InfiniPath: A New High Speed, Low Latency Cluster Interconnect

Greg Lindahl, Distinguished Engineer

PathScale, Inc.

lindahl@pathscale.com

Embedded use of InfiniPath

- Attaches to HyperTransport (16 bit)
  - via HTX slot or directly on a motherboard
- 5 watts
- similar size to AMD 8131
- 1.32 usec 8-byte latency
  - low latency is critical for real-time embedded applications
- 99.99% of packets < 1.74 usec
  - This is with standard Linux; not RT Linux

Good interconnect “hero numbers”

		InfiniBand		Proprietary		10 GbE
		PathScale InfiniPath	Mellanox Ex	Quadrics Elan 4	Myricom E/F-Card	Chelsio
MPI	Latency (μs)	1.32	4.0*	1.4 ~ 2.0*	2.6 ~ 5.5*	10.5
	Bandwidth (MB/s) Unidirectional Bidirectional	952 1,842	970* 1,841*	875 ~ 910* 901*	493* 749*	830
	N ½ Message Size (Bytes)	385	~2,048*	512 ~ 1,024*	1,024 ~ 2,048*	98,000
TCP	Latency (μs)	6.7	20 ~ 30		32	9.6
	Bandwidth (MB/s)	583	199 ~ 425	712	232	988

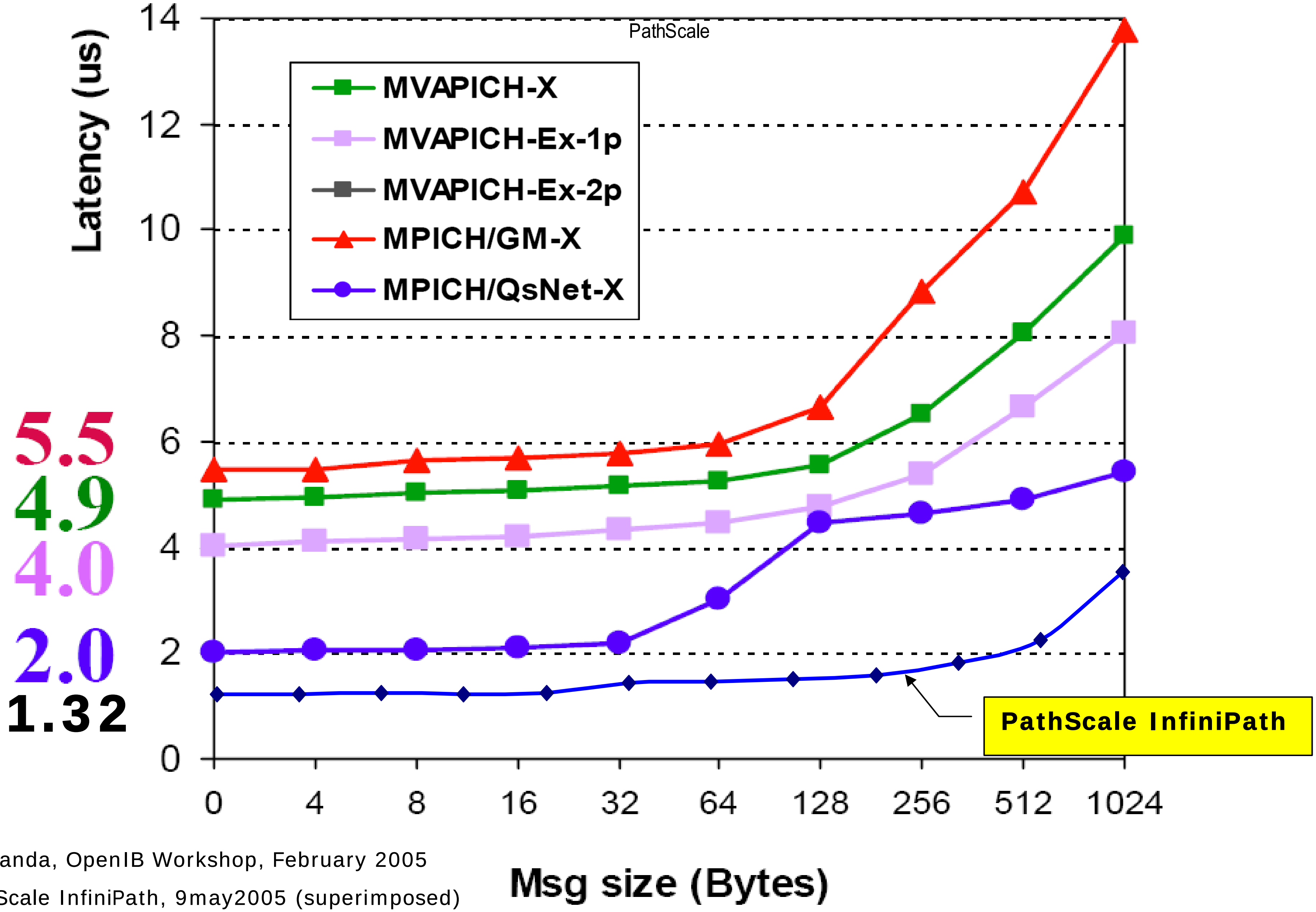
MPI Sources/Notes:

