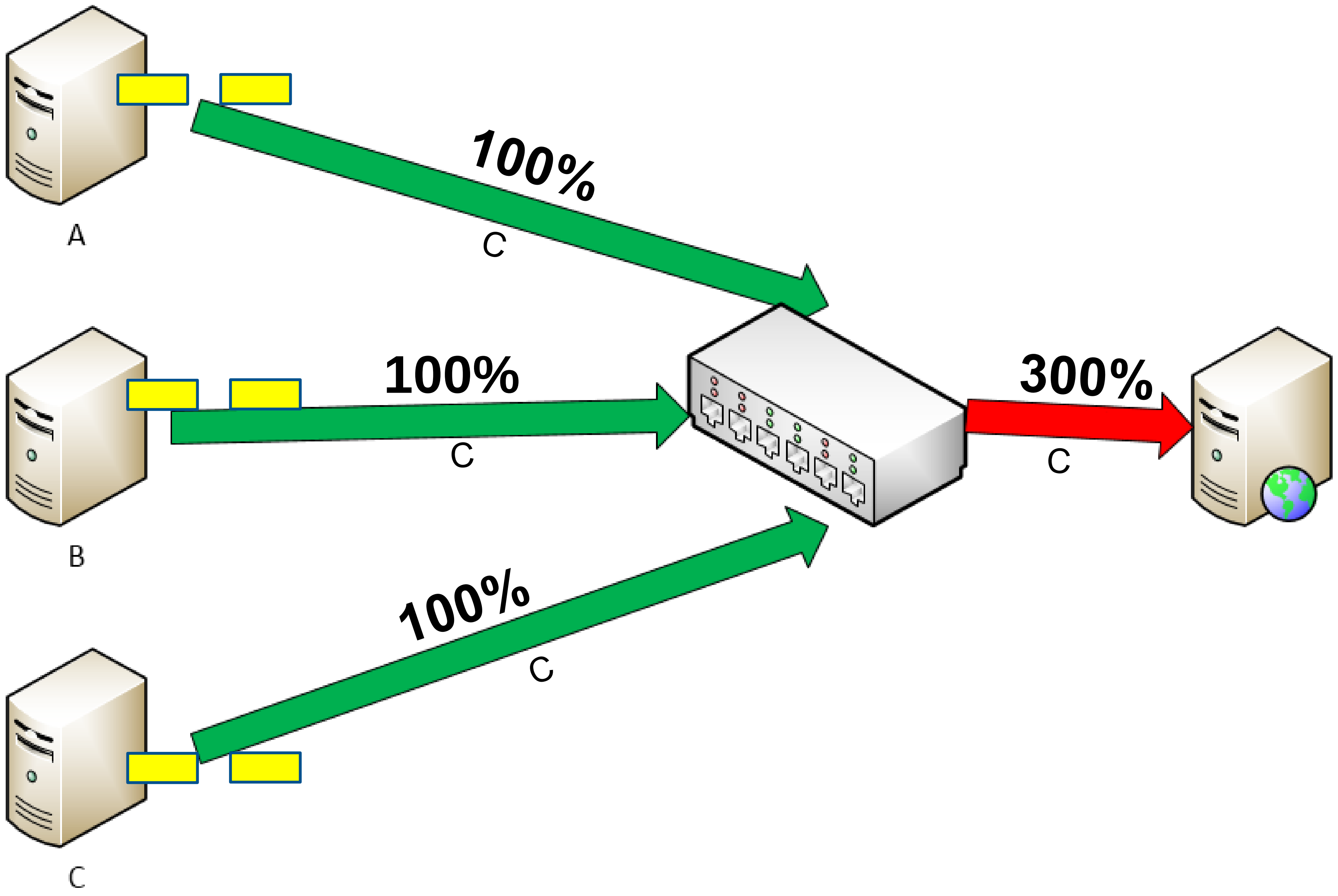
The Buffer Size vs. Link Bandwidth Tradeoff in Lossless Networks

Alex Shpiner, Eitan Zahavi, Ori Rottenstreich

Hot Interconnects, August 2014

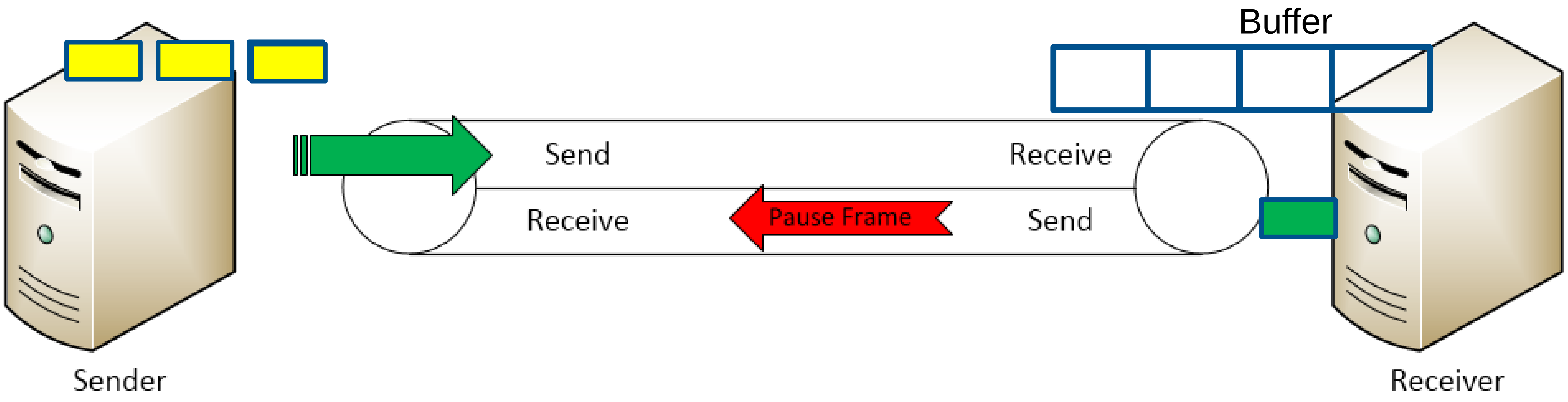
Mellanox TECHNOLOGIES
Connect. Accelerate. Outperform.™

