The Buffer Size vs. Link Bandwidth Tradeoff in Lossless Networks

Alexander Shpiner
Mellanox Technologies
alexshp@mellanox.com

Eitan Zahavi
Mellanox Technologies and Technion
eitan@mellanox.com

Ori Rottenstreich
Mellanox Technologies
orir@mellanox.com

***Abstract*—Data center networks demand high bandwidth switches. These networks also sustain common incast scenarios, which require large switch buffers. Therefore, network and switch designers encounter a buffer-bandwidth tradeoff as follows. Large switch buffers allow absorbing larger incast workload. However, higher switch bandwidth allows both faster buffer draining and more link pausing, which reduces buffering demand for incast. As the two features compete for silicon resources and device power budget, modeling their relative impact on the network is critical.**

In this work our aim is to evaluate this buffer-bandwidth tradeoff. We analyze the worst case incast scenario in the lossless network and find by how much the buffer size can be reduced, while the link bandwidth increased to stand in the same network performance. In addition, we analyze the multi-level incast cascade and support our findings by simulations. Our analysis shows that increasing bandwidth allows reducing the buffering demand by at least the same ratio, while preserving the same network performance. In particular, we show that the switch buffers can be omitted if the links bandwidth is doubled.

I. Introduction

A. Background

As the popularity of cloud and high-performance computing grows, the data center networks demand higher bandwidth devices. In addition, large buffering in the switches is also required in order to sustain the temporary network over-subscription under high load. Conventional lossy networks deal with insufficient buffer sizes by dropping packets, thus bring TCP incast throughput collapse problem, which has been analyzed deeply in the recent literature [1]–[6].

A latest trend in data center networking is to use the Converged Enhanced Ethernet (CEE), whose key feature is the losslessness of the network [7]. Lossless networks do not drop packets, but use Priority Flow Control (PFC) [8] to pause the incoming packets transmission before the buffers that receive the packets fills up. Using PFC incurs a congestion spreading problem, in which flows that do not traverse the congested link, may still suffer from throughput degradation [9]. Since PFC is implemented at the link layer, it does not distinguish between layer 4 flows. Hence, all the flows traversing the paused link are stopped[1], and the incoming link effective bandwidth is decreased for all the flows.

Infiniband [10] and Fiber Channel [11] are two another popular lossless networking technologies in data centers. They use credit-based flow control to implement the losslessness property, by continuously announcing the sender about the exact data amount that can be stored in the receiving buffer, and essentially also suffer from the congestion spreading problem.

Network designers can choose between two methods to deal with temporal network over-provisioning, and a little is known about the tradeoff between these two methods. *They can either increase the buffer sizes or the links bandwidth.* Since these two methods compete for the device area and power budget there is a tradeoff between them that can be expressed. Larger buffers can absorb longer traffic bursts. A higher link bandwidth has two consequences. First, it allows faster buffer draining, and, therefore, requires smaller buffer size to stand the same offered traffic at the same performance. Second, a larger link bandwidth allows to pause the incoming link transmission more frequently and still to achieve the same effective bandwidth under temporal congestion events. Therefore, increasing the link bandwidths reduces the required buffer sizes.

In this work our aim is to evaluate this buffer-bandwidth tradeoff for lossless networks. A similar tradeoff was previously analytically evaluated for lossy networks [12]. However, since the basic behavior of lossless network is to delay packets, instead of dropping them, it has a different effect on the applications performance, hence, requires another analysis. Lossless networks were evaluated by simulations [13], and were shown to perform better than lossy networks in typical data center scenarios. We focus on the incast scenario, since it is the most challenging for the data center networks and, yet, simple to analyze. It consists of multiple concurrent flows that are transmitted on a single link, making it congested. As far as we know, our work is the first to analyze the incast scenario for lossless networks. In the analysis we consider the congestion spreading phenomena. The paper results can be useful for the network and switch designers upon a decision how to allocate their resources. Specifically, we prove that *the switch buffers can be omitted if the links bandwidth is doubled.*

B. Contributions

In this paper we consider an incast scenario in a lossless switch architecture and study *the dependency of the required buffer size in the links bandwidth*. In other words, we answer the question *by how much the buffer size can be reduced while increasing the link bandwidth* to maintain at least the same network performance. To that end we declare the large-buffers

[1]For the simplicity we assume in this work equal-priority flows.

low-bandwidth network as *reference* and model the relative buffer requirements for the higher-bandwidth network.

We first formally describe constraints of the traffic that can be injected to the reference network without causing any link pausing. We show that these constraints are influenced by the bandwidth of the links, switch buffer sizes, and by the number of injected flows.

Next, we analyze the buffer-bandwidth tradeoff. Our model predicts that by increasing the links bandwidth, the switch buffer size can be decreased without degrading the network performance. In particular, we show that if the link bandwidth is doubled, then no buffer is required in the switches.

We generalize our analysis to the case of the multiple-incast cascade network with several congestion stages, where each stage includes switches that inject traffic to the next stage with a smaller number of switches.

We also provide detailed experimental results to assess our analysis. In particular, we examine variable number of injected flows and various factors of link bandwidth acceleration.

C. Analysis Flow

As stated previously, in our buffer-bandwidth tradeoff analysis, we would like to answer the question by how much the buffer size can be reduced while increasing the link bandwidth, and vice versa, to stand with the same network performance.

Our evaluation flow consists of four main steps. First, we assume a *reference network* with C-bandwidth links and B-sized switch buffers. Second, we define the most challenging periodic workload λ that the switch buffers can absorb without pausing the incoming links. In the third step, we increase the links bandwidth by α and reduce the buffer sizes by β.