- PathScale – PathScale measurements with one switch crossing, May 2005
- * Ohio State measurement results – DK Panda, OpenIB Workshop, Feb 2005
- Quadrics – IEEE Micro, to appear 2005
- Myricom – Myricom website 11oct2004, presentation 18may2004
- Benchmarks - OSU MPI Benchmarks 2.0 (streaming)

TCP Sources/Notes:

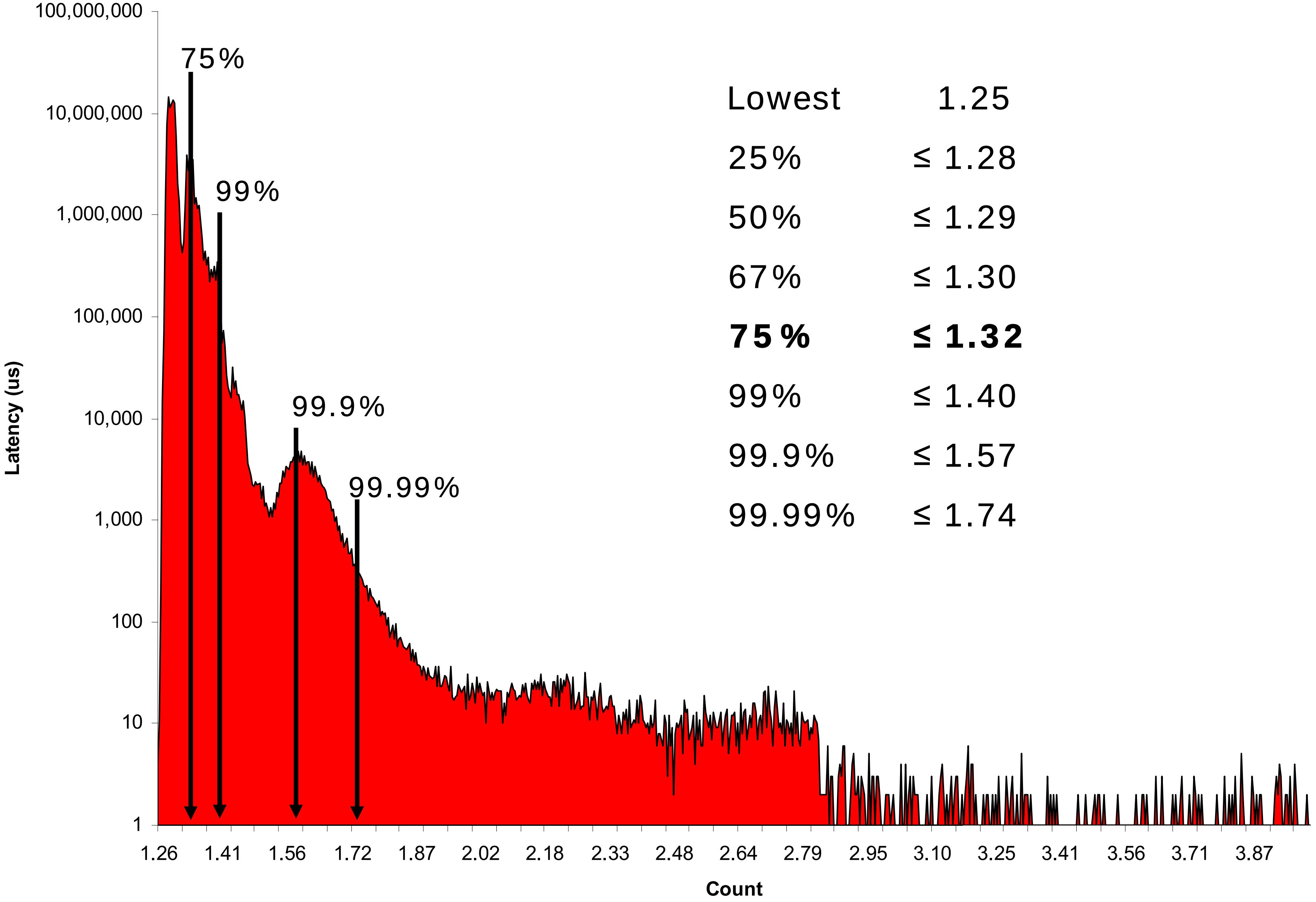
- PathScale – PathScale measurements with one switch crossing, May 2005
- Quadrics – Quadrics website
- Myricom – Myricom website (results for C-card)
- Mellanox – Pathscale measurements, OpenIB Workshop
- 10 GbE – Chelsio T210 Protocol Engine, Scali MPI
- Benchmark – netperf 2.3, one-way latency, goodput

