Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: July 11, 2013

B. Zhang
J. Shi
The University of Arizona
J. Dong
M. Zhang
Huawei

Januray 10, 2013

Power-aware Routing and Traffic Engineering: Requirements, Approaches, and Issues
draft-zhang-greennet-01

Abstract

Energy consumption of network infrastructures is rising fast. There are emerging needs for power-aware routing and traffic engineering, which adjust routing paths to help reduce power consumption network-wide. This document gives a high-level analysis on the basic requirements, approaches, and potential issues in power-aware routing and traffic engineering.

Table of Contents

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. 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 to and heat removal from router components as well as the hosting facilities [Gupta03] [Epps06].

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. Though advances in hardware design have made individual routers more energy efficient over the years, there is still has a long way to go before routers become energy-proportional. Recently researchers have started to look beyond a single router or linecard for network-wide solutions towards energy proportionality. Power-aware routing and traffic engineering has been proposed to improve network's energy efficiency, for example, aggregating traffic onto a subset of links and putting the other links with no traffic into sleep. As demonstrated in several research works, this approach has the potential to save a significant amount of energy [GreenTE] [Nedevschi08] [Chabarek08]. Designing a practical protocol, however, has been challenging, because making routing protocols power-aware brings fundamental changes to the routing system and the entire network, thus it is a complicated task with involvement of hardware support, protocol design and operations, and network monitoring.

This document gives a high-level analysis on power-aware routing and traffic engineering, including solution requirements, existing approaches, and potential issues. Power-aware routing and traffic engineering exploits the over-provisioned feature of networks, so there will be some impact on network performance and resilience. In order to save energy without impacting network performance and resilience too much, certain requirements should be met. When a power-aware approach is implemented, new issues may arise, and should be addressed to make the solution practical. While there are many aspects of energy efficient networks, this document focuses on intra-domain routing within a single ISP.

2. Requirements

The high-level idea of power-aware routing or traffic engineering 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 to a small number of links so that other links can be put to sleep or reduce rate in order to save power. The fundamental requirement for any energy-efficient network solution is to save power without negative impacts on network operations, which can be broken down as follows.

- The network should retain enough resiliency against node and/or link failures.
- The network should have enough spare standby capacity or be able to react quickly enough to traffic spikes in order to minimize packet losses due to links/routers being in low power states.
- QoS metrics such as end-to-end delay should be kept at a desired level.
- The operation of other protocols should not be interrupted, or at least other protocols should be able to adapt without being broken after links/routers change their power states.

While this document focuses on routing and traffic engineering, it requires support from underlying hardware and system for energy management capability, which is the topic of the IETF Energy Management Working Group [EMAN-WG]

3. Approaches

In the last couple of years a number of power-aware protocols have been proposed in research. 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 chassises, 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 approaches 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. Issues

4.1. Hardware Support

In order to save power, routers and switches should support low power states, and make available control primitives to enter or leave low power states. To reduce the impact on network performance, routers and switches should have the ability to change power states quickly. These are the hardware support needed by power-aware protocols.

Sleeping State

Most sleep-based approaches require routers and switches, or a component of them such as a line card, to support sleeping state. While most components can go to sleep very quickly, they also need to be able to wake up quickly. Besides, entering and leaving sleeping state often incurs extra energy draw, which need to be kept small. Different designs may have different requirements of the transition time between power states. In uncoordinated sleeping approach, upstream routers intentionally buffer packets for a very short period of time to allow downstream routers longer sleep time. This approach can only allow a component to sleep for a few milliseconds, otherwise the buffering may cause too much extra delay. Hence this approach requires a very short transition time and low penalty power. In coordinated sleeping approaches, where routers coordinate on which paths to use and when to put links to sleep, a component usually can

sleep much longer, for seconds, minutes or even longer. Therefore their requirement on transition time and power is more relaxed.

Common energy management scheme at the individual component such as line card is sleep-on-idle (SoI) and wake-on-arrival (WoA). When a link is idle for a short period of time, it goes into sleep; when a packet arrives, it wakes up. Power-aware protocols manipulate traffic paths so that some links will have much longer idle time than default routing.

The hardware is also expected to minimize potential packet loss during the transition between power states. Especially in WoA, the first packet is susceptible to loss. The two ends of the link can coordinate, e.g., one end sends a dummy packet to the other end to inform about the link wakeup, or if they don't coordinate, the receiver end should have the capability to buffer incoming packets before the interface wakes up to process these packets.

Multiple Data Rates

CMOS based silicon supports Dynamic Voltage Scaling (DVS), so clocking an interface at a lower frequency, and operating at lower data rate can save considerable amount of energy. This calls for a need for router interfaces to support multiple data rates. If an interface could support more data rates and incur low penalty power on a change, there are more opportunity to save energy. Furthermore, it will also help if an interface supports different sending and receiving data rates.