Finally, we evaluate the relation between α and β that enable to handle the previously defined workload λ. By "handling the workload" we mean that the network can transmit the workload with at least the same rate, while preserving the same effective link bandwidth. We define *effective link bandwidth* as the given bandwidth multiplied by the percentage of time when the link is not paused by the PFC. As a result, we receive an expression of β as a function of α that answers the initially presented question.

II. Model Description

We study a lossless switch network with N identical traffic injectors, and link bandwidths that are initially equal to C, as illustrated in Figure 1. Each switch input has a dedicated buffer of size initially equal to B. The incoming packets are stored in the buffer. The switch serves the buffers in the round-robin manner limited by the outgoing link bandwidth.

The flow control keeps the buffer from overflowing. When the buffer occupancy increases above a pre-defined threshold TX-off, before the buffer fills up, the switch sends to the traffic injector a pause packet. When receiving this packet, the flow source stops injecting traffic. When the available buffer space is reduced below a pre-defined threshold TX-on, an unpause packet is sent by the switch to the traffic injector, allowing it to continue sending new packets.

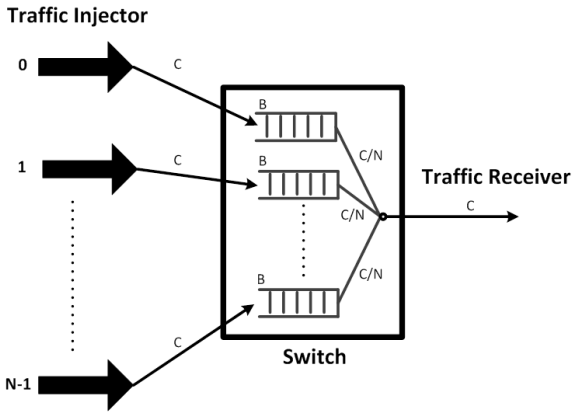


Fig. 1. Incast Workload Network Model: N sources inject traffic to a common destination over C-bandwidth links through B-sized input buffers switch.

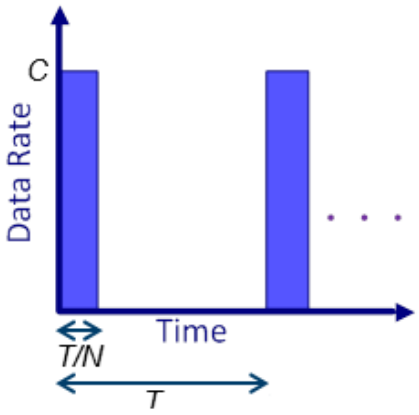


Fig. 2. Traffic pattern: T-periodic traffic with C-rate bursts of length T/N.

For simplicity, our network model presents a part of larger data center network, by considering only the links with flows that directly suffer from congestion. However, in the analysis we consider the congestion spreading phenomena, since the model links can serve also other flows of a larger network that do not traverse the congested link. Those flows are potential congestion spreading victim flows.

Also, for the simplicity of the analysis, we neglect the link propagation times and the pause packet processing times. We also assume bit-sized packets. Under this assumptions the TX-off and TX-on thresholds are set to B and $B-1[bit]$, respectively. Further, in the simulations we use non-zero propagation times and standard packet sizes to show that our results apply to real parameters also.

The workload λ traffic pattern parameters are defined as follows. Each source injects traffic bursts by a periodic pattern with a maximal temporary rate C, as illustrated in Figure 2. Our aim is to define the traffic, such that the congested link is fully utilized, and the network is stable. Moreover, later, the burst length will be defined as function of links bandwidth, buffer size and number of flows, such that it is equal to the

longest time from the period start before the buffer is filled up in the reference system.

Let T be the period length of the traffic pattern. At the first part of a period, all the flows inject traffic at rate C. At the second part of the period the traffic sources are idle. We emphasize that the bursts of the flows are synchronized, meaning that they inject traffic at the same time. This synchronization results in the most challenging constraint on the buffer size requirements. We set the burst length to be equal to $\frac{T}{N}$. Hence, the total injected traffic of all sources in a period is equal to $N \cdot \frac{T}{N} \cdot C = T \cdot C$ and can be transmitted on the output link within the time period of T.

III. Analysis of the Buffer-Bandwidth Tradeoff

A. Overview

As mentioned in Section I, due to the limited buffer size and the flow control, a link can be temporarily in a paused mode. During that mode the traffic injector stops sending packets to the network. In addition, other flows sharing the paused link, but not necessarily directed to the same destination, are also paused. Those flows are called victim flows. This negative phenomena is known as the *congestion spreading*.

There are two directions to avoid the congestion spreading. First, we can increase the size of the buffers. Large enough buffer do not get filled up and thus no pause packets are sent. For a specific traffic pattern and a given bandwidth, in order to completely avoid pause packets, the buffer size should be at least the maximal occupancy of the buffer over a period. According to the model, within each time period all the injected data can be sent over the output link. Hence, the maximal size is well defined and is finite.

The second approach to avoid the congestion spreading is to increase the bandwidth of the switch output link. Higher output link bandwidth allows faster buffer draining. This can slow down the filling rate of the buffer and can sometimes guarantee that it does not become full as long as the same traffic is still injected by the corresponding flow.

For a given traffic pattern, smaller buffers become full faster. Accordingly, they require larger bandwidth of the switch output link to avoid the congestion spreading. This is *the buffer size vs. link bandwidth tradeoff*. We would like to study this tradeoff and describe some of its properties.

B. Reference Settings

As a first step, we would like to set the value of the period length T as a function of buffer size B, link bandwidth C and number of flows N, to the maximum for which link pausing is avoided. Since the period length T defines also the burst length $\frac{T}{N}$, we actually set the burst to be exactly long to fill the reference B-sized buffer, but not to overflow it. Its value is set in the following proposition.