OSU: MPI Small Message Latency



Source: DK Panda, OpenIB Workshop, February 2005
PathScale InfiniPath, 9may2005 (superimposed)

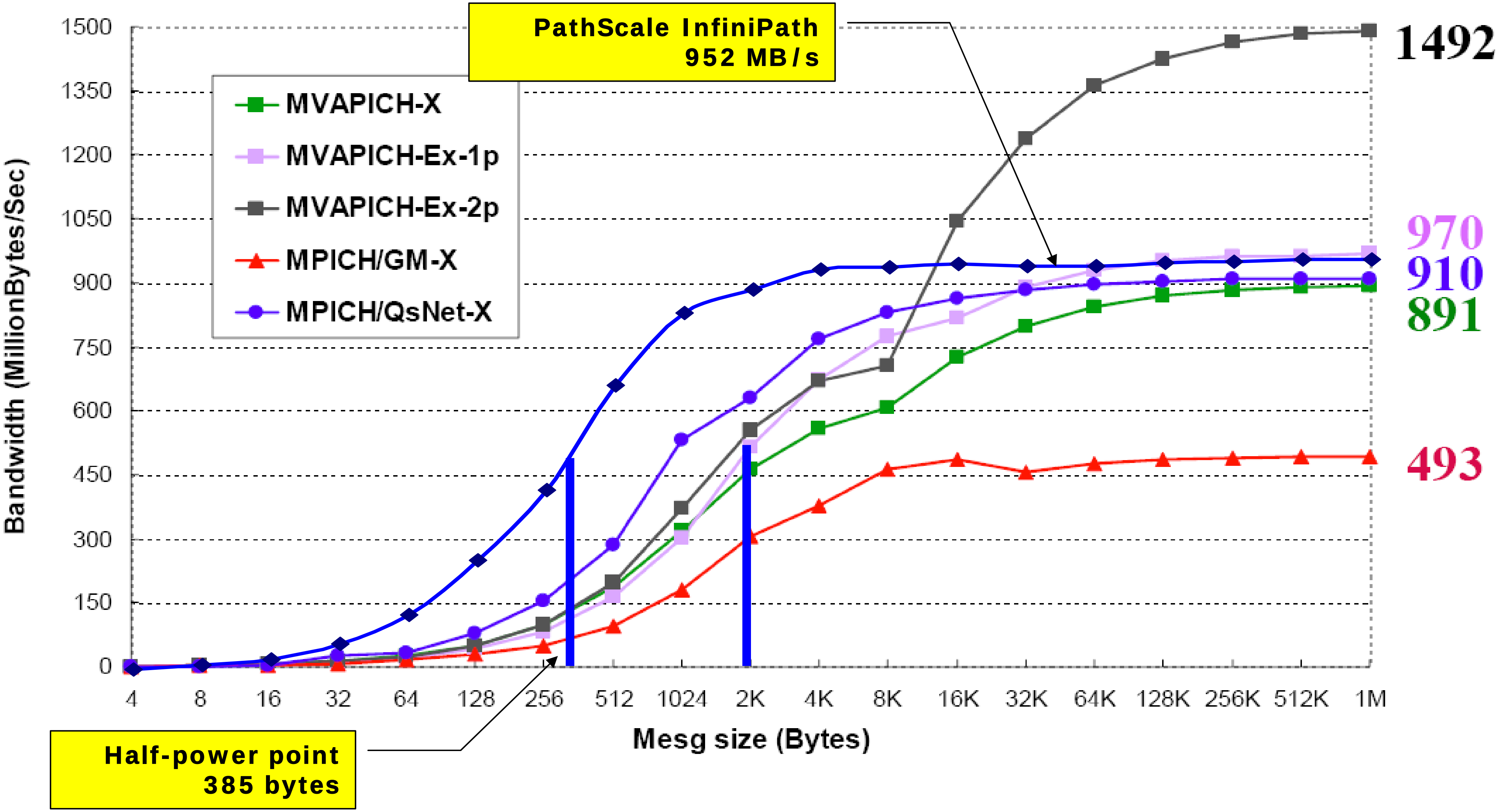
How often we hit “hero” latency in 100M tries



