Issues in The Design of Exascale Networks

HiPINEB

February 5, 2017

NVIDIA

Bill Dally

Chief Scientist and SVP Research, NVIDIA Corporation

Professor, Stanford University

Acknowledgement

This talk presents the work of the NVIDIA Network Research Group

Larry Dennison (Director)

Nir Arad

Matt Blumrich

Hans Eberle

Ted Jiang

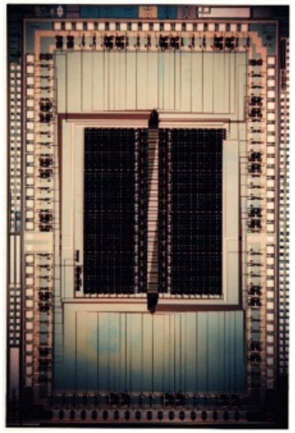
Alex Ishii (DGX Product Group)

Outline

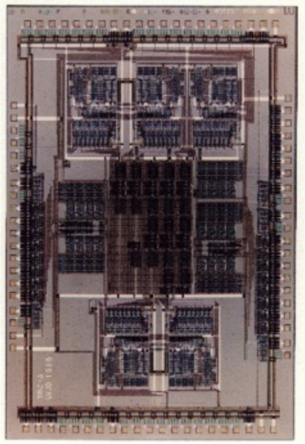
- Desiderata – Exascale network requirements
  - Efficient fine-grain communication, cost effective bandwidth, resilience
- Topology – Engineering to optimize available technology
- Routing
- Flow control and congestion avoidance
- Error control
- Ordering
- The role of photonics
- System sketch

Some History

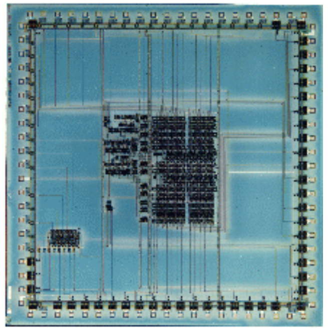
MARS Router
1984



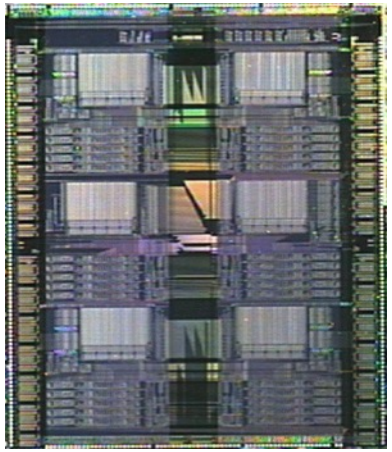
Torus Routing Chip
1985



Network Design Frame
1988

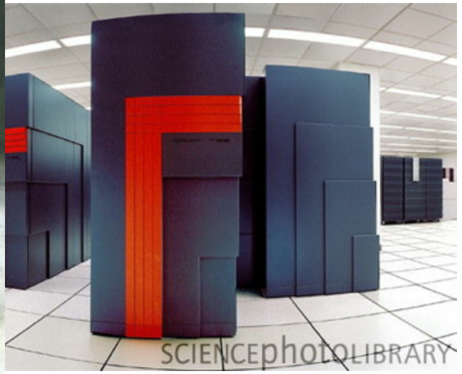


Reliable Router
1994



J-Machine
1992

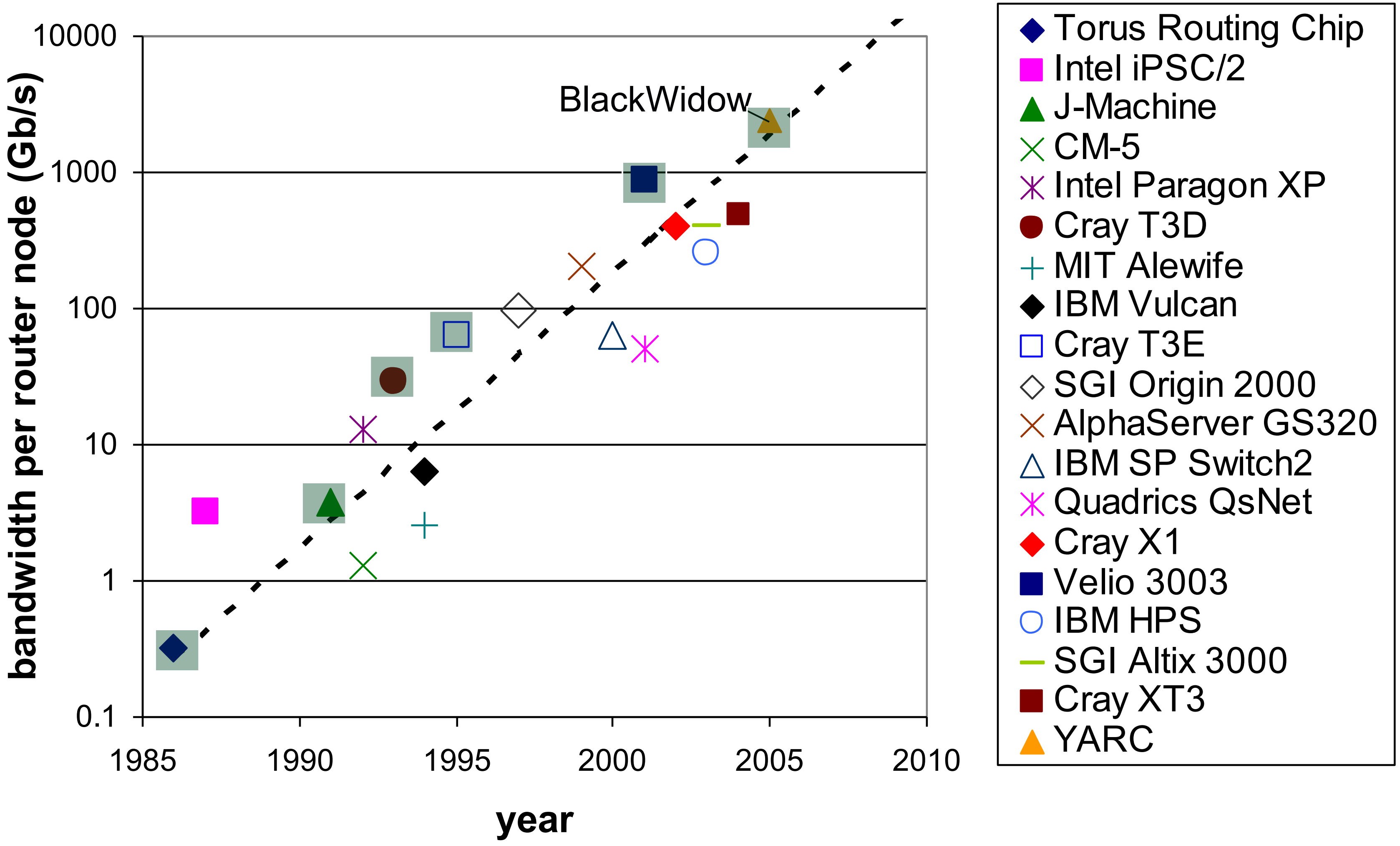
Cray T3D
1992



Cray T3E 1995

Cray Black Widow
2006

Trend Line



bandwidth per router node (Gb/s)
10000
1000
100
10
1
0.1
1985
1990
1995
2000
2005
2010
year
BlackWidow
Torus Routing Chip
Intel iPSC/2
J-Machine
CM-5
Intel Paragon XP
Cray T3D
MIT Alewife
IBM Vulcan
Cray T3E
SGI Origin 2000
AlphaServer GS320
IBM SP Switch2
Quadrics QsNet
Cray X1
Velio 3003
IBM HPS
SGI Altix 3000
Cray XT3
YARC

NVIDIA DGX-1

WORLD'S FIRST DEEP LEARNING SUPERCOMPUTER

170 TFLOPS

8x Tesla P100 16GB

NVLink Hybrid Cube Mesh

Optimized Deep Learning Software

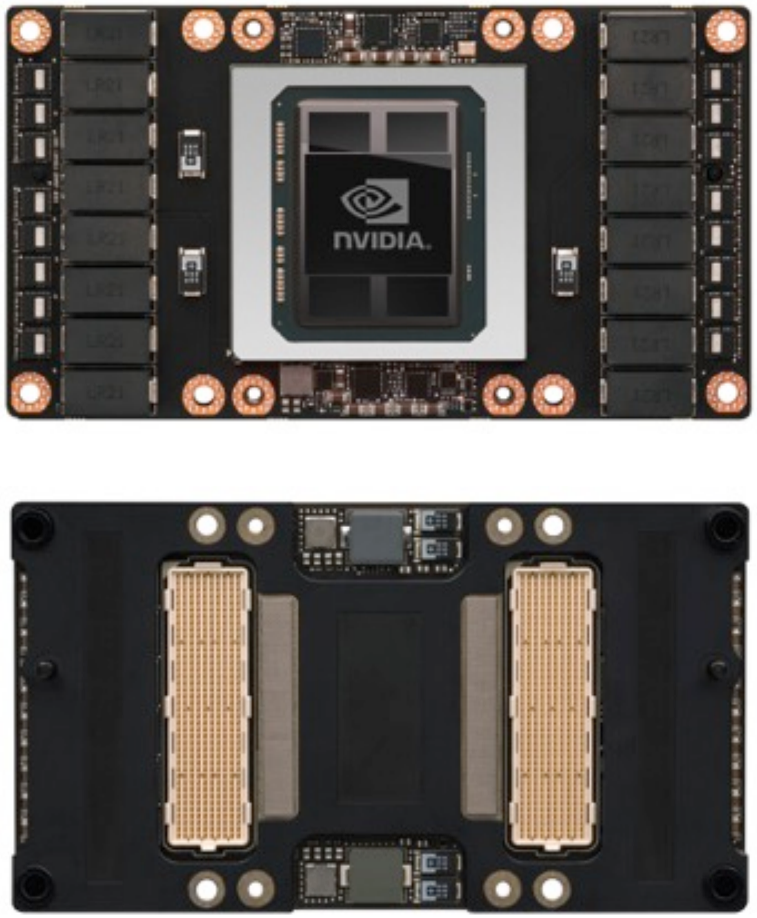
Dual Xeon

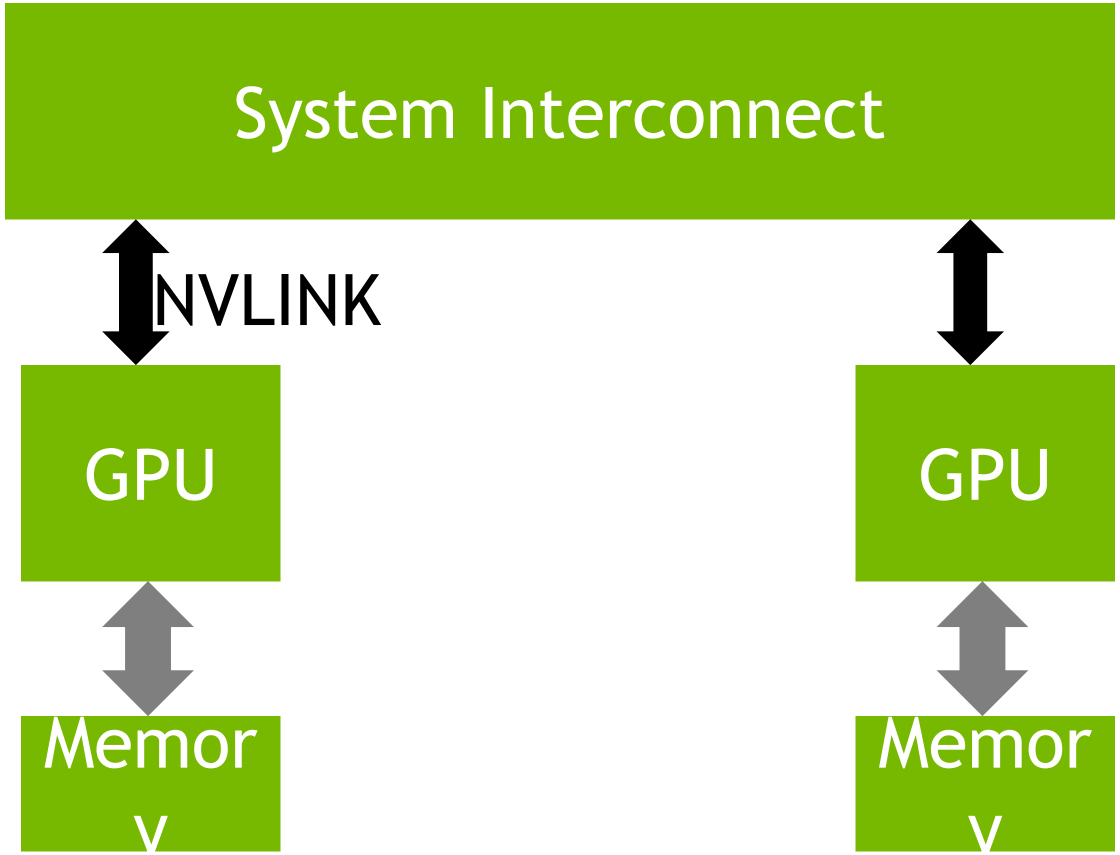
7 TB SSD Deep Learning Cache

Dual 10GbE, Quad IB 100Gb

3RU – 3200W

NVLINK - Enables Fast Interconnect, PGAS Memory





DGX SATURNV

World's Most Efficient AI Supercomputer

TOP 500 The List.

