Zurich Research

IBM

Data Center Interconnects – trends for IB and Ethernet

Ronald Luijten

IBM Zurich Research Lab, Switzerland

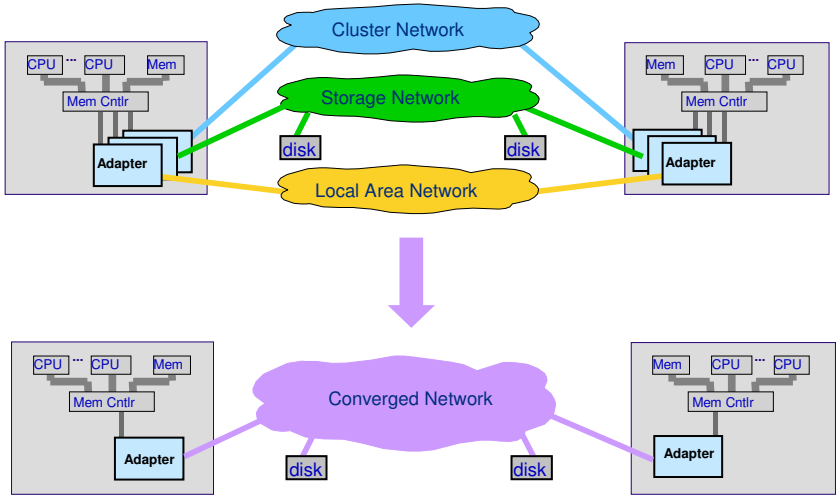
He who controls the network, controls the datacenter.

Data center trends

- **The highest server growth is in bladed servers**
  - IBM calls these servers ‘scale-out’ architecture
  - Yields attractive price/performance
  - Users (re-)write applications for this environment
- **10’000 servers / data center (not HPC!) at top internet companies today**
  - Ethernet is de-facto interconnect solution
- **Infiniband currently has significant cost advantage at 10Gbps**
  - Through bundling serial copper wires at right signaling rates
- **Expect 10GBE on motherboard / server blades soon (BC-H now)**
  - Multi core CPUs coming, additional network bandwidth/blade needed
  - New applications; XML document standard
- **Commodity based**

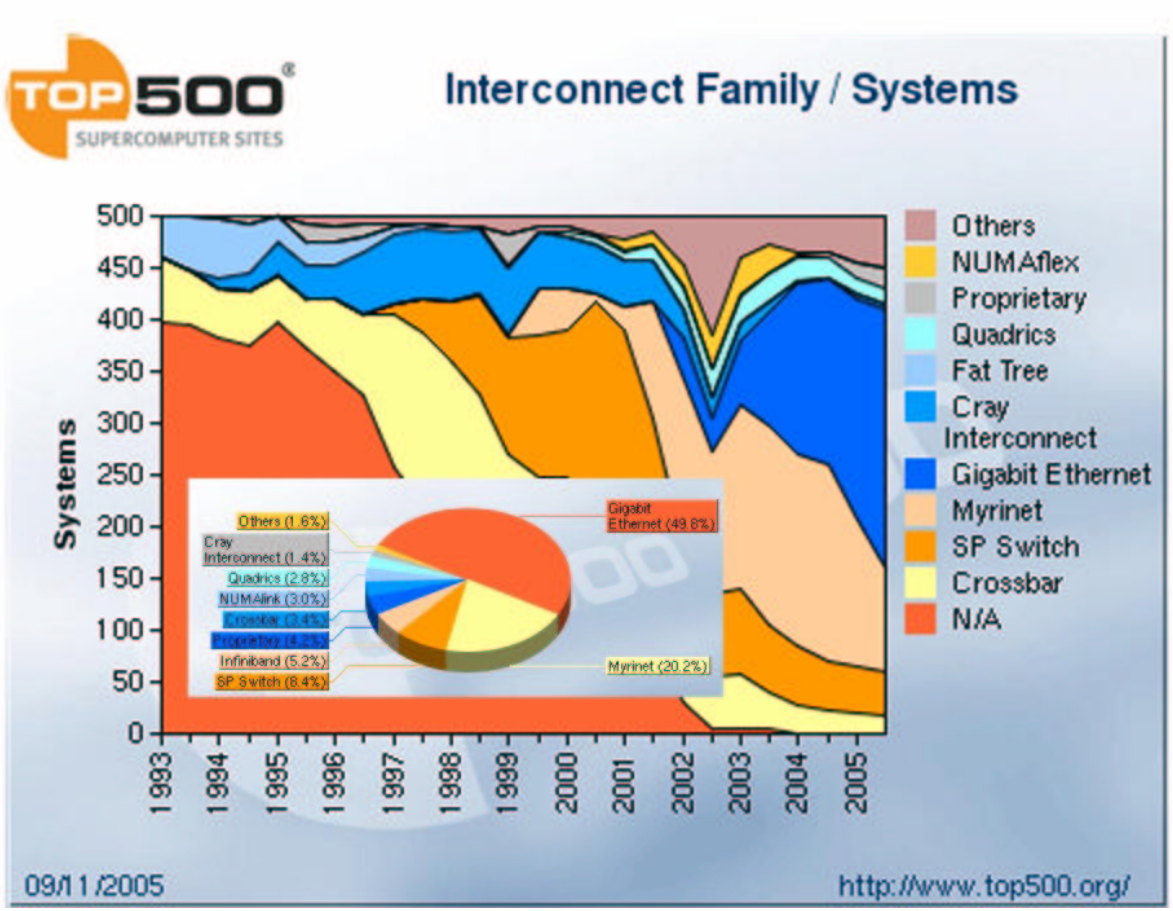
Holy grail: Data Center network convergence

