Cross Stratum Optimization Research Group H. Yang
Internet-Draft YL. Zhao
Intended status: Informational J. Zhang
Expires: October 26, 2018 Beijing University of Posts and Telecommunicati
Y. Lee
Y. Lin
FT. Zhang
Huawei Technologies Co., Ltd.
April 24, 2018

Cross Stratum Optimization Architecture for Optical as a Service
draft-yangh-cso-oaas-14

Abstract

Data centers based applications provide a wide variety of services such as cloud computing, video gaming, grid application and others. Currently application decisions are made with little information concerning underlying network used to deliver those services so that such decisions cannot be the most optimal from both network and application resource utilization and quality of service objectives.

This document presents a novel architecture of Cross Stratum Optimization for application and network resource in dynamic optical networks. Several global load balancing strategies are proposed and demonstrated by experiments in Optical as a Service experimental environment.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

With the emergence of cloud computing and high-bandwidth video applications such as live concerts, sporting events and remote medical surgery, various data center applications become more and more important, some Quality of Service related parameters of which have attracted much attention, such as jitter and latency. Therefore, there is a great need for a joint scheduling of network and application resources, the latter of which mainly refers to computing and storage resource, such as servers of various types and granularities (memory, disk, VMs). Many studies have been focused on traffic awareness in application resource [1], especially cross layer optimization in optical network [2]. However, few of them have been involved in global combined optimization of network and application resources.

This document proposes a novel architecture based on Cross Stratum Optimization (CSO) [3] that enables a joint application/network resource optimization, responsiveness to quickly change demands from/to application to/from network, enhanced service resilience (via cooperative recovery techniques between application and network) and quality of application experience (QoE) enhancement (via better use of current network and application resources). This architecture is intended to enable Optical as a Service (OaaS) by enabling large-bandwidth and multi-granularities applications based on Adaptive Multi-service Optical Networks (AMSON) with an increased resource utilization and resiliency across the application and network stratums. Four strategies including global load balancing (GLB), random based (RB), application resource based (AB) and network resource based (NB) strategies are proposed and validated in our experimental environment. Experimental results show that GLB in CSO architecture performs more effective compared with others.

1.1. Conventions Used in This Document

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

2. Terminologies

AB: Application resource Based.

AC: Application Controller.

AMSON: Adaptive Multi-service Optical Networks.

ARAE: Application Resource Abstract Engine.

ASI: Application-Service Interface.

CSO: Cross Stratum Optimization.

DB: Data Base.

DBM: Data Base Management.

DCN: Data Center Network.

GLB: Global Load Balancing.

GMPLS: General Multi-Protocol Label Switching.

LSA: Link State Advertisement.

MIB: Management Information Base.

NB: Network resource Based.

NRAA: Network Resource Abstraction Algorithm.

NRAE: Network Resource Abstract Engine.

NRDB: Network Resource Database.

NMS: Network Management System.

OaaS: Optical as a Service.

OAM: Operation Administration and Maintenance.

OSPF: Open Shortest Path First.

PA: Protocol Agent.

PCE: Path Computation Element.

QoE: Quality of Experience.

RB: Random Based.

SA: Service Agent.

SA-PCE: Service-Aware PCE enhancement algorithm.

SC: Service Controller.

SCI: Service-Control plane Interface.

SMI: Service-Management Plane Interface.

SSE: Server/VM Selection Engine.

TED: Traffic Engineering Database.

UA: User Agent.

UAI: User-Application Interface.

VM: Virtual Machine.

3. CSO Functional Architecture for OaaS

The CSO functional architecture for OaaS is illustrated in Fig. 1 and Fig. 2.

```
    -------------------------------------                    ----------
   |                -------              |                  |          |
   |               |  SSE  |\            |                  |          |
   |              / -------  \ ------    |                  |          |
   |             /     |      |  DB  |   |   ------------   |          |
   |            /      |       ------    |--| User Plane |--|          |
   |           /       |      /          |   ------------   |          |
   |   ------ /     -------- /           |                  |          |
   |  |  UA  |-----|  ARAE  |       AC   |------------------|          |
   |   ------       --------             |                  |          |
    -----|-------------------------------                   |          |
         |                                                  |          |
         |                                                  |          |
    -----|------------------    -------------------------   |Management|
   | ------        ------   |  |   -----------           |  |  Plane   |
   ||      |      |      |--|--|--| Signaling |          |--|          |
   ||  SA  |------|  PA  |  |  |   -----------           |  |          |
   ||      |\     |      |--|--|-- ---------             |  |          |
   | ------  \     ------   |  |  | OSPF-TE |            |  |          |
   |   |      \       |     |  |   ---------Control Plane|  |          |
   |   |       \      |     |   -------------------------   |          |
   |   |        \     |     |                        |      |          |
   | --------    \ -------  |   ------------------   |      |          |
   ||  NRAE  |----|  PCE  |-|--| Other domain SCs |  |      |          |
   | --------      -------  |   ------------------   |      |          |
   |          \       |     |   -----------------------     |          |
   |           \      |     |  |    Transport Plane    |----|          |
   |            \  -------  |   -----------------------     |          |
   |     SC      \|  DBM  |-|-------------------------------|          |
   |               -------  |                               |          |
    ------------------------                                 ----------
```

