OPENFABRICS ALLIANCE

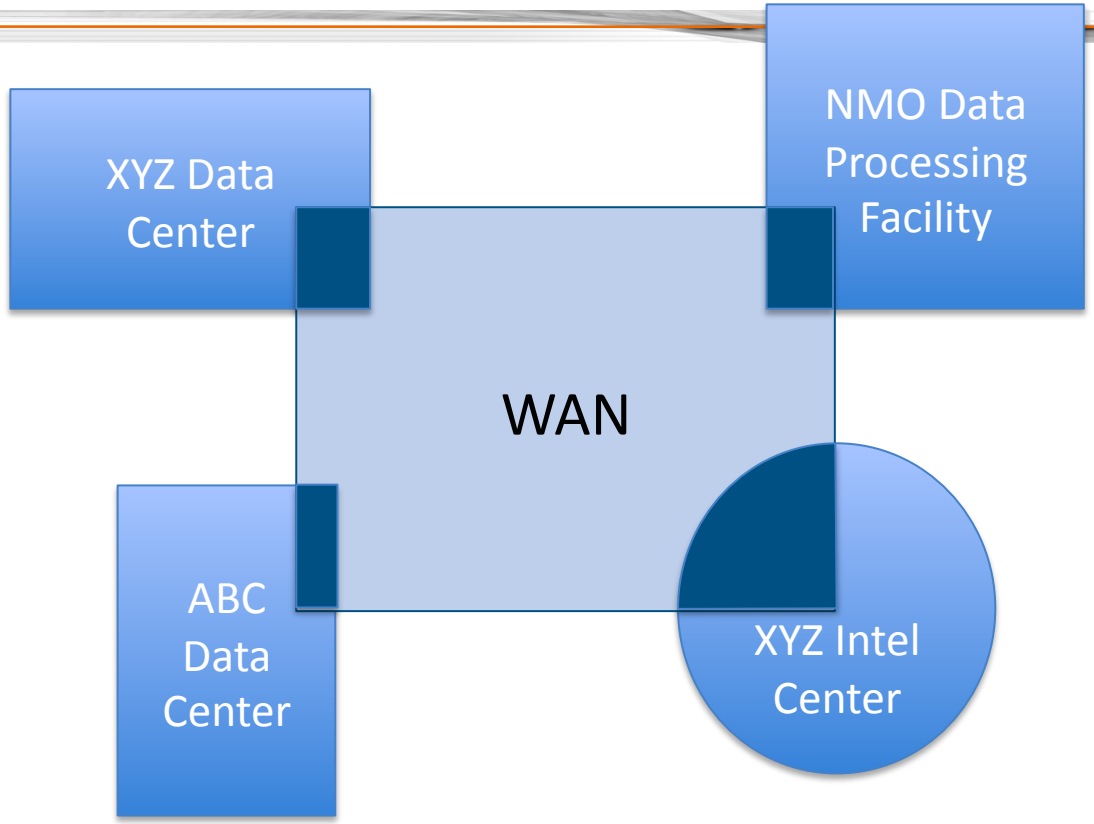
2013 OFA Developer Workshop

RDMA in Wide Area Networks

Linden Mercer, ARL/PSU <lbm2@psu.edu>
Srinath Jayasundera, Exelis <srinath.jayasundera@exelisinc.com>
Eric Kinzie, Exelis <eric.kinzie@exelisinc.com>

Work done under contract to Naval Research Laboratory

Extended Data Center Fabric



Needed: A Resilient, Robust, and Very Efficient Data Center to Data Center WAN

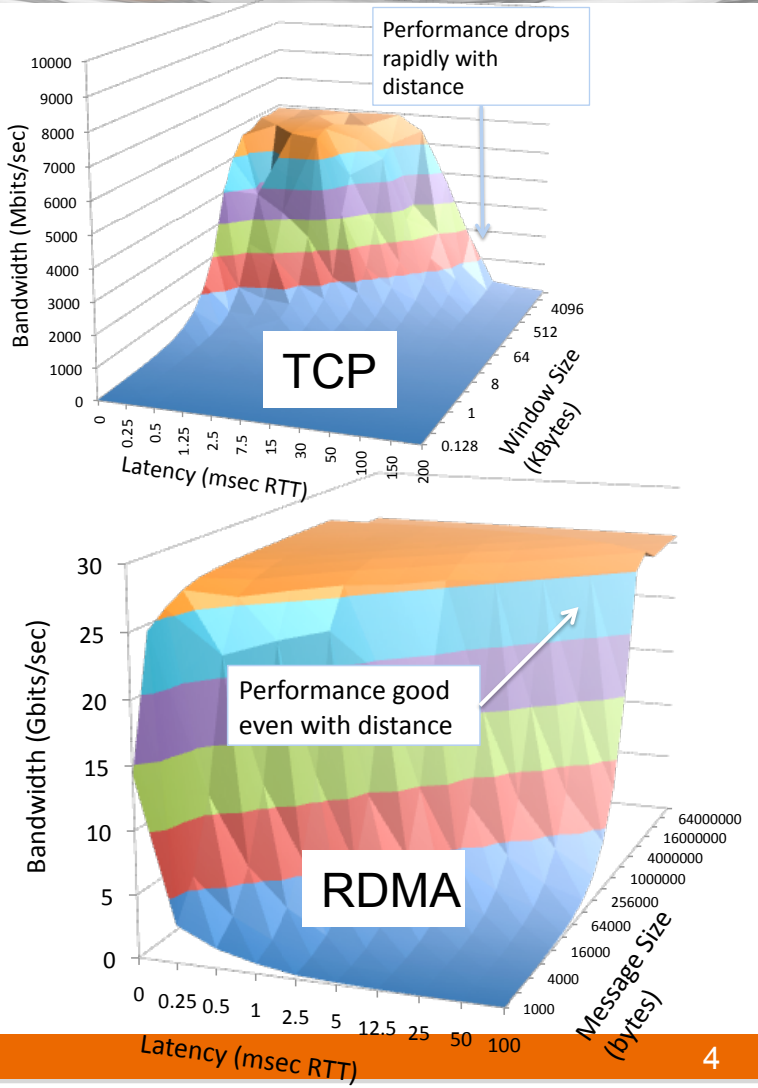
RDMA over Converged Enhanced Ethernet

- Transport layer – RDMA RC and UC, OFED compatible
  - Benefits of RDMA
    - CPU Offload
    - Large effective window size – low time*bandwidth impact on performance
- Network layer – No IP layer, “Lossless” Ethernet
  - Benefits
    - Good data center flow control