- **This is the convergence of:**
  - Communication
  - Storage
  - Clustering



- **Technical capability will be available soon**
- **A proof point: CERN**
  - Running storage over Ethernet – cost driven / huge volumes
  - iSCSI, 10GBE backbone, 1GBE distribution (incl. QoS)
- **Non technical stumbleblocks:**
  - Storage and communication networks are owned by different organizations

Learning from observing HPC market

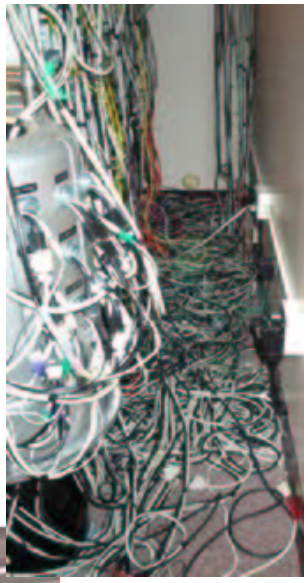


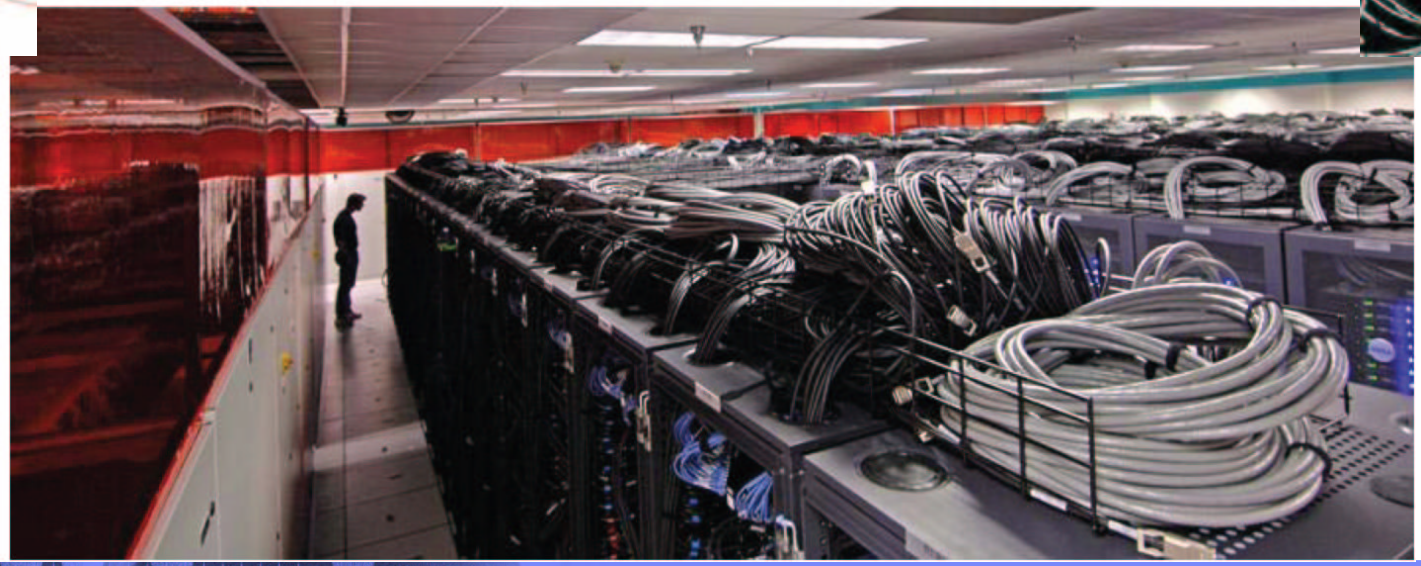
Top100-500 as predictor for commercial market 3-5 years out

