INTERNET-DRAFT
Intended Status: Experimental RFC
Expires: August 2013

Balaji Venkat Venkataswami
Bhargav Bhikkaji
DELL
Shankar Raman
IIT Madras

February 17, 2013

Load-balancing to Data Centers in a L3VPN environment based on Power
draft-balaji-panet-dc-label-semantic-for-pwr-00

Abstract

Data Centers may be spread across different locations for a particular enterprise. Different locations may mean within the same country but across different geographical locations, or outside the country even in a different continent. These data centers may be serving the enterprise or multiple enterprises / tenants wherein the regular enterprise site may request data from a data center site which could be one of the data center sites proximal to the enterprise site. Proximity is usually calculated based on a metric that is bandwidth driven or in terms with regard to the number of hops to reach that data center site hence bringing into play delay characteristics. Assume a topology where the data center sites and the enterprise sites are MPLS based L3VPN sites that are being provided connectivity through a Service Provider deploying Layer 3 VPNs. Given such a topology it is possible that replication of data happens across the data centers in a timely manner to keep the data active and refreshed across all data center sites. Suitable mechanisms for such replication will come into play for this purpose. Thus any of the data centers can cater to the request from a user site.

It is possible that power consumption in each data center may vary according to the load on each data center. It would be prudent to introduce a scheme where the power metric coupled with other metrics such as bandwidth and delay be used by a Provider Edge router in a L3VPN scenario to direct the packets or requests from regular user sites to such data centers with the least such metric. This is in line with the follow-the-moon strategy of directing requests for data and compute to data centers which are power-wise more efficient during the night or during the day. This draft document lays out one such proposal.

Status of this Memo

Copyright and License Notice

Table of Contents

1 Introduction

Data Centers may be spread across different locations for a particular enterprise. Different locations may mean within the same country but across different geographical locations, or outside the country even in a different continent. These data centers may be serving the enterprise or multiple enterprises / tenants wherein the regular enterprise site may request data from a data center site which could be one of the data center sites proximal to the enterprise site. Proximity is usually calculated based on a metric that is bandwidth driven or in terms with regard to the number of hops to reach that data center site hence bringing into play delay characteristics. Assume a topology where the data center sites and the enterprise sites are MPLS based L3VPN sites that are being provided connectivity through a Service Provider deploying Layer 3 VPNs. Given such a topology it is possible that replication of data happens across the data centers in a timely manner to keep the data active and refreshed across all data center sites. Suitable mechanisms for such replication will come into play for this purpose. Thus any such data center site can cater to a request from a user site.

It is possible that power consumption in each data center may vary according to the load on each data center. It would be prudent to introduce a scheme where the power metric coupled with other metrics such as bandwidth and delay be used by a Provider Edge router in a L3VPN scenario to direct the packets or requests from regular user sites to such data centers with the least such metric. This is in line with the follow-the-moon strategy of directing requests for data and compute to data centers which are power-wise more efficient during the night or during the day. This draft document lays out one such proposal.

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].

2.0 Reference Topology

Assume the following topology where Data Center 1 and 2 are geographically dispersed within the county or across different continents and there is sufficient desirable properties of delay and bandwidth allocated to each of them. End User Sites depicted below the CEs would be enterprise sites that would request data and compute from these Data center sites. The DC GWs connecting the Data centers

would play the role of CEs for the Data Centers. The PEs (1-5) are the Provider Edge routers of the ISP Core network that provide regular MPLS based L3VPN services for the inter-connection of the Data Center sites with the End User sites.

During the normal course of events VPN instance labels would be exchanged between the PEs and the ISP core would provide LDP or RSVP based connectivity amongst these PEs. Thus a stack of labels with inner VPN instance label with the outer label being LDP or RSVP would be used to direct traffic from End User Sites to the Data Center sites and even amongst the End User Sites themselves. It is also possible that replication services would run between the Data Center sites as well using this mechanism.

```
                            __________                 ,---------.
                           /                         ,'           `.
                          ;Data Center)              (  Data Center )
                          (    2      '               `.    1      ,'
                          +-----------+                 `-+------+'
                                       \            /
                                     +--+--+   +-+---+
                                     |DC GW|   |DC GW|
                                     +-+---+   +-----+
                                         |       |
                                        PE5     PE4
                                        .--. .--.
                                      (    '    '.--.
                                    .-.' ISP Core     '
                                   (     network      )
                                    (             .'-'
                                     '--'._.'.    )PE3
                                     PE1     '--'  \ \
                                    / /      PE2    \ \
                                   / /        |      \ \
                              +---+--+   +------+  +--+----+
                              | CE1  |   | CE2  |  |  CE3  |
                              +-+--`.+   +-+----+  +-+--+--+
                            __/_                \           \__
                     '--------'                '--------'   '--------'
                     :End User:                :End User:   :End User:
                     : Site   :                : Site   :   : Site   :
                     '--------'                '--------'   '--------'
```

