Network Working Group Greg Bernstein (Grotto)
Internet Draft Young Lee (Huawei)
Intended status: Informational

June 28, 2011

Use Cases for High Bandwidth Query and Control of Core Networks

draft-bernstein-alto-large-bandwidth-cases-00.txt

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/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on December 28, 2011.

Copyright Notice

Copyright (c) 2011 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.

Abstract

This draft describes two generic use-cases that illustrate application layer traffic optimization concepts applied to high bandwidth core networks. For the purposes here high bandwidth will mean bandwidth that is significant with respect to the capacity of a wavelength in a wavelength division multiplexed optical transport system, e.g., 10-40Gbps or more. For each of these generic use cases, we present a generic optimization problem, look at the type of information needed (query interface) to perform the optimization, investigate a reservation interface to request network resources, and also consider enhanced availability and recovery scenarios.

Table of Contents

1. Introduction

Cloud Computing, network applications, software as a service (SaaS), Platform as a service (PaaS), and Infrastructure as a Service (IaaS), are just a few of the terms used to describe situations where multiple computation entities interact with one another across a network.   When the communication resources consumed by these interacting entities is significant compared with link or network capacity then opportunities may exist for more efficient utilization of available computation and network resources if both computation and network stratums cooperate in some way. The application layer traffic optimization (ALTO) working group is tackling the similar problem of "better-than-random peer selection" for distributed applications based on peer to peer (P2P) or client server architectures [16]. In addition, such optimization is important in content distribution networks (CDNs) as illustrated in [17].

General multi-protocol label switching (GMPLS) [18] can and is being applied to various core networking technologies such as SONET/SDH [19] and wavelength division multiplexing (WDM) [20]. GMPLS provides dynamic network topology and resource information, and the capability to dynamically allocation resources (provision label switched paths). Furthermore, the path computation element (PCE) [21] provides for traffic engineered path optimization.

However, neither GMPLS nor PCE provide interfaces that are appropriate for an application layer entity to use for the following reasons:

. GMPLS routing exposes full network topology information which tends to be proprietary to a carrier or require specialized knowledge and techniques to make use of, e.g., the routing and wavelength assignment (RWA) problem in WDM networks [20].

. Core networks typically consist of two or more layers, while applications are typically only know about the IP layer and above. Hence applications would not be able to make direct use of PCE capabilities.

. GMPLS signaling interfaces are defined for either peer GMPLS nodes or via a user network interface (UNI) [22]. Neither of these is appropriate for direct use by an application entity.

In this paper we discuss two general use-cases that can generate core network flows with significant bandwidth and may vary significantly over time. The "cross stratum optimization" problems generated by these use cases are discussed. Finally, we look at interfaces between the application and network "stratums" that can enable overall optimization.

1.1. Computing Clouds, Data Centers, and End Systems

While the definition of cloud computing or compute clouds is somewhat nebulous (or "foggy" if you will) [1], the physical instantiation of compute resources with network connectivity is very real and bounded by physical and logical constraints. For the purposes of this paper we will call any network connected compute resources a data center if its network connectivity is significant compared either to the bandwidth of an individual WDM wavelength or with respect to the network links in which it is located. Hence we include in our definition very large data centers that feature multiple fiber access and consume more than 10MW of power [2], moderate to large content distribution network (CDN) installations located in or near major internet exchange points [3], medium sized business centers, etc...

We will refer to those computational entities that don't meet our bandwidth criteria for a data center as an "end system".

