Network Working Group Young Lee (Huawei)
Internet Draft Greg Bernstein (Grotto)
Intended status: Informational Ning So (University of Texas)
Tae Yeon Kim (ETRI)
Kohei Shiomoto (NTT)
Oscar Gonzalez de Dios (Telefonica)

March 3, 2011

Research Proposal for Cross Stratum Optimization (CSO) between Data Centers and Networks

draft-lee-cross-stratum-optimization-datacenter-00.txt

Status of this Memo

Copyright Notice

Abstract

Data Centers offer various application services to end-users such as video gaming, cloud computing and others. Since the data centers used to provide application services are distributed geographically around a network, many decisions made in the control and management of application services, such as where to instantiate another service instance or to which data center out of several a new client is assigned, can have a significant impact on the state of the network. Conversely the capabilities and state of the network can have a major impact on application performance.

Currently application decisions are made with very little or no information concerning the underlying network used to deliver those services. Hence such decisions may be sub-optimal from both application and network resource utilization and quality of service objectives. This document proposes a research program into cross stratum application/network optimization focusing on the challenges and opportunities presented by data center based applications and carriers networks.

Table of Contents

1. Introduction

This document describes a research program on the automation of certain interactions between data center based distributed

applications and the supporting networking infrastructure. Data center based applications are used to provide a wide variety of services such as video gaming, cloud computing [Nurmi], grid application [GFD-122] and others. High-bandwidth video applications such as remote medical surgery, live concerts and sporting events are also emerging. This document is mainly concerned with data center applications that in aggregate or individually make substantial bandwidth demands on the network. In addition these applications may desire specific bounds on QoS related parameters such as latency and jitter.

Figure 1 shows a network diagram of an example data center based application. Data centers come in an extreme variety of sizes and configurations but all contain compute servers, storage and application control of some sort.

```
                    ,-----.     ---------------
  ----------       / App   \   |         DC 1  |
 | End-user |. . .>( Control )  |      o o o    |
 |          |      \       /   |       \|/     |
  ----------        `-----'    |        O      |
       |                        ----- --|------
       |                                |
       |                                |
       |       -------------------------|--
       |      /                     PE1 |  \
       |     /        ..................O   \     --------------
       |    |       .                        |   | o o o   DC 2 |
       |    | PE4 .                     PE2  |   |  \|/         |
        ----|---O.......................O---|---|---O          |
            |     .                          |   |              |
            |      .          PE3            |    --------------
             \      .........O   Carrier    /
              \              |   Network   /
               --------------|-------------
                             |
                     --------|------
                    |        O      |
                    |       /|\     |
                    |      o o o    |
                    |          DC 3 |
                     ---------------
```

Figure 1. Data center based application architecture example

This research is concerned with a subset of "cross stratum optimization" (CSO) opportunities, e.g., combined optimization of

resources in the application and network stratum. We use the term stratum here to broadly differentiate the layers of most concern to the application and to the network in general.

In the application stratum we are concerned with and limiting the scope of this research to those distributed applications offered via data centers. In particular, this project does not intend to cover applications delivered in a strictly peer to peer manner. Application resources can be roughly categorized into computing resources, i.e., servers of various types and granularities (VMs, memory, disk) and content, e.g., video, audio, databases, large data sets, etc..

By the network stratum we mean the "network layer" (IP) and below, e.g., MPLS, SDH, OTN, WDM. The network stratum has resources that include routers, switches, and links. We are particularly interested in further unleashing the potential presented by MPLS and GMPLS control planes at the lower network layers in response to the high aggregate or individual demands from the application layer.

The four main cross stratum optimization opportunities of this research project are:

1. Resource optimization (application and network)

2. Responsiveness to quickly changing demands

3. Enhanced service resilience (via cooperative recovery techniques between application and network)

4. Quality of application experience (QoE) enhancement (via better use of existing network and application resources)

In the following document we first give a brief overview of data center technology for network oriented readers, describe the current state of application/network integration from the deployment, and standards points of view, an then conclude with a more detailed description of the research thrusts (optimization, resilience, QoE) from the perspective of an IRTF project.

