Network Working Group
Internet-Draft
Intended status: Informational
Expires: December 1, 2025

W. Sun
J. Shao
W. Hu
SJTU

June 1, 2025

Resource Allocation Model for Hybrid Switching Networks
draft-sun-nmrg-hybrid-switching-12.txt

Abstract

The fast increase in traffic volumn within and outside Datacenters is placing an unprecendented challenge on the underline network, in both the capacity it can provide, and the way it delivers traffic. When a large portion of network traffic is contributed by large flows, providing high capacity and slow to change optical circuit switching along side fine-granular packet services may potentially improve network utility and reduce both CAPEX and OpEX. This gives rise to the concept of hybrid switching - a paradigm that seeks to make the best of packet and circuit switching.

However, the full potential of hybrid switching networks (HSNs) can only be realized when such a network is optimally designed and operated, in the sense that "an appropriate amount of resource is used to handle an appropriate amount of traffic in both switching planes." The resource allocation problem in HSNs is in fact complex ineractions between three components: resource allocation between the two switching planes, traffic partitioning between the two switching planes, and the overall cost or performance constraints.

In this memo, we explore the challenges of planning and operating hybrid switching networks, with a particular focus on the resource allocation problem, and provide a high-level model that may guide resource allocation in future hybrid switching networks.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on January 26, 2025.

1. Introduction

In facing rapid increase of network traffic [Gantz12], as well as the number of servers in cloud data centers [Cisco15], new architectures and operation models of Data Center Networks (DCNs) gained wide interests. One concept that attracted considerable and lasting attention is the introduction of optical switching technologies into DCNs, hoping that bypassing some of the traffic without performing per-packet electronic processing will help reducing the Operational Cost (OpEx), as well as the Capital Expenditure (CapEx) of DCNs. This concept of combining electronic packet switching (EPS) and optical switching (often optical circuit switching, OCS), is called hybrid switching [Zukerman89]. In recent years, many hybrid switching schemes have been proposed [Gauger06], and it is reasonable to believe that when a DCN grows beyond a certain scale, the benefit of introducing optical switching will emerge and become more evident as the size of the DC continues to increase.

On the other hand, achieving the benefits of hybrid switching requires careful design at the planning stage, and proper operation during runtime. This poses challenges that goes far beyond the topological or architectural aspects. For instance, at the planning stage, one has to decide how much to invest in the two switching planes, such that each could be fully utilized when the network becomes operational. Under cases when dynamic resource allocation between the two planes are possible, one has to decide how resource is allocated between the two planes, and how traffic should be directed to each of them, such that performance constraints can be satisfied, and operational cost such as power consumption can be minimized.

This memo aims to explore the challenges of planning and operating hybrid switching networks, and provide a high-level model that may guide the resource allocation in future hybrid switching networks. We will use hybrid switching DCN as an example to show one possible application of this model.

2. 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.

3. Overview of Hybrid Switching Networks

Hybrid Switching Networks (HSNs) are networks that employ more than one switching technology. The term started to attract attention when Wavelength Dense Multiplexing (WDM) started to be deployed as a underlying infrastructure of TCP/IP based packet networks [FENG17]. It continued to receive considerable attention, as the research on future-looking optical switching schemes boomed, around and after the begining of the 21th century.

The research on hybrid switching gained momentum again with the rapid growth of cloud data centers. With a clearer context and real-life prototyping efforts, a wider concensus regarding the benefits and feasibility of HSN have been reached.

The challenges of planning and operating hybrid DCNs are rooted in the fundamental differences between EPS and OCS [Farrington10], [WANG10]. In principle, EPS is good at delivering traffic that is bursty and difficult to predict. By aggregating the traffic from a large number of communcating peers, high network utilization can be achieved at modest cost. OCS, on the other hand, is suited for well planed, or highly predictable traffic patterns. One good example is the delivery of bulk flows, which can last up to a few minutes when carried by a wavelength channel at full capacity.

4. Terms used in this document

- Electronic Packet Switching (EPS)
  EPS in this memo refers to the off-the-shelf switching technology. It provides "best-effort" packet delivery service. Since EPS performs fine-granular per-packet processing, it is generally regarded to be best suitable for traffic that is bursty and difficult to predict. Existing researches show that when lightly loaded, the performance of EPS can be rather reliable and predictable. However, when the network is heavily loaded, the

performance of EPS will deteriorate very quickly and result in long queueing delay and high packet loss rate.

