Understanding switch buffer utilization in CLOS data center fabric

Yihua He
Nitin Batta
Igor Gashinsky
hyihua@verizonmedia.com
batta@verizonmedia.com
igor@verizonmedia.com

ABSTRACT

Despite a lot of discussion and research on switch buffer sizing, there is no consensus due to the complicated nature of many factors, such as heterogeneous workload, congestion control protocol, geographical radius, network topology, buffer allocation algorithm and etc. In this paper, we profiled the switch buffer utilization in our 3-tier CLOS data center fabrics, each of which consists of up to 500 racks or 20000 physical compute and storage elements. We found that buffer utilization in data center switches has a tight correlation with their topological positions, port speed profile and data proximity. Based on these findings, we can select a different buffer size for different purposes of the devices to better optimize the cost of the cluster.

1 INTRODUCTION

In the last decade or two, there has been an intense research activity in studying the buffer sizing in the computer networking devices, namely, routers and switches. One commonly-known rule of thumb is to size the buffer in the Internet core routers in relation to delay-bandwidth product[1], as well as by the square root of the number of flows[2]. A number of efforts, such as [3], have been devoted to study the behavior and fine-tune transportation protocols to better utilize the buffer and increase the goodput of given conditions. On the other hand, the authors of [4] looked into SPN-LEF data center network topology and measured their buffer utilization via high-fidelity simulations.

In this paper, we attempt looking into the buffer utilization of our data center network through direct measurement. Our data center network are comprised of a typical 3-stage CLOS network with SPN, LEF and TOR as depicted in Fig 1. We found that buffer utilization in data center switches has a tight correlation with their topological positions, port speed profile and data proximity. Based on these findings, we can select a different buffer size for different purposes of the devices to better optimize the cost of the cluster.

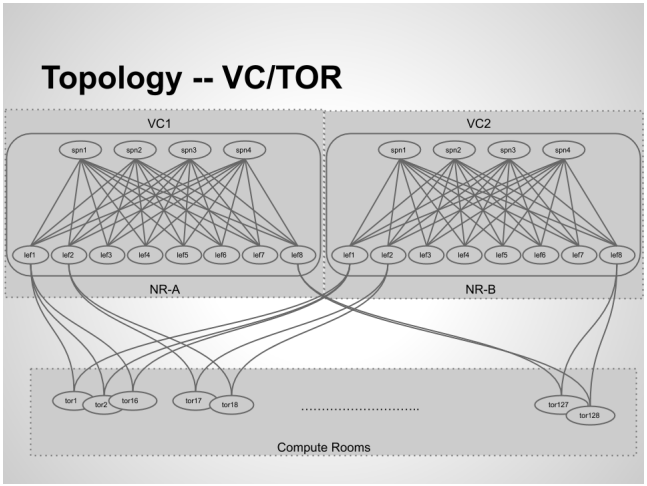

Figure 1: Data center fabric topology

2 EXPERIMENT DETAILS

2.1 Our data center fabric topology

Our data center networks are comprised of multiple 3-stage CLOS clusters. Each of these clusters typically have either 2 or 4 Virtual Chassises (VCs), each of which has a number of spine (SPN) and leaf (LEF) switches. Normally, the SPN and LEF switches are from the same port layout, for example, 32x100G. They are interconnected with Direct Attach Cables (DAC). TORs are uplinked to LEFs of each VC by fiber optics to cover longer distances. When 32-port switches are used as SPNs and LEFs, a typical full cluster has 16 SPNs and 32 LEFs in each virtual chassis, and up to 512 TORs can be connected to the LEFs in the cluster. These numbers can be adjusted by the number of parallel links running between the same pair of SPN and LEF switches. The clusters are run on top of vanilla BGP protocols to the TOR switches. Routing is designed in such a way that traffic always takes the least number of hops — for example, if two TORs happen to be connected with the same LEF, traffic between the two TORs will never travel to any SPNs.

2.2 Buffer utilization tracking

Buffer utilization is measured by utilizing Broadcom's buffer statistic tracking (BST) feature. Max use count (high watermark) of the number of cells of buffer memory is tracked in the registers of the ASIC at per port per queue basis. The values of these registers are read periodically and streamed into a time-series database. Each read operation on the high watermark registers also resets the value of the register back to zero. Therefore if the high watermark value is read once per minute, it represents the true maximum buffer utilization at the port in the last minute.