- **TOP500 supercomputer list represents highly aggressive use of interconnect technologies**
- **Top300-500 (nov 05)**
  - GBE has 51%, growing
  - IB growing
  - All others are declining
- **Top100-300 (nov 05):**
  - GBE has 69%, expected to grow
  - IB has 4%, growing
  - All others are declining
- **Top1 – 100 (nov 05)**
  - Only growing proprietary interconnects: BlueGene/L + Cray
  - GBE declining rapidly
  - IB emerging
  - This is the area where people are willing to pay high premium (‘capability machines’)
  - Interconnect standard conformance not important criteria
- **Lessons:**
  - HPC & server markets have almost fully adopted Ethernet and InfiniBand for interconnect
  - Apparently there is not much willingness to ‘pay extra’ for improved interconnect performance
  - Recommended strategy: focus on Ethernet and IB *standards,* focus on *cost*

Further interconnect observations

- **Significant HW cost is in the cables, connectors, chip pins**
  - So we better start using them!
  - Datacenter bandwidth no longer is ‘free’
  - No longer can afford to throw bandwidth at obtaining QoS
    - Throw know-how at it instead
  - Optical cables will only see volume only when cheaper than copper





2010 Commercial data center Interconnect requirements

- **Standards based**
- **10 Gbps port speeds**
- **High scaling (up to 10'000 ports)**
- **Low port cost: comparable to 1GBE cost today (incl. cable)**
- **High density**
  - Low power consumption
  - NIC on motherboard
  - high port count switches
- **Latency: ~ 10 μs (app to app)**
  - Not as ultra-low as for HPC
  - First commercial applications start to demand guaranteed delays today
    - Finance
    - Web based applications
  - Need lossless operation and QoS function
- **Interconnect management**
  - Standards based

Cost scaling of host facing ports

194 ports @ 1G
~175$ / port

776 ports @ 1G
~525$ / port

Thousands of ports

Scaling and reliability

- **For a 64 way cluster, running with 10Gbps links**
  - One corrupted packet corrupted every 13 minutes

    (128 unidirectional links with BER of 10^{-15})
- **For a 2048 way cluster, running with 30Gbs links**
  - One corrupted packet every 1.6 seconds