- Optical Circuit Switching (OCS)
  OCS in this memo refers to connection oriented network services based on optical switching technologies, such as MEMS or WSS based switches, and the like. The connection oriented nature of OCS requires the establishment of connections through signaling prior to data transfer. The capacity of each connection, for instance, a wavelength channel, often consumes a significant portion of the overall network capacity. Request blocking is thus difficult to eliminate in OCS, if not impossible.

- Hybrid Switching Networks (HSNs)
  HSNs in this memo refers to networks that: i) employ both EPS and OCS, and ii) accept data transfer request in both packet and stream/flow form. Upon entering the network, requests in packets form will be handled by the EPS plane, and requests in flow form will be handled by OCS following the connection provisioning procedures. This differs HSN from IP over WDM networks, where both switching schemes exist, but services start and terminate only on the IP layer, and standalone OCS service is not provided. Note that the boundary between packet and flow requests may not naturally exist. For instance, when flow level information is not available from outside the network, it will be up to the network to decide how traffic should be partitioned and then directed to either EPS or OCS.

5. Performance Measures in Hybrid switching Networks

5.1. Performance Measures in Electronic Packet Switching

Without loss of generality, performance of packet switching networks can be characterized by one or more of the following metrics:

- Packet loss rate - packet loss may happen when congestions occur. Statistically, in a given network, packet loss rate can be seen as a function of network load. Packet loss rarely happen when the traffic load is low. But when the load increases to a certain threshold in the network, or in part of it, packet loss rate may increase quickly as load continues to increase.
  Packet delay and jitter - like packet loss rate, packet delay is mostly stable and jitter is small when the network is lightly loaded. Delay and jitter will increase dramatically when network load increases.
  Flow completion time - flow completion time is a composite metric that relies on both packet loss rate and packet delay.

5.2. Performance Measures in Optical Circuit Switching

Performance of Optical Circuit Switching (OCS) is typically measured by request blocking rate, defined as the number of admitted requests over the total number of arrivals. In theory, blocking in OCS can not be eliminated. The planning of OCS is thus often a tradeoff between performance and cost, as in the case of conventional telephone networks, in which trunk capacity can be dimensioned with the Erlang-B formula.

6. BLOC - the Blocking LOss Curves

6.1. General Idea

To understand the resource allocation in HSNs, it is important to understand the interactions between the three components in the systems:

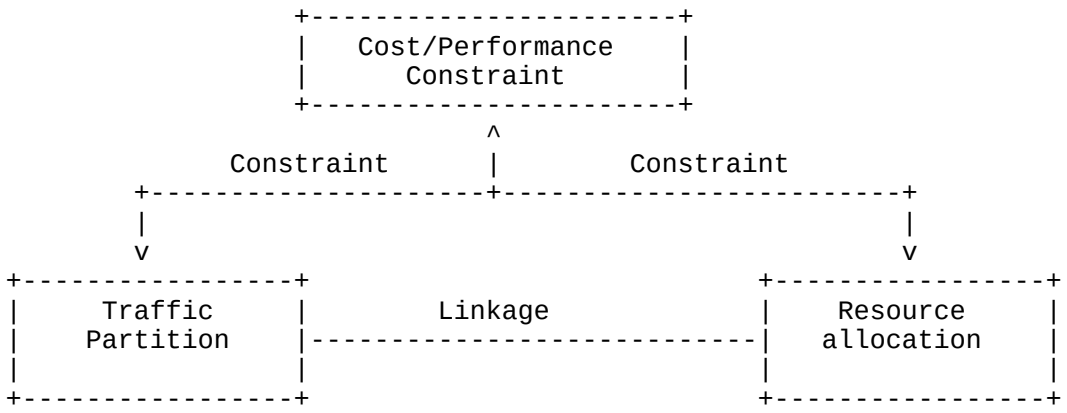
- Traffic partitioning
  Traffic partitioning means the separation of incoming traffic into two parts so that each part can be handled by the one of the two switching planes. In today's networks, there might be many traffic separation/differentiation mechanisms for the purpose of enforcing differentiated policy based on traffic type. Traffic partitioning in the context of HSN, however, aims to realize the optimal separation of flows into the two planes, such that the utility of the network can be maximized.
  One traffic partitioning method is a flow length based method. With a predefined threshold, flows are classified into short flows and large flows, each served with the packet switching plane and the circuit switching plane, respectively.
  Partitioning can be performed according to a priori knowledge, e.g., according to the information provided by the applications that generate the traffic flows. It also can be performed in network during runtime. The details on traffic classification and partitioning may be found in [Cisco15] and are outside the scope of this memo.