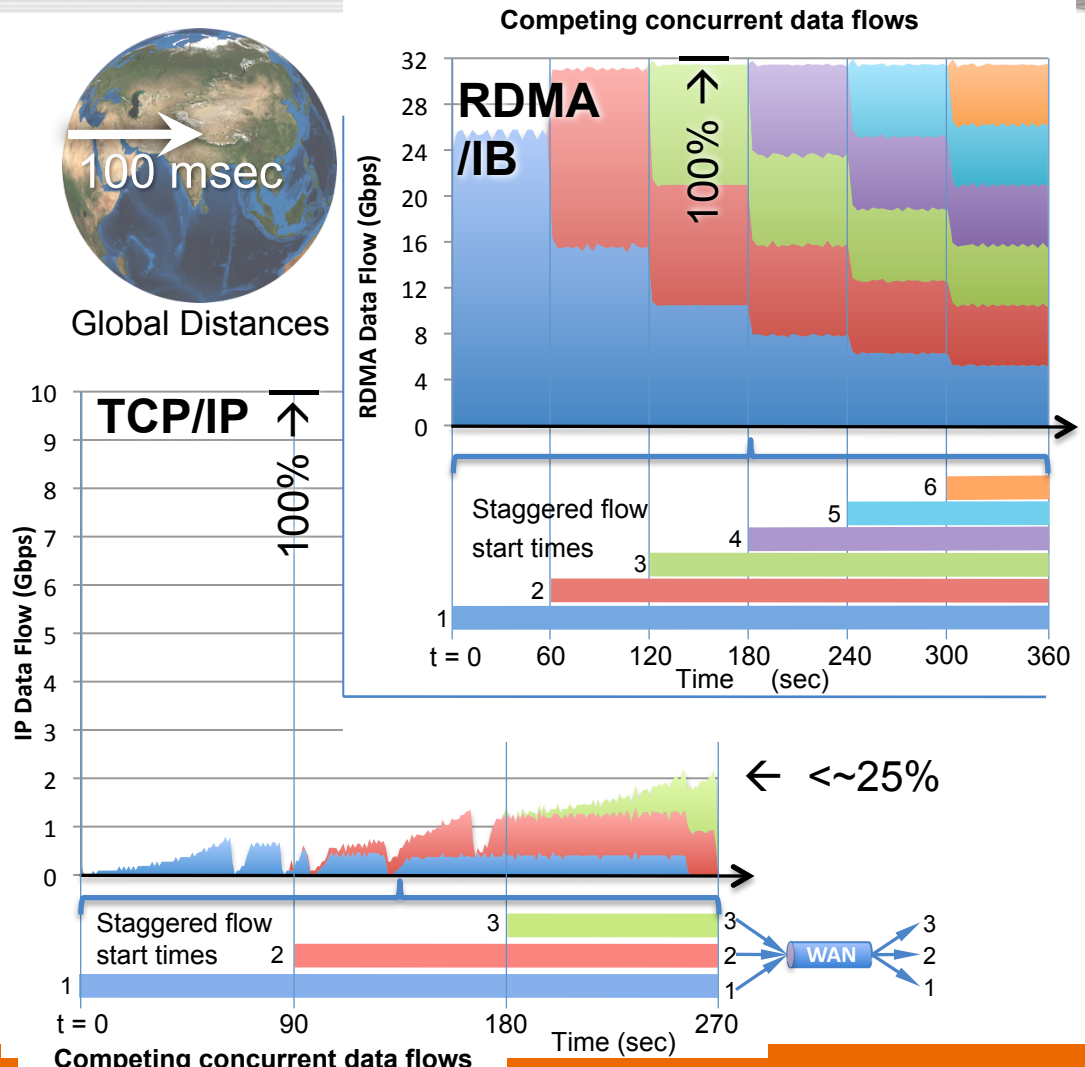
RDMA

- Much more effective long distance transport than TCP/IP
- Proven performance over global distance in LD JCTD using RDMA/InfiniBand
- System integration/testing over 40/100G Ethernet begun
  (ESNet, IU, OSV, NRL, …)
- Performance requires very low loss network

Bandwidth performance vs. window size & Latency (RTT)