Proposition 1. *The largest time period length T for which link pausing is avoided satisfies*

$$T = \frac{B \cdot N^2}{C \cdot (N-1)}.$$

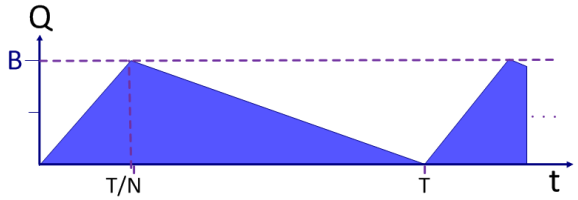


Fig. 3. The buffer occupancy during a time period. It is filled at a fixed rate and becomes full after T/N. It is empty again at the end of the period at time T.

Proof: For the value of T, the buffer becomes full exactly at the completion of the traffic injection of a flow at time T/N. During this time, traffic fills the buffer of size B at rate C. Likewise, the buffer is drained at rate of $1/N$ of the output link bandwidth C in which packets of the N flows are sent. We then have that $\frac{T}{N} = \frac{B}{C - C/N}$ and the result follows. ■

Example 1. *Let $C = 40$ Gbps and $B = 1$ Mb = 128 KB. If $N = 2$ the time period length is $T = \frac{B \cdot N^2}{C \cdot (N-1)} = \frac{1Mb \cdot 2^2}{40Gbps \cdot (2-1)} = \frac{4Mb}{40Gbps} = 100\mu s$. Likewise, if $N = 10$ then T is larger and satisfies $T = \frac{1Mb \cdot 10^2}{40Gbps \cdot (10-1)} \approx 277\mu s$.*

Figure 3 shows the occupancy Q of the buffer of size B as a function of the time within a period. It is empty at the beginning of the time period and during a time span T/N is being filled at a rate of $\frac{C \cdot (N-1)}{N}$, which is the difference between the injection rate C and the service rate C/N. The buffer reaches a maximum occupancy of B after T/N. Then the traffic injection is stopped and its size starts decreasing at rate C/N. It is empty again at the end of the time period after another time span of $\frac{B}{C/N}$.

Now we would like to examine the effect of reducing the buffer size by factor of $\beta \leq 1$ and increasing link bandwidths by factor of $\alpha \geq 1$. Figure 4 shows the buffer occupancy after the above changes. We define three *time intervals* during the T-length period, which differ by the arrival rate to the buffer. We denote the time at which the i-th interval ends by t_i for $i = 1, 2, 3$, which are illustrated in Figure 4.

At the first time interval, traffic is injected at rate C and served with rate $\frac{\alpha C}{N}$ until the βB-sized buffer is filled at time t_1. Since the buffer is now smaller compared to the reference, it might be impossible to inject all the amount of traffic of a period within this interval. Notice that, generally, the time takes to fill the small-sized higher-bandwidth buffer can be either shorter (e.g. for relatively small β) or longer (e.g. for relatively large α) than the burst length $\frac{T}{N}$ (which is the time that takes to fill the reference system switch buffer).

The remained traffic is be injected in the second time interval between the times t_1 and t_2. Since the buffer is kept full during this interval, the traffic injection is turned on and off alternately by pause and unpause packets according to the immediate occupancy of the buffer.

Last, in the third time interval between the times t_2 and t_3 no traffic is injected, since no traffic remained to inject in this period. The buffer is served until it becomes empty at time

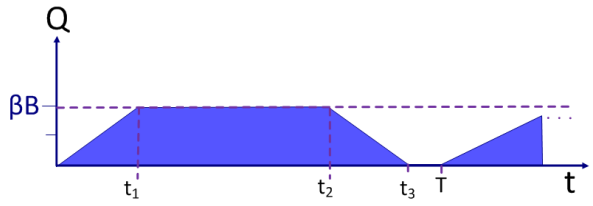


Fig. 4. The occupancy of the smaller buffer with size $\beta \cdot B$ during a time period with a bandwidth of $\alpha \cdot C$. It is filled until time t_1. Then, the traffic injection is paused from time to time until t_2 according to the available buffer space. In the third interval it gets empty without injection. Here, the buffer gets empty and the traffic of a period is completed in shorter time of T/α.

t_3. Between time t_3 and the end of the period, the modeled network is idle.

The length of each of the intervals can be derived using the following proposition.

Proposition 2. *When serving the N flows traffic λ with a buffer of size βB and links of bandwidth αC, the end times of the three time intervals, as defined previously, within a time period T satisfy*
(i) $t_1 = \frac{\beta \cdot B \cdot N}{C \cdot (N-\alpha)}$.
(ii) $t_2 = \frac{T \cdot C - \beta \cdot B \cdot N}{\alpha \cdot C}$.
(iii) $t_3 = \frac{T}{\alpha}$.

Proof: We first calculate t_1 as the time in which the smaller buffer of size βB becomes full. Since traffic is injected at rate C and the buffer is served in an average rate of $\alpha \cdot C/N$, we have that $t_1 = \frac{\beta \cdot B}{C - \alpha \cdot C/N} = \frac{\beta \cdot B \cdot N}{C \cdot (N-\alpha \cdot 1)}$. In addition, the total amount of traffic $T \cdot C$ (for all symmetrical flows) is now sent on the single output link at rate $\alpha \cdot C$ and the process is completed within $t_3 = \frac{T \cdot C}{\alpha \cdot C} = \frac{T}{\alpha}$. To deduce t_2, we first calculate the length of the third interval in which the buffer occupancy is sent without traffic injection. It then takes $\frac{\beta \cdot B}{(\alpha \cdot C)/N} = \frac{\beta \cdot B \cdot N}{\alpha \cdot C}$. Accordingly, $t_2 = t_3 - \frac{\beta \cdot B \cdot N}{\alpha \cdot C} = \frac{T}{\alpha} - \frac{\beta \cdot B \cdot N}{\alpha \cdot C} = \frac{T \cdot C - \beta \cdot B \cdot N}{\alpha \cdot C}$. ■

We will now continue the analysis in three steps. At the first step we analyze the system after reducing switch buffer size only. Next, we increase the links bandwidth while reducing the buffer size to the level at which the link pausing is avoided. Finally, using the observation that while increasing the bandwidth of the links, they can be paused more frequently, we will define the ultimate tradeoff between the buffer size and the links bandwidth.