- Resource allocation
  The resource here can be physical resources such as switch ports, wavelengths or fibers. It also can be abstract resource such as the overall budget.

- Performance/Cost Constraints
  The cost constraint applies when the making of the hybrid switching system is subject to limited budget. For any given traffic demand, the cost and performance of carrying the traffic through either EPS or OCS can be very different. A good design

should, on the first hand, satisfy the performance constraint; on the other, it should also leave space for future traffic demand growth. The performance constraints specify the acceptable worst-case performance of the system, for example, the maximum packet loss rate, highest request-blocking rate, or longest packet delay etc. Given traffic demand, the worst-case performance constraint specifies the least amount of resource that should be allocated to a switching plane.

As can easily be seen, the operation of HSNs involves close interactions between the three components, and is a difficult problem. The interactions can be summarized into the following diagram.



Interactions beween the three components in HSN

6.2. Modeling the curves

In a typical IP network with a given traffic load, the packet loss rate decreases when the network capacity increases and vice versa. Similarly, in circuit switching networks, the request blocking probability will decrease when the bandwidth increases and vice versa. In a hybrid switching system, the overall resource capacity is constant. The resource allocation between EPS and OCS plane will directly affect the network performance of both switching planes. The network performance is also affected directly by how the traffic is partitioned between EPS and OCS planes.

