6. RSVP-TE Admission Control and Preemption

When a change is needed to path selection, a corresponding change is needed in admission control. The same set of sections applies: Section 7.2.1, Section 7.2.2, Section 7.2.5, and Section 7.2.7. Some resource changes such as a link delay change might trigger preemption. The rules of preemption remain unchanged, still based on holding priority.

7.3.7. Flow Identification and Traffic Balance

The following describe either the state of the art in flow identification and traffic balance or propose changes: Section 7.2.4,

Section 7.2.5, Section 7.2.7, and Section 7.2.8.

8. IANA Considerations

This is a framework document and therefore does not specify protocol extensions. This memo includes no request to IANA.

9. Security Considerations

The security considerations for MPLS/GMPLS and for MPLS-TP are documented in [RFC5920] and [RFC6941].

The types protocol extensions proposed in this framework document provide additional information about links, forwarding adjacencies, and LSP requirements. The protocol semantics changes described in this framework document propose additional LSP constraints applied at path computation time and at LSP admission at midpoints LSR. The additional information and constraints provide no additional security considerations beyond the security considerations already documented in [RFC5920] and [RFC6941].

10. Acknowledgments

Authors would like to thank Adrian Farrel, Fred Jounay, Yuji Kamite for his extensive comments and suggestions regarding early versions of this document, Ron Bonica, Nabil Bitar, Eric Gray, Lou Berger, and Kireeti Kompella for their reviews of early versions and great suggestions.

Authors would like to thank Iftekhhar Hussain for review and suggestions regarding recent versions of this document.

In the interest of full disclosure of affiliation and in the interest of acknowledging sponsorship, past affiliations of authors are noted. Much of the work done by Ning So occurred while Ning was at Verizon. Much of the work done by Curtis Villamizar occurred while at Infinera. Infinera continues to sponsor this work on a consulting basis.

11. References

11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC5712] Meyer, M. and JP. Vasseur, "MPLS Traffic Engineering Soft Preemption", RFC 5712, January 2010.
- [RFC6107] Shiomoto, K. and A. Farrel, "Procedures for Dynamically Signaled Hierarchical Label Switched Paths", RFC 6107, February 2011.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.
- [RFC6391] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC 6391, November 2011.

11.2. Informative References

- [DBP] Bertsekas, D., "Dynamic Behavior of Shortest Path Routing Algorithms for Communication Networks", IEEE Trans. Auto. Control 1982.
- [I-D.atlas-mpls-te-express-path]

Atlas, A., Drake, J., Giacalone, S., Ward, D., Previdi, S., and C. Filsfils, "Performance-based Path Selection for Explicitly Routed LSPs", draft-atlas-mpls-te-express-path-02 (work in progress), February 2013.

[I-D.ietf-mpls-multipath-use]

Villamizar, C., "Use of Multipath with MPLS-TP and MPLS", draft-ietf-mpls-multipath-use-00 (work in progress), February 2013.

[I-D.ietf-ospf-te-metric-extensions]

Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", draft-ietf-ospf-te-metric-extensions-04 (work in progress), June 2013.

[I-D.ietf-rtgwg-cl-requirement]

Villamizar, C., McDysan, D., Ning, S., Malis, A., and L. Yong, "Requirements for Advanced Multipath in MPLS Networks", draft-ietf-rtgwg-cl-requirement-11 (work in progress), July 2013.

[I-D.ietf-rtgwg-cl-use-cases]

Ning, S., Malis, A., McDysan, D., Yong, L., and C. Villamizar, "Advanced Multipath Use Cases and Design Considerations", draft-ietf-rtgwg-cl-use-cases-04 (work in progress), July 2013.

[I-D.ospf-cc-stlv]

Osborne, E., "Component and Composite Link Membership in OSPF", draft-ospf-cc-stlv-00 (work in progress), August 2011.

[I-D.previdi-isis-te-metric-extensions]

Previdi, S., Giacalone, S., Ward, D., Drake, J., Atlas, A., and C. Filsfils, "IS-IS Traffic Engineering (TE) Metric Extensions", draft-previdi-isis-te-metric-extensions-03 (work in progress), February 2013.

[I-D.villamizar-mpls-multipath-extn]

Villamizar, C., "Multipath Extensions for MPLS Traffic Engineering", draft-villamizar-mpls-multipath-extn-00 (work in progress), November 2012.

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated

Services", RFC 2475, December 1998.

- [RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and Multicast Next-Hop Selection", RFC 2991, November 2000.
- [RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, November 2000.
- [RFC3260] Grossman, D., "New Terminology and Clarifications for Diffserv", RFC 3260, April 2002.
- [RFC3468] Andersson, L. and G. Swallow, "The Multiprotocol Label Switching (MPLS) Working Group decision on MPLS signaling protocols", RFC 3468, February 2003.
- [RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.
- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.
- [RFC4448] Martini, L., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, April 2006.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, June 2007.
- [RFC5151] Farrel, A., Ayyangar, A., and JP. Vasseur, "Inter-Domain MPLS and GMPLS Traffic Engineering -- Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 5151, February 2008.
- [RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.
- [RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS

Traffic Engineering", RFC 5316, December 2008.

- [RFC5392] Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, January 2009.
- [RFC5441] Vasseur, JP., Zhang, R., Bitar, N., and JL. Le Roux, "A Backward-Recursive PCE-Based Computation (BRPC) Procedure to Compute Shortest Constrained Inter-Domain Traffic Engineering Label Switched Paths", RFC 5441, April 2009.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.
- [RFC6941] Fang, L., Niven-Jenkins, B., Mansfield, S., and R. Graveman, "MPLS Transport Profile (MPLS-TP) Security Framework", RFC 6941, April 2013.

Authors' Addresses

So Ning
Tata Communications

Email: ning.so@tatacommunications.com

Dave McDysan
Verizon
22001 Loudoun County PKWY
Ashburn, VA 20147
USA

Email: dave.mcdysan@verizon.com

Eric Osborne
Cisco

Email: eosborne@cisco.com

Lucy Yong
Huawei USA
5340 Legacy Dr.
Plano, TX 75025
USA

Phone: +1 469-277-5837
Email: lucy.yong@huawei.com

Curtis Villamizar
Outer Cape Cod Network Consulting

Email: curtis@occnc.com

Routing Area Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 22, 2013

S. Litkowski
B. Decraene
Orange
C. Filsfils
K. Raza
Cisco Systems
February 18, 2013

Operational management of Loop Free Alternates
draft-litkowski-rtgwg-lfa-manageability-01

Abstract

Loop Free Alternates (LFA), as defined in RFC 5286 is an IP Fast ReRoute (IP FRR) mechanism enabling traffic protection for IP traffic (and MPLS LDP traffic by extension). Following first deployment experiences, this document provides operational feedback on LFA, highlights some limitations, and proposes a set of refinements to address those limitations. It also proposes required management specifications.

This proposal is also applicable to remote LFA solution.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 22, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Operational issues with default LFA tie breakers	3
2.1. Case 1: Edge router protecting core failures	4
2.2. Case 2: Edge router chosen to protect core failures while core LFA exists	5
2.3. Case 3: suboptimal core alternate choice	6
2.4. Case 4: ISIS overload bit on LFA computing node	7
3. Configuration requirements	7
3.1. LFA enabling/disabling scope	7
3.2. Policy based LFA selection	8
3.2.1. Mandatory criteria	8
3.2.2. Enhanced criteria	9
4. Operational aspects	13
4.1. ISIS overload bit on LFA computing node	13
4.2. Manual triggering of FRR	14
4.3. Required local information	14
4.4. Coverage monitoring	15
5. Security Considerations	15
6. Contributors	15
7. Acknowledgements	15
8. IANA Considerations	15
9. References	16
9.1. Normative References	16
9.2. Informative References	16
Authors' Addresses	17

1. Introduction

Following the first deployments of Loop Free Alternates (LFA), this document provides feedback to the community about the management of LFA.

Section 2 provides real uses cases illustrating some limitations and suboptimal behavior.

Section 3 proposes requirements for activation granularity and policy based selection of the alternate.

Section 4 express requirements for the operational management of LFA.

2. Operational issues with default LFA tie breakers

[RFC5286] introduces the notion of tie breakers when selecting the LFA among multiple candidate alternate next-hops. When multiple LFA exist, RFC 5286 has favored the selection of the LFA providing the best coverage of the failure cases. While this is indeed a goal, this is one among multiple and in some deployment this lead to the selection of a suboptimal LFA. The following sections details real use cases of such limitations.

Note that the use case of per-prefix LFA is assumed throughout this analysis.

2.1. Case 1: Edge router protecting core failures

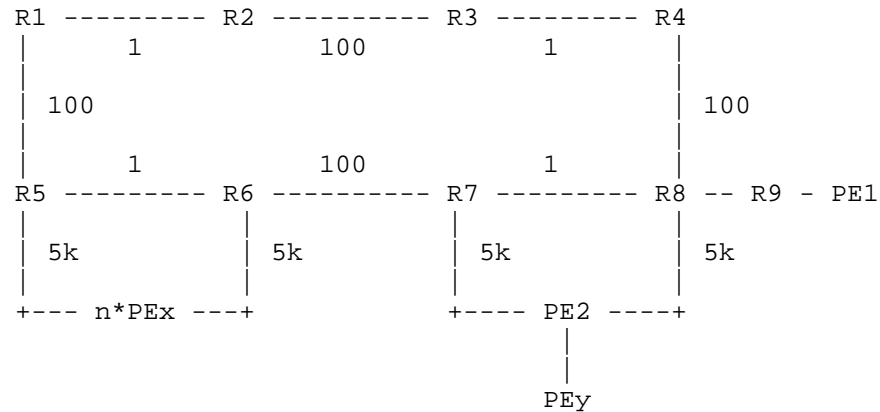


Figure 1

Rx routers are core routers using $n \times 10G$ links. PEs are connected using links with lower bandwidth.

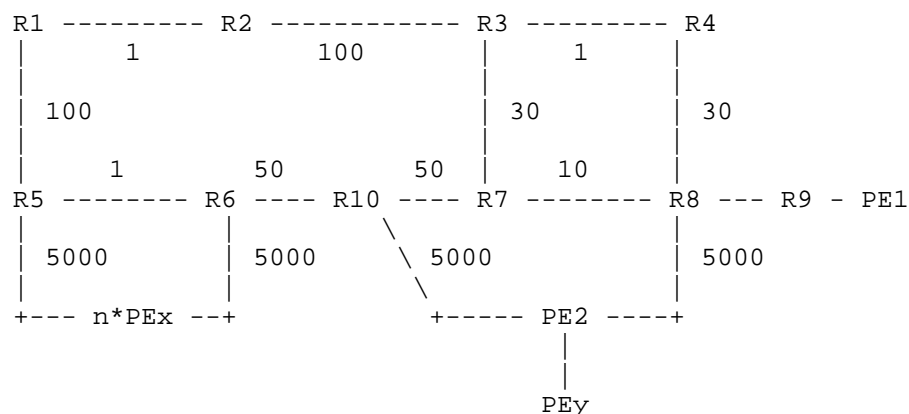
In figure 1, let us consider the traffic flowing from PE1 to PEx. The nominal path is R9-R8-R7-R6-PEx. Let us consider the failure of link R7-R8. For R8, R4 is not an LFA and the only available LFA is PE2.

When the core link R8-R7 fails, R8 switches all traffic destined to all the PEx towards the edge node PE2. Hence an edge node and edge links are used to protect the failure of a core link. Typically, edge links have less capacity than core links and congestion may occur on PE2 links. Note that although PE2 was not directly affected by the failure, its links become congested and its traffic will suffer from the congestion.

In summary, in case of failure, the impact on customer traffic is:

- o From PE2 point of view :
 - * without LFA: no impact
 - * with LFA: traffic is partially dropped (but possibly prioritized by a QoS mechanism). It must be highlighted that in such situation, traffic not affected by the failure may be affected by the congestion.

Besides the congestion aspects of using an Edge router as an alternate to protect a core failure, a service provider may consider this as a bad routing design and would like to prevent it.



In the figure 2, let us consider the traffic coming from PE1 to PEx. Nominal path is R9-R8-R7-R6-PEx. Let us consider the failure of the link R7-R8. For R8, R4 is a link-protecting LFA and PE2 is a node-protecting LFA. PE2 is chosen as best LFA due to its better protection type. Just like in case 1, this may lead to congestion on PE2 links upon LFA activation.

2.3. Case 3: suboptimal core alternate choice

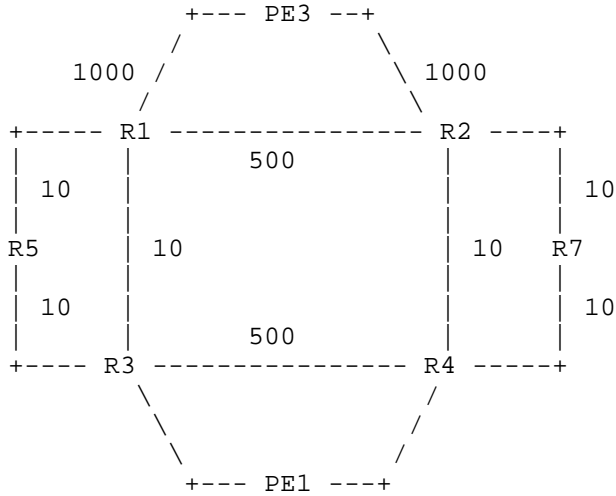


Figure 3

Rx routers are core routers. R1-R2 and R3-R4 links are 1G links. All others inter Rx links are 10G links.

In the figure above, let us consider the failure of link R1-R3. For destination PE3, R3 has two possible alternates:

- o R4, which is node-protecting
- o R5, which is link-protecting

R4 is chosen as best LFA due to its better protection type. However, it may not be desirable to use R4 for bandwidth capacity reason. A service provider may prefer to use high bandwidth links as preferred LFA. In this example, preferring shortest path over protection type may achieve the expected behavior, but in cases where metric are not reflecting bandwidth, it would not work and some other criteria would need to be involved when selecting the best LFA.

2.4. Case 4: ISIS overload bit on LFA computing node

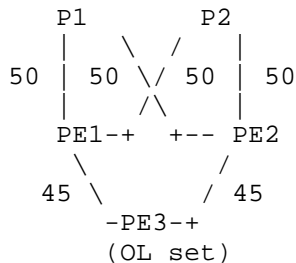


Figure 4

In the figure above, PE3 has its overload bit set (permanently, for design reason) and wants to protect traffic using LFA for destination PE2.

On PE3, the loopfree condition is not satisfied : $100 \nless 45 + 45$. PE1 is thus not considered as an LFA. However thanks to the overload bit set on PE3, we know that PE1 is loopfree so PE1 is an LFA to reach PE2.

In case of overload condition set on a node, LFA behavior must be clarified.

3. Configuration requirements

Controlling best alternate and LFA activation granularity is a requirement for Service Providers. This section defines configuration requirements for LFA.

3.1. LFA enabling/disabling scope

The granularity of LFA activation should be controlled (as alternate nexthop consume memory in forwarding plane).

An implementation of LFA SHOULD allow its activation with the following criteria:

- o Per address-family : ipv4 unicast, ipv6 unicast, LDP IPv4 unicast, LDP IPv6 unicast ...
- o Per routing context : VRF, virtual/logical router, global routing table, ...

- o Per interface
- o Per protocol instance, topology, area
- o Per prefixes: prefix protection SHOULD have a better priority compared to interface protection. This means that if a specific prefix must be protected due to a configuration request, LFA must be computed and installed for this prefix even if the primary outgoing interface is not configured for protection.

3.2. Policy based LFA selection

When multiple alternates exist, LFA selection algorithm is based on tie breakers. Current tie breakers do not provide sufficient control on how the best alternate is chosen. This document proposes an enhanced tie breaker allowing service providers to manage all specific cases:

1. An implementation of LFA SHOULD support policy-based decision for determining the best LFA.
2. Policy based decision SHOULD be based on multiple criterions, with each criteria having a level of preference.
3. If the defined policy does not permit to determine a unique best LFA, an implementation SHOULD pick only one based on its own decision, as a default behavior. An implementation SHOULD also support election of multiple LFAs, for loadbalancing purposes.
4. Policy SHOULD be applicable to a protected interface or to a specific set of destinations. In case of application on the protected interface, all destinations primarily routed on this interface SHOULD use the interface policy.
5. It is an implementation choice to reevaluate policy dynamically or not (in case of policy change). If a dynamic approach is chosen, the implementation SHOULD recompute the best LFAs and reinstall them in FIB, without service disruption. If a non-dynamic approach is chosen, the policy would be taken into account upon the next IGP event. In this case, the implementation SHOULD support a command to manually force the recomputation/reinstallation of LFAs.

3.2.1. Mandatory criteria

An implementation of LFA MUST support the following criteria:

- o Non candidate link: A link marked as "non candidate" will never be used as LFA.
- o A primary nexthop being protected by another primary nexthop of the same prefix (ECMP case).
- o Type of protection provided by the alternate: link protection, node protection. In case of node protection preference, an implementation SHOULD support fallback to link protection if node protection is not available.
- o Shortest path: lowest IGP metric used to reach the destination.
- o SRLG (as defined in [RFC5286] Section 3).

3.2.2. Enhanced criteria

An implementation of LFA SHOULD support the following enhanced criteria:

- o Downstreamness of a neighbor : preference of a downstream path over a non downstream path SHOULD be configurable.
- o Link coloring with : include, exclude and preference based system.
- o Link Bandwidth.
- o Neighbor preference.
- o Neighbor type: link or tunnel alternate. This means that user may change preference between link alternate or tunnel alternate (link preferred over tunnel, or considered as equal).

3.2.2.1. Link coloring

Link coloring is a powerful system to control the choice of alternates. Protecting interfaces are tagged with colors. Protected interfaces are configured to include some colors with a preference level, and exclude others.

Link color information SHOULD be signalled in the IGP. How signalling is done is out of scope of the document but it may be useful to reuse existing admin-groups from traffic-engineering extensions.

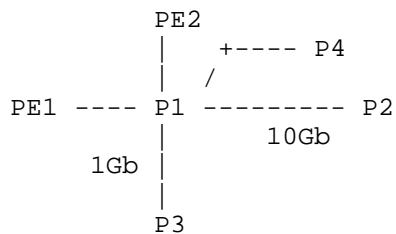


Figure 5

Example : P1 router is connected to three P routers and two PEs.

P1 is configured to protect the P1-P4 link. We assume that given the topology, all neighbors are candidate LFA. We would like to enforce a policy in the network where only a core router may protect against the failure of a core link, and where high capacity links are preferred.

In this example, we can use the proposed link coloring by:

- o Marking PEs links with color RED
- o Marking 10Gb CORE link with color BLUE
- o Marking 1Gb CORE link with color YELLOW
- o Configured the protected interface P1->P4 with :
 - * Include BLUE, preference 200
 - * Include YELLOW, preference 100
 - * Exclude RED

Using this, PE links will never be used to protect against P1-P4 link failure and 10Gb link will be preferred.

The main advantage of this solution is that it can easily be duplicated on other interfaces and other nodes without change. A Service Provider has only to define the color system (associate color with a significance), as it is done already for TE affinities or BGP communities.

An implementation of link coloring:

- o SHOULD support multiple include and exclude colors on a single protected interface.

- o SHOULD provide a level of preference between included colors.
- o SHOULD support multiple colors configuration on a single protecting interface.

3.2.2.2. Bandwidth

As mentionned in previous sections, not taking into account bandwidth of an alternate could lead to congestion during FRR activation. We propose to base the bandwidth criteria on the link speed information for the following reason :

- o if a router S has a set of X destinations primarily forwarded to N, using per prefix LFA may lead to have a subset of X protected by a neighbor N1, another subset by N2, another subset by Nx ...
- o S is not aware about traffic flows to each destination and is not able to evaluate how much traffic will be sent to N1,N2, ... Nx in case of FRR activation.

Based on this, it is not useful to gather available bandwidth on alternate paths, as the router does not know how much bandwidth it requires for protection. The proposed link speed approach provides a good approximation with a small cost as information is easily available.

The bandwidth criteria of the policy framework SHOULD work in two ways :

- o PRUNE : exclude a LFA if link speed to reach it is lower than the link speed of the primary nexthop interface.
- o PREFER : prefer a LFA based on his bandwidth to reach it compared to the link speed of the primary nexthop interface.

3.2.2.3. Neighbor preference

Rather than tagging interface on each node (using link color) to identify neighbor node type (as example), it would be helpful if routers could be identified in the IGP. This would permit a grouped processing on multiple nodes. Some existing IGP extension like SUB-TLV 1 of TLV 135 may be useful for this purpose. As an implementation must be able to exclude some specific neighbors (see mandatory criterions), an implementation :

- o SHOULD be able to give a preference to specific neighbor.

- o SHOULD be able to give a preference to a group of neighbor.
- o SHOULD be able to exclude a group of neighbor.

A specific neighbor may be identified by its interface or IP address and group of neighbors may be identified by a marker like SUB-TLV1 in TLV135. As multiple prefixes may be present in TLVs 135, an heuristic is required to choose the appropriate one that will identify the neighbor and will transport the tag associated with the neighbor preference.

We propose the following algorithm to select the prefix :

1. Select the prefix in TLV#135 that is equal to TLV#134 value (Router ID) and prefix length is 32.
2. Select the prefix in TLV#135 that is equal to TLV#132 value (IP Addresses) and prefix length is 32, it must be noted that TLV#132 may transport multiple addresses and so multiple matches may happen.
3. If multiple prefixes are matching TLV#132 values, choose the highest one.

Consider the following network:

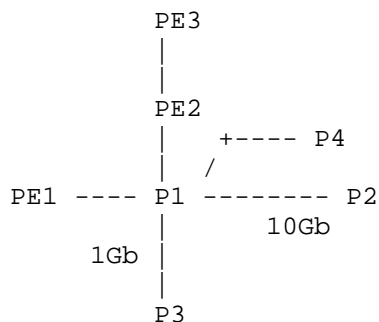


Figure 6

In the example above, each node is configured with a specific tag flooded through the IGP.

- PE1,PE3: 200 (non candidate).

- o PE2: 100 (edge/core).
- o P1,P2,P3: 50 (core).

A simple policy could be configured on P1 to choose the best alternate for P1->P4 based on router function/role as follows :

- o criteria 1 -> neighbor preference: exclude tag 100 and 200.
- o criteria 2 -> bandwidth.

3.2.2.4. Link vs remote alternate

In addition to LFA, tunnels (IP, LDP or RSVP-TE) to distant routers may be used to complement LFA coverage (tunnel tail used as virtual neighbor). When a router has multiple alternate candidates for a specific destination, it may have connected alternates (link alternates) and remote alternates reachable via a tunnel. Link alternates may not always provide an optimal routing path and it may be preferable to select a remote alternate over a link alternate. The usage of tunnels to extend LFA coverage is described in [I-D.ietf-rtgwg-remote-lfa] and [I-D.litkowski-rtgwg-lfa-rsvpte-cooperation].

In figure 1, there is no core alternate for R8 to reach PEs located behind R6, so R8 is using PE2 as alternate, which may generate congestion when FRR is activated. Instead, we could have a remote core alternate for R8 to protect PEs destinations. For example, a tunnel from R8 to R3 would ensure a LFA protection without any impact.

There is a requirement to be able to compare remote alternates (reachable through a tunnel) to link alternates (a remote alternate may provide a better protection than a link alternate based on service provider's criteria). Policy will associate a preference to each alternate whatever their type (link or remote) and will elect the best one.

4. Operational aspects

4.1. ISIS overload bit on LFA computing node

In [RFC5286], Section 3.5, the setting of the overload bit condition in LFA computation is only taken into account for the case where a neighbor has the overload bit set.

In addition to RFC 5286 inequality 1 Loop-Free Criterion

(Distance_opt(N, D) < Distance_opt(N, S) + Distance_opt(S, D)), the IS-IS overload bit of the LFA calculating neighbor (S) SHOULD be taken into account. Indeed, if it has the overload bit set, no neighbor will loop back to traffic to itself.

4.2. Manual triggering of FRR

Service providers often use using manual link shutdown (using router CLI) to perform some network changes/tests. Especially testing or troubleshooting FRR requires to perform the manual shutdown on the remote end of the link as generally a local shutdown would not trigger FRR. To enhance such situation, an implementation SHOULD support triggering/activating LFA Fast Reroute for a given link when a manual shutdown is done.

4.3. Required local information

LFA introduction requires some enhancement in standard routing information provided by implementations. Moreover, due to the non 100% coverage, coverage informations is also required.

Hence an implementation :

- o MUST be able to display, for every prefixes, the primary nexthop as well as the alternate nexthop information.
- o MUST provide coverage information per activation domain of LFA (area, level, topology, instance, virtual router, address family ...).
- o MUST provide number of protected prefixes as well as non protected prefixes globally.
- o SHOULD provide number of protected prefixes as well as non protected prefixes per link.
- o MAY provide number of protected prefixes as well as non protected prefixes per priority if implementation supports prefix-priority insertion in RIB/FIB.
- o SHOULD provide a reason for choosing an alternate (policy and criteria) and for excluding an alternate.
- o SHOULD provide the list of non protected prefixes and the reason why they are not protected (no protection required or no alternate available).

4.4. Coverage monitoring

It is pretty easy to evaluate the coverage of a network in a nominal situation, but topology changes may change the coverage. In some situations, the network may no longer be able to provide the required level of protection. Hence, it becomes very important for service providers to get alerted about changes of coverage.

An implementation SHOULD :

- o provide an alert system if total coverage (for a node) is below a defined threshold or comes back to a normal situation.
- o provide an alert system if coverage of a specific link is below a defined threshold or comes back to a normal situation.

An implementation MAY :

- o provide an alert system if a specific destination is not protected anymore or when protection comes back up for this destination

Although the procedures for providing alerts are beyond the scope of this document, we recommend that implementations consider standard and well used mechanisms like syslog or SNMP traps.

5. Security Considerations

This document does not introduce any change in security consideration compared to [RFC5286].

6. Contributors

Significant contributions were made by Pierre Francois, Hannes Gredler and Mustapha Aissaoui which the authors would like to acknowledge.

7. Acknowledgements

8. IANA Considerations

This document has no action for IANA.

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.

9.2. Informative References

- [I-D.ietf-rtgwg-remote-lfa]
Bryant, S., Filsfils, C., Previdi, S., Shand, M., and S. Ning, "Remote LFA FRR", draft-ietf-rtgwg-remote-lfa-01 (work in progress), December 2012.
- [I-D.litkowski-rtgwg-lfa-rsvpte-cooperation]
Litkowski, S., Decraene, B., Filsfils, C., and K. Raza, "Interactions between LFA and RSVP-TE", draft-litkowski-rtgwg-lfa-rsvpte-cooperation-01 (work in progress), February 2013.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC3906] Shen, N. and H. Smit, "Calculating Interior Gateway Protocol (IGP) Routes Over Traffic Engineering Tunnels", RFC 3906, October 2004.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.
- [RFC5715] Shand, M. and S. Bryant, "A Framework for Loop-Free Convergence", RFC 5715, January 2010.
- [RFC6571] Filsfils, C., Francois, P., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.

Authors' Addresses

Stephane Litkowski
Orange

Email: stephane.litkowski@orange.com

Bruno Decraene
Orange

Email: bruno.decraene@orange.com

Clarence Filsfils
Cisco Systems

Email: cfilsfil@cisco.com

Kamran Raza
Cisco Systems

Email: skraza@cisco.com

Routing Area Working Group
Internet-Draft
Intended status: Standards Track
Expires: February 20, 2014

S. Litkowski
B. Decraene
Orange
C. FilsFils
K. Raza
Cisco Systems
August 19, 2013

Interactions between LFA and RSVP-TE
draft-litkowski-rtgwg-lfa-rsvpte-cooperation-02

Abstract

This document defines the behavior of a node supporting Loopfree Alternates (LFA) when the node has established RSVP TE tunnels. It first describes the decisions to be made by the LFA mechanism with respect to the use of TE tunnels as LFA candidates. Second, it discusses the use of RSVP TE tunnels as a way to complement the LFA coverage, illustrating how these technologies can benefit from each other.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on February 20, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. LFA FRR and MPLS-TE interactions	3
2.1. Use case : using MPLS LSP as LFA candidates	3
2.2. Specifications of interactions between LFA and TE LSP . .	4
2.2.1. Having both a physical interface and a TE tunnel toward a LFA	4
2.2.2. TE ingress LSP as LFA candidate	4
2.2.3. Independence between LFA and TE FRR	5
3. Operational considerations	7
3.1. Relevance of joint LFA FRR and RSVP-TE FRR deployments .	7
3.2. Extending LFA coverage using RSVP-TE tunnels	8
3.2.1. Creating multihop tunnel to extend topology	8
3.2.2. Selecting multihop tunnels to extend topology	9
4. Security Considerations	10
5. Contributors	10
6. IANA Considerations	10
7. References	10
7.1. Normative References	10
7.2. Informative References	10
Authors' Addresses	11

1. Introduction

When a failure occurs in an IP network, the subsequent converge process often leads to traffic disruption. Some mechanisms are available to limit traffic disruptions by pre-computing alternate paths and locally reroute over these as soon as the failure is detected. Such techniques are commonly known as "protection mechanisms". Currently, the protection mechanisms widely used in Service Provider networks are RSVP-TE Fast Reroute [RFC4090] and Loop Free Alternates [RFC5286]. RSVP-TE FRR permits full network coverage but with a quite high complexity in terms of operation, as well as potential scaling issues. On the other hand, LFA offer a very easy,

manageable, and scalable mechanism, but does not provide full coverage.

This document discusses how LFA and RSVP-TE should interact. It first describes how an LFA implementation should deal with existing RSVP TE tunnels established by the LFA node, as well as its behavior with respect to established IGP Shortcut tunnels [RFC3906]. Second, the document suggests the use of RSVP-TE tunnels to extend LFA coverage, and discusses the management and operational aspects of such a practice.

2. LFA FRR and MPLS-TE interactions

This section discusses the various interactions among LFA FRR and MPLS-TE FRR. It starts with a simple example emphasizing the benefits of jointly using of LFA-FRR and MPLS-TE FRR, and then summarizes the requirements for the interactions between LFAs and MPLS-FRR.

2.1. Use case : using MPLS LSP as LFA candidates

In some cases, typically in ring shapped parts of network topologies, links cannot be protected by LFAs. In the following topology, from the point of view of R5, LFAs are able to partialtly protect (49% of the destination routers) from the failure of R3, while the failure of R4 is not covered at all.

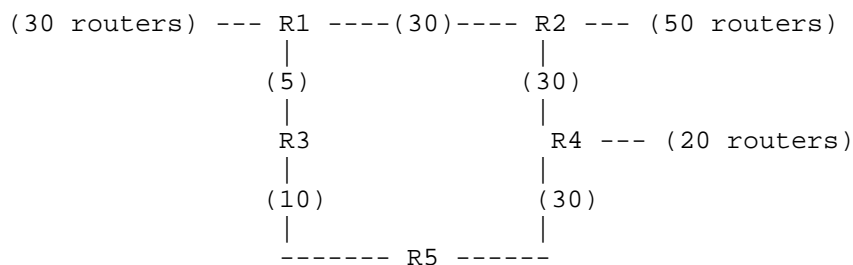


Figure 1

Many networks deploy MPLS tunnels for traffic engineering and resiliency reasons. To extend its benefit, an LFA implementation could take advantage of such existing MPLS tunnels. In the exemple above, if R5 has established TE tunnels bypassing R4 and R3, these could be considerd as LFA candidates respectively protecting links from R5 to R4 and R3.

In the following section, we provide a detailed summary of the behavior to be applied by an LFA implementation which would consider the existence of MPLS TE tunnels to improve its applicability. The explicit configuration of such tunnels with the intent of improving LFA applicability is discussed in later sections.

2.2. Specifications of interactions between LFA and TE LSP

Here we summarize the normative requirements for the interaction between LFA FRR and MPLS TE tunnels.

2.2.1. Having both a physical interface and a TE tunnel toward a LFA

If a node S has both a physical interface and a TE tunnel to reach a LFA, it SHOULD use the physical interface unless :

1. The tunnel has been explicitly configured as an LFA candidate.
2. The tunnel does not pass through the link subject to LFA protection.

In other words, if a node S has an IGP/LDP forwarding entry F1 with outgoing interface i1, and S originates a TE tunnel T2 terminating on direct neighbor N2 (for example : if a TE tunnel is provisionned for link protection), T2 has an outgoing interface i2 and N2 is best LFA for F1, then an implementation MUST NOT use T2 when programming LFA repair for F1 unless T2 is configured as an LFA candidate.

2.2.2. TE ingress LSP as LFA candidate

A TE LSP can be used as a virtual interface to reach a LFA if

1. The TE tunnel has been configured to allow its use as an LFA candidate.
2. The TE tunnel does not pass through the primary outgoing interface of D.

This would permit to extend LFA coverage as described in [I-D.ietf-rtgwg-remote-lfa], in a controlled fashioned, as the tunnels used by the fast reroute mechanism are defined by configuration.

In other words, if a node S has an IGP/LDP forwarding entry F1 with outgoing interface i1 and S originates a TE tunnel T1 terminating at node Y, then an implementation SHOULD support a local policy which instructs node S to consider Y as a virtual neighbor and hence include Y as part of the LFA FRR alternate computation. In such

case, an implementation MUST not use Y as an LFA for F1 if T1's outgoing interface is i1.

2.2.3. Independence between LFA and TE FRR

2.2.3.1. Tunnel head-end case

Similar requirements can be expressed for TE IGP shortcut tunnels.

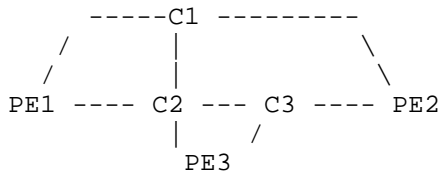


Figure 2

PE to Cx metrics are 50, Cx to Cx are 1

A service provider is often providing traffic-engineered path for specific customer traffic (L3VPN, PW ...) to ensure path diversity or traffic constraints. In the diagram above, we consider a TE tunnel T2 built on a non shortest path as follows : PE1->C2->C3->PE2 and IGP shortcut is activated on PE1 to make traffic to PE2 using T2. Based on operational feedback, some implementations prevent LFA computation to run for an interface where a TE tunnel exists. In our example, if LFA is activated on N, we would not be able to have a protection for PE3 destination as a tunnel exists on the interface. This current observed behavior leads to a very limited coverage for LFA. In the other hand, it is important to keep protection mechanisms independant as much as possible to keep implementation simple. We propose the following approach :

- o If an IP prefix is reachable through a TE tunnel, LFA must not compute a protection for it.
- o If an IP prefix is reachable through a native IP path, LFA MUST compute a protection for it disregarding the presence of a tunnel or not on the primary interface.

In other words, if a node S has an IGP/LDP forwarding entry F1 with outgoing interface i1 and an IGP/LDP forwarding entry F2 with outgoing interface onto a TE tunnel T2 (due to IGP shortcut [RFC3906]) and tunnel T2 has outgoing interface i2, then an implementation MUST support enabling LFA FRR for F1 and using TE FRR for F2 as long as $i1 \neq i2$.

If $i1 == i2$, an implementation SHOULD allow for using LFA FRR backup for F1 and TE FRR backup for F2.

The mechanisms for using TE tunnel as an LFA candidate, and RFC3906 mechanisms MUST be de-correlated -- i.e an implementation MUST support TE tunnel configuration with RFC3906 only, as LFA candidate only, or both at the same time.

2.2.3.2. Tunnel midpoint case

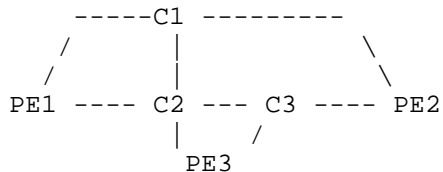


Figure 3

PE to Cx metrics are 50, except PE3-C3 (60), Cx to Cx are 1

In the diagram above, we consider a TE tunnel T2 built on a non shortest path as follows : PE1->C2->C3->PE2 and IGP shortcut is activated on PE1 to make traffic to PE2 using T2. C2 is a TE tunnel midpoint router. In terms of forwarding, C2 has a MPLS TE forwarding entry for T2, as well as an IP forwarding entry to PE2. As explained in previous sections, it would be too restrictive and would limit LFA benefit on C2 if C2 would not be able to compute an LFA for the IP forwarding entry to PE2 due to the presence of a transit tunnel.

We propose the following approach for a midpoint router of a TE tunnel :

- o MPLS TE forwarding entries MUST not be protected by LFA (if an operator wants protection, TE FRR could be enabled).
- o IP forwarding entries MUST be protected by LFA disregarding the presence of a TE tunnel transiting through the primary interface of the destination.

In our example :

- o MPLS TE forwarding entry for T2 (ending on PE2) would be protected by TE-FRR (if enabled).
- o IP forwarding entry for PE2 would be protected by LFA.

In case of failure of C2-C3 :

- o traffic from PE1 to PE2 (encapsulated in T2), would be protected by TE FRR.
- o traffic from PE3 to PE2 (native IP), would be protected by LFA.

In other words, if a node S has an IGP/LDP forwarding entry F1 with outgoing interface i1 and a MPLS TE midpoint forwarding entry F2 with outgoing interface i2, then an implementation MUST support using LFA FRR for F1 and TE FRR for F2 as long as $i1 \neq i2$.

If $i1 == i2$, an implementation SHOULD allow for using LFA FRR backup for F1 and TE FRR backup for F2.

3. Operational considerations

In this section, we first discuss the benefit of considering a joint deployment of LFA and MPLS tunnels to achieve resiliency. We then discuss one approach aiming at defining MPLS tunnels for the purpose of complementing LFA coverage.

3.1. Relevance of joint LFA FRR and RSVP-TE FRR deployments

This section describes the deployment scenarios where it can be beneficial to jointly use LFAs and RSVP-TE FRR.

There are many networks where RSVP-TE is already deployed. The deployment of RSVP-TE is typically for two main reasons :

- o Traffic engineering : a provider wants to route some flows on some specific paths using constraints;
- o Traffic protection using Fast-reroute ability

LFA is a feature that may bring benefits on RSVP-TE enabled networks, with no/minimal operational cost (compared to RSVP-TE FRR global roll out). These benefits include:

- o Should increase protection on network where FRR is not available everywhere. Although it may not provide full coverage, it will increase the protection significantly.
- o May provide better protection in specific cases than RSVP-TE FRR

For IP networks that do not have any traffic protection mechanism, LFA is a very good first step to provide traffic protection even if its coverage is not 100%. Providers may want to increase protection coverage if LFA benefit is not sufficient for some destinations, in some parts of the network. The following sections discusses the use of basic RSVP-TE tunnels to extend protection coverage.

3.2. Extending LFA coverage using RSVP-TE tunnels

We already have seen in previous sections that RSVP-TE tunnels could be established by an operator to complement LFA coverage. The method of tunnel placement depends on what type of protection (link or node) is required, as well as on the set of destinations or network parts which requires better protection than what LFA can provide.

3.2.1. Creating multihop tunnel to extend topology

To extend the coverage, the idea is to use a mechanism extending LFA by turning TE tunnels into LFA candidates. This mechanism is of a local significance only.

When explicitly establishing tunnels for that purpose, choices have to be made for the endpoints of such tunnels, in order to maximize coverage while preserving management simplicity. Requirements are that:

- o Endpoints must satisfy equations from [RFC5286], otherwise it will not be a valide LFA candidate: so when releasing traffic from tunnel, the traffic will go to the destination without flowing through the protected link or node. Depending on which equations are satisfied, node or link protection will be provided by the tunnel hop.
- o Tunnel must not flow though the link or node to be protected, explicit routing of tunnel is recommended to enforce this condition.

The approach to choose tunnel endpoints might be different here when compared to [I-D.ietf-rtgwg-remote-lfa] as endpoint choice is a manual one. Automatic behavior and scaling of [I-D.ietf-rtgwg-remote-lfa] requires:

- o Non null intersection of Extended P-Space and Q-Space
- o Computation of PQ node only for the remote end of the link

Based on this, [I-D.ietf-rtgwg-remote-lfa] may:

- o Not find a tunnel endpoint;
- o Not provide the more efficient protection : -- i.e. provides only link protection, while there is node protection possible for a specific destination

The proposed solution of manual explicitly routed tunnels is a good complement for [I-D.ietf-rtgwg-remote-lfa] and provides more flexibility:

- o Always a possibility to find a tunnel endpoint for a specific destination.
- o Possibility to provide a better protection type (link vs. node).

3.2.2. Selecting multihop tunnels to extend topology

From a manageability point of view, computing a best Q node for each destination could lead to have one different Q node for each destination. This is not optimal in terms of number of tunnels, given that possibly one Q node may be able to serve multiple non covered destinations.

Rather than computing the best Q node per non covered destination, we would prefer to find best compromise Q nodes (best for multiple destinations). To find the best compromise between coverage increase and number of tunnels, we recommend to use a simulator performing the following computations per link:

Step 1 : Compute for each not covered destination (routed on the link) the list of endpoints that are satisfying equations from [RFC5286] (node or link protection equations depending of required level of protection) : nodes in Q-Space

Step 2 : Remove endpoints that are not eligible for repair (Edge nodes, low bandwidth meshed nodes, number of hops ...) : multiple attributes could be specified to exclude some nodes from Q-Space : The example of attributes include router type, metric to node, bandwidth, packet loss, RTD ...

Step 3 : Within the list of endpoints (one list per destination), order the endpoints by number of destination covered

Step 4 : Choose the endpoint that has the highest number of destination covered : some other criteria could be used to prefer an endpoint from another (same type of criteria that excluded some nodes from Q-Space)

Step 5 : Remove destinations covered by this endpoint from non covered list

Step 6 : If non covered list is not empty, restart from Step 1

Multiple endpoint (and so tunnels) could be necessary to have 100% coverage. But the idea is to find a tradeoff between number of tunnels configured (complexity) and number of destination covered, combining with traffic information would also provide a better view.

4. Security Considerations

TBD.

5. Contributors

Significant contribution was made by Pierre Francois which the authors would like to acknowledge.

6. IANA Considerations

This document has no actions for IANA.

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3906] Shen, N. and H. Smit, "Calculating Interior Gateway Protocol (IGP) Routes Over Traffic Engineering Tunnels", RFC 3906, October 2004.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.

7.2. Informative References

[I-D.bryant-ipfrr-tunnels]

Bryant, S., Filsfils, C., Previdi, S., and M. Shand, "IP Fast Reroute using tunnels", draft-bryant-ipfrr-tunnels-03 (work in progress), November 2007.

[I-D.ietf-rtgwg-remote-lfa]

Bryant, S., Filsfils, C., Previdi, S., Shand, M., and S. Ning, "Remote LFA FRR", draft-ietf-rtgwg-remote-lfa-02 (work in progress), May 2013.

Authors' Addresses

Stephane Litkowski
Orange

Email: stephane.litkowski@orange.com

Bruno Decraene
Orange

Email: bruno.decraene@orange.com

Clarence FilsFils
Cisco Systems

Email: cfilsfil@cisco.com

Kamran Raza
Cisco Systems

Email: skraza@cisco.com

Routing Area Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 16, 2016

S. Litkowski
B. Decraene
Orange
C. Filsfils
Cisco Systems
P. Francois
IMDEA Networks
October 14, 2015

Microloop prevention by introducing a local convergence delay
draft-litkowski-rtgwg-uloop-delay-04

Abstract

This document describes a mechanism for link-state routing protocols to prevent local transient forwarding loops in case of link failure. This mechanism Proposes a two-steps convergence by introducing a delay between the convergence of the node adjacent to the topology change and the network wide convergence.

As this mechanism delays the IGP convergence it may only be used for planned maintenance or when fast reroute protects the traffic between the link failure and the IGP convergence.

Simulations using real network topologies have been performed and show that local loops are a significant portion (>50%) of the total forwarding loops.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any

time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 16, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Transient forwarding loops side effects	3
2.1. Fast reroute unefficiency	3
2.2. Network congestion	5
3. Overview of the solution	6
4. Specification	6
4.1. Definitions	7
4.2. Current IGP reactions	7
4.3. Local events	7
4.4. Local delay	8
4.4.1. Link down event	8
4.4.2. Link up event	9
5. Applicability	9
5.1. Applicable case : local loops	9
5.2. Non applicable case : remote loops	10
6. Simulations	10
7. Deployment considerations	11
8. Comparison with other solutions	12
8.1. PLSN	12
8.2. OFIB	13
9. Security Considerations	13
10. Acknowledgements	13
11. IANA Considerations	13
12. References	14
12.1. Normative References	14
12.2. Informative References	14

Authors' Addresses	15
------------------------------	----

1. Introduction

Micro-forwarding loops and some potential solutions are well described in [RFC5715]. This document describes a simple targeted mechanism that solves micro-loops local to the failure; based on network analysis, these are a significant portion of the micro-forwarding loops. A simple and easily deployable solution to these local micro-loops is critical because these local loops cause traffic loss after an advanced fast-reroute alternate has been used (see Section 2.1).

Consider the case in Figure 1 where S does not have an LFA to protect its traffic to D. That means that all non-D neighbors of S on the topology will send to S any traffic destined to D if a neighbor did not, then that neighbor would be loop-free. Regardless of the advanced fast-reroute technique used, when S converges to the new topology, it will send its traffic to a neighbor that was not loop-free and thus cause a local micro-loop. The deployment of advanced fast-reroute techniques motivates this simple router-local mechanism to solve this targeted problem. This solution can be work with the various techniques described in [RFC5715].

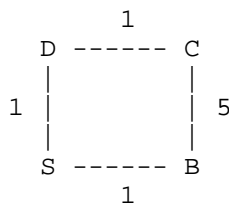


Figure 1

When S-D fails, a transient forwarding loop may appear between S and B if S updates its forwarding entry to D before B.

2. Transient forwarding loops side effects

Even if they are very limited in duration, transient forwarding loops may cause high damage for the network.

2.1. Fast reroute unefficiency

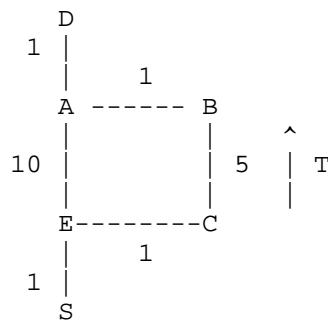
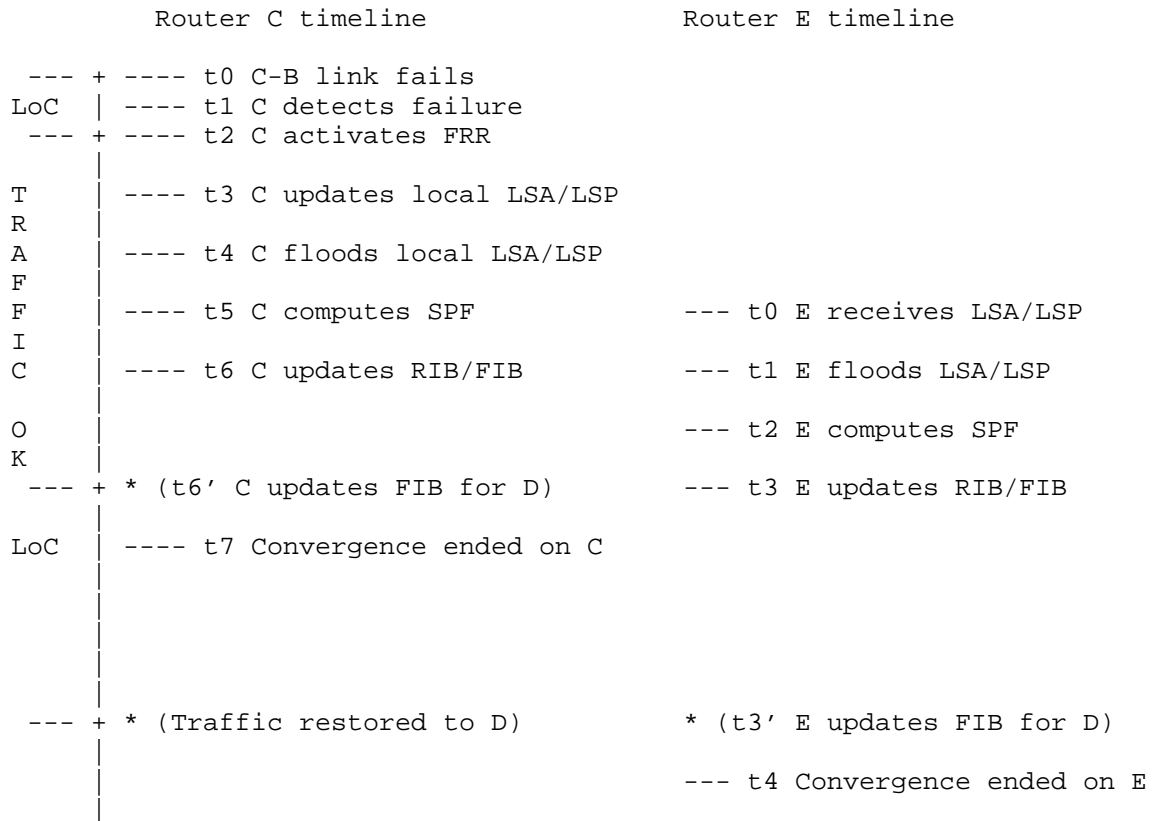


Figure 2 - RSVPTE FRR case

In figure 2, a RSVP-TE tunnel T, provisionned on C and terminating on B, is used to protect against C-B link failure (IGP shortcut activated on C). Primary path of T is C->B and FRR is activated on T providing a FRR bypass or detour using path C->E->A->B. On C, nexthop to D is tunnel T thanks to IGP shortcut. When C-B link fails :

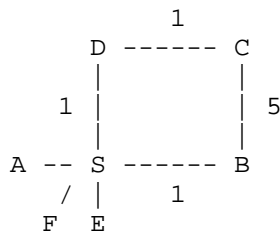
1. C detects the failure, and updates the tunnel path using preprogrammed FRR path, traffic path from S to D is :
S->E->C->E->A->B->A->D .
2. In parallel, on router C, both IGP convergence and TE tunnel convergence (tunnel path recomputation) are occurring :
 - * T path is recomputed : C->E->A->B
 - * IGP path to D is recomputed : C->E->A->D
3. On C, tail-end of the TE tunnel (router B) is no more on SPT to D, so C does not encapsulate anymore the traffic to D using the tunnel T and update forwarding entry to D using nexthop E.

If C updates its forwarding entry to D before router E, there would be a transient forwarding loop between C and E until E has converged.



The issue described here is completely independent of the fast-reroute mechanism involved (TE FRR, LFA/rLFA, MRT ...). Fast-reroute is working perfectly but ensures protection, by definition, only until the PLR has converged. When implementing FRR, a service provider wants to guarantee a very limited loss of connectivity time. The previous example shows that the benefit of FRR may be completely lost due to a transient forwarding loop appearing when PLR has converged. Delaying FIB updates after IGP convergence may permit to keep fast-reroute path until neighbor has converged and preserve customer traffic.

2.2. Network congestion



In the figure above, as presented in Section 1, when link S-D fails, a transient forwarding loop may appear between S and B for destination D. The traffic on S-B link will constantly increase due to the looping traffic to D. Depending on TTL of packets, traffic rate destined to D and bandwidth of link, the S-B link may be congested in few hundreds of milliseconds and will stay overloaded until the loop is solved.

Congestion introduced by transient forwarding loops are problematic as they are impacting traffic that is not directly concerned by the failing network component. In our example, the congestion of S-B link will impact customer traffic that is not directly concerned by the failure : e.g. A to B, F to B, E to B. Class of services may be implemented to mitigate the congestion but some traffic not directly concerned by the failure would still be dropped as a router is not able to identify looped traffic from normal traffic.

3. Overview of the solution

This document defines a two-step convergence initiated by the router detecting the failure and advertising the topological changes in the IGP. This introduces a delay between the convergence of the local router and the network wide convergence. This delay is positive in case of "down" events and negative in case of "up" events.

This ordered convergence, is similar to the ordered FIB proposed defined in [RFC6976], but limited to only one hop distance. As a consequence, it is simpler and becomes a local only feature not requiring interoperability; at the cost of only covering the transient forwarding loops involving this local router. The proposed mechanism also reuses some concept described in [I-D.ietf-rtgwg-microloop-analysis] with some limitation.

4. Specification

4.1. Definitions

This document will refer to the following existing IGP timers:

- o LSP_GEN_TIMER: to batch multiple local events in one single local LSP update. It is often associated with damping mechanism to slowdown reactions by incrementing the timer when multiple consecutive events are detected.
- o SPF_TIMER: to batch multiple events in one single computation. It is often associated with damping mechanism to slowdown reactions by incrementing the timer when the IGP is instable.
- o IGP_LDP_SYNC_TIMER: defined in [RFC5443] to give LDP some time to establish the session and learn the MPLS labels before the link is used.

This document introduces the following two new timers :

- o ULOOP_DELAY_DOWN_TIMER: slowdown the local node convergence in case of link down events.
- o ULOOP_DELAY_UP_TIMER: slowdown the network wide IGP convergence in case of link up events.

4.2. Current IGP reactions

Upon a change of status on an adjacency/link, the existing behavior of the router advertising the event is the following:

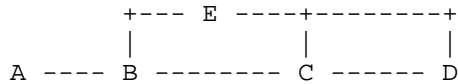
1. UP/Down event is notified to IGP.
2. IGP processes the notification and postpones the reaction in LSP_GEN_TIMER msec.
3. Upon LSP_GEN_TIMER expiration, IGP updates its LSP/LSA and floods it.
4. SPF is scheduled in SPF_TIMER msec.
5. Upon SPF_TIMER expiration, SPF is computed and RIB/FIB are updated.

4.3. Local events

The mechanisms described in this document assume that there has been a single failure as seen by the IGP area/level. If this assumption is violated (e.g. multiple links or nodes failed), then standard IP

convergence MUST be applied. There are three types of single failures: local link, local node, and remote failure.

Example :



Let B be the computing router when the link B-C fails. B updates its local LSP/LSA describing the link B->C as down, C does the same, and both start flooding their updated LSP/LSAs. During the SPF_TIMER period, B and C learn all the LSPs/LSAs to consider. B sees that C is flooding as down a link where B is the other end and that B and C are describing the same single event. Since B receives no other changes, B can determine that this is a local link failure.

[Editor s Note: Detection of a failed broadcast link involves additional complexity and will be described in a future version.]

If a router determines that the event is local link failure, then the router may use the mechanism described in this document.

Distinguishing local node failure from remote or multiple link failure requires additional logic which is future work to fully describe. To give a sense of the work necessary, if node C is failing, routers B,E and D are updating and flooding updated LSPs/LSAs. B would need to determine the changes in the LSPs/LSAs from E and D and see that they all relate to node C which is also the far-end of the locally failed link. Once this detection is accurately done, the same mechanism of delaying local convergence can be applied.

4.4. Local delay

4.4.1. Link down event

Upon an adjacency/link down event, this document introduces a change in step 5 in order to delay the local convergence compared to the network wide convergence: the node SHOULD delay the forwarding entry updates by ULOOP_DELAY_DOWN_TIMER. Such delay SHOULD only be introduced if all the LSDB modifications processed are only reporting down local events . Note that determining that all topological change are only local down events requires analyzing all modified LSP/LSA as a local link or node failure will typically be notified by multiple nodes. If a subsequent LSP/LSA is received/updated and a new SPF computation is triggered before the expiration of ULOOP_DELAY_DOWN_TIMER, then the same evaluation SHOULD be performed.

As a result of this addition, routers local to the failure will converge slower than remote routers. Hence it SHOULD only be done for non urgent convergence, such as for administrative de-activation (maintenance) or when the traffic is Fast ReRouted.

4.4.2. Link up event

Upon an adjacency/link up event, this document introduces the following change in step 3 where the node SHOULD:

- o Firstly build a LSP/LSA with the new adjacency but setting the metric to MAX_METRIC . It SHOULD flood it but not compute the SPF at this time. This step is required to ensure the two way connectivity check on all nodes when computing SPF.
- o Then build the LSP/LSA with the target metric but SHOULD delay the flooding of this LSP/LSA by SPF_TIMER + ULOOP_DELAY_UP_TIMER. MAX_METRIC is equal to MaxLinkMetric (0xFFFF) for OSPF and $2^{24}-2$ (0xFFFFFE) for IS-IS.
- o Then continue with next steps (SPF computation) without waiting for the expiration of the above timer. In other word, only the flooding of the LSA/LSP is delayed, not the local SPF computation.

As as result of this addition, routers local to the failure will converge faster than remote routers.

If this mechanism is used in cooperation with "LDP IGP Synchronization" as defined in [RFC5443] then the mechanism defined in RFC 5443 is applied first, followed by the mechanism defined in this document. More precisely, the procedure defined in this document is applied once the LDP session is considered "fully operational" as per [RFC5443].

5. Applicability

As previously stated, the mechanism only avoids the forwarding loops on the links between the node local to the failure and its neighbor. Forwarding loops may still occur on other links.

