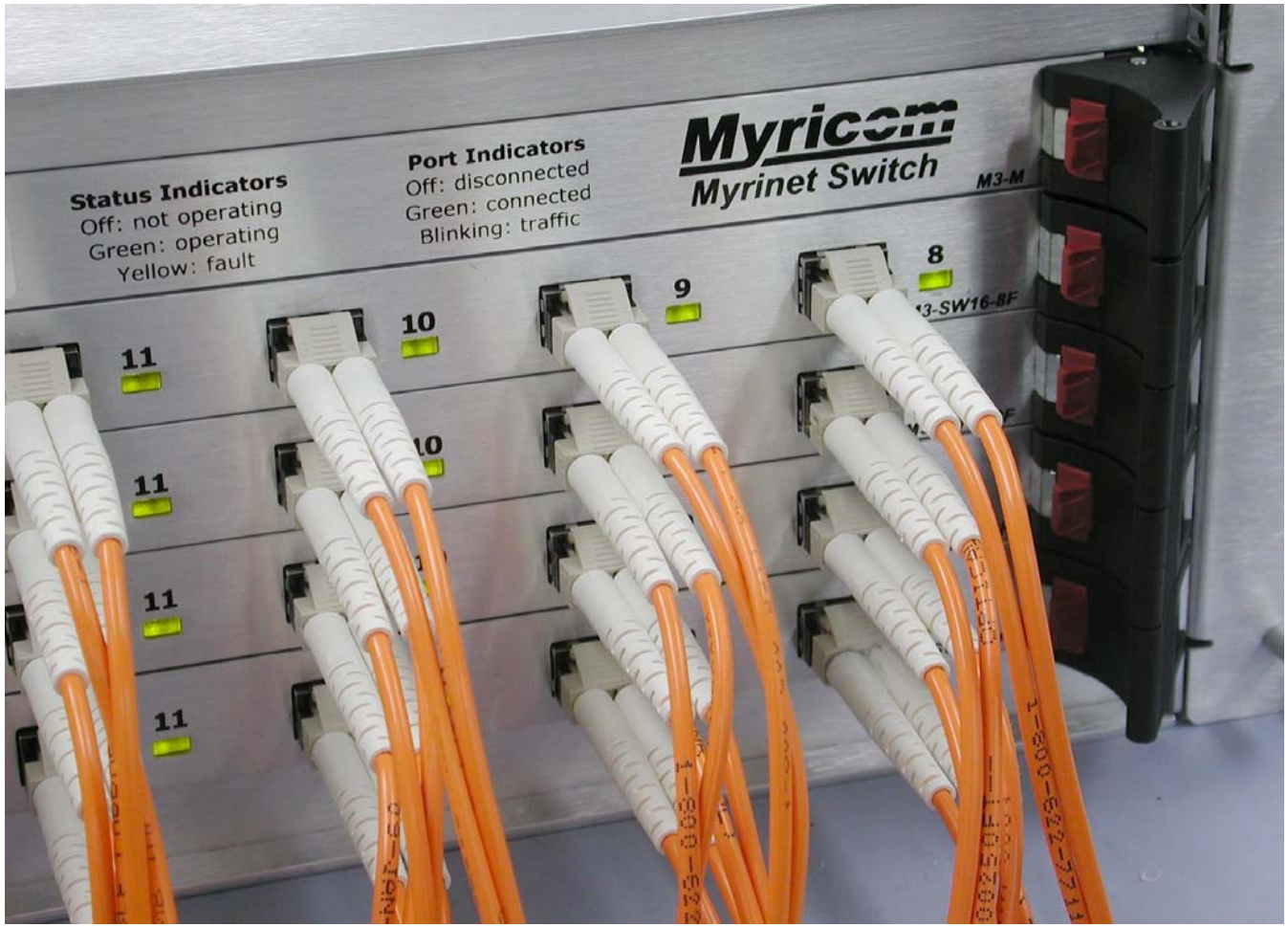
Battle of the Network Stars

Jeff Chase
Duke University

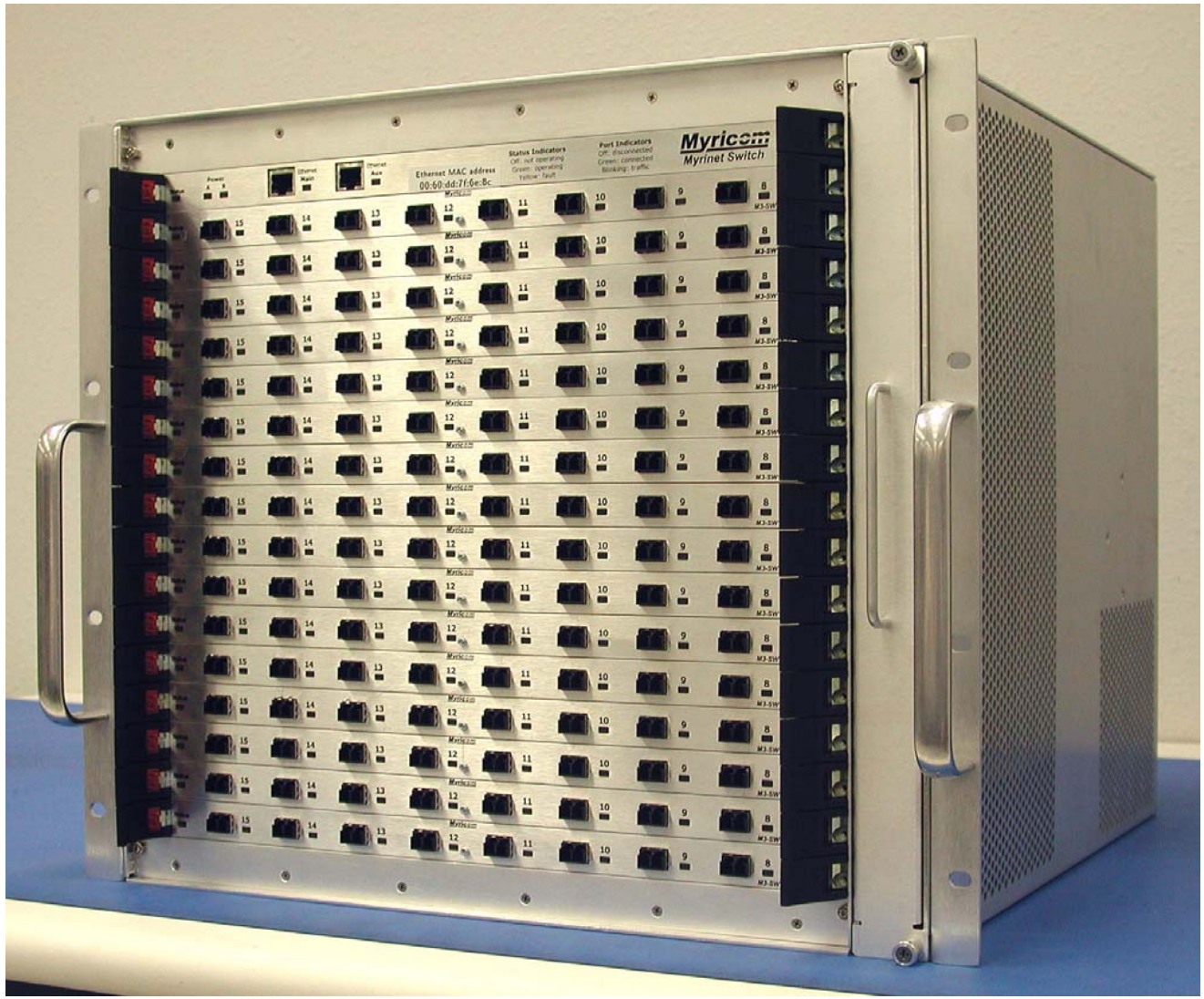
Myrinet-2000 Links, 2+2 Gbit/s, full duplex



Note: The signaling rate on these links is 2.5 GBaud, which, after 8b/10b encoding, yields a data rate of 2 Gbit/s.