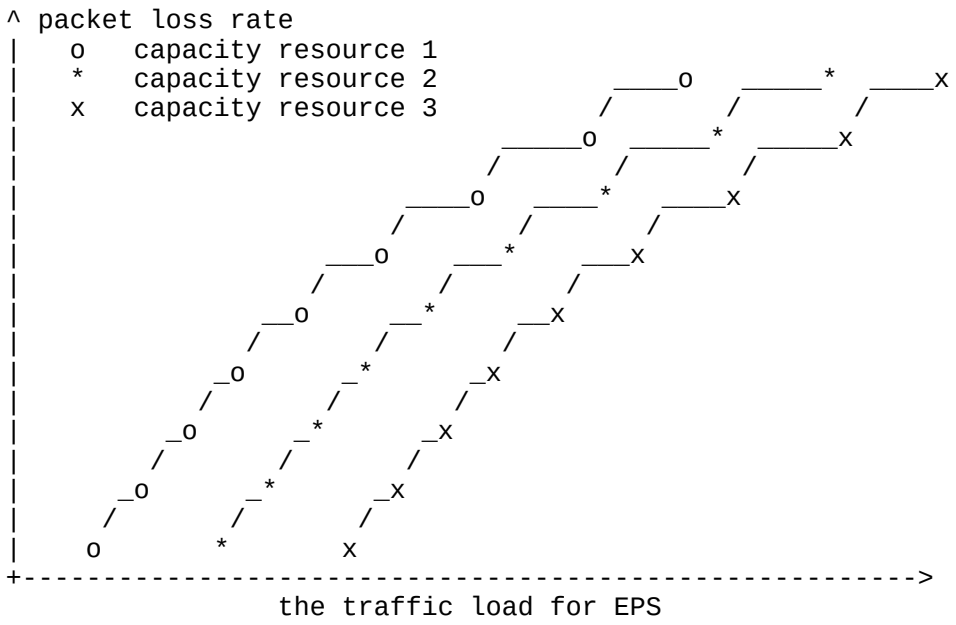


Fig. 1(a) the performance curves for EPS

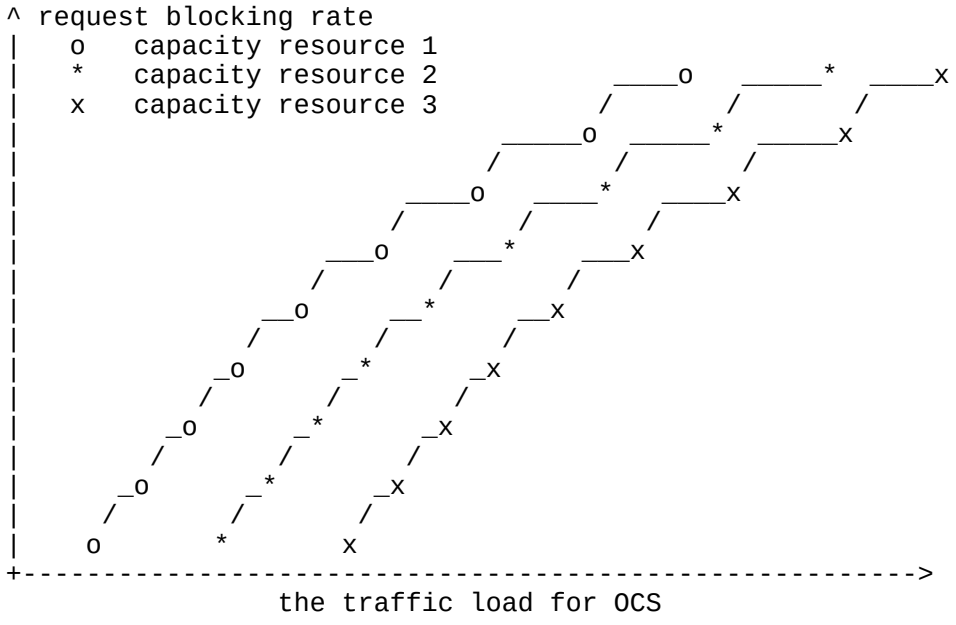


Fig. 1(b) the performance curves for OCS

To clearly classify the influence from the traffic load partition and network capacity allocation, we take Fig. 1(a) and 1(b) to show the performance curves for EPS and OCS with varying traffic loads and network capacities. In Fig. 1(a), we choose the packet loss rate as the performance of EPS. When the traffic load for EPS increases, the packet loss rate becomes worse under the constrained network capacity. The extension of network capacity will bring a promoted packet loss rate for EPS which is classified in Fig. 1(a) with the capacity resource 1 <= capacity resource 2 <= capacity resource 3. The performance curves for OCS is shown in Fig. 1(b) after choosing the request blocking rate as the y-axis, and the relationship among these curves is still resource 1 <= capacity resource 2 <= capacity resource 3. Combining Fig. 1(a) and 1(b), the more network capacities we allocate to EPS or OCS, the better service they will provide under a heavier traffic load transmission.

6.3. The BLOC System

The BLOC framework comprises two types of curves, i.e., loss curves (LCs) and blocking curves (BCs), in the same two-dimensional coordinate system. An LC or a BC in the BLOC framework is a curve that contains a series of points with the same packet loss rate or request blocking probability. Using the percentage of traffic delivered by EPS as the x-axis and the percentage of bandwidth allocated to EPS plane as the y-axis, all of the curves in the BLOC framework are monotonic. Another important component of the BLOC framework is the feasible region. In this paper, "feasible" means that as long as the traffic partitioning and resource allocation fall within this area, the resulting packet loss rate will be smaller than Pmax (the maximum packet loss rate) and the request blocking probability will be lower than Bmax (the maximum request blocking rate). Thus, the feasible region contains all the feasible combinations of resource allocation and traffic partitioning that satisfy the network performance requirements. Different resource allocation strategies in hybrid switching networks can be achieved by choosing a point from the feasible region.