2.3 Buffer Management

The switches used in our data center network are all shallow buffer switches with single slice of shared buffer architecture. The amount of buffer in each type of switch depends on the chosen hardware. For example, in a Broadcom Trident II based switch, the total amount of buffer is 12M bytes and organized as 61440 memory cells of 208 bytes each. There are a small number of reserved cells for control-frames, high-priority CPU queues and etc. The rest of cells can be allocated dynamically on all ports following a predefined allocation algorithm. An important adjustable parameter in the buffer allocation algorithm is α. A single port can obtain as much as $\frac{\alpha}{\alpha+1} \times B$ cells, where B is the total amount of free buffer cells at the time of allocation. We empirically set α to be 8 in our environment. In this setting, the maximum number of cells of buffer that can be used for a single port is around 50k for Broadcom Trident II based switches and 112k for Broadcom Trident III based switches. For different type of workload, we can dynamically change the alpha value in the switches to adapt to the traffic pattern and minimize buffer related impact to the end users.

3 MEASUREMENT

The buffer utilization statistics were taken from a production data center cluster for 24 hours. There are total 2 VC, 32 SPN switches, 32 LEF switches and 200 TORs which connect about 10,000 physical compute nodes. The workload is mainly Hadoop. This cluster has been installed for a few years by now. Most compute nodes use 10Gbps connections to the TORs, and TOR-LEF and LEF-SPN links are 40G. Broadcom Trident II based switches are used in this cluster. There is a lot of incast type of traffic, and it's bursty.

We notice that the buffer consumption on a single switch is not uniform among all the ports. Certain types of ports consume more buffer than other types. We classify the type of ports by the type of switches on both ends. For example, on LEF switches, there are two types of ports: (1) LEF-SPN type and (2) LEF-TOR type. On the other hand, on SPN switches, there is only one type of ports, the SPN-LEF type. On TORs, there are typically two types of ports: (1) TOR-host type and (2) TOR-LEF type.

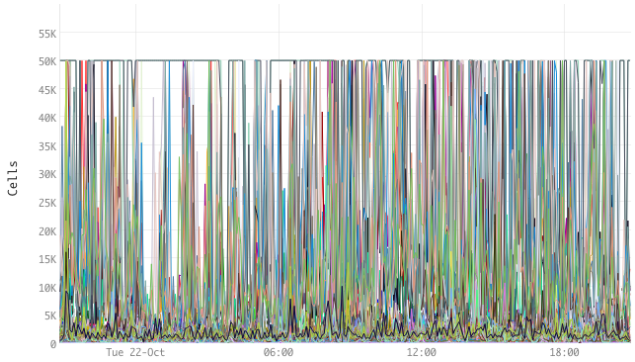

Figure 2: Buffer utilization of LEF-TOR ports on the LEF switches

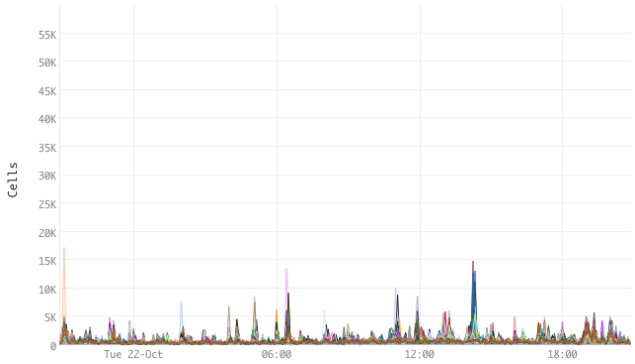

Figure 3: Buffer utilization of LEF-SPN ports on the LEF switches

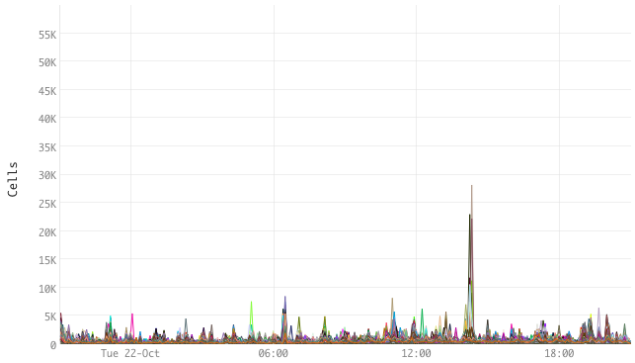

Figure 4: Buffer utilization of SPN-LEF ports on the SPN switches

Fig. [2-6] depicts buffer consumption on various types of ports. Each time slice in Fig. [2-6] has 1024 data points. We take the maximum value within each 5 minutes time slice. There are total 288 such 5 minutes time slices in each of