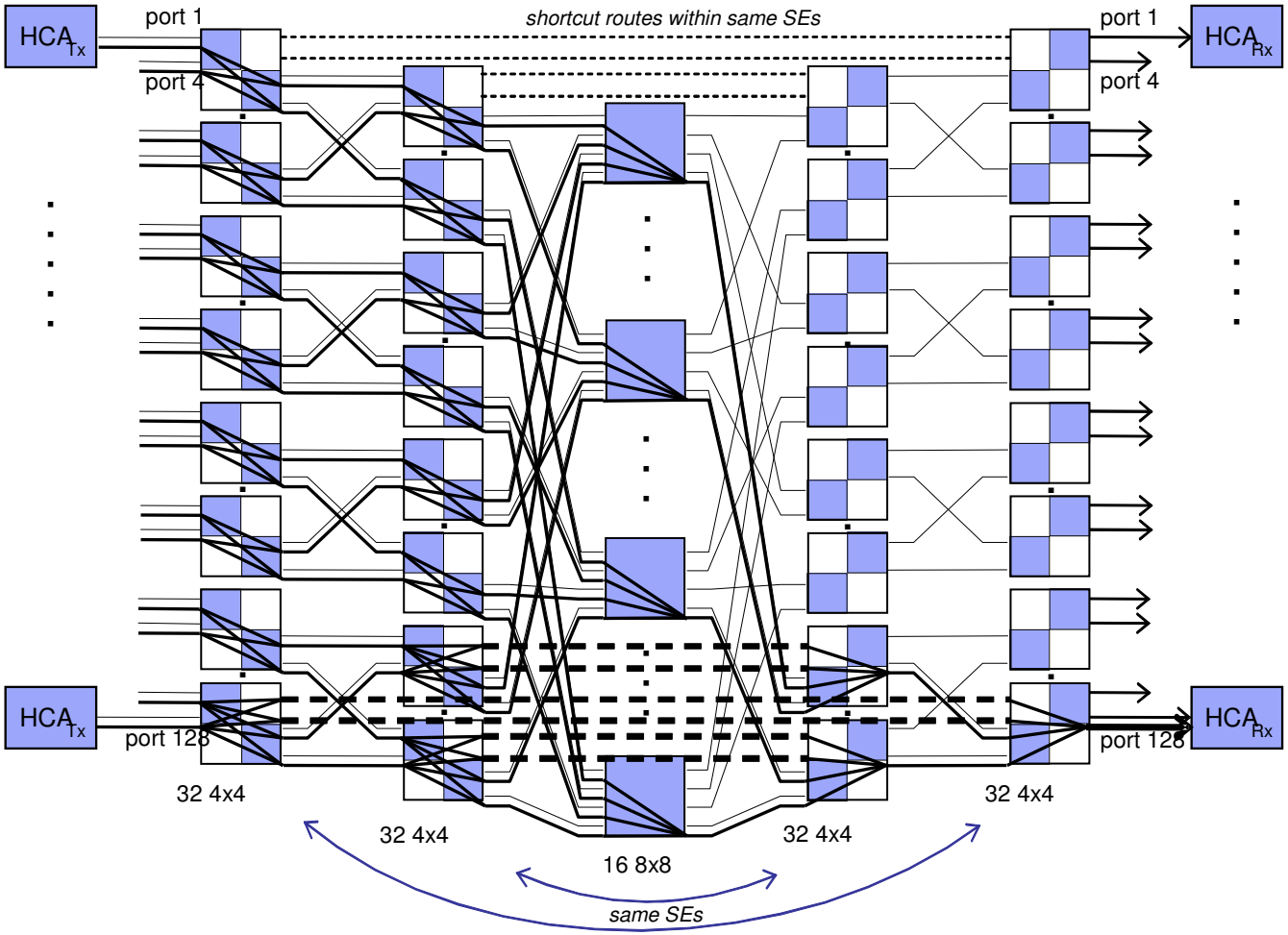
    (using 8X8 switch chips with BER 10^{-15} for on-board and cable links)
  - Optics have higher BER and make things worse
- **Some applications experience hiccup for each packet corruption**
  - Need hardware hop by hop retransmission
  - Think about FEC when using large systems, especially when using optics

The need for large, lossless interconnects

- **Lossless operation required to eliminate latencies incurred due to end-to-end packet retransmission resulting from buffer overflows**
- **High port count networks requires multistage topologies**
- **A lossless multistage network will require congestion control mechanism to avoid performance collapse**
  - or overprovisioning – which is becoming expensive
  - HPC treats congestion as application problem – not feasible for commercial (esp. with virtualization)
- **TCP congestion control only works for lossy networks**
  - TCP optimized for WAN, not datacenter
- **IB is lossless**
  - Credit based flow control
  - Already has congestion control in v1.2 of standard, products emerging now
- **Ethernet way behind IB, but expected to adopt key IB/FC-like function over time**
  - Very slow adoption: Ethernet community still largely biased towards 'throwing bandwidth at problem'
  - 802.3ar: Congestion control group, has defined the rate control mechanism, relying upon 802.1xx to define signalling mechanism. 802.1 has PAR underway
  - 802.1p defined QoS mechanism
  - Must maintain backwards compatibility