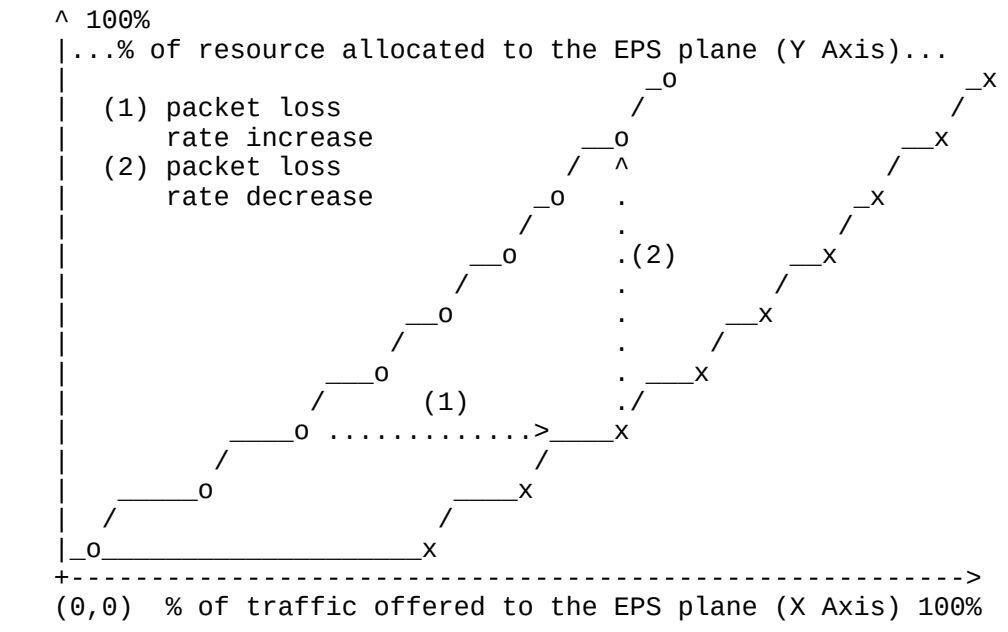


Fig. 2(a) the packet loss curves for EPS plane

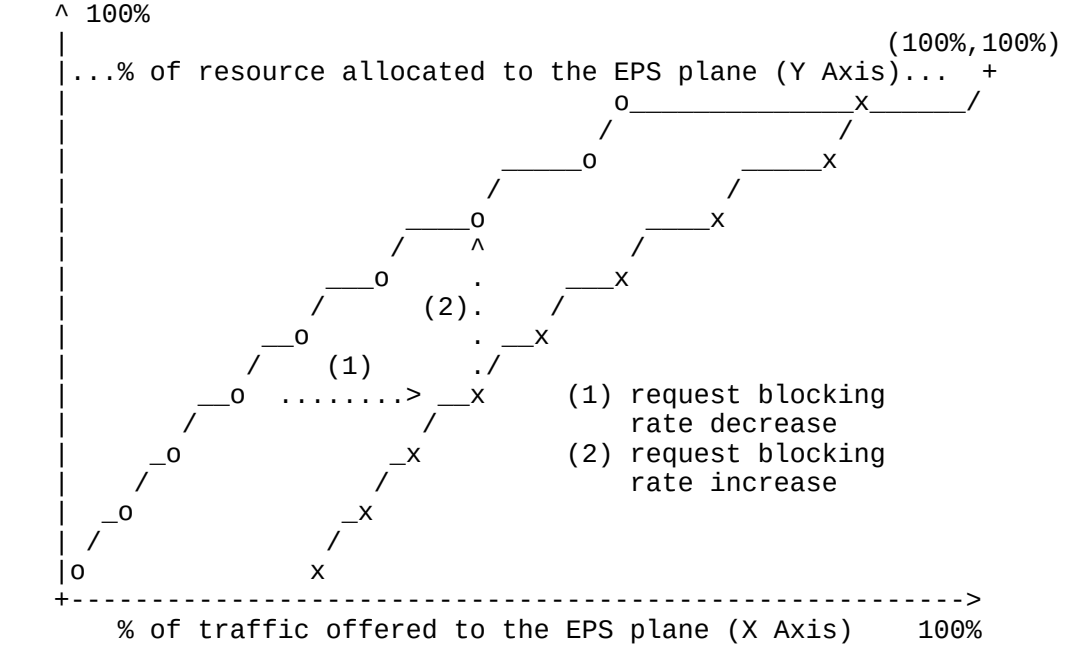


Fig. 2(b) the request blocking curves for OCS plane

Fig. 2(a) and 2(b) show an example of LCs and BCs when the overall hybrid system capacity and the traffic volume are fixed. In Fig. 2(a), when the percentage of traffic to be transmitted by EPS increases, the bandwidth allocated to EPS plane must also be increased so the same packet loss rate can be achieved. Hence, each LC is monotonically increasing. In addition, the LCs with smaller loss rate values require a larger percentage of bandwidth for the same amount of traffic. Therefore, the LCs moves to the top left when the packet loss rate becomes smaller, as shown in Fig. 2(a). All of the LCs pass through the origin (0, 0), so if no bandwidth is allocated to PS plane, it cannot transmit any traffic. Similarly, the BCs move downward to the right when the request blocking probability becomes lower, and all of the BCs converge to the point (100,100), where all of the bandwidth and traffic is assigned to PS plane, as shown in Fig. 2(b).