5.1. Applicable case : local loops

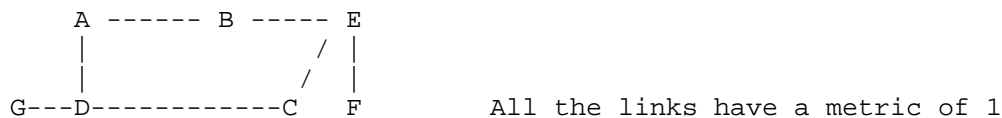
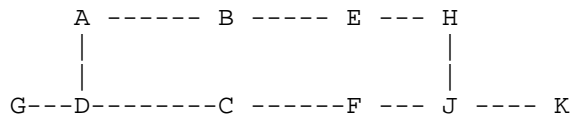


Figure 2

Let us consider the traffic from G to F. The primary path is G->D->C->E->F. When link CE fails, if C updates its forwarding entry for F before D, a transient loop occurs. This is sub-optimal as C has FRR enabled and it breaks the FRR forwarding while all upstream routers are still forwarding the traffic to itself.

By implementing the mechanism defined in this document on C, when the CE link fails, C delays the update of his forwarding entry to F, in order to let some time for D to converge. FRR keeps protecting the traffic during this period. When the timer expires on C, forwarding entry to F is updated. There is no transient forwarding loop on the link CD.

5.2. Non applicable case : remote loops



All the links have a metric of 1 except BE=15

Figure 3

Let us consider the traffic from G to K. The primary path is G->D->C->F->J->K. When the CF link fails, if C updates its forwarding entry to K before D, a transient loop occurs between C and D.

By implementing the mechanism defined in this document on C, when the link CF fails, C delays the update of his forwarding entry to K, letting time for D to converge. When the timer expires on C, forwarding entry to F is updated. There is no transient forwarding loop between C and D. However, a transient forwarding loop may still occur between D and A. In this scenario, this mechanism is not enough to address all the possible forwarding loops. However, it does not create additional traffic loss. Besides, in some cases -such as when the nodes update their FIB in the following order C, A, D, for example because the router A is quicker than D to converge- the mechanism may still avoid the forwarding loop that was occurring.

6. Simulations

Simulations have been run on multiple service provider topologies. So far, only link down event have been tested.

Topology	Gain
T1	71%
T2	81%
T3	62%
T4	50%
T5	70%
T6	70%
T7	59%
T8	77%

Table 1: Number of Repair/Dst that may loop

We evaluated the efficiency of the mechanism on eight different service provider topologies (different network size, design). The benefit is displayed in the table above. The benefit is evaluated as follows:

- o We consider a tuple (link A-B, destination D, PLR S, backup nexthop N) as a loop if upon link A-B failure, the flow from a router S upstream from A (A could be considered as PLR also) to D may loop due to convergence time difference between S and one of his neighbor N.
- o We evaluate the number of potential loop tuples in normal conditions.
- o We evaluate the number of potential loop tuples using the same topological input but taking into account that S converges after N.
- o Gain is how much loops (remote and local) we succeed to suppress.

On topology 1, 71% of the transient forwarding loops created by the failure of any link are prevented by implementing the local delay. The analysis shows that all local loops are obviously solved and only remote loops are remaining.

7. Deployment considerations

Transient forwarding loops have the following drawbacks :

- o Limit FRR efficiency : even if FRR is activated in 50msec, as soon as PLR has converged, traffic may be affected by a transient loop.

- o It may impact traffic not directly concerned by the failure (due to link congestion).

This local delay proposal is a transient forwarding loop avoidance mechanism (like OFIB). Even if it only address local transient loops, , the efficiency versus complexity comparison of the mechanism makes it a good solution. It is also incrementally deployable with incremental benefits, which makes it an attractive option for both vendors to implement and Service Providers to deploy. Delaying convergence time is not an issue if we consider that the traffic is protected during the convergence.

8. Comparison with other solutions

As stated in Section 3, our solution reuses some concepts already introduced by other IETF proposals but tries to find a tradeoff between efficiency and simplicity. This section tries to compare behaviors of the solutions.

8.1. PLSN

PLSN ([I-D.ietf-rtgwg-microloop-analysis]) describes a mechanism where each node in the network tries to avoid transient forwarding loops upon a topology change by always keeping traffic on a loop-free path for a defined duration (locked path to a safe neighbor). The locked path may be the new primary nexthop, another neighbor, or the old primary nexthop depending how the safety condition is satisfied.

PLSN does not solve all transient forwarding loops (see [I-D.ietf-rtgwg-microloop-analysis] Section 4 for more details).

Our solution reuse some concept of PLSN but in a more simple fashion :

- o PLSN has 3 different behavior : keep using old nexthop, use new primary nexthop if safe, or use another safe nexthop, while our solution only have one : keep using the current nexthop (old primary, or already activated FRR path).
- o PLSN may cause some damage while using a safe nexthop which is not the new primary nexthop in case the new safe nexthop does not enough provide enough bandwidth (see [I-D.ietf-rtgwg-lfa-manageability]). Our solution may not experience this issue as the service provider may have control on the FRR path being used preventing network congestion.

- o PLSN applies to all nodes in a network (remote or local changes), while our mechanism applies only on the nodes connected to the topology change.

8.2. OFIB

OFIB ([RFC6976]) describes a mechanism where convergence of the network upon a topology change is made ordered to prevent transient forwarding loops. Each router in the network must deduce the failure type from the LSA/LSP received and compute/apply a specific FIB update timer based on the failure type and its rank in the network considering the failure point as root.

This mechanism permit to solve all the transient forwarding loop in a network at the price of introducing complexity in the convergence process that may require strong monitoring by the service provider.

Our solution reuses the OFIB concept but limits it to the first hop that experience the topology change. As demonstrated, our proposal permits to solve all the local transient forwarding loops that represents a high percentage of all the loops. Moreover limiting the mechanism to one hop permit to keep the network-wide convergence behavior.

9. Security Considerations

This document does not introduce change in term of IGP security. The operation is internal to the router. The local delay does not increase the attack vector as an attacker could only trigger this mechanism if he already has be ability to disable or enable an IGP link. The local delay does not increase the negative consequences as if an attacker has the ability to disable or enable an IGP link, it can already harm the network by creating instability and harm the traffic by creating forwarding packet loss and forwarding loss for the traffic crossing that link.

10. Acknowledgements

We wish to thanks the authors of [RFC6976] for introducing the concept of ordered convergence: Mike Shand, Stewart Bryant, Stefano Previdi, and Olivier Bonaventure.

11. IANA Considerations

This document has no actions for IANA.

12. References

12.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC5443] Jork, M., Atlas, A., and L. Fang, "LDP IGP Synchronization", RFC 5443, DOI 10.17487/RFC5443, March 2009, <<http://www.rfc-editor.org/info/rfc5443>>.
- [RFC5715] Shand, M. and S. Bryant, "A Framework for Loop-Free Convergence", RFC 5715, DOI 10.17487/RFC5715, January 2010, <<http://www.rfc-editor.org/info/rfc5715>>.

12.2. Informative References

- [I-D.ietf-rtgwg-lfa-manageability] Litkowski, S., Decraene, B., Filsfils, C., Raza, K., Horneffer, M., and P. Sarkar, "Operational management of Loop Free Alternates", draft-ietf-rtgwg-lfa-manageability-11 (work in progress), June 2015.
- [I-D.ietf-rtgwg-microloop-analysis] Zinin, A., "Analysis and Minimization of Microloops in Link-state Routing Protocols", draft-ietf-rtgwg-microloop-analysis-01 (work in progress), October 2005.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, DOI 10.17487/RFC3630, September 2003, <<http://www.rfc-editor.org/info/rfc3630>>.
- [RFC6571] Filsfils, C., Ed., Francois, P., Ed., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, DOI 10.17487/RFC6571, June 2012, <<http://www.rfc-editor.org/info/rfc6571>>.
- [RFC6976] Shand, M., Bryant, S., Previdi, S., Filsfils, C., Francois, P., and O. Bonaventure, "Framework for Loop-Free Convergence Using the Ordered Forwarding Information Base (oFIB) Approach", RFC 6976, DOI 10.17487/RFC6976, July 2013, <<http://www.rfc-editor.org/info/rfc6976>>.

[RFC7490] Bryant, S., Filsfils, C., Previdi, S., Shand, M., and N. So, "Remote Loop-Free Alternate (LFA) Fast Reroute (FRR)", RFC 7490, DOI 10.17487/RFC7490, April 2015, <<http://www.rfc-editor.org/info/rfc7490>>.

Authors' Addresses

Stephane Litkowski
Orange

Email: stephane.litkowski@orange.com

Bruno Decraene
Orange

Email: bruno.decraene@orange.com

Clarence Filsfils
Cisco Systems

Email: cfilsfil@cisco.com

Pierre Francois
IMDEA Networks

Email: pierre.francois@imdea.org

PANET Working Group
Internet-Draft
Intended Status: Experimental RFC
Expires: September 2013

Shankar Raman
Balaji Venkat Venkataswami
Gaurav Raina
Vasan Srini
IIT Madras
March 24, 2013

Reducing Power Consumption using BGP path selection
draft-mjsraman-panet-bgp-power-path-02

Abstract

In this paper, we propose a framework to reduce the aggregate power consumption of the Internet using a collaborative approach between Autonomous Systems (AS). We identify the low-power paths among the AS and then use suitable modifications to the BGP path selection algorithm to route the packets along the paths. Such low-power paths can be identified by using the consumed-power-to-available-bandwidth (PWR) ratio as an additional parameter in the BGP Path Selection Algorithm. For re-routing the data traffic through these low-power paths, the power based best path is selected and advertised as per the modified algorithm proposed in this document. Extensions to the Border Gateway Protocol (BGP) can be used to disseminate the PWR ratio metric among the AS thereby creating a collaborative approach to reduce the power consumption. The feasibility of our approaches is illustrated by applying our algorithm to a subset of the Internet. The techniques proposed in this paper for the Inter-AS power reduction require minimal modifications to the existing features of the Internet. The proposed techniques can be extended to other levels of Internet hierarchy, such as Intra-AS paths, through suitable modifications. A recent addition is the use of this method in AIGP domains and also the use of power source data in the calculation of low power paths using the BGP path selection algorithm.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	4
1.1	Low-power routers and switches	4
1.2	Power reduction using routing and traffic engineering	4
1.1	Terminology	5
2.	Methodology	5
2.1	Pre-requisites for the Proposed Method	5
2.1.1	PWR ratio calculation	5
2.1.1.1	Power Sources as additional factor	7
2.1.1.2	Earlier method of computing numerator of PWR ratio.	8
2.2	LOW-POWER PATHS	9
2.2.0.1	Current BGP Best Path Selection Algorithm	10
2.2.0.2	Algorithm 1 on ASBR	12
2.2.0.3	Modified Algorithm 0 on all BGP routers	13
2.3	Implementation notes and Discussion	14
2.4	Applicability within ASes within a single Admin Domain	16
2.4.1	PWR_SESSION	16

2.4.2	Power profiles of Routers and Switches	17
2.4.2.1	Concave and Convex power curves	19
2.4.2.3	Need to advertise both available power and consumed power	20
2.4.3	Conclusion and Future Work	21
2.5	Acknowledgements	22
3	Security Considerations	23
4	IANA Considerations	23
5	References	23
5.1	Normative References	23
5.2	Informative References	23
	Authors' Addresses	24

1 Introduction

Estimates of power consumption for the Internet predict a 300% increase, as access speeds increase from 10 Mbps to 100 Mbps [3], [8]. Access speeds are likely to increase as new video, voice and gaming devices get added to the Internet. Various approaches have been proposed to reduce the power consumption of the Internet such as designing low-power routers and switches, and optimizing the network topology using traffic engineering methods [2].

1.1 Low-power routers and switches

Low-power router and switch design aim at reducing the power consumed by hardware architectural components such as transmission link, lookup tables and memory. In [4] it is shown that the router's link power consumption can vary by 20 Watts between idle and traffic scenarios. Hence the authors suggest having more line cards and running them to capacity: operating the router at full throughput will lead to less power per bit, and hence larger packet lengths will consume lower power. The two important components in routers that have received attention for high power consumption are buffers and TCAMs. Buffers are built using dynamic RAM (DRAM) or static RAM (SRAM). SRAMs are limited in size and consume more power, but have low access times. Guido [1] states that a 40Gb/s line card would require more than 300 SRAM chips and consume 2.5kW. DRAM access times prevent them from being used on high speed line cards. Sometimes the buffering of packets in DRAM is done at the back end, while SRAM is used at the front end for fast data access. But these schemes cannot scale with increasing line speeds. Some variants of TCAMs have been proposed for increasing line speeds and for reduced power consumption [7].

1.2 Power reduction using routing and traffic engineering

At the Internet level, creating a topology that allows route adaptation, capacity scaling and power-aware service rate tuning, will reduce power consumption. In [8] the author has proposed a technique to traffic engineer the data packets in such a way that the link capacity between routers is optimized. Links which are not utilized are moved to the idle state. Power consumption can be reduced by trading off performance related measures like latency. For example, power savings while switching from 1 Gbps to 100 Mbps is approximately 4 W and from 100 Mbps to 10 Mbps around 0.1 Watts. Hence instead of operating at 1 Gbps the link speed could be reduced to a lower bandwidth under certain conditions for reduced power consumption.

Multi layer traffic engineering based methods make use of parameters

such as resource usage, bandwidth, throughput and QoS measures, for power reduction. In [6] an approach for reducing Intra-AS power consumption for optical networks that uses Dijkstra's shortest path algorithm is proposed. The input to this method assumes the existence of a network topology using which an auxiliary graph is constructed. Power optimization is done on the auxiliary graph and traffic is routed through the low-power links. However, the algorithm expects the topology to be available for getting the auxiliary graph. This topology is easy to obtain for Intra-AS scenario, but not for Inter-AS cases. In our approach, we propose a collaborative approach by AS in power reduction. The core of the Internet at the Inter-AS level, uses the BGP best path selection algorithm. The AS use the Border Gateway Protocol (BGP) for exchanging routing related information. One of the attributes of BGP namely, AS-PATH-INFO is used to derive the topology of the Internet at the AS level. In this document we propose that the BGP best path selection algorithm is run in each AS at an appropriate BGP router with the consumed-power-to-available-bandwidth (PWR) ratio as a parameter, to determine the low-power paths from the head-end to the tail-end AS in order to reach a prefix or a set of prefixes. The PWR ratio can be exchanged among the collaborating AS using BGP attributes.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Methodology

<Document text>

2.1 Pre-requisites for the Proposed Method

In this section we discuss the pre-requisites for the implementation of the proposed scheme.

2.1.1 PWR ratio calculation

In this proposal each AS is expected to share its PWR ratio from as many ASBRs (Autonomous System Border Routers) that it has. Intuitively in order to calculate this ratio we need to calculate the consumed power representative of the AS and the maximum bandwidth available with an ASBR on its egress links into the AS. The entry point to the AS is through the ASBRs that advertise the prefixes reachable through the AS. Hence the numerator of the PWR ratio is

calculated for the AS at each ingress ASBR. We first obtain the summation of power consumed at the Provider (P) and the Provider Edge (PE) routers within an AS. The numerator of the PWR ratio is calculated by summing up the consumed power of all the routers to be taken into account and then dividing this sum by the number of routers. A more intuitive approach would be to use a weighted average method by assigning routers to categories and having appropriate coefficients for each of these categories, thus arriving at a weighted average which is more accurate. One of these alternatives can be used to arrive at the numerator of the PWR ratio. Yet another alternative would have been to sum up the total consumed power of all routers in the AS and represent that as the numerator of the PWR ratio.

This average consumed power is divided by the maximum bandwidth available at each of the ASBR's egress link. This step is necessary as the requested bandwidth for any path from the head-end to the tail-end using the ASBR is limited by the bandwidth available in the ASBR's egress links. The highest available bandwidth amongst the egress links of the ASBR is used as the denominator in the PWR ratio computation. If the entry point to the AS is through a different ASBR then the PWR ratio assigned to the ingress link of the ASBR might vary. Hence, an head-end AS might see different PWR ratios for an intermediate AS, if the intermediate AS has different ASBRs as its entry point.

We now illustrate the PWR ratio calculation. Consider an AS X which is one of the AS in the vicinity of another AS Y. Let this ASBR of X have 3 egress links into X denoted as E(1), E(2) and E(3), and 2 ingress links labeled I(1) and I(2). We now calculate the PWR ratio for I(1) and I(2). Assume that the routers in X have average consumed power of 200K Watts per hour. From figure 4 we can calculate the PWR ratio for I(1) and I(2) as $200K \text{ Watts} / (60 * 60 * 1.5 \text{ Gigb}) = 3.7037 * (10 \text{ raised to } -8)$. We could scale this to 0.37087 by multiplying with a base value of 10 raised to the 7th power.

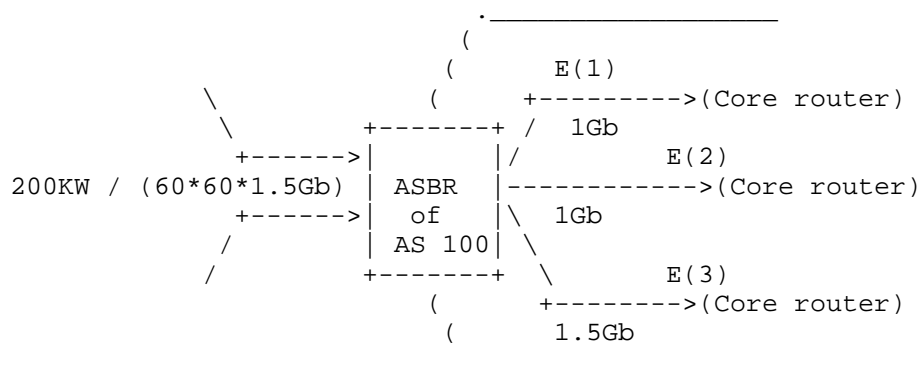


Figure 1: Calculation of PWR ratio by an ASBR associated with an AS. The I represents ingress links and E represents egress links. 200KW is the average consumed power in the AS. 1.5Gb is the maximum available bandwidth of the egress link in an ASBR.

Note that this ratio is actually a mapping function that is defined for each of the ingress links of the ASBR associated with an AS. For the head-end which is the BGP Path selection running AS this mapping function does not exist as there is no ingress link. The PWR ratio can then be advertised to the other neighboring AS using the control plane through BGP extensions. BGP ensures that the information is percolated to other AS beyond the immediate neighbors. On receipt of these power metrics to the AS at the far-ends of the Internet, the overall AS level PWR ratio based Internet topology can be constructed. This view of the Internet is available with each of the routers without using any other complex discovery mechanism. Some sample link weights shown in Figure 1 is obtained by using such a mapping function on the ingress links.

2.1.1.1 Power Sources as additional factor

It is envisaged that the power sources of the Autonomous system using which the routers in the AS are powered should be declared as a metric which is further incorporated in the PWR ratio.

A suitable weight is provided to each type of source and the following table which is not claimed as totally exhaustive can be used to add this metric in the equation to compute the PWR ratio.

A formal classification of power sources and their weights is a topic to be considered later. For now we will deal with 2 main categories. Renewable sources of energy and non-renewable sources. There would be multiple categories under each of these major categories. Each such power source is assigned a weight.

Renewable Sources of Energy :

Wind - HighWeightOne
Solar - HighWeightTwo
Hydro - HighWeightThree
etc...

Non-renewable Sources of Energy :

Natural Gas - LowWeightOne
Petroleum and Diesel - LowWeightTwo
Nuclear - LowWeightThree
etc...

The PWR-SOURCE ratio is calculated in the proportion of how the above sources are combined to power the routers and its coolant systems and ancillary facilities in the AS.

Thus $PWR-RATIO = (Consumed-Power / Available-Bandwidth)$
 $* (1 / Weighted\ Average\ of\ Power\ Sources)$

This compound metric could be used as the PWR metric in the calculations specified in this draft.

2.1.1.2 Earlier method of computing numerator of PWR ratio.

Earlier in the previous versions of this document in order to calculate this PWR ratio we needed to calculate the available power and the maximum bandwidth available with an ASBR. The entry point to the AS is through ASBRs that advertise the prefixes reachable through the AS. Hence, the numerator of the PWR ratio is calculated for the AS at each ingress ASBR. We first obtained the summation of power consumed at the major Provider (P) and Provider Edge (PE) routers within an AS. The average available power is obtained by subtracting the consumed power from the maximum power rating and summing the values for all the routers and then dividing the result by the number of routers. As an alternative, one could use a weighted average for more accuracy depending on the category of the router advertising the consumed power. Yet another alternative is to take the average or sum of the maximum power rating of all the routers within an AS without taking into account the consumed power. One of these alternatives was chosen to calculate the numerator of the PWR ratio.

Intuition however drives us towards consumed power as a better numerator since the lesser the power consumed the lesser the numerator and hence lesser the ratio if enough bandwidth is available at the ingress ASBR. The amount of consumed power per bit of information ought to be low for the shortest path to work out

properly. One more aspect is that lesser the power consumed per available bit of bandwidth it could be a sign that routers are more optimal in their power consumption as they take on more traffic. This is a very crucial point to be considered.

However additional research seems to indicate that both Available and Consumed Power for a router be advertised. The need that arises for such a proposition is that there exist power profiles of routers which is dealt in later sections (section 2.4.2). Please refer section 2.4.2.1 onwards for more analysis and research on this subject.

2.2 LOW-POWER PATHS

In this section we present the low-power path BGP best path selection algorithm. The algorithm consists of two sub-algorithms: the first algorithm is executed by all the ASBRs in the network and the second by all the BGP routers in their respective AS. The algorithms for the ASBRs and BGP routers are given as Algorithm 1 and 2. The algorithm in 2.2.0.1 is the current BGP best path algorithm and is titled Algorithm 0.

2.2.0.1 Current BGP Best Path Selection Algorithm

As taken from [11] the following is the current BGP Best Path Selection Algorithm.

Algorithm 0 : BEGIN

BGP assigns the first valid path as the current best path. BGP then compares the best path with the next path in the list, until BGP reaches the end of the list of valid paths. This list provides the rules that are used to determine the best path:

- 1) Prefer the path with the highest WEIGHT.
- 2) Prefer the path with the highest LOCAL_PREF.
- 3) Prefer the path that was locally originated via a network or aggregate BGP subcommand or through redistribution from an IGP.

Local paths that are sourced by the network or redistribute commands are preferred over local aggregates that are sourced by the aggregate-address command.

- 4) Prefer the path with the shortest AS_PATH.

An AS_SET counts as 1, no matter how many ASs are in the set.

The AS_CONFED_SEQUENCE and AS_CONFED_SET are not included in the AS_PATH length.

- 5) Prefer the path with the lowest origin type.

Note: IGP is lower than Exterior Gateway Protocol (EGP), and EGP is lower than INCOMPLETE.

- 6) Prefer the path with the lowest multi-exit discriminator (MED).

- 7) Prefer eBGP over iBGP paths.

If bestpath is selected, go to Step 9 (multipath).

Note: Paths that contain AS_CONFED_SEQUENCE and AS_CONFED_SET are local to the confederation. Therefore, these paths are treated as internal paths. There is no distinction between Confederation External and Confederation Internal.

- 8) Prefer the path with the lowest IGP metric to the BGP next hop.

Continue, even if bestpath is already selected.

9) Determine if multiple paths require installation in the routing table for BGP Multipath.

Continue, if bestpath is not yet selected.

10) When both paths are external, prefer the path that was received first (the oldest one).

This step minimizes route-flap because a newer path does not displace an older one, even if the newer path would be the preferred route based on the next decision criteria (Steps 11, 12, and 13).

Skip this step if any of these items is true:

You have enabled the `bgp best path compare-routerid` command.

The router ID is the same for multiple paths because the routes were received from the same router.

There is no current best path.

The current best path can be lost when, for example, the neighbor that offers the path goes down.

11) Prefer the route that comes from the BGP router with the lowest router ID.

The router ID is the highest IP address on the router, with preference given to loopback addresses. Also, you can use the `bgp router-id` command to manually set the router ID.

Note: If a path contains route reflector (RR) attributes, the originator ID is substituted for the router ID in the path selection process.

12) If the originator or router ID is the same for multiple paths, prefer the path with the minimum cluster list length.

This is only present in BGP RR environments. It allows clients to peer with RRs or clients in other clusters. In this scenario, the client must be aware of the RR-specific BGP attribute.

13) Prefer the path that comes from the lowest neighbor address.

This address is the IP address that is used in the BGP neighbor configuration. The address corresponds to the remote peer that is

used in the TCP connection with the local router.

Algorithm 0: END

2.2.0.2 Algorithm 1 on ASBR

```
1: Begin
2: if ROUTER == ASBR then
3: /* As part of IGP-TE */
4: Trigger exchange of available bandwidth on bandwidth change,
   to the AS internal neighbors;
5: BEGIN PROCESS 1
6: while PWR ratio changes do
7: Assign the PWR ratio to the Ingress links;
8: Exchange the PWR ratio with its external neighbors;
9: Exchange the PWR ratio with AS's (internal) ASBRs;
10: end while
11: END PROCESS 1
12: End
```

2.2.0.3 Modified Algorithm 0 on all BGP routers

```
1: Begin
2: If ROUTER is Configured with BGP then

3: Run all steps from 1 to 3 in BGP regular path selection algorithm;
  /* when comparing AS_PATHS (MODIFICATION HERE) */
4: Check if there are no multiple AS_PATHS then goto regular step
  (4);
5: if PWR metric based path selection is configured then
6:     For each AS_PATH(1..n) in this set in step (4)
7:         if there exists a PWR metric for all
           elements in AS_PATH then
8:             PWR_SUM[i] = sum the PWR
               metrices for that AS_PATH;
9:         else
10:            ignore the AS_PATH;
11:        endif
12:    endFor

13:    If there exists multiple PWR_SUM[i] then
14:        Choose the AS_PATH / AS_PATHS with
           least PWR_SUM;
15:        if multiple least PWR_SUMs (equal valued)
           exist then
16:            Take up the set of such
               AS_PATHS and goto step 5;
17:        endif
18:    else
19:        if there exist no PWR_SUM because of
           exclusion then
20:            do regular step(4)
               to select best paths;
21:        endif
22:    endif
23: else

24: Do regular step(4);

25: endif

26: Run all steps from 5 to 13 in BGP regular path selection
    algorithm;
27: endif
28: End
```

It is to be noted that the PWR metric based path selection will ensue only if the modified steps are activated as a result of specific user configuration.

2.3 Implementation notes and Discussion

We propose addition of some BGP attributes with no change to the protocol implementation. There may be a time lag when the far ends of the Internet receive the attribute and the time it originated. This however cannot be avoided as with other attributes and metrics.

0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7

Owning 32 bit Autonomous System Number																															

Other 32 bit Autonomous System Number																															

PWR Ratio for the AS (Consumed PWR)																															

PWR Ratio for the AS (Available PWR)																															

Advertising ASBR's IP router ID																															

Peer ASBR's IP router ID																															

64 bit sequence number for restarts, aging																															
and comparison of current PWR Ratio.																															

Figure 2: Proposed PDU format with an added attribute for AS-PATH-POWER-METRIC

The additions to the above Attribute have been added to optimize and correctly correlate the connecting ASes and the inter-AS links among them. For the traffic direction into the Advertising AS the above information will be easier to correlate than the previous version which did not advertise the peer AS which had the ingress links into the advertising Router.

In MPLS-TE for example, when the TE metrics are modified, there is a reliable flooding process within an Interior Gateway Protocol (IGP). Such triggered updates apply to the PWR ratio in BGP as well. The proposed PWR ratio is advertised to the neighboring AS and the information percolated to all the AS, in a AS-PATH-POWER-METRIC attribute. This attribute can be implemented as shown in Figure 2. The frequency of the updates for this attribute should be fixed to avoid network flooding.

The AS-PATH-POWER-METRIC for each ASBR is calculated, and advertised as the PWR ratio for the AS. This AS-PATH-POWER-METRIC is filled into the appropriate transitive non-discretionary attribute and inserted into a unique vector for a set of prefixes advertised from the AS. Such advertised prefixes may have originated from the AS or be the transit prefixes. The filled vector is sent to the ASBR of the neighboring AS and the information propagates to all the ASBRs. If the elements denoting AS in a vector of AS-PATH-INFO is not the same as the ones that need to be advertised in a AS-PATH-POWER-METRIC, then a suitable subset of AS-PATH-POWER-METRIC is identified and sent in the BGP updates. A vector of size 1 also can be employed if the AS in question is the only one for which PWR ratio has changed in the originating AS. The collation can be done depending on availability of such metrics and their mapping to a valid AS-PATH-INFO metric.

The power consumed by each router may fluctuate over short time intervals. In order to dampen these fluctuations which can cause unnecessary updates, power can be measured when falling within intervals of suitable size (say a range of values). This is as opposed to measuring power as a discrete quantity. This method of power measurement reduces the frequency of triggered updates from the routers due to power change.

```

0.1      0.2      0.1
(A) ----> (B) ----> (D)

0.1      0.2      0.02      0.2
(A) ----> (C) ----> (E) ----> (D)

0.1      0.2
(D) ----> (X)

```

Figure 4: Example of strands where more than one PWR ratio is advertised by "D"

```

      0.2      0.1      0.2
(A).....>(B).....>(D).....>(X)
|           ^
|0.2      0.02      | 0.2
+---->(C)----->(E)

```

Figure 5:Choice of low-power path derived using the algorithm which uses lower value of the ingress link but through the same AS

A use case of multiple ASBRs advertising differing PWR ratio shows that an AS may be seen as green through one ingress link and not through the other. Consider the case of multiple ASBRs that belong to

the same AS, advertising PWR ratios that differ. This could lead to power values that belong to different classes of ratios with many intervening classes in between. These advertised PWR ratios could lead to one ASBR being preferred over the other thus taking a different path from head-end to tail-end. This also entails that there may be multiple paths to the AS through these different ASBRs.

Consider Figure 4 which shows a set of strands that derive a topology as in Figure 5. Here D is reachable via two paths but the PWR ratios differ. This illustrates the case where the better metric wins out. The average power consumed would not have an effect but the bandwidth available on these ASBR egress links would definitely influence the path.

2.4 Applicability within ASes within a single Admin Domain

As per [draft-ietf-idr-aigp] there are deployments in which a single administration runs a network which has been sub-divided into multiple, contiguous ASes, each running BGP. There are several reasons why a single administrative domain may be broken into several ASes (which, in this case, are not really "autonomous".) It may be that the existing IGPs do not scale well in the particular environment; it may be that a more generalized topology is desired than could be obtained by use of a single IGP domain; it may be that a more finely grained routing policy is desired than can be supported by an IGP. In such deployments, it can be useful to allow BGP to make its routing decisions based on the IGP metric, so that BGP chooses the "shortest" path between two nodes, even if the nodes are in two different ASes within that same administrative domain. The authors refer to the set of ASes in a common administrative domain as an "AIGP Administrative Domain".

A combination of the AIGP administrative metric and the Path selection algorithm could be combined to arrive at a set of a suitable number of equal k power-shortest paths and then use a tie-break amongst such k power-shortest-paths with the least AIGP metric. This is provided the set of ASes where the decision is being made all fall under a AIGP Administrative domain. This provides a trade-off of power shortest paths and least number of hops (link wise) to get from source to destination across these ASes.

2.4.1 PWR_SESSION

An implementation that supports the PWR attribute CAN support a per-session configuration item, PWR_SESSION, that indicates whether the PWR attribute is enabled or disabled for use on that session.

- The default value of PWR_SESSION, for EBGp sessions, between

providers (distinct operators) CAN be "disabled".

- The default value of PWR_SESSION, for IBGP and confederation-EBGP sessions, MUST be "enabled."

The PWR attribute MUST NOT be sent on any BGP session for which PWR_SESSION is disabled.

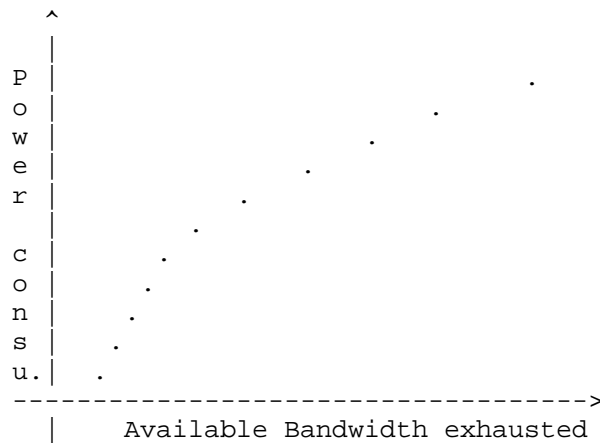
If an PWR attribute is received on a BGP session for which PWR_SESSION is disabled, the attribute MUST be treated exactly as if it were an unrecognized transitive attribute. That is, " The handling of an unrecognized optional attribute is determined by the setting of the Transitive bit in the attribute flags octet. Paths with unrecognized transitive optional attributes SHOULD be accepted. If a path with an unrecognized transitive optional attribute is accepted and passed to other BGP peers, then the unrecognized transitive optional attribute of that path MUST be passed, along with the path, to other BGP peers with the Partial bit in the Attribute Flags octet set to 1. If a path with a recognized, transitive optional attribute is accepted and passed along to other BGP peers and the Partial bit in the Attribute Flags octet is set to 1 by some previous AS, it MUST NOT be set back to 0 by the current AS".

This helps in confining the distribution of the attribute and use in calculation of the power shortest paths only amongst ASes that have trust relationships with other ASes. Of course, this includes and promotes the use of PWR attribute within a AIGP administrative domain.

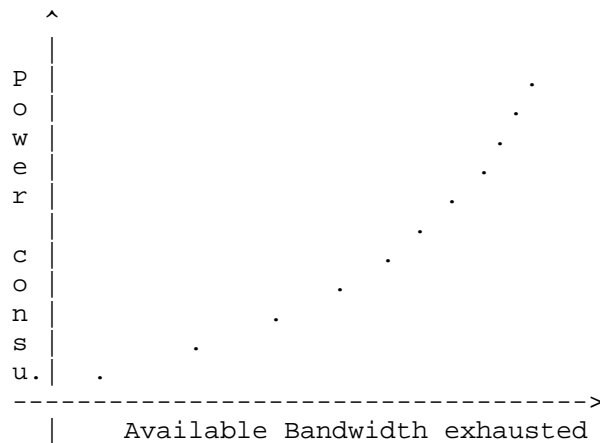
2.4.2 Power profiles of Routers and Switches

It has been experimented and from several sources found that there exist routers which have different power profiles. The power profile of a router is the curve of power consumption to available bandwidth. Mentioned below are a few of these prominent ones that have to be taken into consideration.

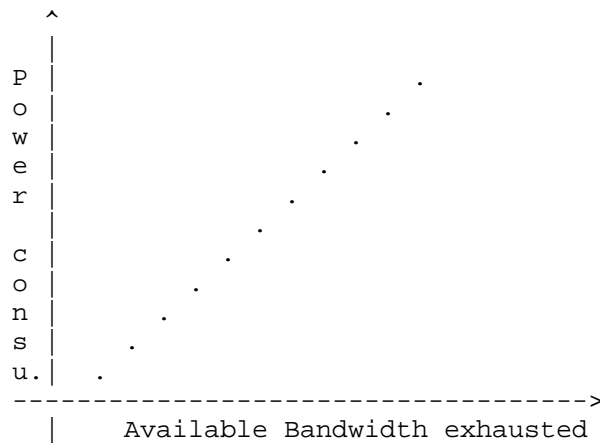
The first profile that we will consider is the flattening curve. The power consumed to available bandwidth curve takes the shape of a steep one initially and then tapers off to a plateau. The point at which it begins to give a delta-C (delta in Power Consumed) to delta-B (Available Bandwidth exhausted) is the inflection point that tapers off to a plateau. Here the delta-C/delta-B begins to slow down or decrease rapidly. The more the traffic that is added onto the device the lesser it draws power.



The second profile that we will consider is the exponential curve. The power consumed to available bandwidth curve takes the shape of an ever increasing steep curve as shown below. Here the $\Delta C / \Delta B$ begins to increase as more traffic is thrown onto it as the Available bandwidth exhausted increases. This power curve beyond a point is intolerable with respect to power guzzling.



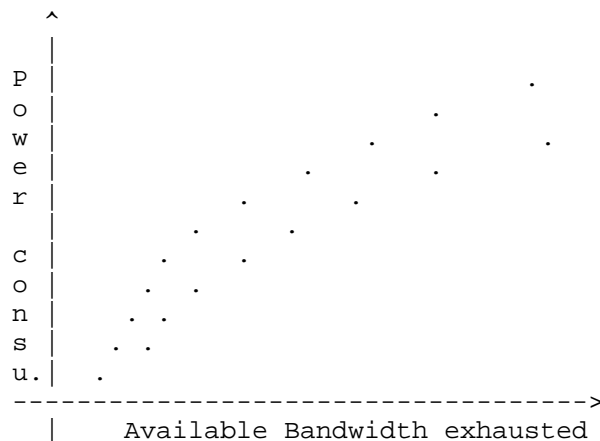
The third profile that we will consider is a linear curve. In other words just a straight line. Here $\Delta C / \Delta B$ is a constant.



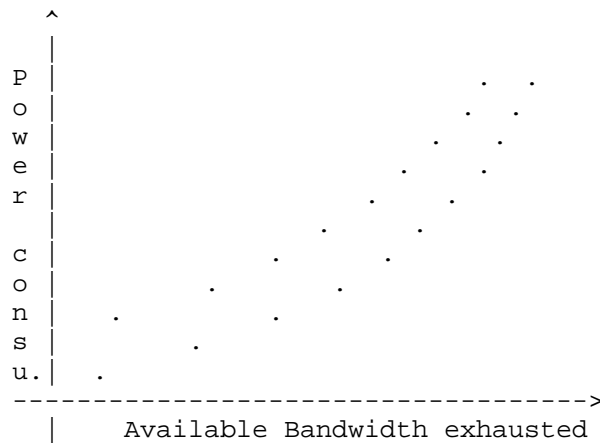
2.4.2.1 Concave and Convex power curves

Given that there are 3 kinds of major profiles in the router power consumption, what line would we like to pick. This is an important point when choosing the metric to pick the low power paths.

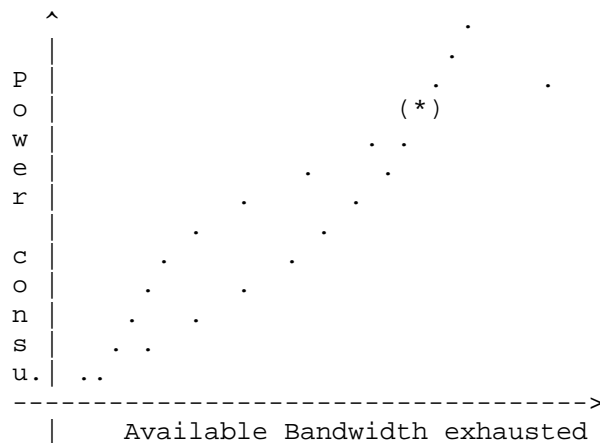
(a) If the confrontation is between 2 first profile routers the lower of the 2 would be considered as shown below. The lower curve offers better power savings for each GB of bandwidth transported.



(b) If the confrontation is between 2 second profile routers the upper curve offers more power savings per GB of bandwidth.



(c) When the confrontation is between a first profile curve and a second profile curve, it would be optimal to pick (as shown below) the lower of the curves because it gives us lesser power consumed for every GB of traffic routed / switched. Here the exponential curve is the one that offers lesser amount of power consumed per GB of traffic is chosen. But when it gets to a point that the two curves intersect it would be more optimal to pick the tapering curve. Thus at the meeting point of the 2 curves the exponential curve becomes more costly and the tapering one gives us more GB for the power buck. Thus this switchover from one curve to the other (in other words from the exponential curve to the tapering one) does the trick in terms of finding an optimal solution.



(*) Metric switchover point from Consumed Power to Available Power.

2.4.2.3 Need to advertise both available power and consumed power

Thus the above sections have shown that both the available power and the consumed power MUST be advertised so that case (c) can be deciphered and the switchover of the curves be done and the appropriate router be chosen for the rest of the bandwidth to be switched over to.

Thus there will exist Consumed-Power to Available Bandwidth ratio and the Available Power to Available Bandwidth ratio. Both the ratios are computed and the lower value chosen. The Available Power can be judged from the calibration process such as the one carried out by independent test organizations as in [12]. An example of their calibration is referred to in [12].

2.4.3 Conclusion and Future Work

In this paper, we proposed a scheme for reducing the power consumption of the Internet using collaborative effort between AS. The BGP best path algorithm is run with suitable modifications in step (4) as described by using the PWR ratio as a parameter. The PWR ratio is advertised through the ingress links of the ASBRs associated with AS using BGP updates. The Modified BGP Best Path Selection Algorithm finds out the low-power consuming AS that can route data packets for a set of prefixes. This entails adopting routes by choosing entry points to an AS that give energy saving paths. Our work complements the current schemes for reducing power consumption within a router such as switching off or bringing to power-idle-state certain select components within the forwarding and lookup mechanisms.

Normally the ASes have SLA agreements between each other to carry X amount of traffic from say a provider A. If the AS representing the ISP then advertises fake figures to carry more traffic than is mandated by the SLA agreement with other providers, then it is to that ISPs detriment since by advertising a better PWR ratio it invites more traffic through it thus getting paid less and carrying more traffic. This is not in the best interest of the ISP. This is so because in the final analysis the Power Shortest Path computed would include it regardless of the amount of traffic to be carried thus causing it to invite more traffic through it than it has accepted, even much more than its capacity. Hence it would be advisable for that ISP to advertise proper PWR ratios and NOT on the lower side of the spectrum. If it advertises HIGHER PWR ratios it would not be chosen, and hence that could be a policy measure NOT to accept any traffic at all since its capacity may be filled up with existing traffic. So advertising on the LOWER side would lead to lesser amount of benefit with respect to dollar per bit transported, and on the HIGHER side would be to exclude it from carrying any traffic that wanted to use the Power Shortest Path.

We also propose that there be a governing body in the IETF or outside it or sponsored by the IETF to verify the power ratios advertised are indeed valid or approximately closer to the actual consumption. A link up for each ISP with a power application level gateway to ensure proper ratios are advertised could be mandated amongst at least the co-operating ISPs (ASes).

The aspect of innovation in this proposal is to use BGP as the piggyback protocol upon which this scheme stands.

When links and switches are gated or put into low-power state within an AS, the power-consumption automatically drops at the aggregate level, as a result of which the PWR ratio would be a lower figure advertised through BGP and thus this AS would attract more Power Shortest Path traffic through it. Thus the links within the AS and the switches within it would function more optimally if it had more traffic that went along paths that were originally put in low-power state thus utilizing the paths more effectively, when attracting traffic.

There exist MIBs today that have object identifier for power consumed in a router. Maybe all the related components within it may NOT be listed with regards to power consumed. But the overall power consumed by the Router / Switch is gettable. Once it is advertised in an opaque Link-State-Advertisement say in the form of a TLV (Type Length Value) and the LSAs (Link State Advertisements) are flooded through the network in an AS, all routers get a uniform picture of which router consumes what power. This method already exists for Traffic engineering Database LSAs that are advertised as LSAs for the purpose of traffic engineering within an AS. We are merely piggybacking on this capability to calculate the PWR ratio at the ASBR which amongst others is yet another Router / Switch of the AS.

Our future work includes looking into computing low-power paths within AS as well.

2.5 Acknowledgements

Shankar Raman would like to acknowledge the support by BT Public Limited (UK) under the BT IITM PhD Fellowship award. Balaji Venkat and Gaurav Raina would like to acknowledge the UK EPSRC Digital Economy Programme and the Government of India Department of Science and Technology (DST) for funding given to the IU-ATC. Vasan Srini would like to thank Dr.(Prof).Kamakoti of the Computer Science and Engineering department for his support.

3 Security Considerations

No specific security considerations apart from the usual considerations with respect to authenticating BGP messages / updates from BGP neighbors is necessary for this scheme.

4 IANA Considerations

A new optional transitive non-discretionary attribute needs to be provided by IANA for carrying the PWR ratio across the Internet in the specified format in BGP.

5 References

5.1 Normative References

TBD

5.2 Informative References

REFERENCES

- [1] G. Appenzeller, Sizing router buffers, Doctoral Thesis, Department of Electrical Engineering, Stanford University, 2005.
- [2] A. P. Bianzino, C. Chaudet, D. Rossi and J. L. Rougier, A survey of green networking research, IEEE Communications and Surveys Tutorials, preprint.
- [3] J. Baliga, K. Hinton and R. S. Tucker, Energy consumption of the internet, Proc. of joint international conference on optical internet, June 2007, pp. 1-3.
- [4] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsang and S. Wright, Power awareness in network design and routing, Proc. of the IEEE INFOCOM 2008, April 2008, pp. 457-465.
- [5] B. Venkat et.al, Constructing disjoint and partially disjoint InterAS TE-LSPs, USPTO Patent 7751318, Cisco Systems, 2010.
- [6] M. Xia et. al., Greening the optical backbone network: A traffic engineering approach, IEEE ICC Proceedings, May 2010, pp. 1-5.

- [7] W. Lu and S. Sahni, Low-power TCAMs for very large forwarding tables, IEEE/ACM Transactions on Computer Networks, June 2010, vol. 18, no. 3, pp. 948-959.
- [8] B. Zhang, Routing Area Open Meeting, Proceedings of the IETF 81, Quebec, Canada, July 2011.
- [9] M.J.S Raman, V.Balaji Venkat, G.Raina, Reducing Power consumption using the Border Gateway Protocol, IARIA conferences ENERGY 2012.
- [10] A.Cianfrani et al., An OSPF enhancement for energy saving in IP Networks, IEEE INFOCOM 2011 Workshop on Green Communications and Networking
- [11] http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a0080094431.shtml, BGP best path selection algorithm.
- [draft-ietf-idr-aigp] P. Mohapatra et.al, The Accumulated IGP metric attribute for BGP, <https://datatracker.ietf.org/doc/draft-ietf-idr-aigp/>, November 2012.

Authors' Addresses

Shankar Raman
Department of Computer Science and Engineering
IIT Madras
Chennai - 600036
TamilNadu
India.

EMail: mjsraman@cse.iitm.ac.in

Balaji Venkat Venkataswami
Department of Electrical Engineering
IIT Madras
Chennai - 600036
TamilNadu
India.

Email: balajivenkat299@gmail.com

Prof.Gaurav Raina
Department of Electrical Engineering
IIT Madras
Chennai - 600036
TamilNadu
India.

Email: gaurav@ee.iitm.ac.in

Vasan Srini
Department of Computer Science and Engineering
IIT Madras
Chennai - 600036
TamilNadu
India.

Email: vasan.vs@gmail.com

PANET Working Group
Internet-draft
Intended Status: Standards Track
Expires: May 9, 2013

Shankar Raman
Balaji Venkat Venkataswami
Gaurav Raina
I.I.T Madras.
November 5, 2012

PIM ECMP Redirect based on Linecard Replication Capacity and Power
draft-mjsraman-panet-ecmp-redirect-power-repl-cap-00

Abstract

This work derives itself from [1] which proposes a ECMP redirect from a PIM upstream neighbor that instructs or advices the PIM downstream neighbor to choose another of its own ECMP links between the former and the latter. What we propose in this document is a criterion based on power consumed in the linecards that comprise the ECMP links between the former and the latter. Also the multicast replication capacity available within the said linecards which form the ECMP links between the two is taken into consideration while making a ECMP redirect.

A PIM router uses RPF procedure to select an upstream interface and router to build forwarding state. When there are equal cost multiple paths (ECMP), existing implementations often use hash algorithms to select a path. Such algorithms do not allow the spread of traffic among the ECMPs according to administrative metrics. This usually leads to inefficient or ineffective use of network resources. This document introduces the ECMP Redirect, a mechanism to improve the RPF procedure over ECMPs. It allows ECMP path selection to be based on administratively selected metrics, such as data transmission delays, path preferences and routing metrics.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	3
1.1	Terminology	3
2.	Conditions where this mechanism applies.	3
3	Security Considerations	6
4	IANA Considerations	6
5	References	6
5.1	Normative References	6
5.2	Informative References	6
	Authors' Addresses	6

1 Introduction

This work derives itself from [1] which proposes a ECMP redirect from a PIM upstream neighbor that instructs or advises the PIM downstream neighbor to choose another of its own ECMP links between the former and the latter. What we propose in this document is a criterion based on power consumed in the linecards that comprise the ECMP links between the former and the latter. Also the multicast replication capacity available within the said linecards which form the ECMP links between the two is taken into consideration while making a ECMP redirect.

A PIM router uses RPF procedure to select an upstream interface and router to build forwarding state. When there are equal cost multiple paths (ECMP), existing implementations often use hash algorithms to select a path. Such algorithms do not allow the spread of traffic among the ECMPs according to administrative metrics. This usually leads to inefficient or ineffective use of network resources. This document introduces the ECMP Redirect, a mechanism to improve the RPF procedure over ECMPs. It allows ECMP path selection to be based on administratively selected metrics, such as data transmission delays, path preferences and routing metrics.

As mentioned earlier this document also proposes the use of the power being consumed by the linecards (if the ECMP links fall on multiple linecards with respect to their ECMP link ports) and the available replication capacity of the linecard within its ASICs as a measure of which ECMP link to which the PIM-Join is to be redirected to. If for example there exist multiple replication engines within different linecards then the lightly loaded replication engine and its corresponding ECMP link on that linecard can be recommended to the PIM downstream neighbor which sends the upstream neighbor a PIM join for a specific (S,G) or (*,G) group.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Conditions where this mechanism applies.

Assume there are multiple ECMP links between a PIM upstream and

downstream router. Lets assume they fall on different linecards on the upstream neighbor with respect to the port placement of these ECMP links. Assume one linecard A, which is already carrying multiple data streams in either the incoming direction on its ports and replicating it to multiple other outgoing linecards through the switch fabric. Assume another linecard B that is lightly loaded with respect to multicast traffic and heavily loaded with respect to the unicast traffic. Assume another linecard C which is lightly loaded with respect to both. Now also assume that these linecards A,B and C are all members of the group of linecards whose ports place themselves within the ECMP links between the upstream and downstream neighbor.

It is possible to decipher from this is that linecard C would be a better point of placement of the replication for the group for which the PIM-Join comes from the downstream neighbor. Assume now that the downstream neighbor selects the linecard A. It is now possible as a result of using the mechanism dictated to in [1], to send a redirect to the downstream neighbor that it would be a better choice to choose linecard C. This it does by recommending the neighbor address field as the port IP address which falls on linecard C. The PDU format for the ECMP redirect specified in [1] and the procedures that go along with it follow.

It is also possible to make this decision in combination with the above or solely on a metric which we will call PWR-REPLIC-CAP. This metric is derived as follows...

PWR-REPLIC-CAP = Power Consumed on that linecard

Available replication capacity on the linecard.

In the metric case the lowest PWR-REPLIC-CAP metric is chosen. This optimizes on the power being spent on replication to the extent possible.

It is important to note that a linecard may have multiple replication engines and ports assigned to each replication engine or to all of them. It is possible that the ECMP links and their ports fall on the same linecard and in that case it would be possible to choose a specific replication engine from among the multiple replication engines available. that has better available replication capacity by choosing a neighbor address belonging to the port that falls in a set that belongs to the superior replication engine (with respect to the available replication capacity at that point in time). If there is no distinction amongst ports with respect to the multiple replication then it actually makes no difference since it would be an internal decision as to where the replication for that port actually happens.

It is also important to note here that the linecards that are multicast capable have well advertised replication capacities which the vendors advice in the linecard data sheets. This could be placed in a variable that can be monitored for shifts within intervals of threshold values. The same goes for the power consumed by the linecards as well.

Pseudo code for the steps to be followed to implement this scheme is as follows...

```

if (multiple ECMP links exist to the PIM neighbor
    from which PIM-Join was received) then

    Get the list of ECMP ports which are members of the ECMP links;

    Get the list of linecards on which these ECMP ports are placed;

    LCA = Consider the least used linecard with respect
    to the replication capacity;

    if (port on LCA is the same on which PIM-Join was received)

        do nothing; return;

    else if (choice is to be made on replication capacity)

        LCA = Consider the least used linecard with respect
        to replication capacity;

    else if (choice is to be made on PWR-REPLIC-CAP)

        LCA = Consider the linecard with the best
        PWR-REPLIC-CAP metric;

    end if

    Send PIM redirect to PIM downstream neighbor
    recommending LCA ECMP link;

end if

```

3 Security Considerations

<Security considerations text>

4 IANA Considerations

<IANA considerations text>

5 References

5.1 Normative References

5.2 Informative References

[1] Yiqun, Cai et.al, Protocol Independent Multicast ECMP Redirect, "draft-ietf-pim-ecmp-02.txt", Work in Progress, October 2011.

[2] Shankar Raman, et.al, Building power optimal Multicast Trees, "draft-mjsraman-rtgwg-pim-power-01.txt", Work in Progress, February 2011.

Authors' Addresses

Shankar Raman
Department of Computer Science and Engineering
I.I.T Madras,
Chennai - 600036
TamilNadu,
India.

EMail: mjsraman@cse.iitm.ac.in

Balaji Venkat Venkataswami
Department of Electrical Engineering,
I.I.T Madras,
Chennai - 600036,
TamilNadu,
India.

EMail: balajivenkat299@gmail.com

Prof.Gaurav Raina
Department of Electrical Engineering,
I.I.T Madras,
Chennai - 600036,
TamilNadu,
India.

EMail: gaurav@ee.iitm.ac.in

PANET Working Group
Internet-Draft
Intended Status: Experimental RFC
Expires: July 2013

Shankar Raman
Balaji Venkat Venkataswami
Gaurav Raina
Kamakoti Veezhinathan
IIT Madras
January 25, 2013

Reducing Power Consumption using BGP with power source data
draft-mjsraman-panet-inter-as-power-source-00

Abstract

In this paper, we propose a framework to reduce the aggregate power consumption of the Internet using a collaborative approach between Autonomous Systems (AS). We identify the low-power paths among the AS and then use Traffic Engineering (TE) techniques to route the packets along the paths. Such low-power paths can be identified by using the consumed-power-to-available-bandwidth (PWR) ratio as an additional constraint in the Constrained Shortest Path First (CSPF) algorithm. For re-routing the data traffic through these low-power paths, the Inter-AS Traffic Engineered Label Switched Path (TE-LSP) that spans multiple AS can be used. Extensions to the Border Gateway Protocol (BGP) can be used to disseminate the PWR ratio metric among the AS thereby creating a collaborative approach to reduce the power consumption. Since calculating the low-power paths can be computationally intensive, a graph-labeling heuristic is also proposed. This heuristic reduces the computational complexity but may provide a sub-optimal low-power path. The feasibility of our approaches is illustrated by applying our algorithm to a subset of the Internet. The techniques proposed in this paper for the Inter-AS power reduction require minimal modifications to the existing features of the Internet. The proposed techniques can be extended to other levels of Internet hierarchy, such as Intra-AS paths, through suitable modifications. The addition to this draft is that the power source of the Autonomous system is broken down to a ratio called PWR-SOURCE Ratio and used in the arrival of the metric to be used for this purpose.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering

Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	4
1.1	Low-power routers and switches	4
1.2	Power reduction using routing and traffic engineering	4
1.1	Terminology	5
2.	Methodology	5
2.1	Pre-requisites for the Proposed Method	6
2.1.1	Constructing network topology using BGP strands	6
2.1.2	PWR ratio calculation	7
2.1.2.1	Power Sources as additional factor	9
2.1.2.2	Earlier method of computing numerator of PWR ratio.	10
2.1.3	Explicit routing using TE-LSPs	11
2.2	LOW-POWER PATHS	11

2.2.0.1	Algorithm 1 ASBR low-power path algorithm	12
2.2.0.2	Algorithm 2 PCE low-power path algorithm	12
2.2.1	Illustration	13
2.2.3	Equivalence class with total ordering	13
2.2.3.1	Algorithm 3 PCE low-power path algorithm with graph labeling	14
2.3	Implementation notes and Discussion	15
2.4	Applicability within ASes within a single Admin Domain . . .	18
2.4.1	PWR_SESSION	18
2.5	Conclusion and Future Work	19
2.5	Acknowledgements	22
3	Security Considerations	23
4	IANA Considerations	23
5	References	23
5.1	Normative References	23
5.2	Informative References	23
	Authors' Addresses	24

1 Introduction

Estimates of power consumption for the Internet predict a 300% increase, as access speeds increase from 10 Mbps to 100 Mbps [3], [8]. Access speeds are likely to increase as new video, voice and gaming devices get added to the Internet. Various approaches have been proposed to reduce the power consumption of the Internet such as designing low-power routers and switches, and optimizing the network topology using traffic engineering methods [2].

1.1 Low-power routers and switches

Low-power router and switch design aim at reducing the power consumed by hardware architectural components such as transmission link, lookup tables and memory. In [4] it is shown that the router's link power consumption can vary by 20 Watts between idle and traffic scenarios. Hence the authors suggest having more line cards and running them to capacity: operating the router at full throughput will lead to less power per bit, and hence larger packet lengths will consume lower power. The two important components in routers that have received attention for high power consumption are buffers and TCAMs. Buffers are built using dynamic RAM (DRAM) or static RAM (SRAM). SRAMs are limited in size and consume more power, but have low access times. Guido [1] states that a 40Gb/s line card would require more than 300 SRAM chips and consume 2.5kW. DRAM access times prevent them from being used on high speed line cards. Sometimes the buffering of packets in DRAM is done at the back end, while SRAM is used at the front end for fast data access. But these schemes cannot scale with increasing line speeds. Some variants of TCAMs have been proposed for increasing line speeds and for reduced power consumption [7].