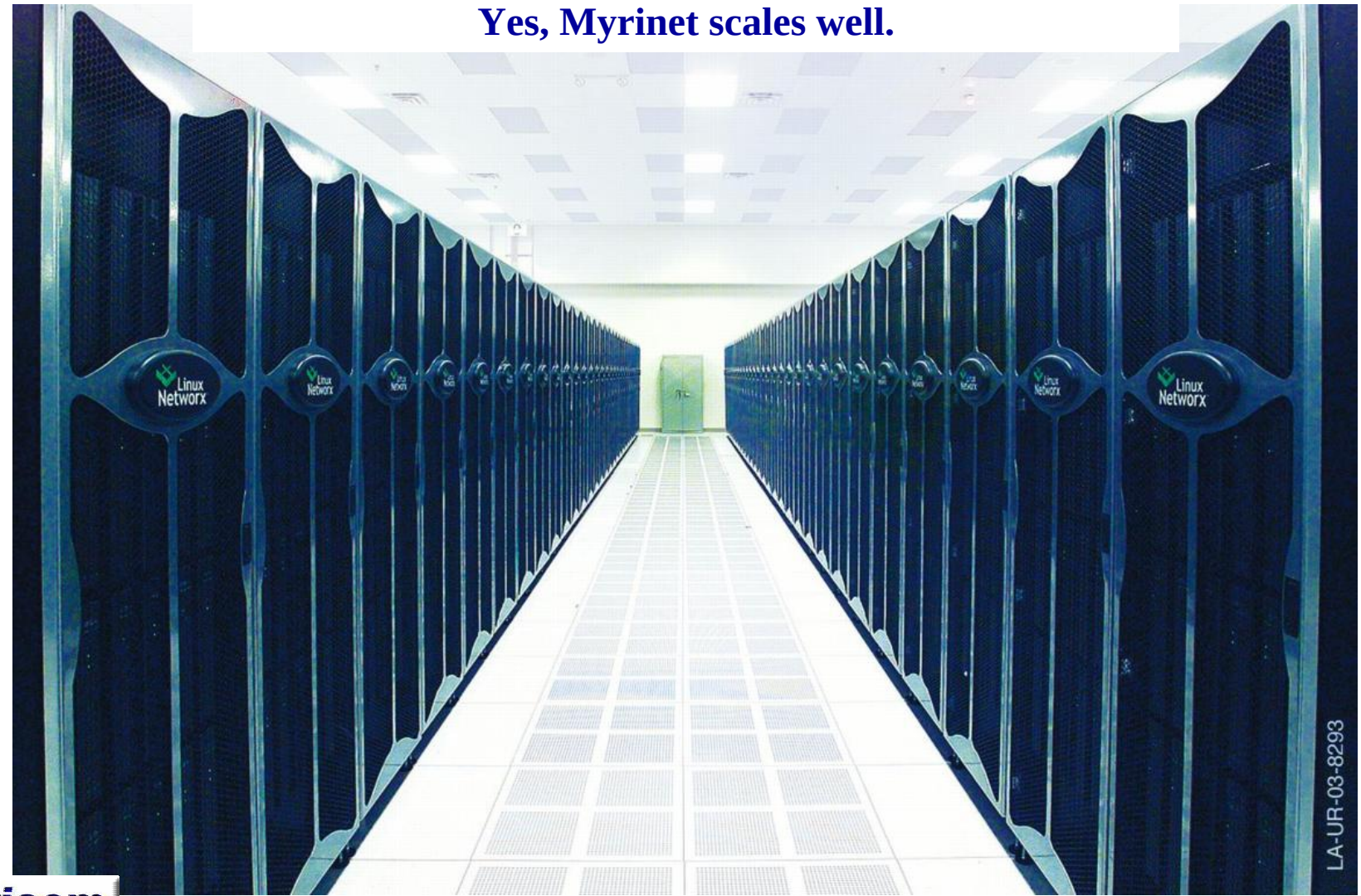
Advantages of fiber: small-diameter, lightweight, flexible cables; reliability; EMC; 200m length; connector size.