C. Reducing Buffer Size

In the first step, we reduce the buffer size without trying to avoid the congestion spreading effect.

We assume now that congestion spreading is allowed and not limited. Under this condition, we show that the traffic workload λ can be served within the time period T without switch buffering. To show that, we observe that we can reduce t_1 to be equal to 0 and increase t_2 to be equal to t_3, i.e., practically eliminating the first and the third time intervals. This happens when we reduce β to be equal to 0. While doing so, the incoming links are paused for a portion of $\frac{N-1}{N}$ and the effective bandwidth reduced from αC to $\frac{\alpha C}{N}$, the rate at which the buffer is served. By making β equal to 0, the traffic λ can still be transmitted until time t_3 within the period.

This property is summarized in the following proposition.

Proposition 3. *If congestion spreading is allowed, while increasing the network link bandwidths by a factor of α, the traffic load λ can be served with any buffer size, for the time duration of $\frac{T}{\alpha}$, particularly with buffer of size 0.*

D. Reducing Buffer Size with Links Bandwidth Acceleration

In the second step, we study the minimal possible buffer sizing, while avoiding congestion spreading completely. Notice that for the fair comparison to the reference system, we keep using the defined traffic pattern λ, meaning that the traffic injector peak rate stays at C, although the link bandwidths are increased to αC.

Now we would like to completely avoid link pausing. To do so, we set the duration of the second interval $t_2 - t_1$ to 0 and examine the minimal value of β for which it holds. Intuitively, an increased link bandwidth allows faster buffer serving, and therefore requires reduced buffering demand.

Proposition 4. *By increasing the network links bandwidth by a factor of α, the network can serve the traffic λ without link pausing with a reduced buffer size of at least βB for $\beta = \frac{N-\alpha}{N-1}$.*

Proof: Based on Proposition 2, we find the value of β for which the equality $t_1 = t_2$ is satisfied. This eliminates the second interval with its pause modes. From the equality $t_1 = \frac{\beta \cdot B \cdot N}{C \cdot (N-\alpha)} = \frac{T \cdot C - \beta \cdot B \cdot N}{\alpha \cdot C} = t_2$, we deduce that $\beta \cdot B \cdot N^2 = N \cdot T \cdot C - \alpha \cdot T \cdot C$ and accordingly $\beta = \frac{(N-\alpha) \cdot T \cdot C}{B \cdot N^2}$. By setting the value of T from Proposition 1, we finally have that $\beta = \frac{(N-\alpha) \cdot C}{B \cdot N^2} \cdot \frac{B \cdot N^2}{C \cdot (N-1)} = \frac{N-\alpha}{N-1}$. ■

Example 2. *We would like to demonstrate the last proposition with several settings. For two flows ($N = 2$), in order to avoid pause modes, if we increase the link bandwidth from 40Gbps to 56Gbps (with $\alpha = \frac{56}{40} = 1.4$)[2], we have that $\beta = \frac{N-\alpha}{N-1} = \frac{2-1.4}{2-1} = 0.6$. Accordingly, the required buffer size is only 60% of its original size and we have 40% buffer saving. Taking the values of B and C from Example 1, i.e. $C = 40$ Gbps and $B = 1$ Mb (that together yield $T = 0.1$ ms) we have $t_1 = \frac{\beta \cdot B \cdot N}{C \cdot (N-\alpha)} = \frac{0.6 \cdot 1 \cdot 10^6 \cdot 2}{40 \cdot 10^9 \cdot (2-1.4)} = 0.05ms$ and $t_2 = \frac{T \cdot C - \beta \cdot B \cdot N}{\alpha \cdot C} = \frac{0.1 \cdot 10^{-3} \cdot 40 \cdot 10^9 - 0.6 \cdot 1 \cdot 10^6 \cdot 2}{1.4 \cdot 40 \cdot 10^9} = 0.05$ ms $= t_1$. Likewise, $t_3 = \frac{T}{\alpha} = 0.1 \cdot 10^{-3}/1.4 = \frac{5}{70}$ ms ≈ 0.071 ms. For larger N, the improvement is smaller. If for instance $N = 10$, we obtain a much smaller improvement and the minimal value of β is $\beta = \frac{N-\alpha}{N-1} = \frac{10-1.4}{10-1} \approx 0.95$, i.e. the improved memory size has to be at least about 95% of its original size.*

E. Reducing Buffer Size with Links Bandwidth Acceleration while Preserving the Effective Bandwidth

In the third step, we study the required buffer size while limiting the total time length of which congestion spreading takes place. We use the observation that while increasing the

[2]These values are chosen, since they represent parameters of real switch products [14].

bandwidth of the links, they can be paused more frequently, to preserve the same effective bandwidth. Intuitively, since now we allow link pausing, the required buffer size will be smaller compared to the previous step.

The impact of the increased bandwidth of the network links by α is two fold. First, as stated previously, it allows faster buffer serving. In addition, during the second interval the link can be paused for longer periods and preserve the initial effective bandwidth.

This results with the following ultimate buffer-bandwidth tradeoff.