1.2 Power reduction using routing and traffic engineering

At the Internet level, creating a topology that allows route adaptation, capacity scaling and power-aware service rate tuning, will reduce power consumption. In [8] the author has proposed a technique to traffic engineer the data packets in such a way that the link capacity between routers is optimized. Links which are not utilized are moved to the idle state. Power consumption can be reduced by trading off performance related measures like latency. For example, power savings while switching from 1 Gbps to 100 Mbps is approximately 4 W and from 100 Mbps to 10 Mbps around 0.1 Watts. Hence instead of operating at 1 Gbps the link speed could be reduced to a lower bandwidth under certain conditions for reduced power consumption.

Multi layer traffic engineering based methods make use of parameters

such as resource usage, bandwidth, throughput and QoS measures, for power reduction. In [6] an approach for reducing Intra-AS power consumption for optical networks that uses Dijkstra's shortest path algorithm is proposed. The input to this method assumes the existence of a network topology using which an auxiliary graph is constructed. Power optimization is done on the auxiliary graph and traffic is routed through the low-power links. However, the algorithm expects the topology to be available for getting the auxiliary graph. This topology is easy to obtain for Intra-AS scenario, but not for Inter-AS cases. In our approach, we propose a collaborative approach by AS in power reduction. The core of the Internet at the Inter-AS level, uses the Multi-Protocol Label Switching (MPLS) technology. MPLS label switched paths that traverse multiple AS carry traffic from a head-end to a tail-end. The AS use the Border Gateway Protocol (BGP) for exchanging routing and topology related information. One of the attributes of BGP namely, AS-PATH-INFO is used to derive the topology of the Internet at the AS level. The CSPF algorithm is run on this AS level topology with the consumed-power-to-available-bandwidth (PWR) ratio as a constraint, to determine the low-power path from the head-end to the tail-end. The PWR ratio can be exchanged among the collaborating AS using BGP. Explicit routing can be achieved between the head-end and the tail-end through the low-power paths connecting the AS using the Inter-AS Traffic Engineered Label Switched Path (TE-LSP) that span multiple AS.

Calculation of such low-power paths can be computationally intensive and hence certain heuristics may be needed to reduce the computation time. A graph-labeling heuristic is proposed to reduce the computation time, which may lead to sub-optimal low-power paths. We illustrate our approaches by applying it to a subset of the Internet topology. The rest of the paper is organized as follows: In Section II, we discuss in detail the pre-requisites for the algorithm. Section III introduces the proposed technique which uses the CSPF algorithm to calculate the low-power paths. We also show that by using a graph-labeling technique, we can reduce the computational complexity of the low-power path algorithm, but may obtain a sub-optimal low-power path. In Section IV, we discuss the implementation issues. We present our conclusion and future work in Section V.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Methodology

<Document text>

2.1 Pre-requisites for the Proposed Method

In this section we discuss the pre-requisites for the implementation of the proposed scheme.

2.1.1 Constructing network topology using BGP strands

The Inter-AS topology can be modeled as a directed graph $G = (V; E; f)$ where the vertices (V) are mapped to AS and the edges (E) map the link that connect the neighboring AS. The direction (f) on the edge, represents the data flow from the head-end to the tail-end AS. To obtain the Inter-AS topology, the approach proposed in [5] is used. In this approach, it is shown that a sub-graph of the Internet topology, can be obtained by collecting several prefix updates in BGP. This is illustrated in Figure 1 which shows the different graph strands of AS that are recorded from the BGP packets. Each vertex in this graph is assigned a weight according to the consumed-power-to-available-bandwidth (PWR) ratio of the AS, as seen by an Autonomous System Border Router (ASBR) that acts as an entry point to the AS. Figure 2 shows the strands merged together to form the topology sub-graph. In this figure, the weight of the vertices are mapped to the ingress edges. A reference AS level topology derived from 100 strands of AS-PATH-INFO received by an AS in the Internet is presented in Figure 3 in [9].

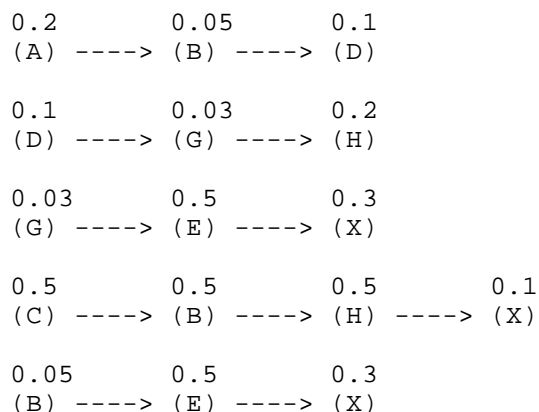


Figure 1: Different strands obtained from BGP updates, where vertices A,B,C,D and G represent the head-end AS. D,H and X form the tail-end AS. The vertex weights refer to the PWR ratio of the AS, and the direction of the link shows the next AS hop.

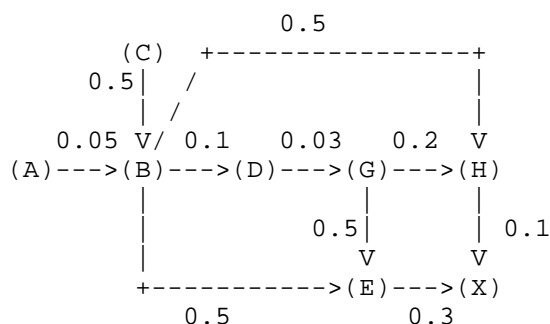


Figure 2: Combining the strands to get the topology of the Internet. The PWR ratio is mapped to the ingress link of the ASBR and not to the AS.

2.1.2 PWR ratio calculation

In the topology sub-graph, each AS is expected to share its PWR ratio. In order to calculate this ratio we need to calculate the consumed power in the AS and the maximum bandwidth available with an ASBR.

In this proposal each AS is expected to share its PWR ratio from as many ASBRs (Autonomous System Border Routers) that it has. Intuitively in order to calculate this ratio we need to calculate the consumed power representative of the AS and the maximum bandwidth available with an ASBR on its egress links into the AS. The entry point to the AS is through the ASBRs that advertise the prefixes reachable through the AS. Hence the numerator of the PWR ratio is calculated for the AS at each ingress ASBR. We first obtain the summation of power consumed at the Provider (P) and the Provider Edge (PE) routers within an AS. The numerator of the PWR ratio is calculated by summing up the consumed power of all the routers to be taken into account and then dividing this sum by the number of routers. A more intuitive approach would be to use a weighted average method by assigning routers to categories and having appropriate coefficients for each of these categories, thus arriving at a weighted average which is more accurate. One of these alternatives can be used to arrive at the numerator of the PWR ratio. Yet another alternative would have been to sum up the total consumed power of all routers in the AS and represent that as the numerator of the PWR ratio.

This average consumed power is divided by the maximum bandwidth available at each of the ASBR's egress link. This step is necessary

as the requested bandwidth for any path from the head-end to the tail-end using the ASBR is limited by the bandwidth available in the ASBR's egress links. The highest available bandwidth amongst the egress links of the ASBR is used as the denominator in the PWR ratio computation. If the entry point to the AS is through a different ASBR then the PWR ratio assigned to the ingress link of the ASBR might vary. Hence, an head-end AS might see different PWR ratios for an intermediate AS, if the intermediate AS has different ASBRs as its entry point.

The PWR ratio must be computed and disbursed much ahead of time before the Inter-AS TE-LSP explicit path or route is computed using the CSPF algorithm. The correctness of this ratio is of importance to compute the Inter-AS TE-LSP route through the green AS. If the entry point to the AS is through a different ASBR then the PWR ratio assigned to the ingress link of the ASBR might vary. Hence, an head-end AS might see different PWR ratios for an intermediate AS, if the intermediate AS has different ASBRs as its entry point.

We now illustrate the PWR ratio calculation. Consider an AS X which is one of the AS in the vicinity of another AS Y . Let this ASBR of X have 3 egress links into X denoted as E(1), E(2) and E(3), and 2 ingress links labeled I(1) and I(2). We now calculate the PWR ratio for I(1) and I(2). Assume that the routers in X have average consumed power of 200K Watts per hour. From figure 4 we can calculate the PWR ratio for I(1) and I(2) as $200K \text{ Watts} / (60 * 60 * 1.5 \text{ Gigb}) = 3.7037 * (10 \text{ raised to } -8)$ We could scale this to 0.37087 by multiplying with a base value of 10 raised to the 7th power.

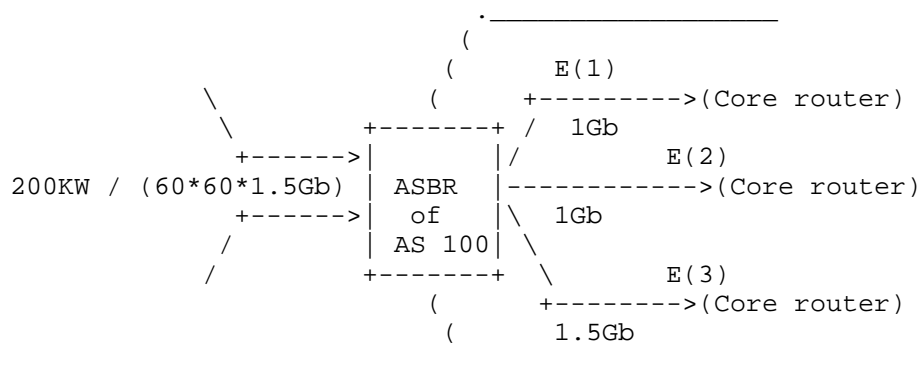


Figure 4: Calculation of PWR ratio by an ASBR associated with an AS. The I represents ingress links and E represents egress links. 200KW is the average consumed power in the AS. 1.5Gb is the maximum available bandwidth of the egress link in an ASBR.

Note that this ratio is actually a mapping function that is defined for each of the ingress links of the ASBR associated with an AS. For the head-end AS this mapping function does not exist as there is no ingress link. The PWR ratio can then be advertised to the other neighboring AS using the control plane through BGP extensions. BGP ensures that the information is percolated to other AS beyond the immediate neighbors. On receipt of these power metrics to the AS at the far-ends of the Internet, the overall AS level PWR ratio based Internet topology can be constructed. This view of the Internet is available with each of the routers without using any other complex discovery mechanism. Some sample link weights shown in Figure 2 is obtained by using such a mapping function on the ingress links.

2.1.2.1 Power Sources as additional factor

It is envisaged that the power sources of the Autonomous system using which the routers in the AS are powered should be declared as a metric which is further incorporated in the PWR ratio.

A suitable weight is provided to each type of source and the following table which is not claimed as totally exhaustive can be used to add this metric in the equation to compute the PWR ratio.

A formal classification of power sources and their weights is a topic to be considered later. For now we will deal with 2 main categories. Renewable sources of energy and non-renewable sources. There would be multiple categories under each of these major categories. Each such power source is assigned a weight.

Renewable Sources of Energy :

Wind - HighWeightOne
Solar - HighWeightTwo
Hydro - HighWeightThree
etc...

Non-renewable Sources of Energy :

Natural Gas - LowWeightOne
Petroleum and Diesel - LowWeightTwo
Nuclear - LowWeightThree
etc...

The PWR-SOURCE ratio is calculated in the proportion of how the above sources are combined to power the routers and its coolant systems and ancillary facilities in the AS.

Thus $PWR-RATIO = (Consumed-Power / Available-Bandwidth)$
 $* (1 / Weighted\ Average\ of\ Power\ Sources)$

This compound metric could be used as the PWR metric in the calculations specified in this draft.

2.1.2.2 Earlier method of computing numerator of PWR ratio.

Earlier in the previous versions of this document in order to calculate this PWR ratio we needed to calculate the available power and the maximum bandwidth available with an ASBR. The entry point to the AS is through ASBRs that advertise the prefixes reachable through the AS. Hence, the numerator of the PWR ratio is calculated for the AS at each ingress ASBR. We first obtained the summation of power consumed at the major Provider (P) and Provider Edge (PE) routers within an AS. The average available power is obtained by subtracting the consumed power from the maximum power rating and summing the values for all the routers and then dividing the result by the number of routers. As an alternative, one could use a weighted average for more accuracy depending on the category of the router advertising the consumed power. Yet another alternative is to take the average or sum of the maximum power rating of all the routers within an AS without taking into account the consumed power. One of these alternatives was chosen to calculate the numerator of the PWR ratio.

Intuition however drives us towards consumed power as a better numerator since the lesser the power consumed the lesser the numerator and hence lesser the ratio if enough bandwidth is available at the ingress ASBR. The amount of consumed power per bit of

information ought to be low for the shortest path to work out properly. One more aspect is that lesser the power consumed per available bit of bandwidth it could be a sign that routers are more optimal in their power consumption as they take on more traffic. This is a very crucial point to be considered.

2.1.3 Explicit routing using TE-LSPs

We assume that the head-end and the tail-end may reside in different AS and the path is along multiple intervening AS. The way to generate this path is by using Traffic Engineered Label Switched Paths (TE-LSPs). TE-LSPs can influence the exact path (at the AS level) that the traffic will pass through. This path can then be realized by providing these set of low-power consuming AS to a protocol like Resource Reservation Protocol (RSVP). RSVP-TE then creates TE-LSPs or tunnels, using its label assigning procedure. The routers use these low-power paths created by the explicit routing method rather than using the conventional shortest path to the destination. By this way, we can influence exclusion of a number of high power AS on the way from the head-end to the tail-end AS. For example, the dotted line in Figure 5 represents the explicit route that is chosen by making use of such TE-LSPs from head-end AS A to the tail-end AS X. Note that if number of hop was the metric used by CSPF, then the route chosen is the path with 3 hops.

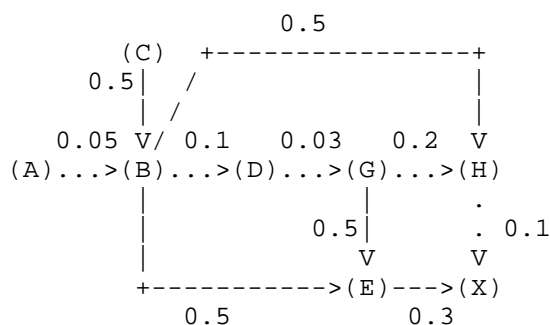


Figure 5: Low-power path is represented by the dotted lines. This low-power path has a longer number of hops than the conventional shortest path.

2.2 LOW-POWER PATHS

In this section we present the low-power path algorithm. The algorithm consists of two sub-algorithms: the first algorithm is

executed by all the ASBRs in the network and the second by all the Path Computation Elements (PCEs) in their respective AS. The algorithms for the ASBRs and PCEs are given as Algorithm 1, 2 and 3.

2.2.0.1 Algorithm 1 ASBR low-power path algorithm

Require: Weighted Topology Graph $T=(AS, E, f)$

```

1: Begin
2: if ROUTER == ASBR then
3: /* As part of IGP-TE */
4: Trigger exchange of available bandwidth on bandwidth change,
   to the AS internal neighbors;
5: BEGIN PARALLEL PROCESS 1
6: while PWR ratio changes do
7: Assign the PWR ratio to the Ingress links;
8: Exchange the PWR ratio with its external neighbors;
9: Exchange the PWR ratio with AS's (internal) ASBRs;
10: end while
11: END PARALLEL PROCESS 1
,br 12: BEGIN PARALLEL PROCESS 2
13: while RSVP packets arrive do
14: Send and Receive TE-LSP reservations in the explicit path;
15: Update routing table with labels for TE-LSP;
16: end while
17: END PARALLEL PROCESS 2
18: end if
19: End

```

2.2.0.2 Algorithm 2 PCE low-power path algorithm

Require: Weighted Topology Graph $T=(AS, E, f)$

Require: Source and Destination for Inter-AS TE LSP with sufficient bandwidth

```

1: Begin
2: if ROUTER == PCE then
3: Calculate the shortest paths from the head-end to the
   tail-end using CSPF with PWR ratio as the metric;
4: if no path available then
5: Signal error;
6: end if
7: if path exists then
8: Send explicit path to head-end to construct path;
9: end if
10: Continue passively listening to BGP updates to update
    $T=(AS, E, f)$ ;
11: end if
12: End

```

2.2.1 Illustration

We now illustrate the proposed technique with a simple example. Consider the AS level topology sub-graph shown in Figure 5 constructed using the strands shown in Figure 1. The PWR ratio calculated at an ASBR which represents the metric for the AS is assigned to the ingress link. For example, AS H has two edges coming into it: one from B and the other from G. Note that the power metrics for the two strands are different as G to H is lower than that of B to H. This means that the lower power metric into H is better if the path from G to H is chosen rather than the one from B to H. This is illustrated in the Figure 5 using dotted lines. To construct a path with A as the head-end AS and X as the tail-end AS, from the AS level topology we see that the path A, B, H, X and A, B, E, X have the shortest number of hops. However by using CSPF with the PWR ratio metric as the constraint, we see that the path A, B, D, G, H, X is power efficient. The routing choice will however be based on the reservation of the bandwidth on this path. Given that available bandwidth exists to setup a TE-LSP, the explicit path A, B, D, G, H, X is chosen. The Resource Reservation Protocol (RSVP) adheres to its usual operation and tries to setup a path. If bandwidth is not available in the low-power path thus calculated, then we may fall back to other paths like A, B, H, X or A, B, E, X provided there is available bandwidth in these paths. The low-power path algorithm given as Algorithm 2 is executed by the PCE. Algorithm 1 prepares the topology and feeds it as input to the PCE as a weighted topology graph. Using the CSPF algorithm to calculate a route from a source to destination could be time consuming for a large networks. But the topology is dynamically updated and hence the computation of the shortest paths can be triggered based on need. We now give a heuristic method based on graph-labeling that reduces the computation time but could trade-off the optimal low-power path.

2.2.3 Equivalence class with total ordering

The heuristic is based on avoiding high PWR ratios. The approach partitions the weighted links into equivalence classes based on a range of PWR values. For each partition a labeling is applied such that each link in the partition has the same label. A total ordering relationship is then defined on the equivalence class. The heuristic then starts including partitions with minimum label value iteratively until we get a connected component, which includes the head-end and tail-end AS. We apply the CSPF algorithm with the weights as label values on this sub-graph to obtain the low-power path. The modified algorithm which uses this scheme is given in Algorithm 3. It should be noted that this algorithm could provide sub-optimal power paths as the intermediate steps carry incomplete Internet topology information.

2.2.3.1 Algorithm 3 PCE low-power path algorithm with graph labeling

Require: Weighted Topology Graph $T=(AS, E, f)$

Require: Source and Destination for Inter-AS TE LSP with sufficient bandwidth

```
1: Begin
2: if ROUTER == PCE then
3: Group the links into N partitions with a label for
  each partition depending on the PWR ratio
4: Sort the labels in ascending order.
5: repeat
6: Include the links that have the least label value;
7: Remove the partition with this label;
8: until there is a path from the head-end to tail-end AS
9: Calculate the low-power path using labels from the
  head-end to the tail-end using CSPF ;
10: if no path available then
11: Signal error;
12: end if
13: if path exists then
14: Send explicit path to head-end to construct path;
15: end if
16: Continue passively listening to BGP updates to
  update  $T=(AS, E)$ ;
17: end if
18: End
```

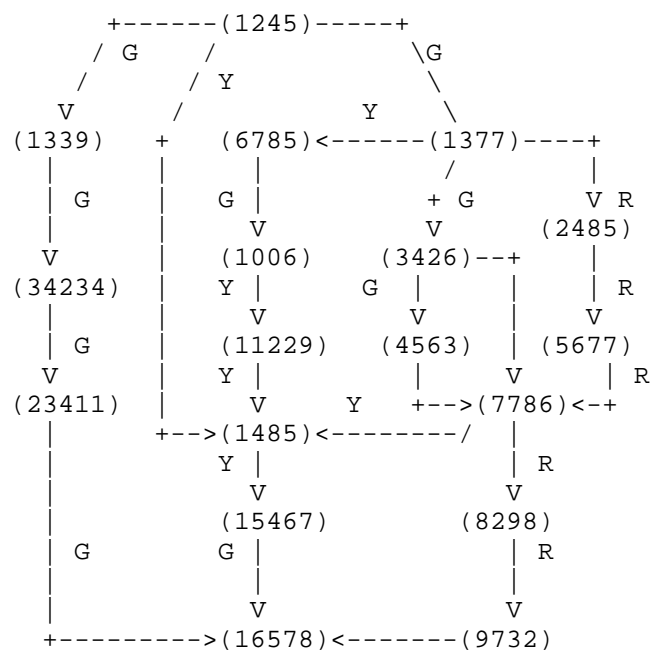


Figure 6: Application of the graph-labeling heuristic. We consider 3 labels "G" < "Y" < "R". Using algorithm 3 the "G" path from the head-end AS 1245 to the tail-end AS 16578 is chosen in the first iteration.

2.2.4 Illustration of graph labeling

We briefly illustrate the graph-labeling algorithm using Figure 6. In this diagram we have categorized the links into three partitions based on the PWR ratio. PWR ratio less than 0:1 are labeled as G, between 0:1 to 0:3 are labeled as Y and the rest as R. The total ordering is defined as $G < Y < R$, where the G links have low PWR ratios than the Y links. The path could be established through the AS that have G as the ingress link; the path being 1245, 1339, 34234, 23411 and 16578.

2.3 Implementation notes and Discussion

In this section we present some notes on feasibility of implementation of our scheme in a live network. First, the requested bandwidth should be available on the low-power path, but the CSPF algorithm is run with multiple constraints, one of which is the bandwidth requirement for the flows to be transported through the TE-LSP. The PWR ratio can then be applied to the available paths thus computing the low-power paths. Second, as we are using traffic

engineering with link state routing protocols, there is a reliable flooding process that are triggered when updates about the change in characteristic arise. We propose addition of some attributes with no change to the protocol implementation. There may be a time lag when the far ends of the Internet receive the attribute and the time it originated. This however cannot be avoided as with other attributes and metrics.

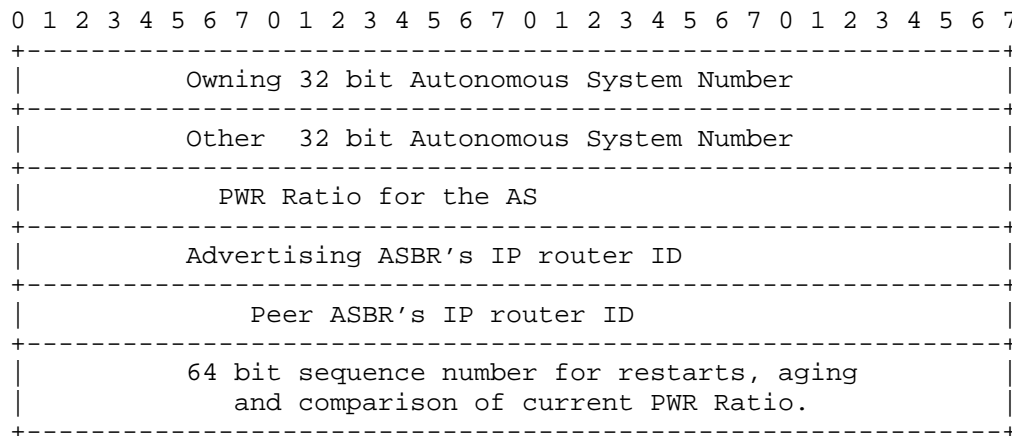


Figure 7: Proposed PDU format with an added attribute for AS-PATH-POWER-METRIC

The additions to the above Attribute have been added to optimize and correctly correlate the connecting ASes and the inter-AS links among them. For the traffic direction into the Advertising AS the above information will be easier to correlate than the previous version which did not advertise the peer AS which had the ingress links into the advertising Router.

In MPLS-TE when the TE metrics are modified, there is a reliable flooding process within an Interior Gateway Protocol (IGP). Such triggered updates apply to the PWR ratio as well. The proposed PWR ratio is advertised to the neighboring AS and the information percolated to all the AS, in a AS-PATH-POWER-METRIC attribute. This attribute can be implemented as shown in Figure 7. The frequency of the updates for this attribute should be fixed to avoid network flooding.

The AS-PATH-POWER-METRIC for each ASBR is calculated, and advertised as the PWR ratio for the AS. This AS-PATH-POWER-METRIC is filled into the appropriate optional transitive non-discretionary attribute and inserted into a unique vector for a set of prefixes advertised from the AS. Such advertised prefixes may have originated from the AS or

be the transit prefixes. The filled vector is sent to the ASBR of the neighboring AS and the information propagates to all the ASBRs. If the elements denoting AS in a vector of AS-PATH-INFO is not the same as the ones that need to be advertised in a AS-PATH-POWER-METRIC, then a suitable subset of AS-PATH-POWER-METRIC is identified and sent in the BGP updates. A vector of size 1 also can be employed if the AS in question is the only one for which PWR ratio has changed in the originating AS. The collation can be done depending on availability of such metrics and their mapping to a valid AS-PATH-INFO metric.

The power consumed by each router may fluctuate over short time intervals. In order to dampen these fluctuations which can cause unnecessary updates, power can be measured when falling within intervals of suitable size (say a range of values). This is as opposed to measuring power as a discrete quantity. This method of power measurement reduces the frequency of triggered updates from the routers due to power change.

```

0.1      0.2      0.1
(A) ----> (B) ----> (D)

0.1      0.2      0.02     0.2
(A) ----> (C) ----> (E) ----> (D)

0.1      0.2
(D) ----> (X)

```

Figure 8: Example of strands where more than one PWR ratio is advertised by "D"

```

      0.2      0.1      0.2
(A).....>(B).....>(D).....>(X)
|           ^
|0.2       0.02   | 0.2
+---->(C)----->(E)

```

Figure 9: Choice of low-power path derived using the algorithm which uses lower value of the ingress link but through the same AS

A use case of multiple ASBRs advertising differing PWR ratio shows that an AS may be seen as green through one ingress link and not through the other. Consider the case of multiple ASBRs that belong to the same AS, advertising PWR ratios that differ. This could lead to power values that belong to different classes of ratios with many intervening classes in between. These advertised PWR ratios could lead to one ASBR being preferred over the other thus taking a different path from head-end to tail-end. This also entails that

there may be multiple paths to the AS through these different ASBRs.

Consider Figure 8 which shows a set of strands that derive a topology as in Figure 9. Here D is reachable via two paths but the PWR ratios differ. This illustrates the case where the better metric wins out. The average power consumed would not have an effect but the bandwidth available on these ASBR egress links would definitely influence the path.

2.4 Applicability within ASes within a single Admin Domain

As per [draft-ietf-idr-aigp] there are deployments in which a single administration runs a network which has been sub-divided into multiple, contiguous ASes, each running BGP. There are several reasons why a single administrative domain may be broken into several ASes (which, in this case, are not really "autonomous".) It may be that the existing IGPs do not scale well in the particular environment; it may be that a more generalized topology is desired than could be obtained by use of a single IGP domain; it may be that a more finely grained routing policy is desired than can be supported by an IGP. In such deployments, it can be useful to allow BGP to make its routing decisions based on the IGP metric, so that BGP chooses the "shortest" path between two nodes, even if the nodes are in two different ASes within that same administrative domain. The authors refer to the set of ASes in a common administrative domain as an "AIGP Administrative Domain".

A combination of the AIGP administrative metric and the graph heuristic algorithm could be combined to arrive at a set of a suitable number k power-shortest paths and then use a tie-break amongst such k power-shortest-paths with the least AIGP metric. This is provided the set of ASes where the decision is being made all fall under a AIGP Administrative domain. This provides a trade-off of power shortest paths and least number of hops (link wise) to get from source to destination across these ASes.

2.4.1 PWR_SESSION

An implementation that supports the PWR attribute CAN support a per-session configuration item, PWR_SESSION, that indicates whether the PWR attribute is enabled or disabled for use on that session.

- The default value of PWR_SESSION, for EBGP sessions, between providers (distinct operators) CAN be "disabled".
- The default value of PWR_SESSION, for IBGP and confederation-EBGP sessions, MUST be "enabled."

The PWR attribute MUST NOT be sent on any BGP session for which PWR_SESSION is disabled.

If an PWR attribute is received on a BGP session for which PWR_SESSION is disabled, the attribute MUST be treated exactly as if it were an unrecognized transitive attribute. That is, " The handling of an unrecognized optional attribute is determined by the setting of the Transitive bit in the attribute flags octet. Paths with unrecognized transitive optional attributes SHOULD be accepted. If a path with an unrecognized transitive optional attribute is accepted and passed to other BGP peers, then the unrecognized transitive optional attribute of that path MUST be passed, along with the path, to other BGP peers with the Partial bit in the Attribute Flags octet set to 1. If a path with a recognized, transitive optional attribute is accepted and passed along to other BGP peers and the Partial bit in the Attribute Flags octet is set to 1 by some previous AS, it MUST NOT be set back to 0 by the current AS".

This helps in confining the distribution of the attribute and use in calculation of the power shortest paths only amongst ASes that have trust relationships with other ASes. Of course, this includes and promotes the use of PWR attribute within a AIGP administrative domain.

2.5 Conclusion and Future Work

In this paper, we proposed a scheme for reducing the power consumption of the Internet using collaborative effort between AS. The topology of the Internet is represented using a graph model and derived using the strands obtained from the AS-PATH attribute of the BGP updates. CSPF algorithm is run on this topology by using the PWR ratio as a constraint. The PWR ratio is advertised through the ingress links of the ASBRs associated with AS using BGP updates. The CSPF algorithm finds out the low-power consuming AS that can route data packets from a head-end to a tail-end. Explicit routing is handled through the use of TE-LSPs. This entails adopting routes by choosing entry points to an AS that give energy saving paths. Since using CSPF can be time consuming a heuristic algorithm to derive the low-power paths using graph-labeling was proposed. Our work complements the current schemes for reducing power consumption within a router such as switching off or bringing to power-idle-state certain select components within the forwarding and lookup mechanisms.

This Power shortest Path calculation can be taken care of a Path Computation Element (PCE) unit that could be either be a process running on a linecard on a ASBR, or even a core router or an offline

engine that is passively listening to the BGP updates within the AS without spitting out any routes of its own. The PCE architecture has already been proposed in the ietf and even has a separate working group for itself.

These offline or linecard engines are currently being sold in the market by the networking majors and other companies that develop hardware and software for the PCE. All the PCE needs to do is to accept configuration and passively listen to BGP updates from various peers or even be a client for a route reflector, thus

- a) Accepting these BGP updates
- b) Extracting the AS PATH information from these updates
- c) Then constructing the inter-AS topology
- d) Apply the PWR metric that comes along in these BGP updates to the edges of the graph
- e) Then compute the power shortest path as required by the configuration.

Normally the ASes have SLA agreements between each other to carry X amount of traffic from say a provider A. If the AS representing the ISP then advertises fake figures to carry more traffic than is mandated by the SLA agreement with other providers, then it is to that ISPs detriment since by advertising a better PWR ratio it invites more traffic through it thus getting paid less and carrying more traffic. This is not in the best interest of the ISP. This is so because in the final analysis the Power Shortest Path computed would include it regardless of the amount of traffic to be carried thus causing it to invite more traffic through it than it has accepted, even much more than its capacity. Hence it would be advisable for that ISP to advertise proper PWR ratios and NOT on the lower side of the spectrum. If it advertises HIGHER PWR ratios it would not be chosen, and hence that could be a policy measure NOT to accept any traffic at all since its capacity may be filled up with existing traffic. So advertising on the LOWER side would lead to lesser amount of benefit with respect to dollar per bit transported, and on the HIGHER side would be to exclude it from carrying any traffic that wanted to use the Power Shortest Path.

We also propose that there be a governing body in the IETF or outside it or sponsored by the IETF to verify the power ratios advertised are indeed valid or approximately closer to the actual consumption. A link up for each ISP with a power application level gateway to ensure

proper ratios are advertised could be mandated amongst at least the co-operating ISPs (ASes).

The points on which this proposal by us innovates is as follows.

a) There has been no effort prior to this to build an inter-AS topology with a weighted graph based on a PWR ratio. On this point it breaks a new path that would lead to inter-AS co-operation that contributes to power reduction overall in the internet. The paper suggested for OSPF by [10] deals with intra-Autonomous-system scenario rather than an inter-AS one. It is also to be noted that the IGP such as OSPF / IS-IS or any other link-state protocol for that matter is expected to capture the energy consumption of each router within the Autonomous system as in paper [10] to help get a hold on the overall average within the AS, or even sum up the total of all the power consumption within the AS with such intra-AS IGP LSA. This contributes to the PWR ratio proposed in our idea. Thus the intra-AS metric contributes to the PWR ratio. [10] proposal deals with primarily paths setup within an AS and not inter-AS paths. Thus the fundamental problem it solves is different while the problem we solve relates to the inter-AS paths which run across ASes from a head-end AS to a tail-end one.

b) The other aspect of innovation is to use BGP as the piggyback protocol upon which this scheme stands. There has been no effort earlier to approach the internet power reduction problem with BGP as the mode of transport of the energy ratios and coupling it with the inter-AS topology built with AS-PATH-INFO information.

The above 2 are key aspects of innovation.

When links and switches are gated or put into low-power state within an AS, the power-consumption automatically drops at the aggregate level, as a result of which the PWR ratio would be a lower figure advertised through BGP and thus this AS would attract more Power Shortest Path traffic through it. Thus the links within the AS and the switches within it would function more optimally if it had more traffic that went along paths that were originally put in low-power state thus utilizing the paths more effectively, when attracting PSP traffic.

There exist MIBs today that have object identifier for power consumed in a router. Maybe all the related components within it may NOT be listed with regards to power consumed. But the overall power consumed by the Router / Switch is gettable. Once it is advertised in a opaque Link-State-Advertisement say in the form of a TLV and the LSAs are flooded through the network in an AS, all routers get a uniform picture of which router consumes what power. This method already

exists for Traffic engineering Database LSAs that are advertised as LSAs for the purpose of traffic engineering within an AS. We are merely piggybacking on this capability to calculate the PWR ratio at the ASBR which amongst others is yet another Router / Switch of the AS.

Our future work includes looking into computing low-power paths within AS as well. Further it can be noted that the proposed algorithms might lead to increased latency as the number of hops increase, which could be critical for time sensitive applications. Since the PWR ratio could vary dynamically with traffic, the impact of traffic on the algorithm would also be of interest.

2.5 Acknowledgements

Shankar Raman would like to acknowledge the support by BT Public Limited (UK) under the BT IITM PhD Fellowship award. Balaji Venkat and Gaurav Raina would like to acknowledge the UK EPSRC Digital Economy Programme and the Government of India Department of Science and Technology (DST) for funding given to the IU-ATC. We would like to acknowledge that a version of this paper has been accepted in IARIA conference ENERGY 2012.

3 Security Considerations

No specific security considerations apart from the usual considerations with respect to authenticating BGP messages / updates from BGP neighbors is necessary for this scheme.

4 IANA Considerations

A new optional transitive non-discretionary attribute needs to be provided by IANA for carrying the PWR ratio across the Internet in the specified format in BGP.

5 References

5.1 Normative References

5.2 Informative References

REFERENCES

- [1] G. Appenzeller, Sizing router buffers, Doctoral Thesis, Department of Electrical Engineering, Stanford University, 2005.
- [2] A. P. Bianzino, C. Chaudet, D. Rossi and J. L. Rougier, A survey of green networking research, IEEE Communications and Surveys Tutorials, preprint.
- [3] J. Baliga, K. Hinton and R. S. Tucker, Energy consumption of the internet, Proc. of joint international conference on optical internet, June 2007, pp. 1-3.
- [4] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsang and S. Wright, Power awareness in network design and routing, Proc. of the IEEE INFOCOM 2008, April 2008, pp. 457-465.
- [5] B. Venkat et.al, Constructing disjoint and partially disjoint InterAS TE-LSPs, USPTO Patent 7751318, Cisco Systems, 2010.
- [6] M. Xia et. al., Greening the optical backbone network: A traffic engineering approach, IEEE ICC Proceedings, May 2010, pp. 1-5.

- [7] W. Lu and S. Sahni, Low-power TCAMS for very large forwarding tables, IEEE/ACM Transactions on Computer Networks, June 2010, vol. 18, no. 3, pp. 948-959.
- [8] B. Zhang, Routing Area Open Meeting, Proceedings of the IETF 81, Quebec, Canada, July 2011.
- [9] M.J.S Raman, V.Balaji Venkat, G.Raina, Reducing Power consumption using the Border Gateway Protocol, IARIA conferences ENERGY 2012.
- [10] A.Cianfrani et al., An OSPF enhancement for energy saving in IP Networks, IEEE INFOCOM 2011 Workshop on Green Communications and Networking
- [draft-ietf-idr-aigp] P. Mohapatra et.al, The Accumulated IGP metric attribute for BGP, <https://datatracker.ietf.org/doc/draft-ietf-idr-aigp/>, November 2012.

Authors' Addresses

Shankar Raman
Department of Computer Science and Engineering
IIT Madras
Chennai - 600036
TamilNadu
India.

EEmail: mjsraman@cse.iitm.ac.in

Balaji Venkat Venkataswami
Department of Electrical Engineering
IIT Madras
Chennai - 600036
TamilNadu
India.

EEmail: balajivenkat299@gmail.com

Prof.Gaurav Raina
Department of Electrical Engineering
IIT Madras
Chennai - 600036
TamilNadu
India.

EMail: gaurav@ee.iitm.ac.in

Prof.Kamakoti Veezhinathan
Department of Computer Science and Engineering
IIT Madras
Chennai - 600036
Tamilnadu
India

Email: kama@cse.iitm.ac.in

PANET Working Group
Internet-Draft
Intended Status: Experimental RFC
Expires: September 25, 2013

Shankar Raman
Balaji Venkat Venkataswami
Gaurav Raina
IIT Madras
March 24, 2013

Reducing Power Consumption using BGP
draft-mjsraman-panet-inter-as-psp-02

Abstract

In this paper, we propose a framework to reduce the aggregate power consumption of the Internet using a collaborative approach between Autonomous Systems (AS). We identify the low-power paths among the AS and then use Traffic Engineering (TE) techniques to route the packets along the paths. Such low-power paths can be identified by using the consumed-power-to-available-bandwidth (PWR) ratio as an additional constraint in the Constrained Shortest Path First (CSPF) algorithm. For re-routing the data traffic through these low-power paths, the Inter-AS Traffic Engineered Label Switched Path (TE-LSP) that spans multiple AS can be used. Extensions to the Border Gateway Protocol (BGP) can be used to disseminate the PWR ratio metric among the AS thereby creating a collaborative approach to reduce the power consumption. Since calculating the low-power paths can be computationally intensive, a graph-labeling heuristic is also proposed. This heuristic reduces the computational complexity but may provide a sub-optimal low-power path. The feasibility of our approaches is illustrated by applying our algorithm to a subset of the Internet. The techniques proposed in this paper for the Inter-AS power reduction require minimal modifications to the existing features of the Internet. The proposed techniques can be extended to other levels of Internet hierarchy, such as Intra-AS paths, through suitable modifications.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	4
1.1	Low-power routers and switches	4
1.2	Power reduction using routing and traffic engineering	4
1.1	Terminology	5
2.	Methodology	5
2.1	Pre-requisites for the Proposed Method	6
2.1.1	Constructing network topology using BGP strands	6
2.1.2	PWR ratio calculation	7
2.1.2.1	Earlier method of computing numerator of PWR ratio.	9
2.1.3	Explicit routing using TE-LSPs	10
2.2	LOW-POWER PATHS	11
2.2.0.1	Algorithm 1 ASBR low-power path algorithm	12
2.2.0.2	Algorithm 2 PCE low-power path algorithm	12
2.2.1	Illustration	12
2.2.3	Equivalence class with total ordering	13
2.2.3.1	Algorithm 3 PCE low-power path algorithm with	

graph labeling	14
2.3 Implementation notes and Discussion	15
2.4 Applicability within ASes within a single Admin Domain . . .	18
2.4.1 PWR_SESSION	18
2.4.2 Power profiles of Routers and Switches	19
2.4.2.1 Concave and Convex power curves	21
2.4.2.3 Need to advertise both available power and consumed power	22
2.5 Conclusion and Future Work	23
2.6 Acknowledgements	26
3 Security Considerations	27
4 IANA Considerations	27
5 References	27
5.1 Normative References	27
5.2 Informative References	27
Authors' Addresses	28

1 Introduction

Estimates of power consumption for the Internet predict a 300% increase, as access speeds increase from 10 Mbps to 100 Mbps [3], [8]. Access speeds are likely to increase as new video, voice and gaming devices get added to the Internet. Various approaches have been proposed to reduce the power consumption of the Internet such as designing low-power routers and switches, and optimizing the network topology using traffic engineering methods [2].

1.1 Low-power routers and switches

Low-power router and switch design aim at reducing the power consumed by hardware architectural components such as transmission link, lookup tables and memory. In [4] it is shown that the router's link power consumption can vary by 20 Watts between idle and traffic scenarios. Hence the authors suggest having more line cards and running them to capacity: operating the router at full throughput will lead to less power per bit, and hence larger packet lengths will consume lower power. The two important components in routers that have received attention for high power consumption are buffers and TCAMs. Buffers are built using dynamic RAM (DRAM) or static RAM (SRAM). SRAMs are limited in size and consume more power, but have low access times. Guido [1] states that a 40Gb/s line card would require more than 300 SRAM chips and consume 2.5kW. DRAM access times prevent them from being used on high speed line cards. Sometimes the buffering of packets in DRAM is done at the back end, while SRAM is used at the front end for fast data access. But these schemes cannot scale with increasing line speeds. Some variants of TCAMs have been proposed for increasing line speeds and for reduced power consumption [7].

1.2 Power reduction using routing and traffic engineering

At the Internet level, creating a topology that allows route adaptation, capacity scaling and power-aware service rate tuning, will reduce power consumption. In [8] the author has proposed a technique to traffic engineer the data packets in such a way that the link capacity between routers is optimized. Links which are not utilized are moved to the idle state. Power consumption can be reduced by trading off performance related measures like latency. For example, power savings while switching from 1 Gbps to 100 Mbps is approximately 4 W and from 100 Mbps to 10 Mbps around 0.1 Watts. Hence instead of operating at 1 Gbps the link speed could be reduced to a lower bandwidth under certain conditions for reduced power consumption.

Multi layer traffic engineering based methods make use of parameters

such as resource usage, bandwidth, throughput and QoS measures, for power reduction. In [6] an approach for reducing Intra-AS power consumption for optical networks that uses Dijkstra's shortest path algorithm is proposed. The input to this method assumes the existence of a network topology using which an auxiliary graph is constructed. Power optimization is done on the auxiliary graph and traffic is routed through the low-power links. However, the algorithm expects the topology to be available for getting the auxiliary graph. This topology is easy to obtain for Intra-AS scenario, but not for Inter-AS cases. In our approach, we propose a collaborative approach by AS in power reduction. The core of the Internet at the Inter-AS level, uses the Multi-Protocol Label Switching (MPLS) technology. MPLS label switched paths that traverse multiple AS carry traffic from a head-end to a tail-end. The AS use the Border Gateway Protocol (BGP) for exchanging routing and topology related information. One of the attributes of BGP namely, AS-PATH-INFO is used to derive the topology of the Internet at the AS level. The CSPF algorithm is run on this AS level topology with the consumed-power-to-available-bandwidth (PWR) ratio as a constraint, to determine the low-power path from the head-end to the tail-end. The PWR ratio can be exchanged among the collaborating AS using BGP. Explicit routing can be achieved between the head-end and the tail-end through the low-power paths connecting the AS using the Inter-AS Traffic Engineered Label Switched Path (TE-LSP) that span multiple AS.

Calculation of such low-power paths can be computationally intensive and hence certain heuristics may be needed to reduce the computation time. A graph-labeling heuristic is proposed to reduce the computation time, which may lead to sub-optimal low-power paths. We illustrate our approaches by applying it to a subset of the Internet topology. The rest of the paper is organized as follows: In Section II, we discuss in detail the pre-requisites for the algorithm. Section III introduces the proposed technique which uses the CSPF algorithm to calculate the low-power paths. We also show that by using a graph-labeling technique, we can reduce the computational complexity of the low-power path algorithm, but may obtain a sub-optimal low-power path. In Section IV, we discuss the implementation issues. We present our conclusion and future work in Section V.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Methodology

<Document text>

2.1 Pre-requisites for the Proposed Method

In this section we discuss the pre-requisites for the implementation of the proposed scheme.

2.1.1 Constructing network topology using BGP strands

The Inter-AS topology can be modeled as a directed graph $G = (V; E; f)$ where the vertices (V) are mapped to AS and the edges (E) map the link that connect the neighboring AS. The direction (f) on the edge, represents the data flow from the head-end to the tail-end AS. To obtain the Inter-AS topology, the approach proposed in [5] is used. In this approach, it is shown that a sub-graph of the Internet topology, can be obtained by collecting several prefix updates in BGP. This is illustrated in Figure 1 which shows the different graph strands of AS that are recorded from the BGP packets. Each vertex in this graph is assigned a weight according to the consumed-power-to-available-bandwidth (PWR) ratio of the AS, as seen by an Autonomous System Border Router (ASBR) that acts as an entry point to the AS. Figure 2 shows the strands merged together to form the topology sub-graph. In this figure, the weight of the vertices are mapped to the ingress edges. A reference AS level topology derived from 100 strands of AS-PATH-INFO received by an AS in the Internet is presented in Figure 3 in [9].

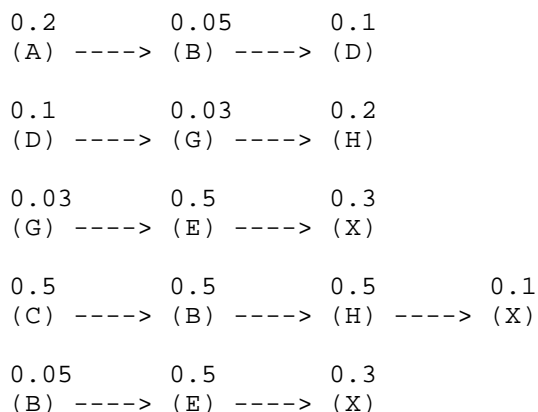


Figure 1: Different strands obtained from BGP updates, where vertices A,B,C,D and G represent the head-end AS. D,H and X form the tail-end AS. The vertex weights refer to the PWR ratio of the AS, and the direction of the link shows the next AS hop.

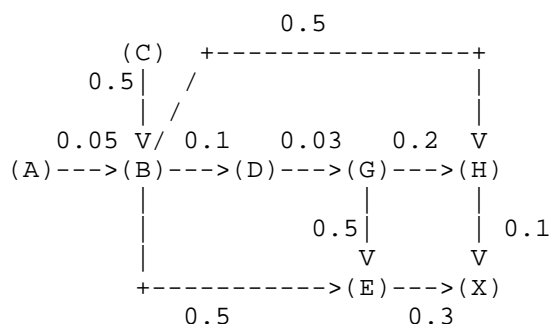


Figure 2: Combining the strands to get the topology of the Internet. The PWR ratio is mapped to the the ingress link of the ASBR and not to the AS.

2.1.2 PWR ratio calculation

In the topology sub-graph, each AS is expected to share its PWR ratio. In order to calculate this ratio we need to calculate the consumed power in the AS and the maximum bandwidth available with an ASBR.

In this proposal each AS is expected to share its PWR ratio from as many ASBRs (Autonomous System Border Routers) that it has. Intuitively in order to calculate this ratio we need to calculate the consumed power representative of the AS and the maximum bandwidth available with an ASBR on its egress links into the AS. The entry point to the AS is through the ASBRs that advertise the prefixes reachable through the AS. Hence the numerator of the PWR ratio is calculated for the AS at each ingress ASBR. We first obtain the summation of power consumed at the Provider (P) and the Provider Edge (PE) routers within an AS. The numerator of the PWR ratio is calculated by summing up the consumed power of all the routers to be taken into account and then dividing this sum by the number of routers. A more intuitive approach would be to use a weighted average method by assigning routers to categories and having appropriate co-efficients for each of these categories, thus arriving at a weighted average which is more accurate. One of these alternatives can be used to arrive at the numerator of the PWR ratio. Yet another alternative would have been to sum up the total consumed power of all routers in the AS and represent that as the numerator of the PWR ratio.

This average consumed power is divided by the maximum bandwidth available at each of the ASBR's egress link. This step is necessary

as the requested bandwidth for any path from the head-end to the tail-end using the ASBR is limited by the bandwidth available in the ASBR's egress links. The highest available bandwidth amongst the egress links of the ASBR is used as the denominator in the PWR ratio computation. If the entry point to the AS is through a different ASBR then the PWR ratio assigned to the ingress link of the ASBR might vary. Hence, an head-end AS might see different PWR ratios for an intermediate AS, if the intermediate AS has different ASBRs as its entry point.

The PWR ratio must be computed and disbursed much ahead of time before the Inter-AS TE-LSP explicit path or route is computed using the CSPF algorithm. The correctness of this ratio is of importance to compute the Inter-AS TE-LSP route through the green AS. If the entry point to the AS is through a different ASBR then the PWR ratio assigned to the ingress link of the ASBR might vary. Hence, an head-end AS might see different PWR ratios for an intermediate AS, if the intermediate AS has different ASBRs as its entry point.

We now illustrate the PWR ratio calculation. Consider an AS X which is one of the AS in the vicinity of another AS Y . Let this ASBR of X have 3 egress links into X denoted as E(1), E(2) and E(3), and 2 ingress links labeled I(1) and I(2). We now calculate the PWR ratio for I(1) and I(2). Assume that the routers in X have average consumed power of 200K Watts per hour. From figure 4 we can calculate the PWR ratio for I(1) and I(2) as $200K \text{ Watts} / (60 * 60 * 1.5 \text{ Gigb}) = 3.7037 * (10 \text{ raised to } -8)$ We could scale this to 0.37087 by multiplying with a base value of 10 raised to the 7th power.

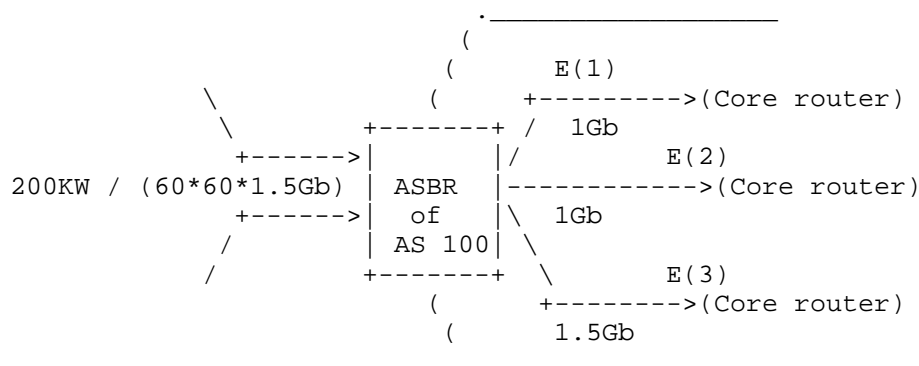


Figure 4: Calculation of PWR ratio by an ASBR associated with an AS. The I represents ingress links and E represents egress links. 200KW is the average consumed power in the AS. 1.5Gb is the maximum available bandwidth of the egress link in an ASBR.

Note that this ratio is actually a mapping function that is defined for each of the ingress links of the ASBR associated with an AS. For the head-end AS this mapping function does not exist as there is no ingress link. The PWR ratio can then be advertised to the other neighboring AS using the control plane through BGP extensions. BGP ensures that the information is percolated to other AS beyond the immediate neighbors. On receipt of these power metrics to the AS at the far-ends of the Internet, the overall AS level PWR ratio based Internet topology can be constructed. This view of the Internet is available with each of the routers without using any other complex discovery mechanism. Some sample link weights shown in Figure 2 is obtained by using such a mapping function on the ingress links.

2.1.2.1 Earlier method of computing numerator of PWR ratio.

Earlier in the previous versions of this document in order to calculate this PWR ratio we needed to calculate the available power and the maximum bandwidth available with an ASBR. The entry point to the AS is through ASBRs that advertise the prefixes reachable through the AS. Hence, the numerator of the PWR ratio is calculated for the AS at each ingress ASBR. We first obtained the summation of power consumed at the major Provider (P) and Provider Edge (PE) routers within an AS. The average available power is obtained by subtracting the consumed power from the maximum power rating and summing the values for all the routers and then dividing the result by the number of routers. As an alternative, one could use a weighted average for more accuracy depending on the category of the router advertising the consumed power. Yet another alternative is to take the average or sum of the maximum power rating of all the routers within an AS without

taking into account the consumed power. One of these alternatives was chosen to calculate the numerator of the PWR ratio.

Intuition however drives us towards consumed power as a better numerator since the lesser the power consumed the lesser the numerator and hence lesser the ratio if enough bandwidth is available at the ingress ASBR. The amount of consumed power per bit of information ought to be low for the shortest path to work out properly. One more aspect is that lesser the power consumed per available bit of bandwidth it could be a sign that routers are more optimal in their power consumption as they take on more traffic. This is a very crucial point to be considered.

However additional research seems to indicate that both Available and Consumed Power for a router be advertised. The need that arises for such a proposition is that there exist power profiles of routers which is dealt in later sections (section 2.4.2). Please refer section 2.4.2.1 onwards for more analysis and research on this subject.

2.1.3 Explicit routing using TE-LSPs

We assume that the head-end and the tail-end may reside in different AS and the path is along multiple intervening AS. The way to generate this path is by using Traffic Engineered Label Switched Paths (TE-LSPs). TE-LSPs can influence the exact path (at the AS level) that the traffic will pass through. This path can then be realized by providing these set of low-power consuming AS to a protocol like Resource Reservation Protocol (RSVP). RSVP-TE then creates TE-LSPs or tunnels, using its label assigning procedure. The routers use these low-power paths created by the explicit routing method rather than using the conventional shortest path to the destination. By this way, we can influence exclusion of a number of high power AS on the way from the head-end to the tail-end AS. For example, the dotted line in Figure 5 represents the explicit route that is chosen by making use of such TE-LSPs from head-end AS A to the tail-end AS X. Note that if number of hop was the metric used by CSPF, then the route chosen is the path with 3 hops.

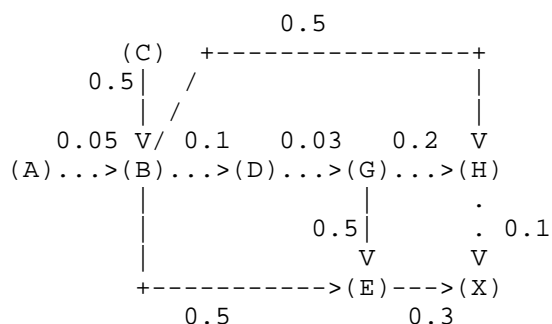


Figure 5: Low-power path is represented by the dotted lines. This low-power path has a longer number of hops than the conventional shortest path.

2.2 LOW-POWER PATHS

In this section we present the low-power path algorithm. The algorithm consists of two sub-algorithms: the first algorithm is executed by all the ASBRs in the network and the second by all the Path Computation Elements (PCEs) in their respective AS. The algorithms for the ASBRs and PCEs are given as Algorithm 1, 2 and 3.

2.2.0.1 Algorithm 1 ASBR low-power path algorithm

Require: Weighted Topology Graph $T=(AS, E, f)$

```
1: Begin
2: if ROUTER == ASBR then
3: /* As part of IGP-TE */
4: Trigger exchange of available bandwidth on bandwidth change,
   to the AS internal neighbors;
5: BEGIN PARALLEL PROCESS 1
6: while PWR ratio changes do
7: Assign the PWR ratio to the Ingress links;
8: Exchange the PWR ratio with its external neighbors;
9: Exchange the PWR ratio with AS's (internal) ASBRs;
10: end while
11: END PARALLEL PROCESS 1
,br 12: BEGIN PARALLEL PROCESS 2
13: while RSVP packets arrive do
14: Send and Receive TE-LSP reservations in the explicit path;
15: Update routing table with labels for TE-LSP;
16: end while
17: END PARALLEL PROCESS 2
18: end if
19: End
```

2.2.0.2 Algorithm 2 PCE low-power path algorithm

Require: Weighted Topology Graph $T=(AS, E, f)$

Require: Source and Destination for Inter-AS TE LSP with sufficient bandwidth

```
1: Begin
2: if ROUTER == PCE then
3: Calculate the shortest paths from the head-end to the
   tail-end using CSPF with PWR ratio as the metric;
4: if no path available then
5: Signal error;
6: end if
7: if path exists then
8: Send explicit path to head-end to construct path;
9: end if
10: Continue passively listening to BGP updates to update
    $T=(AS, E, f)$ ;
11: end if
12: End
```

2.2.1 Illustration

We now illustrate the proposed technique with a simple example.