128-Host Clos Network



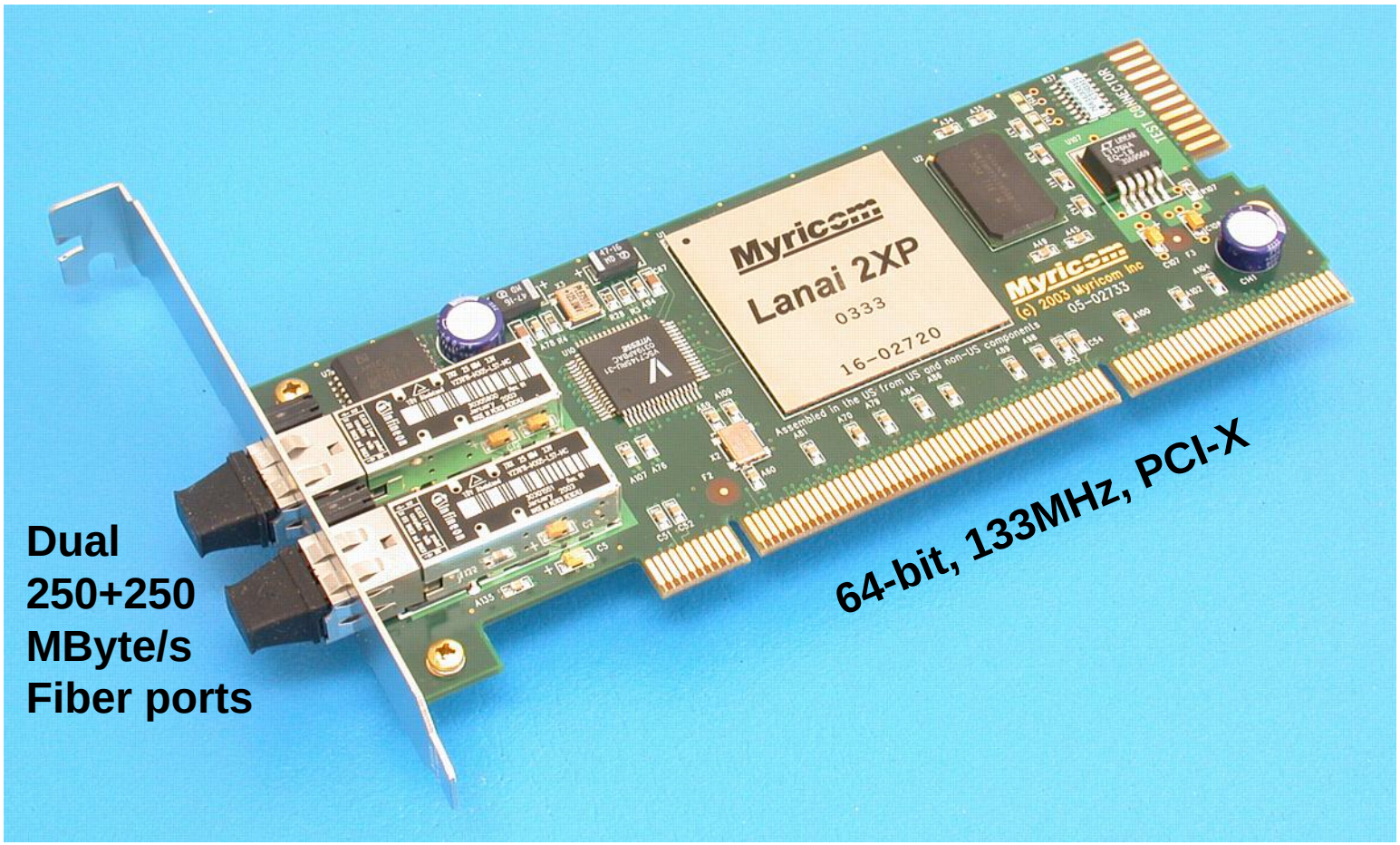
Myricom

LANL Lightning Cluster, 1408 dual Opterons, #6 on the Nov-03 TOP500, 8051 Gflops, 71.5% of peak.
Yes, Myrinet scales well.



Myricom

High-End Myrinet Interface (E Card)