Fastest AI Supercomputer in TOP500

4.9 Petaflops Peak FP64 Performance
19.6 Petaflops DL FP16 Performance
124 NVIDIA DGX-1 Server Nodes

THE GREEN 500

Most Energy Efficient Supercomputer

#1 on Green500 List
9.5 GFLOPS per Watt
2x More Efficient than Xeon Phi System

FACTOIDS

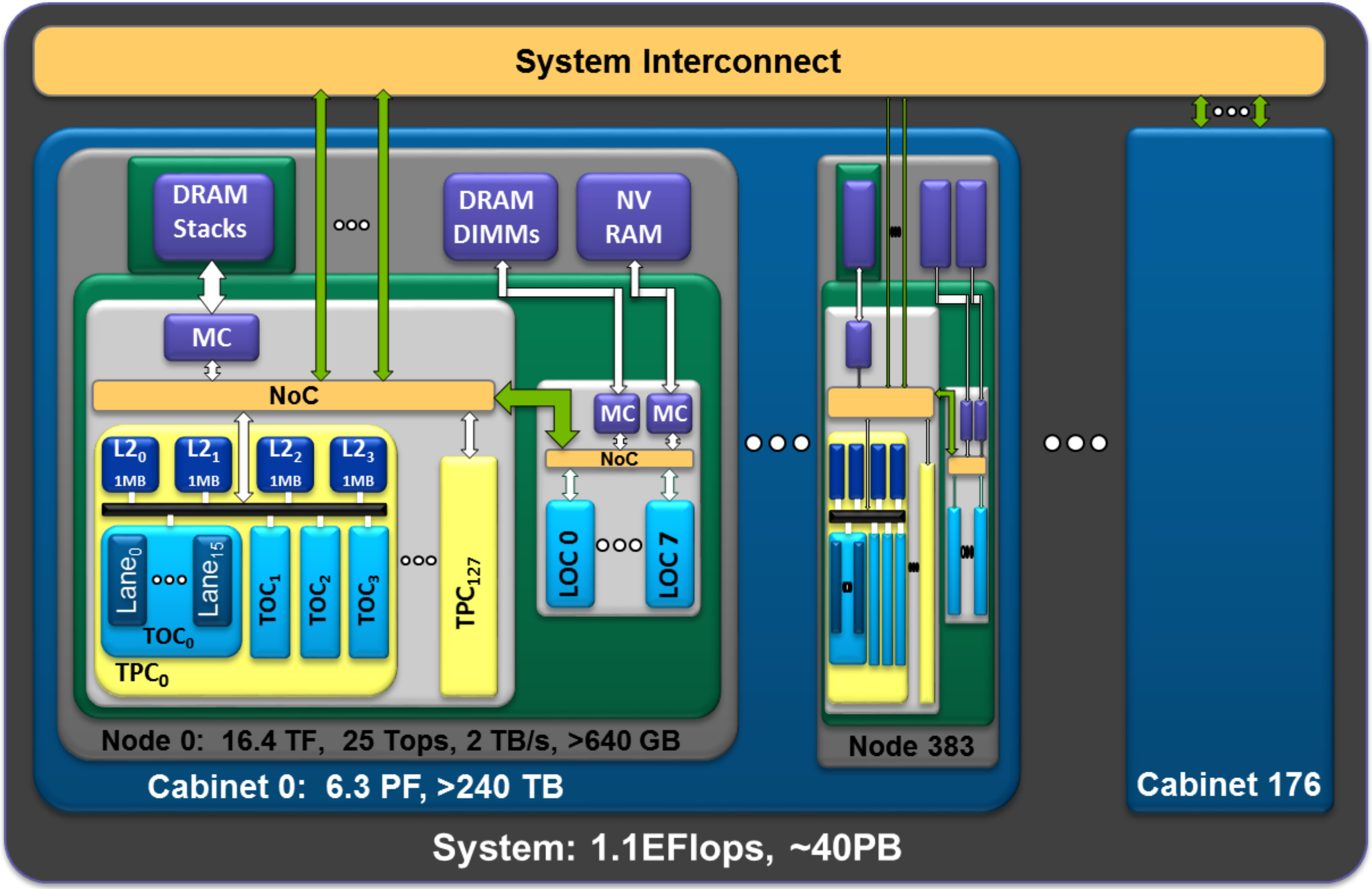
13 DGX-1 Servers in Top500

38 DGX-1 Servers for Petascale supercomputer

55x less servers, 12x less power vs CPU-only supercomputer of similar performance

Desiderata

Exascale System Sketch

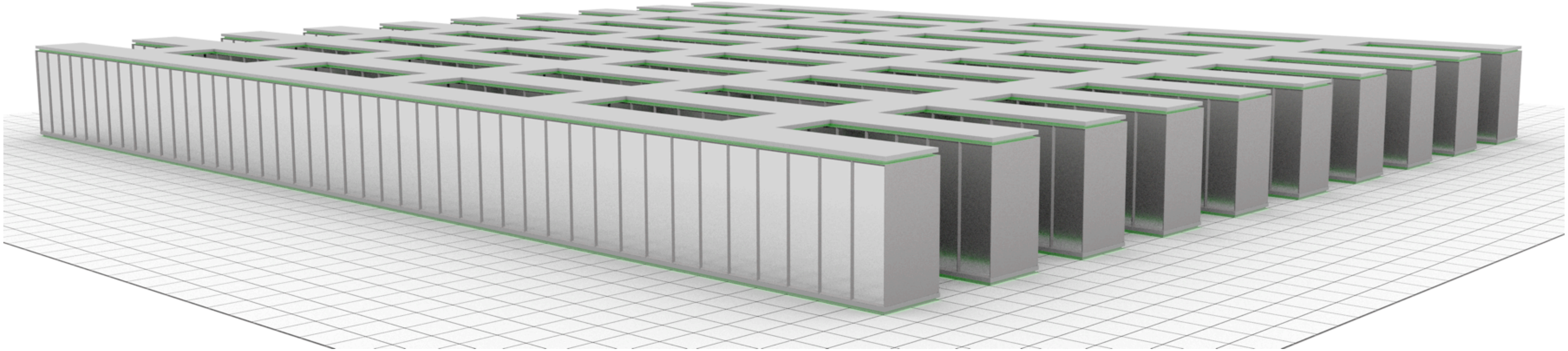


Desiderata

- Scale – 10^5 endpoints
- Cost-effective bandwidth (injection and bisection) – B/s$
- Low latency (dominated by time-of-flight)
- Reliable exactly-once delivery – (BER < 10^{-21})
- Low overhead (latency and occupancy)

- Enable strong scaling
- Low-overhead shared-memory operations
- Highly concurrent operation

Scale – ~10^5 Powerful Endpoints

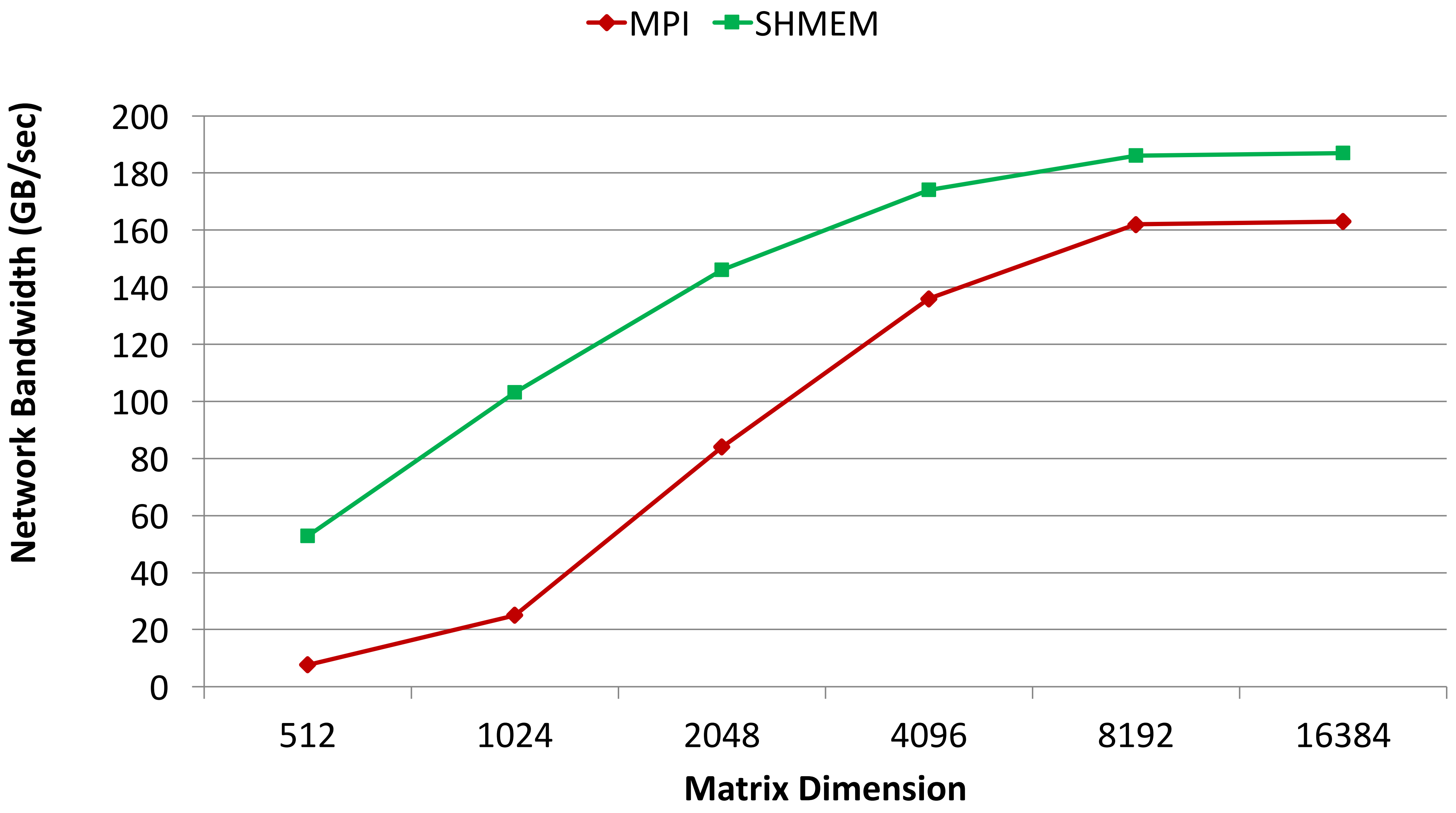


- Each endpoint is a 16.4TFLOPS (DP) GPU with 4TB/s memory bandwidth
- 400GB/s injection bandwidth is 10:1 local memory to neighbor memory
  - Pascal GP100 today is ~5:1 (750GB/s memory: 160GB/s NVLINK)

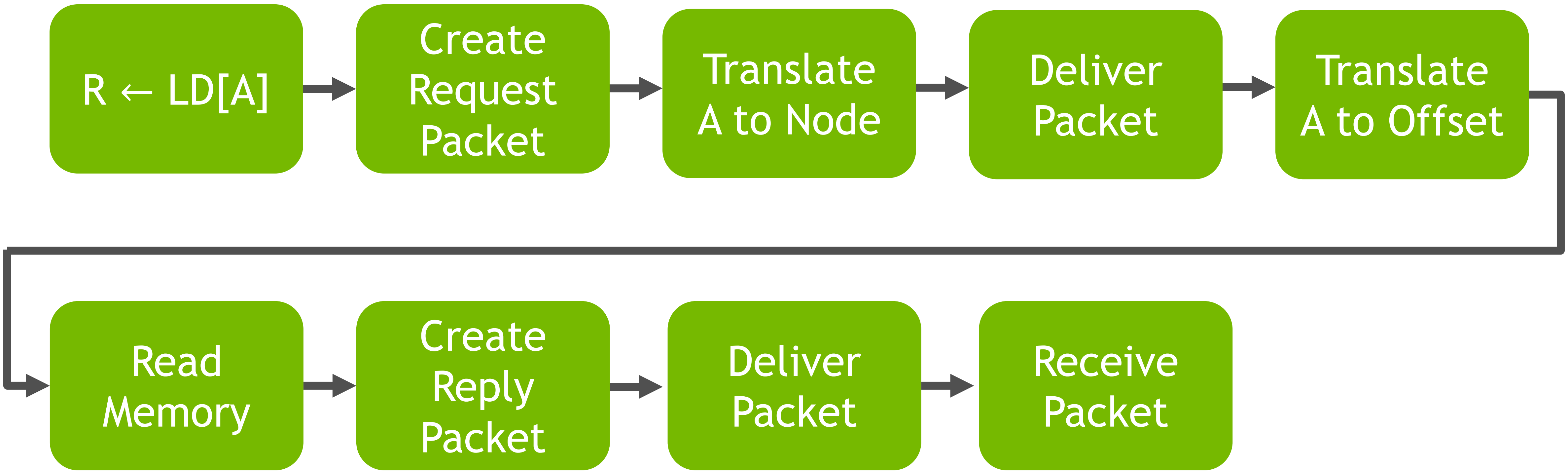
Cost-Efficient Bandwidth (B/s$)

- Network cost dominated by links (AOCs)
- Use minimum number of expensive links per route (1 dragonfly)
- Operate each link near capacity (flow-control and congestion avoidance)
- Make payload a high fraction of bits on the wire
  - For small payloads (16B) as well as large