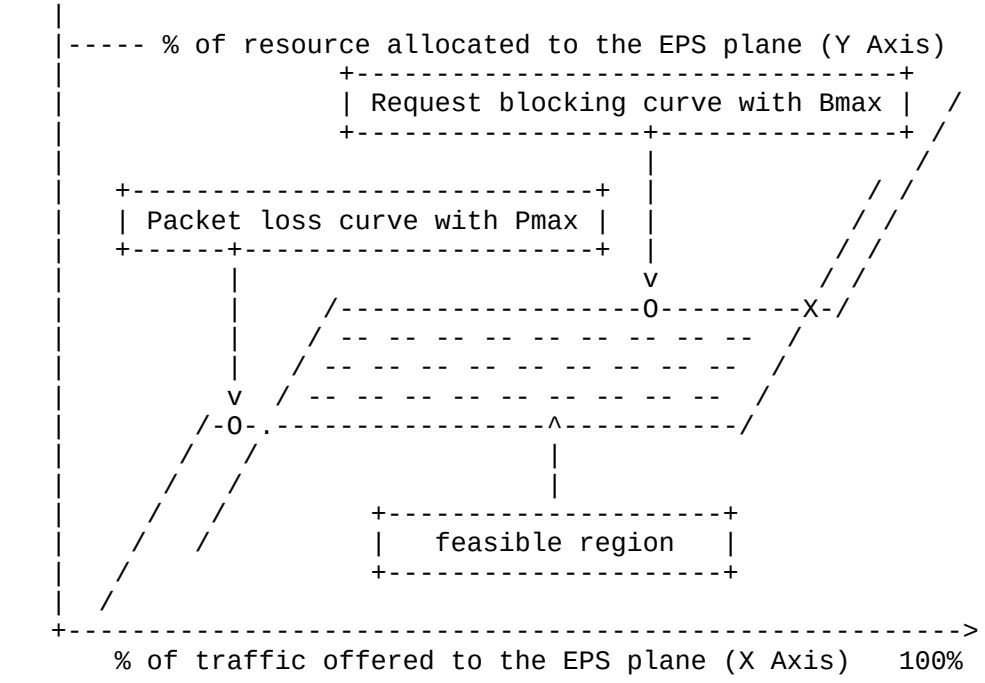


Fig. 3 an examplery BLOC

We now consider a hybrid switching system with the maximal allowed packet loss rate Pmax and the maximal allowed request blocking probability Bmax [FENG16]. Fig. 3 shows a BLOC where the LCs and BCs are placed in the same two-dimensional coordinate system. The hatched area above the LC of Pmax and below the BC of Bmax contains

all of the feasible combinations of traffic partitioning and resource allocation. Choosing a point from the feasible region (i.e., a combination of resource allocation and traffic partitioning) is subject to various optimization objectives. For instance, from an energy consumption perspective, we need to choose the point with the minimal percentage of EPS resources from the feasible region (i.e., the lowest point in the feasible region), so that the overall energy consumption would be minimized. In Section 5, we show that other metrics can also be optimized with the BLOC, such as the packet delay in EPS plane as a function of resource allocation and traffic partitioning.

6.4. An example

A hybrid switching Datacenter network is shown in Fig. 4 [FENG17]. Among all s+p uplink interfaces on each ToR switch, s of them connect the switch to the EPS plane and the rest, p, connect the ToR switch to the OCS network. As the cost of supporting an OCS connection can be very different from that of an EPS port, different combinations of s and p will result in significant difference in building cost. Different combinations of s and p will also lead to different performance and running cost, such as power consumption.