Proposition 5. *By increasing the network links bandwidth by factor of α, the network can serve traffic λ, preserving the same effective bandwidth as the reference network, with a reduced buffer size of at least βB for $\beta = \frac{(N-\alpha)(2\cdot N-\alpha\cdot N-1)}{(N-1)^2}$.*

Proof: Let p be the ratio of the interval length for which an input link is in a pause mode. Effectively, its average injection rate will be $\alpha \cdot C \cdot (1-p) + \frac{\alpha \cdot C}{N} \cdot p$. To obtain an average rate of at least C as the original bandwidth, it is required to satisfy $p \leq \frac{\alpha-1}{\alpha-(\alpha/N)}$.

By Proposition 2, the ratio of the length of the second interval for which the pause modes occur among the period T is given by $p = \frac{t_2-t_1}{T} = (\frac{T\cdot C-\beta\cdot B\cdot N}{\alpha\cdot C} - \frac{\beta\cdot B\cdot N}{C\cdot(N-\alpha)})/(\frac{B\cdot N^2}{C\cdot(N-1)}) = \frac{N-\alpha-\beta\cdot(N-1)}{\alpha\cdot(N-\alpha)}$. In order to satisfy the above upper bound on p, $\frac{N-\alpha-\beta\cdot(N-1)}{\alpha\cdot(N-\alpha)} \leq \frac{\alpha-1}{\alpha-(\alpha/N)}$, the value of β has to satisfy the above lower bound. ∎

Example 3. *For $\alpha = 1.4$ (obtained for instance by increasing the link bandwidth from $40Gbps$ to $56Gbps$) and $N = 2$ connections, the value of β required to keep the same effective bandwidth considering the pause modes is $\beta \geq \frac{(N-\alpha)(2N-\alpha\cdot N-1)}{(N-1)^2} = \frac{(2-1.4)(2\cdot 2-1.4\cdot 2-1)}{(2-1)^2} = 0.6 \cdot 0.2 = 0.12$. For $N = 10$, $\beta = \frac{(10-1.4)(2\cdot 10-1.4\cdot 10-1)}{(10-1)^2} = \frac{43}{81} \approx 0.531$.*

We now demonstrate the lower bound of β by analyzing it as the function of the number of inputs N and the link bandwidths increase α.

Proposition 6. *The value of β satisfies*
(i) $\beta \leq 2-\alpha$ *for* $N \geq 2$.
(ii) $\lim_{N\to\infty} \beta = 2-\alpha$.

Proof: We first show (i). For $1 \leq \alpha \leq 2$ we have $0 \leq (\alpha-1)-(\alpha-1)^2 = 3\cdot\alpha-2-\alpha^2$ and $\alpha-1 \leq 4\cdot\alpha-3-\alpha^2$. Since $N \geq 2$, we also have that $2\cdot(\alpha-1) \leq N\cdot(4\cdot\alpha-3-\alpha^2)$. It then follows that $(N-\alpha)\cdot(2\cdot N-\alpha\cdot N-1) \leq (2-\alpha)\cdot(N-1)^2$ and $\beta = \frac{(N-\alpha)(2\cdot N-\alpha\cdot N-1)}{(N-1)^2} \leq 2-\alpha$.

To show (ii), we rewrite the formula from Proposition 5 as $\beta = \frac{(1-\frac{\alpha}{N})(2-\alpha-\frac{1}{N})}{(1-\frac{1}{N})^2}$ and deduce that $\lim_{N\to\infty} \beta = \frac{1\cdot(2-\alpha)}{1} = 2-\alpha$. ∎

The next corollary follows directly from Proposition 6.

Corollary 1. *The required buffer size can be reduced by the same ratio as the link bandwidths increased to stand in the same network performance and preserve the effective link bandwidths. In particular, if link bandwidths are doubled, then no switch buffering is required.*

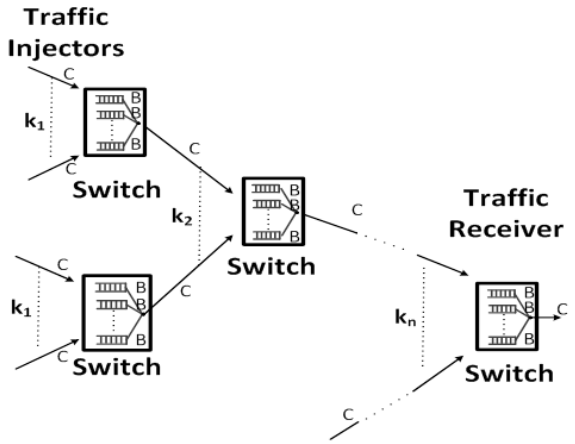


Fig. 5. Ilustration of the multistage switch architecture with n stages. The input traffic of a switch in the i^{th} stage is injected simultaneously as the output traffic of k_i switches in a previous stage.

Note: Above corollary of the buffer size was obtained using ideal settings that were described in Section II. Real implementation of pause-based flow control would also require a minimal buffer size to assure the losslessness property of the network based on the link bandwidth and propagation time, pause frame processing delay and packet size.

IV. Multiple-Incast Cascade

In this section we generalize our study of the buffer size vs. link bandwidth tradeoff to a multiple incast cascade. As illustrated in Figure 5, consider a network of n stages in which the input traffic of a switch in the i^{th} stage is composed of the output traffic of k_i switches in the previous stage. The illustrated network can present a sub-network of a large data-center fat-tree network, by considering only the congested links. Every switch receives traffic simultaneously from its multiple sources and observes an incast congestion. In this network there is a single switch in the last (n^{th}) stage and the number of switch in the $(i-1)^{\text{th}}$ stage is k_i times the number of switches in the next i^{th} stage such that for $i \in [1,n]$ the number of switches in the i^{th} stage is given by $\prod_{j=i+1}^{n} k_j$. For the sake of simplicity we further assume an identical buffer size B per input for all switches in the network. We also assume that each of the k_i flows to the i^{th} stage is served equally at rate $\frac{1}{k_i}$ of the outgoing link capacity, that the flows in the first stage have the same traffic, as well as that $k_i \geq 2$ for $i \in [1,n]$. We would like to mention that by sending the pause and unpause packets, a switch in stage i can stop the transmission from a switch in stage $i-1$, while a switch in stage 1 stops the transmission of the traffic injector.