Throughput collapse in lossless networks

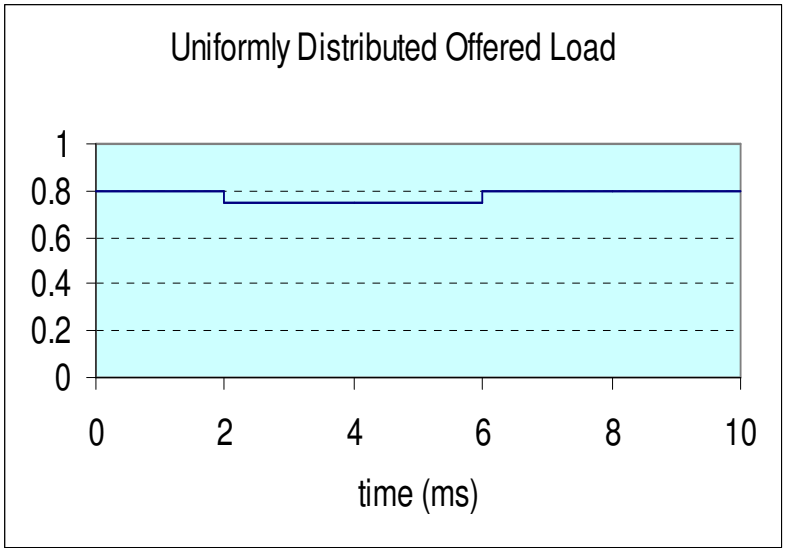
A multistage InfiniBand Fat Tree

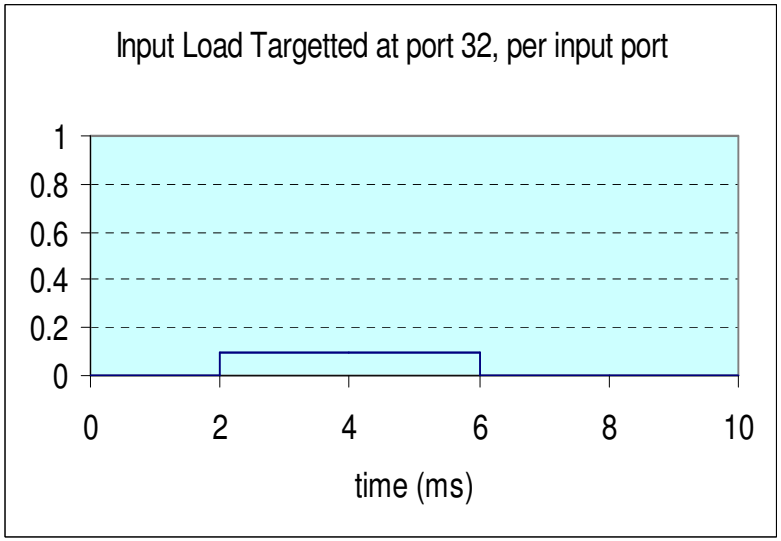


- **Drawn unfolded: Up on left, Down on right.**
- **Dashes & dots are shortcut paths within switches**