Lowest	1.25
25%	≤ 1.28
50%	≤ 1.29
67%	≤ 1.30
75%	**≤ 1.32**
99%	≤ 1.40
99.9%	≤ 1.57
99.99%	≤ 1.74

Source: PathScale InfiniPath, 28jun2005

OSU: MPI Uni-directional Bandwidth



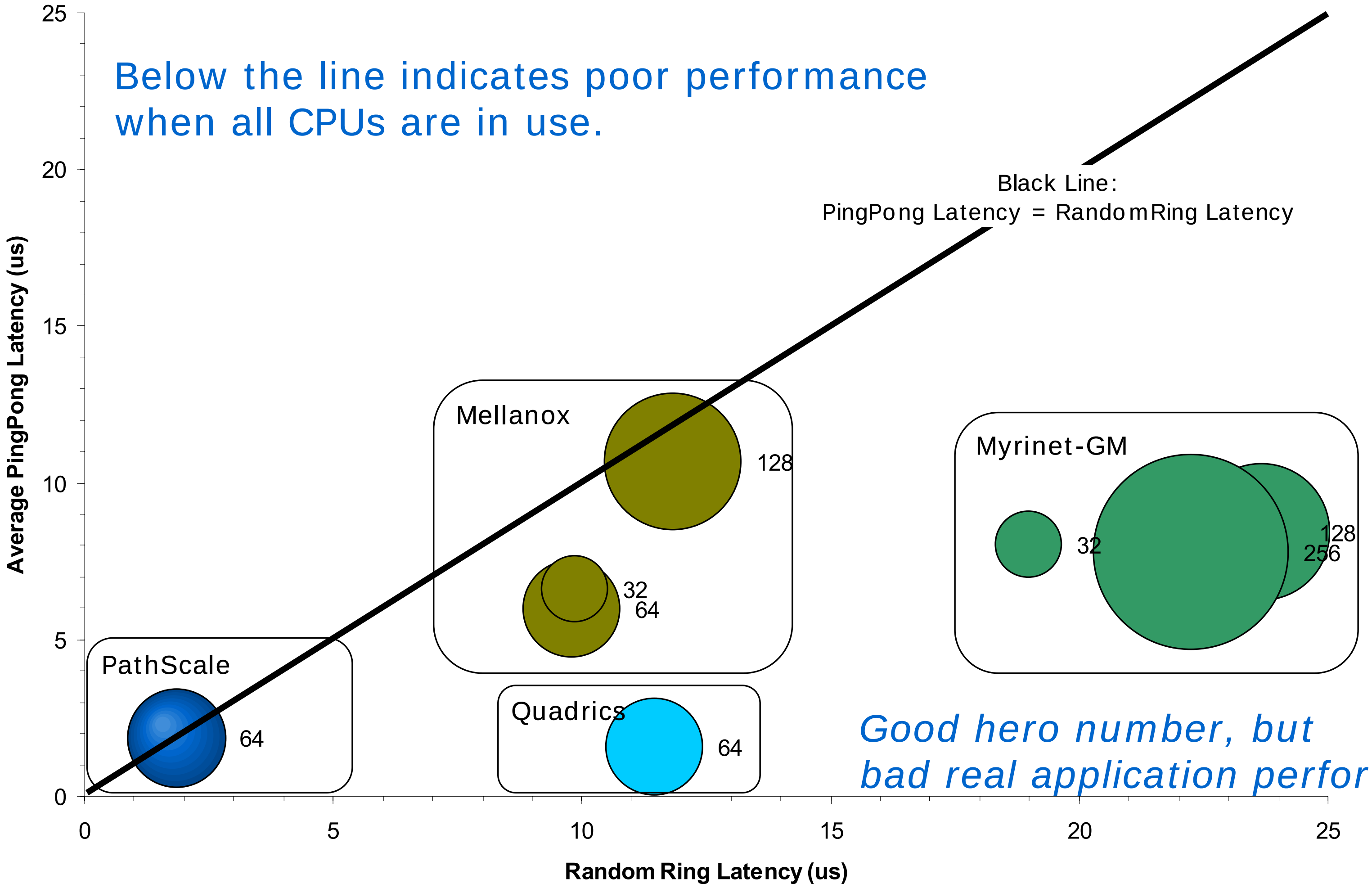
Source: DK Panda, OpenIB Workshop, February 2005