As in the analysis in Section III, first, we would like to set the workload parameters in each of the stages. Clearly, if all the flows inject the same traffic to the switches in the first stage, the symmetry between the switches in a certain stage is preserved by induction also in all other stages. Based on the

previous analysis of a single switch, we can see a switch as transforming k_i bursts with rate C of length T_i into a single burst of the same rate but with longer length of $k_i \cdot T_i$. We will denote the length of the inputs to the first stage by T_1. We will later explain how to calculate this value. Then, the length of the traffic injected to stage i is $T_1 \cdot \prod_{j=1}^{i-1} k_j$.

We prove that under the presented settings, the most congested switch is the switch in stage n. Hence, the value of T_1 will be defined as the maximal length that still avoids a congestion in this switch. To do so, we start by calculating the maximal occupancy in a switch in the i^{th} stage.

Proposition 7. *In the described traffic pattern, the maximal observed occupancy of a switch in the i^{th} stage is given by $Q_i^{max} = \frac{C\cdot(k_i-1)}{k_i} \cdot T_1 \cdot \prod_{j=1}^{i-1} k_j$.*

Proof: As in the simple case of a single switch, the maximal occupancy is obtained exactly when the injection stops, i.e. after $T_1 \cdot \prod_{j=1}^{i-1} k_j$. Until then, traffic to each buffer is injected with rate C and the buffer is drained at rate $\frac{C}{k_i}$. ■

We now show that the maximal buffer occupancy is achieved in the last stage.

Proposition 8. *The maximal observed occupancy in one of the switches in the architecture is obtained in the switch in the last n^{th} stage and equals $Q_n^{max} = \frac{C\cdot(k_n-1)}{k_n} \cdot T_1 \cdot \prod_{j=1}^{n-1} k_j$.*

Proof: We simply show that $Q_{i+1}^{max} \geq Q_i^{max}$ for $i \in [1, n-1]$. By Proposition 7, we have

$$\begin{aligned}\frac{Q_{i+1}^{max}}{Q_i^{max}} &= \frac{C \cdot T_1 \cdot (k_{i+1}-1)/k_{i+1} \cdot \prod_{j=1}^{i} k_j}{C \cdot T_1 \cdot (k_i-1)/k_i \cdot \prod_{j=1}^{i-1} k_j} \\ &= \frac{k_i^2 \cdot (k_{i+1}-1)}{k_{i+1} \cdot (k_i-1)} = \frac{k_{i+1}-1}{k_{i+1}} \Big/ \frac{k_i-1}{k_i^2} \\ &= \Big(1 - \frac{1}{k_{i+1}}\Big) \Big/ \Big(\frac{1}{k_i} - \frac{1}{k_i^2}\Big).\end{aligned}$$

Since $k_i \geq 2$ for $i \in [1, n]$ then $1 - \frac{1}{k_{i+1}} \geq \frac{1}{2}$ and $0 < \frac{1}{k_i} - \frac{1}{k_i^2} < \frac{1}{k_i} \leq \frac{1}{2}$. Accordingly, $Q_{i+1}^{max} \geq Q_i^{max}$ and the maximum occupancy is obtained in the last switch with $i = n$. ■

It follows from the last proposition that in order to avoid the congestion spreading, we should make sure that the maximal occupancy of the last switch is not beyond B. Accordingly, we define the value of T_1 as the length that achieves a maximal occupancy of $Q_n^{max} = B$ and have that $T_1 = \frac{B\cdot k_n}{C\cdot(k_n-1)} / \prod_{j=1}^{n-1} k_j$ and $T_n = T_1 \cdot \prod_{j=1}^{n-1} k_j = \frac{B\cdot k_n}{C\cdot(k_n-1)}$. The length of the time period is $T_n \cdot k_n = \frac{B\cdot k_n^2}{C\cdot(k_n-1)}$.

Next, we analyze the buffer-bandwidth tradeoff on the stage-n switch to learn the potential effect of link bandwidth improvement. The analysis is similar to a single-stage case from Section III, but now the traffic of each flow arrives to a switch at stage i at rate αC, instead of a rate C, and lasts $\frac{T_i}{\alpha}$ time instead of T_i. In particular, in the switch at the last stage in which we concentrate the traffic length is $\frac{T_n}{\alpha}$. We will continue directly to analysis of the case with reduced buffer size, increased link capacities and preserving effective link bandwidth from Section III-E. We clarify the changes between the cases and use again the notations of t_1, t_2, t_3 for the end time of the three obtained time intervals that again differ in their arrival rate.

Since the arrival rate is higher, the buffer fills up faster, which results in the next proposition.

Proposition 9. *While serving the traffic with a buffer of size βB, the end time of the first interval in the stage-n switch equals*

$$t_1 = \frac{\beta B k_n}{\alpha C(k_n-1)}.$$

Proof: We calculate t_1 as the time in which the buffer of size βB becomes full. Since traffic is injected at rate αC and the buffer is served in an average rate of $\alpha C/k_n$, we have that $t_1 = \frac{\beta\cdot B}{\alpha\cdot C - \alpha\cdot C/k_n} = \frac{\beta\cdot B\cdot k_n}{\alpha C\cdot(k_n-1)}$. ■

The time t_3 in which the k_n injections to the switch at the last stage terminates is given by $\frac{T_n\cdot k_n}{\alpha} = \frac{B\cdot k_n}{C\cdot(k_n-1)} \cdot \frac{k_n}{\alpha} = \frac{B\cdot k_n^2}{\alpha\cdot C\cdot(k_n-1)}$.