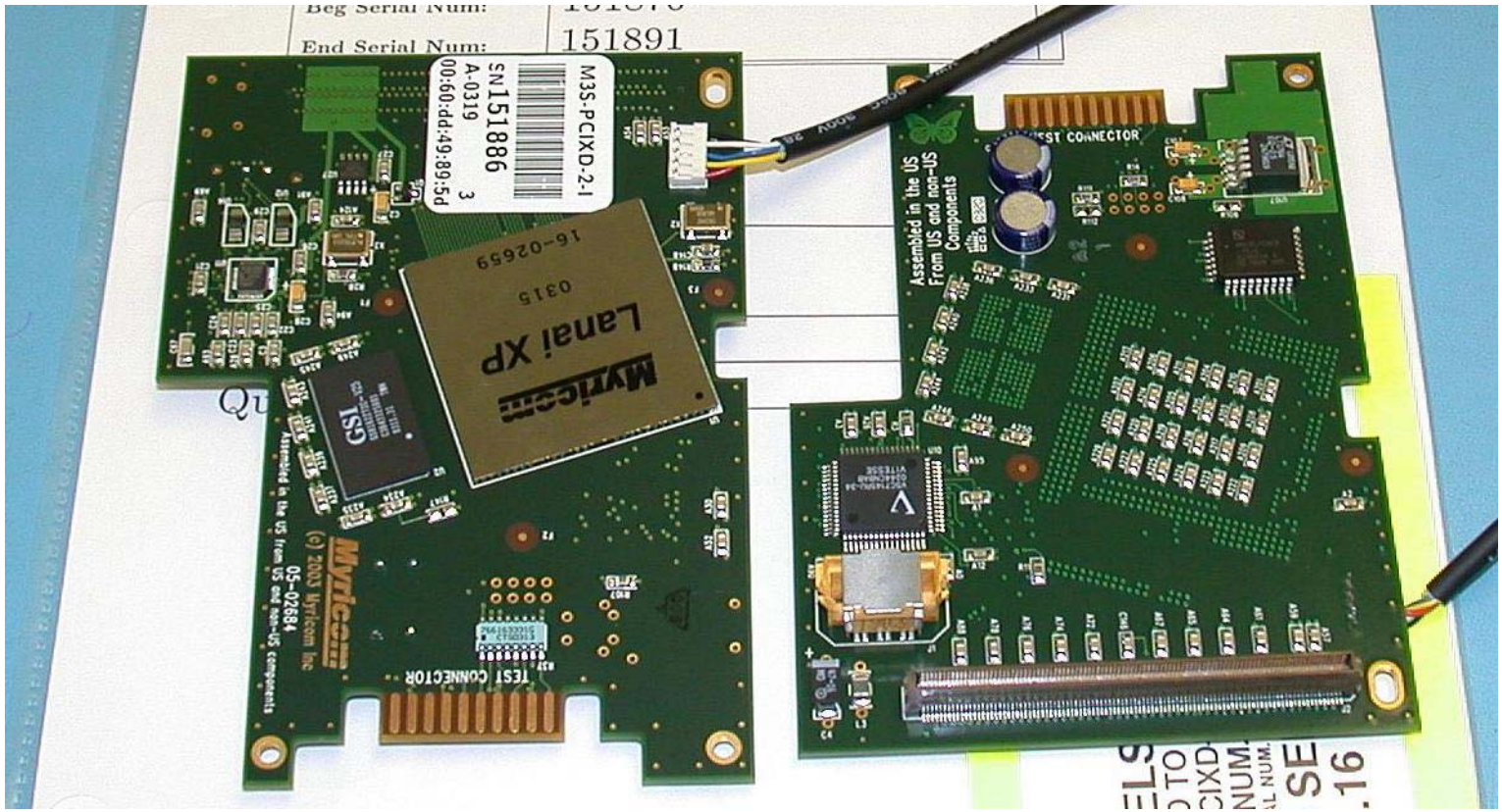


M3F2-PCIXE-2 two-port Myrinet/PCI-X Interface

Myricom

IBM BladeCenter version of the D card