Fig.1 CSO functional architecture for OaaS

```
          -------------------------------------------------------
         |   --------        ------------         ---------      |
         |  |  DCNs  |------|     AC     |-------|  Users  |     |
         |   --------      / ------------ \       ---------      |
         |                /       |        \ Application Stratum |
         |---------------/--------|---------\--------------------|
         |              /         |          \ Network Stratum   |
         |           ------     ------      ------               |
         |          |  SC  |   |  SC  |    |  SC  |              |
         |           ------     ------      ------               |
         |             |          |            |                 |
         |             |          |            |                 |
         |   ------------    ------------    ------------        |
         |  |            |  |            |  |            |       |
         |  |  domain A  |--|  domain B  |--|  domain C  |       |
         |  |            |  |            |  |            |       |
         |   ------------    ------------    ------------        |
          -------------------------------------------------------
```

Fig.2 CSO schematic for OaaS

The application stratum plane, service stratum plane and user plane are introduced in the novel architecture of CSO besides traditional planes, i.e., control plane, management plane and transport plane. The responsibility for centralized application stratum plane is concerned with maintaining application resources in data centers, while service stratum plane provides to application stratum the network resource information abstracted from control plane with NRAA. In addition, GLB computation is implemented based on both the application stratum and network stratum resources, while service stratum will enforce SA-PCE. The responsibilities and interactions among these entities are provided below.

3.1. AC

AC comprises UA, SSE, DB and ARAE. AC is responsible for interacting with user plane and obtaining network and application resource abstract information abstracted from SCs and DCNs. AC completes the GLB computation based on them. UA authenticates the user requests and maintains user information. With GLB computation, SSE chooses the optimal server or VM for users, allocates application resources, and determines the location of the distributed application or where to migrate virtual machines. ARAE provides to GLB computation the suited application resource abstract information obtained from DCNs, such as running state and idle resource of servers or VMs.

3.2. SC

SC is composed of SA, PA, PCE, NRAE and DBM. Three main functional requirements for SC in OaaS architecture are described below. Firstly, SC provides network services to AC. According to the type of services, SC computes the paths and drives control plane to establish the paths so as to implement the concept of OaaS. Secondly, SC offers to AC the resources abstract information including the mapping of application and optical layer, logical topology of optical layer and the status of network transmission for AC decision. Finally, it provides to management plane the database interface so that network administrator can monitor it.

SA communicates to AC with authentication and access control permission of transport network resources through ASI. SA also translates AC profile into connection and service parameters in transport network which contains bandwidth, delay, jitter and others. PA drives the GMPLS signaling of control plane and receives the routing information. PCE enforce SA-PCE while NRAE abstracted from control plane with NRAA. In addition, TED, NRDB, MIB and configuration are contained in DBM.

4. Advantage of CSO Architecture for OaaS

CSO Architecture for OaaS is the spread of traditional three planes, i.e., control plane, management plane and transport plane. The decisions based on CSO architecture for OaaS can be the most optimal and have the least cost from both application and network resource utilization, while the quality of user experience can reach the highest in this architecture. According to various demands and expenses of different server providers, the operator can provide to them abstract topologies with NRAA so that this mechanism guarantees the security between operator and server provider or among server providers. Since the CSO architecture for OaaS is based on new strategies and algorithms, the spread of current network may be just software promotional and the architecture is provided with the higher expansibility and flexibility.