Last, the percentage of time the incoming link is in the paused mode equals the ratio of the time difference $(t_2 - t_1)$ and the length of the time period $T_n \cdot k_n$. Setting the constraint on p as in Section III-E to $p \leq \frac{\alpha-1}{\alpha-(\alpha/k_n)}$ for the case of k_n inputs brings to the next proposition.

Proposition 10. *In the multi-cascade network, by increasing network links bandwidth by a factor of α, the network can serve the traffic λ preserving the effective bandwidth with a reduced buffer size by a factor of β for $\beta \geq \frac{k_n\cdot(2-\alpha)-1}{k_n-1}$.*

Proof: Based on the above explanation we have that $p = \frac{t_2-t_1}{T_n\cdot k_n} = \Big(\frac{Bk_n{}^2}{\alpha C(k_n-1)} - \frac{\beta B k_n}{\alpha C} - \frac{\beta\cdot B\cdot k_n}{\alpha C\cdot(k_n-1)}\Big) \Big/ \frac{B\cdot k_n^2}{C\cdot(k_n-1)} = \frac{B\cdot k_n^2\cdot(1-\beta)}{\alpha\cdot C\cdot(k_n-1)} \Big/ \frac{B\cdot k_n^2}{C\cdot(k_n-1)} = \frac{1-\beta}{\alpha}$. With the upper bound on p from Proposition 5 the result then follows. ■

Example 4. *Consider the mentioned above value of $\alpha = 1.4$ (obtained for instance by increasing the link bandwidth from 40 Gbps to 56 Gbps) and $k_n = 10$ switches that are connected to the single switch in the last stage. Then, the minimal value of β required to keep the same effective bandwidth considering the pause modes is $\beta = \frac{k_n\cdot(2-\alpha)-1}{k_n-1} = \frac{10\cdot(2-1.4)-1}{10-1} \approx 0.55$.*

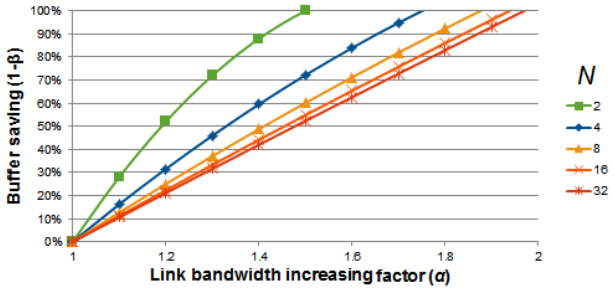
The next corollary follows from Proposition 10.

Corollary 2. *The conclusion of Corollary 1 that the required buffer size can be reduced by the same ratio as the link bandwidths increased to stand in the same network performance and preserve the effective link bandwidths, holds also for the multi-incast cascade case. Moreover, the general result holds even when the sources inject traffic to the congested switch in a peak rate of αC, contrary to our initial restriction in the beginning of Section III-D.*

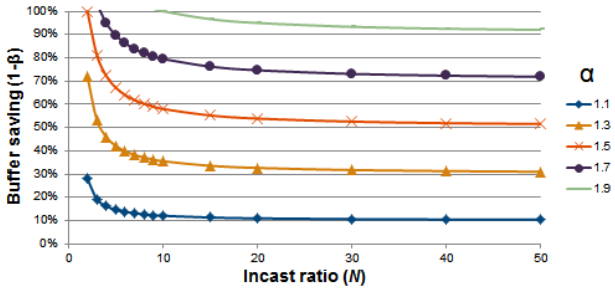
V. Experimental Results

A. Numerical Evaluation

We conduct experiments to evaluate the analysis of the tradeoff. We consider a simple single switch as well as a multiple incast cascade.



(a) Buffer size saving vs. α



(b) Buffer size saving vs. N

Fig. 6. The buffer size saving $1-\beta_0$ as a function of the link bandwidth improvement factor α and the number of inputs N as expressed in Proposition 5. The buffer size saving satisfies $1-\beta_0 \geq \alpha - 1$ for $N \geq 2$ and $\lim_{N\to\infty}(1-\beta_0) = \alpha - 1$, as follows from Proposition 6.

We first consider a single switch. We examine the buffer size saving as a function of the link bandwidth improvement factor α and the number of inputs N as was expressed in Proposition 5. The results are presented in Figure 6. In 6(a), we can see the buffer saving as a function of α for various N. For all values of α, the improvement is more significant for smaller values of N. For instance, if $\alpha = 1.1$ the saving is $1-\beta = 0.280$ for $N = 2$ while it is only $1-\beta = 0.106$ for $N = 32$. Similarly, for $\alpha = 1.5$ it equals 0.524 for $N = 32$ while it equals 1 for $N = 2$, i.e. no buffer is required in this case. Anyway, as indicated in Proposition 6 and Corollary 1, for any $N \geq 2$, the relative improvement in the buffer size $1-\beta_0$ satisfies $1-\beta_0 \geq \alpha - 1$.

Likewise, in 6(b) we present the saving as a function of N for various α. We can see that the asymptotic value of the improvement for large N satisfies $\lim_{N\to\infty}(1-\beta_0) = \alpha - 1$. For instance, if $\alpha = 1.5$, the buffer size saving is 0.580, 0.539 and 0.515 for $N = 10, 20, 50$, respectively.