2. End System Aggregate Networking

In this section we consider the fundamental use case of end systems communicating with data centers as shown in Figure 1. In this figure the "clients" are end systems with relatively small access bandwidth compared to a WDM wavelength, e.g., under 100Mbps. We show these clients roughly partitioned into three network related regions ("A", "B", and "C"). Given a particular network application, in a static network application situation, each client in a region would be associated with a particular data center.

```
                                           Region B
                             +---------+  +------+
                             |  Data   |  |Client|
                             |Center 2 |  |  B1  |+------+
             +------+        +----+----+  +--+---+|Client|
             |Client|             |         /     |  B2  |
             |  A1  `.         _.-+---------+-.   +--+---+
   Region A  +------+ `-.  ,-''               `--.  /   ...
        +------+        ,`:                       `+.     +------+
        |Client|       /                             \    |Client|
        |  A2  +------+                               \---+  BM  |
        +------+     (             Network             )  +------+
         ...        .-'                               /
     +------+   _.-'   \                             `.
     |Client|.-'        `=.                       ,-'  `.
     |  AN  |       _.-''  `--.               _.-\   +---`.----+
     +------+ +----'----+      `----+------+''    \  |  Data   |
              |  Data   |           |       \      | |Center 3 |
              |Center 1 |        +--+---+ +--+---+ \ +---------+
              +---------+        |Client| |Client|  \------+
                                 |  C1  | |  C2  |  |Client|
                                 +------+ +------+  |  CK  |
                                       Region C     +------+
```

Figure 1. End system to data center communications.

2.1. Aggregated Bandwidth Scaling

One of the simplest examples where the aggregation of end system bandwidth can quickly become significant to the "network" is for video on demand (VoD) streaming services. Unlike a live streaming service where IP or lower layer multicast techniques can be generally applied, in VoD the transmissions are unique between the data center and clients. For regular quality VoD we'll use an estimate of 1.5Mbps per stream (assuming H.264 coding), for HD VoD we'll use an estimate of 10Mbps per stream. To fill up a 10Gbps capacity optical wavelength requires either 6,666 or 1,000 clients for regular or high definition respectively. Note that special multicasting techniques such as those discussed in [4] and peer assistance techniques such as provided in some commercial systems [5] can reduce the overall network bandwidth requirements.

With current high speed internet deployment such numbers of clients are easily achieved; in addition demand for VoD services can vary significantly over time, e.g., new video releases, inclement weather (increases number of viewers), etc...

2.2. Cross Stratum Optimization Example

In an ideal world both data centers and networks would have unlimited capacity, however in actuality both can have constraints and possibly varying marginal costs that vary with load or time of day. For example suppose that in Figure 1 that Data Center 3 has been primarily serving VoD to region "C" but that it has, at a particular period in time, run out of computation capacity to serve all the client requests coming from region "C". At this point we have a fundamental cross stratum optimization (CSO) problem. We want to see if we can accommodate additional client request from region "C" by using a different data center than the fully utilized data center #3. To answer this questions we need to know (a) available capacity on other data centers to meet a request, (b) the marginal (incremental) cost of servicing the request on a particular data center with spare capacity, (c) the ability of the network to provide bandwidth between region "C" to a data center, and (d) the incremental cost of bandwidth from region "C" to a data center.

```
                                          Region B
                            +---------+  +------+
                            |  Data   |  |Client|
                            |Center 2 |  |  B1  |+------+
            +------+        +----+----+  +--+---+|Client|
            |Client|             |         /     |  B2  |
            |  A1  `.         _.-+--------+-.    +--+---+
  Region A  +------+ `-.  ,-'' XXXXX     XX  `--.  /   ...
       +------+        ,':       ``---..__ XXXX   `+.    +------+
       |Client|       /  X        |       ```--XX   \    |Client|
       |  A2  +------+..X`.       \              XX--+---+  BM  |
       +------+     (  X   `-/     \                  )  +------+
        ...        .-'     .'       |        +----.X /
    +------+   _.-'   \  X/         \        |    X `.
    |Client|.-'        `=.X          \      XXXX ,-'  `.
    |  AN  |       _.-''  `--.    XXXXXXXXX  _.-\   +---`.----+
    +------+ +----'----+      `----+------+''    \  |  Data   |
             |  Data   |           |       \      | |Center 3 |
             |Center 1 |        +--+---+ +--+---+ \ +---------+
             +---------+        |Client| |Client|  \------+
                                |  C1  | |  C2  |  |Client|
                                +------+ +------+  |  CK  |
                                      Region C     +------+
```

Figure 2. Aggregated flows between end systems and data centers.