Background - Incast



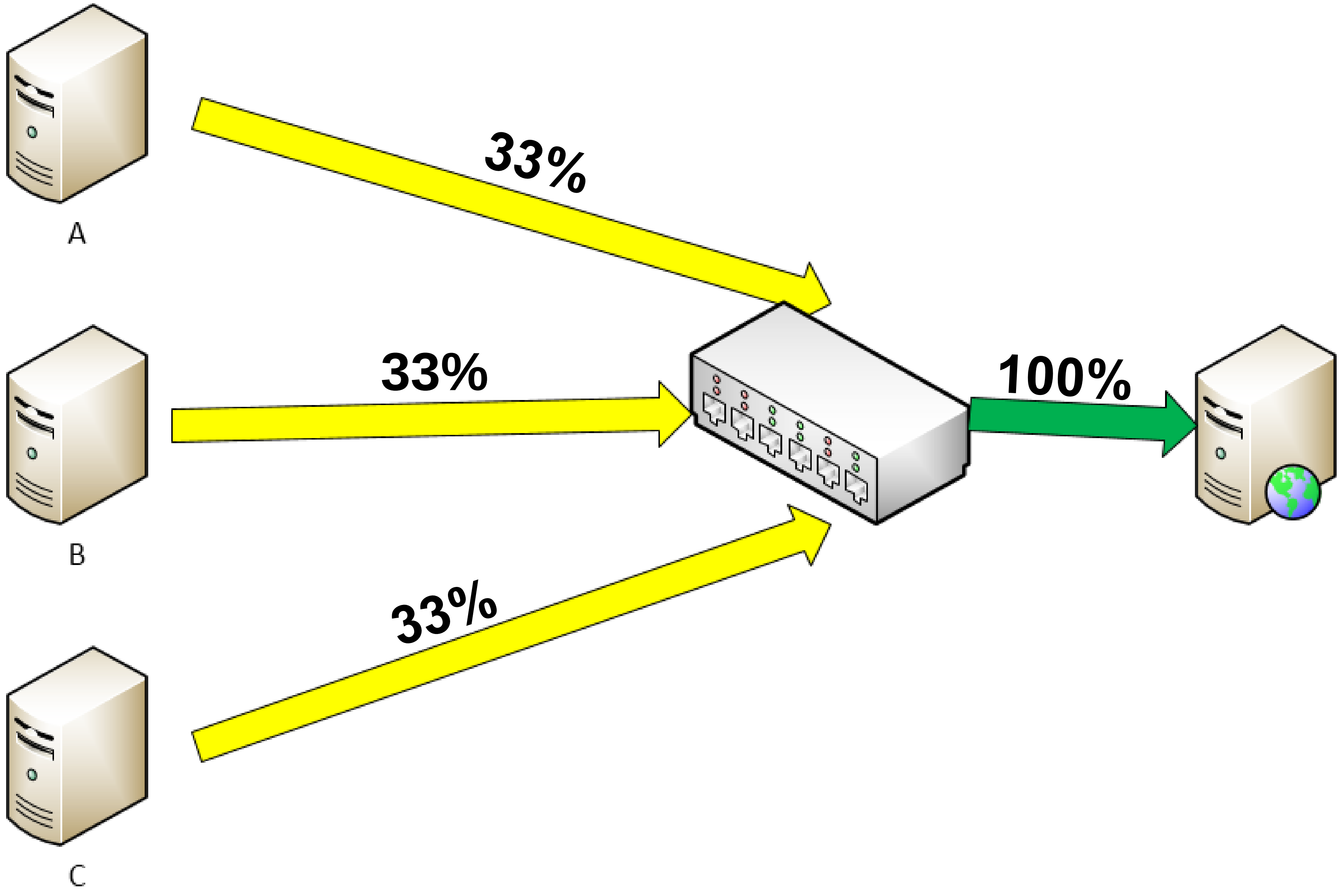
Source: http://theithollow.com/2013/05/flow-control-explained/

Background - Pause Frame Flow Control



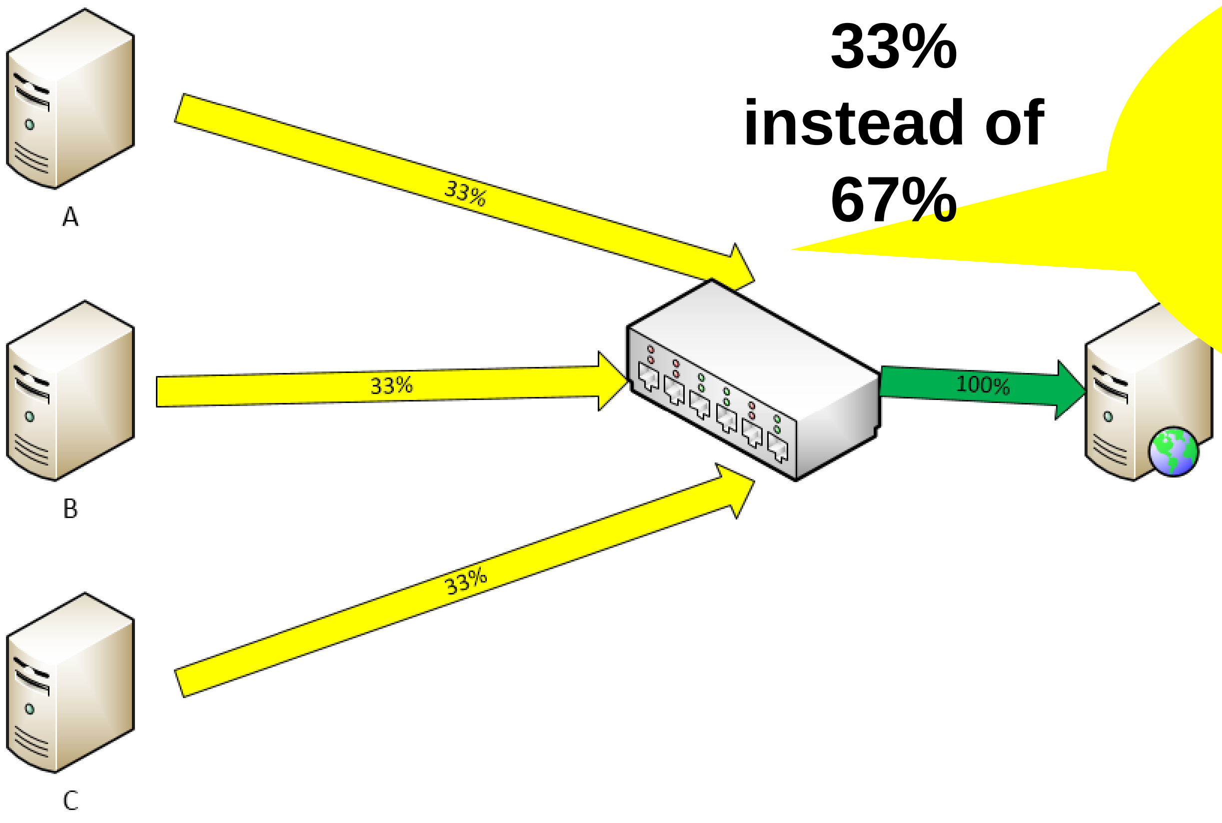
Source: http://theithollow.com/2013/05/flow-control-explained/

Background – Incast with Pause Frame Flow Control



Source: http://theithollow.com/2013/05/flow-control-explained/

Background – Congestion Spreading Problem



33% instead of 67%

This flow is also paused, since the pause control does not distinguish between flows.