(Photograph of 2 HCAs, to show both sides)



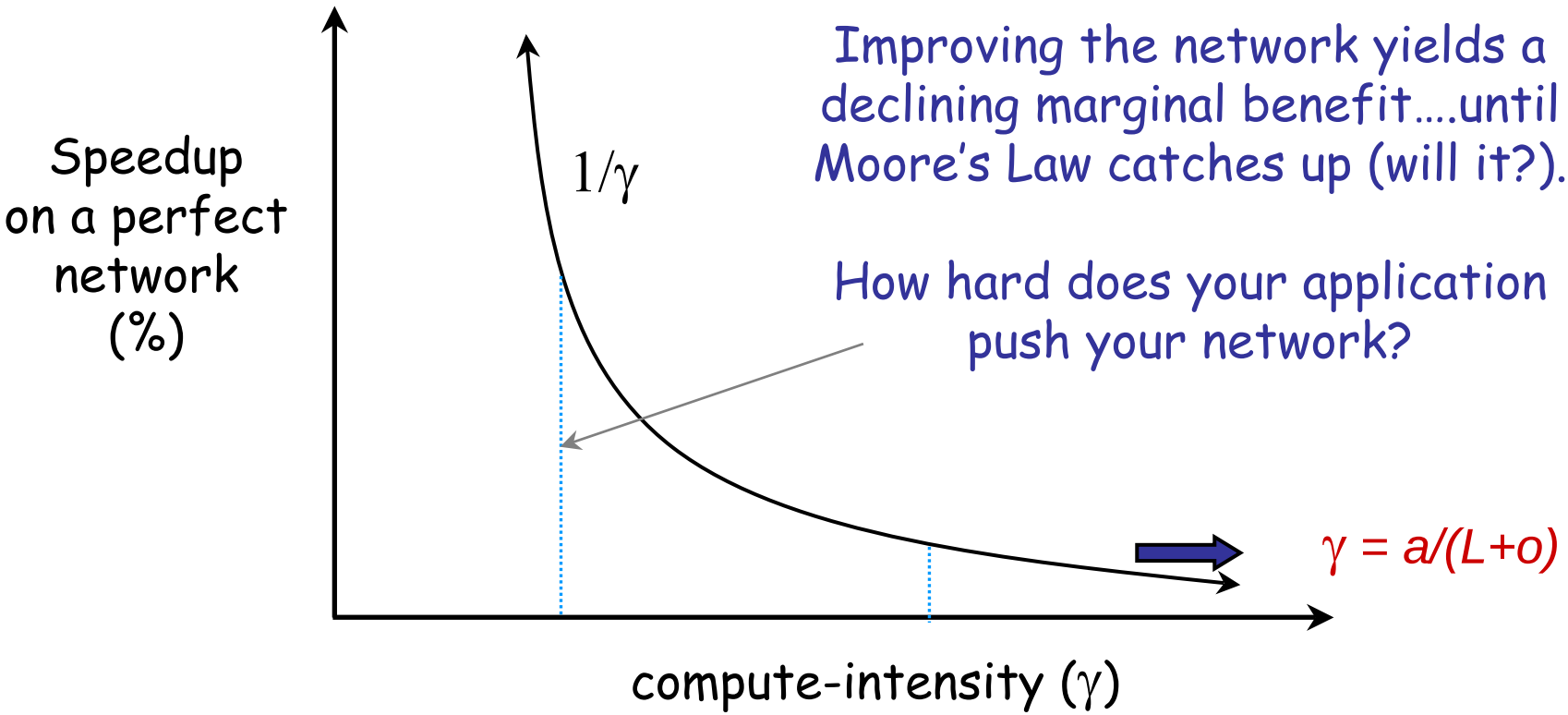
Product announcement http://www.myri.com/news/03909/