In Figure 2 we show a possible result of solving the previously mentioned CSO problem. Here we show the additional client requests from region "C" being serviced by data center #2 across the network. Figure 2 also illustrates the possibility of setting up "express" routes across the network at the MPLS level or below. Such techniques, known as "optical grooming" or "optical bypass" [6], [7] at the optical layer, can result in significant equipment and power savings for the network by "bypassing" higher level routers and switches.

2.3. Data Center and Network Faults and Recovery

Data center failures, whether partial or complete, can have a major impact on revenues in the VoD example previously described. If there is excess capacity in other data centers within the network associated with the same application then clients could be redirected to those other centers if the network has the capacity. Moreover, MPLS and GMPLS controlled networks have the ability to reroute traffic very quickly while preserving QoS. As with general network recovery techniques [8] various combinations of pre-planning and "on the fly" approaches can be used to tradeoff between recovery time and excess network capacity needed for recovery.

In the case of network failures there is the potential for clients to be redirected to other data centers to avoid failed or over utilized links.

2.4. Cross Stratum Control Interfaces

Two types of load balancing techniques are currently utilized in cloud computing. The first is load balancing within a data center and is sometimes referred to as local load balancing. Here one is concerned with distributing requests to appropriate machines (or virtual machines) in a pool based on the current machine utilization. The second type of load balancing is known as global load balancing and is used to assign clients to a particular data center out of a choice of more than one within the network and is our concern here. A number of commercial vendors offer both local and global load balancing products (F5, Brocade, Coyote Point Systems). Currently global load balancing systems have very little knowledge of the underlying network. To make better assignments of clients to data centers many of these systems use geographic information based on IP addresses [9]. Hence we see that current systems are attempting to perform cross stratum optimization albeit with very coarse network information. A more elaborate interface for CSO in the client aggregation case would be:

1. A Network Query Interface - Where the global load balancer can inquire as to the bandwidth availability between "client regions" and data centers.

2. A Network Resource Reservation Interface - Where the global load balancer can make explicit requests for bandwidth between client regions and data centers.

3. A Fault Recovery Interface - For the global load balancer to make requests for expedited bulk rerouting of client traffic from one data center to another.

The network query interface can be considered a superset of the functionality proposed from the ALTO (application layer traffic optimization) servers being standardized in [10]. Note that in the network query and reservation interfaces it would be worthwhile to consider both current resources and resources at a future time, i.e., scheduled resources. Although scheduled reservations are not supported directly by technologies such as MPLS and GMPLS they can be considered in network planning and provisioning systems. For example, a VoD provider knows ahead of time when the latest "blockbuster" film will be available via its service and can make estimates based on historical data on the bandwidth that it will need to deal with the subsequent demand.

3. Data Center to Data Center Networking

There are a number of motivations for data center to data center communications: on demand capacity expansion ("cloud bursting") [11], cooperative exchanges between business partners, offsite data backup, "rent before building"[12], etc... In Figure 3 we show an example where a number of businesses each with an "internal data center" contracts with a large external data center for additional computational (which may include storage) capacity. The data centers may connect to each other via IP transit type services or more typically via some type of Ethernet virtual private line or LAN service.

```
                    +-------------------+
                    |                   |
                    | Large Data Center |
                    |                   |
                    +----------+--------+
                               |
                             _.+-----------.
                        ,--''               '---.
                     ,-'                         '-.
                    ,'                              '.
                 ,'                                   '.
+--------+      ;                Network                :
|Business|  __..+                                       |
| #1 DC  +-'    :                                       ;
+--------+       '.                                   ,'
                   '.                               ;:
                     '-.                         ,-'  \
                        '---.               _.--'   +--'.----+
                             '+-----------''        |Business|
                              /                     | #N DC  |
                             |                      +--------+
                        +----+---+
                        |Business|
                        | #2 DC  |
                        +--------+
```

Figure 3. Basic data center to data center networking.