B. Simulations Results

Next, we conducted simulations with Omnet++ [15] enhanced by INET Framework modules [16] to examine our analysis. We simulated a network with $10ns$ propagation delay links and $1.5KB$-sized packets. The TX-off and TX-on thresholds have been set accordingly for preserving the losslessness of the network, while maximizing the buffer utilization.

We measure the occupancy of the switch buffer within a time period for two possible values link bandwidth. Figure 7 presents

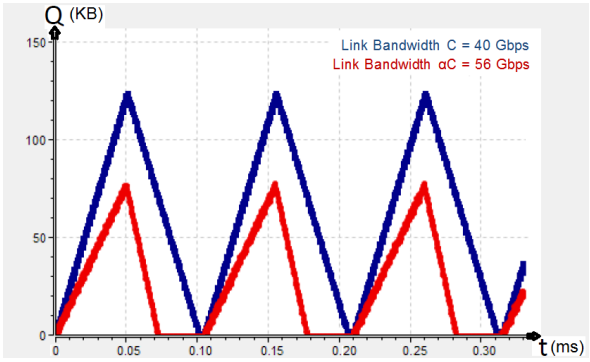


Fig. 7. Comparison of the buffer occupancy with 40 Gbps and 56 Gbps (with $\alpha = 1.4$) and $N = 2$ as in Example 1 and Example 2. The maximal occupancy is $B = 128$ KB and 76.8, respectively with $\beta = 0.6$.

the results. As in Example 1 and Example 2, we assume $N = 2$ with $C = 40$ Gbps. With $T = 0.1ms$ we achieve maximal occupancy of $B = 128KB$ that is obtained after $0.05ms$. With an increased bandwidth of 56 Gbps ($\alpha = 1.4$), the maximal occupancy is only $\beta B = 76.8KB$, i.e. for $\beta = 0.6$ and the service of the injection is served within only $\frac{5}{70} \approx 0.071ms$.

VI. Conclusions

In this paper we introduced and studied the tradeoff between buffer size and the link bandwidth in lossless networks under incast scenario. We presented a formal analysis of the tradeoff and showed that it holds also for the multiple incast cascade. We verified the results using simulations.

Our study yields several major conclusions. First, in lossless networks, the switch buffer sizes can be reduced significantly, while still pushing the same incast traffic. We explained that although the buffer size reduction might result in congestion spreading, this phenomena can be mitigated by increasing the link bandwidths. Specifically, the tradeoff analysis shows that the buffer size saving ratio is equal to the link bandwidth increase ratio. In particular, when doubling the link bandwidths, the switch buffering can be completely avoided, without degrading the network performance.

Acknowledgements

The authors would like to thank Isaac Keslassy, Mark Shifrin, Michael Kagan, Liran Liss and Matty Kadosh for their helpful comments.

References

[1] Y. Chen, R. Griffith, J. Liu, R. H. Katz, and A. D. Joseph, "Understanding TCP Incast throughput collapse in datacenter networks," in *ACM WREN*, 2009.

[2] E. Krevat, V. Vasudevan, A. Phanishayee, D. G. Andersen, G. R. Ganger, G. A. Gibson, and S. Seshan, "On application-level approaches to avoiding TCP throughput collapse in cluster-based storage systems," in *ACM PDSW*, 2007.

[3] A. Phanishayee, E. Krevat, V. Vasudevan, D. G. Andersen, G. R. Ganger, G. A. Gibson, and S. Seshan, "Measurement and analysis of TCP throughput collapse in cluster-based storage systems," in *USENIX FAST*, 2008.

[4] V. Vasudevan, A. Phanishayee, H. Shah, E. Krevat, D. G. Andersen, G. R. Ganger, G. A. Gibson, and B. Mueller, "Safe and effective fine-grained TCP retransmissions for datacenter communication," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 303–314, Aug. 2009.
[5] J. Zhang, F. Ren, and C. Lin, "Modeling and understanding TCP incast in data center networks," in *IEEE INFOCOM*, 2011.
[6] A. Shpiner, I. Keslassy, G. Bracha, E. Dagan, O. Iny, and E. Soha, "A switch-based approach to throughput collapse and starvation in data centers," *Comput. Netw.*, vol. 56, no. 14, pp. 3333–3346, Sep. 2012.
[7] M. Ko, D. Eisenhauer, and R. Recio, "A case for convergence enhanced ethernet: Requirements and applications," in *IEEE ICC*, 2008.
[8] "P802.1Qbb/D1.3 Virtual bridged local area networks - amendment: Priority-based flow control," IEEE Draft Standard, 2010. [Online]. Available: http://www.ieee802.org/1/pages/802.1bb.html
[9] J. Santos, Y. Turner, and G. Janakiraman, "End-to-end congestion control for Infiniband," in *IEEE INFOCOM*, 2003.
[10] "Infiniband architecture volume 1 - general specifications, release 1.2.1," 2008. [Online]. Available: http://www.infinibandta.org/content/pages.php?pg=technology_public_specification/
[11] "Fibre channel features," 2014. [Online]. Available: http://www.fibrechannel.org/fibre-channel-features.html/
[12] A. Rosén and G. Scalosub, "Rate vs. buffer size: Greedy information gathering on the line," in *ACM SPAA*, 2007.
[13] A. S. Anghel, R. Birke, D. Crisan, and M. Gusat, "Cross-layer flow and congestion control for datacenter networks," in *DC-CaVES*, 2011.
[14] "Mellanox scale-out ethernet products," 2014. [Online]. Available: http://www.mellanox.com/page/ethernet_switch_overview/
[15] A. Varga, "Omnet++," 2004. [Online]. Available: http://www.omnetpp.org/
[16] "INET framework," 2006. [Online]. Available: http://inet.omnetpp.org/