2. Key Issues in Data Centers and Clouds

This section provides some key issues related to data centers and cloud computing that motivate the need for cross stratum optimization between applications and networks.

2.1. Some Obstacles of Cloud Computing

There are many drivers for the move towards data center based application services. They include reducing maintenance costs, energy costs, flexibility, scalability, etc...

Reference [Armb] offers a very timely and readable review of cloud computing practice and potential. Though here we do not differentiate between cloud computing and medium and small data center based computing that utilize modern virtualization techniques and possibly other cloud computing techniques [Nurmi]. From their "top ten obstacles and opportunities for cloud computing" we see that over half have significantly involvement of the network.

1. Availability/Business Continuity

3. Data Confidentiality and Auditability

4. Data Transfer Bottlenecks

5. Performance Unpredictability

8. Scaling Quickly

9. Reputation Fate Sharing

2.2. Changes in Network Access from Data Centers and Clouds

At the high side of data center size we begin to see significant changes in network access, e.g., from a drop-off of an optical metro ring (a wavelength or two), to an end destination in a long haul DWDM system (many wavelengths, multiple fibers). These changes have been partly driven by the consolidation effort of existing smaller size data centers into Super Data Centers in the government IT infrastructure and carriers.

Another factor that contributes to high-speed network access is due to emerging applications that require high bandwidth such as sporting events, live converts, 3D video applications, remote medical surgery and so on.

These changes provide still more motivation to enable the application layer to take advantage of the dynamic networking features offered by network capability such as MPLS/GMPLS.

2.3. Virtual Machine Migration

A key enabler for data center cost savings, consolidation, flexibility and application scalability has been the technology of compute virtualization or Virtual Machines (VMs)[XEN]. A VM to the software application looks like a dedicated processor with dedicated memory and dedicated operating system. In modern data centers or "computing clouds" the smallest unit of computing resource is the VM [Nurmi]. In public data centers one can buy computing capacity in terms of VMs for a particular amount of time. Though different VM configurations may be offered that are optimized for different types of processing (e.g., memory intensive, throughput intensive)[EC2].

VMs offer not only a unit of compute power but also as an "application environment" that can be replicated, backed up and moved [Clark]. Although VM migration started in the LAN, Wide area VM migration has also been discussed in the literature, e.g., [Brad]. The impact of VM migration on the network and hence other services has just recently been studied along with some mitigation approaches [Stage].

Virtual machine migration has a variety of modes: (i) scheduled vs. dynamic; (ii) bulk vs. sequential; (iii) point-to-point vs. point-to-multi-point. Network capability can impact virtual machine migration strategy. For certain mission critical applications, bandwidth guarantee as well as performance guarantee must be provided by the network. Make-before-break capability is also critical to support seamless migration.

For certain applications such as disaster recovery, bulk migration is required on the fly, which may necessitate concurrent computation and path setup dynamically.

2.4. Entities Involved

We have the data center provider, a possibly separate application provider, and the user (See Figure 2). Note that the data center provider and the application provider may be potential competitors. In addition network providers may also offer data center services, making them potential competitors to an independent data center provider. Hence, for cross stratum optimization, understanding of various trust relationships is important when developing interfaces application/network interfaces.

Figure 2 illustrates key entities involved.

```
     ------------       ----------------------
    |  End-User  |-----| Application Provider |-----
     ------------       ----------------------      |
           |                      |                 |
           |            ----------------------      |
           |           | Data Center Provider |     |
           |            ----------------------      |
           |                      |                 |
           |            ----------------------      |
            -----------|   Network Provider   |-----
                        ----------------------
```

Figure 2: Key Entities involved in CSO

2.5. Load Balancing

As the application servers are distributed geographically across many Data Centers for various reasons (e.g., load balancing), the decision as which server to select for an application request from end-users has many factors that can negatively affect the quality of experience (QoE) of the users if not done correctly. One of the major drivers for operating multiple Data Centers is allowing the application to be closer to the end-users, so that the overall service performance and the user experience can be enhanced.