Figure 1 : A Generic Architecture for Multiple Data Centers providing Data and Compute services for End user sites.

2.1 Methodology of the new scheme

When a site on the enterprise connected through a L3 VPN intends to access a data center, load-balancing mechanisms may use the nearest data center or that within the nearest country or even in another continent depending on the availability and load being serviced. The method proposed in this document would advertise a power consumption rating along with a L3VPN service instance label for that Data Center site and the enterprise site PE accessing the Data Center may arrive at a decision as to which DC it is to access based on the power consumption rating so advertised in the MP-iBGP update. Trust level between Provider and customer is advised in this case.

The PE devices would receive different VPN instance labels from each of the Data Centers using a MP-iBGP update with the compound power based metric in the attribute information in the update. Suitable extensions to extended community attributes would be done to facilitate the passage of such information. Assume Label 100 is sent from PE4 to the enterprise sites with Compound power metric 1200 and Label 200 is sent from the PE5 to the enterprise sites with Compound power metric 1400 at a certain point in time.

When the End User sites request a service from a data center, the PE (1-3) which have the labels 100 and 200 with their respective compound power metrices 1200 and 1400 respectively will choose which Data Center (1 or 2) has the least compound metric and direct the services towards that PE connected to the data center with that least power metric.

```
                            __________                ,---------.
                           /                        ,'           `.
                          ;Data Center)            (  Data Center )
                          (    2      '             `.    1      ,'
                          +-----------+               `-+------+'
                           pwr = 1400  \            /     pwr = 1200
                                     +--+--+   +-+---+
                                     |DC GW|   |DC GW|
                                     +-+---+   +-----+
                                         |       |
                    (200, pwr = 1400)   PE5     PE4 (100, pwr = 1200)
                                        .--. .--.
                                      (    '    '.--.
                                    .-.' ISP Core     '
                                   (     network      )
                                    (             .'-'
                                     '--'._.'.    )PE3
                                     PE1     '--'  \ \
                                    / /      PE2    \ \
                                   / /        |      \ \
                              +---+--+   +---+--+  +--+----+
                              | CE1  |   | CE2  |  |  CE3  |
                              +-+--`.+   +-+----+  +-+--+--+
                            __/_                \           \__
                     '--------'                 '--------'   '--------'
                     :End User:                 :End User:   :End User:
                     : Site   :                 : Site   :   : Site   :
                     '--------'                 '--------'   '--------'
```

Figure 2: Control plane exchanges between PE5, PE4 and PEs (1-3)

In this case Data Center 1 has the least Compound power metric 1200 and hence the traffic from say CE1 would be sent to Data Center 1. There is a periodic calculation of the compound power metric in each of the Data Center sites and this is exchanged with the PE through eBGP (say for example) between the respective CEs and the PEs. The VPN instance label used would be 100 to get to the Data Center 1 and the reachable PE would be set to PE4.

```
                            __________                 ,---------.
                           /                         ,'           `.
                          ;Data Center)              (  Data Center )
                          (    2      '               `.    1      ,'
                          +------------+                `-+------+'
                           pwr = 1400  \            /     pwr = 1200
                                     +--+--+   +-+---+
                                     |DC GW|   |DC GW|
                                     +-+---+   +-----+
                                         |       |
                                        PE5     PE4
                                        .--. .--.
                                      (    '    '.--.
                                    .-.' ISP Core     '
                                   (     network      )
                                    (             .'-'
                                     '--'._.'.    )PE3
                    (LDP Label, 100) PE1     '--'  \ \
                                    / /      PE2    \ \
                                   / /        |      \ \
                              +---+--+   +---+--+  +--+----+
                              | CE1  |   | CE2  |  |  CE3  |
                              +-+--`.+   +-+----+  +-+--+--+
                            __/_                \           \__
                     '--------'               '--------'   '--------'
                     :End User:               :End User:   :End User:
                     : Site   :               : Site   :   : Site   :
                     '--------'               '--------'   '--------'
```

Figure 3: Data Plane path from say CE1 to PE4 based on Power Metric for DC.