The transition between different data rates needs be quick and on-the-fly. Most Ethernet cards supports auto-negotiation of data rates, which happens when a cable is plugged in and takes hundreds of milliseconds. Auto-negotiation is not suitable for changing data rate to save energy, because buffer would be filled up during the negotiation period and leads to packet loss. A fast mechanism for initiating and agreeing upon a link data rate change is necessary.

Electrical Damage

Many electronic devices are not designed to be turned on and off frequently. When a device is waken up, the in-rush current may damage or destroy a component and related circuits. Hardware that is more friendly to power management is needed.

Optical Component Support

Electrical components consumes much more energy than optical components in network routing infrastructure. Therefore, many power-

aware routing or traffic engineering approaches are designed with electrical devices in mind. However, the number of optical components in ISP networks can also be large. Care should be taken when adopting existing approaches to optical networks. For example, optical receivers cannot be turned off when WoA is needed. Furthermore, there is room for more power saving when optical components are explicitly considered in the approach. It's possible to turn off an optical receiver while maintaining the ability to wake it up when needed, by maintaining another route towards the other end that a control packet could be delivered.

4.2. Software Support

There are many different power-aware approaches. They need different input datasets, and generate different instructions. Software is required to collect necessary input, as well as deliver and execute resulting instructions.

Topology and Traffic

Many power-aware approaches require knowledge of global topology on a centralized server or on each router. It's fairly easy to satisfy this requirement by running a link state routing protocol such as OSPF. If a network running OSPF has OSPF areas configured, power-aware approaches can only be deployed within one area, or some other way to collect global topology is needed.

Many approaches require knowledge of link utilization on a local router, its neighbors, or all routers. Routers may need to maintain necessary counters to calculate this information, and exchange or announce them. Routers then need to categorize packets and maintain a separate set of counters for each interesting category. Some approaches such as GreenTE require network-wide traffic matrix. There are two ways to obtain this information: infer from link utilization, or collect directly. We can infer traffic matrix from global topology and link utilization by using gravity model and tomographic method [TM]. This method requires some computation power, but needs least amount of data exchange, so it is particularly useful when traffic matrix is only needed on a centralized server. However, the accuracy of this method is not guaranteed, especially when traffic engineering is in place that causes traffic pattern to deviate from the gravity model, or multicast is enabled which creates multiple copies of packets. We can also collect traffic matrix directly. There is a cost on ingress routers: an ingress router needs to identify the egress node, and maintain one counter per egress. Identifying egress is not an extra cost in many cases, because many approaches need to know egress to select a feasible route. Maintaining per-egress counters, as well as sending them to

the centralized server in centralized approaches, is a high cost.

Traffic Splitting

Some approaches may route traffic between an ingress-egress pair along multiple paths, according to certain split ratio. To avoid out-of-order arrival which impacts TCP performance, traffic splitting is usually based on the hash of some fields in the packet header, such as source-destination IP pair. In a small network such as a company, there are big flows between some IPs, while there is little traffic on most other IPs. In this case, hash-based splitting has significant bias.

Timeliness of Solutions

Traffic engineering approaches take network topology and traffic information (in the form of link utilization or traffic matrix) as input, and outputs a solution including which links should be sleeping and what rate should links be operated on. Most traffic engineering approaches run on a centralized server. Traffic demand changes over time, and network topology may even change due to link failure. It takes time to collect traffic information from the entire network, and time is also consumed while computing the solution. Thus, the solution, when comes out, is based on network topology and traffic information of sometime earlier, and it may not still be applicable to current situation. Prediction of future traffic information may help in some situations.

4.3. Impacts on Protocols

Power-aware routing and traffic engineering is a tradeoff between energy consumption and network resilience. They save power by turning off or slowing down some links, which were previously over-provisioned to obtain better resilience. Any power-aware approach will cause loss of network resilience to some extent. Sleeping based approaches has another impact. Traditionally, a link is either up or down. An up link can transmit packets, and a down link cannot. A third state, sleeping, is added by power-aware protocols. A sleeping link cannot transmit packets right away, but it can be waken up when needed. The introduction of a third sleeping state has its impact on protocols that maintain their own states about network links.

Congestion after Traffic Surges

Traffic engineering approaches usually take traffic information at certain time, and a solution contains a routing scheme that could accommodate such traffic on a reduce topology with some links sleeping or operating at lower rate. This routing scheme usually

keeps link utilization under certain threshold, so that there is some safe margin in case traffic increases. However, because a solution is computed periodically, congestion is still possible when traffic increases to a level that exceeds the safe margin within one adjustment period. To address this issue, some method of fast readjustment is needed. When a traffic increase is observed, the routing scheme should be slightly changed to accommodate this traffic, probably waking up or increase rate on a few links.