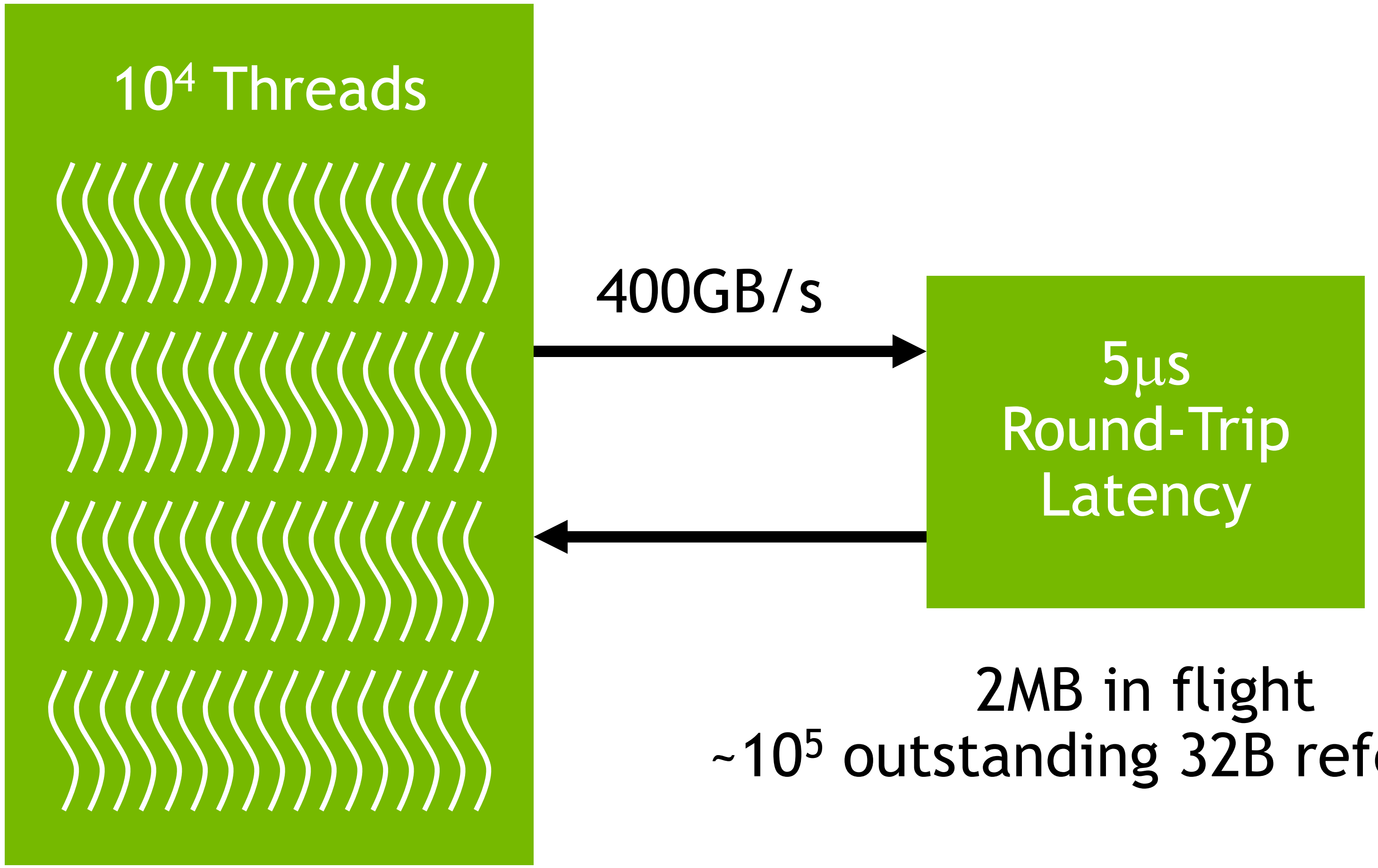
The Need for PGAS



Remote Load/Store



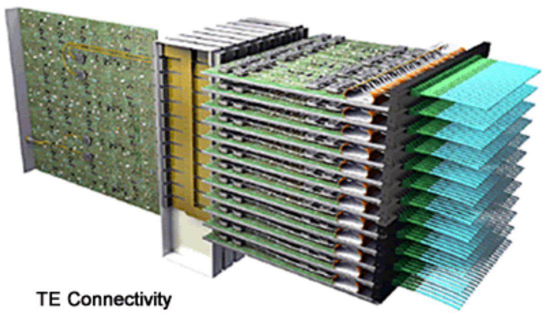
~10^5 Outstanding References per Endpoint



2MB in flight
~10^5 outstanding 32B references

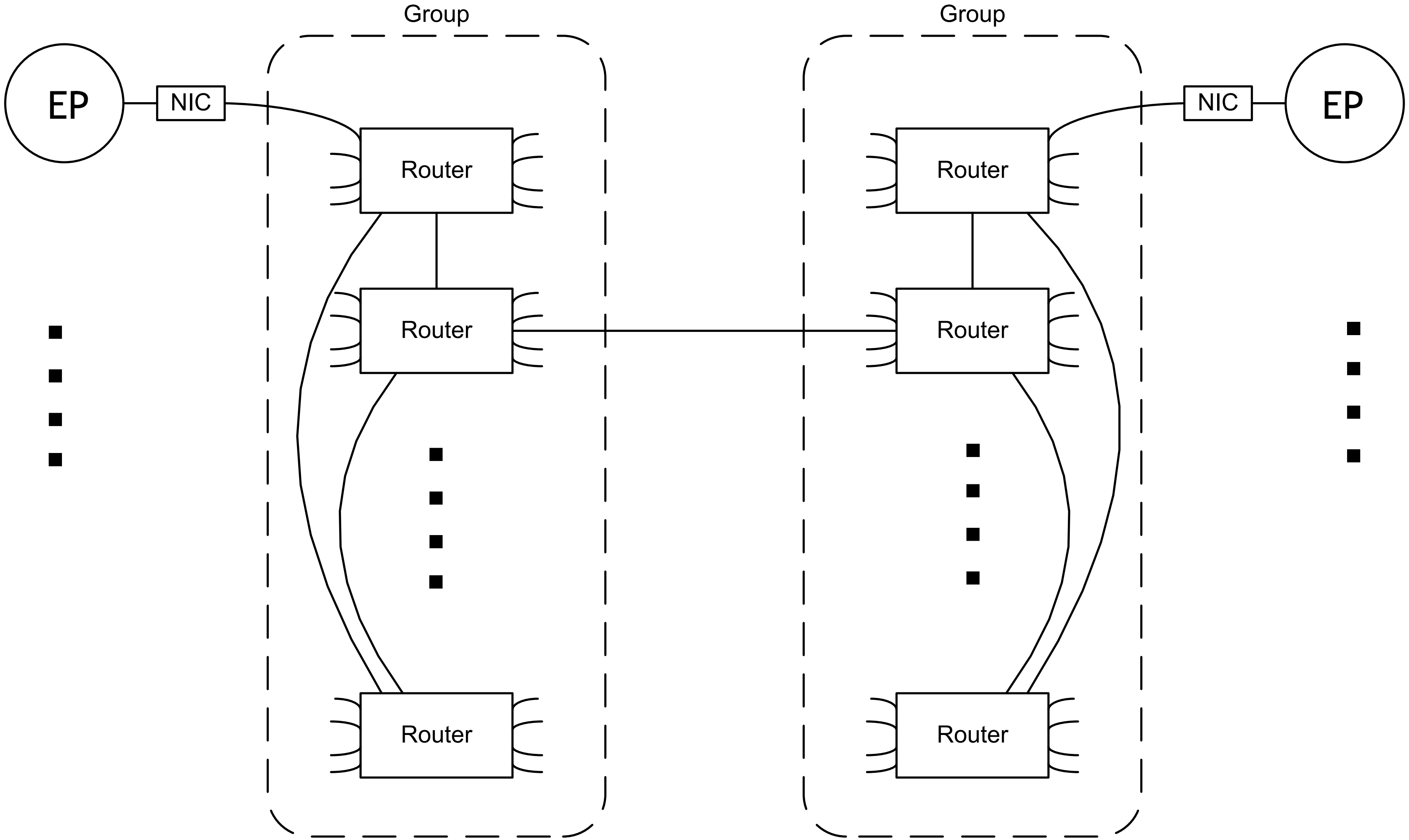
Topology

Cost of 100Gb/s



$500	$50	$10	$5
100m	5m	1m	0.3m

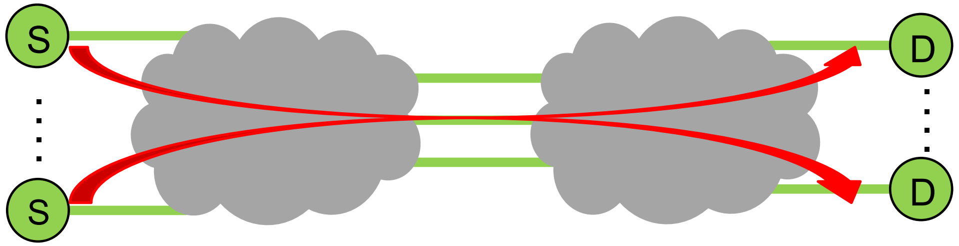
Dragonfly Topology



Kim, J., et al. "Technology-Driven, Highly-Scalable Dragonfly Topology." *ISCA 2008*

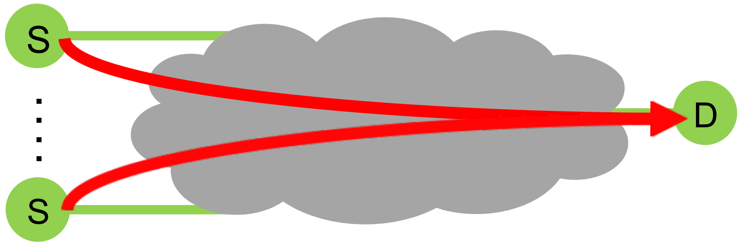
Adaptive Routing

Sources of Congestion



Fabric congestion

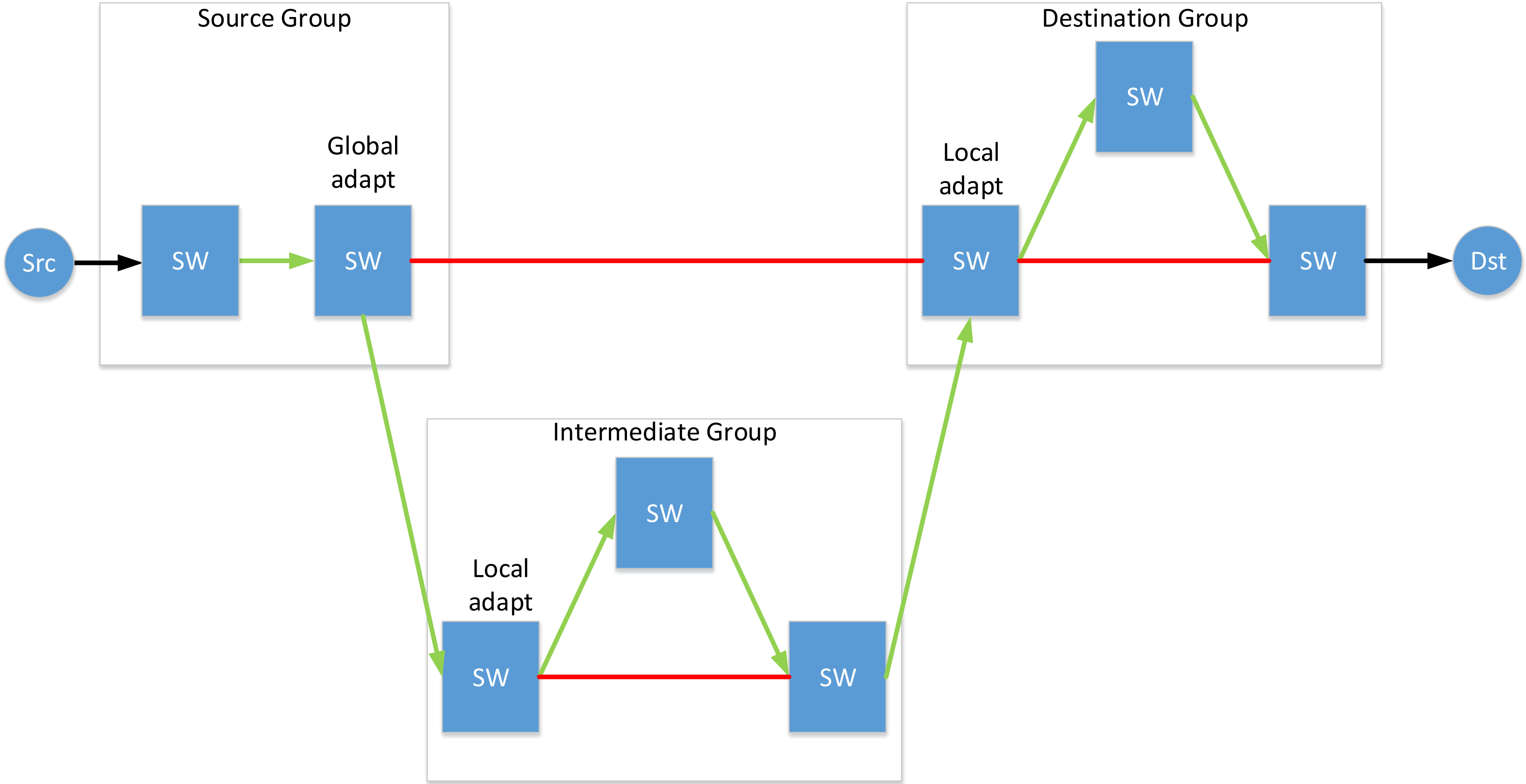
- Cause: low bisection bandwidth or load imbalance
- Solution: add bandwidth, improve load-balance using adaptive routing