Consider the AS level topology sub-graph shown in Figure 5 constructed using the strands shown in Figure 1. The PWR ratio calculated at an ASBR which represents the metric for the AS is assigned to the ingress link. For example, AS H has two edges coming into it: one from B and the other from G. Note that the power metrics for the two strands are different as G to H is lower than that of B to H. This means that the lower power metric into H is better if the path from G to H is chosen rather than the one from B to H. This is illustrated in the Figure 5 using dotted lines. To construct a path with A as the head-end AS and X as the tail-end AS, from the AS level topology we see that the path A, B, H, X and A, B, E, X have the shortest number of hops. However by using CSPF with the PWR ratio metric as the constraint, we see that the path A, B, D, G, H, X is power efficient. The routing choice will however be based on the reservation of the bandwidth on this path. Given that available bandwidth exists to setup a TE-LSP, the explicit path A, B, D, G, H, X is chosen. The Resource Reservation Protocol (RSVP) adheres to its usual operation and tries to setup a path. If bandwidth is not available in the low-power path thus calculated, then we may fall back to other paths like A, B, H, X or A, B, E, X provided there is available bandwidth in these paths. The low-power path algorithm given as Algorithm 2 is executed by the PCE. Algorithm 1 prepares the topology and feeds it as input to the PCE as a weighted topology graph. Using the CSPF algorithm to calculate a route from a source to destination could be time consuming for a large networks. But the topology is dynamically updated and hence the computation of the shortest paths can be triggered based on need. We now give a heuristic method based on graph-labeling that reduces the computation time but could trade-off the optimal low-power path.

2.2.3 Equivalence class with total ordering

The heuristic is based on avoiding high PWR ratios. The approach partitions the weighted links into equivalence classes based on a range of PWR values. For each partition a labeling is applied such that each link in the partition has the same label. A total ordering relationship is then defined on the equivalence class. The heuristic then starts including partitions with minimum label value iteratively until we get a connected component, which includes the head-end and tail-end AS. We apply the CSPF algorithm with the weights as label values on this sub-graph to obtain the low-power path. The modified algorithm which uses this scheme is given in Algorithm 3. It should be noted that this algorithm could provide sub-optimal power paths as the intermediate steps carry incomplete Internet topology information.

2.2.3.1 Algorithm 3 PCE low-power path algorithm with graph labeling

Require: Weighted Topology Graph $T=(AS, E, f)$

Require: Source and Destination for Inter-AS TE LSP with sufficient bandwidth

```
1: Begin
2: if ROUTER == PCE then
3: Group the links into N partitions with a label for
  each partition depending on the PWR ratio
4: Sort the labels in ascending order.
5: repeat
6: Include the links that have the least label value;
7: Remove the partition with this label;
8: until there is a path from the head-end to tail-end AS
9: Calculate the low-power path using labels from the
  head-end to the tail-end using CSPF ;
10: if no path available then
11: Signal error;
12: end if
13: if path exists then
14: Send explicit path to head-end to construct path;
15: end if
16: Continue passively listening to BGP updates to
  update  $T=(AS, E)$ ;
17: end if
18: End
```

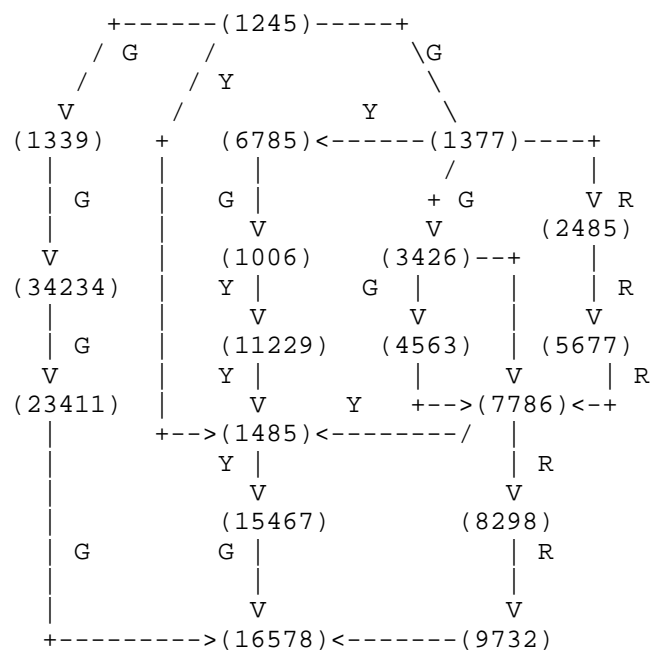


Figure 6: Application of the graph-labeling heuristic. We consider 3 labels "G" < "Y" < "R". Using algorithm 3 the "G" path from the head-end AS 1245 to the tail-end AS 16578 is chosen in the first iteration.

2.2.4 Illustration of graph labeling

We briefly illustrate the graph-labeling algorithm using Figure 6. In this diagram we have categorized the links into three partitions based on the PWR ratio. PWR ratio less than 0:1 are labeled as G, between 0:1 to 0:3 are labeled as Y and the rest as R. The total ordering is defined as G < Y < R, where the G links have low PWR ratios than the Y links. The path could be established through the AS that have G as the ingress link; the path being 1245, 1339, 34234, 23411 and 16578.

2.3 Implementation notes and Discussion

In this section we present some notes on feasibility of implementation of our scheme in a live network. First, the requested bandwidth should be available on the low-power path, but the CSPF algorithm is run with multiple constraints, one of which is the bandwidth requirement for the flows to be transported through the TE-LSP. The PWR ratio can then be applied to the available paths thus computing the low-power paths. Second, as we are using traffic

engineering with link state routing protocols, there is a reliable flooding process that are triggered when updates about the change in characteristic arise. We propose addition of some attributes with no change to the protocol implementation. There may be a time lag when the far ends of the Internet receive the attribute and the time it originated. This however cannot be avoided as with other attributes and metrics.

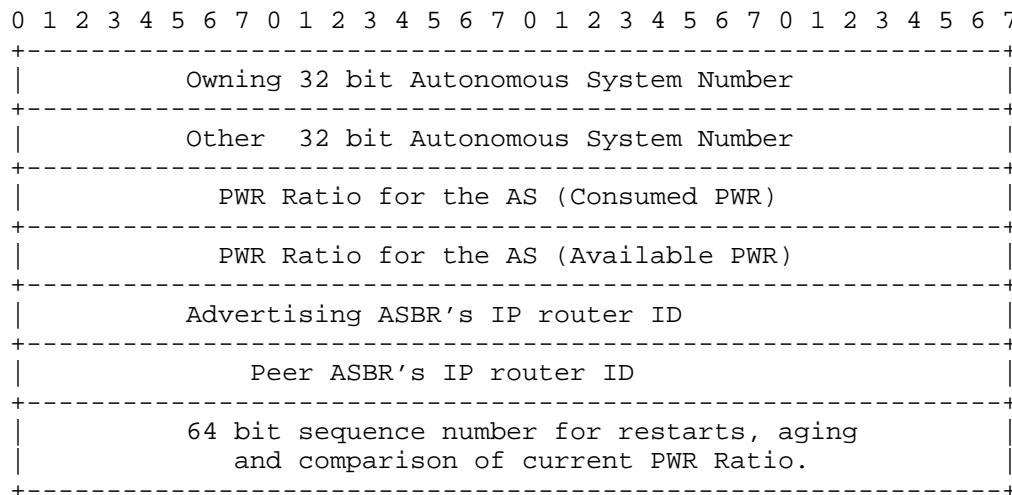


Figure 7: Proposed PDU format with an added attribute for AS-PATH-POWER-METRIC

The additions to the above Attribute have been added to optimize and correctly correlate the connecting ASes and the inter-AS links among them. For the traffic direction into the Advertising AS the above information will be easier to correlate than the previous version which did not advertise the peer AS which had the ingress links into the advertising Router.

In MPLS-TE when the TE metrics are modified, there is a reliable flooding process within an Interior Gateway Protocol (IGP). Such triggered updates apply to the PWR ratio as well. The proposed PWR ratio is advertised to the neighboring AS and the information percolated to all the AS, in a AS-PATH-POWER-METRIC attribute. This attribute can be implemented as shown in Figure 7. The frequency of the updates for this attribute should be fixed to avoid network flooding.

The AS-PATH-POWER-METRIC for each ASBR is calculated, and advertised as the PWR ratio for the AS. This AS-PATH-POWER-METRIC is filled into the appropriate optional transitive non-discretionary attribute and

inserted into a unique vector for a set of prefixes advertised from the AS. Such advertised prefixes may have originated from the AS or be the transit prefixes. The filled vector is sent to the ASBR of the neighboring AS and the information propagates to all the ASBRs. If the elements denoting AS in a vector of AS-PATH-INFO is not the same as the ones that need to be advertised in a AS-PATH-POWER-METRIC, then a suitable subset of AS-PATH-POWER-METRIC is identified and sent in the BGP updates. A vector of size 1 also can be employed if the AS in question is the only one for which PWR ratio has changed in the originating AS. The collation can be done depending on availability of such metrics and their mapping to a valid AS-PATH-INFO metric.

The power consumed by each router may fluctuate over short time intervals. In order to dampen these fluctuations which can cause unnecessary updates, power can be measured when falling within intervals of suitable size (say a range of values). This is as opposed to measuring power as a discrete quantity. This method of power measurement reduces the frequency of triggered updates from the routers due to power change.

```

0.1      0.2      0.1
(A) ----> (B) ----> (D)

0.1      0.2      0.02      0.2
(A) ----> (C) ----> (E) ----> (D)

0.1      0.2
(D) ----> (X)

```

Figure 8: Example of strands where more than one PWR ratio is advertised by "D"

```

      0.2      0.1      0.2
(A).....>(B).....>(D).....>(X)
      |          ^
      |          |
      |0.2      0.02      | 0.2
      +---->(C)----->(E)

```

Figure 9: Choice of low-power path derived using the algorithm which uses lower value of the ingress link but through the same AS

A use case of multiple ASBRs advertising differing PWR ratio shows that an AS may be seen as green through one ingress link and not through the other. Consider the case of multiple ASBRs that belong to the same AS, advertising PWR ratios that differ. This could lead to power values that belong to different classes of ratios with many intervening classes in between. These advertised PWR ratios could

lead to one ASBR being preferred over the other thus taking a different path from head-end to tail-end. This also entails that there may be multiple paths to the AS through these different ASBRs.

Consider Figure 8 which shows a set of strands that derive a topology as in Figure 9. Here D is reachable via two paths but the PWR ratios differ. This illustrates the case where the better metric wins out. The average power consumed would not have an effect but the bandwidth available on these ASBR egress links would definitely influence the path.

2.4 Applicability within ASes within a single Admin Domain

As per [draft-ietf-idr-aigp] there are deployments in which a single administration runs a network which has been sub-divided into multiple, contiguous ASes, each running BGP. There are several reasons why a single administrative domain may be broken into several ASes (which, in this case, are not really "autonomous".) It may be that the existing IGPs do not scale well in the particular environment; it may be that a more generalized topology is desired than could be obtained by use of a single IGP domain; it may be that a more finely grained routing policy is desired than can be supported by an IGP. In such deployments, it can be useful to allow BGP to make its routing decisions based on the IGP metric, so that BGP chooses the "shortest" path between two nodes, even if the nodes are in two different ASes within that same administrative domain. The authors refer to the set of ASes in a common administrative domain as an "AIGP Administrative Domain".

A combination of the AIGP administrative metric and the graph heuristic algorithm could be combined to arrive at a set of a suitable number k power-shortest paths and then use a tie-break amongst such k power-shortest-paths with the least AIGP metric. This is provided the set of ASes where the decision is being made all fall under a AIGP Administrative domain. This provides a trade-off of power shortest paths and least number of hops (link wise) to get from source to destination across these ASes.

2.4.1 PWR_SESSION

An implementation that supports the PWR attribute CAN support a per-session configuration item, PWR_SESSION, that indicates whether the PWR attribute is enabled or disabled for use on that session.

- The default value of PWR_SESSION, for EBGp sessions, between providers (distinct operators) CAN be "disabled".
- The default value of PWR_SESSION, for IBGP and confederation-

EBGP sessions, MUST be "enabled."

The PWR attribute MUST NOT be sent on any BGP session for which PWR_SESSION is disabled.

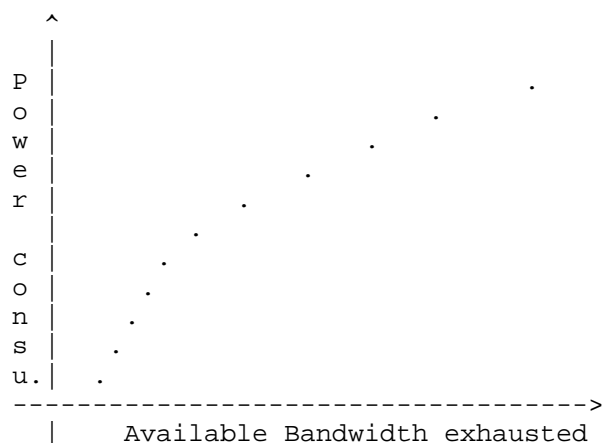
If an PWR attribute is received on a BGP session for which PWR_SESSION is disabled, the attribute MUST be treated exactly as if it were an unrecognized transitive attribute. That is, "The handling of an unrecognized optional attribute is determined by the setting of the Transitive bit in the attribute flags octet. Paths with unrecognized transitive optional attributes SHOULD be accepted. If a path with an unrecognized transitive optional attribute is accepted and passed to other BGP peers, then the unrecognized transitive optional attribute of that path MUST be passed, along with the path, to other BGP peers with the Partial bit in the Attribute Flags octet set to 1. If a path with a recognized, transitive optional attribute is accepted and passed along to other BGP peers and the Partial bit in the Attribute Flags octet is set to 1 by some previous AS, it MUST NOT be set back to 0 by the current AS".

This helps in confining the distribution of the attribute and use in calculation of the power shortest paths only amongst ASes that have trust relationships with other ASes. Of course, this includes and promotes the use of PWR attribute within a AIGP administrative domain.

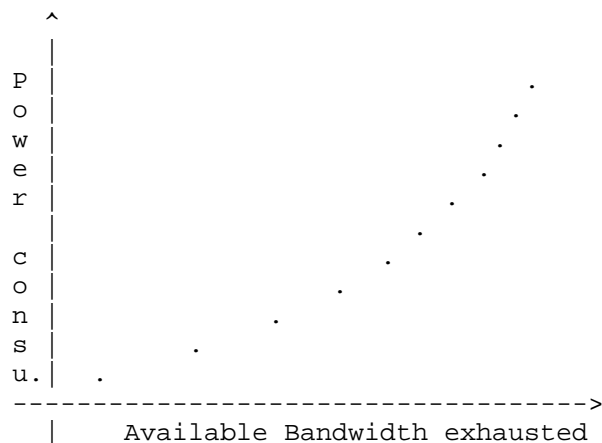
2.4.2 Power profiles of Routers and Switches

It has been experimented and from several sources found that there exist routers which have different power profiles. The power profile of a router is the curve of power consumption to available bandwidth. Mentioned below are a few of these prominent ones that have to be taken into consideration.

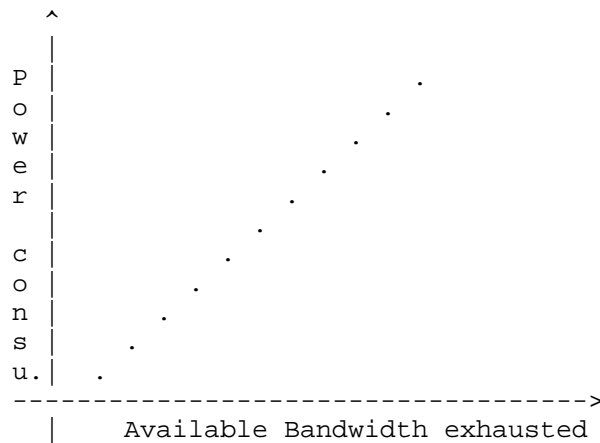
The first profile that we will consider is the flattening curve. The power consumed to available bandwidth curve takes the shape of a steep one initially and then tapers off to a plateau. The point at which it begins to give a delta-C (delta in Power Consumed) to delta-B (Available Bandwidth exhausted) is the inflection point that tapers off to a plateau. Here the delta-C/delta-B begins to slow down or decrease rapidly. The more the traffic that is added onto the device the lesser it draws power.



The second profile that we will consider is the exponential curve. The power consumed to available bandwidth curve takes the shape of an ever increasing steep curve as shown below. Here the $\Delta C / \Delta B$ begins to increase as more traffic is thrown onto it as the Available bandwidth exhausted increases. This power curve beyond a point is intolerable with respect to power guzzling.



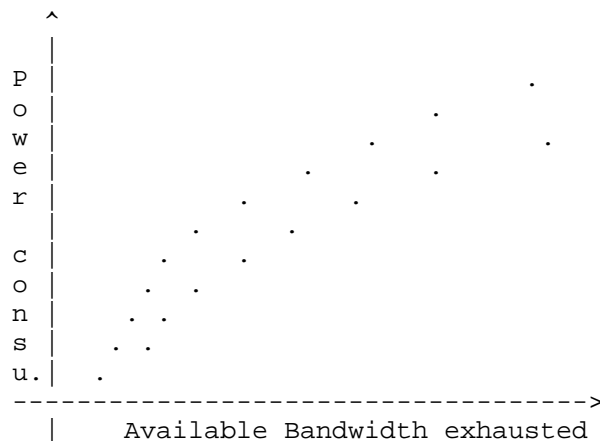
The third profile that we will consider is a linear curve. In other words just a straight line. Here $\Delta C / \Delta B$ is a constant.



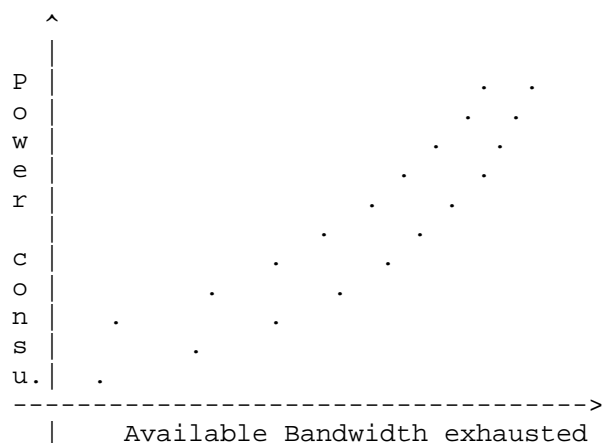
2.4.2.1 Concave and Convex power curves

Given that there are 3 kinds of major profiles in the router power consumption, what line would we like to pick. This is an important point when choosing the metric to pick the low power paths.

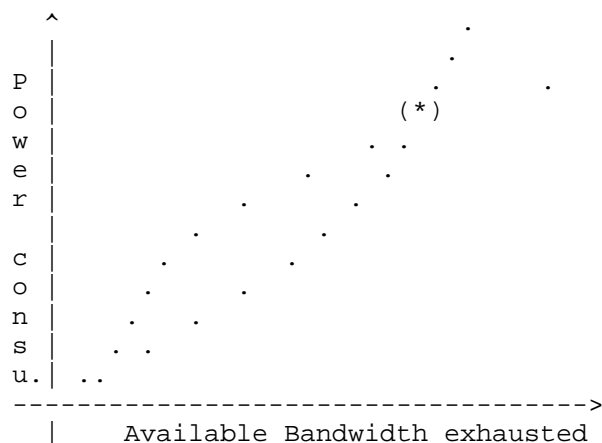
(a) If the confrontation is between 2 first profile routers the lower of the 2 would be considered as shown below. The lower curve offers better power savings for each GB of bandwidth transported.



(b) If the confrontation is between 2 second profile routers the upper curve offers more power savings per GB of bandwidth.



(c) When the confrontation is between a first profile curve and a second profile curve, it would be optimal to pick (as shown below) the lower of the curves because it gives us lesser power consumed for every GB of traffic routed / switched. Here the exponential curve is the one that offers lesser amount of power consumed per GB of traffic is chosen. But when it gets to a point that the two curves intersect it would be more optimal to pick the tapering curve. Thus at the meeting point of the 2 curves the exponential curve becomes more costly and the tapering one gives us more GB for the power buck. Thus this switchover from one curve to the other (in other words from the exponential curve to the tapering one) does the trick in terms of finding an optimal solution.



(*) Metric switchover point from Consumed Power to Available Power.

2.4.2.3 Need to advertise both available power and consumed power

Thus the above sections have shown that both the available power and the consumed power MUST be advertised so that case (c) can be deciphered and the switchover of the curves be done and the appropriate router be chosen for the rest of the bandwidth to be switched over to.

Thus there will exist Consumed-Power to Available Bandwidth ratio and the Available Power to Available Bandwidth ratio. Both the ratios are computed and the lower value chosen. The Available Power can be judged from the calibration process such as the one carried out by independent test organizations as in [12]. An example of their calibration is referred to in [12].

2.5 Conclusion and Future Work

In this paper, we proposed a scheme for reducing the power consumption of the Internet using collaborative effort between AS. The topology of the Internet is represented using a graph model and derived using the strands obtained from the AS-PATH attribute of the BGP updates. CSPF algorithm is run on this topology by using the PWR ratio as a constraint. The PWR ratio is advertised through the ingress links of the ASBRs associated with AS using BGP updates. The CSPF algorithm finds out the low-power consuming AS that can route data packets from a head-end to a tail-end. Explicit routing is handled through the use of TE-LSPs. This entails adopting routes by choosing entry points to an AS that give energy saving paths. Since using CSPF can be time consuming a heuristic algorithm to derive the low-power paths using graph-labeling was proposed. Our work complements the current schemes for reducing power consumption within a router such as switching off or bringing to power-idle-state certain select components within the forwarding and lookup mechanisms.

This Power shortest Path calculation can be taken care of a Path Computation Element (PCE) unit that could be either be a process running on a linecard on a ASBR, or even a core router or an offline engine that is passively listening to the BGP updates within the AS without spitting out any routes of its own. The PCE architecture has already been proposed in the ietf and even has a separate working group for itself.

These offline or linecard engines are currently being sold in the market by the networking majors and other companies that develop hardware and software for the PCE. All the PCE needs to do is to accept configuration and passively listen to BGP updates from various peers or even be a client for a route reflector, thus

- a) Accepting these BGP updates
- b) Extracting the AS PATH information from these updates
- c) Then constructing the inter-AS topology
- d) Apply the PWR metric that comes along in these BGP updates to the edges of the graph
- e) Then compute the power shortest path as required by the configuration.

Normally the ASes have SLA agreements between each other to carry X amount of traffic from say a provider A. If the AS representing the ISP then advertises fake figures to carry more traffic than is mandated by the SLA agreement with other providers, then it is to that ISPs detriment since by advertising a better PWR ratio it invites more traffic through it thus getting paid less and carrying more traffic. This is not in the best interest of the ISP. This is so because in the final analysis the Power Shortest Path computed would include it regardless of the amount of traffic to be carried thus causing it to invite more traffic through it than it has accepted, even much more than its capacity. Hence it would be advisable for that ISP to advertise proper PWR ratios and NOT on the lower side of the spectrum. If it advertises HIGHER PWR ratios it would not be chosen, and hence that could be a policy measure NOT to accept any traffic at all since its capacity may be filled up with existing traffic. So advertising on the LOWER side would lead to lesser amount of benefit with respect to dollar per bit transported, and on the HIGHER side would be to exclude it from carrying any traffic that wanted to use the Power Shortest Path.

We also propose that there be a governing body in the IETF or outside it or sponsored by the IETF to verify the power ratios advertised are indeed valid or approximately closer to the actual consumption. A link up for each ISP with a power application level gateway to ensure proper ratios are advertised could be mandated amongst at least the co-operating ISPs (ASes).

The points on which this proposal by us innovates is as follows.

- a) There has been no effort prior to this to build an inter-AS topology with a weighted graph based on a PWR ratio. On this point it breaks a new path that would lead to inter-AS co-operation that contributes to power reduction overall in the internet. The paper suggested for OSPF by [10] deals with intra-Autonomous-system scenario rather than an inter-AS one. It is also to be noted that the

IGP such as OSPF / IS-IS or any other link-state protocol for that matter is expected to capture the energy consumption of each router within the Autonomous system as in paper [10] to help get a hold on the overall average within the AS, or even sum up the total of all the power consumption within the AS with such intra-AS IGP LSA. This contributes to the PWR ratio proposed in our idea. Thus the intra-AS metric contributes to the PWR ratio. [10] proposal deals with primarily paths setup within an AS and not inter-AS paths. Thus the fundamental problem it solves is different while the problem we solve relates to the inter-AS paths which run across ASes from a head-end AS to a tail-end one.

b) The other aspect of innovation is to use BGP as the piggyback protocol upon which this scheme stands. There has been no effort earlier to approach the internet power reduction problem with BGP as the mode of transport of the energy ratios and coupling it with the inter-AS topology built with AS-PATH-INFO information.

The above 2 are key aspects of innovation.

When links and switches are gated or put into low-power state within an AS, the power-consumption automatically drops at the aggregate level, as a result of which the PWR ratio would be a lower figure advertised through BGP and thus this AS would attract more Power Shortest Path traffic through it. Thus the links within the AS and the switches within it would function more optimally if it had more traffic that went along paths that were originally put in low-power state thus utilizing the paths more effectively, when attracting PSP traffic.

There exist MIBs today that have object identifier for power consumed in a router. Maybe all the related components within it may NOT be listed with regards to power consumed. But the overall power consumed by the Router / Switch is gettable. Once it is advertised in a opaque Link-State-Advertisement say in the form of a TLV and the LSAs are flooded through the network in an AS, all routers get a uniform picture of which router consumes what power. This method already exists for Traffic engineering Database LSAs that are advertised as LSAs for the purpose of traffic engineering within an AS. We are merely piggybacking on this capability to calculate the PWR ratio at the ASBR which amongst others is yet another Router / Switch of the AS.

Our future work includes looking into computing low-power paths within AS as well. Further it can be noted that the proposed algorithms might lead to increased latency as the number of hops increase, which could be critical for time sensitive applications. Since the PWR ratio could vary dynamically with traffic, the impact

of traffic on the algorithm would also be of interest.

2.6 Acknowledgements

Shankar Raman would like to acknowledge the support by BT Public Limited (UK) under the BT IITM PhD Fellowship award. Balaji Venkat and Gaurav Raina would like to acknowledge the UK EPSRC Digital Economy Programme and the Government of India Department of Science and Technology (DST) for funding given to the IU-ATC. We would like to acknowledge that a version of this paper has been accepted in IARIA conference ENERGY 2012.

3 Security Considerations

No specific security considerations apart from the usual considerations with respect to authenticating BGP messages / updates from BGP neighbors is necessary for this scheme.

4 IANA Considerations

A new optional transitive non-discretionary attribute needs to be provided by IANA for carrying the PWR ratio across the Internet in the specified format in BGP.

5 References

5.1 Normative References

5.2 Informative References

REFERENCES

- [1] G. Appenzeller, Sizing router buffers, Doctoral Thesis, Department of Electrical Engineering, Stanford University, 2005.
- [2] A. P. Bianzino, C. Chaudet, D. Rossi and J. L. Rougier, A survey of green networking research, IEEE Communications and Surveys Tutorials, preprint.
- [3] J. Baliga, K. Hinton and R. S. Tucker, Energy consumption of the internet, Proc. of joint international conference on optical internet, June 2007, pp. 1-3.
- [4] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsang and S. Wright, Power awareness in network design and routing, Proc. of the IEEE INFOCOM 2008, April 2008, pp. 457-465.
- [5] B. Venkat et.al, Constructing disjoint and partially disjoint InterAS TE-LSPs, USPTO Patent 7751318, Cisco Systems, 2010.
- [6] M. Xia et. al., Greening the optical backbone network: A traffic engineering approach, IEEE ICC Proceedings, May 2010, pp. 1-5.

- [7] W. Lu and S. Sahni, Low-power TCAMs for very large forwarding tables, IEEE/ACM Transactions on Computer Networks, June 2010, vol. 18, no. 3, pp. 948-959.
- [8] B. Zhang, Routing Area Open Meeting, Proceedings of the IETF 81, Quebec, Canada, July 2011.
- [9] M.J.S Raman, V.Balaji Venkat, G.Raina, Reducing Power consumption using the Border Gateway Protocol, IARIA conferences ENERGY 2012.
- [10] A.Cianfrani et al., An OSPF enhancement for energy saving in IP Networks, IEEE INFOCOM 2011 Workshop on Green Communications and Networking
- [draft-ietf-idr-aigp] P. Mohapatra et.al, The Accumulated IGP metric attribute for BGP, <https://datatracker.ietf.org/doc/draft-ietf-idr-aigp/>, November 2012.

Authors' Addresses

Shankar Raman
Department of Computer Science and Engineering
IIT Madras,
Chennai - 600036
TamilNadu,
India.

EEmail: mjsraman@cse.iitm.ac.in

Balaji Venkat Venkataswami
Department of Electrical Engineering,
IIT Madras,
Chennai - 600036,
TamilNadu,
India.

EEmail: balajivenkat299@gmail.com

Prof.Gaurav Raina
Department of Electrical Engineering,
IIT Madras,
Chennai - 600036,
TamilNadu,
India.

EMail: gaurav@ee.iitm.ac.in

PANET Working Group
Internet-Draft
Intended Status: Experimental RFC
Expires: May 9, 2013

Shankar Raman
Balaji Venkat Venkataswami
Gaurav Raina
Vasan Srini
I.I.T, Madras
November 5, 2012

Constructing inter-AS power shortest protection TE-LSPs using BGP
draft-mjsraman-panet-inter-as-psp-protect-00

Abstract

In this paper, we propose a framework to build protection / backup paths for power shortest primary inter-AS TE-LSPs. The primary path is built within a framework to reduce the aggregate power consumption of the Internet using a collaborative approach between Autonomous Systems (AS). We identify the low-power paths among the AS and then use Traffic Engineering (TE) techniques to route the packets along the paths. Such low-power paths can be identified by using the consumed-power-to-available-bandwidth (PWR) ratio as an additional constraint in the Constrained Shortest Path First (CSPF) algorithm. For re-routing the data traffic through these low-power paths, the Inter-AS Traffic Engineered Label Switched Path (TE-LSP) that spans multiple AS can be used.

Once the primary paths have been built we use the same techniques to build backup power shortest paths in a similar manner except by excluding the nodes (ASes) and links (between these ASes) that are present in the primary path. This way the backup path does not traverse any of the ASes or links between these ASes of the primary path so constructed.

Extensions to the Border Gateway Protocol (BGP) can be used to disseminate the PWR ratio metric among the AS thereby creating a collaborative approach to reduce the power consumption. Since calculating the low-power paths can be computationally intensive, a graph-labeling heuristic is also proposed. This heuristic reduces the computational complexity but may provide a sub-optimal low-power path. The feasibility of our approaches is illustrated by applying our algorithm to a subset of the Internet. The techniques proposed in this paper for the Inter-AS power reduction require minimal modifications to the existing features of the Internet. The proposed techniques can be extended to other levels of Internet hierarchy, such as Intra-AS paths, through suitable modifications.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	4
1.1	Low-power routers and switches	4
1.2	Power reduction using routing and traffic engineering	4
1.1	Terminology	5
2.	Methodology	6
2.1	Pre-requisites for the Proposed Method	6
2.1.1	Constructing network topology using BGP strands	6

2.1.2	PWR ratio calculation	7
2.1.2.1	Earlier method of computing numerator of PWR ratio.	9
2.1.3	Explicit routing using TE-LSPs	10
2.2	LOW-POWER PATHS	11
2.2.0.1	Algorithm 1 ASBR low-power path algorithm	12
2.2.0.2	Algorithm 2 PCE low-power path algorithm	12
2.2.1	Illustration	12
2.2.1.1	Backup path construction	13
2.2.3	Equivalence class with total ordering	14
2.2.3.1	Algorithm 3 PCE low-power path algorithm with graph labeling	15
2.2.4	Illustration of graph labeling	16
2.2.5	Moving traffic when the primary path fails	17
2.3	Implementation notes and Discussion	17
2.4	Conclusion and Future Work	20
2.4.1	Link and node disjoint Backup paths and power constraints	20
2.5	Acknowledgements	23
3	Security Considerations	24
4	IANA Considerations	24
5	References	24
5.1	Normative References	24
5.2	Informative References	24
	Authors' Addresses	25

1 Introduction

Estimates of power consumption for the Internet predict a 300% increase, as access speeds increase from 10 Mbps to 100 Mbps [3], [8]. Access speeds are likely to increase as new video, voice and gaming devices get added to the Internet. Various approaches have been proposed to reduce the power consumption of the Internet such as designing low-power routers and switches, and optimizing the network topology using traffic engineering methods [2].

1.1 Low-power routers and switches

Low-power router and switch design aim at reducing the power consumed by hardware architectural components such as transmission link, lookup tables and memory. In [4] it is shown that the router's link power consumption can vary by 20 Watts between idle and traffic scenarios. Hence the authors suggest having more line cards and running them to capacity: operating the router at full throughput will lead to less power per bit, and hence larger packet lengths will consume lower power. The two important components in routers that have received attention for high power consumption are buffers and TCAMs. Buffers are built using dynamic RAM (DRAM) or static RAM (SRAM). SRAMs are limited in size and consume more power, but have low access times. Guido [1] states that a 40Gb/s line card would require more than 300 SRAM chips and consume 2.5kW. DRAM access times prevent them from being used on high speed line cards. Sometimes the buffering of packets in DRAM is done at the back end, while SRAM is used at the front end for fast data access. But these schemes cannot scale with increasing line speeds. Some variants of TCAMs have been proposed for increasing line speeds and for reduced power consumption [7].

1.2 Power reduction using routing and traffic engineering

At the Internet level, creating a topology that allows route adaptation, capacity scaling and power-aware service rate tuning, will reduce power consumption. In [8] the author has proposed a technique to traffic engineer the data packets in such a way that the link capacity between routers is optimized. Links which are not utilized are moved to the idle state. Power consumption can be reduced by trading off performance related measures like latency. For example, power savings while switching from 1 Gbps to 100 Mbps is approximately 4 W and from 100 Mbps to 10 Mbps around 0.1 Watts. Hence instead of operating at 1 Gbps the link speed could be reduced to a lower bandwidth under certain conditions for reduced power consumption.

Multi layer traffic engineering based methods make use of parameters

such as resource usage, bandwidth, throughput and QoS measures, for power reduction. In [6] an approach for reducing Intra-AS power consumption for optical networks that uses Dijkstra's shortest path algorithm is proposed. The input to this method assumes the existence of a network topology using which an auxiliary graph is constructed. Power optimization is done on the auxiliary graph and traffic is routed through the low-power links. However, the algorithm expects the topology to be available for getting the auxiliary graph. This topology is easy to obtain for Intra-AS scenario, but not for Inter-AS cases. In our approach, we propose a collaborative approach by AS in power reduction and also methods to construct backup power paths for the primary paths so constructed. The core of the Internet at the Inter-AS level, uses the Multi-Protocol Label Switching (MPLS) technology. MPLS label switched paths that traverse multiple AS carry traffic from a head-end to a tail-end. The AS use the Border Gateway Protocol (BGP) for exchanging routing and topology related information. One of the attributes of BGP namely, AS-PATH-INFO is used to derive the topology of the Internet at the AS level. The CSPF algorithm is run on this AS level topology with the consumed-power-to-available-bandwidth (PWR) ratio as a constraint, to determine the low-power path from the head-end to the tail-end. The PWR ratio can be exchanged among the collaborating AS using BGP. Explicit routing can be achieved between the head-end and the tail-end through the low-power paths connecting the AS using the Inter-AS Traffic Engineered Label Switched Path (TE-LSP) that span multiple AS.

Calculation of such low-power paths can be computationally intensive and hence certain heuristics may be needed to reduce the computation time. A graph-labeling heuristic is proposed to reduce the computation time, which may lead to sub-optimal low-power paths. We illustrate our approaches by applying it to a subset of the Internet topology. We then indicate how backup paths can be constructed by avoiding links and nodes in the primary path using both the brute force method and graph-labeling method.

The rest of the paper is organized as follows: In Section 2, we discuss in detail the pre-requisites for the algorithm. Section 2.2 introduces the proposed technique which uses the CSPF algorithm to calculate the low-power primary and backup paths. We also show that by using a graph-labeling technique, we can reduce the computational complexity of the low-power path algorithm, but may obtain a sub-optimal low-power path. In Section 2.3, we discuss the implementation issues. We present our conclusion and future work in Section 2.4.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this

document are to be interpreted as described in RFC 2119 [RFC2119].

2. Methodology

2.1 Pre-requisites for the Proposed Method

In this section we discuss the pre-requisites for the implementation of the proposed scheme.

2.1.1 Constructing network topology using BGP strands

The Inter-AS topology can be modeled as a directed graph $G = (V; E; f)$ where the vertices (V) are mapped to AS and the edges (E) map the link that connect the neighboring AS. The direction (f) on the edge, represents the data flow from the head-end to the tail-end AS. To obtain the Inter-AS topology, the approach proposed in [5] is used. In this approach, it is shown that a sub-graph of the Internet topology, can be obtained by collecting several prefix updates in BGP. This is illustrated in Figure 1 which shows the different graph strands of AS that are recorded from the BGP packets. Each vertex in this graph is assigned a weight according to the consumed-power-to-available-bandwidth (PWR) ratio of the AS, as seen by an Autonomous System Border Router (ASBR) that acts as an entry point to the AS. Figure 2 shows the strands merged together to form the topology sub-graph. In this figure, the weight of the vertices are mapped to the ingress edges. A reference AS level topology derived from 100 strands of AS-PATH-INFO received by an AS in the Internet is presented in Figure 3 in [9].

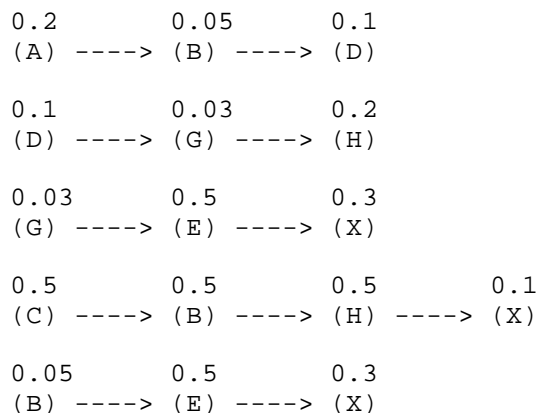


Figure 1: Different strands obtained from BGP updates, where vertices

A,B,C,D and G represent the head-end AS. D,H and X form the tail-end AS. The vertex weights refer to the PWR ratio of the AS, and the direction of the link shows the next AS hop.

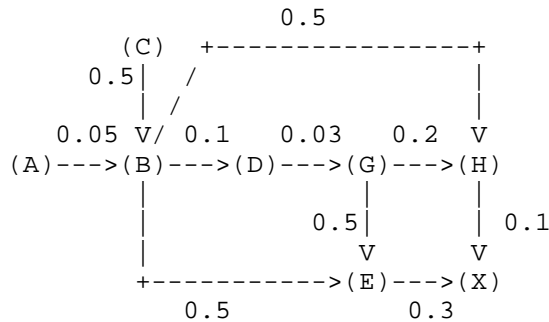


Figure 2:Combining the strands to get the topology of the Internet. The PWR ratio is mapped to the the ingress link of the ASBR and not to the AS.

2.1.2 PWR ratio calculation

In the topology sub-graph, each AS is expected to share its PWR ratio. In order to calculate this ratio we need to calculate the consumed power in the AS and the maximum bandwidth available with an ASBR.

In this proposal each AS is expected to share its PWR ratio from as many ASBRs (Autonomous System Border Routers) that it has. Intuitively in order to calculate this ratio we need to calculate the consumed power representative of the AS and the maximum bandwidth available with an ASBR on its egress links into the AS. The entry point to the AS is through the ASBRs that advertise the prefixes reachable through the AS. Hence the numerator of the PWR ratio is calculated for the AS at each ingress ASBR. We first obtain the summation of power consumed at the Provider (P) and the Provider Edge (PE) routers within an AS. The numerator of the PWR ratio is calculated by summing up the consumed power of all the routers to be taken into account and then dividing this sum by the number of routers. A more intuitive approach would be to use a weighted average method by assigning routers to categories and having appropriate co-efficients for each of these categories, thus arriving at a weighted average which is more accurate. One of these alternatives can be used to arrive at the numerator of the PWR ratio. Yet another alternative would have been to sum up the total consumed power of all routers in

the AS and represent that as the numerator of the PWR ratio.

This average consumed power is divided by the maximum bandwidth available at each of the ASBR's egress link. This step is necessary as the requested bandwidth for any path from the head-end to the tail-end using the ASBR is limited by the bandwidth available in the ASBR's egress links. The highest available bandwidth amongst the egress links of the ASBR is used as the denominator in the PWR ratio computation. If the entry point to the AS is through a different ASBR then the PWR ratio assigned to the ingress link of the ASBR might vary. Hence, an head-end AS might see different PWR ratios for an intermediate AS, if the intermediate AS has different ASBRs as its entry point.

The PWR ratio must be computed and disbursed much ahead of time before the Inter-AS TE-LSP explicit path or route is computed using the CSPF algorithm. The correctness of this ratio is of importance to compute the Inter-AS TE-LSP route through the green AS. If the entry point to the AS is through a different ASBR then the PWR ratio assigned to the ingress link of the ASBR might vary. Hence, an head-end AS might see different PWR ratios for an intermediate AS, if the intermediate AS has different ASBRs as its entry point.

We now illustrate the PWR ratio calculation. Consider an AS X which is one of the AS in the vicinity of another AS Y . Let this ASBR of X have 3 egress links into X denoted as E(1), E(2) and E(3), and 2 ingress links labeled I(1) and I(2). We now calculate the PWR ratio for I(1) and I(2). Assume that the routers in X have average consumed power of 200K Watts per hour. From figure 4 we can calculate the PWR ratio for I(1) and I(2) as $200K \text{ Watts} / (60 * 60 * 1.5 \text{ Gigb} = 3.7037 * (10 \text{ raised to } -8)$ We could scale this to 0.37087 by multiplying with a base value of 10 raised to the 7th power.

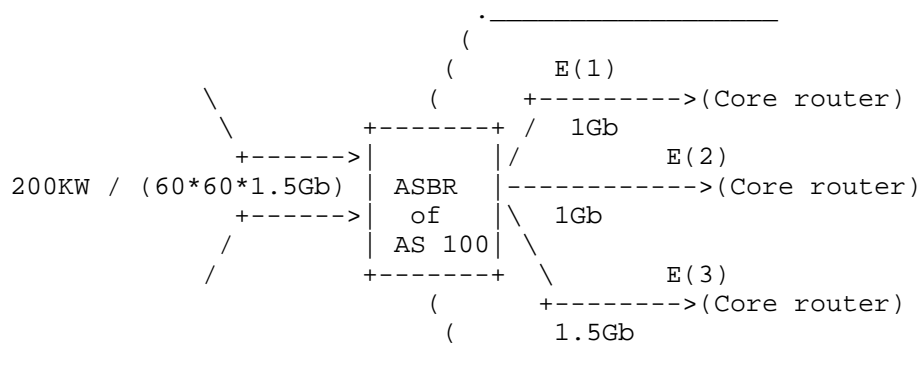


Figure 4: Calculation of PWR ratio by an ASBR associated with an AS. The I represents ingress links and E represents egress links. 200KW is the average consumed power in the AS. 1.5Gb is the maximum available bandwidth of the egress link in an ASBR.

Note that this ratio is actually a mapping function that is defined for each of the ingress links of the ASBR associated with an AS. For the head-end AS this mapping function does not exist as there is no ingress link. The PWR ratio can then be advertised to the other neighboring AS using the control plane through BGP extensions. BGP ensures that the information is percolated to other AS beyond the immediate neighbors. On receipt of these power metrics to the AS at the far-ends of the Internet, the overall AS level PWR ratio based Internet topology can be constructed. This view of the Internet is available with each of the routers without using any other complex discovery mechanism. Some sample link weights shown in Figure 2 is obtained by using such a mapping function on the ingress links.

2.1.2.1 Earlier method of computing numerator of PWR ratio.

Earlier in the previous versions of this document in order to calculate this PWR ratio we needed to calculate the available power and the maximum bandwidth available with an ASBR. The entry point to the AS is through ASBRs that advertise the prefixes reachable through the AS. Hence, the numerator of the PWR ratio is calculated for the AS at each ingress ASBR. We first obtained the summation of power consumed at the major Provider (P) and Provider Edge (PE) routers within an AS. The average available power is obtained by subtracting the consumed power from the maximum power rating and summing the values for all the routers and then dividing the result by the number of routers. As an alternative, one could use a weighted average for more accuracy depending on the category of the router advertising the consumed power. Yet another alternative is to take the average or sum of the maximum power rating of all the routers within an AS without

taking into account the consumed power. One of these alternatives was chosen to calculate the numerator of the PWR ratio.

Intuition however drives us towards consumed power as a better numerator since the lesser the power consumed the lesser the numerator and hence lesser the ratio if enough bandwidth is available at the ingress ASBR. The amount of consumed power per bit of information ought to be low for the shortest path to work out properly. One more aspect is that lesser the power consumed per available bit of bandwidth it could be a sign that routers are more optimal in their power consumption as they take on more traffic. This is a very crucial point to be considered.

2.1.3 Explicit routing using TE-LSPs

We assume that the head-end and the tail-end may reside in different AS and the path is along multiple intervening AS. The way to generate this path is by using Traffic Engineered Label Switched Paths (TE-LSPs). TE-LSPs can influence the exact path (at the AS level) that the traffic will pass through. This path can then be realized by providing these set of low-power consuming AS to a protocol like Resource Reservation Protocol (RSVP). RSVP-TE then creates TE-LSPs or tunnels, using its label assigning procedure. The routers use these low-power paths created by the explicit routing method rather than using the conventional shortest path to the destination. By this way, we can influence exclusion of a number of high power AS on the way from the head-end to the tail-end AS. For example, the dotted line in Figure 5 represents the explicit route that is chosen by making use of such TE-LSPs from head-end AS A to the tail-end AS X. Note that if number of hop was the metric used by CSPF, then the route chosen is the path with 3 hops.

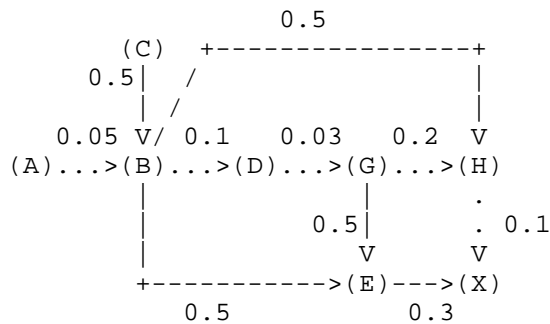


Figure 5: Low-power path is represented by the dotted lines. This low-power path has a longer number of hops than the conventional shortest path.

2.2 LOW-POWER PATHS

In this section we present the low-power path algorithm. The algorithm consists of two sub-algorithms: the first algorithm is executed by all the ASBRs in the network and the second by all the Path Computation Elements (PCEs) in their respective AS. The algorithms for the ASBRs and PCEs are given as Algorithm 1, 2 and 3.

2.2.0.1 Algorithm 1 ASBR low-power path algorithm

Require: Weighted Topology Graph $T=(AS, E, f)$

```

1: Begin
2: if ROUTER == ASBR then
3: /* As part of IGP-TE */
4: Trigger exchange of available bandwidth on bandwidth change,
   to the AS internal neighbors;
5: BEGIN PARALLEL PROCESS 1
6: while PWR ratio changes do
7: Assign the PWR ratio to the Ingress links;
8: Exchange the PWR ratio with its external neighbors;
9: Exchange the PWR ratio with AS's (internal) ASBRs;
10: end while
11: END PARALLEL PROCESS 1
,br 12: BEGIN PARALLEL PROCESS 2
13: while RSVP packets arrive do
14: Send and Receive TE-LSP reservations in the explicit path;
15: Update routing table with labels for TE-LSP;
16: end while
17: END PARALLEL PROCESS 2
18: end if
19: End

```

2.2.0.2 Algorithm 2 PCE low-power path algorithm

Require: Weighted Topology Graph $T=(AS, E, f)$

Require: Source and Destination for Inter-AS TE LSP with sufficient bandwidth

```

1: Begin
2: if ROUTER == PCE then
3: Calculate the shortest paths from the head-end to the
   tail-end using CSPF with PWR ratio as the metric;
4: if no path available then
5: Signal error;
6: end if
7: if path exists then
8: Send explicit path to head-end to construct path;
9: end if
10: Continue passively listening to BGP updates to update
    $T=(AS, E, f)$ ;
11: end if
12: End

```

2.2.1 Illustration

We now illustrate the proposed technique with a simple example.

Consider the AS level topology sub-graph shown in Figure 5 constructed using the strands shown in Figure 1. The PWR ratio calculated at an ASBR which represents the metric for the AS is assigned to the ingress link. For example, AS H has two edges coming into it: one from B and the other from G. Note that the power metrics for the two strands are different as G to H is lower than that of B to H. This means that the lower power metric into H is better if the path from G to H is chosen rather than the one from B to H. This is illustrated in the Figure 5 using dotted lines. To construct a path with A as the head-end AS and X as the tail-end AS, from the AS level topology we see that the path A, B, H, X and A, B, E, X have the shortest number of hops. However by using CSPF with the PWR ratio metric as the constraint, we see that the path A, B, D, G, H, X is power efficient. The routing choice will however be based on the reservation of the bandwidth on this path. Given that available bandwidth exists to setup a TE-LSP, the explicit path A, B, D, G, H, X is chosen. The Resource Reservation Protocol (RSVP) adheres to its usual operation and tries to setup a path. If bandwidth is not available in the low-power path thus calculated, then we may fall back to other paths like A, B, H, X or A, B, E, X provided there is available bandwidth in these paths. The low-power path algorithm given as Algorithm 2 is executed by the PCE. Algorithm 1 prepares the topology and feeds it as input to the PCE as a weighted topology graph. Using the CSPF algorithm to calculate a route from a source to destination could be time consuming for a large networks. But the topology is dynamically updated and hence the computation of the shortest paths can be triggered based on need. We later give a heuristic method based on graph-labeling that reduces the computation time but could trade-off the optimal low-power path.

2.2.1.1 Backup path construction

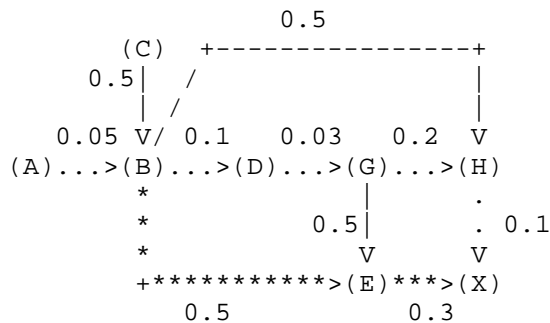
Assume there is a requirement to construct one or more backup paths that excludes all nodes and links in the primary path as constructed above. That is none of the nodes and links in the primary path constructed should be present in the backup path to be constructed.

This is only possible to the extent possible since there may not be available paths that are completely disjoint node and link wise from the primary path previously constructed. In that case only the minimum of the originally used links and nodes in the primary path should be included in the backup path. That is if it is not possible to construct a completely node and link disjoint path for the backup path, then only the necessary nodes and links from the primary path should be included in the backup path.

In the below figure the path from A to X was constructed with the primary path going through A,B,D,G,H and X. For the backup path there

exist no other path power shortest wise other than A,B,E,X. If we were to think of the next Power shortest path through A,B,H,X it includes A->B and H->X which are members of the primary path as well. Though the A,B,H and X path have shorter power than the A,B,E,X path the number of links included from the primary path for the former are 2 while the latter has only one. This would advice us to choose the latter since it has only one link and nodes (A->B) that coincide with the primary path. Here the * dotted line shows the backup path computed.

It is possible to request for more than one backup path to be constructed and section 2.2.3 would provide a better time complexity algorithm to do the same. Even for the first backup path the equivalence class heuristic would be a better choice.



2.2.3 Equivalence class with total ordering

The heuristic is based on avoiding high PWR ratios. The approach partitions the weighted links into equivalence classes based on a range of PWR values. For each partition a labeling is applied such that each link in the partition has the same label. A total ordering relationship is then defined on the equivalence class. The heuristic then starts including partitions with minimum label value iteratively until we get a connected component, which includes the head-end and tail-end AS. We apply the CSPF algorithm with the weights as label values on this sub-graph to obtain the low-power path. The modified algorithm which uses this scheme is given in Algorithm 3. It should be noted that this algorithm could provide sub-optimal power paths as the intermediate steps carry incomplete Internet topology information.

2.2.3.1 Algorithm 3 PCE low-power path algorithm with graph labeling

Require: Weighted Topology Graph $T=(AS, E, f)$

Require: Source and Destination for Inter-AS TE LSP with sufficient bandwidth

```

1: Begin
2: if ROUTER == PCE then
3: Group the links into N partitions with a label for
  each partition depending on the PWR ratio
4: Sort the labels in ascending order.
5: repeat
6: Include the links that have the least label value;
7: Remove the partition with this label;
8: until there is a path from the head-end to tail-end AS
9: Calculate the low-power path using labels from the
  head-end to the tail-end using CSPF ;
10: if no path available then
11: Signal error;
12: end if
13: if path exists then
14: Send explicit path to head-end to construct path;
15: end if
15.1 if backup path needs to be constructed then
15.2 Repeat 3 to 15 after excluding the links and nodes in
15.3 the primary path constructed, and by including next
15.3.1 available label grade if necessary.
15.4 end if
15.5 if no path exists then add the least power links and
15.6 nodes in the primary path back into the graph and repeat
15.7 the process.
15.8 end if
16: Continue passively listening to BGP updates to
  update  $T=(AS, E)$ ;
17: end if
18: End

```

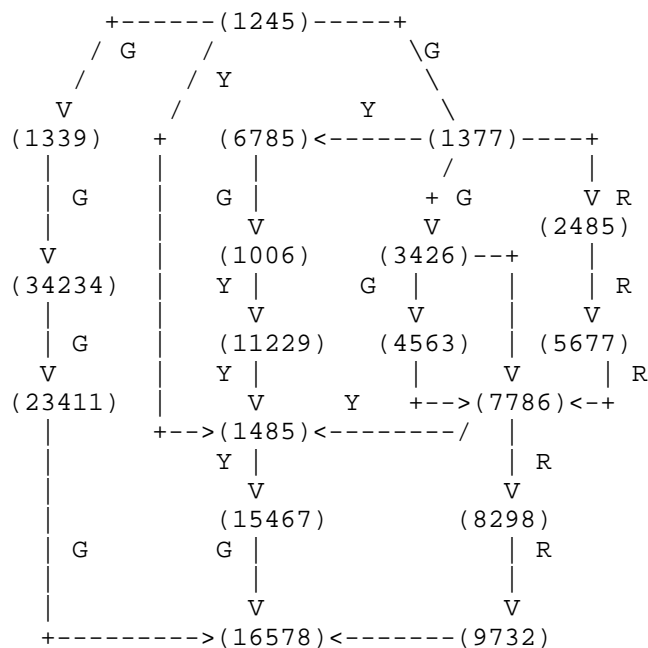
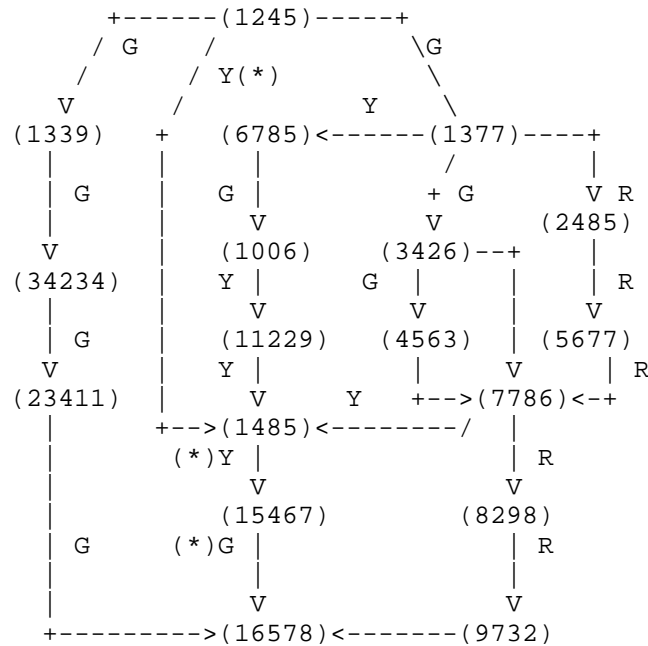


Figure 6: Application of the graph-labeling heuristic. We consider 3 labels "G" < "Y" < "R". Using algorithm 3 the "G" path from the head-end AS 1245 to the tail-end AS 16578 is chosen in the first iteration.

2.2.4 Illustration of graph labeling

We briefly illustrate the graph-labeling algorithm using Figure 6. In this diagram we have categorized the links into three partitions based on the PWR ratio. PWR ratio less than 0:1 are labeled as G, between 0:1 to 0:3 are labeled as Y and the rest as R. The total ordering is defined as $G < Y < R$, where the G links have low PWR ratios than the Y links. The path could be established through the AS that have G as the ingress link; the path being 1245, 1339, 34234, 23411 and 16578.

Once the primary path is constructed then the backup paths may be constructed by excluding to the maximum extent possible the nodes (ASes) and the links (links between ASes) of the primary path from the backup paths so constructed. The links and nodes of the primary path are excluded first. The graph heuristic algorithm is run again. If the path is not available the least power consuming links in the primary path are added back and the construction method repeated again and again till a backup path is constructed.



In the example given above there exist no paths that are GREEN all the way for a backup path as it is in the primary path. So we choose the links having YELLOW as the next category to be included. So the backup path would be from 1245, 1485, 15467 and 16578. This is best secondary / backup path with shortest number of hops that can be constructed with a combination of the GREEN and YELLOW category links. The path is marked by (*) against the label for better illustration.