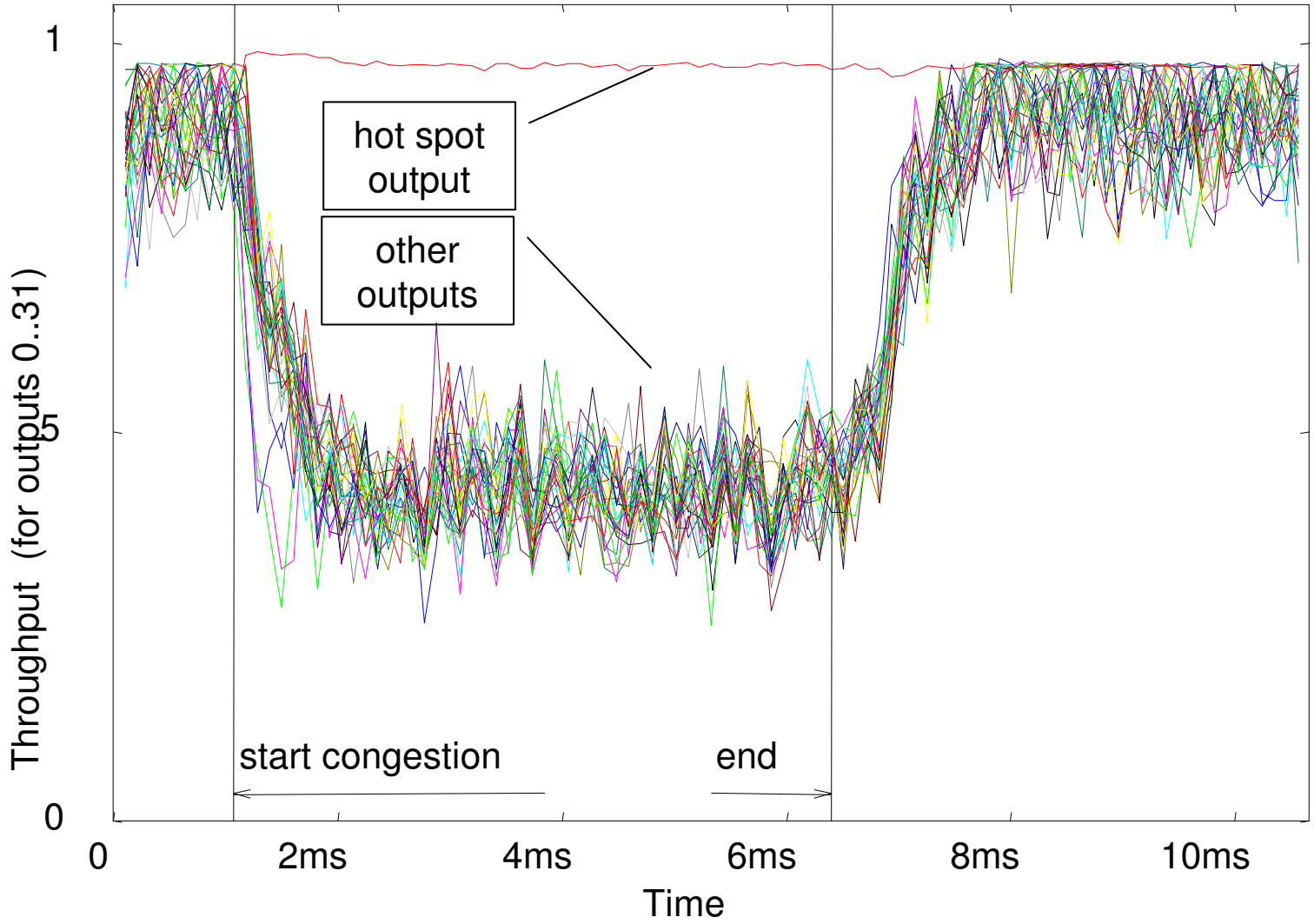
Congestion performance simulation

- **We simulate a 32 port multistage built with 8X8 switches**
- **Run for a 2 ms at 80% load, with destinations uniformly distributed from each source to each destination.**
- **Now that fabric reached steady state, inputs 1..32 each target 9% load to port 32**
  - Lower uniformly-distributed load to keep aggregate load constant.
- **4 ms later, go back to original uniform load.**