Effective link bandwidth=
Link bandwidth * %unpaused

- Small buffers ⇒ Link pauses ⇒ Congestion spreading ⇒ Effective link bandwidth decrease
- To deal with Incast we can:
  - Increase buffers
  - Increase link bandwidth

Tradeoff

Source: http://theithollow.com/2013/05/flow-control-explained/

Buffer-Bandwidth Tradeoff

- Higher bandwidth allows:
  - Faster buffer draining
  - More link pausing, but achieving same effective bandwidth

Effective bandwidth =
Link bandwidth * %unpaused

⇒ reduced buffering demand
  - to handle same incast scenario without congestion spreading

- Aim: evaluate the buffer-bandwidth tradeoff

- Assumptions:
  - Lossless network
  - Congestion spreading avoidance is desired

Evaluation Flow

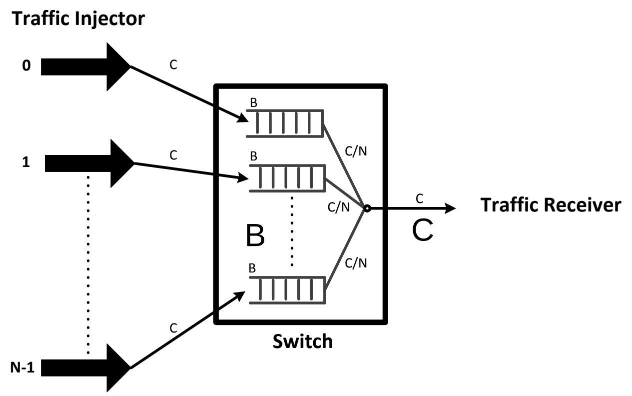
1). **Assume** network with links of **bandwidth C** and **buffers** of size **B**

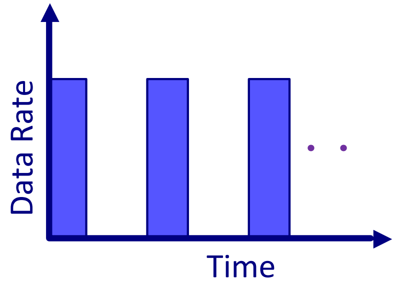
2). **Define** the most challenging **workload** the network can handle without congestion spreading

3). **Increase** links **bandwidth** by **α** and **reduce buffer** size by **β**

4). **Evaluate** the relation between **α and β** that able to handle the workload from step 2.

Transmit the workload at the same rate, while keeping the same effective link bandwidth