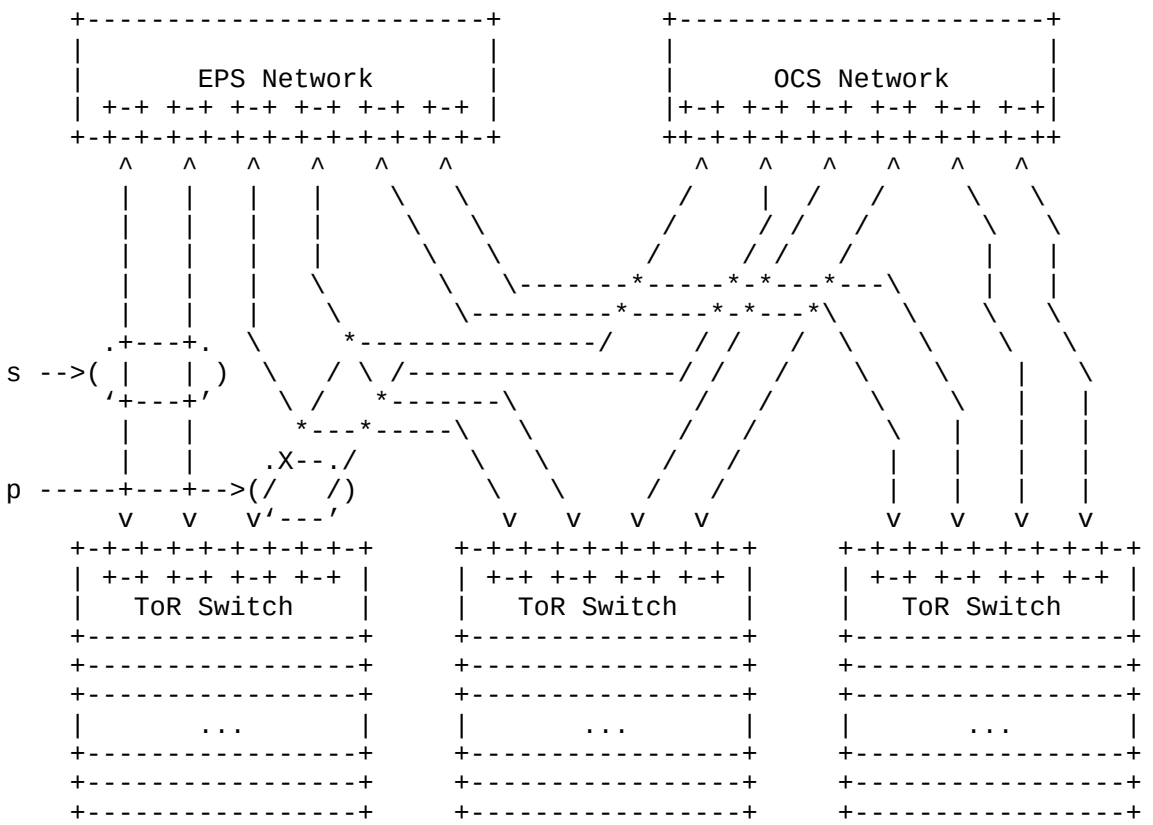


Fig. 4 switch ports allocation in hybrid DCN

The costs of network interconnecting devices in the EPS and OCS networks are determined by allocation of uplink interfaces. Thus, for each ToR, the cost constraint can be presented as Cp(s) + Cc(p)<= C, in which Cp(s) stands for the cost of EPS with s uplinks, and Cc(p) stands for the OCS cost with p uplinks.

The total volume of flows to be transmitted on a ToR switch is V. The traffic is carried by either EPS or OCS:Vp + Vc = V.

The performance requirements specify the acceptable worst-case performance of the system, such as the longest flow completion time and the highest request blocking probability. A proper resource allocation and traffic partitioning should satisfy the performance requirements in both EPS and OCS networks: Tp(s,Vp) <= Tmax, Bc(p,Vc) <= Bmax, where Tmax and Bmax are the flow completion time and request blocking probability requirements in EPS and OCS, respectively.