It is possible that more than one backup path may be required to be constructed. In that case the process is repeated (through iterations) to construct the required number of backup paths.

2.2.5 Moving traffic when the primary path fails

The regular methods of switching traffic from the primary path to the backup path is followed when the primary path fails. This is subject to the regular methods proposed by other documents in MPLS-TE in the ietf.

2.3 Implementation notes and Discussion

In this section we present some notes on feasibility of implementation of our scheme in a live network. First, the requested bandwidth should be available on the low-power path, but the CSPF

algorithm is run with multiple constraints, one of which is the bandwidth requirement for the flows to be transported through the TE-LSP. The PWR ratio can then be applied to the available paths thus computing the low-power paths. Second, as we are using traffic engineering with link state routing protocols, there is a reliable flooding process that are triggered when updates about the change in characteristic arise. We propose addition of some attributes with no change to the protocol implementation. There may be a time lag when the far ends of the Internet receive the attribute and the time it originated. This however cannot be avoided as with other attributes and metrics.

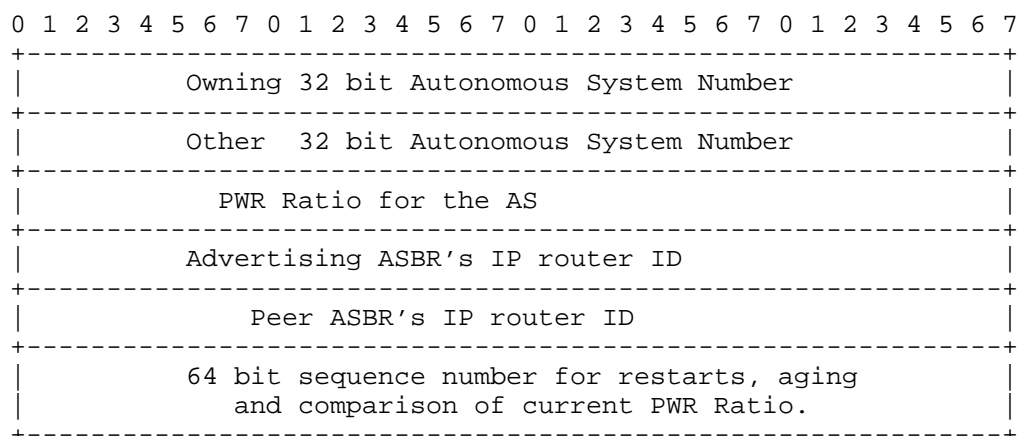


Figure 7: Proposed PDU format with an added attribute for AS-PATH-POWER-METRIC

The additions to the above Attribute have been added to optimize and correctly correlate the connecting ASes and the inter-AS links among them. For the traffic direction into the Advertising AS the above information will be easier to correlate than the previous version which did not advertise the peer AS which had the ingress links into the advertising Router.

In MPLS-TE when the TE metrics are modified, there is a reliable flooding process within an Interior Gateway Protocol (IGP). Such triggered updates apply to the PWR ratio as well. The proposed PWR ratio is advertised to the neighboring AS and the information percolated to all the AS, in a AS-PATH-POWER-METRIC attribute. This attribute can be implemented as shown in Figure 7. The frequency of the updates for this attribute should be fixed to avoid network flooding.

The AS-PATH-POWER-METRIC for each ASBR is calculated, and advertised

as the PWR ratio for the AS. This AS-PATH-POWER-METRIC is filled into the appropriate transitive non-discretionary attribute and inserted into a unique vector for a set of prefixes advertised from the AS. Such advertised prefixes may have originated from the AS or be the transit prefixes. The filled vector is sent to the ASBR of the neighboring AS and the information propagates to all the ASBRs. If the elements denoting AS in a vector of AS-PATH-INFO is not the same as the ones that need to be advertised in a AS-PATH-POWER-METRIC, then a suitable subset of AS-PATH-POWER-METRIC is identified and sent in the BGP updates. A vector of size 1 also can be employed if the AS in question is the only one for which PWR ratio has changed in the originating AS. The collation can be done depending on availability of such metrics and their mapping to a valid AS-PATH-INFO metric.

The power consumed by each router may fluctuate over short time intervals. In order to dampen these fluctuations which can cause unnecessary updates, power can be measured when falling within intervals of suitable size (say a range of values). This is as opposed to measuring power as a discrete quantity. This method of power measurement reduces the frequency of triggered updates from the routers due to power change.

```

0.1      0.2      0.1
(A) ----> (B) ----> (D)

0.1      0.2      0.02      0.2
(A) ----> (C) ----> (E) ----> (D)

0.1      0.2
(D) ----> (X)

```

Figure 8: Example of strands where more than one PWR ratio is advertised by "D"

```

      0.2      0.1      0.2
(A).....>(B).....>(D).....>(X)
|               ^
|0.2      0.02      | 0.2
+---->(C)----->(E)

```

Figure 9:Choice of low-power path derived using the algorithm which uses lower value of the ingress link but through the same AS

A use case of multiple ASBRs advertising differing PWR ratio shows that an AS may be seen as green through one ingress link and not through the other. Consider the case of multiple ASBRs that belong to the same AS, advertising PWR ratios that differ. This could lead to

power values that belong to different classes of ratios with many intervening classes in between. These advertised PWR ratios could lead to one ASBR being preferred over the other thus taking a different path from head-end to tail-end. This also entails that there may be multiple paths to the AS through these different ASBRs.

Consider Figure 8 which shows a set of strands that derive a topology as in Figure 9. Here D is reachable via two paths but the PWR ratios differ. This illustrates the case where the better metric wins out. The average power consumed would not have an effect but the bandwidth available on these ASBR egress links would definitely influence the path.

2.4 Conclusion and Future Work

In this paper, we proposed a scheme for reducing the power consumption of the Internet using collaborative effort between AS. The topology of the Internet is represented using a graph model and derived using the strands obtained from the AS-PATH attribute of the BGP updates. CSPF algorithm is run on this topology by using the PWR ratio as a constraint. The PWR ratio is advertised through the ingress links of the ASBRs associated with AS using BGP updates. The CSPF algorithm finds out the low-power consuming AS that can route data packets from a head-end to a tail-end. Explicit routing is handled through the use of TE-LSPs. This entails adopting routes by choosing entry points to an AS that give energy saving paths. Since using CSPF can be time consuming a heuristic algorithm to derive the low-power paths using graph-labeling was proposed. Our work complements the current schemes for reducing power consumption within a router such as switching off or bringing to power-idle-state certain select components within the forwarding and lookup mechanisms.

2.4.1 Link and node disjoint Backup paths and power constraints

This document in particular stresses the need for building protection and backup paths for primary paths in such a way that both primary and backup paths (either in a 1:1 or 1:N manner) are derived based on the power constraints that are specified in the CSPF algorithm. It is important to note that network operators may require a backup path which is as power conservative as the primary one. Or even to the extent that all backup paths have power constraints. Also it is important that backup paths are completely node and link disjoint with respect to the primary path. In case of failure in the primary path nodes the backup path would satisfy the same power constraints and also avoid the nodes and links which have failed in the primary completely. However if completely disjoint paths are not available a suitable heuristic to include only those necessary nodes and links

from the primary path should be taken. Then these backup paths become partially disjoint to the minimal extent possible.

This Power shortest Path calculation can be taken care of a Path Computation Element (PCE) unit that could be either be a process running on a linecard on a ASBR, or even a core router or an offline engine that is passively listening to the BGP updates within the AS without spitting out any routes of its own. The PCE architecture has already been proposed in the ietf and even has a separate working group for itself. These offline or linecard engines are currently being sold in the market by the networking majors and other companies that develop hardware and software for the PCE. All the PCE needs to do is to accept configuration and passively listen to BGP updates from various peers or even be a client for a route reflector, thus

- a) Accepting these BGP updates
- b) Extracting the AS PATH information from these updates
- c) Then constructing the inter-AS topology
- d) Apply the PWR metric that comes along in these BGP updates to the edges of the graph
- e) Then compute the power shortest path as required by the configuration.

Normally the ASes have SLA agreements between each other to carry X amount of traffic from say a provider A. If the AS representing the ISP then advertises fake figures to carry more traffic than is mandated by the SLA agreement with other providers, then it is to that ISPs detriment since by advertising a better PWR ratio it invites more traffic through it thus getting paid less and carrying more traffic. This is not in the best interest of the ISP. This is so because in the final analysis the Power Shortest Path computed would include it regardless of the amount of traffic to be carried thus causing it to invite more traffic through it than it has accepted, even much more than its capacity. Hence it would be advisable for that ISP to advertise proper PWR ratios and NOT on the lower side of the spectrum. If it advertises HIGHER PWR ratios it would not be chosen, and hence that could be a policy measure NOT to accept any traffic at all since its capacity may be filled up with existing traffic. So advertising on the LOWER side would lead to lesser amount of benefit with respect to dollar per bit transported, and on the HIGHER side would be to exclude it from carrying any traffic that wanted to use the Power Shortest Path.

We also propose that there be a governing body in the IETF or outside it or sponsored by the IETF to verify the power ratios advertised are indeed valid or approximately closer to the actual consumption. A link up for each ISP with a power application level gateway to ensure proper ratios are advertised could be mandated amongst at least the co-operating ISPs (ASes).

The points on which this proposal by us innovates is as follows.

a) There has been no effort prior to this to build an inter-AS topology with a weighted graph based on a PWR ratio. On this point it breaks a new path that would lead to inter-AS co-operation that contributes to power reduction overall in the internet. The paper suggested for OSPF by [10] deals with intra-Autonomous-system scenario rather than an inter-AS one. It is also to be noted that the IGP such as OSPF / IS-IS or any other link-state protocol for that matter is expected to capture the energy consumption of each router within the Autonomous system as in paper [10] to help get a hold on the overall average within the AS, or even sum up the total of all the power consumption within the AS with such intra-AS IGP LSA. This contributes to the PWR ratio proposed in our idea. Thus the intra-AS metric contributes to the PWR ratio. [10] proposal deals with primarily paths setup within an AS and not inter-AS paths. Thus the fundamental problem it solves is different while the problem we solve relates to the inter-AS paths which run across ASes from a head-end AS to a tail-end one.

b) The other aspect of innovation is to use BGP as the piggyback protocol upon which this scheme stands. There has been no effort earlier to approach the internet power reduction problem with BGP as the mode of transport of the energy ratios and coupling it with the inter-AS topology built with AS-PATH-INFO information.

The above 2 are key aspects of innovation.

When links and switches are gated or put into low-power state within an AS, the power-consumption automatically drops at the aggregate level, as a result of which the PWR ratio would be a lower figure advertised through BGP and thus this AS would attract more Power Shortest Path traffic through it. Thus the links within the AS and the switches within it would function more optimally if it had more traffic that went along paths that were originally put in low-power state thus utilizing the paths more effectively, when attracting PSP traffic.

There exist MIBs today that have object identifier for power consumed in a router. Maybe all the related components within it may NOT be listed with regards to power consumed. But the overall power consumed

by the Router / Switch is gettable. Once it is advertised in a opaque Link-State-Advertisement say in the form of a TLV and the LSAs are flooded through the network in an AS, all routers get a uniform picture of which router consumes what power. This method already exists for Traffic engineering Database LSAs that are advertised as LSAs for the purpose of traffic engineering within an AS. We are merely piggybacking on this capability to calculate the PWR ratio at the ASBR which amongst others is yet another Router / Switch of the AS.

Our future work includes looking into computing low-power paths within AS as well. Further it can be noted that the proposed algorithms might lead to increased latency as the number of hops increase, which could be critical for time sensitive applications. Since the PWR ratio could vary dynamically with traffic, the impact of traffic on the algorithm would also be of interest.

2.5 Acknowledgements

Shankar Raman would like to acknowledge the support by BT Public Limited (UK) under the BT IITM PhD Fellowship award. Balaji Venkat and Gaurav Raina would like to acknowledge the UK EPSRC Digital Economy Programme and the Government of India Department of Science and Technology (DST) for funding given to the IU-ATC. Vasan would like to thank Prof.Kamakoti in the Department of Computer Science and engineering for his support.

3 Security Considerations

No specific security considerations apart from the usual considerations with respect to authenticating BGP messages / updates from BGP neighbors is necessary for this scheme.

4 IANA Considerations

A new optional transitive non-discretionary attribute needs to be provided by IANA for carrying the PWR ratio across the Internet in the specified format in BGP.

5 References

5.1 Normative References

5.2 Informative References

REFERENCES

- [1] G. Appenzeller, Sizing router buffers, Doctoral Thesis, Department of Electrical Engineering, Stanford University, 2005.
- [2] A. P. Bianzino, C. Chaudet, D. Rossi and J. L. Rougier, A survey of green networking research, IEEE Communications and Surveys Tutorials, preprint.
- [3] J. Baliga, K. Hinton and R. S. Tucker, Energy consumption of the internet, Proc. of joint international conference on optical internet, June 2007, pp. 1-3.
- [4] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsang and S. Wright, Power awareness in network design and routing, Proc. of the IEEE INFOCOM 2008, April 2008, pp. 457-465.
- [5] B. Venkat et.al, Constructing disjoint and partially disjoint InterAS TE-LSPs, USPTO Patent 7751318, Cisco Systems, 2010.
- [6] M. Xia et. al., Greening the optical backbone network: A traffic engineering approach, IEEE ICC Proceedings, May 2010, pp. 1-5.

- [7] W. Lu and S. Sahni, Low-power TCAMs for very large forwarding tables, IEEE/ACM Transactions on Computer Networks, June 2010, vol. 18, no. 3, pp. 948-959.
- [8] B. Zhang, Routing Area Open Meeting, Proceedings of the IETF 81, Quebec, Canada, July 2011.
- [9] M.J.S Raman, V.Balaji Venkat, G.Raina, Reducing Power consumption using the Border Gateway Protocol, IARIA conferences ENERGY 2012.
- [10] A.Cianfrani et al., An OSPF enhancement for energy saving in IP Networks, IEEE INFOCOM 2011 Workshop on Green Communications and Networking

Authors' Addresses

Shankar Raman
Department of Computer Science and Engineering
I.I.T Madras,
Chennai - 600036
TamilNadu,
India.

EMail: mjsraman@cse.iitm.ac.in

Balaji Venkat Venkataswami
Department of Electrical Engineering,
I.I.T Madras,
Chennai - 600036,
TamilNadu,
India.

EMail: balajivenkat299@gmail.com

Prof.Gaurav Raina
Department of Electrical Engineering,
I.I.T Madras,
Chennai - 600036,
TamilNadu,

India.

EMail: gaurav@ee.iitm.ac.in

Vasan Srin
Department of Computer Science and Engineering
I.I.T Madras,
Chennai - 600036,
TamilNadu,
India.

Email: vasan.vs@gmail.com

PANET Working Group
Internet-Draft
Intended Status: Experimental RFC
Expires: November 4, 2013

Shankar Raman
Balaji Venkat Venkataswami
Gaurav Raina
IIT Madras
May 3, 2013

Building power shortest inter-Area TE LSPs using pre-computed paths
draft-mjsraman-panet-intra-as-psp-te-leak-01

Abstract

In this paper, we propose a framework to reduce the aggregate power consumption of an Autonomous System (AS) using a collaborative approach between areas within an AS. We identify the low-power paths within non-backbone areas and then use Traffic Engineering (TE) techniques to route the packets along the stitched paths from non-backbone areas / backbone area to other non-backbone areas. Such low-power paths can be identified by using the power-to-available-bandwidth (PWR) ratio as an additional constraint in the Constrained Shortest Path First (CSPF) algorithm. For routing the data traffic through these low-power paths, the Inter-Area Traffic Engineered Label Switched Path (TE-LSP) that spans multiple areas can be used. Extensions to the Interior Gateway Protocols like OSPF and IS-IS that support TE extensions can be used to disseminate information about low-power paths in the respective areas (backbone or non-backbone) that minimize the PWR ratio metric on the links within the areas and between the areas thereby creating a collaborative approach to reduce the power consumption.

The feasibility of our approaches is illustrated by applying our algorithm to an AS with a backbone area and several non-backbone areas. The techniques proposed in this paper for the Inter-Area power reduced paths require a few modifications to the existing features of the IGPs supporting TE extensions. The proposed techniques can be extended to other levels of Internet hierarchy, such as Inter-AS paths, through suitable modifications as in [11].

When link state routing protocols like OSPF or ISIS are used to discover TE topology, there is the limitation that traffic engineered paths can be set up only when the head and tail end of the label switched path are in the same area. There are solutions to overcome this limitation either using offline Path Computation Engine (PCE) that attach to multiple areas and know the topology of all areas. This document proposes an alternative approach that does not require any centralized PCE and uses selective leaking of low-power TE path information from one area into other areas.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	4
1.1	Terminology	4
1.1	Low-power routers and switches	4
1.2	Power reduction using routing and traffic engineering	4
2.	Methodology of the proposal	6
2.1	ABR Operation	6
2.1.1	Methodology	7

2.1.2	ERRATA	11
2.1.3	Power Bias	11
2.1.4	Advertising Available POWER	12
2.1.5	ECMP links	12
2.1.6	Dampening the side effects of constant change	12
2.1.7	Calculating power shortest paths in an Area	12
2.1.8	Power profiles of Routers and Switches	13
2.1.8.1	Concave and Convex power curves	15
2.1.8.2	Need to advertise both available power and consumed power	17
2.1.9	Power to Available Bandwidth ratio in a TLV	17
2.2	TE Path Head-end Operation	20
2.2	Suppression of Frequent updates owing to fluctuation in power and bandwidth	22
2.3	Advantages	23
3	Security Considerations	24
4	IANA Considerations	24
5	References	24
5.1	Normative References	24
5.2	Informative References	24
	Authors' Addresses	25

1 Introduction

Estimates of power consumption for the Internet predict a 300% increase, as access speeds increase from 10 Mbps to 100 Mbps [3], [8]. Access speeds are likely to increase as new video, voice and gaming devices get added to the Internet. Various approaches have been proposed to reduce the power consumption of the Internet such as designing low-power routers and switches, and optimizing the network topology using traffic engineering methods [2].

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

1.1 Low-power routers and switches

Low-power router and switch design aim at reducing the power consumed by hardware architectural components such as transmission link, lookup tables and memory. In [4] it is shown that the router's link power consumption can vary by 20 Watts between idle and traffic scenarios. Hence the authors suggest having more line cards and running them to capacity: operating the router at full throughput will lead to less power per bit, and hence larger packet lengths will consume lower power. The two important components in routers that have received attention for high power consumption are buffers and TCAMs. Buffers are built using dynamic RAM (DRAM) or static RAM (SRAM). SRAMs are limited in size and consume more power, but have low access times. Guido [1] states that a 40Gb/s line card would require more than 300 SRAM chips and consume 2.5kW. DRAM access times prevent them from being used on high speed line cards. Sometimes the buffering of packets in DRAM is done at the back end, while SRAM is used at the front end for fast data access. But these schemes cannot scale with increasing line speeds. Some variants of TCAMs have been proposed for increasing line speeds and for reduced power consumption [7].

1.2 Power reduction using routing and traffic engineering

At the Internet level, creating a topology that allows route adaptation, capacity scaling and power-aware service rate tuning, will reduce power consumption. In [8] the author has proposed a technique to traffic engineer the data packets in such a way that the link capacity between routers is optimized. Links which are not utilized are moved to the idle state. Power consumption can be reduced by trading off performance related measures like latency. For

example, power savings while switching from 1 Gbps to 100 Mbps is approximately 4 W and from 100 Mbps to 10 Mbps around 0.1 Watts. Hence instead of operating at 1 Gbps the link speed could be reduced to a lower bandwidth under certain conditions for reduced power consumption.

Multi layer traffic engineering based methods make use of parameters such as resource usage, bandwidth, throughput and QoS measures, for power reduction. In [6] an approach for reducing Intra-AS power consumption for optical networks that uses Dijkstra's shortest path algorithm is proposed. The input to this method assumes the existence of a network topology using which an auxiliary graph is constructed. Power optimization is done on the auxiliary graph and traffic is routed through the low-power links. However, the algorithm expects the topology to be available for getting the auxiliary graph. This topology is easy to obtain for Intra-AS scenario, but by using a centralized PCE (Path Computation Element) as in a hierarchical PCE approach. Here for each area a PCE is assigned and each such PCE calculates the path from a head-end router to a tail-end router, both falling within the same area. When TE paths have to be stitched across several areas then the hierarchical PCE which may be one level up from the respective area PCEs is contacted for such a stitching.

In our approach, we propose a collaborative approach by the respective areas in calculating low-power paths that result in power reduction within an AS. This document proposes an alternative approach that does not require any centralized PCE and uses selective leaking of low-power TE path information from one area into other areas. The core of most ISP ASes use the Multi-Protocol Label Switching (MPLS) technology. MPLS label switched paths that traverse multiple areas carry traffic from a head-end to a tail-end that can be situated in different areas within the AS. The AS uses the Interior Gateway Protocol (IGP) for exchanging routing related information. The topology of one area is not revealed to the other in OSPF-TE and IS-IS-TE.

The CSPF algorithm as proposed here is run on a specific area with the available power-to-bandwidth (PWR) ratio as a constraint, to determine "k" (where k is a suitable number) low-power-paths from the head-end to the tail-end within the same area. The low-cost power paths that minimize the PWR ratio can be exchanged among the collaborating areas using IGP-TE TLVs that we propose in this document. Explicit routing using RSVP-TE (for signalling) then can be achieved between the head-end and the tail-end routers traversing multiple areas through these low-power paths connecting the head-end and tail-end using the Inter-Area Traffic Engineered Label Switched Path (TE-LSP) that span multiple areas.

2. Methodology of the proposal

There are three known solutions to inter-area TE

- (a) hop expansion at area boundaries where the head end can only choose the path to area boundary rather than right to tail end,
- (b) centralized PCE is attached to all areas and is aware of entire topology, and
- (c) path stitching by designating ABRs acting as BGP route reflectors.

It is of course possible to build out low-power paths through the above techniques but they suffer limitations such as not knowing for certain whether the path exists a-priori. This document proposes a technique where a-priori low-power paths are pre-computed in the various areas and are leaked into other areas so that provisioning these paths is done much more quicker than is otherwise possible.

Assume $\{N\}$ as the set of nodes in a network running link state routing protocol and $\{N'\}$ be the set of nodes that are known to be the endpoints of the traffic engineered paths. The topology $\{N, E\}$ has been divided into hierarchical areas with backbone area as the second level that connects first level of all non-backbone areas. We assume the network runs either OSPF-TE or ISIS-TE for establishing TE paths. The set of nodes $\{N'\}$ can be situated in any non-backbone area or the backbone area. Nodes in $\{N'\}$ may become aware of being potential endpoints through offline configuration.

Once the nodes in $\{N'\}$ become aware of being TE endpoints, they advertise themselves in a special TLV in TE link state information. We would term this "TE Endpoint TLV". In OSPF, they would advertise a newly defined TLV in TE LSA and in ISIS, they would advertise a newly defined TLV in TE LSP. Apart from nodes in $\{N'\}$ the area border routers or ABRs advertise another newly defined TLV that we would term as "Area Border TLV".

2.1 ABR Operation

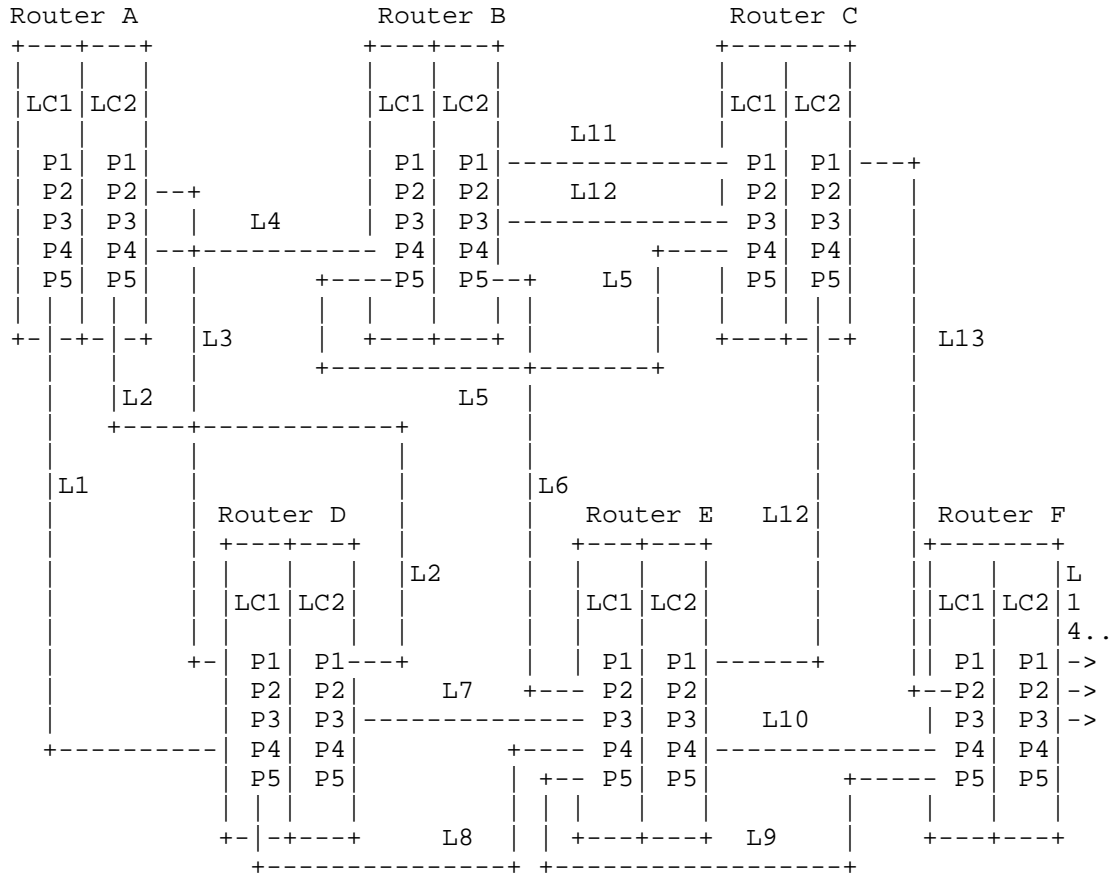
Apart from standard OSPF/ISIS ABR functions, each ABR should discover the TE endpoints in every area attached to it. Assume for an ABR, let the set discovered be $\{A_i, N_j\}$. The ABR should compute k-power-shortest-paths to every element in $\{A_i, N_j\}$ based on the constraints applicable to the network. The constraint applied here is the minimization of the PWR ratio which is defined as follows.

For a given router that is an ABR for an area (straddling the backbone and non-backbone), a set of k-shortest paths that can be potentially be used as a link towards a TE endpoint are identified.

2.1.1 Methodology

For each router / switch there exist linecards and each linecard has a set of ports or sometimes just one port of high capacity. This usually applies on routers and switches that are either single chassis or multi-chassis in their characterisation. By single chassis we mean that there exists a single chassis and slots for the Route Processor Card (one or more of these) typically upto to two of them, and one or more slots for linecards each having their respective characteristics such as number of ports (port density), type of such ports (SONET, ethernet, ATM etc..) usually depending on the link layer technology they support. Links are connections between ports on these linecards to other ports on linecards of other single chassis or multi-chassis system. A multi-chassis system is one that has multiple such chassis interconncted amongst each other to form a single logical view of the system. Both single and multi-chassis have linecards and respective ports on these linecards. Multi-chassis typically have a switch fabric chassis which connects each of these chassis to each other or to chassis of other multi-chassis or single chassis systems.

Consider the following topology as one that falls within an area...



The table of links between the various routers (which are assumed to be single chassis systems) is as follows...

Links	Routers	LC <> LC	Port Conn.	Capacity	Available Bandwidth
L1	A <> D	LC1<>LC1	P5<>P4	10G	7.5
L2	A <> D	LC2<>LC2	P5<>P1	10G	6.0
L3	A <> D	LC2<>LC1	P2<>P1	10G	4.0
L4	A <> B	LC2<>LC1	P4<>P4	10G	3.0
L5	B <> C	LC1<>LC1	P5<>P4	10G	3.5
L6	B <> E	LC1<>LC1	P6<>P2	10G	1.0
L7	D <> E	LC2<>LC1	P3<>P3	10G	6.0
L8	D <> E	LC1<>LC1	P5<>P4	10G	1.5
L9	E <> F	LC1<>LC2	P5<>P5	100G	20.0
L10	E <> F	LC2<>LC1	P4<>P4	10G	2.5
L11	B <> C	LC2<>LC1	P1<>P1	10G	3.0
L12	E <> C	LC2<>LC2	P1<>P5	10G	2.0
L13	C <> F	LC2<>LC1	P1<>P2	10G	1.0
L14	F <> OA	LC2<>	P1<>		

In the above topology assume all point-to-point links between the routers. For now we will deal with P2P links alone and not venture into Broadcast Multi-access links or Non-Broadcast Multi-access links etc.. It is suffice to show how the scheme works for P2P links and then move more specifically to other types of networks to demonstrate this method of calculating the power topology of the network in the figure.

Each linecard consumes a certain amount of power and it is vendor dependent as to how the power consumed relates to the Available Bandwidth on any of the links to which the linecard connects to. It is possible that the said topology of routers come from one vendor or from multiple vendors. It is assumed that the algorithm proposed will have the power consumed by a linecard available as a readable value in terms of W or kW or whichever measurable metric that is provided by the vendor.

It is possible that some of the Linecards are more capable than the others. Consider that Router A is a more capable router with more powerful linecards with higher port density. This is not shown in the figure, but assume so. LC1, LC2 on Router A could be consuming more power than the other Linecards on other routers. The main reason could be that LC1 and LC2 may have higher port density or higher

speed ports than the other routers. In order to calculate the power consumed on a link by a linecard it is important that we normalize the power as power consumed per port. Here the ports are normalized to lowest common denominator. If all links in the topology have 10G port capacity then the power calculated should be in terms power consumed per 10G port.

Assuming we have done this normalization we go on to calculate the POWER metric for each of the ports involved in a link which is derived as follows...

$$\text{POWER metric for a given Port on a LC} = \frac{\text{Power consumed per XG (normalized bandwidth) port}}{\text{Available Bandwidth on that port}}$$

Assume link L1. The ports concerned are both 10G and the ports are P5 on Router A and P4 on Router D. For calculating the POWER metric for a link which we will call PWRLINK we calculate the POWER metric for each side of the link and average the two to get PWRLINK.

$$\text{So PWRLINK for L1} = \frac{\text{POWER for P5 on LC1 on Router A} + \text{Power for P4 on LC1 on Router D}}{2}$$

The above can also be weighted if there is a multi-capacity port on one side of the link and not on the other. A multi-capacity link is one which provides multiple bandwidth capabilities such (1G/10G/100G) for example but auto-negotiates with other end to provide a lesser than highest capacity service.

The PWRLINK metrics once calculated are flooded in already defined OSPF-TE-LSA as an adapted TE-metric and is typically flooded as a link characteristic.

It is important to note that the denominator for POWER metric is Available Bandwidth instead of Available Bandwidth on that port. The Available Bandwidth is measured in terms of intervals and not as discrete quantities. This is in order not to flood PWRLINK metrics into the OSPF area in LSAs very frequently as Bandwidth may constantly change. The same applies to POWER metric as well.

Once the LSAs have been flooded the Routers run CSPF on the graph of the topology with PWRLINKs assigned to the links and calculate the PWRLINK based paths which consume the least power. The shortest power paths based on this topology can be used for forwarding high bandwidth streams and to optimally use power within the area.

The Available Bandwidth column shows the Available bandwidth of the link corresponding to the row and column intersection. This figure is used as the numerator in the POWER metric computation for that port.

2.1.2 ERRATA

ERRATA : Previously the experiments were carried out with Available Utilization since only 10G and 100G ports were considered. This baselines the metric to 10G ports and proportionality thereof. But in reality the actual Available Bandwidth needs to be considered for real world experiments. Hence this draft has been changed to reflect the Available Bandwidth to be taken as the denominator of the formula thereof.

In our previous experiments the 100G link if it showed a utilization of 0.2 would end up as a high POWER metric and hence would be totally avoided. In reality this link may have been a more power optimal link given that if it had a first power profile (Please refer section on Power Profiles). Dividing the Power consumed or Available Power by the Available Bandwidth gives a better picture of how much power cost per Gb is consumed and normalizes the metric amongst links of varying bandwidth.

An earlier version of this document rev-00 contained a different algorithm to compute the k-shortest-power-paths. From the experimental results gathered it was seen that the said algorithm was prone to errors with respect to direction of traffic and unnecessarily complex for the solution. Hence it has been set aside for a more simple yet better one mentioned in this revision.

2.1.3 Power Bias

Assume in the figure that there exist Routers A and D and that there is a bias on the link L1 in such a way that Router D computes a POWER metric of 10 and the Router A computes a POWER metric of 2 on the ports P5 and P4 respectively. Now the PWRLINK would be 6 for that link L1. Thus even if one side is excessively power guzzling then the PWRLINK moves up and thus is less preferred in the CSPF algorithm and path computation based on the Power topology.

If there is no bias and both the sides of the link are optimal in their power usage then the metric stays low even if more streams are sent on it. This is the main objective that is set out for router and switch manufacturers in the single chassis and multi-chassis world, in that they are incentivised to manufacture linecards that are not power hungry even if the number of packets flowing through them is high and thus the Bandwidth Available is also reasonably on the higher side compared to other routers.

For those manufacturers who set a high power value for even minimal traffic, the vendors that don't would win out in the end.

2.1.4 Advertising Available POWER

Please see section 2.1.8 for more information on why Available POWER plays a crucial role in determining the choice of routers based on the Power metric.

2.1.5 ECMP links

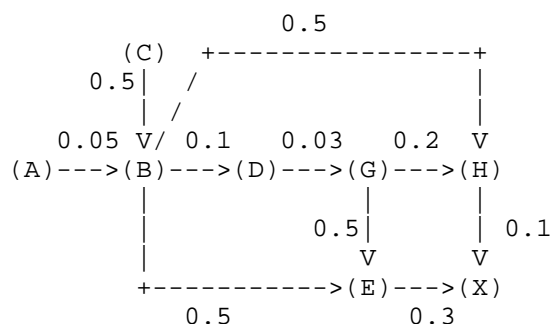
It is possible that multiple links would have the same PWRLINK metric after a computation cycle. In such a case load-balancing techniques can be used to keep the ECMP links in a steady state with respect to each other. Depending on the Available Bandwidth thereafter it is possible that the ECMP links may no longer be Equal cost but UCMP or Unequal Cost Paths.

2.1.6 Dampening the side effects of constant change

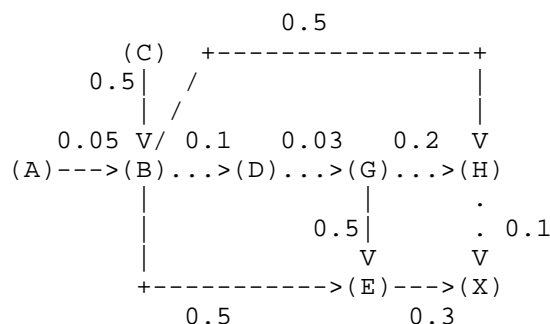
It is recommended in this draft that the implementation of the proposal be adaptive, infrequent in computation to the extent possible without sacrificing adapting to the dynamism and also reduce any frequent oscillations. The actual methods to adopt for this computation are outside the scope of this document.

2.1.7 Calculating power shortest paths in an Area

Assume the following topology where A,B,C etc.. are routers and corresponding labelled edges with weights are the links. These weights are the current values of the PWRLINK attribute that has been flooded in the LSAs through the Area concerned. Assume B is the ABR for Area 1 and the routers A and C are the Area 0 core routers. The rest of the routers are assumed to be in Area 1. Once the power topology of the Area 1 has been calculated as shown below with the PWRLINK attributes being assigned to the links, Constrained shortest path can be run from the ABR to any of the other routers say H, E, X etc.. The CSPF algorithm takes the constraint in terms of the PWRLINK attributes along with other attributes to construct a power shortest path from say router B to other routers in Area 1.



Once the path has been computed it is possible to use RSVP-TE to construct the power shortest path with the TE-LSP being instantiated with the labels appropriately placed in the routers on the power shortest path. In this topology, assume one would want to construct a path from B to X then the dotted path shows the path constructed and to be used by a set of flows or streams of packets belonging to multiple flows as seen fit by the router B. If the PWRLINK metrics change after due course of time then another power shortest path that possibly traverses the same path (if the SUM of PWRLINKs doesn't exceed any other path's metrics' SUM) or some other path would be constructed. Specifically this method makes use of traffic-engineering signalling protocols as the method to place the streams from point X to point Y (where X and Y are routers).

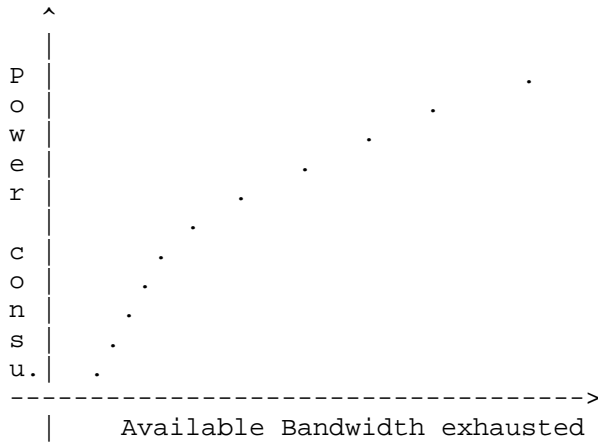


2.1.8 Power profiles of Routers and Switches

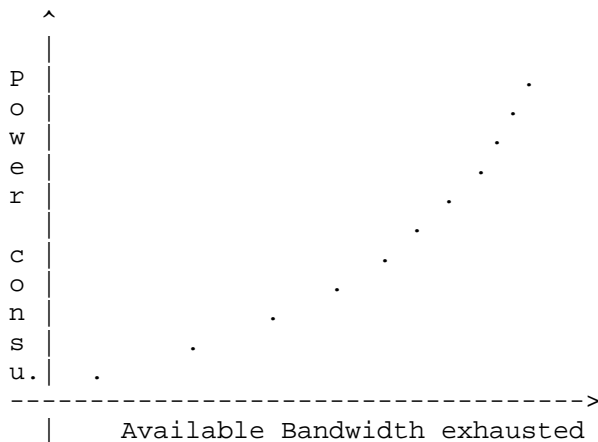
It has been experimented and from several sources found that there exist routers which have different power profiles. The power profile of a router is the curve of power consumption to available bandwidth. Mentioned below are a few of these prominent ones that have to be taken into consideration.

The first profile that we will consider is the flattening curve. The power consumed to available bandwidth curve takes the shape of a

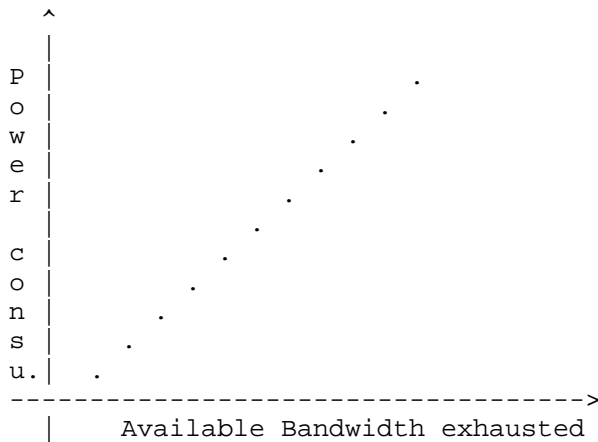
steep one initially and then tapers off to a plateau. The point at which it begins to give a ΔC (Δ in Power Consumed) to ΔB (Available Bandwidth exhausted) is the inflection point that tapers off to a plateau. Here the $\Delta C/\Delta B$ begins to slow down or decrease rapidly. The more the traffic that is added onto the device the lesser it draws power.



The second profile that we will consider is the exponential curve. The power consumed to available bandwidth curve takes the shape of an ever increasing steep curve as shown below. Here the $\Delta C/\Delta B$ begins to increase as more traffic is thrown onto it as the Available bandwidth exhausted increases. This power curve beyond a point is intolerable with respect to power guzzling.



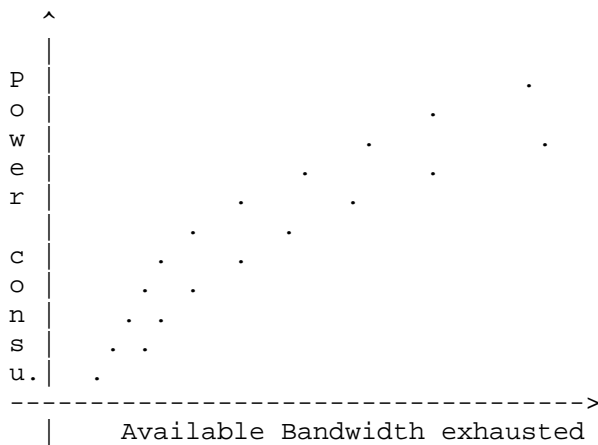
The third profile that we will consider is a linear curve. In other words just a straight line. Here $\Delta C / \Delta B$ is a constant.



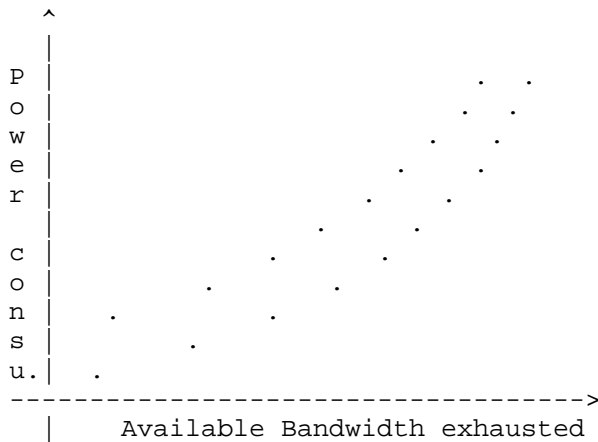
2.1.8.1 Concave and Convex power curves

Given that there are 3 kinds of major profiles in the router power consumption, what line would we like to pick. This is an important point when choosing the metric to pick the low power paths.

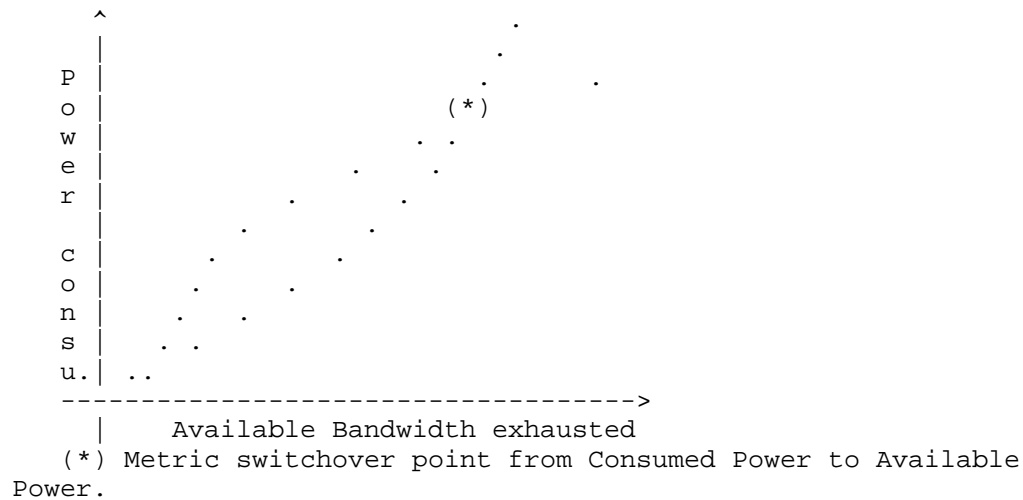
(a) If the confrontation is between 2 first profile routers the lower of the 2 would be considered as shown below. The lower curve offers better power savings for each GB of bandwidth transported.



(b) If the confrontation is between 2 second profile routers the upper curve offers more power savings per GB of bandwidth.



(c) When the confrontation is between a first profile curve and a second profile curve, it would be optimal to pick (as shown below) the lower of the curves because it gives us lesser power consumed for every GB of traffic routed / switched. Here the exponential curve is the one that offers lesser amount of power consumed per GB of traffic is chosen. But when it gets to a point that the two curves intersect it would be more optimal to pick the tapering curve. Thus at the meeting point of the 2 curves the exponential curve becomes more costly and the tapering one gives us more GB for the power buck. Thus this switchover from one curve to the other (in other words from the exponential curve to the tapering one) does the trick in terms of finding an optimal solution.



2.1.8.2 Need to advertise both available power and consumed power

Thus the above sections have shown that both the available power and the consumed power MUST be advertised so that case (c) can be deciphered and the switchover of the curves be done and the appropriate router be chosen for the rest of the bandwidth to be switched over to.

Thus there will exist Consumed-Power to Available Bandwidth ratio and the Available Power to Available Bandwidth ratio. Both the ratios are computed and the lower value chosen. The Available Power can be judged from the calibration process such as the one carried out by independent test organizations as in [12]. An example of their calibration is referred to in [12].

Here given below is the formula for calculating the Available Power to Available Bandwidth ratio also called the Available POWER metric.

$$\text{Available POWER metric} = \frac{\text{Available Power consumed per XG (normalized bandwidth) port}}{\text{Available Bandwidth on that port}}$$

for a given Port on a LC

2.1.9 Power to Available Bandwidth ratio in a TLV

As per [RFC3630] the Link TLV can be used to carry this power to available Bandwidth ratio with an additional sub-TLV of the link TLV. The sub-type number 11 is recommended to be defined for this purpose.

[RFC 3630] states in section 2.2.1 and we QUOTE ...

2.1.10 Link TLV

The Link TLV describes a single link. It is constructed of a set of sub-TLVs. There are no ordering requirements for the sub-TLVs.

Only one Link TLV shall be carried in each LSA, allowing for fine granularity changes in topology.

The Link TLV is type 2, and the length is variable.

The following sub-TLVs of the Link TLV are defined:

- 1 - Link type (1 octet)
- 2 - Link ID (4 octets)
- 3 - Local interface IP address (4 octets)
- 4 - Remote interface IP address (4 octets)
- 5 - Traffic engineering metric (4 octets)
- 6 - Maximum bandwidth (4 octets)
- 7 - Maximum reservable bandwidth (4 octets)
- 8 - Unreserved bandwidth (32 octets)
- 9 - Administrative group (4 octets)
- 10 - Power-to-Multicast-replication-capacity (4 octets)
- 11 - Consumed-Power-to-Available-Bandwidth (4 octets)
- 12 - Available-Power-to-Available-Bandwidth (4 octets)

This memo defines sub-Types 1 through 9. See the IANA Considerations in [RFC3630] section for allocation of new sub-Types.

The Link Type and Link ID sub-TLVs are mandatory, i.e., must appear exactly once. All other sub-TLVs defined here may occur at most once. These restrictions need not apply to future sub-TLVs. Unrecognized sub-TLVs are ignored.

Various values below use the (32 bit) IEEE Floating Point format. For quick reference, this format is as follows:

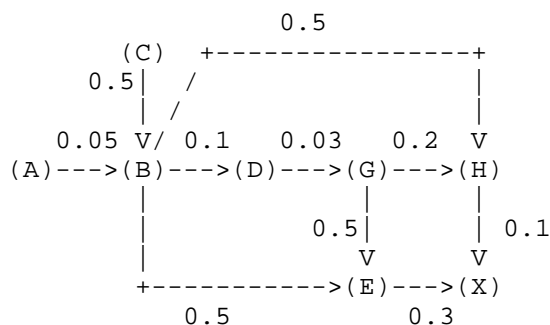
0	1	2	3
0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1
+-----+			
S	Exponent	Fraction	
+-----+			

S is the sign, Exponent is the exponent base 2 in "excess 127" notation, and Fraction is the mantissa - 1, with an implied binary point in front of it. Thus, the above represents the value:

$$(-1)^{(S)} * 2^{(Exponent-127)} * (1 + Fraction)$$

It is proposed that we use the Power-to-Available-Bandwidth ratio as a 32 bit IEEE floating Point format field for the purpose of this document.

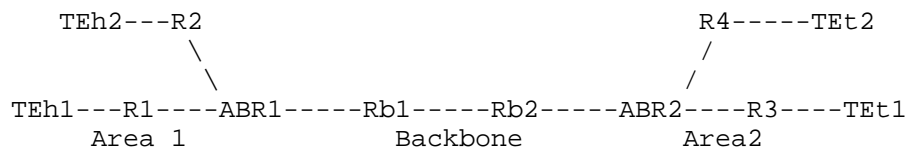
Assume the following topology in a non-backbone area after calculating the PWR ratio in a given stage of the algorithm.



Here (B) is a Area Border Router and has to ingress links into it from (C) and (A) which are in the backbone area. Connectivity within the backbone area are not shown here. Assume (C) and (A) are connected in some way with other routers in the backbone area. Routers (D), (G), (E), (H), (X) are routers in the non-backbone area. Routers (H), (E) and (X) are potential TE endpoints. The PWR metrics shown here on the edges within the area represent metrics for a specific TE endpoint. The metrics on edges (C)->(B) and (A)->(B) are for any traffic ingressing through (B) into the non-backbone area heading towards any TE endpoint (H), (E) or (X).

The number of constraints is likely to be few and the most widely used constraints are TE metric, link groups and bandwidth. But no restriction is assumed on use of other constraints. Thus here we add the PWR metric of a link as an additional constraint. Once the ABR computes k-power-shortest-paths to every {Ai, Nj} it has topology information about, it advertises the k-power-shortest-paths as a reachability vector in a newly defined "TE Reachability Vector TLV".

Consider an example network show below. TEh is head-end and TEt is tail-end of a TE path, ABR1 and ABR2 are area border routers.



In this example, ABR1's TE Reachability vector TLV for area 1 and area 0 are given below.

```
{ ABR1, [<TEh1, <Path info 1>>, <TEh2, <Path info 2>>]}
{ ABR1, ABR2, [<Tet1, <Path info 3>>, <Tet2, <Path info 4>>]}
```

Here the vector TLVs are arranged as per increasing PWR metric associated with each path. That is the summation of all PWR metrics of the links in the path is done and the vector TLVs are ordered in increasing order of PWR metric sums. So the lowest-cost-power path is listed first and so on. If the least cost power path is to be chosen then the path in the first TLV is chosen.

Similarly ABR2's TE Reachability vector TLV for area 2 and area 0 are given below.

```
{ ABR2, [<Tet1, <Path info 3>>, <Tet2, <Path info 4>>]}
{ ABR2, ABR1, [<TEh1, <Path info 1>>, <TEh2, <Path info 2>>]}
```

The first thing to be noted is that head-ends are also considered as TE-endpoints. Essentially this means any head-end or tail-end of a inter-area TE-LSP can be considered as tail-end or head-end respectively.

Note that the reachability vector advertised by ABR1 also contains the reachability vector of ABR2. For example, if ABR2 is brought up first, then it is likely that ABR1 would only have the following as TE Reachability vector TLV for area 0 before ABR2 computes path to the TE endpoints in area 2. { ABR1, ABR2 }

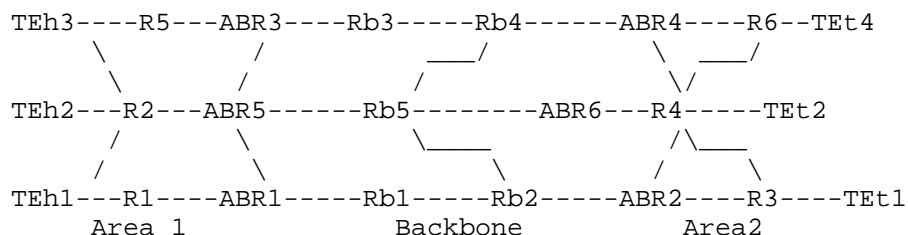
Note that <Path info> TLV would only contain the aggregate of link attributes namely cost, bandwidth etc and most importantly the PWR metric as well but not the complete path of intermediate nodes. For example, <Path info 1> may be a set of <2, admin-group-1|admin-group-2, 1Gbps> (where the 1Gbps could be the minimum bw available along the path). The above example topology has only one path from ABRs to TE endpoints. The number of path info "k" may have a default value or can be configured by the operator on all nodes.

2.2 TE Path Head-end Operation

When any TE application requests TE path to be setup to an endpoint

that is not present in the same area, the head-end scans the TE Reachability vector TLVs advertised by ABRs and selects the path using the <Path info> contained in the vector TLVs.

Here is an example with multiple paths in area 1, backbone and area 2 called Figure 2.0



In this topology in figure 2.0 taking the tail-ends represented in the diagram, it is noted that TET4 is reachable via ABR4, ABR6 and ABR2 as well. The TE reachability TLVs advertised by ABR6 for area 2 would be multiple to each tail-end since there exist multiple paths to reach at least most of them in area 2 once a packet reaches any of the ABRs in area 2.

Here again the least cost power shortest path is listed first and so on.

```
{ ABR6, [<TEt4, <Path info 1>>, <TEt4, <Path info 2>>, <TEt2, <Path
Info 3>>, etc.. }
```

For area 0 the TE reachability TLV would be

```
{ ABR6, ABR1, [<TEh1, <Path info 4>>, <TEh1, <Path info 5>>...]}
{ ABR6, ABR5, [<TEh1, <Path info 6>>, <TEh1, <Path info 7>>...]}
{ ABR6, ABR3, [<TEh1, <Path info 8>>, <TEh1, <Path info 9>>...]}
```

For the sake of brevity we do not enumerate all path information possible as it would be quite extensive.

It is possible that there may be already setup LSPs which are being used for transit traffic on the backbone or in other non-backbone areas. It is also feasible to advertize already set up LSPs in the path info; no additional TLV is required for that purpose. The case where this may be useful would be if such transport LSPs exist in the backbone area and there is a willingness to provide higher preference to these LSPs to carry transit LSPs over backbone.

There can be selective suppression of advertisements to other areas

(backbone or non-backbone) of LSPs if these are existing LSPs setup along a path which are utilized to a greater degree. If underutilized with respect to the PWR metric a more favourable metric could be advertized to other areas.

For example, backbone area transport LSPs will be advertized as transit LSPs which would provide connectivity to LSP sections lying in non-backbone areas and would be updated more frequently since they facilitate inter-Area TE.

Once a path in the TLV has been used for reserving bandwidth for traffic over that path, then it is withdrawn from the advertisements so that it becomes unusable. Another path may be computed over the same path but with possibly a different PWR metric sum since it is possible that the traffic over that path could have changed the PWR metrics in the edges along that path.

2.2 Suppression of Frequent updates owing to fluctuation in power and bandwidth

Using the power consumed and the bandwidth available as discrete quantities will result in frequent oscillations. Such a step would result will result in frequent re-computations of the shortest power paths. For the sake of suppression of such frequent updates, it is possible to handle the PWR metric as falling within reasonable intervals of thresholds. If the interval in which PWR metric lies is moved out of and another interval is reached then the update is sent out in the IGP-TE mechanism. Otherwise if the interval in which the PWR metric lies is not moved out of then the updates are not sent. Suitable thresholds can be arrived at after suitable calibration through tests.

Routers may have step levels in which they increase power consumption when they additively are loaded with more large bandwidth consuming multicast or unicast streams. Calibrating these levels may be useful for implementing this scheme. It is possible that such calibrated thresholds can be used for advertising the PWRLINK ratios in the OSPF LSA advertisements. This would be useful for bringing down the frequency of updates or advertisements from a line-card about its PWRLINK ratio. When power consumption meanders within a certain given interval these ratios need not be re-advertised even if further unicast and/or multicast streams are added to it. The incentive is to recognize a linecard that does not drastically change power consumption even if large bandwidth streams are added onto it for forwarding and thus give it credit for its power optimal functioning. If a router tends to consume the highest level of power even when carrying low amounts of unicast and multicast streams on its line card, it would automatically have a poor ratio when compared to a

router that efficiently uses power when considering the Available Bandwidth being observed. The best case would be a low power consuming line-card or a router filled with such line cards that does not leave its power interval no matter how much ever capacity is sought to be used on it. But that would be an ideal condition but it is definitely an idealistic scenario towards which the router manufacturers should look at.

2.3 Advantages

- 1) The TE Reachability vector TLV contains the aggregate of all link attributes along with TE constraints and so the head-end of the TE path can explicitly select the ABR that connects the destination area even though it does not know the complete topology of the backbone area.
- 2) As the TE reachability vector contains only the aggregate attributes of k-power-shortest-paths, the flooding overhead to support the mechanism is limited.
- 3) Centralized path computation element is not required for supporting inter-area power-shortest-path TE. The additional overhead of computing k-power-shortest-paths on ABR can be solved by offloading the computation overhead to additional processor in multi-core platforms.

3 Security Considerations

None.

4 IANA Considerations

New TLV types for OSPF and IS-IS for the new TLVs that have been introduced need to be assigned.