Comparing speed

- All of these nets are fast.
- Bandwidth
  - PCI-X 64@133 MHz can deliver 1 GB/s aggregate I/O
  - Infiniband or 10+GE can saturate it in either direction
  - Myrinet: two links (XE) saturate for bidirectional communication…which matters most for computing.
  - Quadrics: nominal 400 MB/s; less for bidirectional
- Latency
  - Small messages: gain/lose a μsec here and there
  - Large messages: see "bandwidth" above

How much do these speed differences really matter?

- Microbenchmarks can lead us astray
- Your Mileage May Vary



Overhead

- If your application pegs the CPU, then it's overhead that matters, not latency or bandwidth.
- The host/NIC interface determines overhead.
- Myrinet benefits from previous cycles of innovation on their open-source + programmable NICs.
  - MyriAPI ⇒ GM ⇒ MX
  - 'Bazaar' academics add research: AM, FM, Trapeze, etc.
- Fluid host interface ⇒ free to innovate up to the API
  - Ohio benchmark results (and others) show Myrinet/GM is still the leader in low-overhead network I/O.
- Myrinet MX is designed for low-overhead MPI

…And the Competition

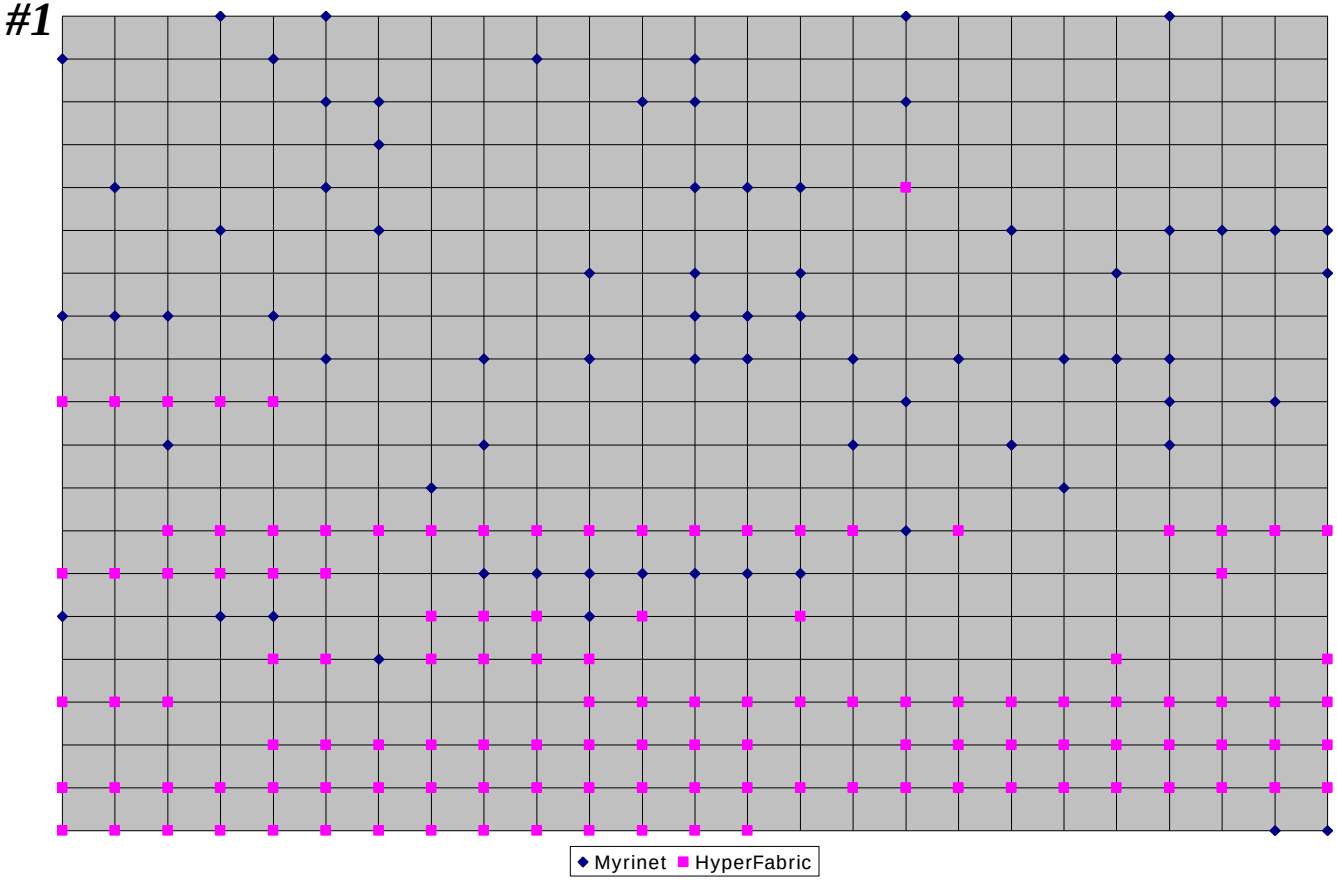
- 10+/-GE
  - Ethernet means IP protocols, which were designed to push all the overhead to the end hosts. It won't be competitive without…
- IP-DDP and IP-RDMA ⇒ transport protocol offload.