5. CSO Procedure in CSO Architecture for OaaS

When the UA in AC receives the application request from user plane, it will forward this request to SSE after authenticating the user requests. The certified request is analyzed via SSE and transmitted to ARAE for the application resource information. SSE receives the network abstract information from SC via AC gateway upon request. ARAE responds to SSE the suited application resource abstract information obtained from DCNs, according to the analysis result from it. Upon completing the GLB computation based on application and

network abstract resource, and SSE chooses the most optimal server or VM for users, allocates application resources, and determines the location of the distributed application or where to migrate virtual machines. According to service type, resources occupancy rate and QoE, UA performs accounting function and transmits the application requirements to SC via ASI. UA receives the responses to NRAE and returns to UA. Rating the service based on the distribution of resources and returning the feedback, UA provides to user stratum the resources at last. When SA receives the location of the server/VM and the service type, it will translate this profile into connection and service parameters in transport network which contains bandwidth, delay, jitter and others after authentication and access control permission to this requirement. SA also forwards the network resource profile to PCE at the same time. Completing SA-PCE computation that factors in the connection and service parameters constraints, SA-PCE provides the explicit route to PA. Then using the RSVP signaling protocol, PA drives control plane to establish the path through SCI. After the path is setup successfully, it will conserve the information of the path into DBM and return overall results including transport network resource to AC. After receiving the OSPF LSA from control plane, PA provides it to DBM for network resources synchronization. AC obtains application and network information periodically or based on event-based trigger. Meanwhile, NRAE interacts with network TE topology information base and DBM for abstracting network resource. NRAE provides abstract information to the authorized AC using NRAA.

6. Different Application Scenarios

6.1. Network Resource Acquirement

SCs receive the OSPF LSA from control plane to obtain the completely TE topology information network and provide it to DBM for network resources synchronization. AC obtains application and network information periodically or based on event-based trigger. Based on NRAA, SCs computes the abstract topology and feedback to AC.

6.2. Virtual Migration Request

Due to the insufficiency of network or servers/VMs resource, or the abrupt emergency to servers or network, or the requirement of saving energy consumption, Virtual migration request becomes significant in reality application. Virtual migration migrates to the destination server with multi-granularities and the choice of destination one follows the procedure of CSO in OaaS architecture.

6.3. Exception Handling

When unexpected error happens in the process of CSO, SC will receive GMPLS OAM from control plane and provide the alarm information to AC and saves into DBM. SC needs to route again as the service delivery process.

7. Definition of New Interfaces in CSO Architecture for OaaS

Due to additional planes in OaaS architecture, new interfaces between themselves, which contain ASI, UAI, and which between them and traditional planes in GMPLS containing SCI, SMI is to be defined in this section. Nevertheless, only functional requirement will be demonstrated for each of above-mentioned interfaces, by which service of OaaS and Cross Stratum Optimization could work well.

7.1. Functional Requirement for UAI

UAI is the interface between user plane and application plane, which conveys the user's application request from user plane to application plane and the reply information. Such user denotes the general users who apply for the application, not only includes the particular clients asking for video service, but also revolves the service provider managing the application resource such as virtual migration. In other words, managers of the service provider access the application Plane by the same interface, even if the permission will differ common users.

Whatever kinds of application request is submitted, UAI should transmit the request information transparently, which consists of the user identity, request type, specified information.

7.2. Functional Requirement for ASI

ASI is the interface between service plane and application plane, which conveys the request for optical service of all application, containing path establishment request and network resource abstract request. The latter is foundation to CSO, because the replied abstract information will be referred to for application plane to make a judgment, such as selecting a proper datacenter for a user or to which migrating virtual machines. Therefore, the interface from SC to AC should convey the whole abstraction information, which is abstracted and packed by abstracting module in SC, as well as optical service reply.

As to the common request for optical service, the request information must include the service style, such as VOD and virtual migration, and the source and destination node in optical layer of this service.

The reply of which also contains the path establishment result and if it is failure, the reason should be given.

7.3. Functional Requirement for SCI

SCI is the interface between service plane and control plane. The message transmitted through this interface is standard GMPLS including OSPF and RSVP messages, which is easily compatible to GMPLS control plane.

7.4. Functional Requirement for SMI

SMI is the interface between service plane and management plane. The database of the network information maintained by SC, could supply some detailed network operating condition for management plane to make decision, and management plane also can issue OAM commands to SC. Both state information and OAM message will be defined by SMI.

8. CSO Strategies and Algorithms

Based on functional architecture of CSO-OaaS described above, we propose four strategies including GLB strategy based on CSO, RB, AB and NB strategies. These strategies and related algorithms are described in detail below.

With RB strategy, the destination node of data center server is randomly selected by control plane when the application request comes. With AB strategy, according to the CPU, memory, disk utilization and I/O scheduling, control plane chooses the server node having the minimum application utilization as the destination. NB strategy selects the node which has the path of the minimum network hop from the source to the destination. With GLB strategy, as described in previous sections, AC selects the server node and the DC location based on the application status collected from data center networks and the network condition provided by SCs dynamically.