Endpoint congestion

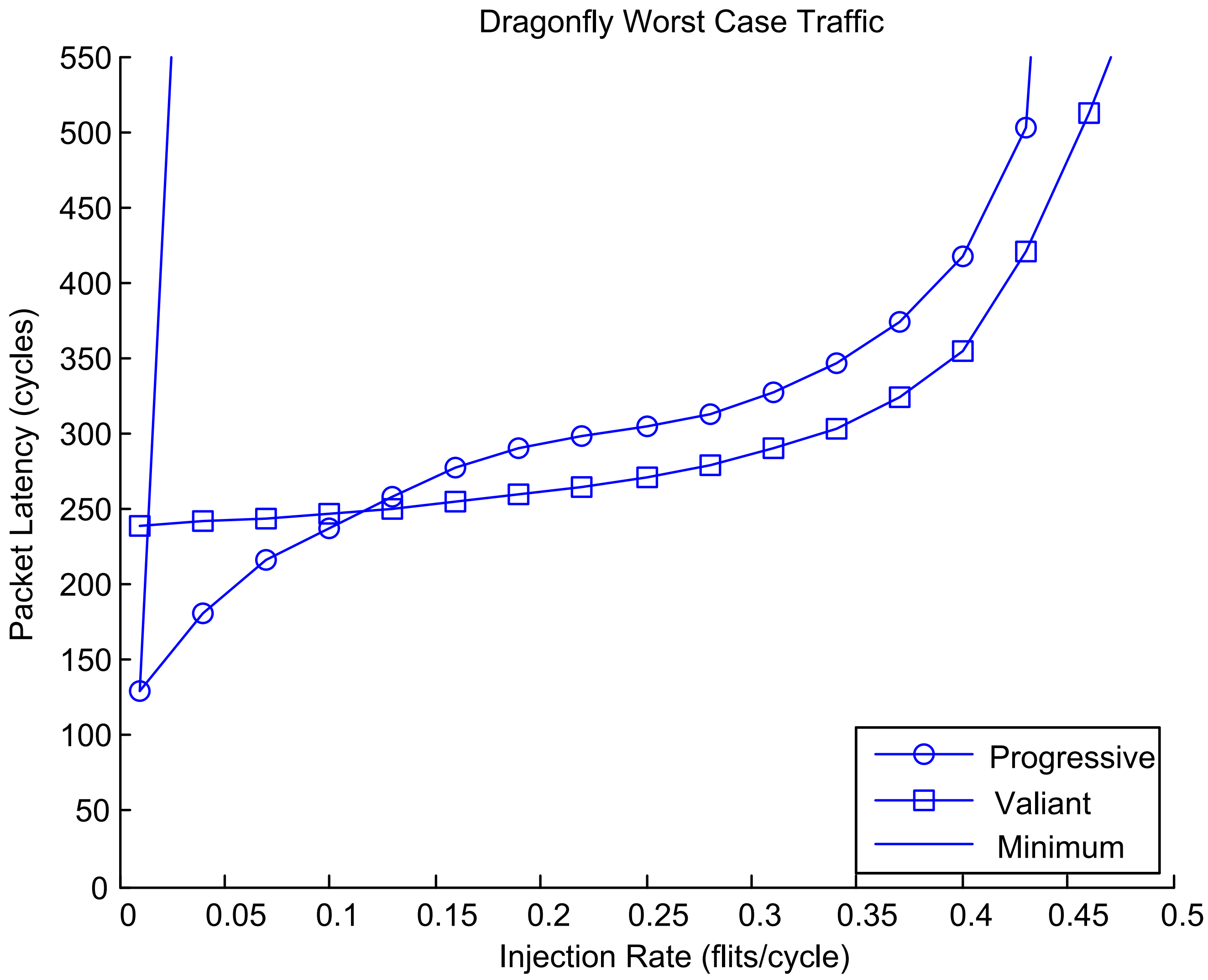
- Cause: endpoint bandwidth over-subscription
- Solution: reduce bandwidth demand by throttling traffic sources

Progressive Adaptive Routing with Local misroute

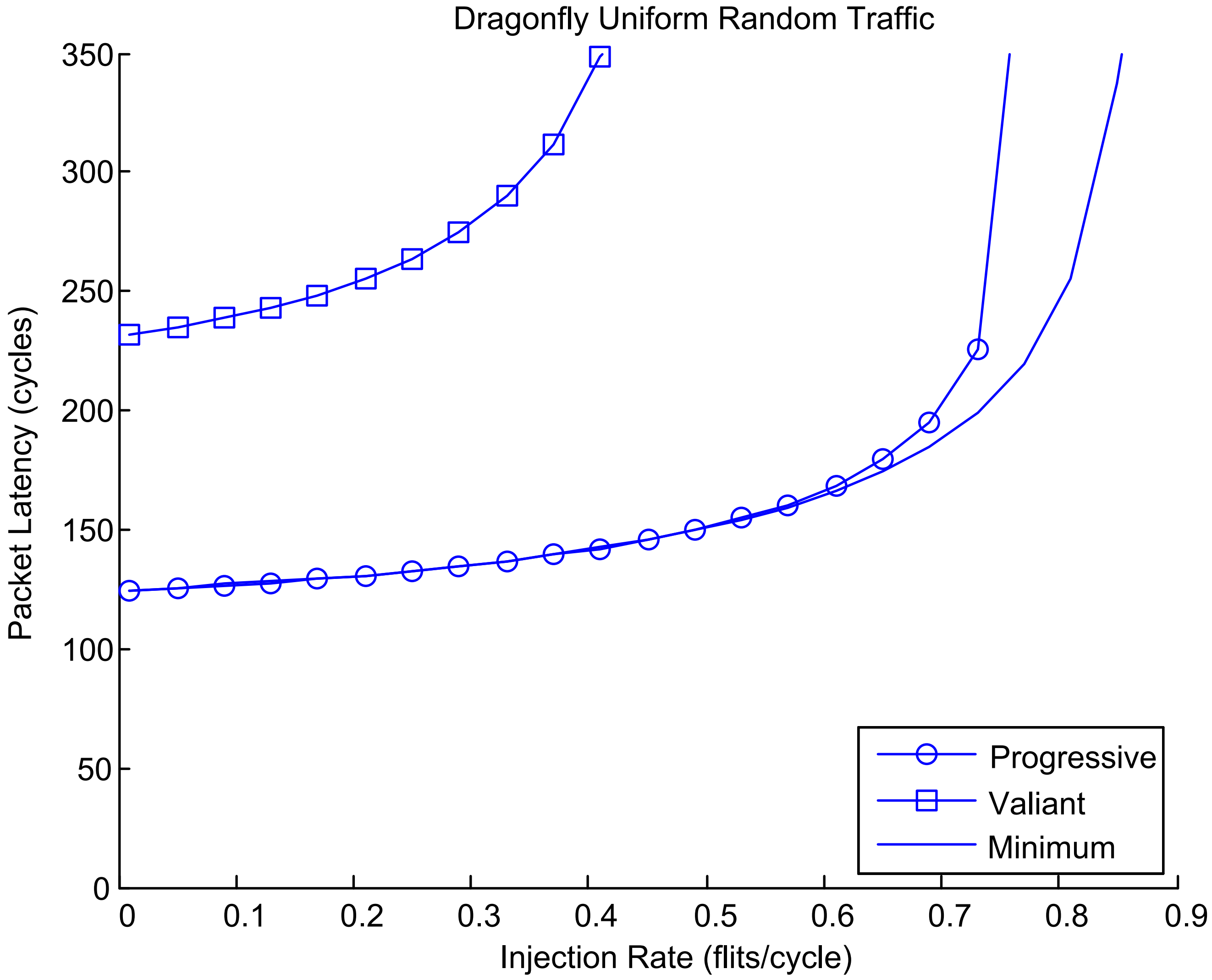


Singh, A., 2005. *Load-balanced routing in interconnection networks* (Doctoral dissertation, Stanford University).

Jiang, Nan, William J. Dally, and John Kim. "Indirect Adaptive Routing on Large Scale Interconnection Networks," *ISCA 2009*



Dragonfly Worst Case Traffic
Packet Latency (cycles)
Injection Rate (flits/cycle)
0
50
100
150
200
250
300
350
400
450
500
550
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Progressive
Valiant
Minimum


Dragonfly Uniform Random Traffic
Packet Latency (cycles)
350
300
250
200
150
100
50
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Injection Rate (flits/cycle)
Progressive
Valiant
Minimum

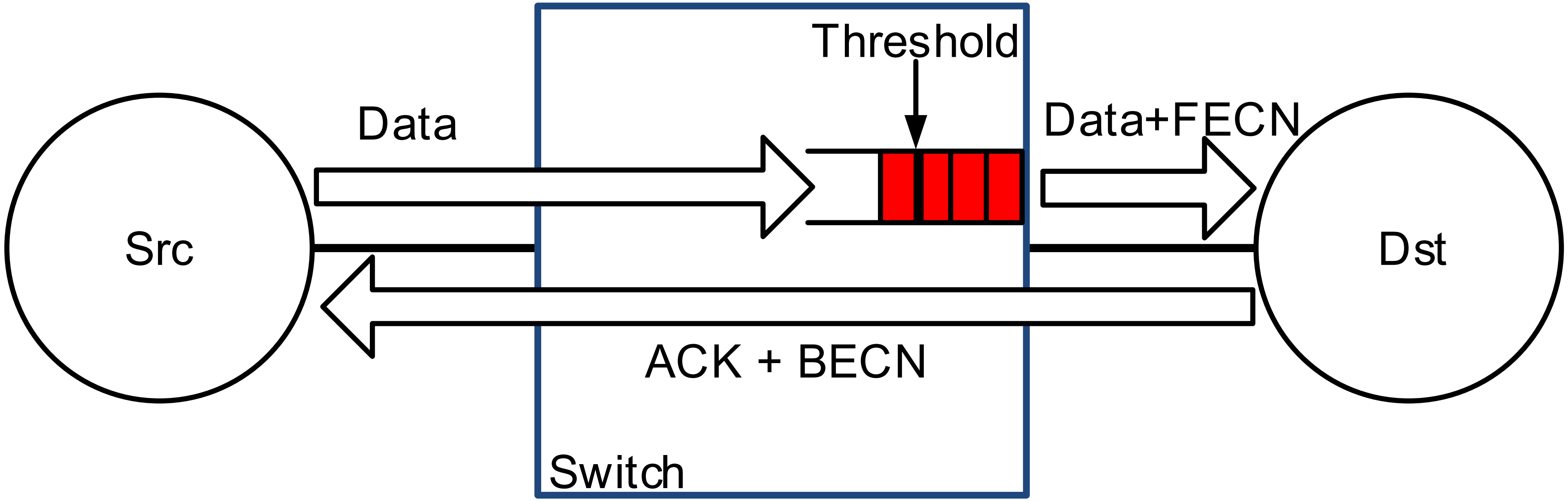
Congestion Avoidance

LHRP

Congestion Notification

150GB/s * 3us RTT = 450 KB inflight before first notification

Reaction is too late, slow response time, large transient



Last-Hop Reservation Protocol

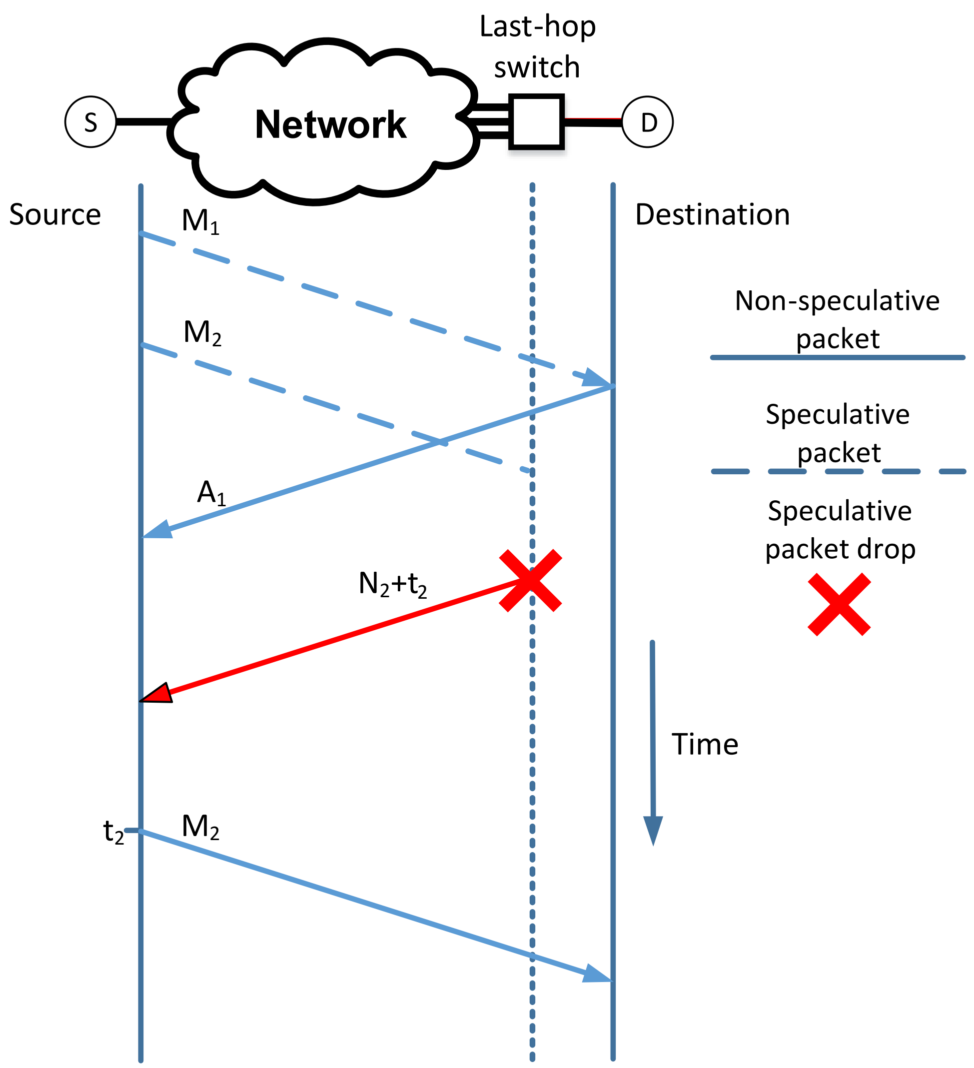
Observation: preserve ejection channel bandwidth for data packets