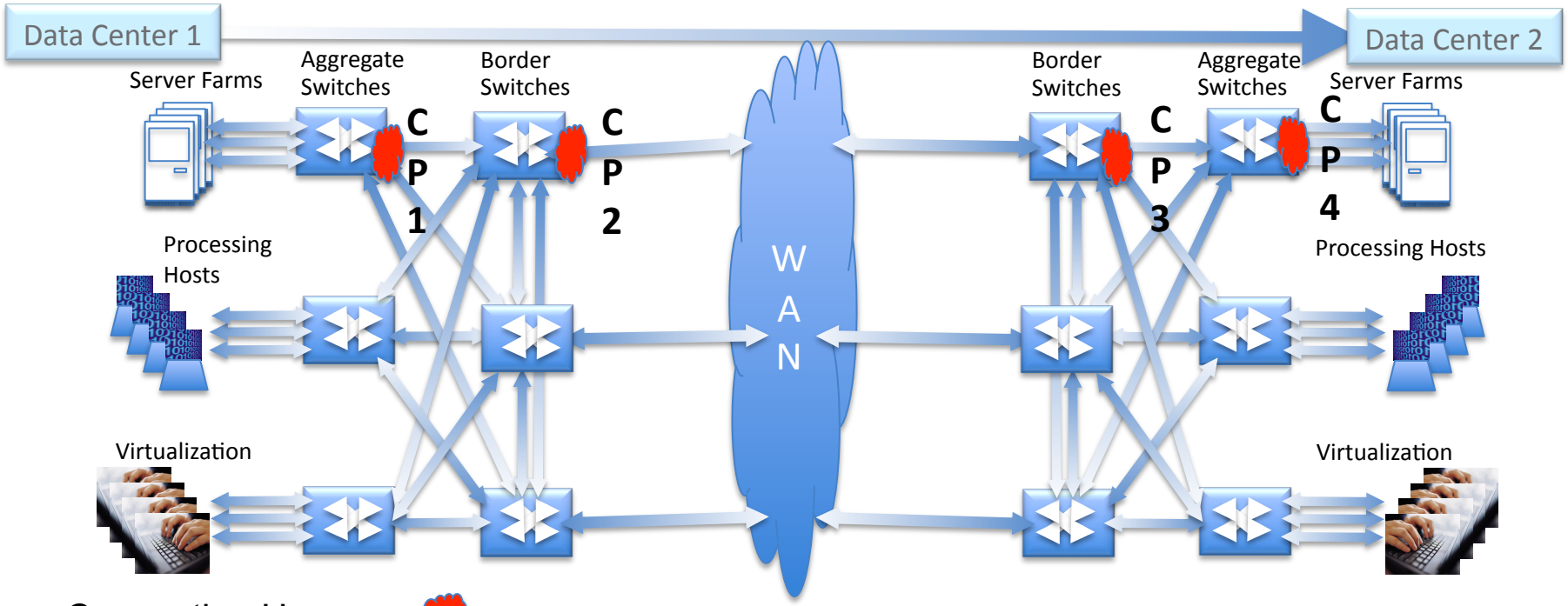


NRL Large Data Networking



- Near 100% Efficient (vs. 25-40% IP efficiency)
- Robust, resilient response to surge loads, outages
- Like Google G-scale but *without* central management
- Ready for IP/MPLS, 40G, Carrier Ethernet WAN
- Scalable and extensible
  – 10 to 100x faster

Data Center to Data Center Layer 2 WAN



Congestion Happens

- LAN
  - Handled by good data center networking
- LAN to WAN
  - Concurrent flows to a common WAN link
- WAN to LAN
  - Concurrent flows from multiple WAN links to a common resource

How Can RoCE be "Lossless" Over the WAN

- "Lossless" network is the result of good flow control options but those features are not currently available beyond data center distance
  - No flow control in Carrier Ethernet
  - Unnamed network expert: "you cannot use 802.3X outside of data center distances since it was not designed for that, i.e. this would result in poor link utilization."
- Work around options
  - Engineer paths with near lossless data delivery up to some specific data rate
    AND
  - Rate limit sources to never exceed available bandwidth
  - Use RoCE point to point so congestion won't happen in the network path (destination system can still be unavailable)
- Solution – Extend good flow control over WAN

Ethernet Flow Control - Introduction

Pause frame contains a field signaling the duration of pause it is requesting from the link transmitter.

This field is a 16-bit field and each bit represents a quanta which is equal to 512 bit times.

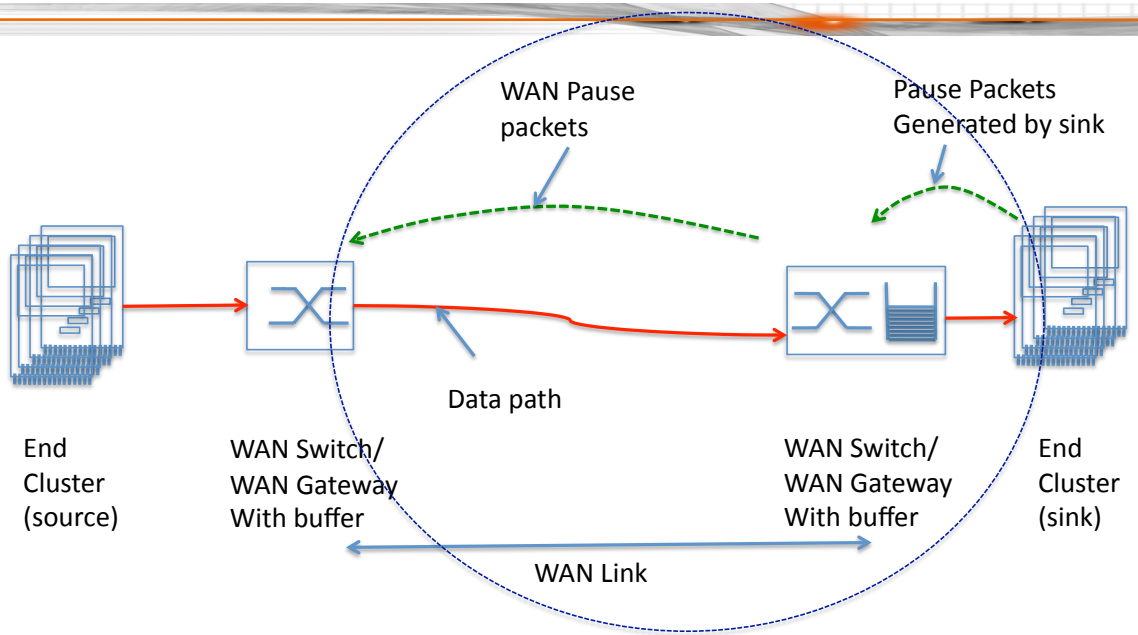
Line rate	bit time	Quanta	Max Pause time
1 Gbps	1 ns	512 ns	33.5 ms
10 Gbps	0.1 ns	51.2 ns	3.35 ms
40 Gbps	0.025 ns	17.92 ns	0.84ms
100 Gbps	0.01 ns	5.12 ns	0.335 ms

Hence maximum pause time is a function of the line rate (and it is much shorter than WAN RTT).