Among the key factors to be considered in choosing the server for an application or instantiating VM include:

- The utilization of the servers;
- The underlying network loading conditions within a data center (LAN);
- The underlying network loading conditions between data centers (MAN/WAN);
- The underlying network conditions between the end-user and data center.

2.6. End-user capability and communication

As there are plethora of end-user terminal types (e.g., desktop device, PDA, mobile phones, etc.), it is important for application to capture end-user device capability and preference. For some applications, the same user may have multiple devices. In such case, seamless device to device transition needs to be provided by application providers to ensure acceptable QoE to the end-users.

For other applications, codec capability and/or terminal screen dimension of end-user devices may also have impact on QoS and bandwidth requirements.

Hence, the interface between end-user and application may need to be enhanced to capture these aspects.

3. Deployed Applications, Services, and Products

Most current methods are associated with IP networks. For instance, Akamai and other content distribution networks (CDN) carriers, have used some IP network knowledge to optimize their application overlay network usage. When selecting the surrogate (cache or mirror) location from the client location, many CDN providers use network latency via a probing technique or proximity based on static configuration to determine the optimal surrogate location. These overlays are not closely integrated with carrier's network real load condition such as link bandwidth utilization and availability. For many current and emerging applications that require stringent QoS and bandwidth guarantee, current CDN infrastructure is not well suited for meeting such service need.

The IETF ALTO WG has focused on overlay optimization among peers by utilizing information about topological proximity and appropriate geographical locations of the underlay networks. With this method, the optimization generally occurs in selecting peer location which will help reduce IP traffic unnecessarily traversing IP service providers. Current scope of this work does not address general problems this document has been discussing such as the selection of application servers based on resource availability and usage of the underlying networks.

In some cases, application controllers can estimate network load based on ping latency, and network topology based on trace routes in the Internet, based on the assumption that the underlying transport network is an IP network, and the routing is based on simple IP forwarding.

In regards to load balancing, DNS redirect technique is currently used to redirect end-user request to certain servers that host end-user application.

In the current Intra-Data Center network, the server selection for an application/VM is done by load-balancer. The load balancer is aware of a certain level of server usage data (e.g., the number

simultaneous instances of the application usage) and distributes the application requests based on that data.

However, the current load balancing technology is insufficient in providing an optimal decision across multiple VLANs and multiple Data Centers. This capability is often referred to as global load balancing.

First of all, there is no good mechanism for the communication exchange among load balancers located in different Data Centers. This implies that load balancers from different vendors cannot communicate to each other.

Secondly, load balancers know little about the underlying network conditions listed in the previous section. Nor is it user condition aware.

When migrating existing VMs/applications from one data center to another, the underlying network load condition in LAN/MAN/WAN can be constraining factors. Migration of VMs/applications, for instance, typically requires a high-speed data transfer across LAN/MAN/WAN to minimize service impact. Application controllers responsible for this operation is not aware of LAN/MAN/WAN network conditions.

4. Research Program

In the previous sections we have looked at key issues in Data Center and Cloud Computing and some commercial service deployments on a variety of cross layer optimization problems.

A common theme to the previous work was that sharing information between the application and network stratums can lead to more optimal solutions to the challenges facing distributed applications. In addition to sharing information, both the application layer and network may possess capabilities that are can very useful to each other if appropriate access can be arranged, e.g., the dynamic high bandwidth services that are enabled by MPLS/GMPLS.