  - Difficult engineering challenges and an uphill fight against a hostile entrenched IETF oligarchy.
  - But we'll get there some sunny day.
- Infiniband
  - Implementations still maturing (e.g., sensitive to buffer locality), host interface standard ⇒ designed by committee.
- Quadrics
  - Also programmable NICs

Other criteria

- Collective communication and topology
  - Point-to-point measures give no insight into hot spots, head-of-line blocking, etc.
  - Myrinet: modular switches, source routing and multipath dispersion
- Manageability and TCO
  - Predictability, stability, interoperability
  - Your application might not care…but you will!
  - Myrinet: Ethernet emulation, automatic network mapping, fault detection

Myrinet in the November 2003 TOP500®

194 ***of the*** TOP500 SUPERCOMPUTER SITES ***use Myrinet technology***



73 Myrinet clusters, up from 51 5 months ago

34 new or upgraded TOP500 clusters

2 in the top 10

121 HyperFabric, which is HP's brand for Myrinet

Myricom

Interface	M3F-PCIXD "D card"	M3F2-PCIXE "E card"
Interface chip	Lanai XP	Lanai 2XP
Myrinet ports	1	2
Lanai & memory clock	225MHz	333MHz
Local memory data rate	1800 MB/s	2664 MB/s
Peak bidirectional Myrinet data rate	500 MB/s	1000 MB/s
Peak PCI data rate	1067 MB/s	1067 MB/s
Myricom software support	GM 2 & MX	GM 2 & MX
MX and MPI/MX bidirectional throughput	**490 MB/s**	**950 MB/s**
MX and MPI/MX latency	**4µs**	**3.5µs**