Ethernet WAN Flow Control

- Basics
  - Simple control loop
  - Link receiver must be able to buffer a full Round Trip Time (RTT)
- Objective
  - Link should deliver data at nearly the same rate that the network element (or end system) is able to consume it
  - Link receiver must be able to signal for anywhere from 0 to 100% line rate
- Option
  - New flow control signaling for Wide Area “Lossless” Ethernet
  - Use the existing PAUSE signal but send PAUSE according to an algorithm that will accomplish WAN flow control
- NRL analysis and simulation shows current IEEE Ethernet standard flow control signaling can work over global distances at high speeds

WAN Network Model



Unidirectional WAN network model used for simulation

- A random pause is generated by “End Cluster” (uniform distribution function)
- Source side always has data available
- Both source and sink rates are adjustable
- Each time step denotes ‘maximum pause’ time and hence is related to line rate

ON/OFF Pause Control

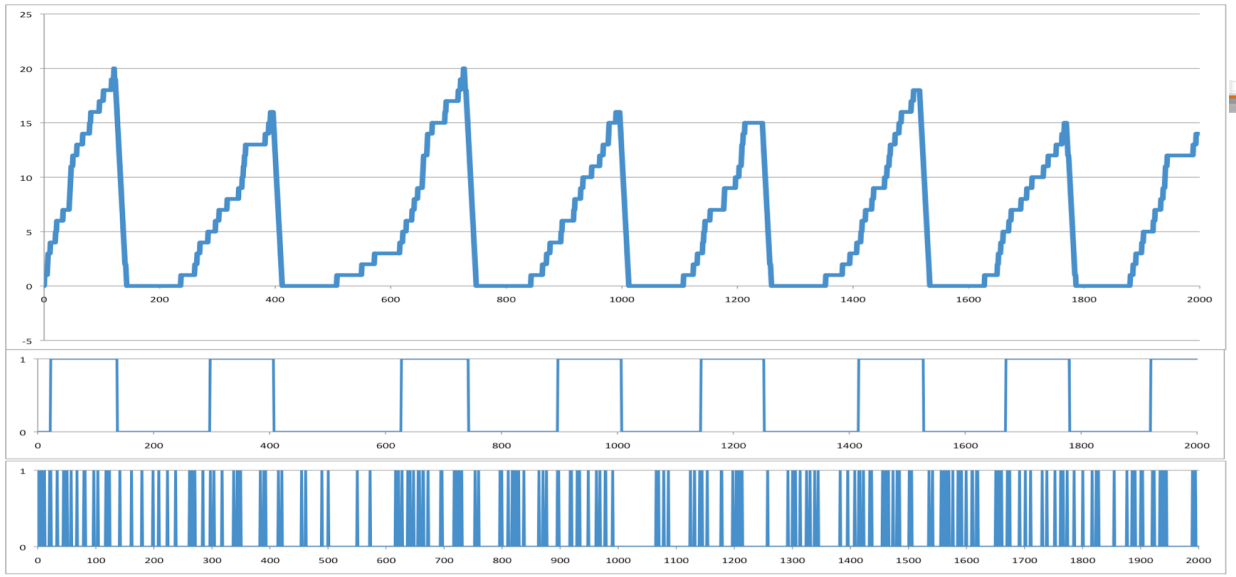
- Algorithm
  - A Max Pause packet (0xFFFF) is generated every ‘maximum pause’ time interval if buffer level is greater than threshold
  - If buffer level is less than threshold do nothing
- The buffer size should be at least Line Rate*RTT plus threshold

ON/OFF Pause Control

Buffer Occupancy

Pause to source

Pause from sink

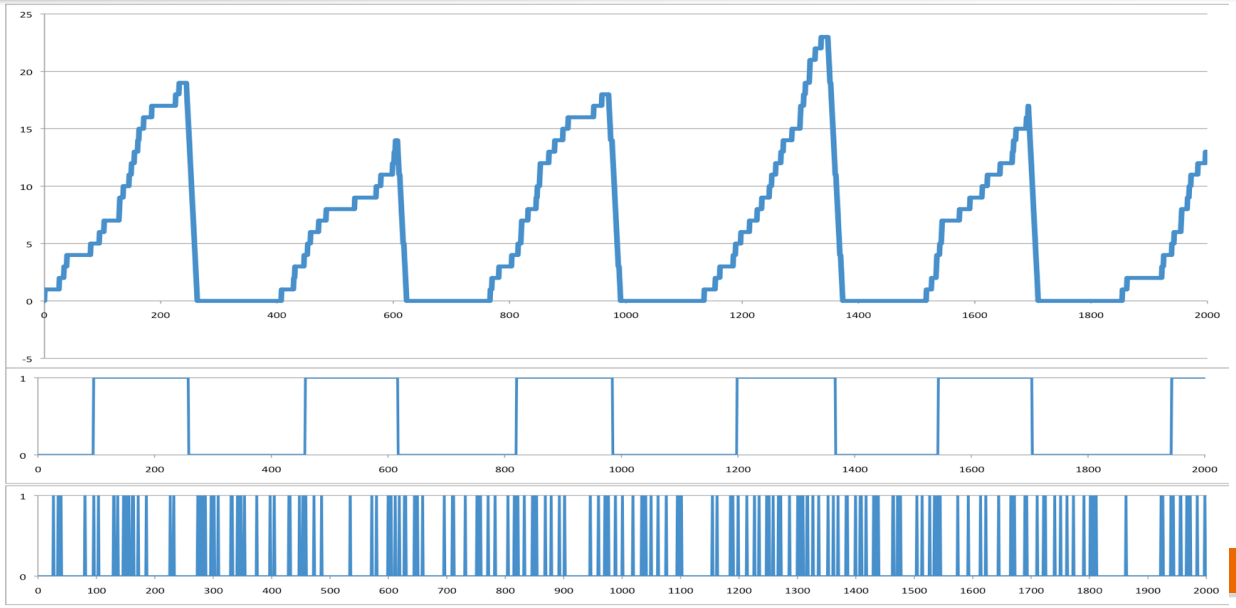


Equal source and sink line rates
Threshold = 5
RTT = 100 max PAUSE
(For 100G RTT = 33.5 ms)