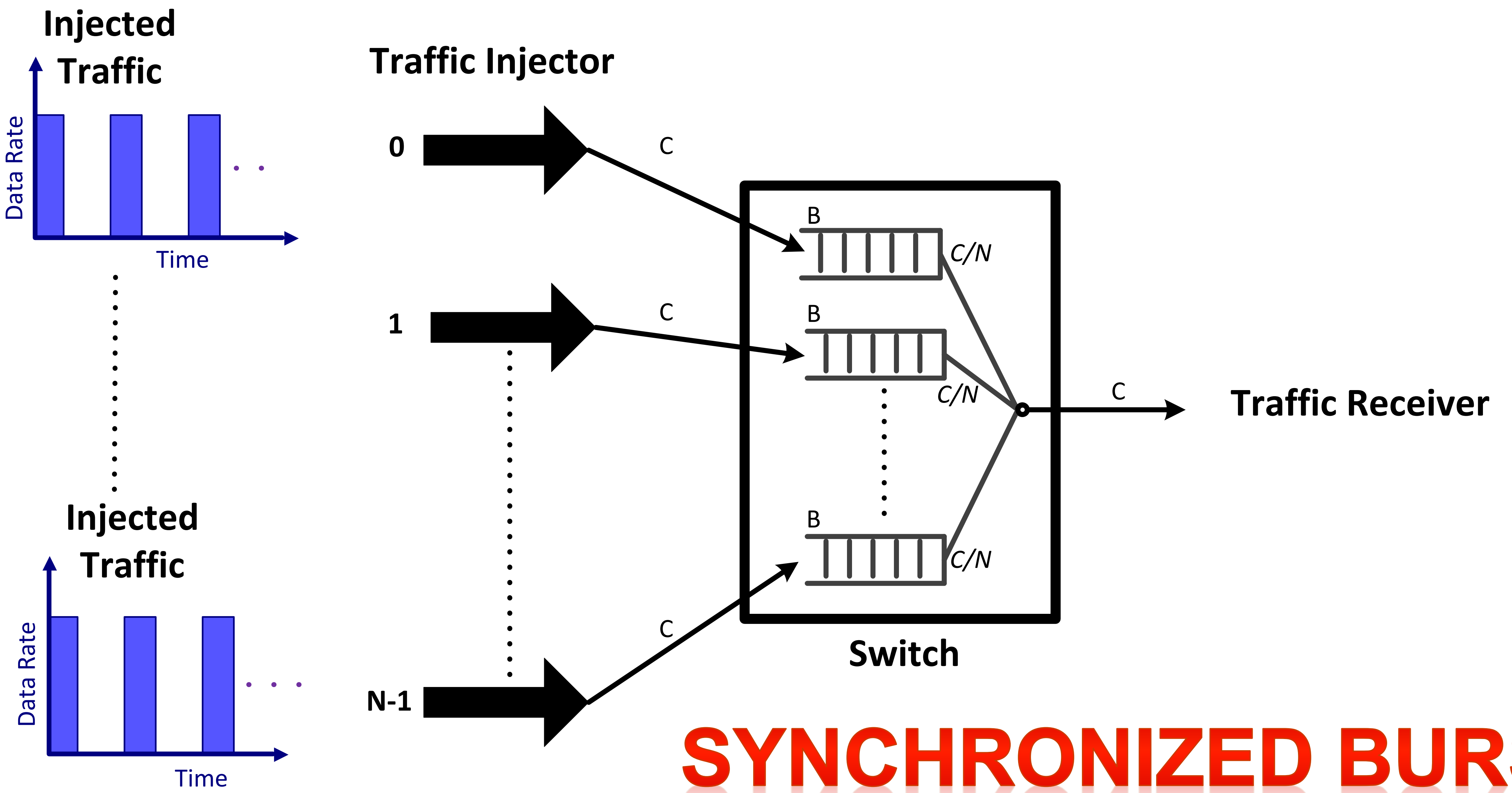


$$C_{new} = \alpha C, \alpha > 1$$
$$B_{new} = \beta B, \beta < 1$$

$$\beta = f(\alpha)$$

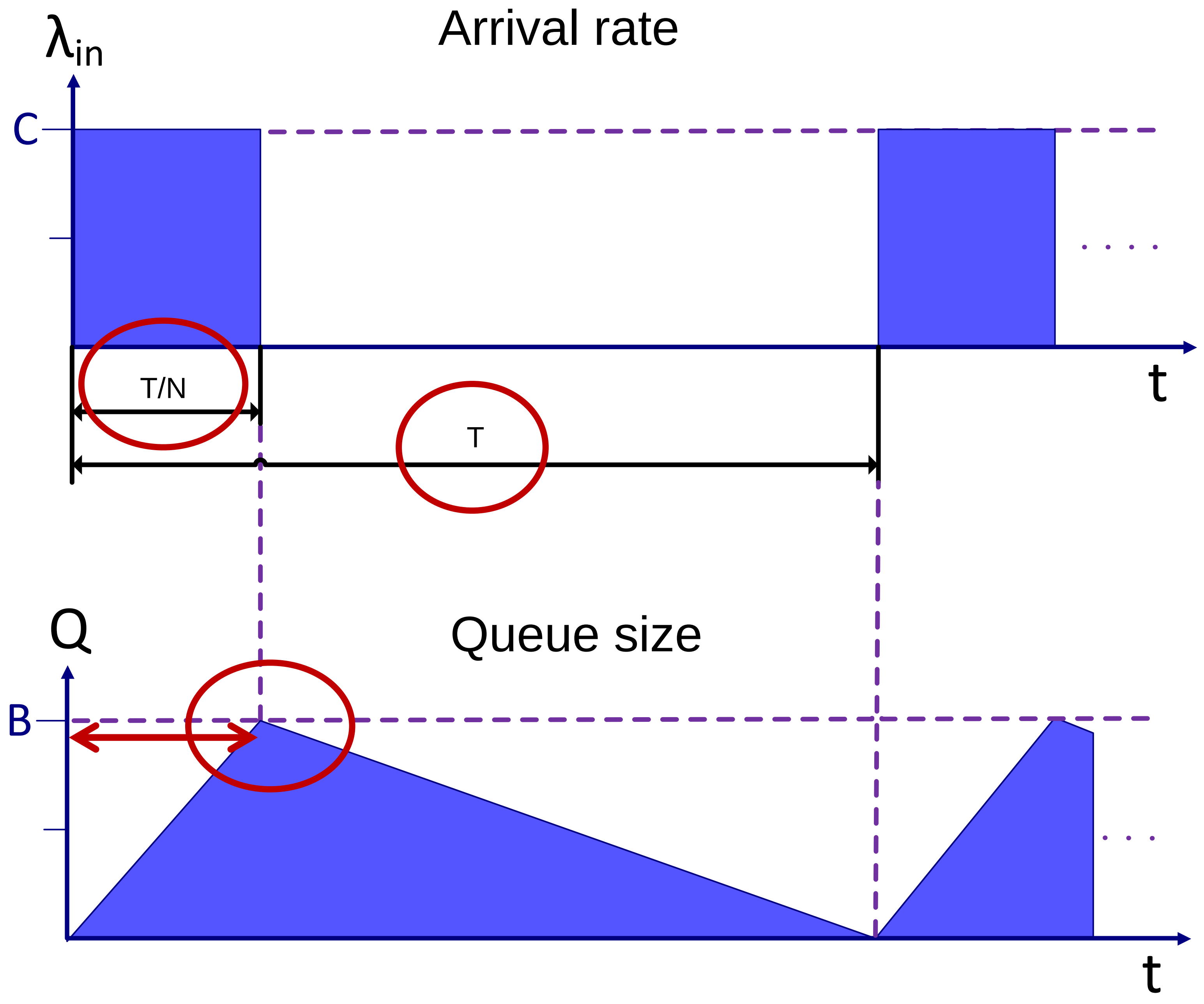
Network Model

The most challenging workload:



SYNCHRONIZED BURSTS

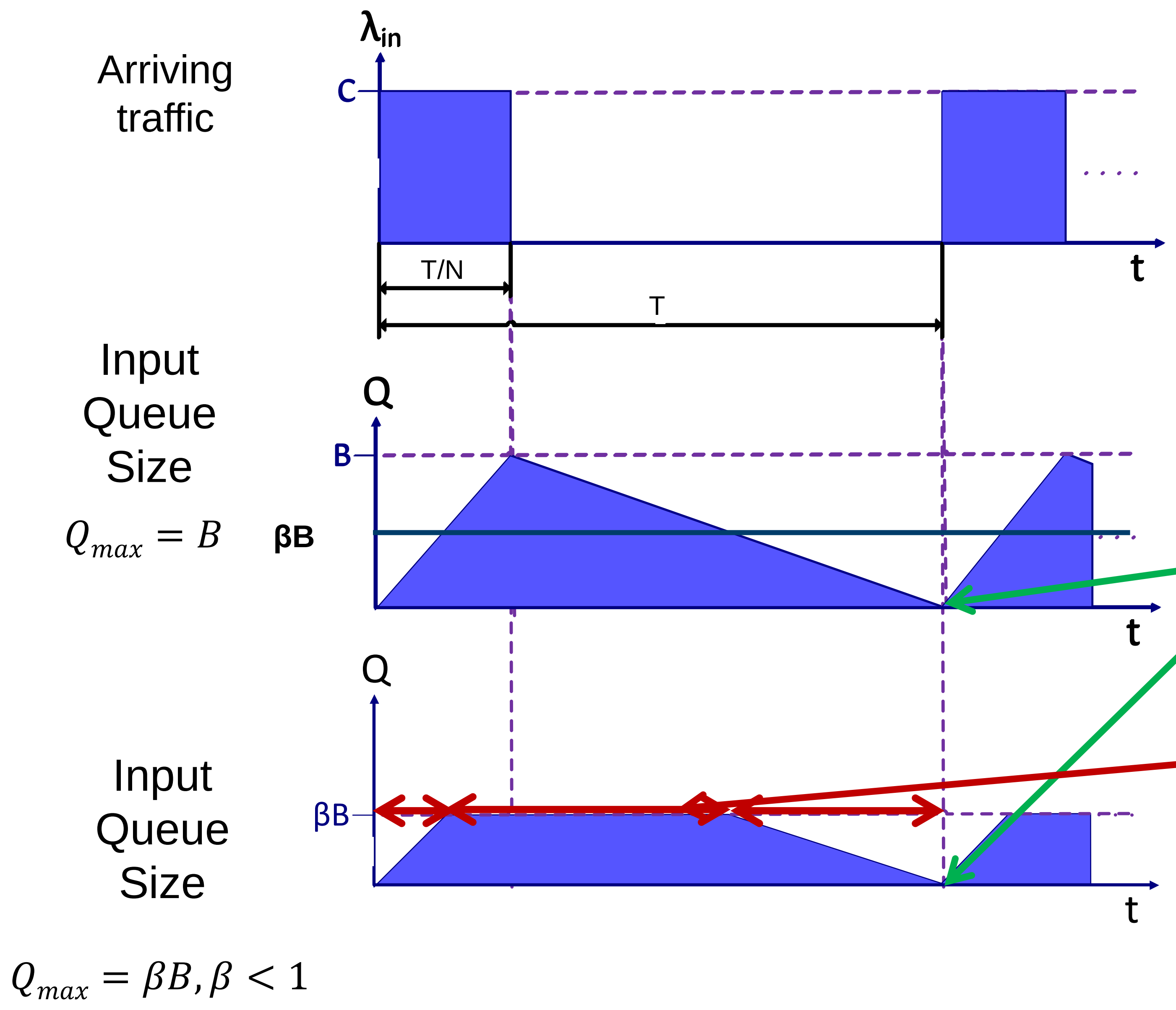
Determining the Workload Parameters



- Assumption: output link is 100% utilized
  - Burst length = T/N (N senders)
- $T = ?$
  - Assume no congestion spreading.
  - It takes $t = {}^{T}/_{N}$ to fill buffer of size B at arrival rate C and departure rate C/N:

$$\frac{T}{N} = \frac{B}{C - {}^{C}/_{N}} \Rightarrow T = \frac{N^2 B}{C(N-1)}$$

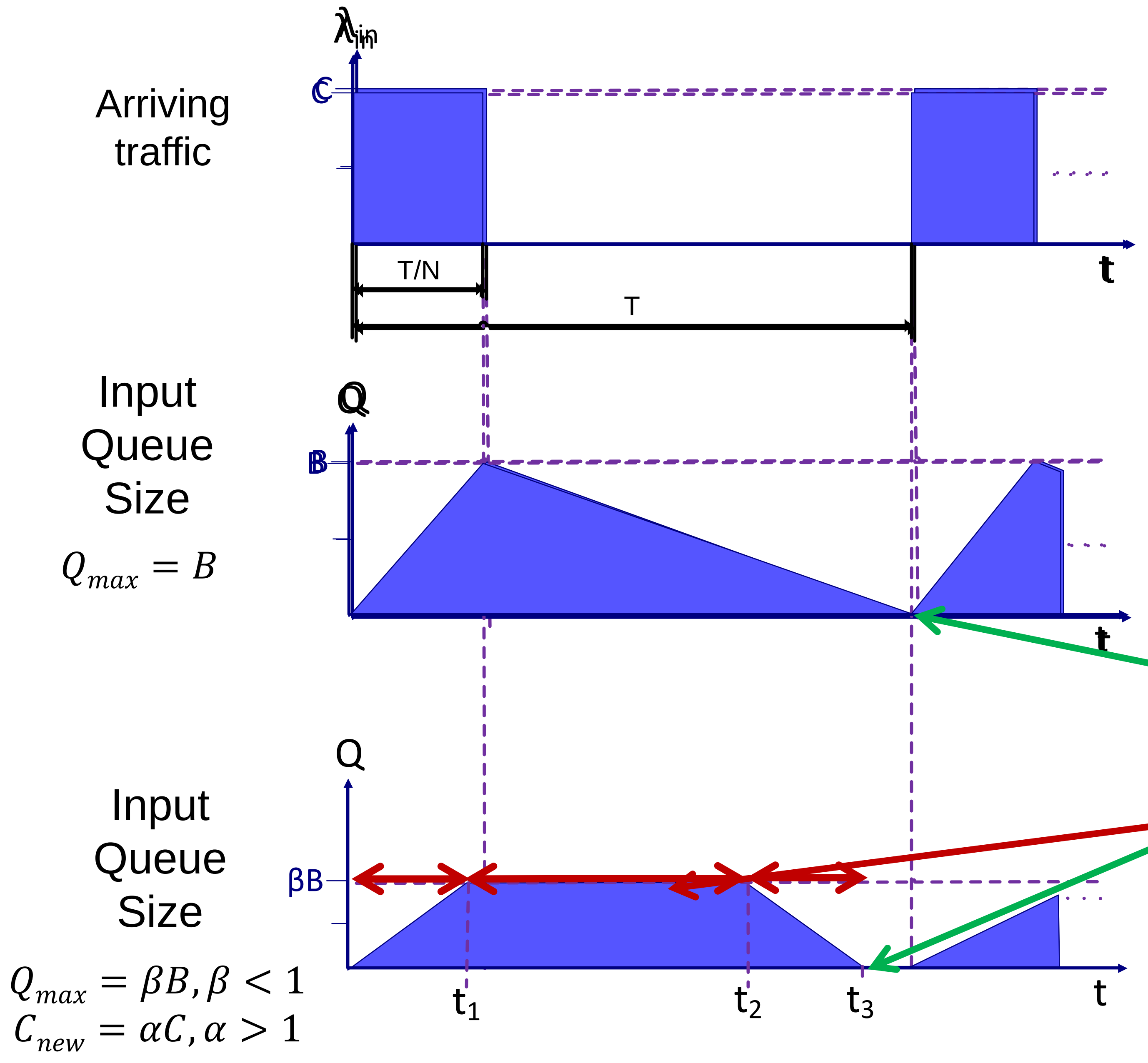
Effect of Buffer Size Reduction



Conclusions:

- With small switch buffers, we are able to push the same traffic
- But, we pay with:
  - Congestion spreading (pause frames)
  - Buffers at the traffic sources (NICs)
    (or application suspending)

Effect of Buffer Size Reduction with Link Acceleration



$$t_3 = \frac{T}{\alpha}$$

$$t_2 = t_3 - \frac{\beta B}{\alpha C/N}$$

$$t_1 = \frac{\beta B}{C(1-\alpha/N)}$$

Conclusions:

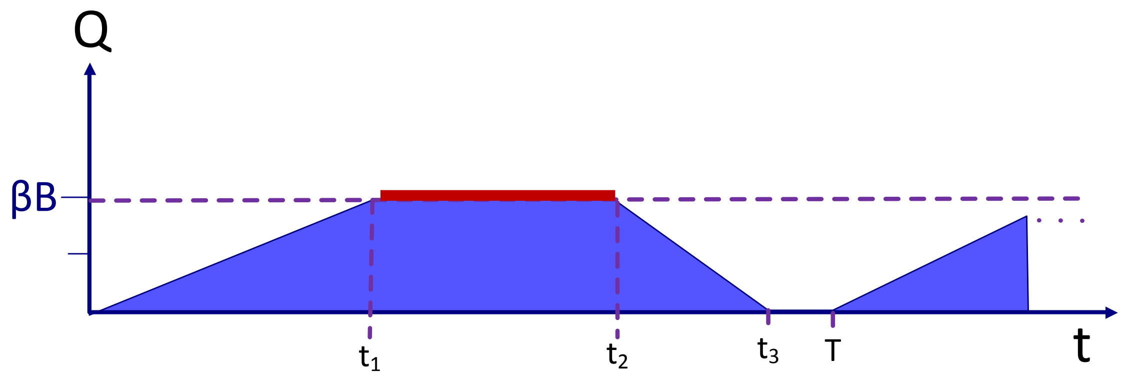
- We can push more traffic.
- When the buffer is full, the link is in *paused* mode: $C_{eff} = {}^{\alpha C}/_{N}$ (congestion spreading)

Buffer Saving vs. Link Acceleration Analysis

- When buffer is full, the link is in *paused* mode:

$$C_{effective} = {}^{\alpha C}/_{N}$$

  - Paused mode ⇒ congestion spreading



$$t_1 = \frac{\beta B}{C(1-\alpha/N)};\ t_2 = t_3 - \frac{\beta B}{\alpha C/N};\ t_3 = \frac{T}{\alpha}$$

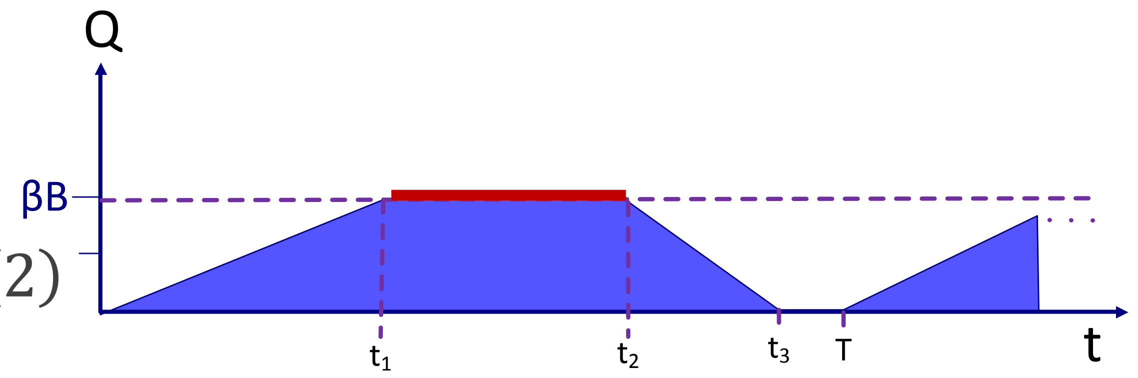
- $\%paused = \frac{t2-t1}{T} = \cdots = \frac{N-\alpha-\beta(N-1)}{\alpha(N-\alpha)}$ (1)

- By how much the buffer can be reduced (β) to avoid congestion spreading ($\%paused = 0$) ?
  - For $\%paused = 0, \alpha = \frac{56\ Gbps}{40\ Gbps} = 1.4, N = 2$ (incast 2 → 1):
    - $\beta = 0.6$ ⇒ 40% of buffer saving!!! ☺
  - For $\%paused = 0, \alpha = \frac{56}{40} = 1.4, N = 10$ (incast 10 → 1):
    - $\beta = 0.95$ ⇒ only 5% of buffer saving ☹

Buffer Saving vs. Link Acceleration Analysis– cont.

- BUT, we are allowed to pause the link, since we increased the link capacity.
  - $C = \alpha C(1 - \%paused) + \frac{\alpha C}{N} \cdot \%p[\ldots]ed \Rightarrow \%paused = \frac{\alpha - 1}{\alpha - \frac{\alpha}{N}}$ (2)
    - For $\alpha = 1.4, \;\; N = 10 : \%pau[\ldots]$

Effective bandwidth =
Link bandwidth * %unpaused



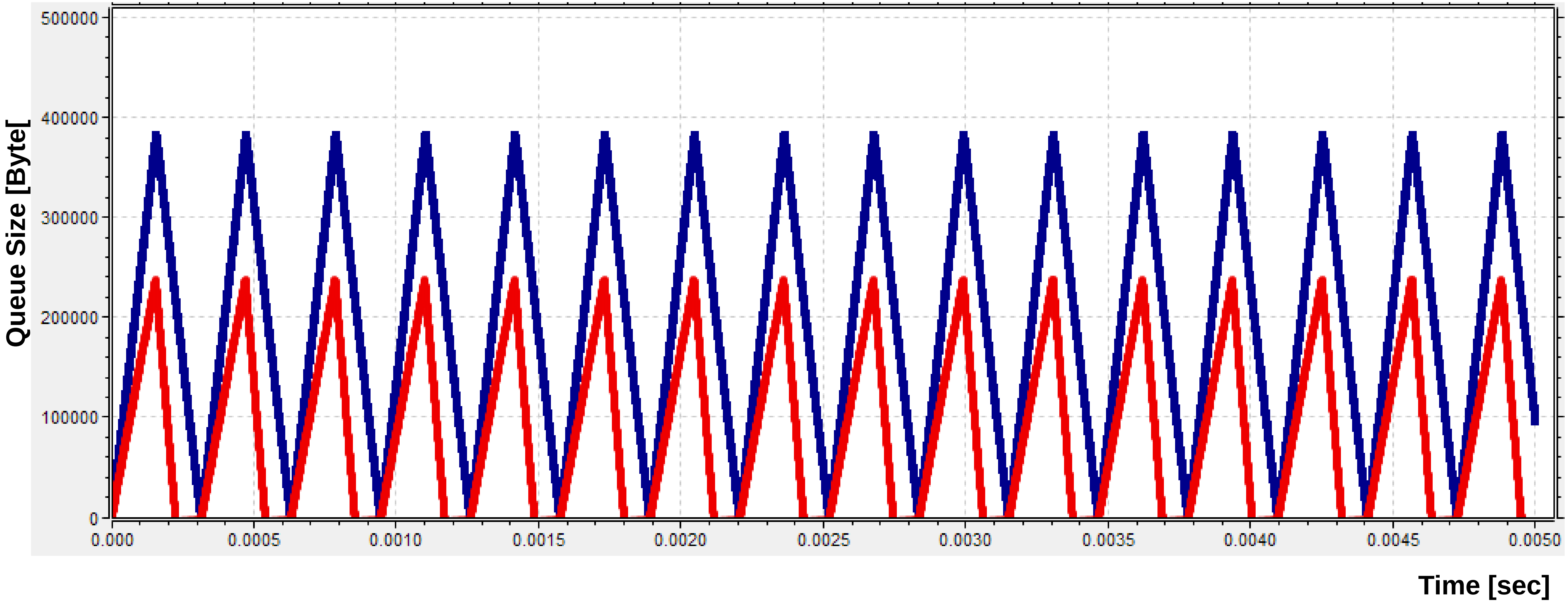
$$t_1 = \frac{\beta B}{C(1-\alpha/N)};\; t_2 = t_3 - \frac{\beta B}{\alpha C/N};\; t_3 = \frac{T}{\alpha}$$