Network Partition on Link or Node Failure

Many sleep based approaches will result in a topology with very low redundancy level. These reduced topologies are vulnerable to link and node failures, which are quite common in large networks. Those approaches should be improved by adding a constraint of redundancy level. A redundancy level of 2, which could protect from single link failure, is a reasonable value. It's possible to incorporate power aware feature into MRT to achieve energy saving while remain the network 2-disjoint [I-D.ietf-rtgwg-mrt-frr-architecture]. Once a partition is detected, it's easy to repair by waking up all sleeping links. But this causes a sudden increase on power consumption, which is sometimes undesirable. A local algorithm to select a subset of sleeping links that could repair the partition is needed. The selection doesn't need to be optimal, because waking up a small subset is much better than waking up all sleeping links.

Sleeping Link State in Routing Protocols

Sleeping links should be handled separately in routing protocols. A sleeping link should be advertised as up, probably with a tag stating it's sleeping. No HELLO messages should be sent over a sleeping link, so no HELLO messages could be received from a sleeping link. Missed HELLO messages on a sleeping link should not cause the link to be treated as down state. As a consequence, if a sleeping link fails, the failure would not be detected until the router attempts to wake it up. To detect a failure earlier, it may be desirable to wake up the link and probe it periodically (using a long interval such as every hour). No control message should be sent over a sleeping link. This may cause the network to converge slower than usual, because LSA flooding takes more hops. Fortunately most power-aware approaches have network diameter constraints, so convergence time should be comparable.

IP Multicast on Reduced Topology

IP Multicast works by building one or more trees on available links. If any link in a multicast tree goes to sleep, some receivers cannot receive multicast packets for a noticeable period of time, until IP

multicast automatically repairs the tree. So, if a link is part of a multicast tree, it should not be put to sleep.

One solution is keeping all links that are contained in multicast trees active. If there are many multicast trees that don't have much overlap, a major portion of links would be forced active by multicast, and power saving potential is greatly limited. Another solution is explicitly modifying multicast trees in a power-aware approach. This is not an easy way to go. There should be a delay constraint on each multicast tree, and there're possibly a large number of multicast trees. After a multicast tree is modified, utilization of multiple links will change.

A third solution is making IP multicast power-aware. When a multicast tree is being built, energy consumption is taken into account, such that IP multicast would attempt to use as few links as possible as long as delay constraint could be satisfied. After that, these links used by IP multicast will not go to sleep.

4.4. Network Monitoring

Network operators demand a monitoring solution when deploying anything. The most important metrics are: How much energy is saved? How much impact is there on network performance? Measurement of Energy Consumption. The IETF eman WG [EMAN-WG] is working on defining energy objects in network devices, and monitoring and controlling their states. When a device is running on full power state, the power demand is recorded as full power demand. When a power-aware approach is deployed, actual energy consumption is measured. The amount of saved energy is the full power demand multiplied by elapsed time during the measurement of actual energy consumption subtracted by actual energy consumption.

In centralized periodical adjustment approaches, the centralized server should have knowledge of current applied solution (which is based on previous traffic information) and current traffic information. It can then calculate what link utilization and delay would be when this traffic is routed on current applied solution, as well as the performance as if this traffic is routed without power-aware consideration. It's not trivial to measure the impact in other cases.

SNMP MIBs are needed to standardize monitoring. Software for operations products such as System Center Operations Manager needs to integrate power-aware routing and traffic engineering to existing IT monitoring architecture.

5. Summary

Power-aware routing and traffic engineering has great potential to improve network energy efficiency while maintain network services at desired levels. Its effectiveness, however, depends on various supports from hardware and software, and more importantly, protocol designs that address operational issues. This document is a first step towards developing practical power-aware protocols.

6. IANA Considerations

This document has no actions for IANA.

7. 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.

8. Informative References

[Chabarek08]
Chabarek, J. and et al. , "Power Awareness in Network Design and Routing", IEEE INFOCOM 2008.

[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>.

[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.

[I-D.ietf-rtgwg-mrt-frr-architecture]
Atlas, A., Kebler, R., Envedi, G., Csaszar, A.,

Konstantynowicz, M., White, R., and M. Shand, "An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees", draft-ietf-rtgwg-mrt-frr-architecture-01 (work in progress), March 2012.

[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/>.

[TM] Roughan, M., Thorup, M., and Y. Zhang, "Traffic Engineering with Estimated Traffic Matrices", IMC 2003.

Authors' Addresses

Beichuan Zhang
The University of Arizona

Email: bzhang@cs.arizona.edu

Junxiao Shi
The University of Arizona

Email: shijunxiao@cs.arizona.edu

Jie Dong
Huawei

Email: jie.dong@huawei.com

Mingui Zhang
Huawei

Email: zhangmingui@huawei.com