Buffer Occupancy

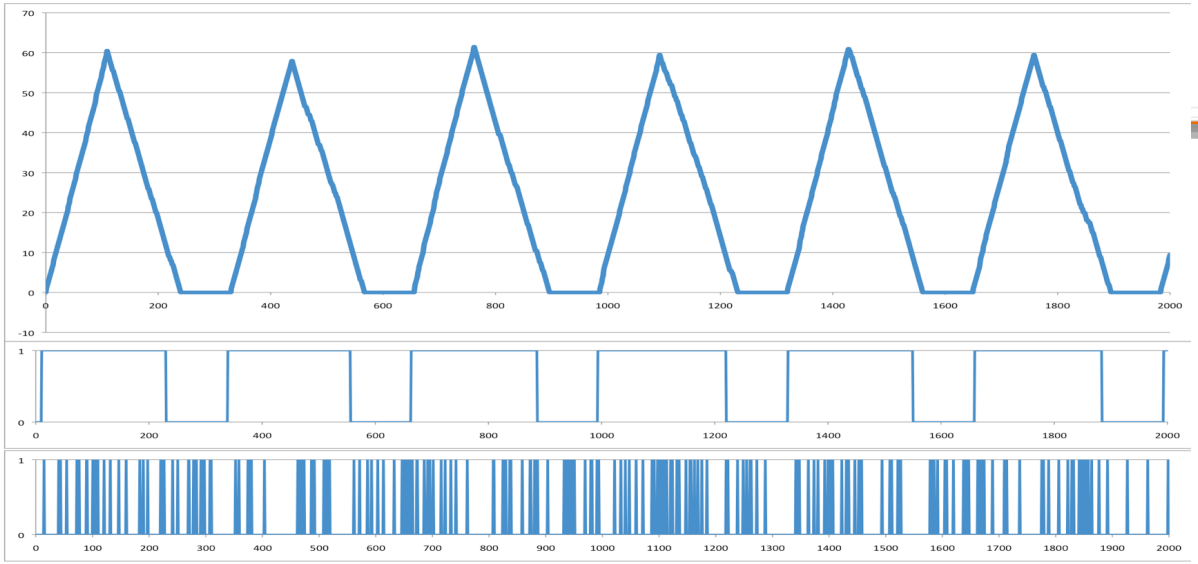
Pause to source

Pause from sink

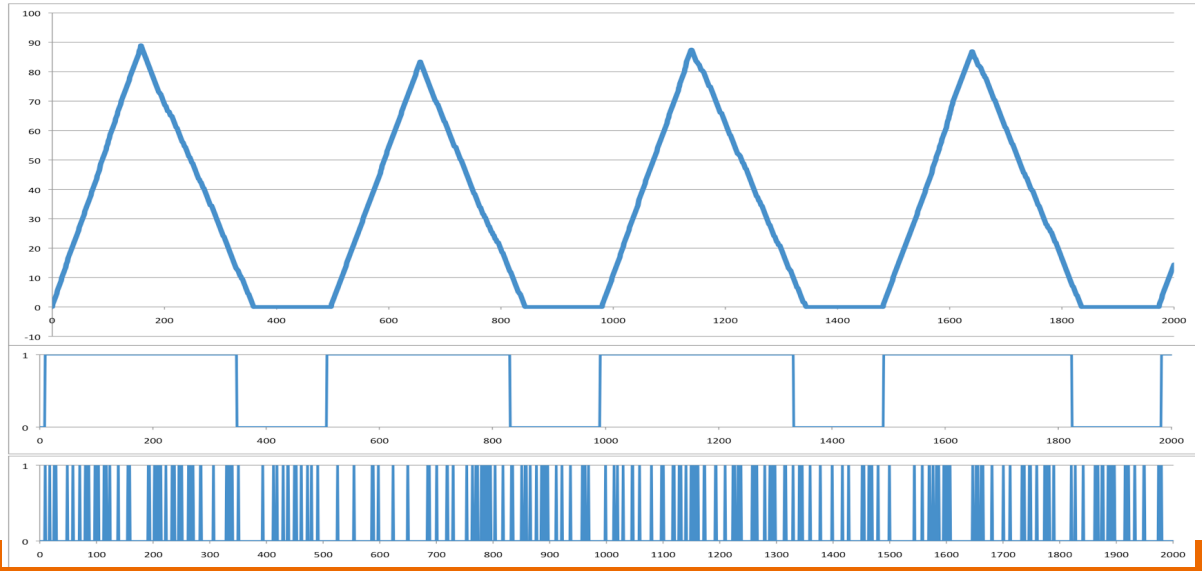


Equal source and sink line rates
Threshold = 5
RTT = 150 max pause times
(For 100G RTT = 50 ms)

ON/OFF Pause Control



Threshold = 5
RTT = 100
i.e. For 100G
RTT = 100x0.335ms
= 33.5 ms



Threshold = 5
RTT = 150

Time Based Pause Control

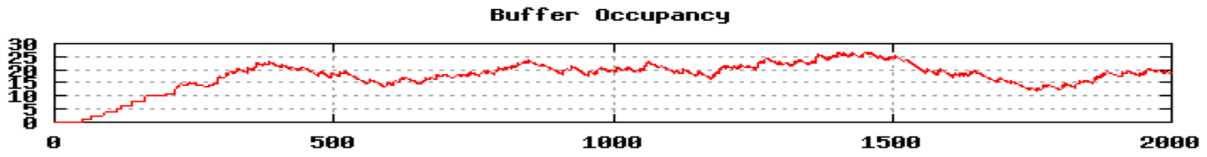
- Algorithm
  - Pause Packet is generated at every max pause interval according to the following equation
  - $$Pause\_time\ (t) = \frac{Buffer\ Occupancy\ (t)}{(Buffer\_height - Characteristic\_Constant) \leq 1}$$