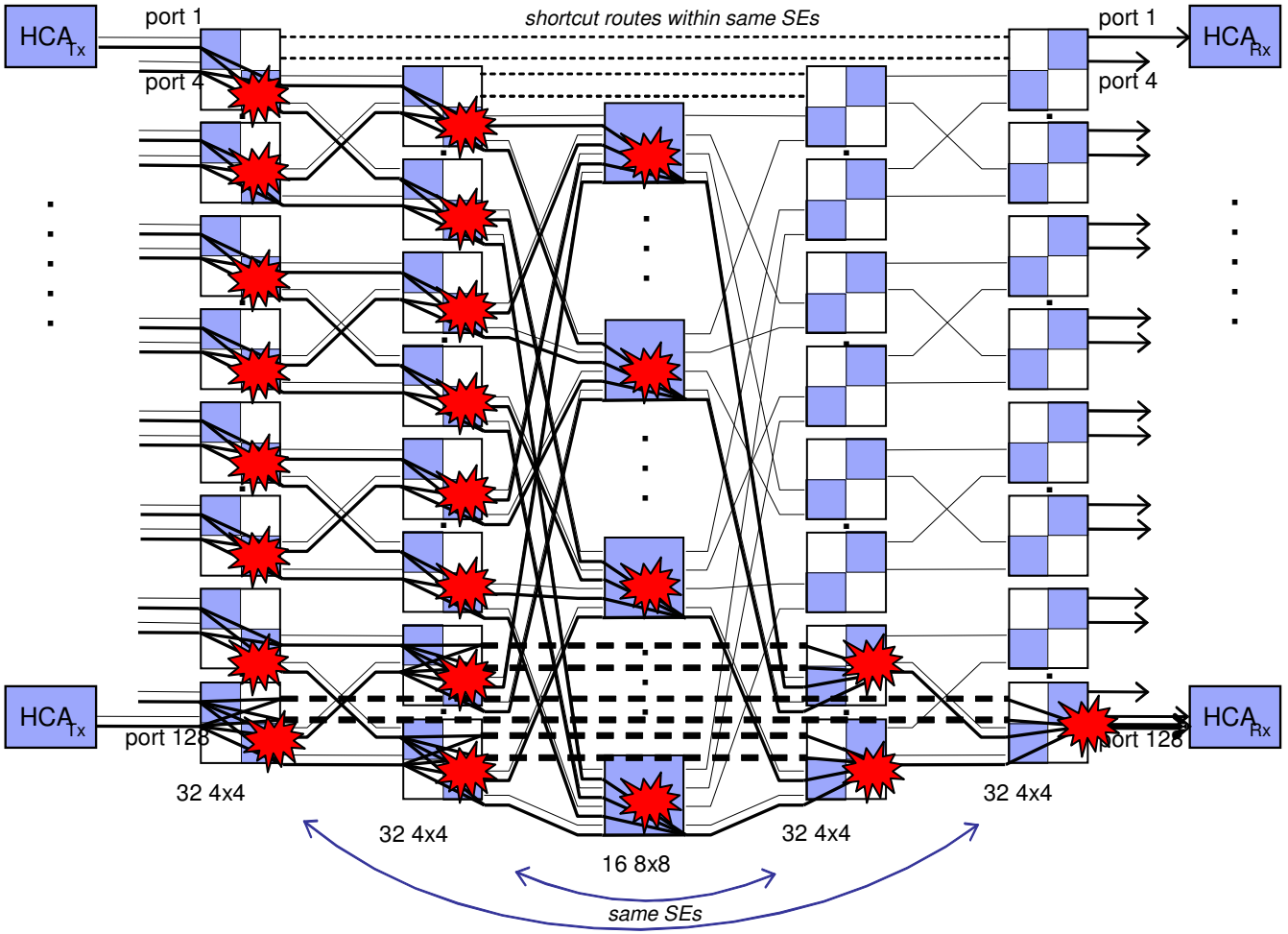


Result: Global Catastrophic Loss of Throughput



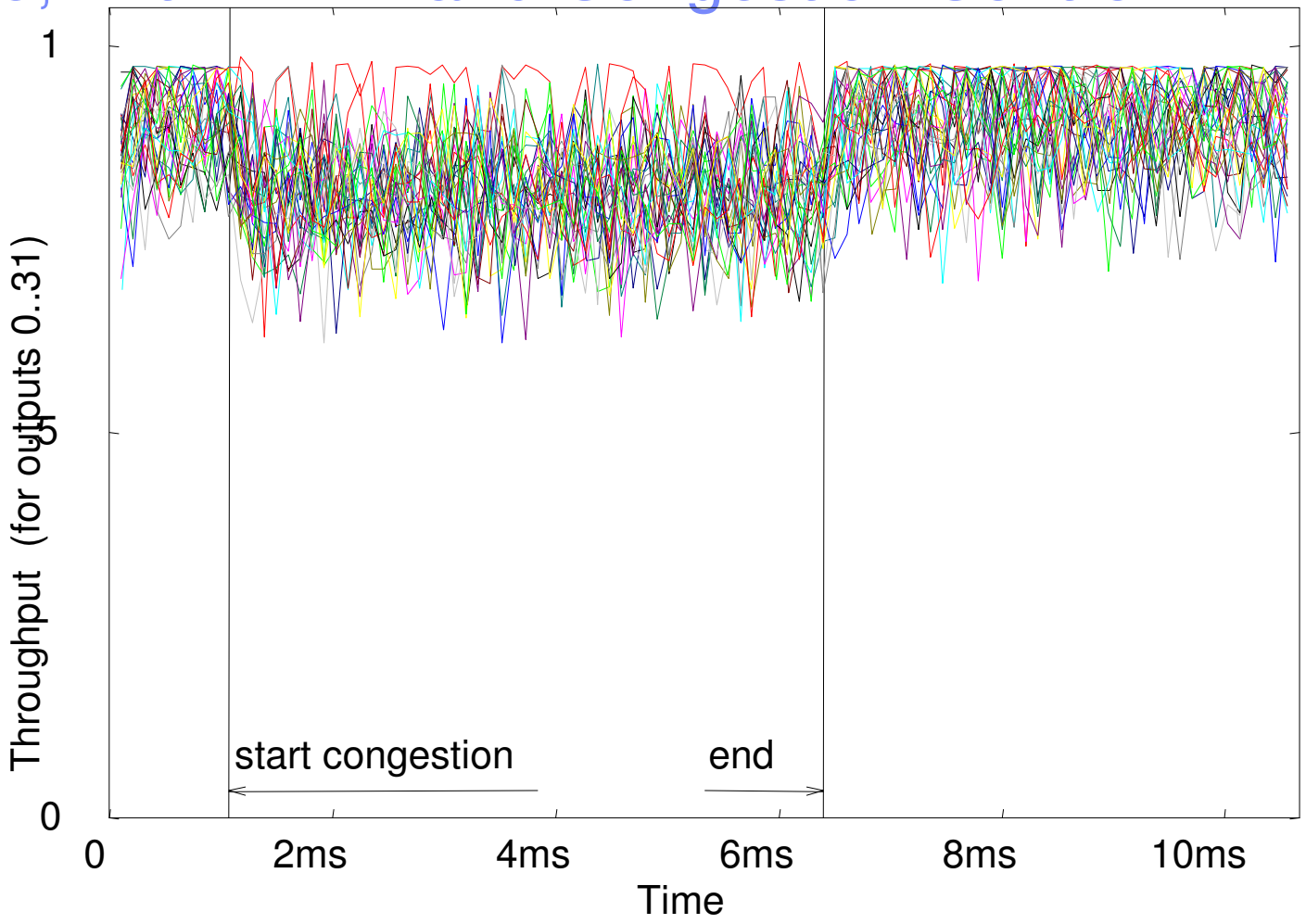
- **Traffic to *one* port messes up other ports**