  - Using (1) and (2) $\Rightarrow \beta = \frac{(N-\alpha)(2N-\alpha N-1)}{(N-1)^2}$.
    - For $N = 10, \alpha = \frac{56}{40} = 1.4 \Rightarrow \;\beta = 53\%$!!!
- We can save 47% of buffer size with 40% of link rate increase, to get the same performance!
- And we can also push more data (56Gbps vs. 40Gbps)
  - With the congestion spreading cost

Simulation Results

- Omnet++ simulator with Inet framework
- 2 → 1 Incast

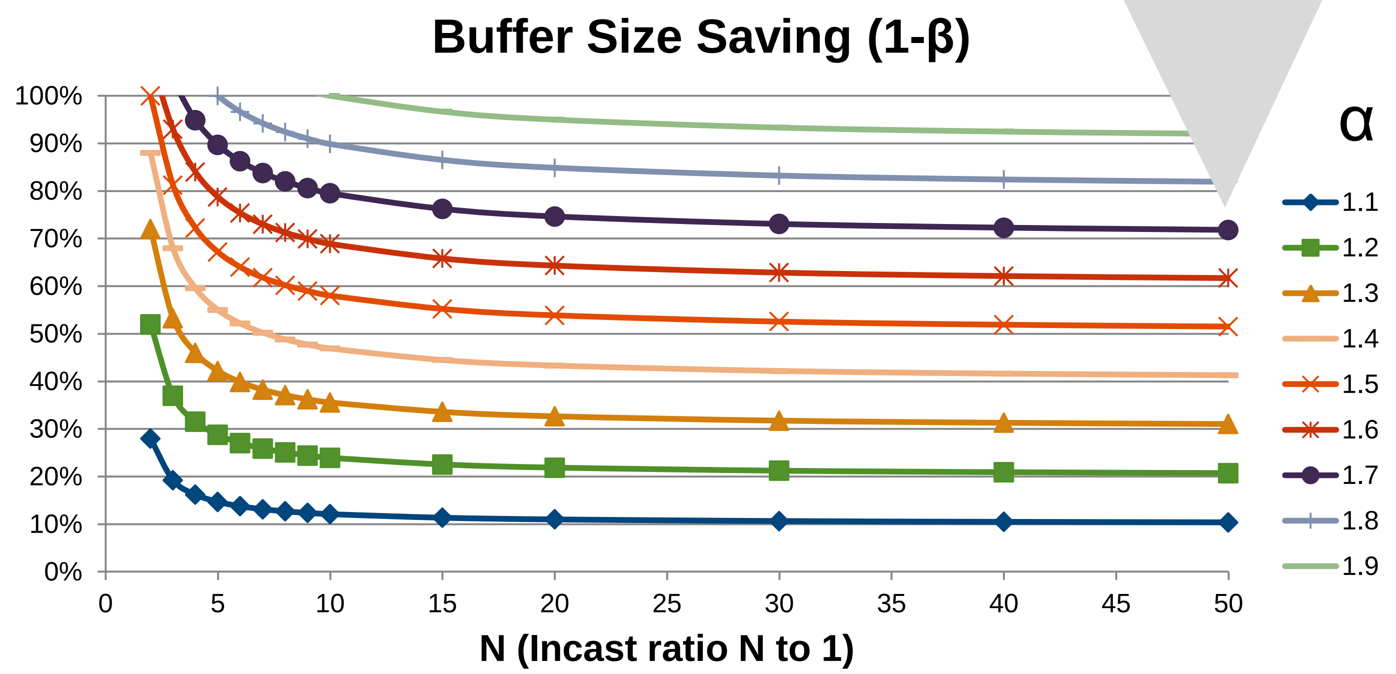
Cout=40 Gbps
Cout=56 Gbps

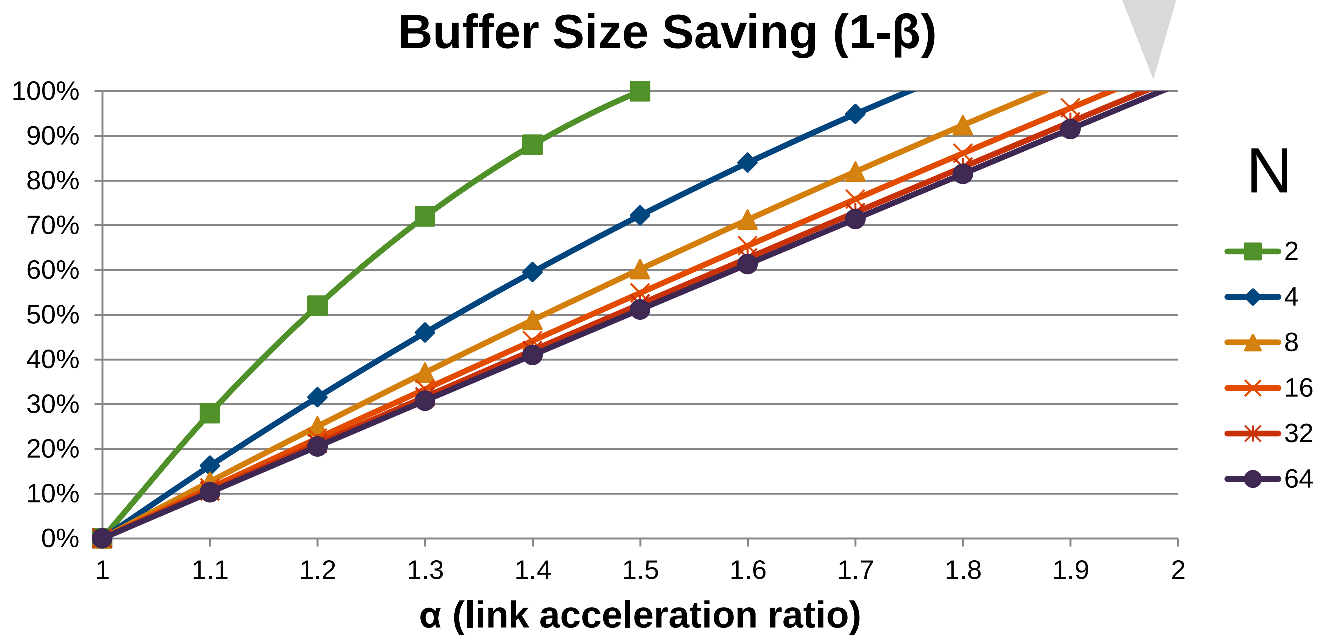


Asymptotic Analysis

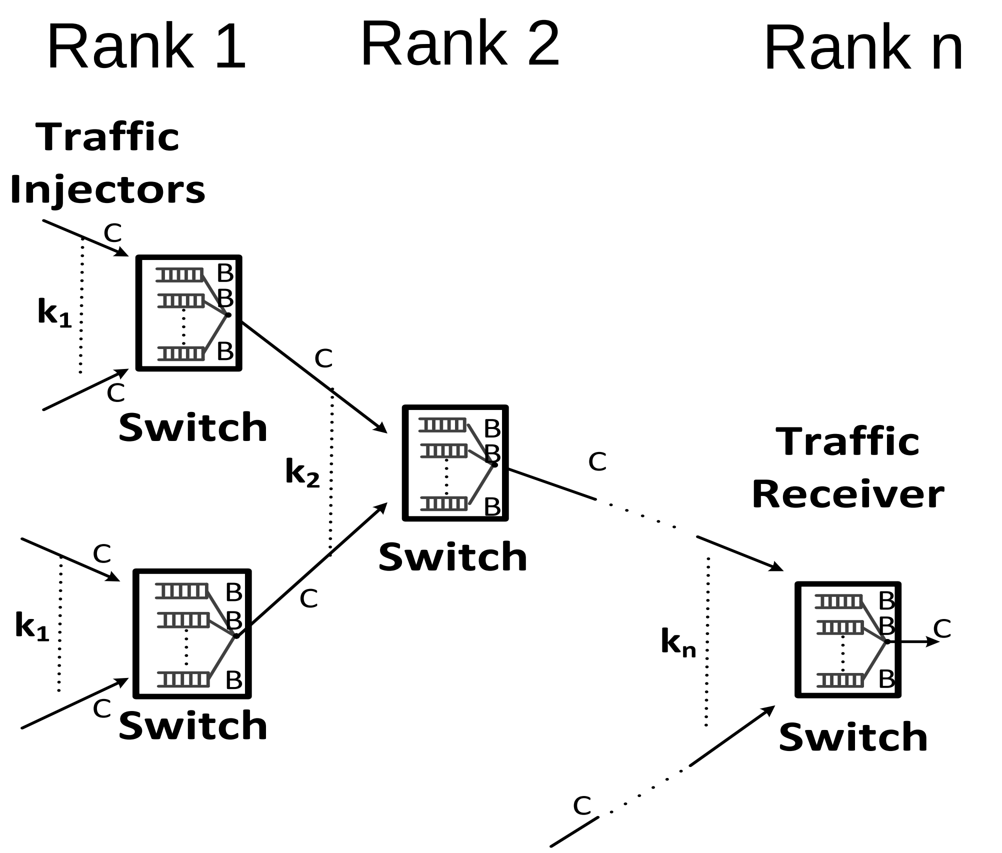
By increasing link bandwidth by X%
the buffer saving is at least X%
for any incast load!

For $\alpha \geq 2$
no buffering
is required in
the switches!

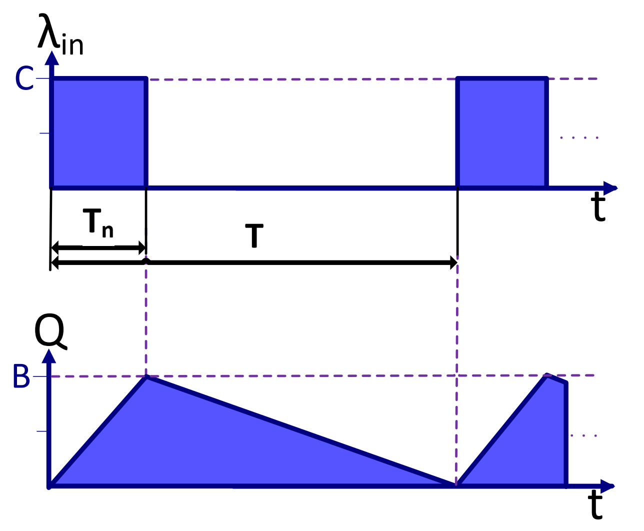




Multiple Incast Cascade Analysis



Last rank defines the workload parameters:

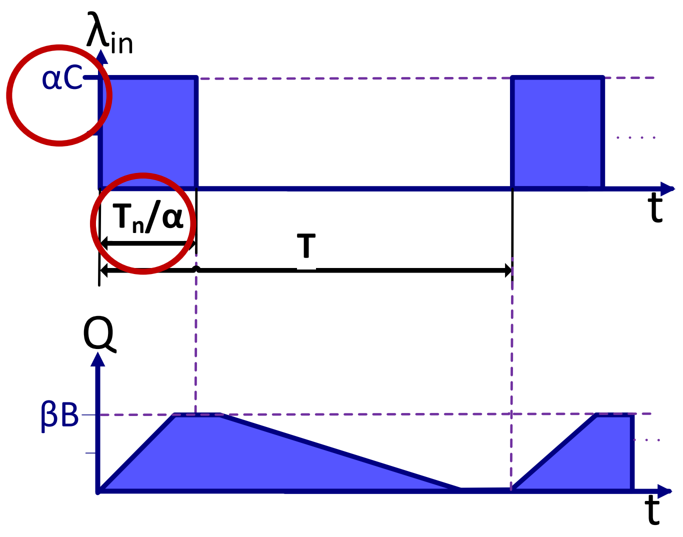


$$T_n = \frac{B \cdot k_n}{C(k_n - 1)};$$

$$T_{n-1} = {T_n}/{k_{n-1}};$$

$$T_1 = \frac{T_n}{\prod_{i=1}^{n-1} k_i}$$

Rank n, α>1, β<1:



The analysis is similar to a single-rank case, but now the traffic arrives at rate αC

$$\boldsymbol{\beta = 1 - \alpha * \%paused}$$

$$\beta(\alpha = 1.4, \%paused = 32\%) = 0.55$$

Conclusions

- We presented a method for analyzing the buffer-bandwidth tradeoff based on the Incast scenario in lossless networks.

- We can reduce switch buffer size, while still pushing the same traffic.
  - But, we pay with:
    - Congestion spreading (pause frames)
    - Buffers at the traffic sources (NICs) or suspending application.

- By increasing the links bandwidth, we can reduce the congestion spreading.
  - And push more traffic.

- We can save X% of buffer size with X% of link rate increase (for any incast).

- When increasing the links bandwidth by a factor of at least 2 ($\boldsymbol{\alpha \geq 2}$) no buffering is required at the switches.

- The results hold also for the multiple incast cascade.

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™