    Send no PAUSE when Buffer Occupancy (t) < Threshold

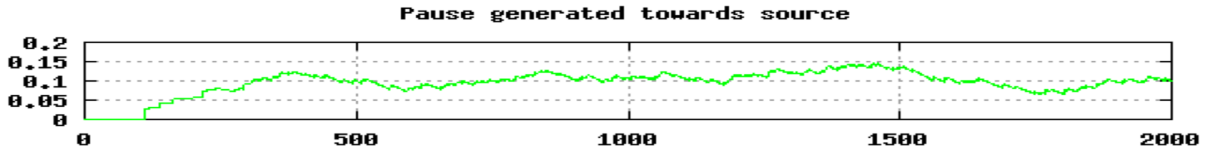
  - The buffer size should be at least Line Rate*RTT plus threshold
  - Characteristic constant try to capture desired behavior of the channel according to the data type
  - Sends a pause at each interval allowing data to be sent between 0 and 100%

Time Based Pause Control

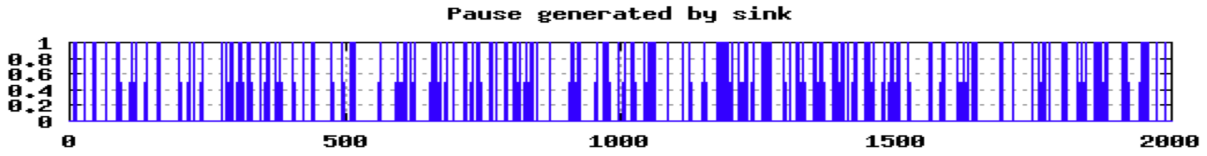
Buffer Occupancy



Pause to source



Pause from sink

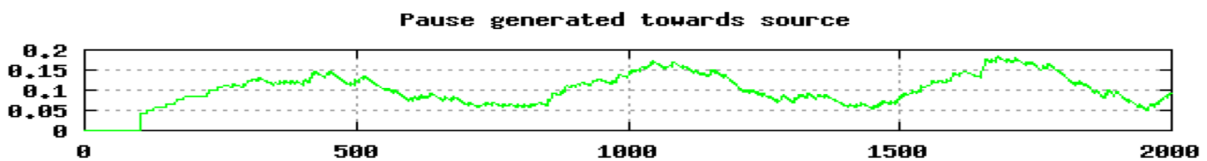


Threshold = 5
RTT = 100
i.e. For 100G
RTT = 100x0.335ms
= 33.5 ms

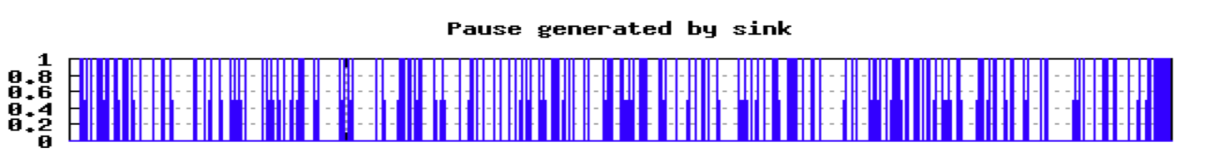
Equal source and sink rates (Maximum)