Why: Tree Saturation / Congestion Spreading



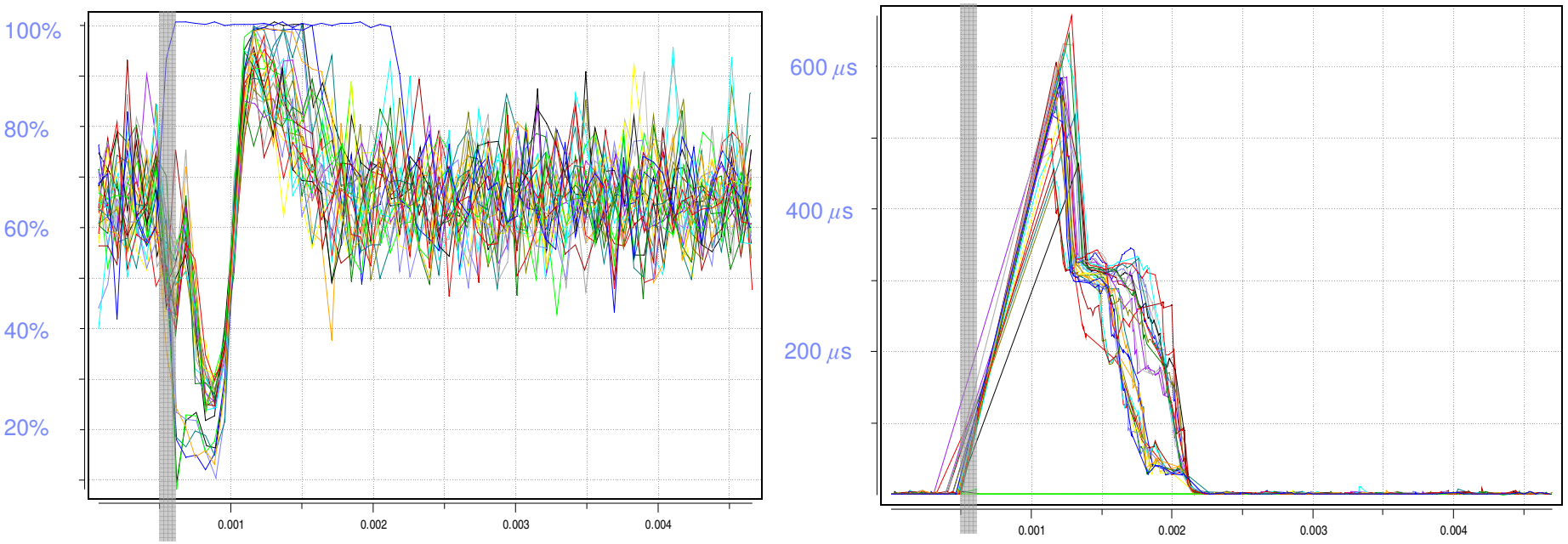
- **Hot output link saturates; link-level FC fills queuing of next stage**
- **Exhausts all storage in switch; backs up to next stage; etc., until all traffic blocked (high order head-of-line blocking).**

Same, With InfiniBand Congestion Control



- **Throughput drop = reduction in load keeping aggregate load constant.**
- **Simulations closely modeled product-purposed hardware designs.**

Another example of congestion



Overall load = 67%, from 0.0005 to 0.0006 3 inputs direct 67% each to single destination

Conclusions

- **Data centers with thousands of servers expected**
  - Will be main driver for high port count interconnects
- **Will need standards based, scalable, low-cost lossless interconnect solutions**
  - QoS, lower latencies required
  - Ethernet and IB probable candidates
  - Scaling requirements (retransmission, FEC, cost)
- **IB currently has cost advantage at 10Gbps**
- **IB has standardized Congestion control mechanism**
- **Ethernet is behind on congestion control**
  - Is being addressed in IEEE 802.
  - BTW we are working on CC mechanism for Ethernet