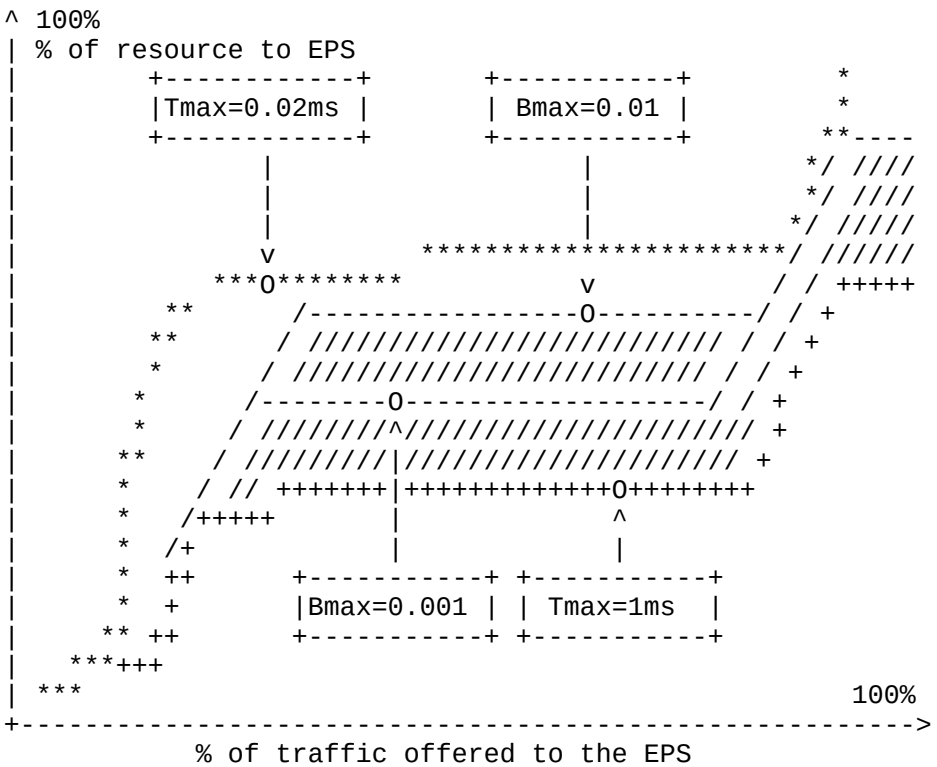


Fig. 5 BLOC for hybrid DCN

Fig. 5 shows the BLOC with different network performance requirements. When Tmax equals 1 ms and Bmax equals 0.01, there is a feasible region between the curves with Tmax and Bmax. When the performance requirements are higher (i.e., smaller Tmax and Bmax), the feasibleregion will be smaller or may even disappear. For example, when Tmax and Bmax decrease respectively to 0.02 ms and 0.001, the feasible region cannot be found. That means it is not possible to find a resource allocation that can satisfy Tmax and Bmax simultaneously. As the hybrid switching system is an interaction between the three components, when the network performance requirements cannot be satisfied, the system should have a greater budget or carry less traffic to obtain a feasible resource allocation.

7. Security Considerations

This document does not impose any new challenges to the current Internet.

8. IANA Considerations

This document makes no requests for IANA action.

9. Acknowledgements

We are grateful to the valuable discussions and inputs from the community. We thank the support from NSFC.

10. Informative References

[Cisco15] Cisco, Cisco., "Cisco global cloud index: Forecast and methodology, 2015-2020. white paper", http://www.cisco.com /en/US/solutions/collateral/ns341/ns525/ns537/ns705/ ns1175/Cloud_Index_White_Paper.html#wp9000816 1-29, 2015.

[Farrington10] Farrington, Nathan., Porter, George., Radhakrishnan, Sivasankar., Bazzaz,, Hamid., Subramanya, Vikram., Fainman, Yeshaiahu., Papen, George., and Amin. Vahdat, "Helios: a hybrid electrical/optical switch architecture for modular data centers", SIGCOMM’10 339-350, DOI 10.1145/1851182.1851223, August 2010.

[FENG16] Feng, Z., Sun, W., and W. Hu, "BLOC: A Generic Resource Allocation Framework for Hybrid Packet/Circuit-Switched Networks", J. Opt. Commun. Netw. 8, 689-700, DOI 10.1364/JOCN.8.000689, August 2016.

[FENG17] Feng, Z., Sun, W., Zhu, J., Shao, J., and W. Hu, "Resource Allocation in Electrical/Optical Hybrid Switching Data Center Networks", J. Opt. Commun. Netw. 9, 648-657, DOI 10.1364/JOCN.8.000689, August 2017.

[Gantz12] Gantz, John. and David. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east", International Data Corporation 1414_v2, December 2012.

[Gauger06] Gauger, C., Kuhn, P., Breusegem, E., Pickavet, M., and P. Demeester, "Hybrid optical network architectures: Bringing packets and circuits together", IEEE Commun. Mag. 44(8), 36-42, DOI 10.1109/MCOM.2006.1678107, August 2006.

[WANG10] Wang, Guohui., Andersen, David., Kaminsky, Michael., Papagiannaki, Konstantina., Eugene Ng, T., Kozuch, Michael., and Michael. Ryan, "c-Through: part-time optics in data centers", SIGCOMM’10 327-338, DOI 10.1145/1851182.1851223, August 2010.

[Zukerman89] Zukerman, M., "Bandwidth allocation for bursty isochronous traffic in a hybrid switching system", IEEE Transactions on Communications 37(12), 1367-1371, DOI 10.1109/26.44208, December 1989.

Authors’ Addresses

Weiqiang Sun
Shanghai Jiao Tong University
800 Dongchuan Road
Shanghai 200240
China

Phone: +86 21 3420 5359
EMail: sun.weiqiang@gmail.com

Junyi Shao
Shanghai Jiao Tong University
800 Dongchuan Road
Shanghai 200240
China

EMail: shaojunyi@sjtu.edu.cn

Weisheng Hu
Shanghai Jiao Tong University
800 Dongchuan Road
Shanghai 200240
China

EMail: wshu@sjtu.edu.cn