Buffer Occupancy

Pause to source



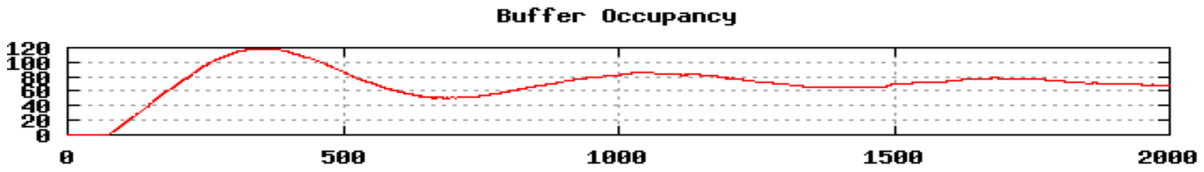
Pause from sink



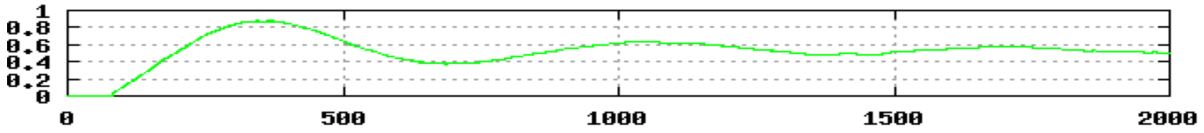
Threshold = 5
RTT = 150

Time Based Pause Control

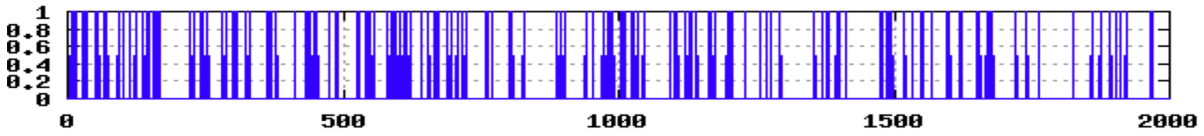
Buffer Occupancy



Pause to source



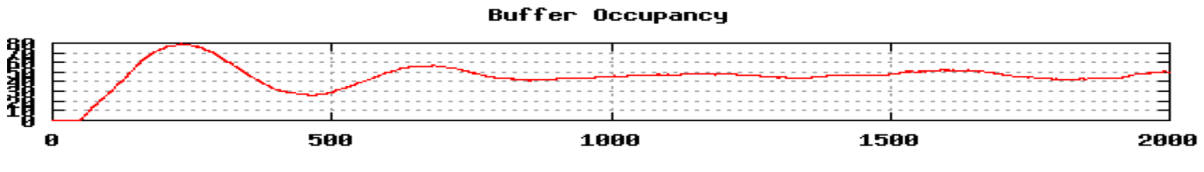
Pause from sink



Threshold = 5
RTT = 100
i.e. For 100G
RTT = 33.5 ms

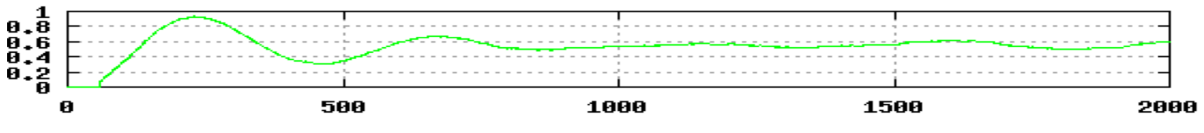
Ratio of 2 for source and sink rates (Maximum)

Buffer Occupancy

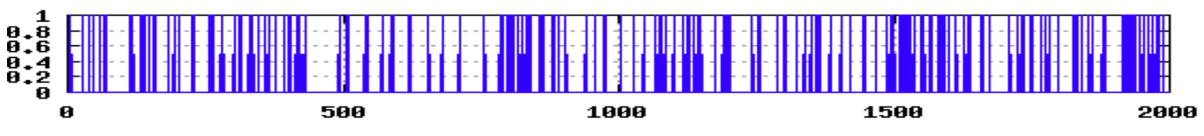


Threshold = 5
RTT = 150

Pause to source



Pause from sink

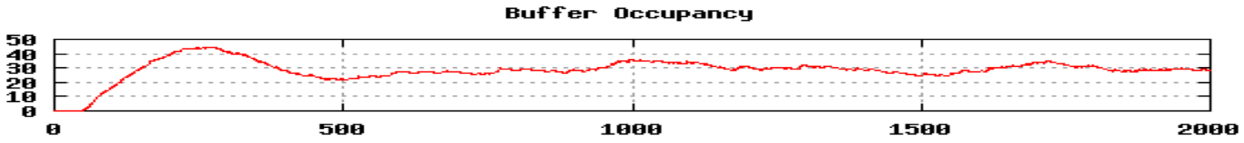


Time Based Pause Control

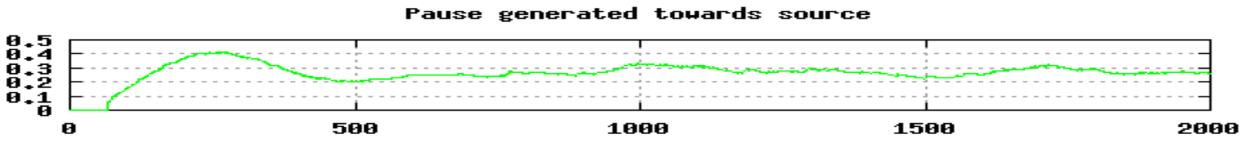
- Analysis of Characteristics constant
- HIGH value should be used if data latency is the primary criterion
- LOW value should be used if data starvation is the primary criterion – i.e. non real-time video transmission