3.1. Cross Stratum Optimization Examples

In the DC-to-DC example of Figure 3 we can have computational constraints/limits at both local and remote data centers; fixed and marginal computational costs at local and remote data centers; and network bandwidth costs and constraints between data centers. Note that computing costs could vary by the time of day along with the cost of power and demand. Some cloud providers such as Amazon [13] have quite sophisticated compute pricing models including: reserved, on demand, and spot (auction) variants.

In addition, to possibly dynamically changing pricing, traffic loads between data centers can be quite dynamic. In addition, data movement between data centers is another source of large network usage variation. Such peaks can be due to scheduled daily or weekly offsite data backup, bulk VM migration to a new data center, periodic virtual machine migration [14], etc...

3.2. Network and Data Center Faults and Reliability

For networked applications that require high levels of reliability/availability the network diagram of Figure 4 could be enhanced with redundant business locations and external data centers as shown in Figure 4. For example cell phone subscriber databases and financial transactions generally require what is called geographic database replication [15] and results in extra communication between sites supporting high availability. For example if business #1 in Figure 4 required a highly available database related service then there would be an additional communication flows from the data center "1a" to data center "1b". Furthermore, if business #1 has outsourced some of its computation and storage needs to independent data center X then for resilience it may want/need to replicate (hot-hot redundancy) this information at independent data center Y.

```
              +-------------+              +-------------+
              |Independent  |              |Independent  |
              |Data Center X|              |Data Center Y|
              +-----+-------+              +------+------+
                     \                           /
                      '.      _.-----------.   .'
                        \--''               '-+-.
                     ,-'                         '-.      +--------+
                   ,'                               '.   .'Business|
                 ,'                                   '.-' |#N DC-a |
                ;                Network                :  +--------+
    +--------+  |                                       |
    |Business+---                                       ;
    |#1 DC-a |   '.                                   +:
    +--------+     '.                               ;/  \
                     '-.                         ,-'     '.
                      .''---.               _.--'       +--'.----+
        +--------+   /       '+-+---------\'            |Business|
        |Business| .'           |          \            |#N DC-a |
        |#1 DC-b .'             /           \           +--------+
        +--------+             |             \
                          +----+---+    +--------+
                          |Business|    |Business|
                          |#2 DC-a |    |#2 DC-b |
                          +--------+    +--------+
```

Figure 4. Data center to data center networking with redundancy.

3.3. Cross Stratum Control Interfaces

Similar to the end system aggregation case we can decompose cross stratum interfaces into three general types: (a) network query, (b) network reservation, and (c) recovery. However for DC-to-DC interfaces we are interested in network resources between data centers rather than between "client regions" and data centers.

For network resource queries we may be concerned with (a) current bandwidth availability, (b) bandwidth availability at a future time, or (c) bandwidth for a bulk data transfer of a given amount that must take place within a given time window. A network reservation interface with both current and advanced reservation capability would complement the query interface.

A simple recovery interface for data center based faults could be based on unused backup paths between data centers that are reserved but not activated unless a request is received from the application stratum that recovery action is requested.

4. Conclusion

In this draft we have discussed two generic use cases that motivate the usefulness of general interfaces for cross stratum optimization in the network core. In our first use case network resource usage became significant due to the aggregation of many individually unique client demands. While in the second use case where data centers were communicating with each other bandwidth usage was already significant enough to warrant the use of private line/LAN type of network services.

Both use cases result in optimization problems that trade off computational versus network costs and constraints. Both featured scenarios where advanced reservation, on demand, and recovery type service interfaces could prove beneficial. Many concepts from recent standardization work at the IETF [10] such as location identifiers, and endpoint properties could be reused in defining such interfaces.

5. Security Considerations

TBD

6. IANA Considerations

This informational document does not make any requests for IANA action.

7. References

7.1. Informative References

[1] M. Armbrust et al., "A view of cloud computing," Communications of the ACM, vol. 53, p. 50-58, Apr. 2010.

[2] "Location Information | DuPont Fabros Technology." (Online). Available: http://www.dft.com/data-centers/location-information.

[3] "Amazon CloudFront." (Online). Available: http://aws.amazon.com/cloudfront/.