Move the endpoint reservation scheduler to the last-hop switch

Messages are first transmitted speculatively

No congestion: speculative messages will arrive successfully

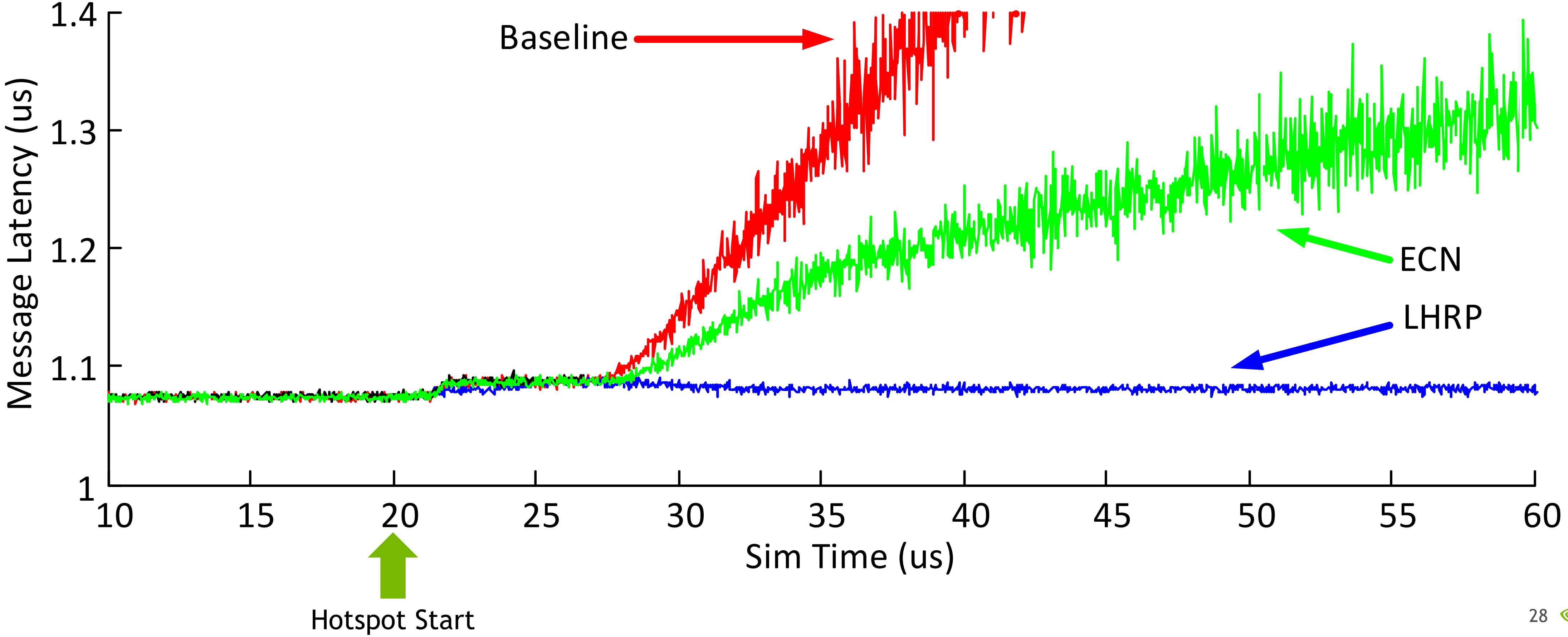
With congestion: speculative message is dropped by the last-hop switch, reservation is sent back with the nack



Jiang, Nan, Larry Dennison, and William J. Dally. "Network endpoint congestion control for fine-grained communication," *SC*, 2015.

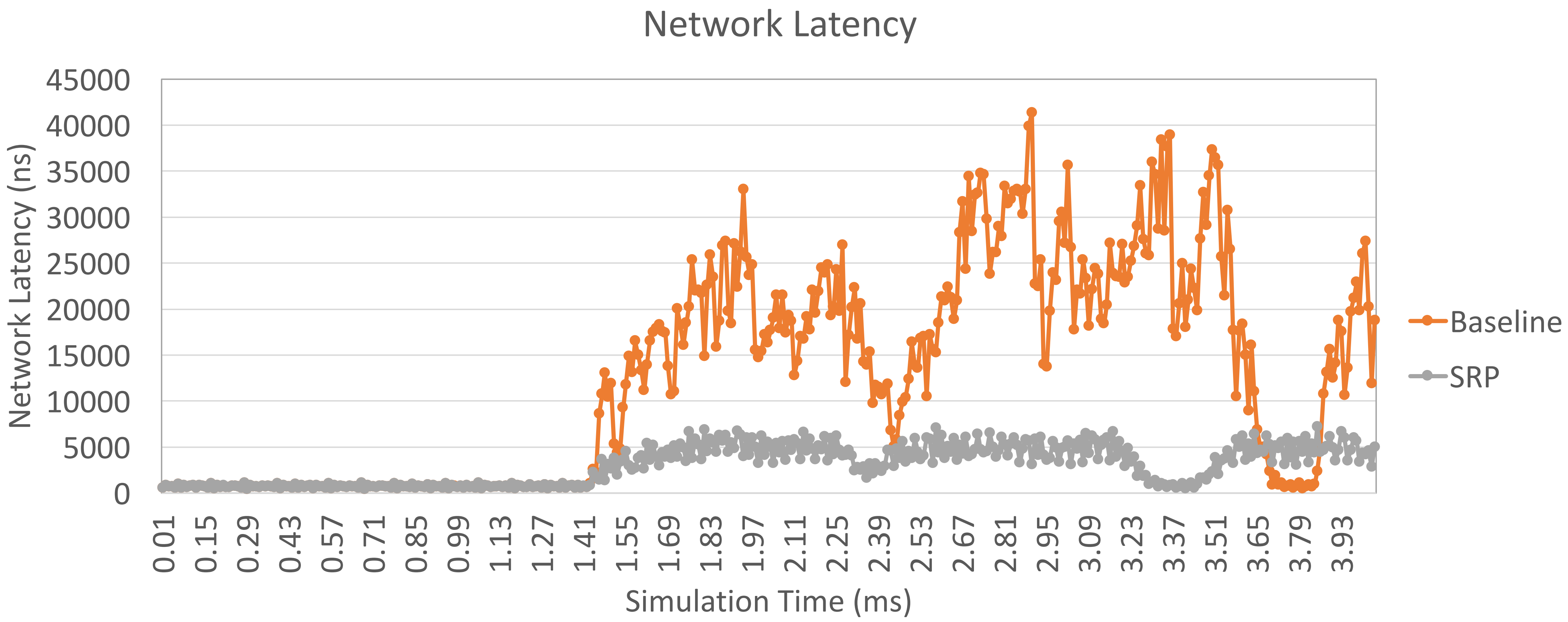
Initial Congestion Response

40% uniform random + 60:4 Hotspot @ 20us – 4 flits/message



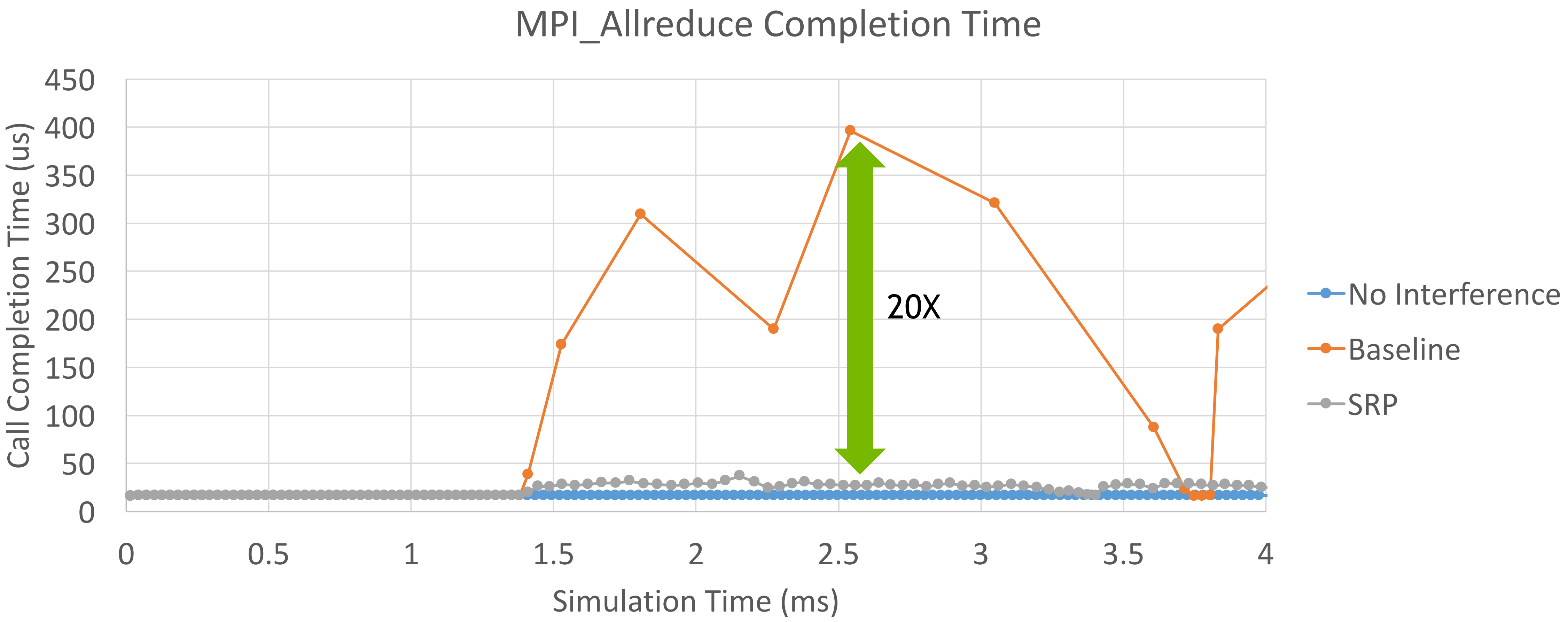
Impact of Congestion on Network Performance

1024-rank BIGFFT MPI trace on 2.5K-node dragonfly



Impact of Congestion Interference on Network Operations

1024-rank MPI_Allreduce + 1024-rank BIGFFT on 2.5K-node dragonfly



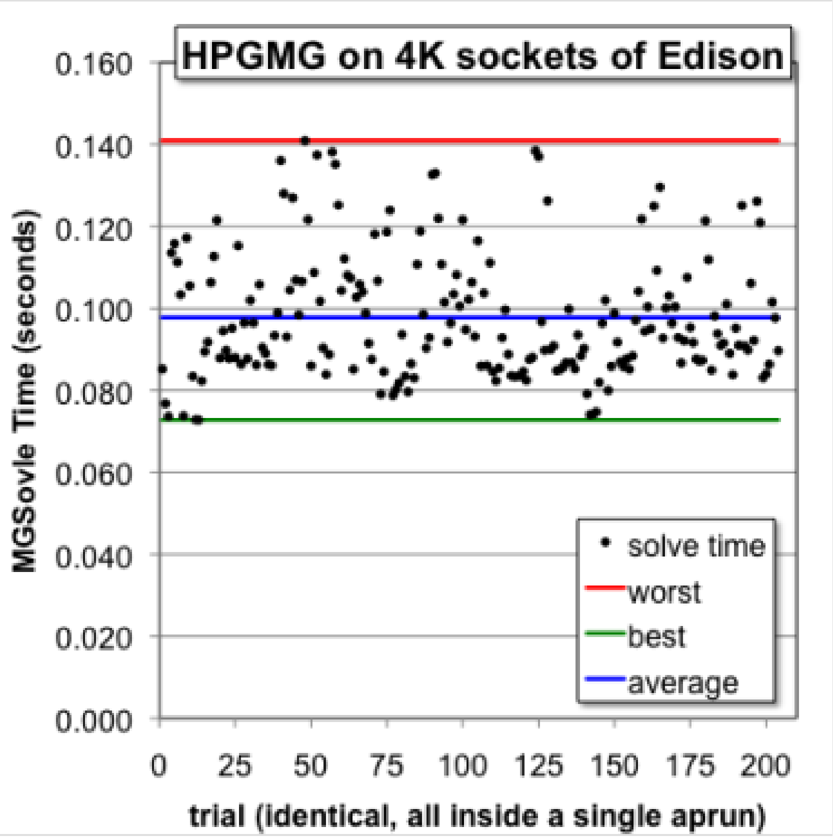
Interference, An Open Problem

Interference

Performance Variability

F U T U R E T E C H N O L O G I E S G R O U P

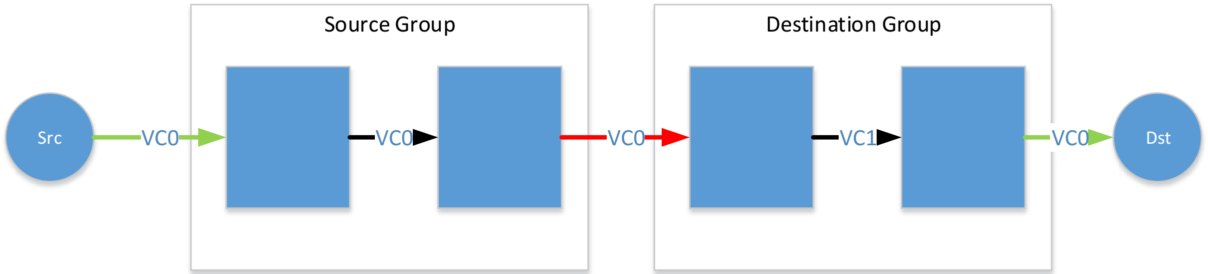
- Performance is highly variable at scale on Edison (Dragonfly)
- Actively investigating the cause with NERSC/Cray.
- Figure to the right shows the individual solves times obtained when solving
  - the exact same problem
  - always using 4K sockets (and 32K cores)
  - with the same decomposition
  - within a single aprun (while loop in single execution)
- **Average performance is 33% lower than best**