Here the power metric is not discrete but in intervals of thresholds. Only the threshold interval is thus exchanged between the CEs and the PEs. This way dampening frequent fluctuations or oscillations within a given power metric interval is taken care of.

Inter-DC connectivity and replication may also benefit from this scheme. DC A would choose to replicate with DC B at a time when the power consumption rating in both sites or in DC B is the lowest.

2.2 Calculation of the Compound Power Metric

Factors such as compute load, cooling power consumption and network bandwidth within the Data Center would be used to compute the compound power metric within a Data Center to be advertised to the PEs within the Layer 3 VPN. The actual calculation is out of scope of this document and there exists sufficient literature to suggest a

suitable method for such calculation. For now, this draft proposes a scheme for exchanging such power metrices and using them for load-balancing within a L3VPN scenario.

2.2.1 Power metric calculation for the DC as a whole and per tenant

Many schemes exist for calculating the power consumed per tenant based on the occupancy of each tenant in a DC and for calculating the power consumed by the DC as a whole. Appropriate labels can be exchanged per tenant with the respective power metrices factored in per tenant if necessary. The specific end user site that is using the data and compute power of a data center may belong to a particular tenant and may wish to direct its traffic to the data center based on the power consumed for that specific tenant Identifier and hence use the appropriate label for the same in the PE. If the Data Center cater to a single tenant and are owned by the tenant then the overall power consumed by the DC will be used in the MP-iBGP update.

2.3 Extending the labeling scheme to the CE

It is possible to extend the label imposition from the CE itself towards the ISP core in order to ensure that the CE can be made aware of such a scheme being available and use appropriate labels to indicate to the PEs that the CE requires power metric based load-balancing. Two different labels could be used by the CEs one for conventional methods of requesting services and the other the power metric based method where the PEs would consult the power metrices available and direct the request towards the low power consuming data center.

If trust levels are not to be adhered to the label may be propogated along with the power consumption ratio to the CE and the CE would make the appropriate decision.

2.4 Extensions to MP-iBGP for this scheme.

A future version of the draft will outline the actual extensions to the BGP protocol and its attributes with regard to how the compound power metric is carried in the actual BGP exchanges.

3 Security Considerations

Trust levels between the Provider Edge and the customer edge should be proper in order that the Data Center's power metrices are exchanged between the PE and CE. eBGP as a PE-CE protocol could be adhered to for this purpose. Appropriate security mechanisms would have to be taken into account if the Data Center is serving multiple tenants. The computed Compound Power Metric may be calculated for each tenant and mechanisms should be adopted that one tenant's compound metric is not shared with other tenants. Appropriate label exchanges with each tenant's Label information and corresponding power metrices should be done with such separation in mind. If there is a collated power metric for all the tenants put together then the PE device should make sure that other Data Center provider's information is held separately in it's tables.

4 IANA Considerations

Suitable IANA considerations for extending the BGP extended community attribute for accommodating the power metric information in the MP-iBGP update are to be taken into account. This will be made more clear in subsequent versions of the document.

5 References

5.1 Normative References

Please see Appendix A.

5.2 Informative References

Please see Appendix A.

APPENDIX - A : References for power saving related material

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- greennet (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-psp-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-psp-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

Balaji Venkat Venkataswami
DELL
Plot #1 SIDCO Estate
Olympia Tech Park
Guindy
Chennai
India

EMail: balaji_venkat_venkat@dell.com

Bhargav Bhikkaji
DELL
350 Holger Way
San Jose, CA
U.S.A

Email: Bhargav_Bhikkaji@dell.com

Shankar Raman
Department of Computer Science and Engineering
IIT Madras
Chennai - 600036
TamilNadu
India

EMail: mjsraman@cse.iitm.ac.in