Time Based Pause Control

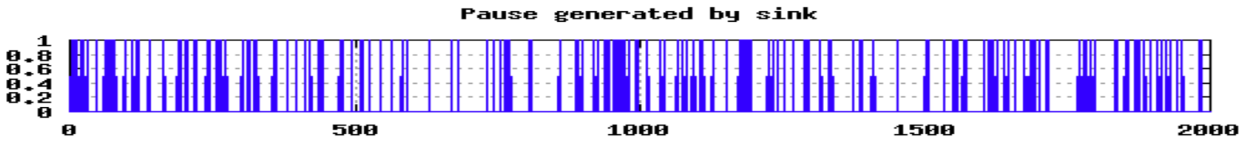
Buffer Occupancy



Pause to source

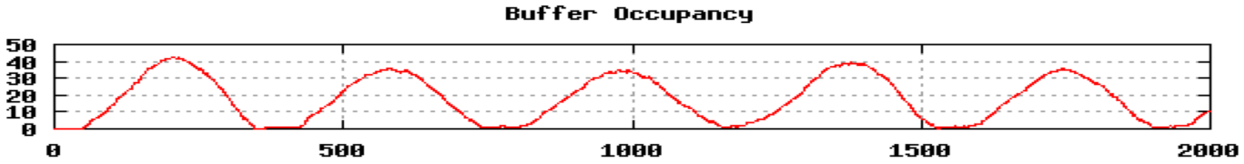


Pause from sink

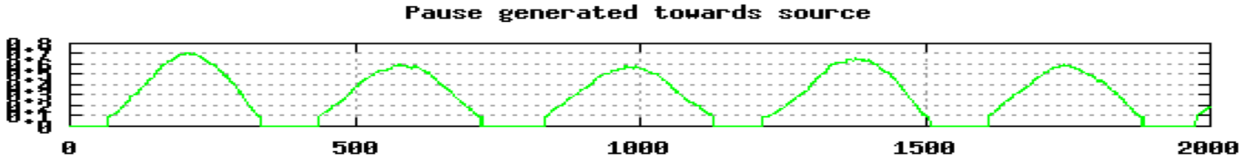


C = 2 and 20 for RTT = 100 and buffer height = 110

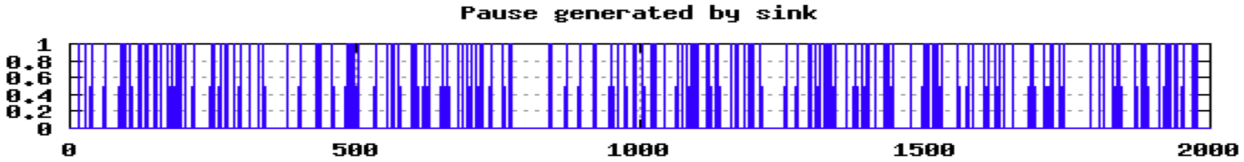
Buffer Occupancy



Pause to source

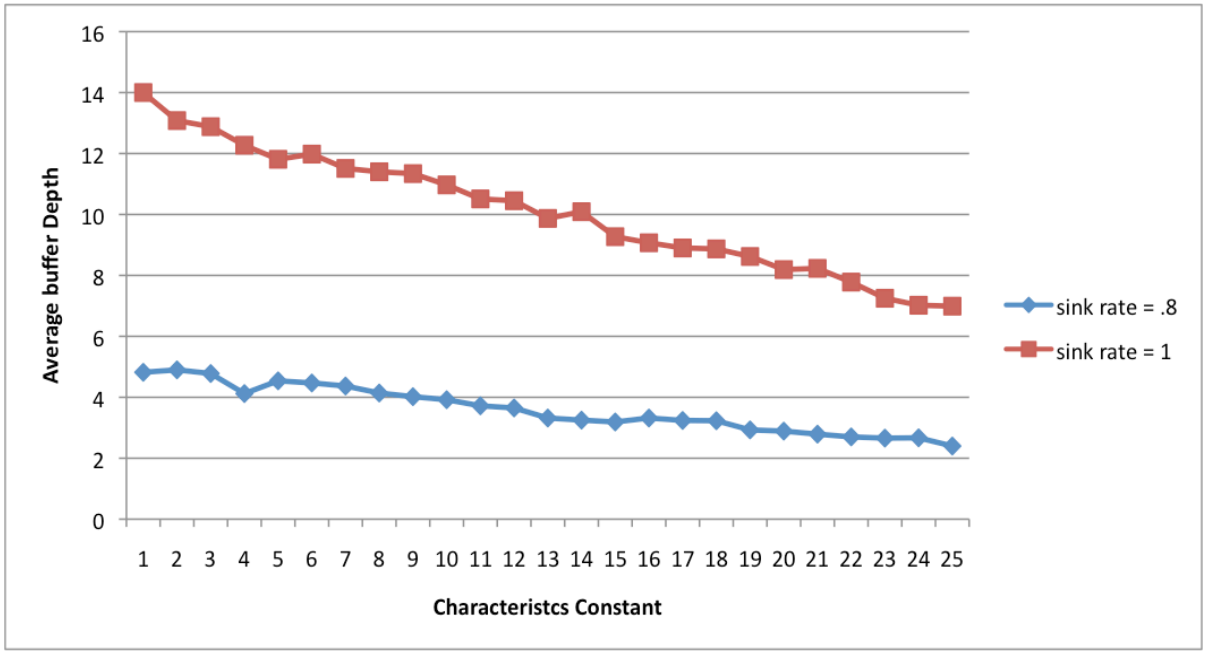


Pause from sink



Time Based Pause Control

Behavior of buffer depth as a function of Characteristics constant



802.3X Flow Control for WAN

- Pro-active rather than re-active
- Does not break the standard
- Simple simulation shows several options
  - On/Off PAUSE frames on max pause interval
  - Timed PAUSE
  - Other options (Interval based, more complex, ...)
  - Implementation should be per priority consistent with PFC
- Implementation options
  - Software (white box)
  - Hardware prototype
  - Commercial switches?

Thank You

OPENFABRICS ALLIANCE

Backup

Computer Room Fabric Extension History

- HiPPI over ATM over SONET(HAS)
- GSN (SuperHiPPI) ATM Network Adapter (GANA) – NRL at SC2000
- Fiber Channel over SONET (FC-BB, Not NRL) NRL at SC2002
- InfiniBand over WAN (NRL's IDE, Obsidian, Bay) SC2004, 2005,… and LD, RR, …
- RoCE over WAN (IU, LBNL, StonyBrook, …)

Approach Applies to IB over WAN as well

- The same basic strategy can be applied to IB flow control over the WAN
- Sending a certain number of credits is equivalent to sending the max size pause minus that number of credits.
- Should send credit messages more often (interval equal to or less than the time corresponding to max credits) and not strictly send credits equal to buffer available
- Requires some addition (adjustment?) to standard since IB specifies more about the algorithm for sending credits (Ethernet standard leaves more of this open). IB includes a Flow Control Total Blocks Sent (FCTBS) value and details for calculating Flow Control Credit Limit (FCCL)

InfiniBand Equivalent

OPENFABRICS ALLIANCE

- IB signals “send” credits instead of pause times but can be equivalent
- Flow control packet contains a FCCL field which extends “permission” to send some quantity of data.
- This field is a 12-bit field and each bit represents a “symbol” which is 8 bits of data.
- *Hence maximum “send” time is a function of the line rate.*
- *For WAN IB flow control, the interval between Flow Control Packets should be less than or equal to max send time (about 104K Flow Control Packets/second for EDR – workable but a larger quanta would make this nicer)*

InfiniBand Equivalent

OPENFABRICS ALLIANCE

Line rate	bit time	Flow Control Block (FCB)	Max Send time (symbol*2^12)
SDR (8 Gbps)	125 ps	64 ns	131 µs
DDR (16 Gbps)	62.5 ps	0.5 ns	65.5 µs
QDR (32 Gbps)	31.25 ps	0.25 ns	32.8 µs
FDR (54.3 Gbps)	18.4 ps	9.43 ns	19.3 µs
EDR (108.6 Gbps)	9.2 ps	4.71 ns	9.65 µs


NAVAL RESEARCH LABORATORY
WASHINGTON, DC


OPENFABRICS
ALLIANCE