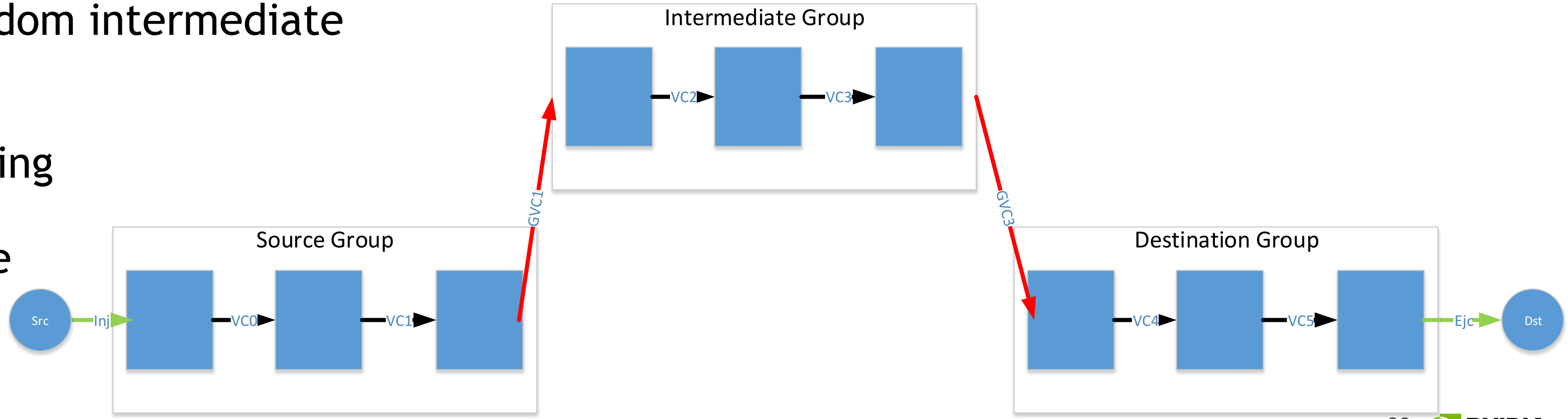


Sam Williams
Designforward Tech Talk July 2014

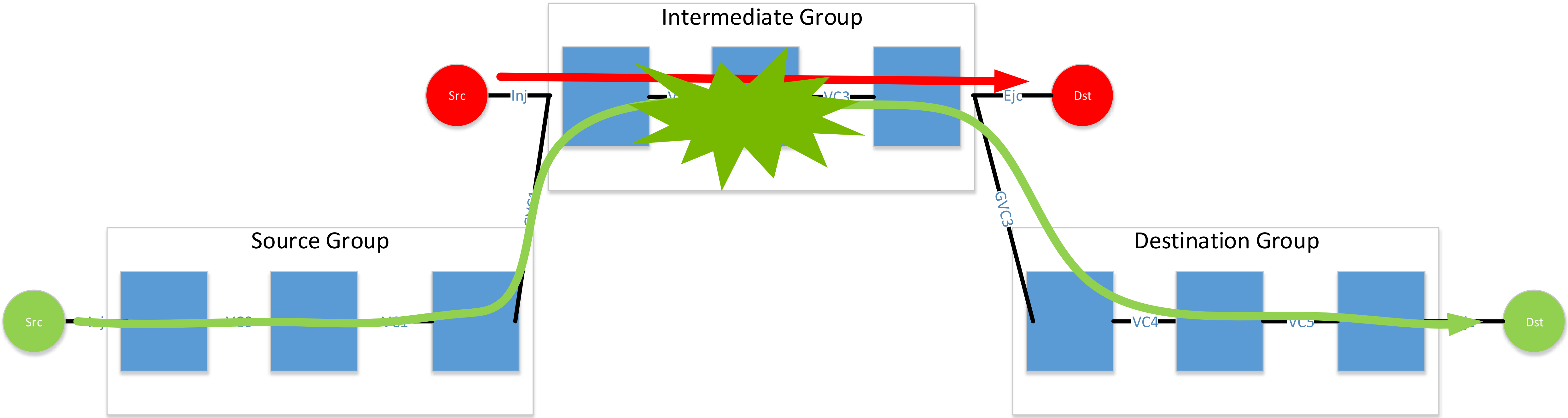
ADAPTIVE ROUTING

- Minimal routing is insufficient for all traffic patterns
  - Two switches shares a single local channel
  - Two groups shares a single global channel
- Utilize non-minimal network paths
  - “Bounce” off of a random intermediate switch/group
  - Creates resource sharing
  - Source of interference



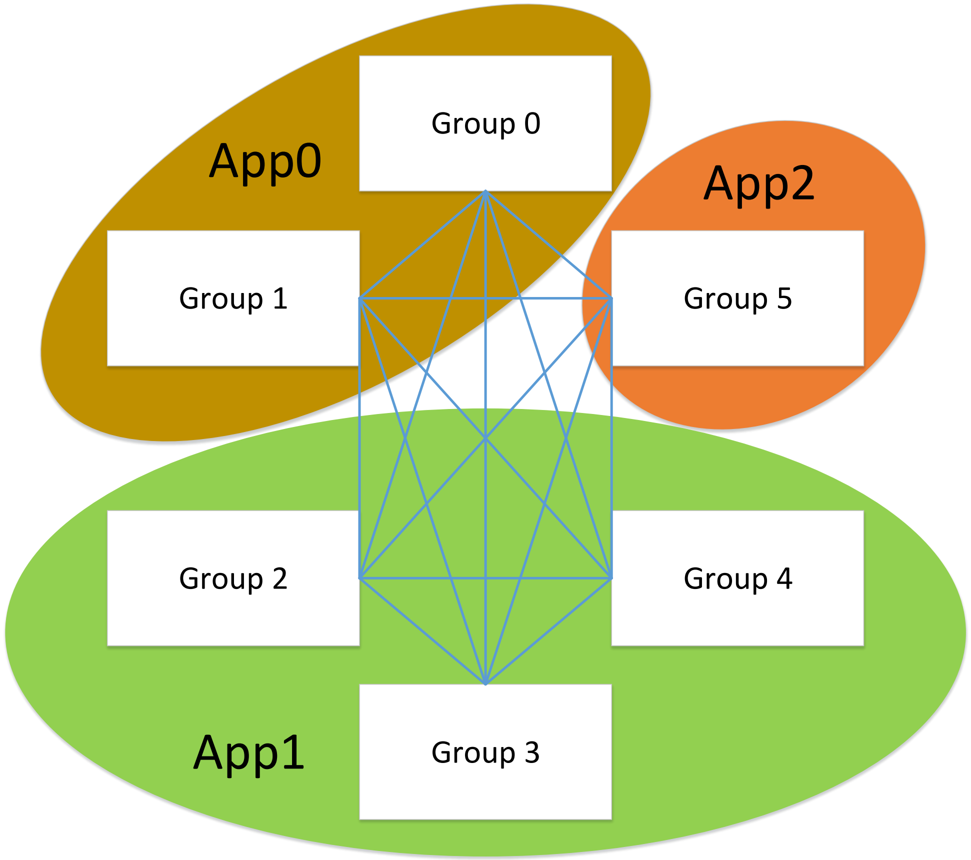


Adaptive Routing Interference



Bandwidth Partitioning

- Multiple applications running on a network each in a partition
- Each partition
  - Monitor the fraction of intra-partition adaptive traffic on the global links
  - Adjust the adaptive routing bias to maintain 50% fraction
  - Simpler alternative to physically partitioning network
- Partitions are not perfectly isolated
  - Subject to transient traffic variations
  - Adaptive routing reaction time