5 References

5.1 Normative References

5.2 Informative References

- [1] G. Appenzeller, Sizing router buffers, Doctoral Thesis, Department of Electrical Engineering, Stanford University, 2005.
- [2] A. P. Bianzino, C. Chaudet, D. Rossi and J. L. Rougier, A survey of green networking research, IEEE Communications and Surveys Tutorials, preprint.
- [3] J. Baliga, K. Hinton and R. S. Tucker, Energy consumption of the internet, Proc. of joint international conference on optical internet, June 2007, pp. 1-3.
- [4] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsang and S. Wright, Power awareness in network design and routing, Proc. of the IEEE INFOCOM 2008, April 2008, pp. 457-465.
- [5] B. Venkat et.al, Constructing disjoint and partially disjoint InterAS TE-LSPs, USPTO Patent 7751318, Cisco Systems, 2010.
- [6] M. Xia et. al., Greening the optical backbone network: A traffic engineering approach, IEEE ICC Proceedings, May 2010, pp. 1-5.
- [7] W. Lu and S. Sahni, Low-power TCAMs for very large forwarding tables, IEEE/ACM Transactions on Computer Networks, June 2010, vol. 18, no. 3, pp. 948-959.
- [8] B. Zhang, Routing Area Open Meeting, Proceedings of the IETF 81, Quebec, Canada, July 2011.

- [9] M.J.S Raman, V.Balaji Venkat, G.Raina, Reducing Power consumption using the Border Gateway Protocol, IARIA conferences ENERGY 2012.
- [10] A.Cianfrani et al., An OSPF enhancement for energy saving in IP Networks, IEEE INFOCOM 2011 Workshop on Green Communications and Networking
- [11] Shankar Raman et al., draft-mjsraman-rtgwg-inter-as-psp-01.txt, Work in Progress, February 2012.

Authors' Addresses

Shankar Raman
Department of Computer Science and Engineering
IIT Madras
Chennai - 600036
TamilNadu
India

Email: mjsraman@cse.iitm.ac.in

Balaji Venkat Venkataswami
Department of Electrical Engineering
IIT Madras
Chennai - 600036
TamilNadu
India

Email: balajivenkat299@gmail.com

Prof.Gaurav Raina
Department of Electrical Engineering
IIT Madras
Chennai - 600036
TamilNadu
India

Email: gaurav@ee.iitm.ac.in

PANET Working Group
INTERNET-DRAFT
Intended Status: Experimental RFC
Expires: September 28, 2013

Shankar Raman
Balaji Venkat Venkataswami
Gaurav Raina
Vasan Srini
IIT Madras
March 27, 2013

Power Based Topologies and TE-Shortest Power Paths in OSPF
draft-mjsraman-panet-ospf-power-topo-02

Abstract

In a Interior Gateway Protocol like OSPF (Open Shortest Path First) the computation of the Constrained shortest path to destinations is computed for an area say a backbone or a non-backbone area using the TE-metrics advertised in the area. With importance given to the reduction of power within a network it becomes important to provide a solution that reduces the power consumed amongst routers and links that make up the network (in this case an area or a collection of areas including the backbone and non-backbone areas). This proposal aims at providing such a solution by producing a power topology of the area / areas. This power topology is constructed by assigning metrics to links based on the power consumed by the linecards (and hence their respective ports in an indirect way) of adjacent routers that are interconnected by each such link.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1 Terminology	3
1.2 Low-power routers and switches	3
1.3 Power reduction using routing and traffic engineering	3
2. Methodology	4
2.0.1 ERRATA	8
2.1 Power Bias	8
2.1.1 Advertising Available POWER	8
2.2 ECMP links	9
2.3 Dampening the side effects of constant change	9
2.4 Calculating power shortest paths in an Area	9
2.4.1 Power profiles of Routers and Switches	10
2.4.1.3 Need to advertise both available power and consumed power	14
2.4.2 Power to Available Bandwidth ratio in a TLV	14
3. Conclusion	16
3 Security Considerations	17
4 IANA Considerations	17
5 References	17
5.1 Normative References	17
5.2 Informative References	17
Authors' Addresses	18

1. Introduction

Estimates of power consumption for the Internet predict a 300% increase, as access speeds increase from 10 Mbps to 100 Mbps [3], [8]. Access speeds are likely to increase as new video, voice and gaming devices get added to the Internet. Various approaches have been proposed to reduce the power consumption of the Internet such as designing low-power routers and switches, and optimizing the network topology using traffic engineering methods [2].

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

1.2 Low-power routers and switches

Low-power router and switch design aim at reducing the power consumed by hardware architectural components such as transmission link, lookup tables and memory. In [4] it is shown that the router's link power consumption can vary by 20 Watts between idle and traffic scenarios. Hence the authors suggest having more line cards and running them to capacity: operating the router at full throughput will lead to less power per bit, and hence larger packet lengths will consume lower power. The two important components in routers that have received attention for high power consumption are buffers and TCAMs. Buffers are built using dynamic RAM (DRAM) or static RAM (SRAM). SRAMs are limited in size and consume more power, but have low access times. Guido [1] states that a 40Gb/s line card would require more than 300 SRAM chips and consume 2.5kW. DRAM access times prevent them from being used on high speed line cards. Sometimes the buffering of packets in DRAM is done at the back end, while SRAM is used at the front end for fast data access. But these schemes cannot scale with increasing line speeds. Some variants of TCAMs have been proposed for increasing line speeds and for reduced power consumption [7].

1.3 Power reduction using routing and traffic engineering

At the Internet level, creating a topology that allows route adaptation, capacity scaling and power-aware service rate tuning, will reduce power consumption. In [8] the author has proposed a technique to traffic engineer the data packets in such a way that the link capacity between routers is optimized. Links which are not utilized are moved to the idle state. Power consumption can be reduced by trading off performance related measures like latency. For

example, power savings while switching from 1 Gbps to 100 Mbps is approximately 4 W and from 100 Mbps to 10 Mbps around 0.1 Watts. Hence instead of operating at 1 Gbps the link speed could be reduced to a lower bandwidth under certain conditions for reduced power consumption.

Multi layer traffic engineering based methods make use of parameters such as resource usage, bandwidth, throughput and QoS measures, for power reduction. In [6] an approach for reducing Intra-AS power consumption for optical networks that uses Dijkstra's shortest path algorithm is proposed. The input to this method assumes the existence of a network topology using which an auxiliary graph is constructed. Power optimization is done on the auxiliary graph and traffic is routed through the low-power links. However, the algorithm expects the topology to be available for getting the auxiliary graph. While [6] handles optical networks and their corresponding power consumption, it does not take into account other link layer technologies. It is specialized for optical and not for heterogeneous links that will exist in common OSPF domains.

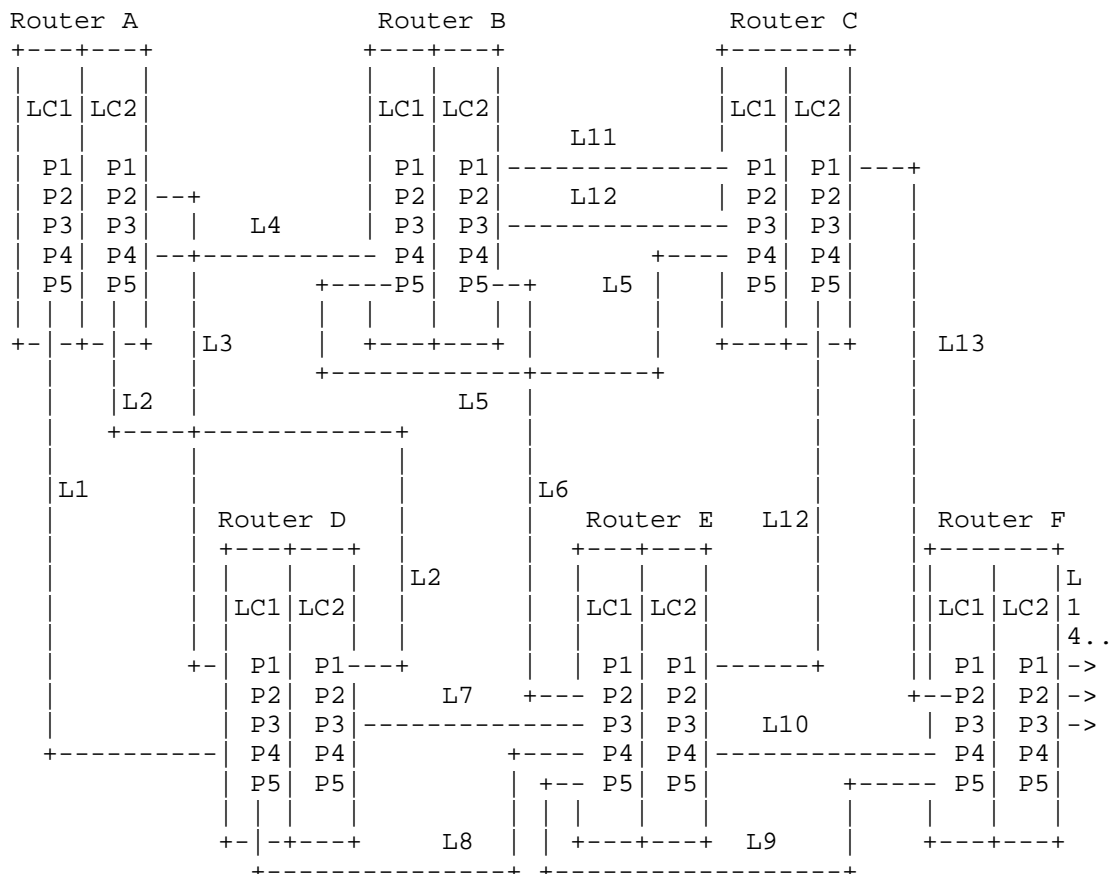
The proposal we make in this document indicates ways to solve the power reduction problem, by calculating a POWER metric whose importance is highlighted in the below mentioned sections. This POWER metric is obtained by including the factors such as power consumed by a linecard on a single chassis or multi-chassis router and consequently a port on that linecard by proportionally calculating power consumed for that port and hence for the link. The other factor that is taken into account is the Available Bandwidth on that port and hence on that link.

2. Methodology

For each router / switch there exist linecards and each linecard has a set of ports or sometimes just one port of high capacity. This usually applies on routers and switches that are either single chassis or multi-chassis in their characterisation. By single chassis we mean that there exists a single chassis and slots for the Route Processor Card (one or more of these) typically upto to two of them, and one or more slots for linecards each having their respective characteristics such as number of ports (port density), type of such ports (SONET, ethernet, ATM etc..) usually depending on the link layer technology they support. Links are connections between ports on these linecards to other ports on linecards of other single chassis or multi-chassis system. A multi-chassis system is one that has multiple such chassis interconnected amongst each other to form a single logical view of the system. Both single and multi-chassis have

linecards and respective ports on these linecards. Multi-chassis typically have a switch fabric chassis which connects each of these chassis to each other or to chassis of other multi-chassis or single chassis systems.

Consider the following topology...



The table of links between the various routers (which are assumed to be single chassis systems) is as follows...

Links	Routers	LC <> LC	Port Conn.	Capacity	Available Bandwidth
L1	A <> D	LC1<>LC1	P5<>P4	10G	7.5
L2	A <> D	LC2<>LC2	P5<>P1	10G	6.0
L3	A <> D	LC2<>LC1	P2<>P1	10G	4.0
L4	A <> B	LC2<>LC1	P4<>P4	10G	3.0
L5	B <> C	LC1<>LC1	P5<>P4	10G	3.5
L6	B <> E	LC1<>LC1	P6<>P2	10G	1.0
L7	D <> E	LC2<>LC1	P3<>P3	10G	6.0
L8	D <> E	LC1<>LC1	P5<>P4	10G	1.5
L9	E <> F	LC1<>LC2	P5<>P5	100G	20.0
L10	E <> F	LC2<>LC1	P4<>P4	10G	2.5
L11	B <> C	LC2<>LC1	P1<>P1	10G	3.0
L12	E <> C	LC2<>LC2	P1<>P5	10G	2.0
L13	C <> F	LC2<>LC1	P1<>P2	10G	1.0
L14	F <> OA	LC2<>	P1<>		

In the above topology assume all point-to-point links between the routers. For now we will deal with P2P links alone and not venture into Broadcast Multi-access links or Non-Broadcast Multi-access links etc.. It is suffice to show how the scheme works for P2P links and then move more specifically to other types of networks to demonstrate this method of calculating the power topology of the network in the figure.

Each linecard consumes a certain amount of power and it is vendor dependent as to how the power consumed relates to the Available Bandwidth on any of the links to which the linecard connects to. It is possible that the said topology of routers come from one vendor or from multiple vendors. It is assumed that the algorithm proposed will have the power consumed by a linecard available as a readable value in terms of W or kW or whichever measurable metric that is provided by the vendor.

It is possible that some of the Linecards are more capable than the others. Consider that Router A is a more capable router with more powerful linecards with higher port density. This is not shown in the figure, but assume so. LC1, LC2 on Router A could be consuming more power than the other Linecards on other routers. The main reason could be that LC1 and LC2 may have higher port density or higher

speed ports than the other routers. In order to calculate the power consumed on a link by a linecard it is important that we normalize the power as power consumed per port. Here the ports are normalized to lowest common denominator. If all links in the topology have 10G port capacity then the power calculated should be in terms power consumed per 10G port.

Assuming we have done this normalization we go on to calculate the POWER metric for each of the ports involved in a link which is derived as follows...

$$\text{POWER metric for a given Port on a LC} = \frac{\text{Power consumed per XG (normalized bandwidth) port}}{\text{Available Bandwidth on that port}}$$

Assume link L1. The ports concerned are both 10G and the ports are P5 on Router A and P4 on Router D. For calculating the POWER metric for a link which we will call PWRLINK we calculate the POWER metric for each side of the link and average the two to get PWRLINK.

$$\text{So PWRLINK for L1} = \frac{\text{POWER for P5 on LC1 on Router A} + \text{Power for P4 on LC1 on Router D}}{2}$$

The above can also be weighted if there is a multi-capacity port on one side of the link and not on the other. A multi-capacity link is one which provides multiple bandwidth capabilities such (1G/10G/100G) for example but auto-negotiates with other end to provide a lesser than highest capacity service.

The PWRLINK metrics once calculated are flooded in already defined OSPF-TE-LSA as an adapted TE-metric and is typically flooded as a link characteristic.

It is important to note that the denominator for POWER metric is Available Bandwidth instead of Available Bandwidth on that port. The Available Bandwidth is measured in terms of intervals and not as discrete quantities. This is in order not to flood PWRLINK metrics into the OSPF area in LSAs very frequently as Bandwidth may constantly change. The same applies to POWER metric as well.

Once the LSAs have been flooded the Routers run CSPF on the graph of the topology with PWRLINKs assigned to the links and calculate the PWRLINK based paths which consume the least power. The shortest power paths based on this topology can be used for forwarding high bandwidth streams and to optimally use power within the area.

The Available Bandwidth column shows the Available bandwidth of the link corresponding to the row and column intersection. This figure is used as the numerator in the POWER metric computation for that port.

2.0.1 ERRATA

ERRATA : Previously the experiments were carried out with Available Utilization since only 10G and 100G ports were considered. This baselines the metric to 10G ports and proportionality thereof. But in reality the actual Available Bandwidth needs to be considered for real world experiments. Hence this draft has been changed to reflect the Available Bandwidth to be taken as the denominator of the formula thereof.

In our previous experiments the 100G link if it showed a utilization of 0.2 would end up as a high POWER metric and hence would be totally avoided. In reality this link may have been a more power optimal link given that if it had a first power profile (Please refer section on Power Profiles). Dividing the Power consumed or Available Power by the Available Bandwidth gives a better picture of how much power cost per Gb is consumed and normalizes the metric amongst links of varying bandwidth.

2.1 Power Bias

Assume in the figure that there exist Routers A and D and that there is a bias on the link L1 in such a way that Router D computes a POWER metric of 10 and the Router D computes a POWER metric of 2 on the ports P5 and P4 respectively. Now the PWRLINK would be 6 for that link L1. Thus even if one side is excessively power guzzling then the PWRLINK moves up and thus is less preferred in the CSPF algorithm and path computation based on the Power topology.

If there is no bias and both the sides of the link are optimal in their power usage then the metric stays low even if more streams are sent on it. This is the main objective that is set out for router and switch manufacturers in the single chassis and multi-chassis world, in that they are incentivised to manufacture linecards that are not power hungry even if the number of packets flowing through them is high and thus the Bandwidth Available is also reasonably on the higher side compared to other routers.

For those manufacturers who set a high power value for even minimal traffic, the vendors that dont would win out in the end.

2.1.1 Advertising Available POWER

Please see section 2.4.1 for more information on why Available POWER

plays a crucial role in determining the choice of routers based on the Power metric.

2.2 ECMP links

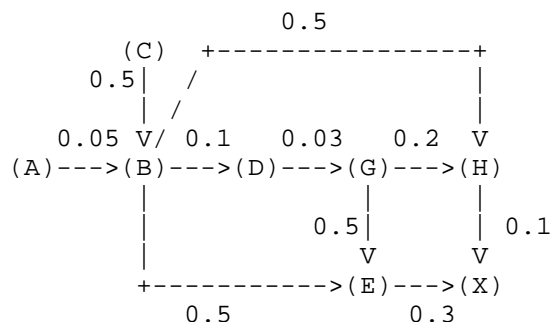
It is possible that multiple links would have the same PWRLINK metric after a computation cycle. In such a case load-balancing techniques can be used to keep the ECMP links in a steady state with respect to each other. Depending on the Available Bandwidth thereafter it is possible that the ECMP links may no longer be Equal cost but UCMP or Unequal Cost Paths.

2.3 Dampening the side effects of constant change

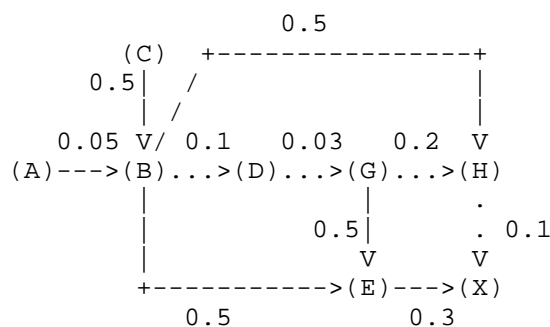
It is recommended in this draft that the implementation of the proposal be adaptive, infrequent in computation to the extent possible without sacrificing adapting to the dynamism and also reduce any frequent oscillations. The actual methods to adopt for this computation are outside the scope of this document.

2.4 Calculating power shortest paths in an Area

Assume the following topology where A,B,C etc.. are routers and corresponding labelled edges with weights are the links. These weights are the current values of the PWRLINK attribute that has been flooded in the LSAs through the Area concerned. Assume B is the ABR for Area 1 and the routers A and C are the Area 0 core routers. The rest of the routers are assumed to be in Area 1. Once the power topology of the Area 1 has been calculated as shown below with the PWRLINK attributes being assigned to the links, Constrained shortest path can be run from the ABR to any of the other routers say H, E , X etc.. The CSPF algorithm takes the constraint in terms of the PWRLINK attributes along with other attributes to construct a power shortest path from say router B to other routers in Area 1.



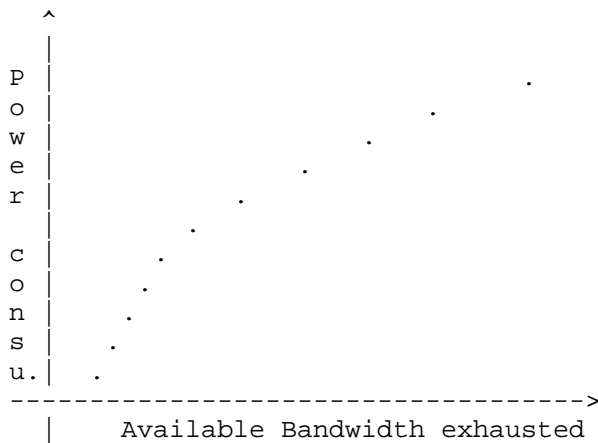
Once the path has been computed it is possible to use RSVP-TE to construct the power shortest path with the TE-LSP being instantiated with the labels appropriately placed in the routers on the power shortest path. In this topology, assume one would want to construct a path from B to X then the dotted path shows the path constructed and to be used by a set of flows or streams of packets belonging to multiple flows as seen fit by the router B. If the PWRLINK metrics change after due course of time then another power shortest path that possibly traverses the same path (if the SUM of PWRLINKs doesn't exceed any other path's metrics' SUM) or some other path would be constructed. Specifically this method makes use of traffic-engineering signalling protocols as the method to place the streams from point X to point Y (where X and Y are routers).



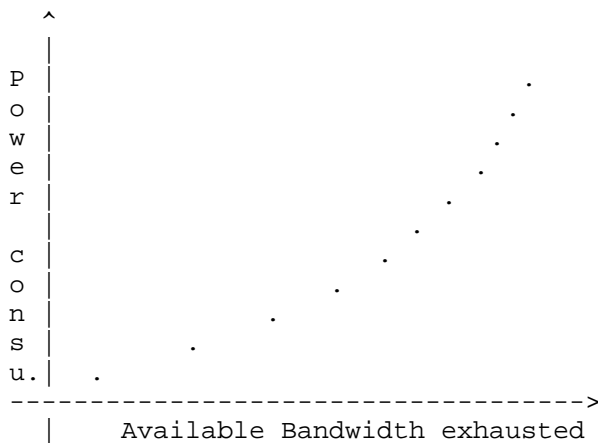
2.4.1 Power profiles of Routers and Switches

It has been experimented and from several sources found that there exist routers which have different power profiles. The power profile of a router is the curve of power consumption to available bandwidth. Mentioned below are a few of these prominent ones that have to be taken into consideration.

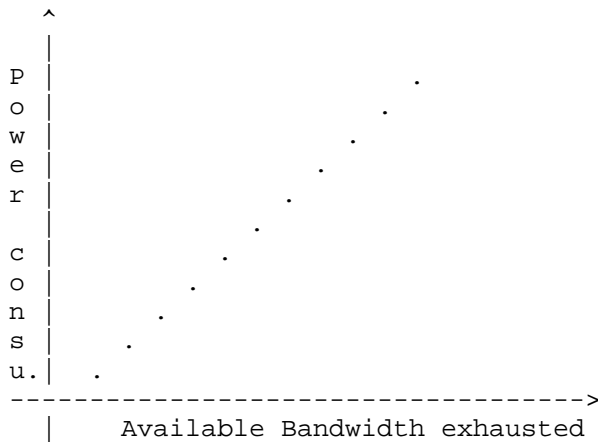
The first profile that we will consider is the flattening curve. The power consumed to available bandwidth curve takes the shape of a steep one initially and then tapers off to a plateau. The point at which it begins to give a delta-C (delta in Power Consumed) to delta-B (Available Bandwidth exhausted) is the inflection point that tapers off to a plateau. Here the delta-C/delta-B begins to slow down or decrease rapidly. The more the traffic that is added onto the device the lesser it draws power.



The second profile that we will consider is the exponential curve. The power consumed to available bandwidth curve takes the shape of an ever increasing steep curve as shown below. Here the $\Delta C / \Delta B$ begins to increase as more traffic is thrown onto it as the Available bandwidth exhausted increases. This power curve beyond a point is intolerable with respect to power guzzling.



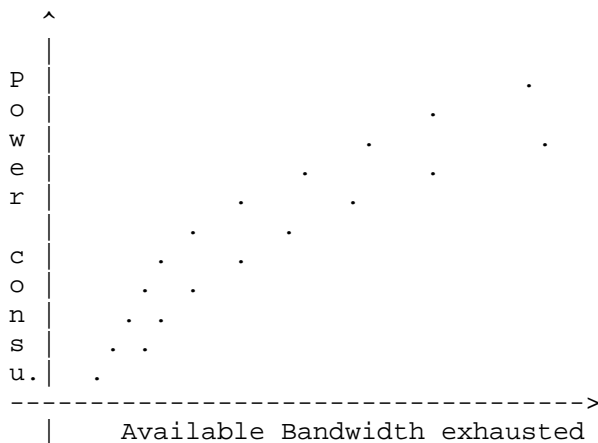
The third profile that we will consider is a linear curve. In other words just a straight line. Here $\Delta C / \Delta B$ is a constant.



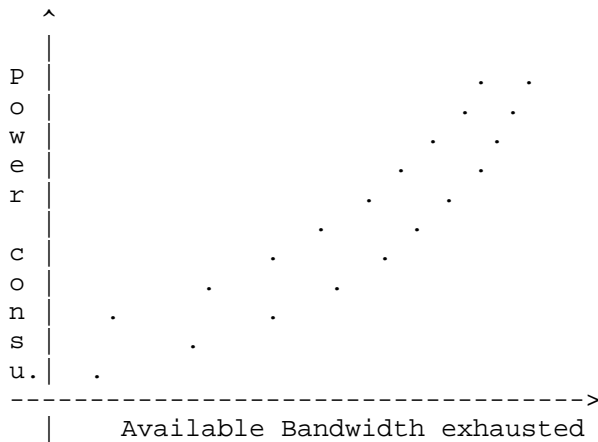
2.4.1.1 Concave and Convex power curves

Given that there are 3 kinds of major profiles in the router power consumption, what line would we like to pick. This is an important point when choosing the metric to pick the low power paths.

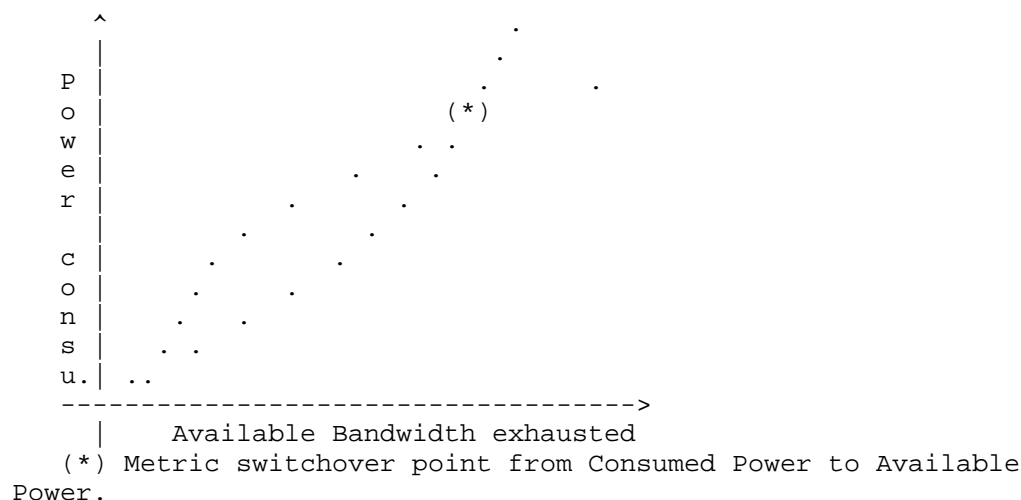
(a) If the confrontation is between 2 first profile routers the lower of the 2 would be considered as shown below. The lower curve offers better power savings for each GB of bandwidth transported.



(b) If the confrontation is between 2 second profile routers the upper curve offers more power savings per GB of bandwidth.



(c) When the confrontation is between a first profile curve and a second profile curve, it would be optimal to pick (as shown below) the lower of the curves because it gives us lesser power consumed for every GB of traffic routed / switched. Here the exponential curve is the one that offers lesser amount of power consumed per GB of traffic is chosen. But when it gets to a point that the two curves intersect it would be more optimal to pick the tapering curve. Thus at the meeting point of the 2 curves the exponential curve becomes more costly and the tapering one gives us more GB for the power buck. Thus this switchover from one curve to the other (in other words from the exponential curve to the tapering one) does the trick in terms of finding an optimal solution.



2.4.1.3 Need to advertise both available power and consumed power

Thus the above sections have shown that both the available power and the consumed power MUST be advertised so that case (c) can be deciphered and the switchover of the curves be done and the appropriate router be chosen for the rest of the bandwidth to be switched over to.

Thus there will exist Consumed-Power to Available Bandwidth ratio and the Available Power to Available Bandwidth ratio. Both the ratios are computed and the lower value chosen. The Available Power can be judged from the calibration process such as the one carried out by independent test organizations as in [12]. An example of their calibration is referred to in [12].

Here given below is the formula for calculating the Available Power to Available Bandwidth ratio also called the Available POWER metric.

$$\text{Available POWER metric} = \frac{\text{Available Power consumed per XG (normalized bandwidth) port}}{\text{Available Bandwidth on that port}}$$

for a given Port on a LC

2.4.2 Power to Available Bandwidth ratio in a TLV

As per [RFC3630] the Link TLV can be used to carry this power to available Bandwidth ratio with an additional sub-TLV of the link TLV. The sub-type number 11 is recommended to be defined for this purpose.

[RFC 3630] states in section 2.2.1 and we QUOTE ...

2.2.1 Link TLV

The Link TLV describes a single link. It is constructed of a set of sub-TLVs. There are no ordering requirements for the sub-TLVs.

Only one Link TLV shall be carried in each LSA, allowing for fine granularity changes in topology.

The Link TLV is type 2, and the length is variable.

The following sub-TLVs of the Link TLV are defined:

- 1 - Link type (1 octet)
- 2 - Link ID (4 octets)
- 3 - Local interface IP address (4 octets)
- 4 - Remote interface IP address (4 octets)
- 5 - Traffic engineering metric (4 octets)
- 6 - Maximum bandwidth (4 octets)
- 7 - Maximum reservable bandwidth (4 octets)
- 8 - Unreserved bandwidth (32 octets)
- 9 - Administrative group (4 octets)
- 10 - Power-to-Multicast-replication-capacity (4 octets)
- 11 - Consumed-Power-to-Available-Bandwidth (4 octets)
- 12 - Available-Power-to-Available-Bandwidth (4 octets)

This memo defines sub-Types 1 through 9. See the IANA Considerations in [RFC3630] section for allocation of new sub-Types.

The Link Type and Link ID sub-TLVs are mandatory, i.e., must appear exactly once. All other sub-TLVs defined here may occur at most once. These restrictions need not apply to future sub-TLVs. Unrecognized sub-TLVs are ignored.

Various values below use the (32 bit) IEEE Floating Point format. For quick reference, this format is as follows:

0	1	2	3
0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1
+-----+	+-----+	+-----+	+-----+
S	Exponent	Fraction	
+-----+	+-----+	+-----+	+-----+

S is the sign, Exponent is the exponent base 2 in "excess 127" notation, and Fraction is the mantissa - 1, with an implied binary point in front of it. Thus, the above represents the value:

$$(-1)**(S) * 2**(Exponent-127) * (1 + Fraction)$$

It is proposed that we use the Power-to-Available-Bandwidth ratio as a 32 bit IEEE floating Point format field for the purpose of this document.

3. Conclusion

Routers may have step levels in which they increase power consumption when they additively are loaded with more large bandwidth consuming multicast or unicast streams. Calibrating these levels may be useful for implementing this scheme. It is possible that such calibrated thresholds can be used for advertising the PWRLINK ratios in the OSPF LSA advertisements. This would be useful for bringing down the frequency of updates or advertisements from a line-card about its PWRLINK ratio. When power consumption meanders within a certain given interval these ratios need not be re-advertised even if further unicast and/or multicast streams are added to it. The incentive is to recognize a linecard that does not drastically change power consumption even if large bandwidth streams are added onto it for forwarding and thus give it credit for its power optimal functioning. If a router tends to consume the highest level of power even when carrying low amounts of unicast and multicast streams on its line card, it would automatically have a poor ratio when compared to a router that efficiently uses power when considering the Available Bandwidth being observed. The best case would be a low power consuming line-card or a router filled with such line cards that does not leave its power interval no matter how much ever capacity is sought to be used on it. But that would be an ideal condition but it is definitely an idealistic scenario towards which the router manufacturers should look at.

3 Security Considerations

<Security considerations text>

4 IANA Considerations

New requirements are required from IANA for a new type in the Link TLV in order to carry the PWRLINK metric as well. This is needed for both Consumed Power Ratio and Available Power Ratio.

5 References

5.1 Normative References

5.2 Informative References

- [1] G. Appenzeller, Sizing router buffers, Doctoral Thesis, Department of Electrical Engineering, Stanford University, 2005.
- [2] A. P. Bianzino, C. Chaudet, D. Rossi and J. L. Rougier, A survey of green networking research, IEEE Communications and Surveys Tutorials, preprint.
- [3] J. Baliga, K. Hinton and R. S. Tucker, Energy consumption of the internet, Proc. of joint international conference on optical internet, June 2007, pp. 1-3.
- [4] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang and S. Wright, Power awareness in network design and routing, Proc. of the IEEE INFOCOM 2008, April 2008, pp. 457-465.
- [5] B. Venkat et.al, Constructing disjoint and partially disjoint InterAS TE-LSPs, USPTO Patent 7751318, Cisco Systems, 2010.
- [6] M. Xia et. al., Greening the optical backbone network: A traffic engineering approach, IEEE ICC Proceedings, May 2010, pp. 1-5.
- [7] W. Lu and S. Sahni, Low-power TCAMs for very large forwarding tables, IEEE/ACM Transactions on Computer Networks, June 2010, vol. 18, no. 3, pp. 948-959.

[8] B. Zhang, Routing Area Open Meeting, Proceedings of the IETF 81, Quebec, Canada, July 2011.

[9] M.J.S Raman, V.Balaji Venkat, G.Raina, Reducing Power consumption using the Border Gateway Protocol, IARIA conferences ENERGY 2012.

[10] A.Cianfrani et al., An OSPF enhancement for energy saving in IP Networks, IEEE INFOCOM 2011 Workshop on Green Communications and Networking

Authors' Addresses

Shankar Raman
Department of Computer Science and Engineering,
IIT Madras
Chennai - 600036
TamilNadu
India.

EEmail: mjsraman@cse.iitm.ac.in

Balaji Venkat Venkataswami
Department of Electrical Engineering
IIT Madras
Chennai - 600036
TamilNadu
India.

EEmail: balajivenkat299@gmail.com

Prof.Gaurav Raina
Department of Electrical Engineering
IIT Madras
Chennai - 600036
TamilNadu
India.

EEmail: gaurav@ee.iitm.ac.in

Vasan Srini
Department of Computer Science and Engineering
IIT Madras
Chennai - 600036
TamilNadu
India.

EMail: vasan.vs@gmail.com

PANET Working Group
Internet-Draft
Intended Status: Experimental RFC
Expires: May 2013

Shankar Raman
Balaji Venkat Venkataswami
Gaurav Raina
I.I.T Madras
November 5, 2012

Constructing power optimal P2MP TE-LSPs within an AS
draft-mjsraman-panet-pce-power-mcast-replic-00

Abstract

Power consumption in multicast replication operations is an area of concern and choosing suitable replication points that can decrease power consumption overall assumes importance. Multicast replication capacity is an attribute of every line card of major routers and multi-layer switches that support multicast in the core of an Internet Service Provider (ISP) or an enterprise network.

Currently multicast replication points on Point-to-Multipoint Traffic Engineering Label-Switched-Paths (P2MP TE-LSPs) consume power while delivering multiple output streams of data from a given input stream. The multicast distribution trees are constructed without any regard for a proper placement of the replication points and consequent optimal power consumption at these points.

This results in overloading certain routers while under-utilizing others. An optimal usage of these replication resources could substantially reduce power consumption on these routers. In this paper, we propose a mechanism by which P2MP TE-LSPs are constructed for carrying multicast traffic across multiple areas within a given AS. We propose that these LSPs be built by using the advertisements of the power-replication capacity ratio advertised by fine grained components such as multicast capable line-cards of routers and multi-layer switches deployed within an AS.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	3
1.1	Terminology	3
2.	Methodology of the proposal	4
2.1	Discussion of this scheme	6
2.2	Power to available multicast replication capacity ratio in a TLV	8
3	Conclusion	10
3	Security Considerations	11
4	IANA Considerations	11
5	References	11
5.1	Normative References	11
5.2	Informative References	11
	Authors' Addresses	12

1 Introduction

Multicast traffic across multiple areas within a given AS, may be carried using P2MP TE-LSPs. The traffic may be carried from a ingress Provider Edge (PE) router to several egress PEs, example in a multicast Virtual Private Network (MVPN) case. The autonomous system (AS) may comprise of multiple areas involving a backbone area and several non-backbone areas connected to each other through the backbone. If several such multicast streams are to be carried in the AS, it would be most useful to have such P2MP TE-LSPs constructed such that they have optimal power to available replication capacity ratios on the routers' linecards that they traverse from source to destinations. The intent is to provide a solution whereby several such P2MP TE-LSPs can be laid out in such a way that the set of routers that replicate multicast traffic traversed by the P2MP TE-LSPs are most optimal in the utilization of the power provided to them given that there is sufficient replication capacity available. This we believe would essentially lead to a equilibrium of power to available replication capacity ratios amongst all routers in the topology which in turn would optimize and reduce the overall ratios for the AS.

Each router and its respective linecards deployed in the AS have an advertised capability for replication. Most multi-layer switches and routers from vendors advertise in their respective data sheets a certain capability for replication for each type of linecard deployable on the box. Replication consumes power and delivers multiple streams of data from a given input stream. It is status quo that P2MP (Point-to-Multipoint) Label Switched Paths are constructed without taking into account the power to available replication capacity ratios of such routers thus overloading certain routers while underutilizing the others. An optimal usage of these resources could reduce power consumption on these routers / multi-layer switches. This equilibrium could be arrived at by using a capability to advertise from each router a Traffic Engineering Database Link State Advertisement (TED-LSA) that carries the power to available replication capacity ratio of each of the said router's line cards, depending on the current utilization of its replication capacity and power consumption.

This paper is organized as follows; In section 2, we deal with the scheme that we propose. In section 2.1, we discuss some examples of the scheme at work, and in section 3 we conclude with future areas of study that may be useful to undertake.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT",

"SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Methodology of the proposal

The key metric under consideration is the power consumed DIVIDED BY available replication capacity on each of the linecards of a router in the AS, which is eligible to be used as a node atop which multicast traffic can be carried. Once an advertisement about the said metric has been sent in the regular flooding process in Link State routing protocols such as OSPF-TE or ISIS-TE, it would be possible for a head-end router for a P2MP TE-LSP to compute the TE-LSP through the AS from the ingress PE to all egress PEs of that multicast stream in such a way that the power to available replication capacity ratios at the replication points are minimal on that path. The Constrained Shortest Path First (CSPF) algorithm could be modified to compute the least cost power to available replication capacity ratio path and thus cause an equilibrium shift to be caused. This path would be supplied to the RSVP-TE component of the head-end and that would set up the path with appropriate labels. Once RSVP-TE establishes the path and traffic is carried across it, the reduced replication capacity of the routers in the P2MP TE-LSP path would be re-advertised again, which in turn would be useful for computation of the other paths from the instance that the replication capacity changed on these routers.

Assume that the following router topology in the vicinity of the sender / senders is computed.

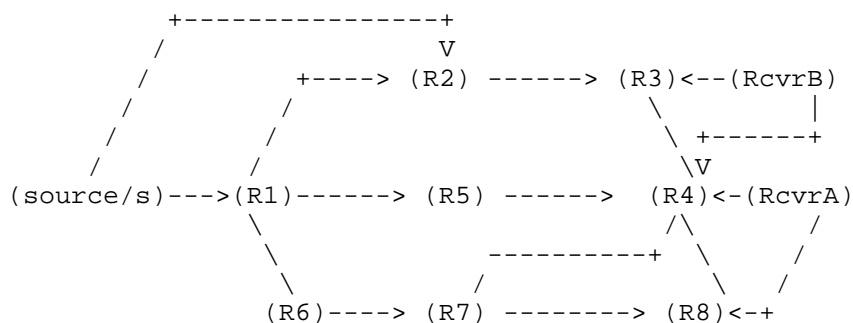
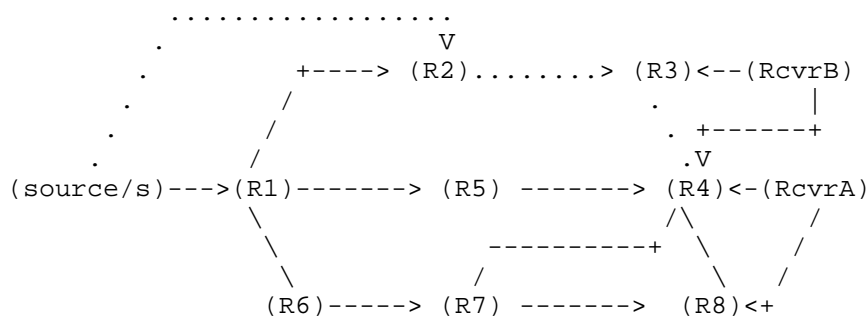


Figure 1: Topology within a given AS with coloring for Power-replication ratios

computed for these new streams would possibly utilize the same path as computed before. If the old streams reduce the replication capacity to an extent such that routers through which they pass can no longer be used since these routers' power to available replication capacity has become poor when compared to other paths then a different path may be computed from the ingress PE to the egress PEs in such a way as to avoid those routers which have such poor ratios.

For example, assume R6, R7, R8 and R4 have exhausted their capacity, or guzzle more power as a result of them carrying the 4GB stream that was originally placed atop them. then a different path would be chosen as follows. The path followed as shown in the Figure is R2,R3 and R4. Given that R4 is the only choice since it has connectivity to both Receivers, in this case the branch point is placed atop R3, one branch to get to RcvrB and the other to get to RcvrA through R4. Policy decisions could guide the placement in case of a tie. Here the the only choice has been to drive the end replication to RcvrA through R4 and RcvrB through R3 owing to topology constraints.

It is to be noted that the power consumed by the linecard is divided by the available replication capacity to arrive at a ratio and that ratio is assigned as a weight to all of the links ingressing on that linecard. It is possible that one might take a weighted average by dividing a weighted co-efficient sum by the weighted sum of ingress links on a linecard and the metrics so assigned be used as the metric for calculation.



Legend : dotted lines represent path computed.

Figure 3: Instantiating a subsequent optimal power consuming distribution tree

2.1 Discussion of this scheme

It is to be noted that our scheme applies to centralized schemes of path calculations. What is being calculated is a tree of nodes that

form a P2MP tree where each node can be conceptualized as a router (read also multi-layer switches) and each edge the link connecting one or more ports on a line card to another linecard on a downstream router to carry multicast traffic from a source located at the head end ingress router to several receiver nodes connected to egress routers. We will call this calculated tree as a P2MP tree. The tree is calculated by the PCE in the head end / ingress router through which sources connect. The PCE calculates the intra-AS P2MP path (the literal P2MP TE-LSP within the AS) within that AS.

The calculated power to available replication capacity ratio is assigned to each of the ingress links on a linecard on a router en-route to egress links through which the multicast stream is replicated on the same router. Thus all ingress links to a router through a linecard are assigned the same metric as the power ratio so calculated. The egress links would in continuity connect to a unicast tunnel or another branch-point in the tunnel towards the receivers which are represented as the egress routers. The egress routers would in turn be replication points or direct connections to the actual receivers. This method could be applied for multicast traffic to be transported through MVPNs. The method of egress routers' discovery is left to existing mechanisms. The primary input to the invention proposed is an ingress router and their respective egress routers. The other input to the construction of P2MP tree is the router level topology with the metrics for the power to available replication capacity ratio.

It is to be noted that this CSPF calculation can be hastened in terms of time complexity by dividing the weights into equivalence classes. First we divide the nodes into graph colored nodes with the least ratio nodes marked as green as shown in the figure and given that there exists a path that is all green from source to egress PEs, one of such paths is chosen. If after coloring the nodes a path which is disconnected exists, we incrementally add the next best colored nodes to the graph to see if we get a connected path from source to egresses. These steps are repeated until we find a connected path. This will hasten the algorithm to a conclusion rather than use a brute force method which may take inordinate amount of time. R4 being used in the 6GB case is an example of this. Because of topology restrictions the R4 node had to be chosen in spite of the fact that it is not green after carrying the 4GB stream.

Routers may have step levels in which they increase power consumption when they additively are loaded with more large bandwidth consuming multicast streams. Calibrating these levels may be useful for implementing this scheme. It is possible that such calibrated thresholds can be used for advertising the power to available replication capacity ratios in the IGP-TE advertisements. This would

be useful for bringing down the frequency of updates or advertisements from a line-card about its ratios. When power consumption meanders within a certain given interval these ratios need not be readvertised even if further multicast streams are added to it. The incentive is to recognize a linecard that does not drastically change power consumption even if large bandwidth streams are added onto it for replication and thus give it credit for its power optimal functioning. If a router tends to consume the highest level of power even when carrying low amounts of multicast streams and replicating them on its line card, it would automatically have a poor ratio when compared to a router that efficiently uses power when considering the replication capacity being used. The best case would be a low power consuming line-card or a router filled with such line cards that does not leave its power interval no matter how much ever replication capacity is sought to be used on it. But that would be an ideal condition but it is definitely an idealistic scenario towards which the router manufacturers should look at.

It is possible that several multicast streams may be aggregated onto a single P2MP-TE-LSP representing the given multicast tree that encompasses the union of all the egress PEs of the several multicast streams. The Ingress PE router is however common for all the multicast streams so covered. Aggregation of these several multicast streams from a given Ingress PE to several egress PEs is a common occurrence to save the amount of state in the core of the network. By aggregating these streams onto a single P2MP tree, it is possible to amortize the cost of replication amongst a particular set of ingress linecards / ports on those line cards while taking into account the current power consumption and replication capacity available at the time of computing the P2MP TE-LSP.

The dynamic nature of the multicast tree and the egress PEs that join into it and leave it based on whether there are multicast listeners in that VPN site attached to the said egress PE/ PEs, makes it important to position the replication points in such a way that there is maximum leverage on optimization in the ratios overall for the AS which are computed. When aggregating multiple multicast streams over a single P2MP TE-LSP it is important to keep this in mind.

So the key point is to aggregate multiple streams with a set theoretical approach in mind so that there is maximum overlap of egress PEs for these streams and position these streams atop a P2MP TE-LSP in such a way that ratios are most optimal for that set of streams (with the overall AS power consumption in mind).

2.2 Power to available multicast replication capacity ratio in a TLV

As per [RFC3630] the Link TLV can be used to carry this power to

available multicast replication capacity ratio with an additional sub-TLV of the link TLV. The sub-type number 10 is recommended to be defined for this purpose.

[RFC 3630] states in section 2.2.1 and we QUOTE ...

2.2.1 Link TLV

The Link TLV describes a single link. It is constructed of a set of sub-TLVs. There are no ordering requirements for the sub-TLVs.

Only one Link TLV shall be carried in each LSA, allowing for fine granularity changes in topology.

The Link TLV is type 2, and the length is variable.

The following sub-TLVs of the Link TLV are defined:

- 1 - Link type (1 octet)
- 2 - Link ID (4 octets)
- 3 - Local interface IP address (4 octets)
- 4 - Remote interface IP address (4 octets)
- 5 - Traffic engineering metric (4 octets)
- 6 - Maximum bandwidth (4 octets)
- 7 - Maximum reservable bandwidth (4 octets)
- 8 - Unreserved bandwidth (32 octets)
- 9 - Administrative group (4 octets)
- 10 - Power-to-Multicast-replication-capacity (4 octets)

This memo defines sub-Types 1 through 9. See the IANA Considerations in [RFC3630] section for allocation of new sub-Types.

The Link Type and Link ID sub-TLVs are mandatory, i.e., must appear exactly once. All other sub-TLVs defined here may occur at most once. These restrictions need not apply to future sub-TLVs. Unrecognized sub-TLVs are ignored.

Various values below use the (32 bit) IEEE Floating Point format. For quick reference, this format is as follows:

0	1	2	3
0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9	0 1
+--+--+--+--+	+--+--+--+--+	+--+--+--+--+	+--+--+--+--+
S	Exponent		Fraction
+--+--+--+--+	+--+--+--+--+	+--+--+--+--+	+--+--+--+--+

S is the sign, Exponent is the exponent base 2 in "excess 127" notation, and Fraction is the mantissa - 1, with an implied binary

point in front of it. Thus, the above represents the value:

$$(-1)^{(S)} * 2^{(Exponent-127)} * (1 + Fraction)$$

It is proposed that we use the Power-to-multicast-replication-capacity ratio as a 32 bit IEEE floating Point format field for the purpose of this document.

3 Conclusion

Here we propose a scheme that takes into account the power to available replication capacity ratios as weights for the edges and compute a low cost power path for multicast replication. This scheme could be extended to inter-AS multicast streams or to inter-AS multicast streams where the multicast stream is sought to be carried over multiple ASes. This is an area of future study which would be most conducive in terms of bringing about optimal power usage and thus incentivising vendors to manufacture low power consuming equipment. Compelled to bring about radical change in the thinking relating to power consumption vendors manufacturing networking equipment will drive down power consumption since the scheme proposed chooses or gives priority to low power guzzling linecards.

3 Security Considerations

The security considerations for this proposal are the same as any NEW opaque LSA introduced in an IGP like OSPF, IS-IS.

4 IANA Considerations

IANA would need to assign a NEW opaque LSA type to carry power and multicast replication capacity such that this information can be carried in the TE-LSAs within an AS.

5 References

5.1 Normative References

5.2 Informative References

- [1] G. Appenzeller, Sizing router buffers, Doctoral Thesis, Department of Electrical Engineering, Stanford University, 2005.
- [2] A. P. Bianzino, C. Chaudet, D. Rossi and J. L. Rougier, A survey of green networking research, IEEE Communications and Surveys Tutorials, preprint.
- [3] J. Baliga, K. Hinton and R. S. Tucker, Energy consumption of the internet, Proc. of joint international conference on optical internet, June 2007, pp. 1993.
- [4] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsang and S. Wright, Power awareness in network design and routing, Proc. of the IEEE INFOCOM 2008, April 2008, pp. 457-465.
- [5] M. Xia et. al., Greening the optical backbone network: A traffic engineering approach, IEEE ICC Proceedings, May 2010, pp. 1995.
- [6] W. Lu and S. Sahni, Low-power TCAMs for very large forwarding tables, IEEE/ACM Transactions on Computer Networks, June 2010, vol. 18, no. 3, pp. 948-959.
- [7] B. Zhang, Routing Area Open Meeting, Proceedings of the IETF 81, Quebec, Canada, July 2011.

Authors' Addresses

Shankar Raman
Department of Computer Science and Engineering
I.I.T Madras,
Chennai - 600036
TamilNadu,
India.

EMail: mjsraman@cse.iitm.ac.in

Balaji Venkat Venkataswami
Department of Electrical Engineering,
I.I.T Madras,
Chennai - 600036,
TamilNadu,
India.

EMail: balajivenkat299@gmail.com

Prof.Gaurav Raina
Department of Electrical Engineering,
I.I.T Madras,
Chennai - 600036,
TamilNadu,
India.

EMail: gaurav@ee.iitm.ac.in

PANET Working Group
Internet-Draft
Intended Status: Experimental RFC
Expires: May 2013

Shankar Raman
Balaji Venkat Venkataswami
Gaurav Raina
Vasan Srini
I.I.T Madras
November 5, 2012

Building power optimal Multicast Trees
draft-mjsraman-panet-pim-power-00

Abstract

Power consumption in multicast replication operations is an area of concern and choosing suitable replication points that can decrease power consumption overall assumes importance. Multicast replication capacity is an attribute of every line card of major routers and multi-layer switches that support multicast in the core of an Internet Service Provider (ISP) or an enterprise network.

Currently multicast replication points on Point-to-Multipoint Multicast Distribution trees consume power while delivering multiple output streams of data from a given input stream. The multicast distribution trees are constructed without any regard for a proper placement of the replication points and consequent optimal power consumption at these points.

This results in overloading certain routers while under-utilizing others. An optimal usage of these replication resources could reduce power consumption on these routers bringing power consumption to optimality. In this paper, we propose a mechanism by which Multicast Distribution Trees are constructed for carrying multicast traffic across multiple routers within a given network. We propose that these Multicast Distribution Trees be built by using the information pertaining to power-replication capacity ratio available with fine grained components such as multicast capable line-cards of routers and multi-layer switches deployed within a network.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as

Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	3
1.1	Terminology	4
2.	Methodology of the proposal	4
2.1	Discussion of this scheme	7
2.2	Pseudo code for the proposed changes	8
2.3	Port Choice on same Linecard	8
3	Conclusion	8
3	Security Considerations	10
4	IANA Considerations	10
5	References	10
5.1	Normative References	10
5.2	Informative References	10
	Authors' Addresses	11

1 Introduction

Multicast traffic across multiple areas within a given network such as an ISP or a Campus Environment Network, may be carried using Multicast Distribution Trees. The traffic may be carried from a ingress router to several egress routers, example in a Campus Environment network. The Network under consideration may comprise of multiple areas involving a backbone area and several non-backbone areas connected to each other through the backbone. If several such multicast streams are to be carried in the network, it would be most useful to have such Multicast Distribution Trees constructed such that they have optimal power to available replication capacity ratios on the routers' linecards that they traverse from source to destinations. The intent is to provide a solution whereby several such Distribution Trees can be laid out in such a way that the set of routers that replicate multicast traffic traversed by the trees are most optimal in the utilization of the power provided to them given that there is sufficient replication capacity available. This we believe would essentially lead to a equilibrium of power to available replication capacity ratios amongst all routers in the topology which in turn would optimize and reduce the overall ratios for the network.

Each router and its respective linecards deployed in the network have an advertised capability for replication. Most multi-layer switches and routers from vendors advertise in their respective data sheets a certain capability for replication for each type of linecard deployable on the box. Replication consumes power and delivers multiple streams of data from a given input stream. It is status quo that (Point-to-Multipoint) P2MP trees are constructed without taking into account the power to available replication capacity ratios of such routers thus overloading certain routers while underutilizing the others. An optimal usage of these resources could reduce power consumption on these routers / multi-layer switches. This equilibrium could be arrived at by using a capability to choose from each downstream PIM router the most power optimal path to the selected (through current mechanisms) PIM upstream neighbor in the PIM-based Multicast Distribution Tree which may be a shared tree or a Shortest Path Tree as the case may be. The metric used to select the upstream PIM neighbor could be the power to available replication capacity ratio of each of the said router's line cards that are part of the ECMP set of paths to the upstream neighbor if such ECMP paths do exist. The metric comparison is done for all ECMP paths and the line cards involved therein depending on their current utilization of their replication capacity and power consumption.

This paper is organized as follows; In section 2, we deal with the scheme that we propose. In section 2.1, we discuss some examples of the scheme at work, and in section 3 we conclude with further areas

of study that may be useful to undertake.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Methodology of the proposal

The key metric under consideration is the power consumed DIVIDED BY available replication capacity on each of the linecards of a router in the network whose constituent ports form part of a ECMP set of paths to a PIM upstream neighbor. The said ports on the different line cards that form the ECMP set of links are eligible to be used as a linecard:port atop which multicast traffic on that tree can be carried. When choosing the path from a ECMP set of paths to a PIM upstream neighbor, the said downstream PIM neighbor calculates the power to multicast replication capacity ratio for each of the line cards that are eligible to be chosen as the linecard:port combination to be used in that section of the distribution tree. The lowest ratio decides which linecard is chosen and if there exist multiple ports within that linecard that connect to the said PIM upstream neighbor the usual algorithm is used to select one of those ports. The key proposal that this document recommends is the use of the power-multicast-replication-capacity ratio to choose from among the different linecards. The choice of port is left to the standard method.

Assume that the following router topology in the vicinity of the sender / senders is computed.

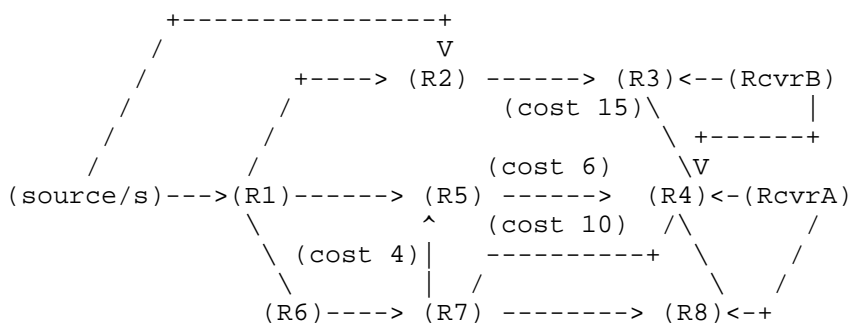
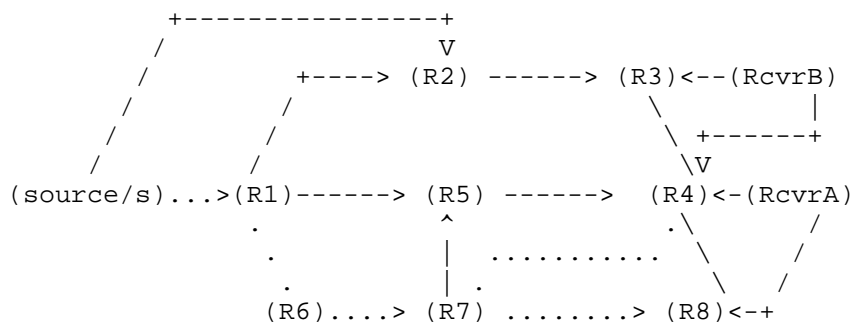


Figure 1: Topology within a given network with an upstream ECMP link from R4 to R7

In the above diagram you can see that the source/sources are connected using a multi homed connections to the same ISP through Routers R1 and R2. Similarly there are two Receiver sites RcvrA and RcvrB that are multihomed to TWO Routers RcvrB to R3 and R4 and for RcvrA to R4 and R8 respectively. You can also observe that R4 is connected to R7 through multiple paths. Assuming that both these paths are Equal Cost then this gives rise to a situation where ECMP paths exist for the PIM downstream router R4 to the PIM upstream router R7.