Hence this research project is focused on the interfaces and services that could be used between the application and network stratum to address the four main problem thrusts of:

1. Joint application/network Resource optimization (global load balancing)

2. Responsiveness to quickly changing demands from/to application to/from network

3. Enhanced service resilience (via cooperative recovery techniques between application and network)

4. Quality of application experience (QoE) enhancement (via better use of existing network and application resources)

Even though algorithms play a big part of optimization, in thrust (1) we are concerned with the information that could be shared to promote optimization and various optimization criteria rather than specific algorithms. Note that this is similar to the approach taken with MPLS-TE, GMPLS and PCE where specific algorithms are not standardized.

4.1. Tentative Research Deliverables

a) Baseline network/application model - general enough to include most cases of interest but no more.

b) Survey the various "trust or lack of" in the relationships between various key players in both the application and network stratum. Include a survey of various "summarization", "abstraction", or other techniques that can reduce the level of "trust" needed at an interface.

c) Survey of the data center/cloud based applications - investigate the commonality and differences with respect to their impact on network infrastructure.

d) Define key interfaces and their functionality and relate these to current standards and potential future standards.

e) Investigation and report on the role of TE based network infrastructure (MPLS, GMPLS) in providing support to dynamic application loads, scaling and QoE enhancement.

f) Report on mechanisms for application level support for network recovery and network support for application recovery.

g) Investigate the time frames and responsiveness of interest to application/network interaction. For example what do various applications need, what can the network provide, can other techniques such as time based "load shifting" be utilized.

5. Security Considerations

TBD

6. IANA Considerations

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

7. References

7.1. Informative References

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

[Brad]   R. Bradford, E. Kotsovinos, A. Feldmann, and H. Schioberg, "Live wide-area migration of virtual machines including local persistent state," in Proceedings of the 3rd international conference on Virtual execution environments, San Diego, California, USA, 2007, pp. 169-179.

[Carter] R. L. Carter and M. E. Crovella, "Server selection using dynamic path characterization in wide-area networks," in INFOCOM '97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE, 1997, vol. 3, pp. 1014-1021 vol.3.

[Chamber] C. Chambers and W.-chang Feng, "Patch scheduling for on-line games," in Proceedings of 4th ACM SIGCOMM workshop on Network and system support for games, Hawthorne, NY, 2005, pp. 1-6.

[Clark]   C. Clark et al., "Live migration of virtual machines," in Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation - Volume 2, 2005, pp. 273-286.

[CostCloud] A. Greenberg, J. Hamilton, D. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," ACM SIGCOMM, Vol. 39, Number1, January 2009.

[GFD-122] Tiziana Ferrari (editor), "Grid Network services Use Cases from the e-Science Community", GFD-I-122, Open Grid Forum, December 12, 2007.

[Gargo] S. Gargolinski, C. S. Pierre, and M. Claypool, "Game server selection for multiple players," in Proceedings of 4th ACM SIGCOMM workshop on Network and system support for games, Hawthorne, NY, 2005, pp. 1-6.

[Grampin] E. Grampin, A. Castro, M. German, F. Rodriguez, G. Tejera, and M. Sanguinetti, "A PCE-based Connectivity Provisioning Management Framework," in Network Operations and Management Symposium, 2007. LANOMS 2007. Latin American, 2007, pp. 76-83.

[Habib] I.W. Habib, Qiang Song, Zhaoming Li, and N. S. V. Rao, "Deployment of the GMPLS control plane for grid applications in experimental high-performance networks," Communications Magazine, IEEE, vol. 44, no. 3, pp. 65-73, 2006.

[Kris] P. Krishnan, D. Raz, and Y. Shavitt, "The cache location problem," Networking, IEEE/ACM Transactions on, vol. 8, no. 5, pp. 568-582, 2000.

[Krishna] B. Krishnamurthy, C. Wills, and Y. Zhang, "On the use and performance of content distribution networks," in Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement, San Francisco, California, USA, 2001, pp. 169-182.