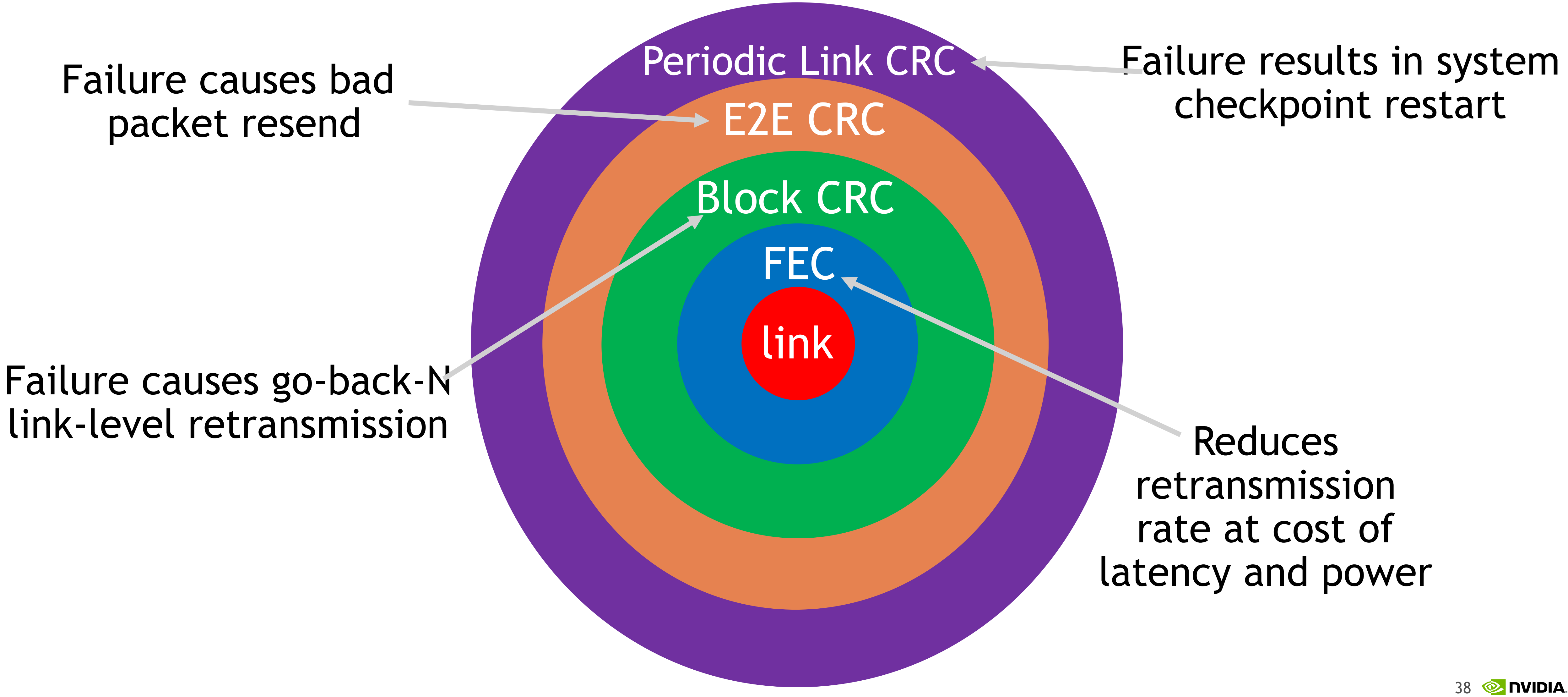


Error Control

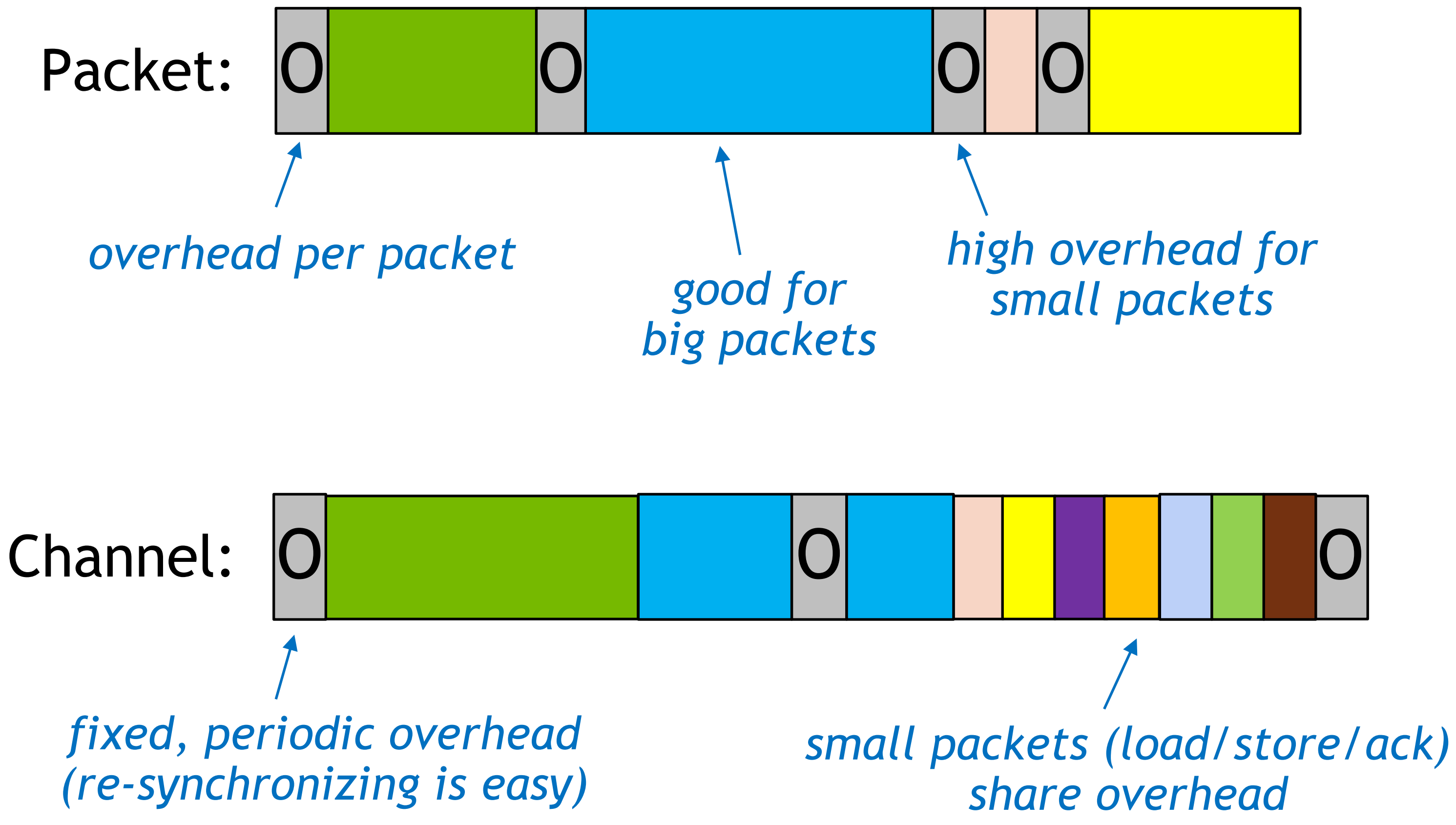
Error Control Problem

- ~10^6 links in an Exascale system
- Bandwidth of 2 x 10^{11} b/s each (total bandwidth of ~10^{17} b/s)
- Bit error rate of 10^{-4} to 10^{-12} (total error rate of 10^5-10^{13} errors/s)
- System wide error rate of 10^{-4} errors/s (1 week MTBF)
- Spec network error rate at 10^{-5} errors/s (order of magnitude less)
- Requires bit error rate of 10^{-22}

Layers of hardware protection



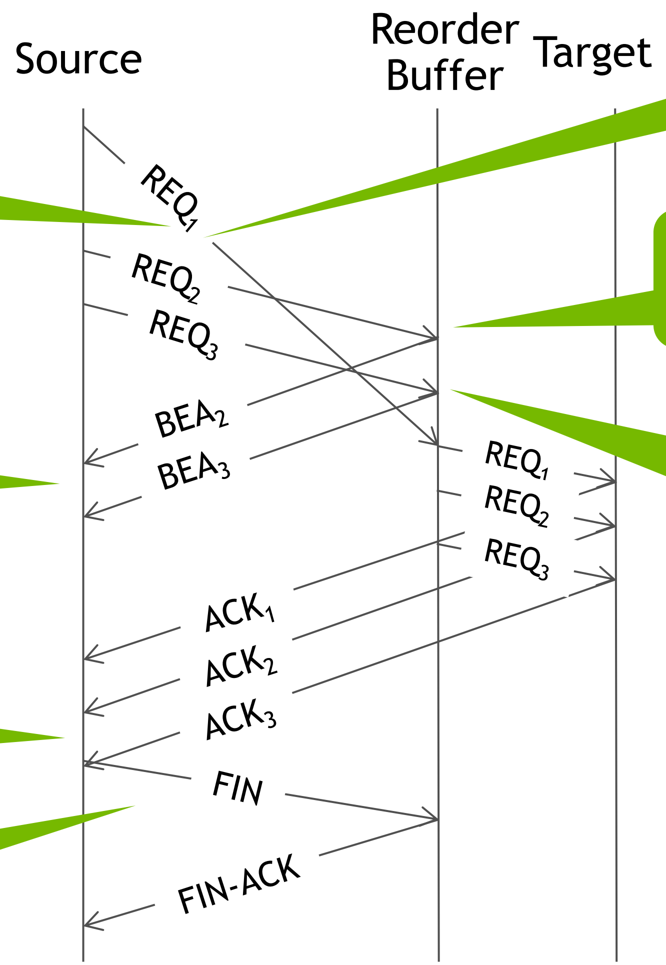
Channel vs packet protocol



Ordering

Ordered Transfer Protocol

Lightweight Connection with Low-Overhead Setup and Teardown



Sequence numbers order packets

Sender sends burst of REQs (using a connection)

First REQ opens connection

Periodic beacon packets suppress retransmissions

Out-of-order REQs are held in re-order buffer until they can be forwarded in order

ACKs acknowledge delivery of REQs to target device

Connection is closed when no more waiting REQs

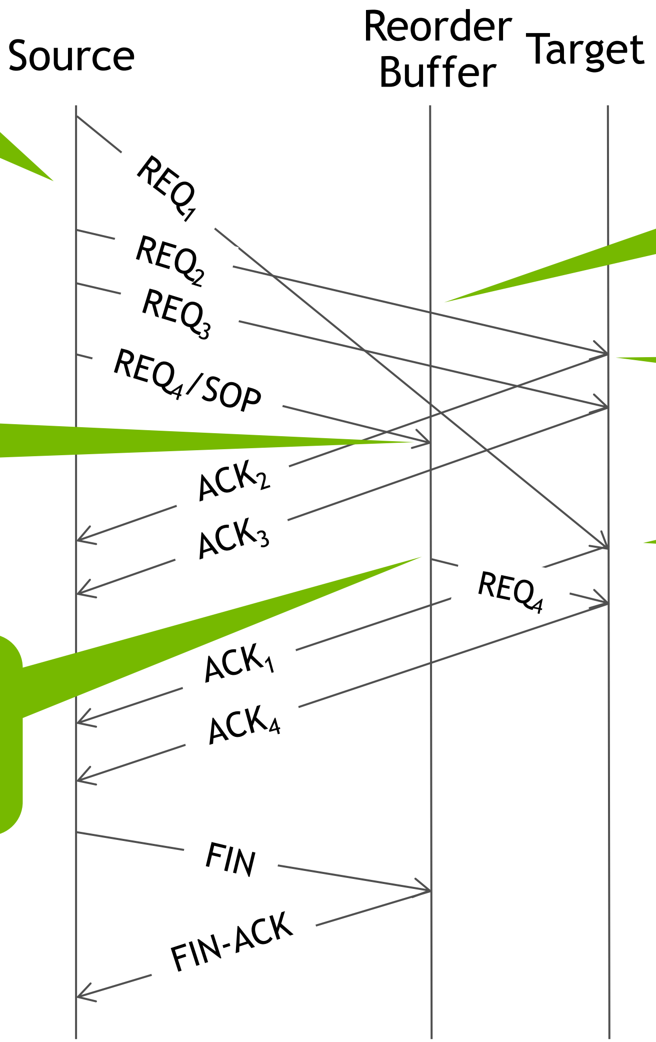
Synchronized Bulk Transfer

Strict Ordering is not Required for many Producer/Consumer Exchanges

$REQ_{1..3}$: Bulk data
REQ_4: Synchronization operation

1-element reorder buffer holds back synchronization operation

Minimal delay between completion of bulk data transfer and delivery of synchronization operation



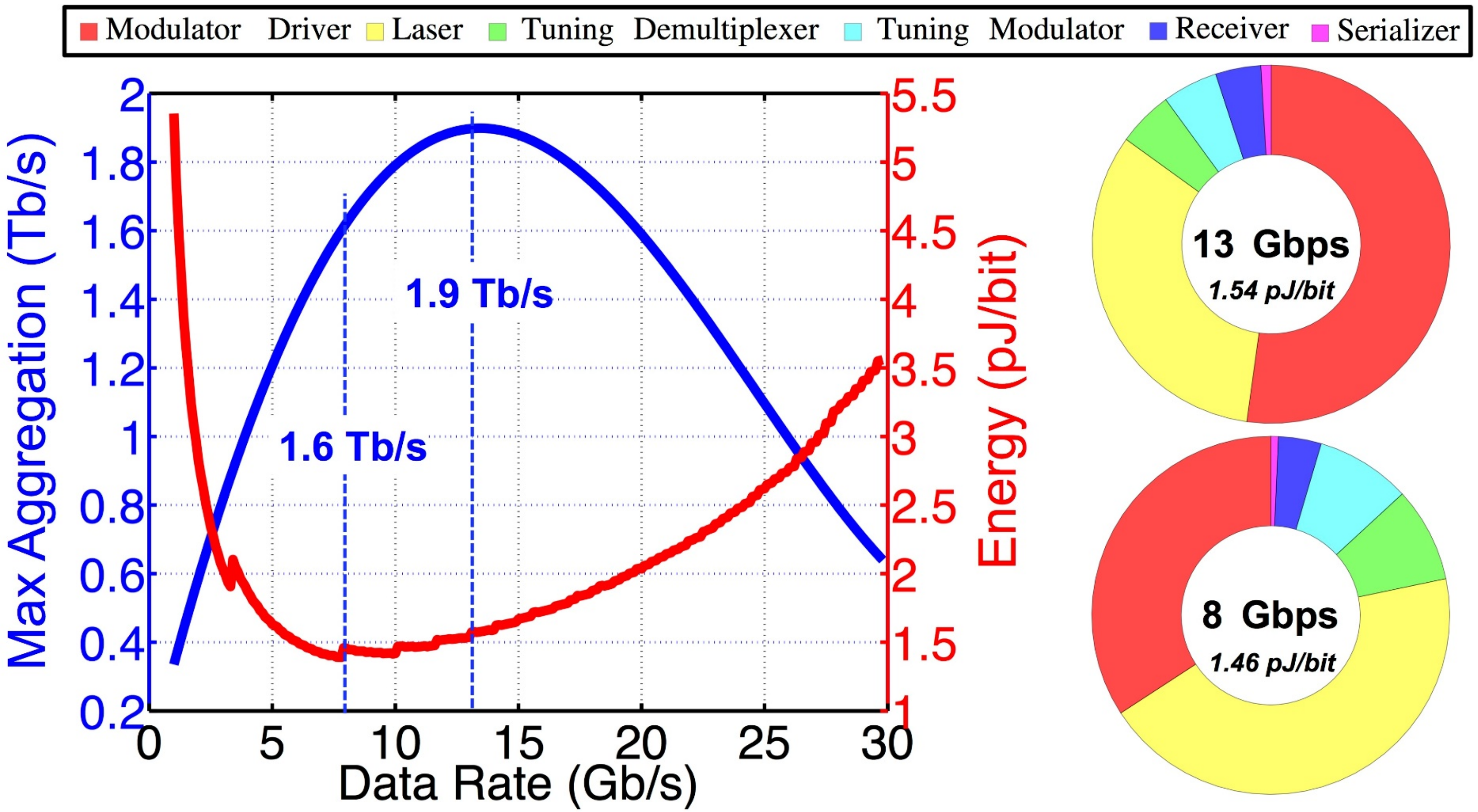
First arriving REQ opens connection (allocates counter for ACKs)

Bulk data REQs are forwarded to target (possibly out of order)