[4]   K. A. Hua and S. Sheu, "Skyscraper broadcasting: a new broadcasting scheme for metropolitan video-on-demand systems," in Proceedings of the ACM SIGCOMM ’97 conference on Applications, technologies, architectures, and protocols for computer communication, Cannes, France, 1997, pp. 89-100.

[5]   "Adobe Flash Media Server 4.0 * Building peer-assisted networking applications." (Online). Available: http://help.adobe.com/en_US/flashmediaserver/devguide/WSa4cb076 93d123884520b86f312a354ba36d-8000.html.

[6]   Rudra Dutta and George N. Rouskas, "Traffic grooming in WDM networks: Past and future," IEEE Network, vol. 16, no. 6, pp. 46 -56, 2002.

[7]   Keyao Zhu and B. Mukherjee, "Traffic grooming in an optical WDM mesh network," Selected Areas in Communications, IEEE Journal on, vol. 20, no. 1, pp. 122-133, 2002.

[8]   G. Bernstein, B. Rajagopalan, and D. Saha, Optical Network Control: Architecture, Protocols, and Standards. Addison-Wesley Professional, 2003.

[9]   "Our IP Geolocation Products | Quova, Inc." (Online). Available: http://www.quova.com/what/products/.

[10] "draft-ietf-alto-reqs-09." (Online). Available: http://datatracker.ietf.org/doc/draft-ietf-alto-reqs/.

[11] "Cloud Computing’s Tipping Point -- InformationWeek." (Online). Available: http://www.informationweek.com/news/government/cloud-saas/229401691.

[12] "Lessons From FarmVille: How Zynga Uses The Cloud -- InformationWeek." (Online). Available: http://www.informationweek.com/news/global-cio/interviews/229402805#.

[13] "Amazon EC2 Pricing." (Online). Available: http://aws.amazon.com/ec2/pricing/.

[14] Dynamic Workload Balancing with EMC VPLEX and Ciena Networking. EMC, 2010.

[15] "MySQL.:: MySQL Cluster Features." (Online). Available: http://www.mysql.com/products/cluster/features.html#geo.

[16] Seedorf, J. and E. Burger, "Application-Layer Traffic Optimization (ALTO) Problem Statement", RFC 5693, October 2009.

[17] B. Niven-Jenkins (Ed.), G. Watson, N. Bitar, J. Medved, S. Previdi, "Use Cases for ALTO within CDNs", work in progress, draft-jenkins-alto-cdn-use-cases.

[18] E. Mannie, Ed., "GMPLS Framework Generalized Multi-Protocol Label Switching (GMPLS) Architecture" RFC 3945, October 2004.

[19] G. Bernstein, E. Mannie, V. Sharma, E. Gray, "Framework for Generalized Multi-Protocol Label Switching (GMPLS)-based Control of Synchronous Digital Hierarchy/Synchronous Optical Networking (SDH/SONET) Networks", RFC 4257, December 2005.

[20] Y. Lee, Ed., G. Bernstein, Ed., W. Imajuku, "WSON Framework Framework for GMPLS and Path Computation Element (PCE) Control of Wavelength Switched Optical Networks (WSONs)", RFC6163, April 2011.

[21] A. Farrel, J.-P. Vasseur, J. Ash, "PCE Framework A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.

[22] G. Swallow, J. Drake, H. Ishimatsu, Y. Rekhter, "Generalized Multiprotocol Label Switching (GMPLS) User-Network Interface (UNI): Resource ReserVation Protocol-Traffic Engineering(RSVP-TE) Support for the Overlay Model" RFC 4208, October 2005.

Author's Addresses

Greg M. Bernstein
Grotto Networking
Fremont California, USA
Phone: (510) 573-2237
Email: gregb@grotto-networking.com

Young Lee
Huawei Technologies
1700 Alma Drive, Suite 500
Plano, TX 75075
USA
Phone: (972) 509-5599
Email: ylee@huawei.com

Intellectual Property Statement

The IETF Trust takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in any IETF Document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights.

Copies of Intellectual Property disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement any standard or specification contained in an IETF Document. Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

All IETF Documents and the information contained therein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE

REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION THEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.