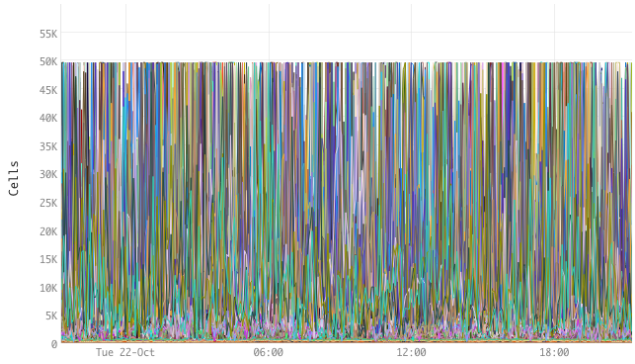

Figure 5: Buffer utilization of TOR-host ports on the TOR switches

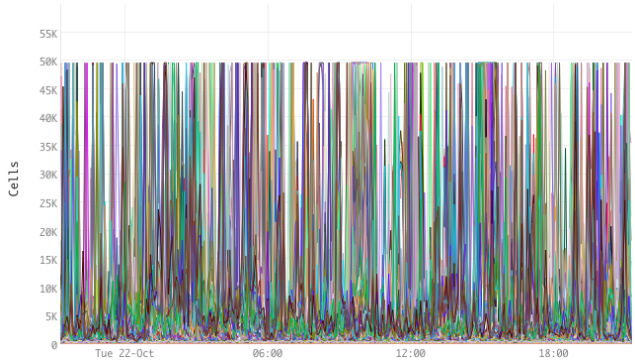

Figure 6: Buffer utilization of TOR-LEF ports on the TOR switches

the figures. The number of TOR-host ports in the cluster is much higher than that of other types of ports. Therefore, we randomly picked 1024 TOR-host ports from the cluster and plotted in Fig 5. This is to keep the number of data points the same across Fig. [2-6] for comparison purpose. A number of observations can be made from these graphs. (1) On LEF switches, LEF-TOR ports (Fig. 2) are significantly busier and consume a lot more buffers than LEF-SPN ports (Fig. 3). A few LEF-TOR ports flatlined 50k limit on the number of shared buffer cells. (2) On SPN switches(Fig. 4), switch buffers are under-utilized. The peak consumption among all SPN-LEF ports never reaches half of the 50k limit. (3) On TOR switches, buffer consumption is high on both TOR-host type (Fig. 5) and TOR-LEF type (Fig. 6) of ports, while TOR-hosts ports being noticeably higher between the two types.

We provide a number of intuitive explanation of the trait on the demand of switches buffers, among different types of ports. First, topologically, there is no oversubscription on the LEFs among SPN-facing ports. Even applications such as Hadoop can still drive in-cast type of traffic on non-oversubscribed ports, the intensity of such traffic is much less. In addition, the traffic heading to the SPNs from the LEF is only portion of the traffic received from the TORs — this is because the traffic among TORs connected at the same LEF will never be routed to any SPNs. In the TORs, the port speeds are different between TOR-LEF (40G) and TOR-host (10G) type of ports. Typically, when in-cast traffic goes through a speed reduction, more buffers are consumed because the drain rate at the egress often can not keep up with the ingress rate on the higher speed ports.

In east-west traffic dominated data center networks, shallow buffer switches are often desired. This is not only because shallow buffer switches are generally cheaper to build but also because of an important characteristic of intra data center traffic: the RTT within a single data center is extremely low. Transportation protocols such as TCP can easily deal with small amount of packet discards by retransmitting without too much penalty for waiting for the (lack of) acknowledgement from the other end before retries (retransmission). On the other hand, even among shallow buffer switches, today's merchant silicon vendors have provided a number of varieties that range from very shallow buffer or mid-shallow buffer chip-sets. Based on the results shown in this paper, we may select different profile of switches for different layers in a CLOS network. For example, we now know that we could select even lower buffered switches as a cost-efficient SPNs without performance loss. We may also want to eliminate speed changes as much as we could in the topology so that not to create buffer bottlenecks.

4 CONCLUSION

In this paper, we attempt looking into the buffer utilization of our data center network through direct measurement. We separate the buffer measurement by the type of ports in their topological uses. We find that buffer utilization in data center switches has a tight correlation with their topological positions, port speed profile and data proximity. Based on these findings, we can select a different buffer size for different purposes of the devices to better optimize the cost of the cluster.

REFERENCES

[1] C. Villamizar and C. Song. High performance tcp in ansnet. *ACM SIGCOMM Computer Communication Review*, 24(5), 1994.

[2] G. Appenzeller, I. Keslassy, and N. Meckeown. Sizing router buffers. *Proceeding of ACM SIGCOMM*, pages 281–292, 2004.

[3] M. Al-Fares, K. Elmeleegy, B. Reed, and I. Gashinsky. Overclocking the yahoo! cdn for faster web page loads. *Proceeding of ACM SIGCOMM IMC*, 2009.

[4] M. Alizadeh and T. Edsall. On the data path performance of leaf-spine datacenter fabrics. *Proceeding of IEEE 21st Annual Symposium on High-Performance Interconnects 2013*, 2013.