Counter determines completion of bulk transfer

The Role of Photonics

Power/Bandwidth Density

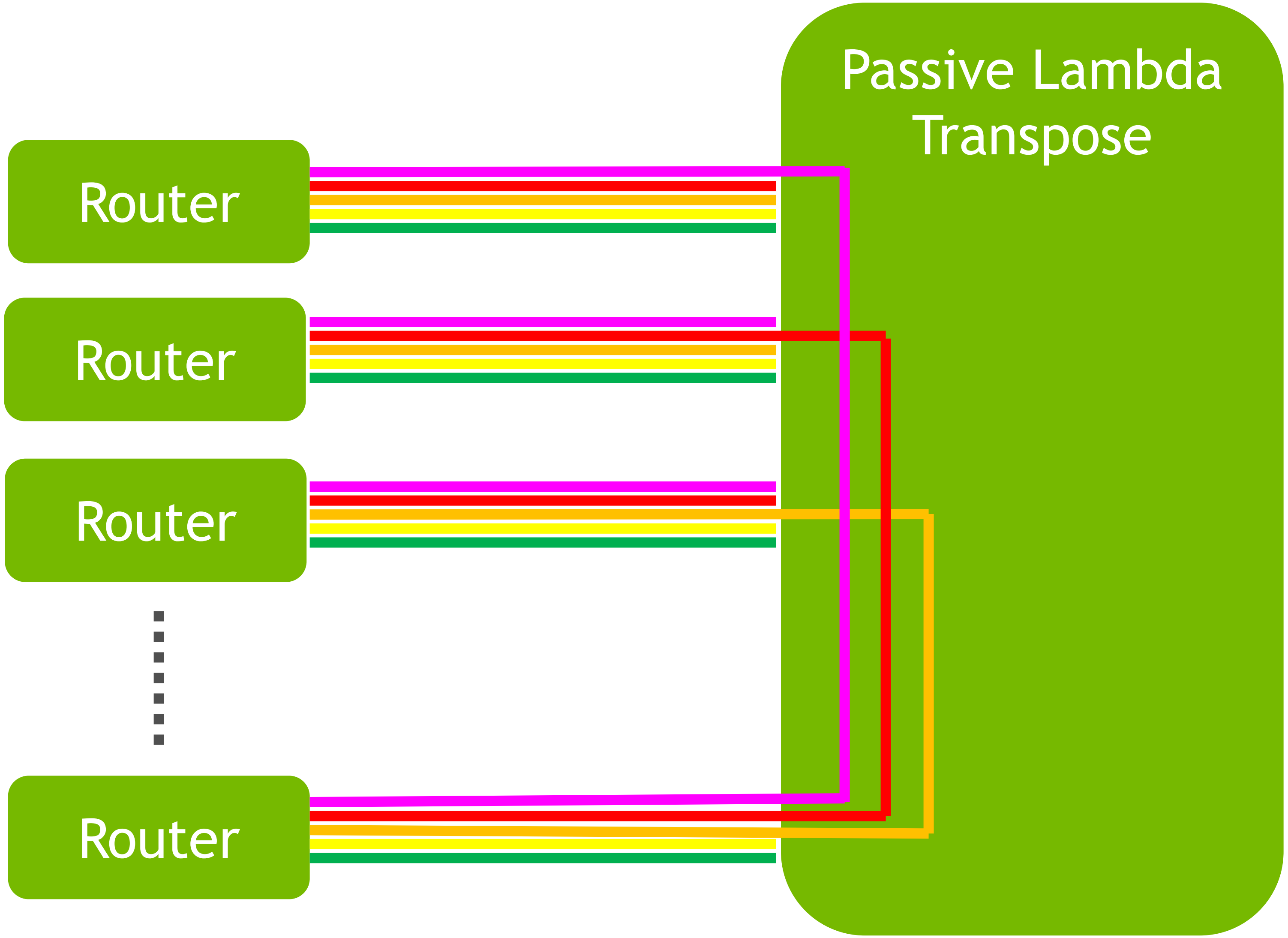


200 lambdas @ 8Gb/s

146 lambdas @ 13Gb/s

M. Bahadori, R. Polster, S. Rumley, Y. Thonnart, J.-L. Gonzalez-Jimenez, K. Bergman, "Energy-Bandwidth Design Exploration of Silicon Photonic Interconnects in 65nm CMOS," IEEE Optical Interconnects Conference (OI) (May 2016)

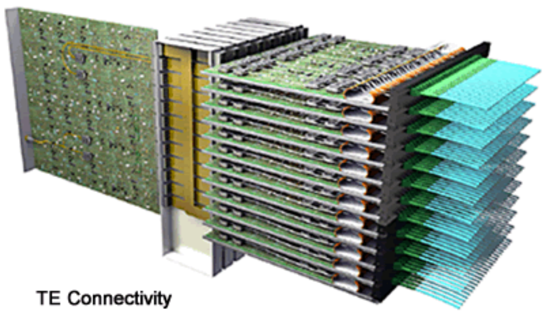
Photonic Dragonfly Concept



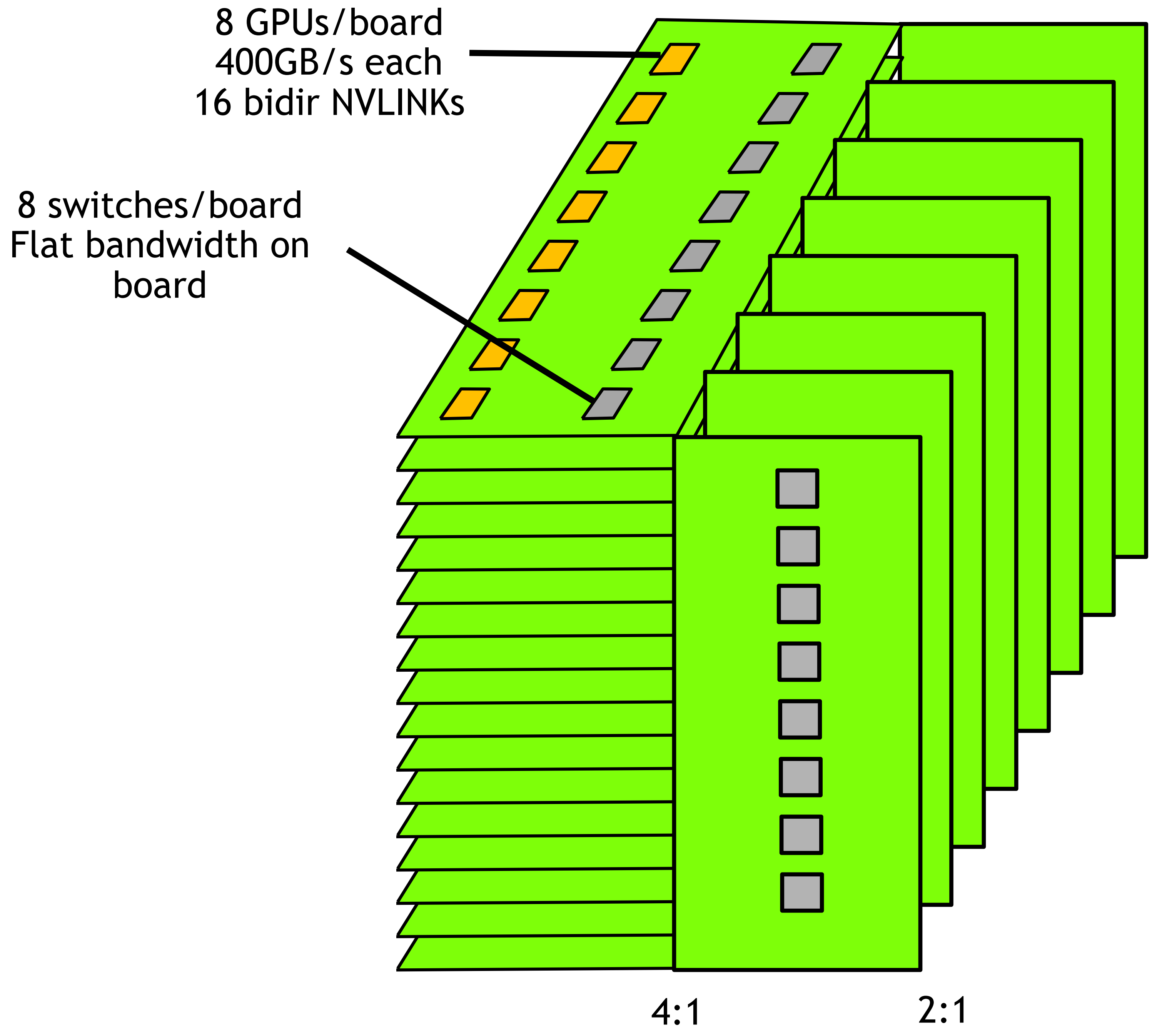
- 256 Groups
- 16 Fiber Bundles per group
- 16n wavelengths per fiber
- 16 Central AWGRs
- Much simpler cable management
- Technology not sufficiently mature for Exascale (2021)
  - Maybe by 2025

Overall System Sketch

Recall Costs

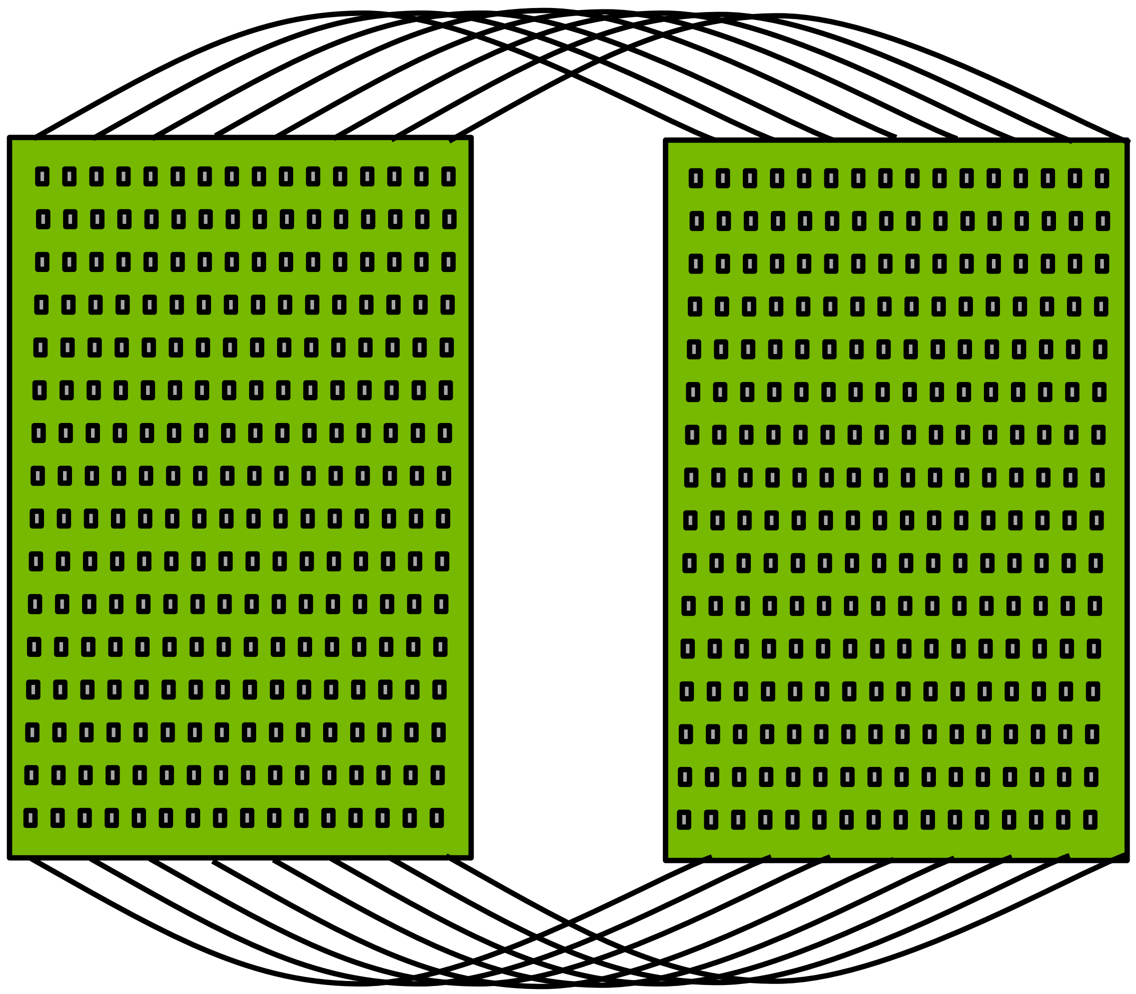


$500	$50	$10	$5
100m	5m	1m	0.3m



Cabinet

- 128 GPUs, 50kW
- 16 GPU x 8 switch boards
- 400GB/s bidir per GPU
  - 3.2TB/s per board
- 192 switches
  - 8 per board
- Flat bandwidth on board
- 100GB/s at crosspoints
  - 32 pairs
  - 12.8TB/s aggregate
- 100GB/s per GPU within cabinet
- All connections electrical



Group

- 1 or 2 cabinets
- Electrical Flex cables between cabinets
- 12.8TB/s between cabinets
  - 512 NVLinks
  - 4096 pairs
  - 64 cables, 64 pairs each
- 256 NVLinks out back of each cabinet – 2 per GPU
  - 50GB/s per GPU global bandwidth
  - 6.4TB/s per cabinet
- Up to 513 groups
  - 131,328 GPUs

System Sketch

- Cost dominated by AOCs – 50GB/s per GPU $2K per endpoint
- Taper by leaving half the cables out – 25GB/s global bandwidth per GPU
    - Limits maximum system size to 64k GPUs
- Routing – progressive adaptive routing with local misroute (6VCs per class)
- Flow control – flit-level flow control with LHRP
- Error control
    - Channel-level CRC for link
    - Packet-level CRC for ETE
    - FEC for optical cables only (where needed)
- Ordering – per packet or bulk sync
- PGAS support, with two-stage address translation

Conclusion

Conclusion

- An **Exascale** network is not business as usual
- Need fine-grain communication (**PGAS**) for strong scaling
  - **Two-stage** address translation
  - **10^5 outstanding references** per endpoint
  - **Ordering** as needed
- Cost-efficient bandwidth
  - **Topology** driven by communication cost – Dragonfly
  - High **payload efficiency** for small packets (32B)
  - **Congestion avoidance** & **adaptive routing** allows links to operate near capacity
- Error control
  - **Channel-level**, not packet-level CRC
- System sketch
  - 8:2:1 – Board:Cabinet:Global bandwidth taper
  - Cost is dominated by AOCs

Backup - Not in main Talk

[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]