[Kurc] A. R. Kurc, D. Raz, and Y. Shavitt, with Cheng Jin, "Constrained mirror placement on the Internet," Selected Areas in Communications, IEEE Journal on, vol. 20, no. 7, pp. 1369-1382, 2002.

[Martini] B. Martini, V. Martini, F. Baroncelli, K. Torkman, and P. Castoldi, "Application-Driven Control of Network Resources in Multiservice Optical Networks," Optical Communications and Networking, IEEE/OSA Journal of, vol. 1, no. 2, p. A270-A283, 2009.

[Meng] X. Meng, V. Pappas, and L. Zhang, "Improving the Scalability of Data Center Networks with Traffic-aware Virtual Machine Placement," in 2010 Proceedings IEEE INFOCOM, San Diego, CA, USA, 2010, pp. 1-9.

[Nurmi] D. Nurmi et al., "The Eucalyptus Open-Source Cloud-Computing System," in Cluster Computing and the Grid, 2009. CCGRID '09. 9th IEEE/ACM International Symposium on, pp. 124-131, 2009.

[Qiu] V. N. Padmanabhan and G. M. Voelker, with Lili Qiu, "On the placement of Web server replicas," in INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, 2001, vol. 3, pp. 1587-1596 vol.3.

[Quax] P. Quax, J. Dierckx, B. Cornelissen, G. Vansichem, and W. Lamotte, "Dynamic server allocation in a real-life deployable communications architecture for networked games," Proceedings of the 7th ACM SIGCOMM Workshop on Network and System Support for Games, Worcester, Massachusetts: ACM, 2008, pp. 66-71.

[Ratnas] S. Ratnasamy, M. Handley, R. Karp, and S. Shenker, "Topologically-aware overlay construction and server selection," in INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, 2002, vol. 3, pp. 1190-1199 vol.3.

[RFC2261] D. Harrington, et al., "An Architecture for Describing SNMP Management Frameworks," January, 1998.

[RFC2265] B. Wijnen, et al., "View-based Access Control Model (VACM) for the Simple Network Management Protocol (SNMP)," January, 1998.

[Stage] A. Stage and T. Setzer, "Network-aware migration control and scheduling of differentiated virtual machine workloads," in Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing, 2009, pp. 9-14.

[Tang] Xueyan Tang, "On replica placement for QoS-aware content distribution," in INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies, 2004, vol. 2, pp. 806-815 vol.2.

[WoWHrs] P. Tarng, K. Chen, and P. Huang, "An analysis of WoW players' game hours," Proceedings of the 7th ACM SIGCOMM Workshop on Network and System Support for Games, Worcester, Massachusetts: ACM, 2008, pp. 47-52.

[WoWAct] M. Suznjevic, M. Matijasevic, and O. Dobrijevic, "Action specific Massive Multiplayer Online Role Playing Games traffic analysis: case study of World of Warcraft," Proceedings of the 7th ACM SIGCOMM Workshop on Network and System Support for Games, Worcester, Massachusetts: ACM, 2008, pp. 106-107.

[XEN] P. Barham et al., "Xen and the art of virtualization," in Proceedings of the nineteenth ACM symposium on Operating systems principles, pp. 164-177, 2003.

[Y.1541] Network performance objectives for IP-based services, February, 2002.

[Y.2011] General principles and general reference model for Next Generation Networks, October, 2004.

[Y.2012] Functional Requirements and architecture of the NGN, April, 2010.

Author's Addresses

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

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

Ning So (Editor)
Univerity of Texas at Dallas
Email: ningso@yahoo.com

Tae Yeon Kim
ETRI
tykim@etri.or.kr

Kohei Shiomoto
NTT
Email : shiomoto.kohei@lab.ntt.co.jp

Oscar Gonzalez de Dios
Telefonica
Email : ogondio@tid.es

Intellectual Property Statement

Acknowledgment

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