Consider that RcvrA sends an IGMP join to R4. R4 now needs to send a PIM join towards the upstream router R7. Assume this is a shared tree with Rendezvous Point (RP) as R7. There are 2 equal cost paths to R7 from R4 each with cost 10 ((R7->R5->R4 = 6 + 4 = 10) and (R7 -> R4 = 10)). Assume that each of these paths from R4 to R5 onto R7 and from R4 to R7 directly are on different linecards in the chassis R4. Normally one of them would be chosen and power-to-multicast-replication-capacity would not be a consideration in that decision. What this document proposes is that R4 consider the metric PWR which is a ratio formed by dividing the power consumed on each of the linecards by their respective current multicast replication capacity.

Obviously one of them would have to be chosen. In the metric comparison the linecard that has the lower PWR metric wins and is selected for consideration to send a PIM join to R7 (the PIM upstream neighbor and in this case the RP as well).



Legend : dotted lines represent path computed.

Figure 2: Instantiating an optimal power consuming distribution tree

In our example as in Figure 2 we find that the direct link to R4 and R7 wins out as the link to be used in the distribution tree.

The one exception that SHOULD be considered in this decision is that if the Outgoing Interface List consists of ports on linecard X on

which R4's downstream PIM neighbors have sent their respective PIM joins and if the ECMP set of paths to the router R7 consist of linecard X and Y, it would be preferable to choose linecard X without taking into consideration the PWR metric. This is in light of the fact that if majority of the OIF list's port members lie on linecard X and the ingress port were also to be placed on linecard X then the replication would be more optimal as it would not have to traverse say the switch fabric to get to the majority of the OIF list. Other localization conditions could also be considered as exceptions to the PWR metric based rule.

This document assumes that the power used by each linecard and the multicast replication utilization and advertised capacity are available as data readable from the hardware on the router chassis under consideration. Please note that unicast traffic already being carried on the linecard may also contribute to the power being consumed at the router's linecards under consideration.

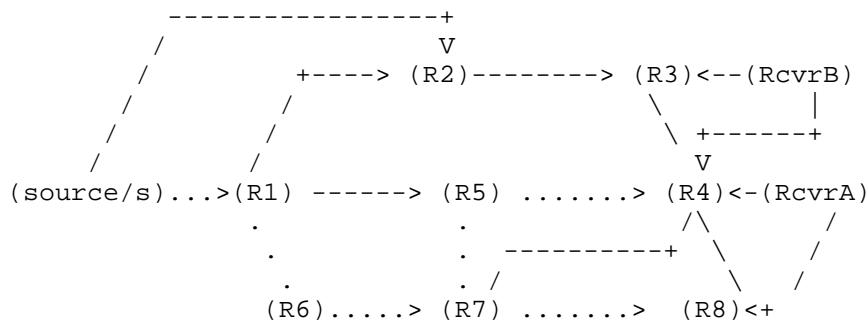
If ECMP paths dont exist then there is no choice to make hence the default selection of the link to be used to send a PIM join to the upstream neighbor is followed.

As a result of this decision to include the PWR metric the paths in the tree where ECMP links occur have the least power to available replication capacity ratios at the time of computation.

Assume the following path is computed as per the least power to available replication capacity ratios. Paths are computed through R6, R7, R8, R4, and say the multicast stream occupies 4GB of traffic along this tree so constructed and the available capacity of these routers reduces to 6GB assuming all of them have a base capacity of 10GB. Subsequent paths constructed would have to take into account the newly computed power to current replication capacity ratio in the topology for multicast streams / trees yet to come. Now the linecard connecting R4 to R7 directly will have reduction of a quantum of 4GB capacity. It would reduce to 6GB as its available capacity.

Assume another 6GB worth of traffic is loaded onto this topology in terms of a multicast stream / multiple streams then the new path computed for these new streams would NOT possibly utilize the same path as computed before since the power utilization and the available replication capacity would have been changed to create a higher PWR ratio. If the old streams reduce the replication capacity to an extent such that routers through which they pass can no longer be used since these routers' power to available replication capacity has become poor when compared to other paths then a different path may be computed from the ingress router to the egress router in such a way as to avoid those routers which have such poor ratios. This again

applies only in ECMP sections of the distribution tree.



Legend : dotted lines represent path computed.

Figure 3: Instantiating a subsequent optimal power consuming distribution tree

Here R4 would now have to choose the path to R7 (which is also the RP) through R5 since the PWR metric on R4 to R7 direct link would have increased as a result of carrying the old stream.

Dynamism in multicast trees is another important point to consider as PIM-Prunes and other PIM-joins may happen with respect to the replication point under consideration. Suitable modifications to the algorithm may be proposed to take into consideration such dynamic conditions without causing major interruption to the multicast flows.

2.1 Discussion of this scheme

This scheme applies to PIM-SM, PIM-SSM. Applicability to PIM-Bidir is also possible but currently not discussed in this document in detail.

Routers may have step levels in which they increase power consumption when they additively are loaded with more large bandwidth consuming multicast streams. Calibrating these levels may be useful for implementing this scheme. It is possible that such calibrated thresholds can be used for calculating the power to available replication capacity ratios in the Multicast environments. This would be useful for bringing down the frequency of calculations on a line-card about its ratios. When power consumption meanders within a certain given interval these ratios need not be calculated even if further multicast streams are added to it. The incentive is to recognize a linecard that does not drastically change power consumption even if large bandwidth streams are added onto it for replication and thus give it credit for its power optimal

functioning. If a linecard on a router tends to consume the highest level of power even when carrying low amounts of multicast streams and replicating them on its line card, it would automatically have a poor ratio when compared to a linecard that efficiently uses power when considering the replication capacity being used. The best case would be a low power consuming line-card or a router filled with such line cards that does not leave its power interval no matter how much ever replication capacity is sought to be used on it. But that would be an ideal condition but it is definitely an idealistic scenario towards which the router manufacturers should look at.

2.2 Pseudo code for the proposed changes

```

If (there exist ECMP paths to a PIM upstream NBR)
    AND (No localized conditions exist)
then
    Calculate PWR ratio for each LC;
    PWR per LC = power consumed by LC /
                AvailableMCastReplicCap;
    Choose the Lowest PWR;
    Select that LC for the link to send PIM Join;
Endif

```

2.3 Port Choice on same Linecard

In case in the set of ECMP links to the upstream PIM NBR there exist ports from the same line card and there is a tie breaking mechanism required amongst these ports the following changes are recommended.

```

If (there exist ports on the same linecard which
    constitute ECMP paths to a PIM upstream NBR)
    AND (No localized conditions exist)
then
    Choose the Lowest Utilized port;
    Select that port in LC for the link to send PIM Join;
Endif

```

3 Conclusion

Here we propose a scheme that takes into account the power to available replication capacity ratios as weights for the edges which are the ECMP set of paths to a PIM upstream neighbor and compute a low cost power path for multicast replication. This is an area of future study which would be most conducive in terms of bringing about optimal power usage and thus incentivising vendors to manufacture low power consuming equipment. Compelled to bring about radical change in the thinking relating to power consumption vendors manufacturing

networking equipment will drive down power consumption since the scheme proposed chooses or gives priority to low power guzzling linecards.

3 Security Considerations

None.

4 IANA Considerations

None.

5 References

5.1 Normative References

5.2 Informative References

- [1] G. Appenzeller, Sizing router buffers, Doctoral Thesis, Department of Electrical Engineering, Stanford University, 2005.
- [2] A. P. Bianzino, C. Chaudet, D. Rossi and J. L. Rougier, A survey of green networking research, IEEE Communications and Surveys Tutorials, preprint.
- [3] J. Baliga, K. Hinton and R. S. Tucker, Energy consumption of the internet, Proc. of joint international conference on optical internet, June 2007, pp. 1993.
- [4] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsang and S. Wright, Power awareness in network design and routing, Proc. of the IEEE INFOCOM 2008, April 2008, pp. 457-465.
- [5] M. Xia et. al., Greening the optical backbone network: A traffic engineering approach, IEEE ICC Proceedings, May 2010, pp. 1995.
- [6] W. Lu and S. Sahni, Low-power TCAMs for very large forwarding tables, IEEE/ACM Transactions on Computer Networks, June 2010, vol. 18, no. 3, pp. 948-959.
- [7] B. Zhang, Routing Area Open Meeting, Proceedings of the IETF 81, Quebec, Canada, July 2011.

Authors' Addresses

Shankar Raman
Department of Computer Science and Engineering
I.I.T Madras,
Chennai - 600036
TamilNadu,
India.

EMail: mjsraman@cse.iitm.ac.in

Balaji Venkat Venkataswami
Department of Electrical Engineering,
I.I.T Madras,
Chennai - 600036,
TamilNadu,
India.

EMail: balajivenkat299@gmail.com

Prof.Gaurav Raina
Department of Electrical Engineering,
I.I.T Madras,
Chennai - 600036,
TamilNadu,
India.

EMail: gaurav@ee.iitm.ac.in

Vasan Srini,
Department of Electrical Engineering,
I.I.T Madras,
Chennai - 600036,
TamilNadu,
India.

Email: vasan.vs@gmail.com

PANET Working Group
INTERNET-DRAFT
Intended Status: Experimental RFC

Shankar Raman
Balaji Venkat Venkataswami
Kamakoti Veezhinathan
Gaurav Raina
IIT Madras
May 4, 2013

Expires: November 5, 2013

TCAM power reduction and optimization in Routers
draft-mjsraman-panet-tcam-power-efficiency-02

Abstract

This specification entails enabling power and performance management in Routers (multi-chassis and single chassis) with respect to TCAM and SRAM usage on intelligent router line cards that implement Virtual Aggregation of routes.

The version 00 of this document had several errata that have been addressed in this version. This scheme relates to unicast alone since multicast entries are programmed only after consultation with the unicast entries in the software plane in the Route processor. Hence this scheme does not apply to multicast.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	3
1.1	Terminology	4
2.	Methodology	4
2.1	Packet flow and its consequences	6
2.2	Switching off the TCAM banks which are unused	7
2.2.1	Algorithm for switching off and switching on TCAM banks	7
2.2.2	Prefix entry population latency	10
2.3	Using Aggregate routes	10
2.4	Route Processor as the DLC/BDLC	10
2.5	Advantages of this scheme	10
3	Security Considerations	11
4	IANA Considerations	11
5	References	11
5.1	Normative References	11
5.2	Informative References	11
	Authors' Addresses	11

1 Introduction

This specification / draft entails enabling power and performance management in Routers (multi-chassis and single chassis) with respect to TCAM / SRAM usage on intelligent router line cards that implement Virtual Aggregation.

Distributed line cards in a routers have intelligence to forward packet without interference from the central control plane, once their TCAMs and associated SRAMs are populated. Since memory and TCAM are the most power hungry components in these line cards, we should effectively manage the line card power usage by optimally using their SRAM memory and TCAM banks.

When a packet enters a distributed line card, the packet switching logic extracts information from the header. It then looks up the entry in the appropriate TCAM (CAM in L2 switches or both in Multi-layer switches), and then passes the packet to the outgoing line card. The packet is then placed onto the outgoing port. The TCAM banks used in the line cards typically carry the entire forwarding table as tabulated by the central control processor card/cards (when in plurality used for redundancy). All line cards do not need the entire forwarding table as each line card may serve a set of sources and destinations. This is true especially in smaller networks but may also apply to routers in the Internet, especially ASBRs and POP border routers facing the customers of the ISP. Switching on the entire set of TCAMs (in the worst case) and downloading the entire forwarding table atop each of the line cards in the chassis leads to a sub-optimal way of switching packets

Current status quo on this (apart from VPN route localization) is a proof that as far as Internet destinations go, all line cards on the backbone routers or even within a small campus or a medium sized ISP, carry the entire forwarding table built by the routing table manager on these routers. In order to obviate the necessity of carrying routes which are unused (by this term, we mean those that are unreferenced for making forwarding decisions) and to reduce TCAM space (applicable to switching as well which involves CAMs), we suggest a solution where a couple of Linecards are considered to be Designated and Backup Designated linecards.

Designated Linecards are filled with all the entries from the control plane. The rest of the Linecards are provided with entries leftover from aging other entries which were not used over a period of time. An entry may not be available in these Linecards after they have been aged out (because of not being referenced and/or modified), these non- DLC linecards (as they are called), refer to the DLC/BDLC linecards for the first packet of a flow and retrieve the prefixes

associated with the hit on the DLC/BDLC. They populate these retrieved entries in their respective TCAM banks. Periodically on the non-DLC linecards a Route aggregation similar to the S-VA or the simple Virtual Aggregate with exception entries is generated to further reduce the TCAM occupation. The TCAM banks which are not occupied as a result of this scheme may be switched off. This in turn can switch off their associated SRAM banks as well which contain the rewrite information. The document also opens up the specification to simulations which help strengthen the argument for this scheme.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Methodology

Distributed line cards in a routers have intelligence to forward packet without interference from the central control plane, once their TCAMs and associated SRAMs are populated.

Since memory and TCAM are the most power hungry components in these line cards, we should effectively manage the line card power usage by optimally using their SRAM memory and TCAM banks.

When a packet enters a distributedline card, the packet switching logic extracts information from the header. It then looks up the entry in the appropriate TCAM (CAM in L2 switches or both in Multi-layer switches), and then passes the packet to the outgoing line card. It is then placed onto the outgoing port. The TCAM banks used in the line cards typically carry the entire forwarding table as tabulated by the central control processor card/cards (when in plurality used for redundancy). All line cards do not need the entire forwarding table as each line card may serve a set of sources and destinations. This is true especially in smaller networks but may also apply to routers in the internet, especially ASBRs and POP border routers facing the customers of the ISP. Switching on the entire set of TCAMs (in the worst case) and downloading the entire forwarding table atop each of the line cards in the chassis leads to a sub-optimal way of switching packets.

Current status quo on this (apart from VPN route localization) is a proof that as far as internet destinations go, all line cards on the backbone routers or even within a small campus or a medium sized ISP, carry the entire forwarding table built by the routing table manager on these routers. In order to obviate the necessity of

carrying routes which are unused (by this term. we mean those that are unreferenced for making forwarding decisions) and to reduce TCAM space (applicable to switching as well which involves CAMs), we suggest the following solution:

a) At initialization time the entire forwarding table that is constructed in the control processor is downloaded to all line cards. This is the current implementation in most routers today.

b) Begin a clock algorithm on these routes using a referenced bit and modified bit that is appended to each TCAM entry in the line card.

a. When a TCAM entry is accessed for forwarding decision in the router on a specific line card, the referenced bit associated with that entry is set.

b. When a TCAM entry is modified as a result of a routing entry being modified either with respect to its next-hop or any other reason the modified bit is set.

c. A timer called the ReferenceTimer is started with a suitable interval.

d. When the ReferenceTimer clock ticks down to 0, the TCAM entries in the line card are scanned and those that have the following values are not disturbed.

1. Referenced bit = 1, Modified bit = 1
2. Referenced bit = 0, Modified bit = 1

e. A timer called ModifiedDecayTimer is also started using a suitable time interval.

f. The ModifiedDecayTimer is a single global timer which is started whenever a route change that modifies the nexthop to a set of route entries is downloaded to the line card (which is the status quo as of now). Each line card has its own respective ModifiedDecayTimer.

g. On expiry of the ModifiedDecayTimer the TCAM entries with modified bit = 1 are set to modified bit = 0.

h. Similarly each TCAM entry has a 32 bit counter that counts down to 0, which is in effect a ReferencedDecayTimer. A suitable value may be appended to this counter at the time of initialization and when the TCAM entry is referenced for forwarding. This counter is active only for the active banks of TCAM. A suitable alternative with or without the counter would be an LRU policy that removes and adds entries as appropriate.

i. When the ReferencedDecayTimer counts down to 0, the referenced bit for that TCAM entry is cleared.

j. Continuing from (d) the TCAM entries with Modified bit = 0 and Referenced bit = 0 are removed from the TCAM when the ReferencedTimer expires. Further optimization on the TCAM may be done at regular intervals whenever the ReferencedTimer expires. This would involve re-arranging the TCAM to optimize on storage. The specific method relates to understanding the spread of TCAM entries with different mask lengths. Usually the TCAM entries are arranged in descending order of mask lengths with enough gaps to accommodate for different prefixes in their respective mask lengths. The algorithms in current software for populating the TCAM entries with enough gaps to accommodate for new additions are many. These however are out of scope of this document. Such algorithms allow for a fair amount of spread of the TCAM entries across the TCAM space thus allowing for several banks of TCAMS to be left empty if not used.

2.1 Packet flow and its consequences

We will consider the initial condition on how these entries get populated. If a packet enters a line card on one of its ports and the switching logic finds that the TCAM entry is absent or has been cleared from its entry in the TCAM. We use a mechanism by which such a packet is first forwarded to one of the DLC (or designated Line cards which may be elected from the set of line cards in the chassis) where the entire forwarding table is always stored. There may be a Primary DLC and a backup DLC for redundancy purposes. When the packet hits the ingress line card which does not have the required TCAM entry for forwarding, it routes the packet within the chassis to the DLC. The DLC receives the packet or just the packet header and does the following:

- i) Looks up its full forwarding table,
- ii) Picks up the result and along with it the TCAM entry or entries which accumulate to all the set of related routes (perhaps all the routes with the same prefix that is rounded off to the nearest major class network boundary) and
- iii) sends them across to the ingress line card in question.

The ingress line card uses the result to forward the packet and populates the TCAM with the group of entries dispatched to it by the DLC or the Backup DLC. One could use per-destination load-balancing within the ingress line card to distribute the lookup in such a way that the load is effectively shared by picking the DLC OR the Backup DLC. Once the set of routes sent back from the DLC is accumulated

and loaded into the TCAM, it may be periodically scanned and compressed even at a later point in time using the S-VA mechanism as explained in later sections.

Now the ingress line card has the required entries. It is thus made possible that the flows that go to the desired destinations from then on would hit the TCAM entries that are populated according to the method discussed above.

Thus, long lived TCP / UDP flows would have TCAM entries populated for the duration of their existence in the respective scheme-deployed linecards. Smaller timed flows too would have entries populated but this could be throttled by the timer mechanisms that have been provided.

With respect to multicast routes, all multicast routing entries are programmed in the hardware after consultation with their respective unicast companions in the Route Processor. So the lookup for programming the (*,G) and (S,G) entries in the hardware are done in the software plane in the Route Processor. Hence the multicast entries are not affected by this scheme. This specification relates to unicast routing alone and TCAM entries that relate to it.

Additionally a set threshold called QueryThreshold is configured on the DLC slots. If this is exceeded in terms of rate of packets coming to DLCs from other linecards in the form of the first packet query for a TCAM entry lookup, a periodic full download of the entire forwarding table is done to the those linecards which deploy this scheme. A further purge is then done using the Timers mentioned above. It is also possible to think of deploying this scheme selectively on a set of line cards. In this method, the rest of the line cards can act as TCAM entry route servers. Such an arrangement can help in load balancing the packets involved in the first query process on to the route server line cards in a sticky fashion for DLC lookup. A suitable hashing algorithm can be used to do this load-balancing.

2.2 Switching off the TCAM banks which are unused

The TCAM banks which are not occupied as a result of this scheme may be switched off completely along with their associated SRAM banks as well which contain the rewrite information. The specification opens up the idea to simulations which help strengthen the argument for this scheme.

2.2.1 Algorithm for switching off and switching on TCAM banks

It is important to note that one cannot switch off all TCAM banks that are empty. If there is a sudden rush / burst of traffic at points in time which may be pretty regular for most deployments if the exact number of used TCAM banks alone are set on there would be a problem with respect to not being able to accommodate all the new flows in the switched on TCAM banks. Thus there would be an overflow and that would mean switching TCAM banks that are switched off to ON when it is too late. The latency of switching ON and off TCAMs will be an issue. If it takes too much time to do it packet loss could occur owing to non-availability of TCAM banks.

In order to alleviate the problem specified above the following algorithm among several methods can be used.

```
Begin
1  // At Initialization time

2  All TCAM banks are filled with entries in the non-DLC linecards;

2.1 // When Timers start working

3  Timer mechanisms are used to age out entries that are not used;

4  When TCAM banks become empty NOT all of them are switched off;

5  Say X % of the TCAM banks are switched off;

6  X is determined by the number of active TCAM banks used;

7  if (number of TCAM banks used is greater than say 75%)

8  then

9      X = 0; // This would mean switching on all TCAM banks
              // even those that are empty

10 elseif (number of TCAM banks is lesser than 75 but
11          closer to 50%)
12      X = 25%; // This would mean switching on the rest
                // of the empty TCAM banks

              // TCAM banks switched off = literal 25%. Rest of the
              // TCAM banks are switched ON.

13 elseif (number of TCAM banks is much lesser than 50%
14          but greater than 35%)
15      X = 30%;

              // TCAM banks switched off = literal 30%. Rest of the
              // TCAM banks are switched ON.

16 elseif (number of TCAM banks is lesser than 35%)
17      X = 45%;

              // TCAM banks switched off = literal 45%. Rest of the
              // TCAM banks are switched ON.

18 endif
End
```

The percentages are rounded off to the ceiling of the TCAM banks

existent on the linecard.

Note : TCAM banks that are empty but switched ON are ready to be populated with entries if the bursty traffic trigger the first-packet lookup the DLC/BDLC linecards.

This could be seen as a conservative approach in making sure there are no major events owing to overflow. Underflow is always to be considered to be a good thing. This loop is run periodically and more frequently when the first packet re-direction to the DLC/BDLC is happening frequently.

Thus there is always a buffer of TCAM banks that are ready to be populated in this conservative approach. Simulation results will be published in the next version of the draft.

2.2.2 Prefix entry population latency

When the first-packet is sent to the DLC/BDLC linecards it is possible that there is a latency in populating the resultant prefix + mask entry in the linecard requesting the lookup. It is to be understood that till the lookup entry is populated for the requesting linecard all preceding lookups are forwarded to the DLC/BDLC linecards.

2.3 Using Aggregate routes

Also periodically on the non-DLC linecards a Route aggregation process similar to the S-VA or the simple Virtual Aggregate with exception entries is used to further reduce the TCAM occupation.

2.4 Route Processor as the DLC/BDLC

It is possible that the linecard which is designated as the DLC/Backup DLC can be the Route processor itself. In such a case the same Network Processor Unit chipset that is available on the linecard would be available on the Route Processor itself. There could be more than one Route Processor each with the NPU chipset. Thus the DLC / BDLC could be the Route Processor and backup Route Processor.

2.5 Advantages of this scheme

The TCAM banks unused as a result of this scheme save power when they are empty since they are switched off. Existing commercial chipsets provide the capability of banks of TCAM to be switched off. Also the upcoming chipsets provide for TCAM banks which are smaller than a set of a few large TCAM banks.

Power consumption of related SRAM banks that map to these TCAM banks , which store datum connected to route entries represented in the respective TCAM entries are also saved from being refreshed. The advantage is realized if the SRAM banks are also granularized to be in the form of sets of smaller banks than a set of few large banks.

A finer granularity of TCAM banks with respect to their size and the number of TCAM entries that occupy these banks can be achieved to derive maximum benefit from this scheme.

3 Security Considerations

There are no security considerations with regard to this specification.

4 IANA Considerations

No IANA considerations need to be considered as part of this document.

5 References

5.1 Normative References

None.

5.2 Informative References

None.

Authors' Addresses

Shankar Raman
Department of Computer Science and Engineering
IIT Madras
Chennai - 600036
TamilNadu
India

Email: mjsraman@cse.iitm.ac.in

Balaji Venkat Venkataswami
Department of Electrical Engineering
IIT Madras
Chennai - 600036
TamilNadu
India

EMail: balajivenkat299@gmail.com

Prof.Kamakoti Veezhinathan
Department of Computer Science and Engineering
IIT Madras
Chennai - 600036
TamilNadu
India

Email: kama@cse.iitm.ac.in

Prof.Gaurav Raina
Department of Electrical Engineering
IIT Madras
Chennai - 600036
TamilNadu
India

EMail: gaurav@ee.iitm.ac.in

PANET Working Group
INTERNET-DRAFT
Intended Status: Experimental RFC
Expires: August 2013

Shankar Raman
Balaji Venkat Venkataswami
Prof.Kamakoti Veezhinathan
IIT Madras
February 3, 2013

Computing Power Saving Paths using TCAM Power Ratio
draft-mjsraman-panet-tcam-power-ratio-02

Abstract

A power saving scheme for switching of TCAM banks of fine granularity is discussed in [ID-TCAM-POWER-EFF]. This scheme switches of TCAM banks of fine granularity and their corresponding SRAM banks depending on the occupancy of routes and their rewrites within the TCAM of a intelligent router line card (where one or more such TCAM entities along with their SRAM banks may reside) by using an algorithm specified in [ID-TCAM-POWER-EFF]. This takes care of switching off TCAM and SRAM banks within a router's line card and correspondingly saves power when the traffic matrix is not large with respect to the routes the packets lookup where the occupancy of such routes is low in the line card. This is with reference to both single and multi-chassis devices. The algorithm specified therein tends to switch off banks which are not in use. The aim of this draft is to use the device level characteristic in the form of a TCAM Power Ratio and disseminate it as a metric within an area or an Autonomous system (where multiple such areas exist within an AS) to compute power saving paths through the set of routers where the TCAM-POWER-RATIO is low. The TCAM-POWER-RATIO is arrived at by dividing the number of bits in the TCAM banks that are switched off by the Total Available TCAM bank space in bits. This ratio is then subjected to the CSPF algorithm as an additional constraint or the only constraint as the case may be in a link state protocol like OSPF or IS-IS with Traffic Engineering Extensions with available utilization on the links also being a factor. This way those routers that have more TCAM banks switched off are avoided and those with higher power consumption but with available bandwidth are used hence not increasing the total power consumed within the area and hence within the AS. This could be achieved by using a PCE like entity that calculates the power shortest paths using this TCAM-POWER-RATIO.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	4
1.1	Terminology	4
2.	Methodology of the Proposal	4
2.1	Discussion	7
3	Security Considerations	8
4	IANA Considerations	8
5	References	8
5.1	Normative References	8
5.2	Informative References	8
	APPENDIX - A : References for power saving related material . . .	8
	Authors' Addresses	9

1 Introduction

A power saving scheme for switching of TCAM banks of fine granularity is discussed in [ID-TCAM-POWER-EFF]. This scheme switches of TCAM banks of fine granularity and their corresponding SRAM banks depending on the occupancy of routes and their rewrites within the TCAM of a intelligent router line card (where one or more such TCAM entities along with their SRAM banks may reside) by using an algorithm specified in [ID-TCAM-POWER-EFF]. This takes care of switching of TCAM and SRAM banks within a router's line card and correspondingly saves power when the traffic matrix is not large with respect to the routes the packets lookup where the occupancy of such routes is low in the line card. This is with reference to both single and multi-chassis devices. The algorithm specified therein tends to switch off banks which are not in use. The aim of this draft is to use the device level characteristic in the form of a TCAM Power Ratio and disseminate it as a metric within an area or an Autonomous system (where multiple such areas exist within an AS) to compute power saving paths through the set of routers where the TCAM-POWER-RATIO is low. The TCAM-POWER-RATIO is arrived at by dividing the number of bits in the TCAM banks that are switched off by the Total Available TCAM bank space in bits. This ratio is then subjected to the CSPF algorithm as an additional constraint or the only constraint as the case may be in a link state protocol like OSPF or IS-IS with Traffic Engineering Extensions with available utilization on the links also being a factor. This way those routers that have more TCAM banks switched off are avoided and those with higher power consumption but with available bandwidth are used hence not increasing the total power consumed within the area and hence within the AS. This could be achieved by using a PCE like entity that calculates the power shortest paths using this TCAM-POWER-RATIO.

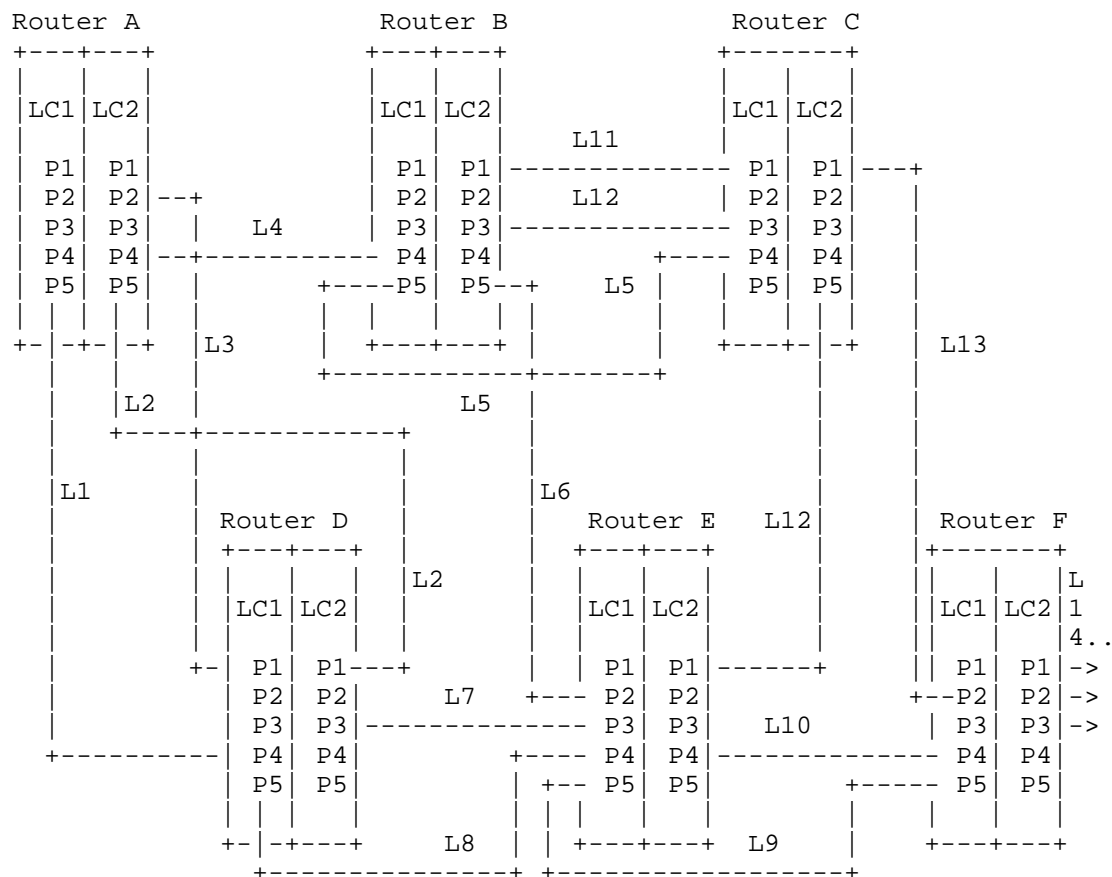
1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Methodology of the Proposal

Consider a topology shown in Figure 1.0 that comprises several chassis devices (in this case single chassis ones) that have multiple line cards within each of them each of them having ports and respective TCAM banks along with their corresponding SRAM banks that contain the next-hop information and necessary re-write information. The TCAM banks contain the routes where packets are subject to longest prefix match lookup.

Consider the following topology...



It is given that the power-saving scheme in [ID-TCAM-POWER-EFF] is in vogue in each of these routers. It is possible that some of them are not running the algorithm specified in the [ID-TCAM-POWER-EFF]. Each of these routers consolidate the total number of bits in the TCAM banks that have been switched off and the total number of bits available in the TCAM banks.

The above 2 pieces of information or data are used to compute the following TCAM-POWER-RATIO.

$$\text{TCAM-POWER-RATIO} = \frac{\text{Bits in TCAM Banks switched off}}{\text{Total number of Bits in TCAM Banks.}}$$

The said ratio is further fed into a computation which is as follows.

Link-Power-Metric = TCAM-POWER-RATIO

Available bandwidth of the said link on the line
card on which TCAM-POWER-RATIO is calculated.

The Link-Power-Metric is calculated for all the links on the line cards within a router and advertised using a suitable TLV in the opaque LSA or LSP packet in the IGP (Interior Gateway Protocol) such as OSPF or IS-IS as a link metric.

The advertised Link-Power-Metric has 2 attached end-points for a link in the above mentioned topology. The overall Link-Power-Metric for a link is arrived at by assigning the Link-Power-Metric between the 2 attached end-points in the ingress direction towards the router which is a candidate next-hop towards the prefix to be reached.

This is then used in the CSPF algorithm to choose the least cost power path in the topology through which the packet can travel from the head-end to tail-end within the area of the AS. This is primarily an area specific calculation which ends up calculating the paths through those routers in the area that are in lesser power saving mode compared to other routers who are in more power saving mode and have shut off more than their fair share of TCAM banks and their corresponding SRAM banks.

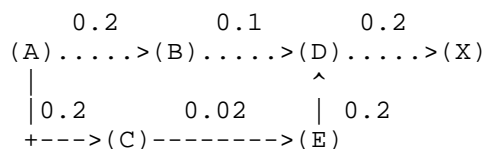
Those routers that do not employ this scheme are given a TCAM-POWER-RATIO of 1 since they are not employing the power saving scheme as outlined in [ID-TCAM-POWER-EFF].

It is important to note that the TCAM banks switched off are not used in the ratio as a discrete quantity but rather as in an interval of thresholds. The Available bandwidth as well are calculated not in discrete quantities but in intervals of thresholds. These intervals are uniformly to be undertaken in all the routers where the scheme is deployed.

2.1 Discussion

Consider the following topology with the appropriate Link-Power-Metric assigned to the links contained within it.

The shortest path is through A,B,D,X rather than through A,C,E,D and X. This primarily owing to the reason that the Link-Power-Metric that takes the TCAM-POWER-RATIO and the Available bandwidth on those links is the least in the CSPF path.



It is to be noted that if the prefix for the traffic whose Longest prefix match entry is not loaded in the TCAM bank which is in ON state in any of the chosen routers along the path, the usual scheme in [ID-TCAM-POWER-EFF] kicks in and loads the prefix and its associated re-write information on the traffic hitting the router in the chosen power saving path. For more information read the reference for scheme in [ID-TCAM-POWER-EFF].

3 Security Considerations

There are no new security considerations within the scope of this document.

4 IANA Considerations

A suitable TLV to carry the Link-Power-Metric is to be defined for this purpose.

5 References

5.1 Normative References

5.2 Informative References

[ID-TCAM-POWER-EFF] Shankar Raman et.al, "TCAM power reduction and optimization in Routers", draft-mjsraman-panet-tcam-power-efficiency-00 (work in progress), 2012.

APPENDIX - A : References for power saving related material

S.Raman, B.V. Venkataswami, K. Veezhinathan, G. Raina, "TCAM power reduction and optimization in Routers", draft-mjsraman-panet-tcam-power-efficiency-00 (work in progress)

M. Zhang, J. Dong, B. Zhang, "Use Cases for Power-Aware Networks", draft-zhang-panet-use-cases (work in progress)

B. Nordman, K. Christensen, "Nanogrids", draft-nordman-nanogrids-00 (work in progress)

T. Suzuki, T. Tarui, "Requirements for an Energy-Efficient Network System", draft-suzuki-eens-requirements (work in progress)

Z. Cao, "Synchronization Layer: an Implementation Method for Energy Efficient Sensor Stack", draft-cao-lwig-syn-layer (work in progress)

A. Junior, R. Sofia, "Energy-awareness metrics global applicability guideline", draft-ajunior-energy-awareness-00 (work in progress)

B. Zhang, J. Shi, M. Zhang, J. Dong, "Power-aware Routing

and Traffic Engineering: Requirements, Approaches, and Issues", draft-zhang- greenet (work in progress)

T. Suganuma, N. Nakamura, S. Izumi, H. Tsunoda, M. Matsuda, K. Ohta, "Green Usage Monitoring Information Base", draft-suganuma-greenmib (work in progress)

S. Raman, B. V. Venkataswami, G. Raina, V. Srini, "Power Based Topologies and TE-Shortest Power Paths in OSPF", draft-mjsraman- rtgwg-ospf-power-topo-01 (work in progress)

S. Raman, B. V. Venkataswami, G. Raina, V. Srini, "Building power optimal Multicast Trees", draft-mjsraman-rtgwg-pim-power-02 (work in progress)

S. Raman, B. V. Venkataswami, G. Raina, "Reducing Power Consumption using BGP", draft-mjsraman-rtgwg-inter-as-ppsp-03 (work in progress)

S. Raman, B. V. Venkataswami, G. Raina, "Building power shortest inter-Area TE LSPs using pre-computed paths", draft-mjsraman-rtgwg- intra-as-ppsp-te-leak-02 (work in progress)

S. Raman, B. V. Venkataswami, G. Raina, V. Srini, "Reducing Power Consumption using BGP path selection", draft-mjsraman-rtgwg-bgp- power-path-02 (work in progress)

Authors' Addresses

Shankar Raman
Department of Computer Science and Engineering
IIT Madras
Chennai - 600036
TamilNadu
India

EMail: mjsraman@cse.iitm.ac.in

Balaji Venkat Venkataswami
Department of Electrical Engineering,
IIT Madras
Chennai - 600036
TamilNadu
India

Email: balajivenkat299@gmail.com

Prof.Kamakoti Veezhinathan
Department of Computer Science and Engineering
IIT Madras
Chennai - 600036
TamilNadu
India

Email: kama@cse.iitm.ac.in

PCE Working Group
Internet-Draft
Intended Status: Experimental RFC
Expires: August 2013

Shankar Raman
Balaji Venkat Venkataswami
Gaurav Raina
IIT Madras
February 18, 2013

Constructing power optimal P2MP TE-LSPs within an AS
draft-mjsraman-pce-power-replic-02

Abstract

Power consumption in multicast replication operations is an area of concern and choosing suitable replication points that can decrease power consumption overall assumes importance. Multicast replication capacity is an attribute of every line card of major routers and multi-layer switches that support multicast in the core of an Internet Service Provider (ISP) or an enterprise network.

Currently multicast replication points on Point-to-Multipoint Traffic Engineering Label-Switched-Paths (P2MP TE-LSPs) consume power while delivering multiple output streams of data from a given input stream. The multicast distribution trees are constructed without any regard for a proper placement of the replication points and consequent optimal power consumption at these points.

This results in overloading certain routers while under-utilizing others. An optimal usage of these replication resources could substantially reduce power consumption on these routers. In this paper, we propose a mechanism by which P2MP TE-LSPs are constructed for carrying multicast traffic across multiple areas within a given AS. We propose that these LSPs be built by using the advertisements of the power-replication capacity ratio advertised by fine grained components such as multicast capable line-cards of routers and multi-layer switches deployed within an AS.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	3
1.1	Terminology	3
2.	Methodology of the proposal	4
2.1	Discussion of this scheme	6
2.2	Power to available multicast replication capacity ratio in a TLV	8
3	Conclusion	10
3	Security Considerations	11
4	IANA Considerations	11
5	References	11
5.1	Normative References	11
5.2	Informative References	11
	Authors' Addresses	12

1 Introduction

Multicast traffic across multiple areas within a given AS, may be carried using P2MP TE-LSPs. The traffic may be carried from a ingress Provider Edge (PE) router to several egress PEs, example in a multicast Virtual Private Network (MVPN) case. The autonomous system (AS) may comprise of multiple areas involving a backbone area and several non-backbone areas connected to each other through the backbone. If several such multicast streams are to be carried in the AS, it would be most useful to have such P2MP TE-LSPs constructed such that they have optimal power to available replication capacity ratios on the routers' linecards that they traverse from source to destinations. The intent is to provide a solution whereby several such P2MP TE-LSPs can be laid out in such a way that the set of routers that replicate multicast traffic traversed by the P2MP TE-LSPs are most optimal in the utilization of the power provided to them given that there is sufficient replication capacity available. This we believe would essentially lead to a equilibrium of power to available replication capacity ratios amongst all routers in the topology which in turn would optimize and reduce the overall ratios for the AS.

Each router and its respective linecards deployed in the AS have an advertised capability for replication. Most multi-layer switches and routers from vendors advertise in their respective data sheets a certain capability for replication for each type of linecard deployable on the box. Replication consumes power and delivers multiple streams of data from a given input stream. It is status quo that P2MP (Point-to-Multipoint) Label Switched Paths are constructed without taking into account the power to available replication capacity ratios of such routers thus overloading certain routers while underutilizing the others. An optimal usage of these resources could reduce power consumption on these routers / multi-layer switches. This equilibrium could be arrived at by using a capability to advertise from each router a Traffic Engineering Database Link State Advertisement (TED-LSA) that carries the power to available replication capacity ratio of each of the said router's line cards, depending on the current utilization of its replication capacity and power consumption.

This paper is organized as follows; In section 2, we deal with the scheme that we propose. In section 2.1, we discuss some examples of the scheme at work, and in section 3 we conclude with future areas of study that may be useful to undertake.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT",

"SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Methodology of the proposal

The key metric under consideration is the power consumed DIVIDED BY available replication capacity on each of the linecards of a router in the AS, which is eligible to be used as a node atop which multicast traffic can be carried. Once an advertisement about the said metric has been sent in the regular flooding process in Link State routing protocols such as OSPF-TE or ISIS-TE, it would be possible for a head-end router for a P2MP TE-LSP to compute the TE-LSP through the AS from the ingress PE to all egress PEs of that multicast stream in such a way that the power to available replication capacity ratios at the replication points are minimal on that path. The Constrained Shortest Path First (CSPF) algorithm could be modified to compute the least cost power to available replication capacity ratio path and thus cause an equilibrium shift to be caused. This path would be supplied to the RSVP-TE component of the head-end and that would set up the path with appropriate labels. Once RSVP-TE establishes the path and traffic is carried across it, the reduced replication capacity of the routers in the P2MP TE-LSP path would be re-advertised again, which in turn would be useful for computation of the other paths from the instance that the replication capacity changed on these routers.

Assume that the following router topology in the vicinity of the sender / senders is computed.

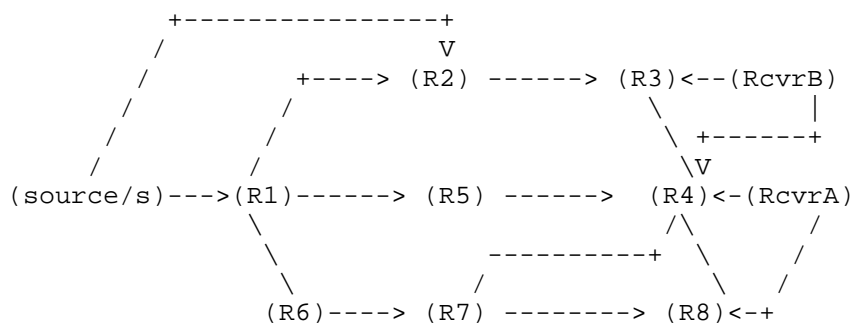
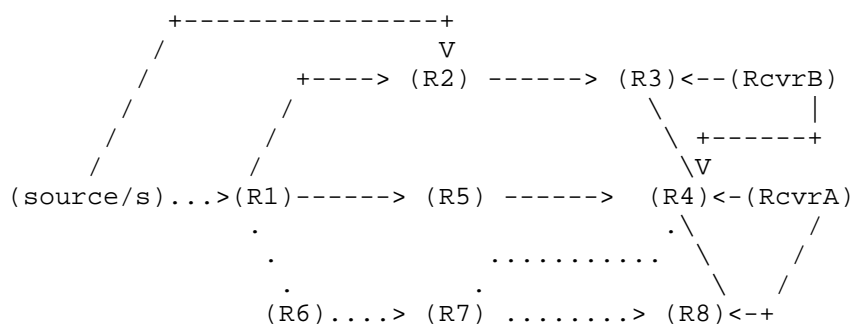


Figure 1: Topology within a given AS with coloring for Power-replication ratios

In the above diagram you can see that the source/sources are connected using a multi homed connections to the same ISP through Routers R1 and R2. Similarly there are two Receiver sites RcvrA and RcvrB that are multihomed to TWO Routers RcvrB to R3 and R4 and for RcvrA to R4 and R8 respectively.



Legend : dotted lines represent path computed.

Figure 2: Instantiating an optimal power consuming distribution tree

Given that the path calculation engine at the head-end R1 is given this topology and along with other TED-LSA packets the current power to available replication capacity ratios are advertised through the IGP-TE extensions to the head-end R1, the paths with the least power to available replication capacity ratios are computed and the paths setup from head-end PEs to the tail-end PEs where the receivers are connected. It is to be noted that the ratios computed for power to available replication capacity on the topology are examined and the replication points are setup on those routers that have the least power to available replication capacity ratio. If branching points are not required at certain points, these are anyways placed on least cost power ratio routers that are the next best location to setup a non-branching point.

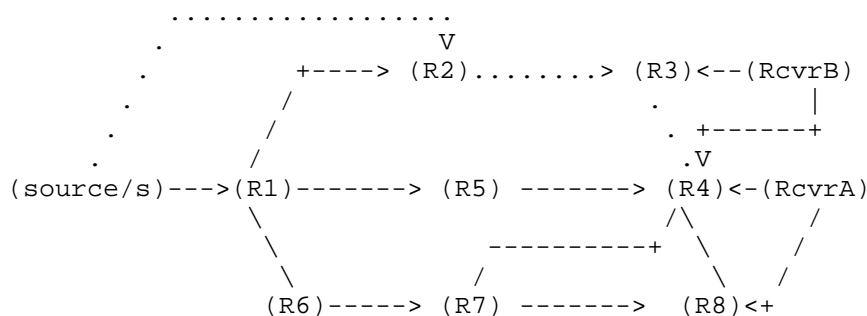
Assume the following path is computed as per the least power to available replication capacity ratios. Paths are computed through R6, R7, R8, R4, and say the multicast stream occupies 4GB of traffic along this tree so constructed and the available capacity of these routers reduces to 6GB assuming all of them have a base capacity of 10GB. Subsequent paths constructed would have to take into account the newly computed power to current replication capacity ratio in the topology and construct new P2MP TE-LSPs for multicast streams yet to come.

Assume another 6GB worth of traffic is loaded onto this topology in terms of a multicast stream / multiple streams then the new path

computed for these new streams would possibly utilize the same path as computed before. If the old streams reduce the replication capacity to an extent such that routers through which they pass can no longer be used since these routers' power to available replication capacity has become poor when compared to other paths then a different path may be computed from the ingress PE to the egress PEs in such a way as to avoid those routers which have such poor ratios.

For example, assume R6, R7, R8 and R4 have exhausted their capacity, or guzzle more power as a result of them carrying the 4GB stream that was originally placed atop them. then a different path would be chosen as follows. The path followed as shown in the Figure is R2,R3 and R4. Given that R4 is the only choice since it has connectivity to both Receivers, in this case the branch point is placed atop R3, one branch to get to RcvrB and the other to get to RcvrA through R4. Policy decisions could guide the placement in case of a tie. Here the the only choice has been to drive the end replication to RcvrA through R4 and RcvrB through R3 owing to topology constraints.

It is to be noted that the power consumed by the linecard is divided by the available replication capacity to arrive at a ratio and that ratio is assigned as a weight to all of the links ingressing on that linecard. It is possible that one might take a weighted average by dividing a weighted co-efficient sum by the weighted sum of ingress links on a linecard and the metrics so assigned be used as the metric for calculation.



Legend : dotted lines represent path computed.

Figure 3: Instantiating a subsequent optimal power consuming distribution tree

2.1 Discussion of this scheme

It is to be noted that our scheme applies to centralized schemes of path calculations. What is being calculated is a tree of nodes that

form a P2MP tree where each node can be conceptualized as a router (read also multi-layer switches) and each edge the link connecting one or more ports on a line card to another linecard on a downstream router to carry multicast traffic from a source located at the head end ingress router to several receiver nodes connected to egress routers. We will call this calculated tree as a P2MP tree. The tree is calculated by the PCE in the head end / ingress router through which sources connect. The PCE calculates the intra-AS P2MP path (the literal P2MP TE-LSP within the AS) within that AS.

The calculated power to available replication capacity ratio is assigned to each of the ingress links on a linecard on a router en-route to egress links through which the multicast stream is replicated on the same router. Thus all ingress links to a router through a linecard are assigned the same metric as the power ratio so calculated. The egress links would in continuity connect to a unicast tunnel or another branch-point in the tunnel towards the receivers which are represented as the egress routers. The egress routers would in turn be replication points or direct connections to the actual receivers. This method could be applied for multicast traffic to be transported through MVPNs. The method of egress routers' discovery is left to existing mechanisms. The primary input to the invention proposed is an ingress router and their respective egress routers. The other input to the construction of P2MP tree is the router level topology with the metrics for the power to available replication capacity ratio.

It is to be noted that this CSPF calculation can be hastened in terms of time complexity by dividing the weights into equivalence classes. First we divide the nodes into graph colored nodes with the least ratio nodes marked as green as shown in the figure and given that there exists a path that is all green from source to egress PEs, one of such paths is chosen. If after coloring the nodes a path which is disconnected exists, we incrementally add the next best colored nodes to the graph to see if we get a connected path from source to egresses. These steps are repeated until we find a connected path. This will hasten the algorithm to a conclusion rather than use a brute force method which may take inordinate amount of time. R4 being used in the 6GB case is an example of this. Because of topology restrictions the R4 node had to be chosen in spite of the fact that it is not green after carrying the 4GB stream.

Routers may have step levels in which they increase power consumption when they additively are loaded with more large bandwidth consuming multicast streams. Calibrating these levels may be useful for implementing this scheme. It is possible that such calibrated thresholds can be used for advertising the power to available replication capacity ratios in the IGP-TE advertisements. This would

be useful for bringing down the frequency of updates or advertisements from a line-card about its ratios. When power consumption meanders within a certain given interval these ratios need not be readvertised even if further multicast streams are added to it. The incentive is to recognize a linecard that does not drastically change power consumption even if large bandwidth streams are added onto it for replication and thus give it credit for its power optimal functioning. If a router tends to consume the highest level of power even when carrying low amounts of multicast streams and replicating them on its line card, it would automatically have a poor ratio when compared to a router that efficiently uses power when considering the replication capacity being used. The best case would be a low power consuming line-card or a router filled with such line cards that does not leave its power interval no matter how much ever replication capacity is sought to be used on it. But that would be an ideal condition but it is definitely an idealistic scenario towards which the router manufacturers should look at.

It is possible that several multicast streams may be aggregated onto a single P2MP-TE-LSP representing the given multicast tree that encompasses the union of all the egress PEs of the several multicast streams. The Ingress PE router is however common for all the multicast streams so covered. Aggregation of these several multicast streams from a given Ingress PE to several egress PEs is a common occurrence to save the amount of state in the core of the network. By aggregating these streams onto a single P2MP tree, it is possible to amortize the cost of replication amongst a particular set of ingress linecards / ports on those line cards while taking into account the current power consumption and replication capacity available at the time of computing the P2MP TE-LSP.

The dynamic nature of the multicast tree and the egress PEs that join into it and leave it based on whether there are multicast listeners in that VPN site attached to the said egress PE/ PEs, makes it important to position the replication points in such a way that there is maximum leverage on optimization in the ratios overall for the AS which are computed. When aggregating multiple multicast streams over a single P2MP TE-LSP it is important to keep this in mind.

So the key point is to aggregate multiple streams with a set theoretical approach in mind so that there is maximum overlap of egress PEs for these streams and position these streams atop a P2MP TE-LSP in such a way that ratios are most optimal for that set of streams (with the overall AS power consumption in mind).

2.2 Power to available multicast replication capacity ratio in a TLV

As per [RFC3630] the Link TLV can be used to carry this power to

available multicast replication capacity ratio with an additional sub-TLV of the link TLV. The sub-type number 10 is recommended to be defined for this purpose.

[RFC 3630] states in section 2.2.1 and we QUOTE ...

2.2.1 Link TLV

The Link TLV describes a single link. It is constructed of a set of sub-TLVs. There are no ordering requirements for the sub-TLVs.

Only one Link TLV shall be carried in each LSA, allowing for fine granularity changes in topology.

The Link TLV is type 2, and the length is variable.

The following sub-TLVs of the Link TLV are defined:

- ```

1 - Link type (1 octet)
2 - Link ID (4 octets)
3 - Local interface IP address (4 octets)
4 - Remote interface IP address (4 octets)
5 - Traffic engineering metric (4 octets)
6 - Maximum bandwidth (4 octets)
7 - Maximum reservable bandwidth (4 octets)
8 - Unreserved bandwidth (32 octets)
9 - Administrative group (4 octets)
10 - Power-to-Multicast-replication-capacity (4 octets)

```

point in front of it. Thus, the above represents the value:

$$(-1)**(S) * 2**(Exponent-127) * (1 + Fraction)$$

It is proposed that we use the Power-to-multicast-replication-capacity ratio as a 32 bit IEEE floating Point format field for the purpose of this document.

### 3 Conclusion

Here we propose a scheme that takes into account the power to available replication capacity ratios as weights for the edges and compute a low cost power path for multicast replication. This scheme could be extended to inter-AS multicast streams or to inter-AS multicast streams where the multicast stream is sought to be carried over multiple ASes. This is an area of future study which would be most conducive in terms of bringing about optimal power usage and thus incentivising vendors to manufacture low power consuming equipment. Compelled to bring about radical change in the thinking relating to power consumption vendors manufacturing networking equipment will drive down power consumption since the scheme proposed chooses or gives priority to low power guzzling linecards.

### 3 Security Considerations

The security considerations for this proposal are the same as any NEW opaque LSA introduced in an IGP like OSPF, IS-IS.

### 4 IANA Considerations

IANA would need to assign a NEW opaque LSA type to carry power and multicast replication capacity such that this information can be carried in the TE-LSAs within an AS.

### 5 References

#### 5.1 Normative References

#### 5.2 Informative References

- [1] G. Appenzeller, Sizing router buffers, Doctoral Thesis, Department of Electrical Engineering, Stanford University, 2005.
- [2] A. P. Bianzino, C. Chaudet, D. Rossi and J. L. Rougier, A survey of green networking research, IEEE Communications and Surveys Tutorials, preprint.
- [3] J. Baliga, K. Hinton and R. S. Tucker, Energy consumption of the internet, Proc. of joint international conference on optical internet, June 2007, pp. 1993.
- [4] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang and S. Wright, Power awareness in network design and routing, Proc. of the IEEE INFOCOM 2008, April 2008, pp. 457-465.
- [5] M. Xia et. al., Greening the optical backbone network: A traffic engineering approach, IEEE ICC Proceedings, May 2010, pp. 1995.
- [6] W. Lu and S. Sahni, Low-power TCAMs for very large forwarding tables, IEEE/ACM Transactions on Computer Networks, June 2010, vol. 18, no. 3, pp. 948-959.
- [7] B. Zhang, Routing Area Open Meeting, Proceedings of the IETF 81, Quebec, Canada, July 2011.

Authors' Addresses

Shankar Raman  
Department of Computer Science and Engineering  
IIT Madras  
Chennai - 600036  
TamilNadu  
India  
  
EMail: mjsraman@cse.iitm.ac.in

Balaji Venkat Venkataswami  
Department of Electrical Engineering  
IIT Madras  
Chennai - 600036  
TamilNadu  
India  
  
EMail: balajivenkat299@gmail.com

Prof.Gaurav Raina  
Department of Electrical Engineering  
IIT Madras  
Chennai - 600036  
TamilNadu  
India  
  
EMail: gaurav@ee.iitm.ac.in

Routing Area Working Group  
Internet-Draft  
Intended status: Informational  
Expires: April 27, 2015

A. Retana  
Cisco Systems, Inc.  
R. White  
Ericsson  
M. Paul  
Deutsche Telekom AG  
October 24, 2014

A Framework and Requirements for Energy Aware Control Planes  
draft-retana-rtgwg-eacp-03

Abstract

There has been, for several years, a rising concern over the energy usage of large scale networks. This concern is strongly focused on campus, data center, and other highly concentrated deployments of network infrastructure. Given the steadily increasing demand for higher network speeds, always-on service models, and ubiquitous network coverage, it is also of growing importance for telecommunication networks both local and wide area in scope. One of the issues in moving forward to reduce energy usage is to ensure that the network can still meet the performance specifications required to support the applications running over it.

This document provides an overview of the various areas of concern in the interaction between network performance and efforts at energy aware control planes, as a guide for those working on modifying current control planes or designing new control planes to improve the energy efficiency of high density, highly complex, network deployments.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 27, 2015.

## Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

|                                                                   |    |
|-------------------------------------------------------------------|----|
| 1. Introduction . . . . .                                         | 4  |
| 2. Requirements Language . . . . .                                | 4  |
| 3. Background . . . . .                                           | 5  |
| 3.1. Scope . . . . .                                              | 5  |
| 3.2. Business Drivers . . . . .                                   | 6  |
| 3.3. Application Drivers . . . . .                                | 6  |
| 4. Framework . . . . .                                            | 7  |
| 4.1. Modes of Reducing Energy Usage . . . . .                     | 7  |
| 4.1.1. Example Network . . . . .                                  | 8  |
| 4.1.2. Examples of Energy Reduction . . . . .                     | 8  |
| 4.2. Global Verses Local Decisions . . . . .                      | 9  |
| 5. Considerations and Requirements . . . . .                      | 9  |
| 5.1. Energy Efficiency and Bandwidth Reduction . . . . .          | 9  |
| 5.1.1. An Example of Lowered Bandwidth . . . . .                  | 10 |
| 5.1.2. Requirements . . . . .                                     | 10 |
| 5.2. Energy Efficiency and Stretch . . . . .                      | 10 |
| 5.2.1. An Example of Stretch . . . . .                            | 11 |
| 5.2.2. Requirements . . . . .                                     | 11 |
| 5.3. Energy Efficiency and Fast Recovery . . . . .                | 12 |
| 5.3.1. An Example of Impact on Fast Recovery . . . . .            | 12 |
| 5.3.2. Requirements . . . . .                                     | 12 |
| 5.4. Introducing Jitter Through Microsleeps . . . . .             | 13 |
| 5.4.1. An Example of Microsleeps to Reduce Energy Usage . . . . . | 13 |
| 5.4.2. Requirements . . . . .                                     | 14 |
| 5.5. Other Operational Aspects . . . . .                          | 14 |
| 5.5.1. An Example of Operational Impact . . . . .                 | 14 |
| 5.5.2. Requirements . . . . .                                     | 14 |
| 6. Security Considerations . . . . .                              | 14 |
| 7. Acknowledgements . . . . .                                     | 15 |
| 8. References . . . . .                                           | 15 |
| 8.1. Normative References . . . . .                               | 15 |
| 8.2. Informative References . . . . .                             | 15 |
| Appendix A. Change Log . . . . .                                  | 15 |
| A.1. Changes between the -00 and -01 versions. . . . .            | 15 |
| A.2. Changes between the -01 and -02 versions. . . . .            | 16 |
| A.3. Changes between the -02 and -03 versions. . . . .            | 16 |
| Authors' Addresses . . . . .                                      | 16 |



## 1. Introduction

As energy prices continue to increase, and energy awareness becomes a watchword for most large companies, places where the network infrastructure used a good deal of power have come under increased scrutiny for savings. There is a concern, however, in saving energy at the cost of network operations --to reduce performance along with energy consumption, negatively impacting the operation of a network and the applications reliant on that network. This concern is primarily focused on the network control plane, but will necessarily apply to network performance and energy usage overall.

This document provides a background, a framework for understanding and managing the tradeoffs between modifications made to network protocols to conserve energy and network performance metrics and requirements, and a set of requirements for protocol designers to consider in proposals for new control plane protocols or modifications to existing control plane protocols. It is intended to encourage work on mechanisms that will reduce network energy usage while providing perspective on balancing energy usage against performance. The ultimate goal is to provide the tools and knowledge necessary for protocol designers to modify network protocols to best balance efficiency against performance, and to provide the background information network operators will need to intelligently deploy and use protocol modifications to network protocols.

The document is organized as follows. Section 3 provides material the reader needs to understand to appreciate the challenges inherent in balancing energy reduction with effective network performance. This section includes subsections considering the application and business requirements that are the basis of the rest of the document. Section 4 provides a framework for understanding mechanisms common to all energy management schemes proposed to date in general terms. Section 5 provides an analysis of the areas highlighted, including an explanation of how the specific area interacts with energy management, and example of the interaction, and, finally, a set of requirements protocol designers should consider when proposing either new protocols or modifications to existing protocols to reduce energy usage.

## 2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

### 3. Background

The background covered here describes the underlying business and application drivers for the consideration and requirements sections below. This section also contains a small example network used throughout the remainder of this document for explaining various mechanisms and technical points.

#### 3.1. Scope

The reader should differentiate between radio based and wireline (or rather, "plugged in"), networks. Radio based networks designed for rapid deployment for highly mobile users (often called Mobile Ad Hoc Networks, or MANETs [MANET]), and sensor networks designed for low power, processing, and memory (such as those described in [ROLL]), are not the target of this document. Readers should refer to the groups working within those areas for energy management requirements based on those specialized environment. While protocol developers for those environments may draw useful information from this document, this work is not intended to address those specialized networks specifically. Mobile cellular networks however are similarly affected by excess energy consumption as wireline networks and seek to save energy by methods as described in the following (see e.g. [3GPP]).

The reader should also differentiate between intradomain and interdomain applications. Interdomain applications require more work in policy than in technical and business considerations, and therefore fall outside the scope of this document. Intradomain control planes are (intuitively) where most energy savings will be attained, at any rate. Most high concentrations of routers, such as data centers and campus networks, are under a single administrative domain. Therefore, placing interdomain control planes outside the scope of this document does not limit its usefulness in any meaningful way.

The reader should further differentiate between the components of an energy management system, namely energy monitoring and energy control. Energy monitoring deals with the collection of information related to energy utilization and characteristics, as described in [EMAN]. Energy control relates to directly influencing the optimization and/or efficiency of devices in the network. The focus of this document is on understanding the tradeoffs between modifications made to network protocols to conserve energy and network performance metrics and requirements, not on the functions, steps or procedures required for energy monitoring.

### 3.2. Business Drivers

Networks are primarily built to support both broad and narrow business requirements. Broad business requirements might include general communication requirements, such as providing email service between internal and external personnel, or providing general access to the World Wide Web for research and business support. Narrow requirements would relate to specific applications, such as supporting a particular financial application in the case of a bank or other financial enterprise, or supporting customer traffic in the case of a service provider. Application requirements will be considered in greater detail in the next section.

Another class of requirements business place on networks can be called operational requirements. These include (but are not limited to), capital expense, operational expense, and the restrictions the network architecture places on the growth and operation of the business itself. These, in turn, drive requirements such as change management, total uptime (availability), and the ability of the network to be easily and quickly modified to meet new business demands, or to shed old business demands. Operational expense is the primary area this document covers in relation to business requirements, because this is where energy management most obviously overlaps with network performance.

### 3.3. Application Drivers

Applications drivers provide the background for each of the technical sections below. When approaching a specific application, there are only a small number of questions network and protocol designers need to fully understand to shape networks and protocols so a specific application can be supported. The first two questions revolve around bandwidth; how much bandwidth will the application consume, and is this bandwidth consumption fairly steady, or highly variable? For instance, applications such as streaming video tend to have long lasting flows with high bandwidth requirements, file transfers tend to produce shorter flows requiring high bandwidth, and HTML traffic tends to be bursty, with much lower bandwidth requirements.

The next question a protocol or network designer might ask about a specific application is it's tolerance to jitter. Real time applications, such as voice and video conferencing, have a very low toleration for jitter. File transfers and streaming video, on the other hand, can often handle large variations in packet arrival times. If packets are delayed long enough, the application may actually time out, shutting down sessions. Users will often "hang up" after a short period of time, as well, causing loss of revenue and productivity.

Delay is another crucial factor in the performance of many applications. Many server virtualization protocols, for instance, have very low tolerance for delay, having been written with a short wire local broadcast segment in mind. Applications such as stock and commodity trading, remote medical, and collaborative video editing also exhibit very little tolerance for delay.

These last two application drivers, jitter and delay, are normally the result of two underlying causes within a network's control plane: stretch and convergence. Stretch (defined more fully in the section considering stretch below) causes longer paths to be taken through the network. Each hop in the network path adds serialization into and out of a set of queues in device memory, along with the delays of various queuing mechanisms implemented on that device. Each hop in the network increases delay directly, and has the potential to increase jitter as packets pass into and out of the additional devices.

Network convergence will also show up as jitter in an application's stream; if packets are held up or looped for hundreds of milliseconds during a network convergence event, applications running over the converging topology will see this convergence time as a massive jitter event, or a short term delay in the delivery of packets.

Jitter and delay can also be introduced directly into the packet stream by reducing the throughput of individual links, or putting devices and/or links into energy reduced modes for very short periods of time (microsleeps). If a link is asleep when the first and third packets from a flow arrive at the head end of the link, and not when the second packet from that same flow arrives, each packet is going to be processed differently, and hence will have a different delay across the path.

The specific technical problems addressed in the following sections, then, are bandwidth reduction, increasing stretch, network convergence, and introducing jitter through microsleeps.

#### 4. Framework

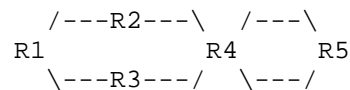
##### 4.1. Modes of Reducing Energy Usage

Regardless of whether the control plane is centralized (such as some form of centrally computed traffic engineering or software defined network), or distributed (traditional routing protocols), there are four primary ways in which energy usage can be reduced:

- o Removing redundant links from the network topology
- o Removing redundant network equipment from the network topology
- o Reducing the amount of time equipment or links are operational
- o Reducing the link speed or processing rate of equipment

#### 4.1.1. Example Network

To illustrate the impacts of link and device removal throughout the rest of this document, the following network is used.



This network is overly simplistic so the impact of removing various links and devices from the topology can be more clearly illustrated. More complex topologies will often exhibit these same impacts without being so obvious.

#### 4.1.2. Examples of Energy Reduction

In the example network above, several different modes of energy reduction might be:

- o Shutting down one of the two links between R4 and R5
- o Shutting down one of the two links between R4 and R5, and shutting down any line cards (or part of the nodes themselves) associated with the removal of these links
- o Shutting down R2 or R3, since these represent alternate paths to reach the same set of destinations
- o Shutting down the link between R2 and R4, since similar connectivity is provided through R1->R3->R4
- o Shutting down all links and devices for fractions of time in a coordinated fashion
- o Shutting down individual links as traffic or the control plane permits for fractions of time (here the momentary shutdown of various links is not coordinated, but undertaken hop by hop)
- o Reducing the speed of all links and devices for fractions of time in a coordinated fashion

- o Reducing the speed of individual links as traffic or the control plane permits for fractions of time (here the momentary slowdown of various links is not coordinated, but undertaken hop by hop)

#### 4.2. Global Verses Local Decisions

Independent of whether the control plane is centralized or distributed, the scope considered when making a decision about energy efficiency may affect the result and effectiveness of the system. There are clearly two extreme options when looking at the scope of the information used to make decisions. The first extreme is that of every device in the network considering only local conditions, and determining the proper local state from that information. An example of this mode of operation might be a local link where the devices on either side of that link measure the link utilization, and independently decide to automatically shut the link down when utilization reaches a specific threshold. An example of the other end of the spectrum might be a network control plane in which all the nodes involved agree before taking a specific action; in the case of two parallel links, the devices on each end not only would have similar configured policies, but would coordinate if one of the links was to be turned off. It is outside the scope of this document to determine which of these two options may be optimal or "best."

There are some considerations and tradeoffs which need to be outlined in considering the global versus local decisions in relation to energy efficiency. System designers should take note of the difficulties with preventing pathological conditions when purely localized decisions are made. For instance, in the example network, assume R1 determines to put the R1->R2 link into an energy saving mode, while R4 determines to put the R4->R3 link into an energy saving mode. In this case, no path will remain available through the network. It is also possible for the opposite to occur, that is for no links or devices to be placed into a reduced energy state because R1 and R4 don't agree through the control plane which links and devices should be removed from the topology.

Protocol designers should consider these tradeoffs in proposals for energy aware control planes.

### 5. Considerations and Requirements

#### 5.1. Energy Efficiency and Bandwidth Reduction

Bandwidth is an important consideration in high density networks; most data centers are designed to provide a specific amount of bandwidth into and out of each server and to facilitate virtual

server movement among physical devices. In campus and core networks bandwidth is finely coupled with quality of service guarantees for applications and services. It should be obvious that removing links or devices from a network topology will adversely affect the amount of available bandwidth, which could, in turn, cause well thought out quality of service mechanisms to degrade or fail.

What might not be so obvious is the relationship between available bandwidth and jitter, or other network quality of service measures. If higher speed links are removed from the topology in order to continue using lower speed (and therefore presumably lower power) links, then serialization delays will have a larger impact on traffic flow. Longer serialization delays can cause input queues to back up, which impacts not only delay but jitter, and possibly even traffic delivery.

#### 5.1.1. An Example of Lowered Bandwidth

In the network illustrated above, one of the two links between R4 and R5 could be an obvious candidate for removal from the network. Especially if the network load can easily be transferred to the remaining link without failure, and without serious consequences for delay or jitter in the network, there is a strong case to be made for doing so --particularly if the accompanying line cards could also be shut down to add to the energy savings.

#### 5.1.2. Requirements

Modifications to control plane protocols to achieve network energy efficiency SHOULD provide the ability to set the minimal bandwidth, jitter, and delay through the network, and not shut down links or devices that would violate those minimal requirements.

#### 5.2. Energy Efficiency and Stretch

In any given network, there is a shortest path between any source and any destination. Network protocols discover these paths from the destination's perspective --routing draws traffic along a path, rather than driving along a path. Along with the shortest path, there are a number of paths that can also carry traffic from a given source to a given destination without the packets passing along the same logical link, or through the same logical device, more than once. These are considered loop-free alternate [RFC5714] paths.

The primary difference between the shortest path and the loop-free alternate paths is the total cost of using the path. In simple terms, this difference can be calculated as the number of links and devices a packet must pass through when being carried from the source

to the destination --the hop count. While most networks use much more sophisticated metrics based on bandwidth, congestion, and other factors, the hop count will stand in as the only metric used throughout this document.

When the control plane causes traffic to pass from the source to the destination along a path which is longer than the shortest path, the network is said to have stretch (see [Krioukov] for a more in depth explanation of network stretch). To measure stretch, simply subtract the metric of the shortest path from the metric of the longer path. For example, in hop count terms, if the best path is three hops, and the current path is four hops, the network exhibits a stretch of 1.

#### 5.2.1. An Example of Stretch

In the network illustrated above, if a modification is made to the control plane in order to remove the link between R1 and R4 in order to save energy, all the destinations shown in the diagram remain reachable. However, from the perspective of R1, the best path available to reach R2 has increased in length by two hops. The original path is R1->R2, the new path is R1->R3->R4->R2. This represents a stretch of 2.

Along with this increased stretch will most likely also come increased delay through the network; each hop in the network represents a measurable amount of delay. This increased stretch might also represent an increased amount of jitter, as there are more queues and more serialization events in the path of each packet carried. There will also be the modifications in jitter as the network switches between the optimal performance configuration and an energy efficient configuration.

#### 5.2.2. Requirements

Designers who propose modifications to control plane protocols to achieve network energy efficiency SHOULD analyze the impact of their mechanisms on the stretch in typical network topologies, and SHOULD include such analysis when explaining the applicability of their proposals. This analysis may include an examination of the absolute, or maximum, stretch caused by the modifications to the control plane as well as analysis at the 95th percentile, the average stretch increase in a given set of topologies, and/or the mean increase in stretch.

Mechanisms that could impact the stretch of a network SHOULD provide the ability for the network administrator to limit the amount of stretch the network will encounter when moving into a more energy efficient mode.



### 5.3. Energy Efficiency and Fast Recovery

A final area where modifications to the control plane for energy efficiency is fast convergence or fast recovery. Many networks are now designed to recover from failures quickly enough to only cause a handful of traffic to be lost; recovery on the order of half a second is not an uncommon goal. It should be obvious that removing redundant links and devices from the network to reduce energy consumption could adversely affect these goals.

#### 5.3.1. An Example of Impact on Fast Recovery

In the network shown, assume R2 and its associated links are removed from the topology in order to save energy. Rather than this second path being available for immediate recovery on the failure of the R1->R3 link, some process must be followed to bring R2 and its associated links back up, reinject them into the topology, and finally begin routing traffic across this path.

In many situations, only links and devices which are a "third point of failure" may be acceptable as removal candidates in order to conserve energy.

#### 5.3.2. Requirements

Modifications to the control plane in order to remove links or nodes to conserve energy SHOULD entail the ability to choose the level of redundancy available after the network topology has been trimmed. For instance, it might be acceptable in some situations to move to single points of failure throughout the network, or in specific sections of the network, for certain periods of time. In other situations, it may only be acceptable to reduce the network to a double point of failure, and never to a single point of failure.

The complete removal of nodes or links from the network topology has several impacts on the control plane which must be considered. In these cases, the control plane must:

- o Modify the network topology so removed links or devices are not used to forward traffic
- o Remember that such links exist, possibly including the neighbors and destinations reachable through those links or devices

#### 5.4. Introducing Jitter Through Microsleeps

One proposed mechanism to reduce energy usage in a network is to sleep links or devices for very short periods of time, called microsleeps. For instance, if a particular link is only used at 50% of the actual available bandwidth, it should be possible to place the link in some lower power state for 50% of the time, thus reducing energy usage by something percentage.

Such schemes introduce delay and jitter into the network path directly; if a packet arrives while the link to the next hop, or the next hop itself, is in a reduced energy state, the packet must wait until the link or next hop device enter a normal operational mode before it can be forwarded. Most of the time the proposed sleep states are so small as to be presumably inconsequential on overall packet delay, but multiple packets crossing a series of links, each encountering different links in different states, could take very different amounts of time to pass along the path.

One possible way to resolve this somewhat random accrual of delays on a per packet basis is to coordinate these sleep states such that packets accepted at the entry of the network are consistently passed through the network when all links and devices are in a normal operating mode, and simply delaying all packets at the entry point into the network while the devices in the network are in some energy reduced state. This solution still introduces some amount of jitter; some packets will be delayed by the sleep state at the edge of the network, while others will not. This solution also requires coordinated timers at the speed of forwarding itself to effectively control the sleep and wake cycles of the network.

##### 5.4.1. An Example of Microsleeps to Reduce Energy Usage

In the example network, assume the bandwidth utilization along the path R1->R2->R4->R5 is 50% of the actual available bandwidth along this path. It is possible to consider a scheme where R1->R2, R2->R4, and R4->R5 are all put into some energy reduced operational mode 50% of the time, since packets are only available to send 50% of the time. A packet entering at R1 may encounter a short delay at R1->R2, at R2->R4, and at R4->R5, or it might not. Even if these delays are very small, say 200ms at each hop, the accumulated delay through the network due to sleep states may be 0ms (all links and devices awake) or 600ms (all links and devices asleep) as the packet passes through the network.

As network paths lengthen to more realistic path lengths in real deployments, the jitter introduced varies more widely, which could cause problems for the operation of a number of applications.

#### 5.4.2. Requirements

Protocol designers SHOULD analyze the impact of accumulated jitter when proposing mechanisms that rely on microsleeps in either equipment or links. This analysis SHOULD include both worst case and best case scenarios, as well as an analysis of how coordinated clocks are to be handled in the case of coordinated sleep states.

#### 5.5. Other Operational Aspects

Modification of the network topology in order to save energy needs to consider the operational needs of the network as well as application requirements. Change management, operational downtime, and business usage of the network need to be considered when determining which links and nodes should be placed into a low energy state. Energy provisions have to be assigned and changed for nodes and links, optimally according to network usage profiles over the time of day.

Control plane protocol operation, in terms of operational efficiency on the wire, also needs to be considered when modifying protocol parameters. Any changes that negatively impact the operation of the protocol, in terms of the amount of traffic, the size of routing information transmitted over the network, and interaction with network management operations need to be carefully analyzed for scaling and operational implications.

##### 5.5.1. An Example of Operational Impact

Time of day is an important consideration in business operations. During normal operational hours, the network needs to be fully available, including all available redundancy and bandwidth. During holidays, night hours, and other times when a campus might not be used, or when there are lower traffic and resiliency demands on the network, network elements can be removed to reduce energy usage.

##### 5.5.2. Requirements

Protocol designers SHOULD analyze operational requirements, such as time of day and network traffic load considerations, and explain how proposed protocols or modifications to protocols will interact with these types of requirements. Protocol designers SHOULD analyze increases in network traffic and the operational efficiency impact of proposed changes or protocols.

#### 6. Security Considerations

None.

## 7. Acknowledgements

The authors of this document would like to acknowledge the suggestions and ideas provided by Sujata Banerjee, Puneet Sharma and Dirk Von Hugo.

## 8. References

### 8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

### 8.2. Informative References

- [3GPP] 3GPP, "3GPP TR 25-927 Solutions for energy saving within UTRA Node B", 2011,  
<<http://3gpp.org/ftp/Specs/html-info/25927.htm>>.
- [EMAN] IETF, "Energy Management Working Group Charter", 2012,  
<<http://datatracker.ietf.org/wg/eman/charter/>>.
- [Krioukov] Krioukov, D., "On Compact Routing for the Internet", 2007,  
<[http://www.caida.org/publications/papers/2007/compact\\_routing/](http://www.caida.org/publications/papers/2007/compact_routing/)>.
- [MANET] IETF, "Mobile Ad Hoc Networks Charter", 2012,  
<<http://datatracker.ietf.org/wg/manet/charter/>>.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.
- [ROLL] IETF, "Routing Over Low power and Lossy networks Charter", 2012,  
<<http://datatracker.ietf.org/wg/roll/charter/>>.

## Appendix A. Change Log

### A.1. Changes between the -00 and -01 versions.

- o Updated authors' contact information.
- o Modified some of the rfc2119 keywords.

A.2. Changes between the -01 and -02 versions.

- o Updated authors' contact information.

A.3. Changes between the -02 and -03 versions.

- o Updated authors' contact information.

Authors' Addresses

Alvaro Retana  
Cisco Systems, Inc.  
7025 Kit Creek Rd.  
Raleigh, NC 27709  
USA

Email: aretana@cisco.com

Russ White  
Ericsson

Email: russw@riw.us

Manuel Paul  
Deutsche Telekom AG  
Winterfeldtstr. 21-27  
Berlin 10781  
Germany

Email: Manuel.Paul@telekom.de



Internet Engineering Task Force  
Internet-Draft  
Intended status: Informational  
Expires: April 18, 2014

B. Zhang  
J. Shi  
Univ. of Arizona  
J. Dong  
M. Zhang  
Huawei  
M. Boucadair  
France Telecom  
October 15, 2013

Power-Aware Networks (PANET): Problem Statement  
draft-zhang-panet-problem-statement-03

Abstract

Energy consumption of network infrastructures is growing fast due to exponential growth of data traffic and the deployment of increasingly powerful equipment. There are emerging needs for power-aware routing and traffic engineering, which adapt routing paths to traffic load in order to reduce energy consumption network-wide. This document outlines the design space and problem areas for potential IETF work.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 29, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

|                                            |   |
|--------------------------------------------|---|
| 1. Introduction . . . . .                  | 3 |
| 2. Motivation and Problem Scope . . . . .  | 3 |
| 3. Potential Solution Approaches . . . . . | 4 |
| 4. Problem Areas for IETF . . . . .        | 6 |
| 5. Security Considerations . . . . .       | 7 |
| 6. Informative References . . . . .        | 7 |
| Authors' Addresses . . . . .               | 9 |



## 1. Introduction

Driven by exponential growth of Internet traffic, networks worldwide are expanding their infrastructures at a fast pace by deploying more high-capacity, power-hungry routers, which also leads to increasing energy consumption. For example, in the US, the energy bill for powering the wired network reaches up to 2.4 billion dollars per year [Doverspike10]. Telecom Italia, the largest ISP in Italy, is now the second largest consumer of electricity after the National Railway system [Pileri07]. As one of the biggest energy consumers in the United Kingdom, British Telecom consumed about 0.7% of the entire nation's electricity in 2007 [Bollal1]. In Japan, predictions say that routers will consume 9% of the total electricity by 2015 [Nakamura07]. Besides operational costs and environmental impacts, the ever-increasing energy consumption has become a limiting factor to long-term growth of network infrastructure due to challenges in power delivery and heat removal for both router components and hosting facilities [Gupta03] [Epps06].

Traditionally energy efficiency is improved at the device level or the link level. For example, energy management techniques can be applied to adjust router CPU's power status or CPU frequency in response to different CPU workload; Links can be put to sleep mode when it has been idle for a while. More recently, there have been a number of research work that look beyond a single router or linecard for network-wide solutions towards energy proportionality.

The purpose of this document is to discuss the problem scope, outline potential approaches, and problem areas for IETF work on power-aware networks.

## 2. Motivation and Problem Scope

Today's ISP networks have redundant routers and links, over-provisioned link capacity, and load-balancing traffic engineering. As a result, routers and links operate at full capacity all the time with low average usage, typically less than 40% of link utilization. This practice makes networks resilient to traffic spikes and component failures, but also makes networks far from energy-efficient.

Power-aware routing and traffic engineering have been proposed to improve network's energy efficiency, for example, by aggregating traffic onto a subset of links and putting other links with no traffic into sleep. Data from various sources (e.g., [Heddeghem12] [Chabarek08]) have shown that line cards are a significant source of router's power consumption, accounting for 40% - 70% of total power consumption. Most of the energy is consumed even in standby state,

and forwarding packets at full speed only increases the energy consumption by a small percentage. This implies that being able to put links into sleep mode can potentially save a lot of energy. In face, this has been demonstrated in several research works such as [GreenTE] [Nedevschi08] [Chabarek08].

Designing practical protocols, however, has been challenging, because making routing protocols power-aware brings significant changes to the routing system and the entire network, thus it involves hardware support, protocol design, network monitoring, and operational practices. These issues often depend on the specific network environments under discussion. In order to focus on protocol-related issues, we suggest that as the first step we limit the scope of the discussion to intra-domain routing within one administrative domain, to avoid inter-domain policy issues. This includes transit networks as well as edge networks. We leave data center networks out of this draft since that usually requires concerted efforts beyond network protocols.

### 3. Potential Solution Approaches

The high-level idea of power-aware networks is to adjust routing paths based on traffic level. When traffic level is high, use more links to carry the traffic; when traffic level is low, merge traffic onto a subset of all links so that other links can be put to sleep or reduce rate in order to save power. This needs to be done without significantly impacting network QoS, network resiliency, and interoperation with other protocols.

In the last few years a number of power-aware network designs have emerged. Instead of listing them individually, here we categorize the solutions along three different dimensions.

#### Link Sleep vs. Rate Adaptation

Sleeping and rate adaptation are two major ways to save energy in computer systems. Many hardware, including line cards and chassis, consumes a significant amount of power when they stand by without doing any actual work. When put into sleep mode, they will consume only a little power. Thus putting an idle component to sleep is a common way to save energy. If there is a need to use this component, it can be waken up and become usable after a transition time. The longer a component is in sleep mode, the more power saved. A power-aware protocol adjusts routing paths to increase the sleep time for certain links in the network.

A network interface often supports multiple data rates. Operating at a lower data rate usually consumes less energy, though the actual

rate-power curve varies from device to device. Rate-adaptation-based approaches operate interfaces at lower data rates when the traffic demand is low and increase the data rate when traffic demand is high. Thus the routers can save power during low utilization period.

These two approaches are also related in the case of "bundled links" [Fisher10]. A bundled link is a virtual link comprised of multiple physical links. A sleep-based approach can put some physical links into sleep to save power, which is same as conducting rate adaptation on the virtual link with adjustment unit of a physical link.

#### Configured vs. Adaptive

The key in power-aware routing and traffic engineering is to adjust routing paths in response to traffic changes, so that the power state of routers (or router components) will also change accordingly to achieve energy saving. Different approaches differ at the granularity of the adjustment.

Some approaches take the long-term traffic average as input, and output a routing configuration that is applied to the network regardless of short-term traffic variation. This is mostly useful when network traffic exhibits a stable, clear pattern, e.g., diurnal pattern where traffic is high during work hours and low during off hours. It can only exploit the target traffic pattern; it cannot react dynamically to short-term traffic changes to either save energy (by putting links to sleep) or avoid congestion (by waking links up), but the design and implementation should be simple.

Another type of approach is to adapt to traffic changes dynamically on much smaller time granularity. This approach may be able to save more energy and have better performance because it is more responsive, but the design and implementation usually are more complicated. This approach needs to continuously collect traffic data in order to adjust routing dynamically. The adjustment may be done periodically or whenever significant traffic changes are observed.

#### Distributed vs. Centralized

In distributed solutions, routers make power-aware adjustment decisions, such as link sleep/wake-up and rate increase/decrease, locally without a central controller. These routers need to exchange information in order to achieve consistent network states. Distributed approach fits the Internet operation model well but its design is the most challenging. Traditional routing does not respond to traffic variation while power-aware routing does, and it needs to do so without causing loops or congestions.

In centralized solutions, a controller computes the routing paths considering the network topology and traffic demand, and informs routers how to adjust their routing paths. A centralized server usually has more complete information, more computation power, and more memory and storage than routers, thus it may make better decisions than distributed approach. The server locates in the network NOC and can be backed up by server replicas. Nevertheless, this approach requires high reliability of the server.

Both distributed and centralized solutions may find their places in ISP networks. For example, centralized solution can be integrated into the Path Computation Element (PCE) framework [PCE-WG]. There can also be hybrid designs, e.g., using a centralized solution based on long-term traffic pattern, and distributed mechanisms to handle short-term traffic variations.

#### 4. Problem Areas for IETF

Power-aware networks have great potentials to improve network energy efficiency while maintaining network services at desired levels. Its effectiveness, however, depends on various supports from hardware and software, especially protocol designs that address operational issues. In this section we list a few problem areas that will benefit from additional input from the IETF community, or have the potential to become work items in related IETF working groups.

##### Motivation and Problem Scope

- o What are the motivations for Power-Aware Networking (PANET)?
- o To what extent power consumption is a key factor for Internet scaling?
- o To what extent power-aware system at router level and link level are not sufficient to reduce the overall energy consumption of networks?

##### Technical Development

- o What are the technical requirements for an efficient PANET solution?
- o What are the technical tracks to reduce the overall power consumption at the level of an IP network?
- o How protocols can be designed to be power-aware and still maintain enough network resiliency?

- o What are the technical challenges for deploying efficient PANET solutions?
- o How routing protocols (e.g., OSPF) can be extended to disseminate power-related information?
- o How PCE architecture can be used to compute power-aware paths?
- o How PANET can be deployed in centralized or in distributed model?

#### Operation Practice

- o What will be the impacts of PANET to network operations?
- o What will be the guidelines for deploying PANET systems?

#### 5. Security Considerations

This draft is a discussion on the Internet's necessity to follow an evolutionary path towards the future. There is no direct impact on the Internet security.

#### 6. Informative References

- [Bollal1] Bolla, R. and et al. , "Energy Efficiency in the Future Internet: A Survey of Existing Approaches and Trends in Energy-Aware Fixed Network Infrastructures", IEEE Communications Surveys and Tutorials, 2011.
- [Chabarek08] Chabarek, J. and et al. , "Power Awareness in Network Design and Routing", IEEE INFOCOM 2008.
- [Doverspike10] Doverspike, R., Ramakrishnan, K., and C. Chas, "Structural overview of ISP networks", Guide to Reliable Internet Services and Applications, Springer, 2010.
- [EMAN-WG] "IETF Energy Management Working Group", 2012, <<https://datatracker.ietf.org/wg/eman/>>.
- [Epps06] Epps, G. and et al. , "System Power Challenges", 2006, <[http://www.slidefinder.net/c/cisco routing research/ seminar august 29/1562106](http://www.slidefinder.net/c/cisco%20routing%20research/seminar%20august%2029/1562106)>.
- [Fisher10] Fisher, W. and et al. , "Greening Backbone Networks: Reducing Energy Consumption by Shutting Off Cables in Bundled Links", Green Networking 2010.
- [GreenTE] Zhang, M. and et al. , "GreenTE: Power-Aware Traffic

Engineering", ICNP 2010.

[Gupta03] Gupta, M. and S. Singh, "Greening the Internet", ACM SIGCOMM 2003.

[Heddeghem12] Van Heddeghem, W. and F. Idzikowski, "Equipment power consumption in optical multilayer networks - source data", IBCN Technical Report 2012.

[Nakamura07] Nakamura, M., "Advanced photonic technologies for the information era", Nature Photonics Technology conference, 2007.

[Nedevschi08] Nedevschi, S. and et al. , "Reducing Network Energy Consumption via Sleeping and Rate- Adaptation", USENIX NSDI 2008.

[PCE-WG] "IETF Path Computation Element Working Group", 2012, <<https://datatracker.ietf.org/wg/pce/>>.

[Pileri07] Pileri, S., "Energy and communication: engine of the human progress", 2007.

[TM] Roughan, M., Thorup, M., and Y. Zhang, "Traffic Engineering with Estimated Traffic Matrices", IMC 2003.

Authors' Addresses

Beichuan Zhang  
Univ. of Arizona

Email: bzhang@cs.arizona.edu

Junxiao Shi  
Univ. of Arizona

Email: shijunxiao@cs.arizona.edu

Jie Dong  
Huawei

Email: jie.dong@huawei.com

Mingui Zhang  
Huawei

Email: zhangmingui@huawei.com

Mohamed Boucadair  
France Telecom

Email: mohamed.boucadair@orange.com

INTERNET-DRAFT  
Intended Status: Proposed Standard  
Expires: April 15, 2014

Mingui Zhang  
Jie Dong  
Huawei  
Beichuan Zhang  
The University of Arizona  
Bithika Khargharia  
Extreme Networks  
October 12, 2013

Use Cases for Power-Aware Networks  
draft-zhang-panet-use-cases-03.txt

Abstract

Power Aware Network (PANET) has attracted strong interest from both carriers and vendors. Several use cases are investigated in this document to exhibit the potential usage of PANET in both backbone and data center networks.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal



Provisions Relating to IETF Documents  
(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

|                                                                 |    |
|-----------------------------------------------------------------|----|
| 1. Introduction . . . . .                                       | 3  |
| 1.1. Conventions used in this document . . . . .                | 3  |
| 1.2. Terminology . . . . .                                      | 3  |
| 2. Power Awareness in Backbone Networks . . . . .               | 3  |
| 2.1. Use Case 1: Sleeping Links . . . . .                       | 4  |
| 2.1.1 Aware of Sleeping Links at the Management Plane . . . . . | 5  |
| 2.1.2 Gathering Information for Decision Making . . . . .       | 6  |
| 2.2. Use Case 2: Composite Links . . . . .                      | 7  |
| 2.3. Coordinating L2 and L3 Sleeping Links . . . . .            | 8  |
| 3. Power Aware in Data Center Networks . . . . .                | 9  |
| 3.1. Use Case 3: Server Consolidation . . . . .                 | 9  |
| 3.2. Use Case 4: Elastic Infrastructure . . . . .               | 11 |
| 3.3. Use Case 5: Job Scheduling Among Multiple Sites . . . . .  | 12 |
| 6. Security Considerations . . . . .                            | 13 |
| 7. Summary . . . . .                                            | 13 |
| 8. IANA Considerations . . . . .                                | 13 |
| 9. References . . . . .                                         | 13 |
| 9.1. Normative References . . . . .                             | 13 |
| 9.2. Informative References . . . . .                           | 14 |
| Author's Addresses . . . . .                                    | 16 |

## 1. Introduction

Networks are usually provisioned for peak hours and potential network failures. Network devices are powered on all the time without consideration on energy efficient. In practice, however, the traffic load of a network is low most of the time and redundant network equipments are used for failure recovery occasionally.

In the past years, vendors had paid a great effort on improving the network energy efficiency at the device level: when the traffic load is low, a network equipment should accordingly operate with less power draw. However, network equipments have never become fully power proportional. Even few or no traffic is carried, a powered-on network device draws a considerable amount of power, which means energy is being wasted. There is an explicit gap that idle network devices are shut down or put into sleeping state to save more energy. In order to fill this gap, the network control plane and management system should become power aware (i.e., Power Aware NETWORK, PANET) to coordinate network devices therefore the sleeping or powered-off network devices do not bring service disruption to the network.

The design space and problem areas of PANET is outlined in [PANET-problem]. The requirements for PANET is given in [PANET-requirements]. This documents investigated use cases on PANET which include both backbone networks and data center networks. As for the energy efficiency of backbone networks, only intra-domain use cases are considered. Trying to be energy efficient in the inter-domain scale seems technically feasible. But currently, energy efficient solutions can easily end up with lack of business motivation. This document leaves them for future study.

### 1.1. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

### 1.2. Terminology

PANET: Power Aware NETWORK

## 2. Power Awareness in Backbone Networks

The IETF Energy Management (eman) Working Group works on the management of power-aware network devices. Basically, the power states of power-aware network devices are reported and recorded in MIB. However, there is a gap on how to make use of this kind of data to achieve energy efficient networks. With energy aware control plane

[power-control], it becomes possible to make use of these measurements and power control ability to achieve the energy efficiency of a whole network. This section lists several use cases for backbone networks.

Take a router system as an example, the start-up of it may take several minutes and the stabilization of it may take much longer time. It is unrealistic to switch off and on a whole node in backbone networks frequently to achieve energy efficiency, so this document only investigates the cases in which links (i.e., links' attached components) are shut-down or put into sleeping state for energy conservation.

### 2.1. Use Case 1: Sleeping Links

The power draw on line-cards occupies a great portion in the total power consumption of a whole routing system. For high-end routers, this portion may be higher than 50%.

Network devices and their processing capacity are provisioned for worst cases such as traffic burst and busy hours. Most of the time, the network is lightly loaded. Unfortunately, the power consumption of network devices is not proportional to the traffic load on them. An extreme case is that even there is no load on them, there is still a considerable base power consumption. Unlike personal PCs which can be shut down or enter power saving modes (such as sleeping), network devices are powered on and running even there is no load on them. This reality means that the network is wasting lots of power.

The conception that "a link is put into sleep state" is frequently mentioned in various technical documents. In this document, this conception is formalized as follows. The coupled end-points (such as interfaces, NPU or whole line-cards) attached to a idle link enter the sleeping mode to save energy. Similarly, the wake up of a link also means the wake up of those coupled end-points.

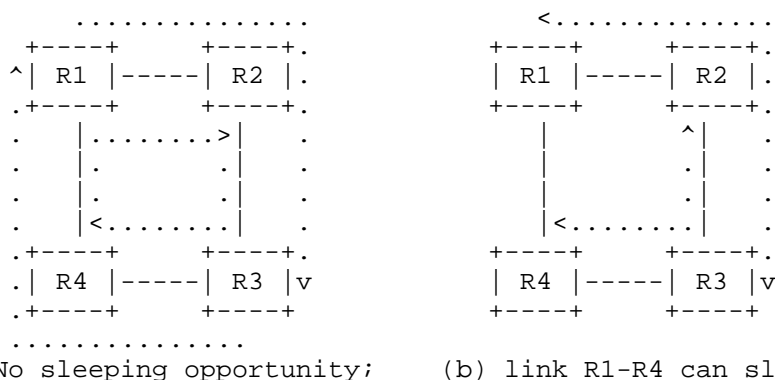


Figure 2.1: Aggregate traffic to create opportunities for link sleeping

Traffic aggregation are used to create the opportunity for more links to become idle. This process can be automated through the control plane [power-control]. Traffic Engineering technique is able to achieve this kind of traffic aggregation [GreenTE]. Take Figure 2.1 as an example, suppose R1, R2, R3 and R4 is sending traffic to R3, R4, R1 and R2 and R4 respectively. In Figure 2.1 (a), paths R1-R2-R3, R2-R3-R4, R3-R4-R1 and R4-R1-R2 are being used. All links are active. In Figure (b), paths R1-R2-R3, R2-R3-R4, R3-R2-R1 and R4-R3-R2 are being used. Link R1-R4 is idle, therefore it can be put into sleeping state to save energy.

Different from traditional Traffic Engineering techniques which aim to balance traffic load around the whole network to avoid hot spots, green Traffic Engineering try to create more idle links which can be put into sleeping state. However, this kind of traffic aggregation should be restricted. At least, green Traffic Engineering SHOULD NOT achieve energy conservation at the expense of apparently downgrade the network performance. It is recommended that the QoS metric should be take into consideration as constraints of green Traffic Engineering. The traffic aggregation which can violate these constraints should be avoid.

With the traffic load fluctuating, the green Traffic Engineering should be periodically performed. When the network is lightly loaded, the GreenTE should be able to put more links to be idle. When the network is heavily loaded, GreenTE should remain more links in active state to absorb more traffic in order to avoid traffic jam.

#### 2.1.1 Aware of Sleeping Links at the Management Plane

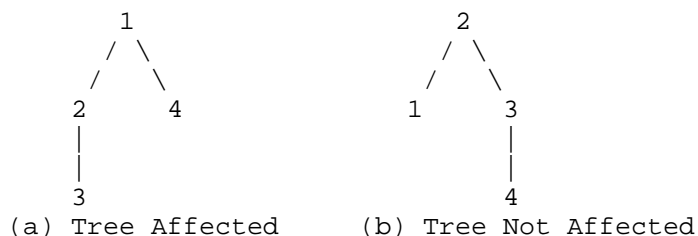


Figure 2.2: Multicast Forwarding Plane

It's possible to wake up a link or put it into sleeping state at the management plane by operators. However, the state change impacts the data plane of the network. Take Figure 2.1 as an example, if link R1-R4 is sleeping, it will cut off the path R3-R4-R1 and R4-R1-R2. Therefore, the paths in Figure 2.1 (b) should be used. Data plane for multicast traffic may also be affected. Take Figure 2.2 as an example, in order to avoid being affected by the sleeping link, the multicast tree in Figure 2.2 (b) should be used rather than that in Figure 2.2 (a). It requires a knob to achieve the adjustment of data plane via the control plane. Before a link is to be put in to sleep state, operators may increase its link metric to disperse the traffic load on it. This will cause the update of FIBs of routers [RFC6976]. The mechanism defined in [RFC6976] can be adopted.

Sleeping links are different from failed links since they can be waken up to relieve the traffic jam when it becomes necessary. This difference creates the necessity for the network to remember these links in order to make decision to wake them up at a proper time.

#### 2.1.2 Gathering Information for Decision Making

The decision of the Traffic Engineering may be centrally made on a NOC (Network Operation Center) or distributedly carried out by each router. However, when decisions are distributedly made, the consistence of decisions MUST be guaranteed. It requires that the algorithm of the Traffic Engineering always generate the same result.

Where ever the decision point locates, the traffic demand of the network is the necessary input for Traffic Engineering. This information should be measured and maintained. For example, network elements (routers and switches) can gather flow data and export it to the collectors using Netflow [RFC3954]. For another example, a cyclic number of bytes transmitted and received on each of those interfaces of a line-card is maintained in the Management Information Base (MIB) via Simple Network Management Protocol (SNMP). The Network Management System (NMS) may periodically poll these numbers to compute the

links' load. The traffic matrix of the network can be further estimated from those links' load [TMEstimated].

Compared to traditional Traffic Engineering, power aware Traffic Engineering additionally requires the information about the power consumption of network devices of the network. It is requires a MIB module for monitoring energy consumption and power states of energy-aware devices [eman].

The essentials of this use case:

- o Devices to be Power Aware: Routers, NOC, etc.
- o What actions to take: NMS (Network Management System) polls the traffic load and power consumption profile of each link; Routers execute the green TE algorithm; Routers send out signals to trigger the sleeping/wake-up transition of a NPU on a line card.

## 2.2. Use Case 2: Composite Links

A composite link is a logical link composed of multiple physical [I-D.ietf-rtgwg-cl-requirement] links. The composite link attached end-points are responsible to map traffic onto the component links and maintain the state of the composite link. Power awareness can be applied to composite links as well. When the traffic volume on a composite link is low, some component links can be shut down to conserve energy consumption. When the traffic volume becomes high, the sleeping component links can be waken up to absorb the traffic load.

Compared to use case 1, the advantage of executing energy saving for composite link is that the connectivity of the composite link does not suffer unless all the component links are sleeping. In this way, the control plane of the component link is not disrupted. When the end points of the composite link execute the energy conservation action, they can do it in a distributed way and decisions are made locally.

The essentials of this use case:

- o Devices to be Power Aware: Composite links attached end-points.
- o What actions to take: NMS measures the traffic load and power profile of component links; Attached end-points adaptively put component links into sleeping state or wake them up according to the traffic load on the composite link.

Use case 1 and use case 2 may be combined in a real network to

achieve more energy saving.

### 2.3. Coordinating L2 and L3 Sleeping Links

Networks devices are usually redundantly provided. However, they may spend a lot of time operating at its full rate while caring few or no traffic, especially at the edge of the network. Manufacturers have put a lot of effort to produce energy efficient network devices. However, fully power proportional network devices are hard, even impossible, to be implemented. Take switches of the day for example, when they are left idle, there is only a low teens of power reduction compared to the peak power [Sx700]. This means a considerable amount of energy is being wasted when no traffic is being delivered. Modern processors support a number of states, which enables various network components to sleep [C-Sleep]. When a network component becomes idle, it's reasonable for them to enter the sleep mode to save energy.

Energy Efficient Ethernet (EEE) provides a mechanism and standard for Ethernet links to operate in an energy-efficient way [802.3az]. The PHY connecting an Ethernet link can enter Low Power Idle mode (i.e., sleep mode) to reduce the energy consumption when no data packet is being sent on the link. Normally, PHY need be refreshed periodically during the LPI. When the signaling protocol indicates the PHY to wake up, it SHOULD resume in a pre-defined delay (Time to Wake,  $T_w$ ). This pre-defined delay is the time the transmitter transmitting a maximum length Ethernet frame. For example,  $T_w$  for 1000BASE-T is 16.5uS, which is the same time that it takes to transmit a 2000-byte Ethernet frame [C.I.EEE]. The transmitter can buffer packets during the time that the receiver transiting from the LPI mode to the active mode. This Wake on Arrival mechanism guarantees that EEE does not interfere with upper layer application protocols (e.g., ISIS). Beyond PHY, other network equipments (such as NPUs, line-cards, or whole switches/routers) may also enter sleep mode, which usually means a much longer time to wake up but a much higher energy saving. The EEE standard also supports PHYs to enter a deep sleep mode through negotiating a larger value of  $T_w$  using the link-layer discovery protocol (LLDP) [802.1ab].

Energy Efficient Ethernet (EEE) protocol makes use of gaps in the data stream to put transceivers attached to an Ethernet link into sleep state to save energy consumption [802.3az]. Without expectation of the arrival of packets, the L2 sleeping is actually an opportunistic sleeping. This kind of opportunistic sleeping is performed locally without coordination of nodes across the network wide. At L3, we can plan the sleeping which help to achieve energy savings beyond physical layer transceiver (PHY) of Ethernet links.

With traffic aggregation realized through network-wide green routing

and traffic engineering, idle links of an L3 network can be scheduled to enter sleep mode for a while to save energy consumption. If a link is scheduled to sleep at L3, no packets will be sent on this link for a longer time (minutes or hours) [GreenTE]. It's feasible for this link to negotiate a longer value of  $T_w$  at L2 and enter the deep sleep mode to save more energy. Compared to a normal L2 link, the scheduled sleeping link has a longer LPI and less wake up actions, which also leads to more energy savings.

On the other hand, if a link is scheduled to be active at L3, the opportunity for this link to enter LPI mode may become slim. It may be better to disable the sleep mode of this link at L2 to avoid the frequent transitioning between LPI and active mode, therefore avoids extra energy consumption and instability of the network.

### 3. Power Aware in Data Center Networks

Servers and network devices (ICT equipments) are intensively placed in Data Centers. In comparison with ISP backbone networks, the operating of Data Center Networks are more power hungry. The growing amount of energy consumed by a Data Center has led to high operating costs. The work load of a data center varies due to the effect of "follow-the-sun". There is a significant opportunity to conserve the energy consumption through right-sizing the ICT infrastructure when the work load is low.

Although non-ICT equipments, such as lighting and air conditioners, in a Data Center consumes a notable large amount of energy as well, this section concentrate on talking about right-sizing the ICT infrastructure for energy conservation. Energy conservation of non-ICT equipments are out of the scope of this document.

#### 3.1. Use Case 3: Server Consolidation



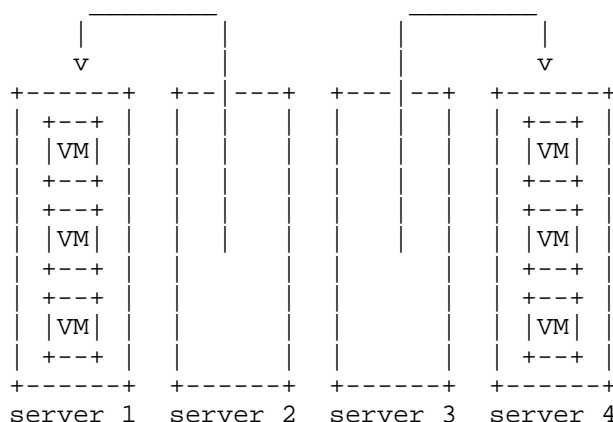


Figure 3.1: VM Placement and Consolidation

With Operating System virtualization, one physical server can provide tens of Virtual Machines (VM) at the same time. Network Virtualization technology makes the placement and migration easy for VMs [nvo3usecase]. With live migration, VM can change its host server (location) without disruption of services running on it [VMmove]. As shown in Figure 3.1, VMs migrate out from server 2, server 3 to server 1, server 4 respectively. When Server 2 and server 3 are idled, they can be put into sleep state to save energy consumption.

With virtualization technology, VMs of a Data Center or multiple Data Centers can be consolidated to fewer physical servers while idled servers can be put into power saving mode or turned off to achieve energy conservation. Virtualization technology allows the administration of a Data Center Network respond rapidly to the fluctuating capacity requirements.

Through monitoring of the work load and power profile, the Data Center Network Management System (Orchestrator) can judge in which hours workload is high and in which hours workload is low. For example, nights are generally off-peak hours in which workload is at low level. Virtual machines can be moved to fewer servers therefore idle servers can be powered off or put into sleep to save energy. Before peak hours (e.g., in the morning), sleeping or powered off servers should be waken up to accommodate more active virtual machines (VMs).

The essentials of this use case:

- o Devices to be Power Aware: All servers in a data center.

- o What actions to take: NMS measures the work load and power profile of servers; The orchestrator of a Data Center Network adaptively triggers the actions of VM migration, the power-off and power-on of servers according to the workload.

### 3.2. Use Case 4: Elastic Infrastructure

Traffic load of a data center is generated by the work load on servers and applied on the network infrastructure. The changing work load determines that the traffic load varies as time goes on. However, network devices are always powered on even though the traffic load fluctuates, which wastes energy inevitably when the traffic load is low.

Ideally, the network infrastructure is elastic and can fit the traffic pattern with minimum subset to minimize the energy consumption of the network infrastructure. For now, Data Center Networks generally work at layer 2. So this use case should be realized through manipulating switching paths, in comparison with the power aware routing at layer 3. Openflow switches may be utilized to achieve this goal [ElasticTree].

The essentials of this use case:

- o Devices to be Power Aware: All network equipments in a data center.
- o What actions to take: Network devices report their traffic load and power consumption profile to the NMS. The orchestrator (NMS) of a Data Center Network adaptively build the switching paths upon the network infrastructure. The idled links are put into power saving mode (e.g., sleeping), so that the network infrastructure becomes energy efficient.

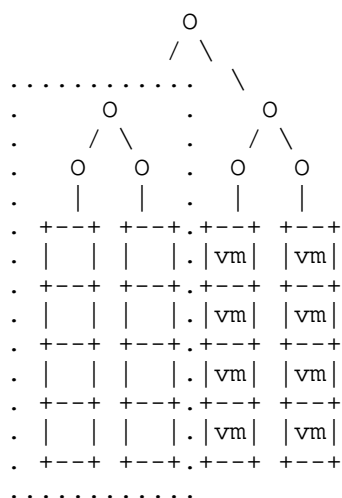


Figure 3.2: Elastic Data Center ICT Infrastructure

As shown in Figure 3.2, when servers are idled and put into sleeping state, their up connected switches may also become idle. Therefore, use case 3 and use case 4 can be combined to achieve more energy saving.

### 3.3. Use Case 5: Job Scheduling Among Multiple Sites

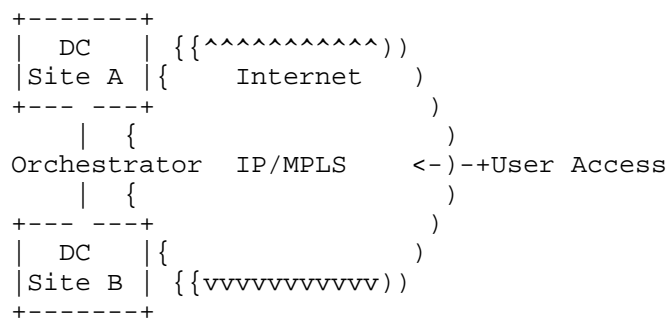


Figure 3.3: Adaptive Job Scheduling

An cloud service provider may have multiple data centers which spread out in different geographic locations. Generally, the ICT resources in these data centers are well replicated and a job can be directed to any of them for execution. These data centers form a large distributed Internet scale systems and the price of power supply for them varies between two different locations. The operating cost of such a system highly depends on the load scheduling scheme. Being

power aware, the system can map requests to locations where energy price is cheaper.

This use case makes use of the difference of the prices of power draw in different locations. The orchestrator of data centers (the NMS) is responsible for monitoring the power profile and work load of the ICT devices located in different data centers, and adaptively schedule the jobs among these data centers. Orchestration protocols are necessary to support the job scheduling [Orchestration].

As shown in Figure 3.3, the orchestrator map the user request (job) to with an economic attachment to either Site A or Site B, while the user is unaware of the executing point of the job. In this way, the job scheduling becomes a trade-off between OPEX and performance. The orchestrator should guaranteed that there is no apparent service performance degradation. Complex SLA fulfillment may be designed to embody the response time, throughput, latency, etc.

The essentials of this use case:

- o Devices to be Power Aware: All ICT-equipments in a data center.
- o What actions to take: ICT devices report their work load and power consumption profile to NMS. The orchestrator (NMS) of the Data Center Networks adaptively map the request onto sites in consideration of reducing the overall power bill of the system.

## 6. Security Considerations

This document raises no new security issues.

## 7. Summary

The document describes some basic potential use cases of Power Aware Network.

## 8. IANA Considerations

No new registry is requested to be assigned by IANA. RFC Editor: please remove this section before publication.

## 9. References

### 9.1. Normative References

[PANET-problem] B. Zhang, J. Shi, J. Dong and M. Zhang, "draft-zhang-panet-problem-statement-02.txt", work in progress.

- [PANET-requirements] J. Dong, M. Zhang and B. Zhang, "draft-dong-panet-requirement-02.txt", work in progress.
- [power-control] A. Retana, R. White, M. Paul, "A Framework and Requirements for Energy Aware Control Planes", draft-retana-rtgwg-eacp-01.txt, work in progress.
- [TMEstimate] Y. Zhang, M. Roughan, N. Duffield and A. Greenberg, "Fast Accurate Computation of Large-Scale IP Traffic Matrices from Link Loads", SIGMETRICS 2003.
- [RFC3954] Siemborski, R., Ed., and A. Melnikov, Ed., "SMTP Service Extension for Authentication", RFC 4954, July 2007.
- [eman] IETF, "Energy Management Working Group Charter", 2012, <<http://datatracker.ietf.org/wg/eman/charter/>>.
- [I-D.ietf-rtgwg-cl-requirement] Villamizar, C., McDysan, D., Ning, S., Malis, A., and L. Yong, "Requirements for MPLS Over a Composite Link", draft-ietf-rtgwg-cl-requirement-11.txt, work in progress.
- [Orchestration] Dalela, A. and M. Hammer, "Service Orchestration Protocol (SOP) Requirements", draft-dalela-orchestration-00.txt, work in progress.
- [oFIB] M. Shand, S. Bryant, S. Previdi, C. Filsfils, P. Francois, O. Bonaventure, "Framework for Loop-Free Convergence Using the Ordered Forwarding Information Base (oFIB) Approach", RFC 6976, July 2013.

## 9.2. Informative References

- [GreenTE] Zhang, M. and et al. , "GreenTE: Power-Aware Traffic Engineering", ICNP 2010.
- [Sx700] "Huawei Enterprise Sx700 Series Switch Product", 2013. [http://enterprise.huawei.com/ilink/enenterprise/download/HW\\_200802](http://enterprise.huawei.com/ilink/enenterprise/download/HW_200802).
- [C-Sleep] "Power and Thermal Management in the Intel Core Duo Processor". Intel Technology May, 15, 2006.
- [C.I.EEE] "IEEE 802.3az Energy Efficient Ethernet: Build Greener Networks", 2011. [http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps4324/white\\_paper\\_c11-676336.pdf](http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps4324/white_paper_c11-676336.pdf).

- [802.1ab] IEEE P802.1ab, "Station and Media Access Control Connectivity Discovery", 2005.  
<http://www.ieee802.org/1/pages/802.1ab.html>
- [802.3az] IEEE P802.3az, "Energy Efficient Ethernet Task Force", 2010. <http://grouper.ieee.org/groups/802/3/az>.
- [Rate-Adaptation] S. Nedevschi, L. Popa, G. Iannaccone, S. Ratnasamy, and D. Wetherall, "Reducing Network Energy Consumption via Sleeping and Rate-Adaptation," in Proceedings of USENIX NSDI, 2008.
- [nvo3usecase] L. Yong, M. Toy, and et al., "Use Cases for DC Network Virtualization Overlays", draft-ietf-nvo3-use-case-02.txt, work in progress.
- [VMmove] Y. Rekhter, W. Henderickx and et al, "Network-related VM Mobility Issues", draft-ietf-nvo3-vm-mobility-issues-01.txt, work in progress.
- [ElasticTree] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, "ElasticTree: Saving Energy in Data Center Networks," in Proceedings of USENIX NSDI, 2010.

## Author's Addresses

Mingui Zhang  
Huawei Technologies Co.,Ltd  
Huawei Building, No.156 Beiqing Rd.  
Beijing 100095 P.R. China

Email: zhangmingui@huawei.com

Jie Dong  
Huawei Technologies Co.,Ltd  
Huawei Building, No.156 Beiqing Rd.  
Beijing 100095 P.R. China

Email: jie.dong@huawei.com

Beichuan Zhang  
The University of Arizona

Email: bzhang@cs.arizona.edu

Bithika Khargharia  
Extreme Networks

bithika@gmail.com