We define alpha as the joint optimization factor to measure the balance between the network and application resources, which contains the application and network parameters. Three application parameters, current memory utilization Ur which models RAM, CPU usage Uc and the utilization of I/O scheduling Us describe the current usage of data center application resource. The network parameters are comprised of the TE weight Bl and delay tl which is related to traffic cost and delay of the current link and the hop Hp of the candidate path. These parameters are normalized to meet the linear relationship between them. The application function with application parameters of current each node is expressed as dimensionless overall function fac(Ur,Uc,Us,k) = kc*Uc+kr*Ur+ks*Us, kc+kr+ks=1, kc>=0,

kr>=0, ks>=0, where kc,kr,ks are adjustable evaluation rank rate among CPU, RAM utilization and I/O scheduling. Initially, the evaluation rank of CPU is the highest of all, while the rank of RAM is higher than I/O scheduling. At this point, evaluation ranks satisfy the expressions as follows: kc=Ra, kr=Rb, ks=Rc, Ra+Rb+Rc=1, Ra>=Rb>=Rc, where Ra,Rb,Rc are constants and their priorities decrease increasingly. That means the higher utilization corresponds to higher priority. Once Ur or Us exceeds Uc, for instance Ur>=Uc>=Us, the evaluation rank of them will adjust according to this change as follows: kc=Rb, kr=Ra, ks=Rc. By parity of reasoning, kc,kr,ks will modify dynamically based on the feedback of utilization variation. In addition, network function with parameters of current each node is expressed as dimensionless overall function fbc(Bl,Hp,tl) = kB*(B1+B2+...+Bl+...+BHp)/B*Hp+kt*(t1+t2+...+tl+...+tHp)/t*Hp, which the candidate path is calculated by the network stratum resources with candidate server destination nodes chosen by AC. fa1, fa2,...,fak are the application functions with parameters among the K candidate server nodes and fb1, fb2,...,fbk are the network functions with parameters associated with the K candidate paths. So the joint optimization factor alpha meets the formula as follows. In this formula, beta is the dynamic weight between the network and application parameter, which associates with the variance of application parameters from each server node. The variance is related to DC load balancing degree, while the larger variance represents balancing degree becomes worse in DCs. Based on the formula described below, the application utilization weight changes dynamically according to the feedback of load balancing degree. At first, the weight of application utilization is relatively smaller due to the lower application parameters variance. With the increasing of application parameters variance, the application utilization weight turns into higher, which miu is normalizing factor of beta. The formula is alpha = [fac(Ur,Uc,Us,k)/max(fa1,fa2,...,fak)]*beta + [fbc(Bl,Hp,tl)/max(fb1,fb2,...,fbk)]*(1-beta), beta = miu*sqrt{var(fa)/max[var(fa1),var(fa2),...,var(fak)]}.

According to application utilization, AC first chooses the K candidate server nodes in application stratum, which can provide this type of application. In network stratum, the node with minimum alpha value based on the joint optimization factor will be selected from the K candidates. In all schemes, the path will be reserved and setup through signalling protocol between the source and destination node after the choice of the node.

9. CSO Experiment and Demonstration

9.1. CSO Experimental Environment

Experimental environment is built to support the architecture of CSO and deployed in five servers, while each server mounts virtual machines created by VMware software running at servers. Since each virtual machine has the operation system and its own computation resource, the virtual OS technology makes it easy to set up experiment topology based upon NSFNET with 14 control plane nodes. In addition, Network Management System (NMS) is placed to monitor and initialize the transport plane elements, while NMS is an inseparable management system which manages the overall network.[4] The service application usage is selected randomly from 1% to 0.1% for each application demand and network bandwidth required for each application is assumed one wavelength equivalent. Each node supports 40 wavelengths with no wavelength conversion or 3R regeneration capability.

9.2. CSO Experimental Results

Based on CSO functional architecture described above, GLB strategy based on the cross-stratum optimization is implemented and experimentally compared with RB, AB and NB strategies in CSO Experimental environment. The experimental results are shown in Tab. 1-4. Tab. 1 illustrates load balancing degree resulting from RB, AB, NB and GLB strategy. The load balancing degree is defined as the variance of application utilization in each data center server. The higher load balancing degree is, the worse the effect of load balancing is. As shown, GLB strategy leads to much lower load balancing degree than RB and NB strategy, but higher than AB strategy. In fact, AB strategy computes the node only considered application utilization, the path may not be able to setup because it does not have enough wavelength resource. In Tab. 2, GLB has less network blocking probability than RB and AB strategies. Tab. 3 shows that GLB approach has less average hop than RB and AB strategies obviously, for it factors the latency. With the increase of offered load, the curve of GLB scheme gets closer to NB. In Tab. 4, global blocking probability measures both the network and application blocking situation measured by CPU and memory overflow. Though AB approach has lower load balancing degree and similar average hop is computed through NB scheme, GLB strategy has significantly lower integrated blocking probability than all other approaches.

Traffic load	Load balancing degree			
	RB	AB	NB	GLB
100	0.00594	7.65E-5	0.05639	0.00333
200	0.00951	7.77E-5	0.10181	0.00361
300	0.01286	7.85E-5	0.12019	0.0036
400	0.01409	7.49E-5	0.12352	0.00334
500	0.01198	7.8E-5	0.12043	0.00303

Tab.1 Load balance factor of four strategies

Traffic load	Network blocking probability			
	RB	AB	NB	GLB
100	0.00002	6.5E-4	7.6E-4	5E-5
200	0.01902	0.01866	0.02152	5.2E-4
300	0.08462	0.09368	0.05992	0.03628
400	0.15036	0.17944	0.08968	0.12418
500	0.19862	0.25528	0.10462	0.18104

Tab.2 Network blocking probability of four strategies

Traffic load	Average hop			
	RB	AB	NB	GLB
100	5.50661	5.5058	3.604	4.2668
200	5.48937	5.4813	3.59706	4.25557
300	5.42255	5.40668	3.56946	4.23117
400	5.34908	5.31895	3.5374	4.1668
500	5.28607	5.21635	3.50851	4.0981

Tab.3 Average hop of four strategies

Traffic load	Global blocking probability			
	RB	AB	NB	GLB
100	2E-5	6.5E-4	0.00162	5E-5
200	0.02902	0.01866	0.06412	5.2E-4
300	0.0975	0.09368	0.1776	0.03628
400	0.18458	0.17944	0.2843	0.12864
500	0.27046	0.25528	0.36988	0.19704

Tab.4 Global blocking probability of four strategies

10. Security Considerations

TBD

11. Acknowledgments

The RFC text was produced using Marshall Rose's xml2rfc tool.

12. References

12.1. Normative References

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

12.2. Informative References

[Ref1] Meng, Xiaoqiao., Pappas, V., and Li. Zhang, "Improving the Scalability of Data Center Networks with Traffic-aware Virtual Machine Placement", May 2010.

[Ref2] Christodoulopoulos, K., Manousakis, K., and E. Varvarigos, "Cross Layer Optimization of Static Lightpath Demands in Transparent WDM Optical Networks", July 2009.

[Ref3] Lee, Young., Bernstein, Greg., So, Ning., Kim, Tae., Shiomoto, Kohei., and Oscar. Dios, "draft-lee-cross-stratum-optimization-datacenter-00", March 2011.

[Ref4] Zhang, Jie., Chen, Xue., and Yuefeng. Ji, "Experimental Demonstration of a DREAM-based Optical Transport Network with 1000 Control Plane Nodes, ECOC2011", September 2011.

Authors' Addresses

Hui Yang
Beijing University of Posts and Telecommunications
No.10,Xitucheng Road,Haidian District
Beijing 100876
P.R.China

Phone: +8613466774108
Email: yang.hui.y@126.com
URI: http://www.bupt.edu.cn/

Yongli Zhao
Beijing University of Posts and Telecommunications
No.10,Xitucheng Road,Haidian District
Beijing 100876
P.R.China

Phone: +8613811761857
Email: yonglizhao@bupt.edu.cn
URI: http://www.bupt.edu.cn/

Jie Zhang
Beijing University of Posts and Telecommunications
No.10,Xitucheng Road,Haidian District
Beijing 100876
P.R.China

Phone: +8613911060930
Email: lgr24@bupt.edu.cn
URI: http://www.bupt.edu.cn/

Young Lee
Huawei Technologies Co., Ltd.
Huawei Base,Bantian,Longgang District,Shenzhen
Shenzhen 518129
P.R.China

Email: leeyoung@huawei.com
URI: http://www.huawei.com/

Yi Lin
Huawei Technologies Co., Ltd.
Huawei Base,Bantian,Longgang District,Shenzhen
Shenzhen 518129
P.R.China

Email: yi.lin@huawei.com
URI: http://www.huawei.com/

Fatai Zhang
Huawei Technologies Co., Ltd.
Huawei Base,Bantian,Longgang District,Shenzhen
Shenzhen 518129
P.R.China

Email: zhangfatai@huawei.com
URI: http://www.huawei.com/