PathScale InfiniPath, 9may2005 (superimposed)

Better than just good hero numbers

- □ The typical “latency” measurement is a 2-node, 2-cpu ping-pong
- □ It’s much more realistic to use all the cpus, and have each one talk to more than 1 neighbor
- □ The HPC Challenge Random Ring benchmark does this... and it searches for the worst latency, too

Most interconnects score poorly

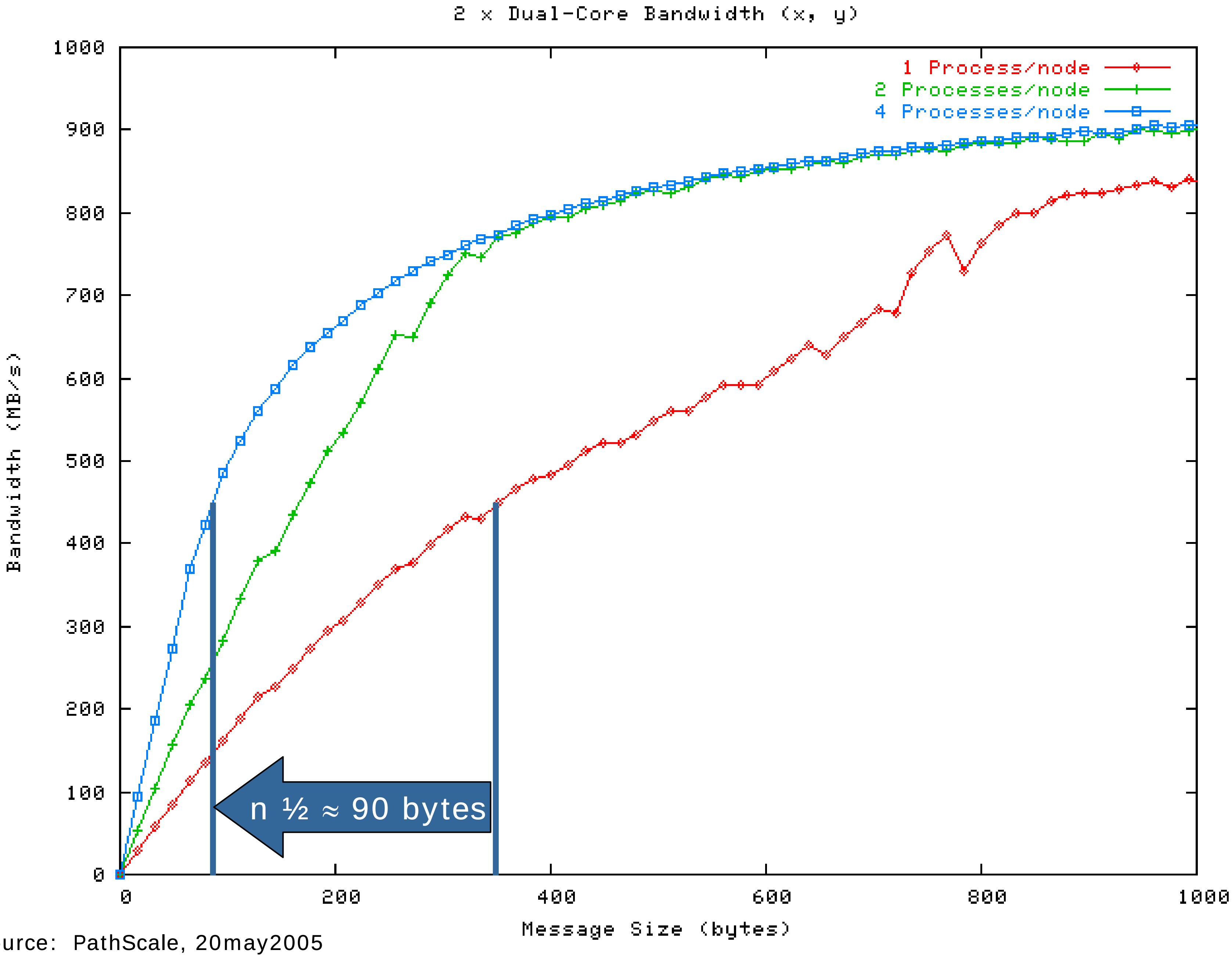


Below the line indicates poor performance when all CPUs are in use.

Good hero number, but bad real application performance!

Sources: HPC Challenge Website (June 11, 2005); PathScale measured by PathScale (May 2005; 2.0 Ghz cpu). Size of circle indicates cpu count.

Our performance increases with additional cpus



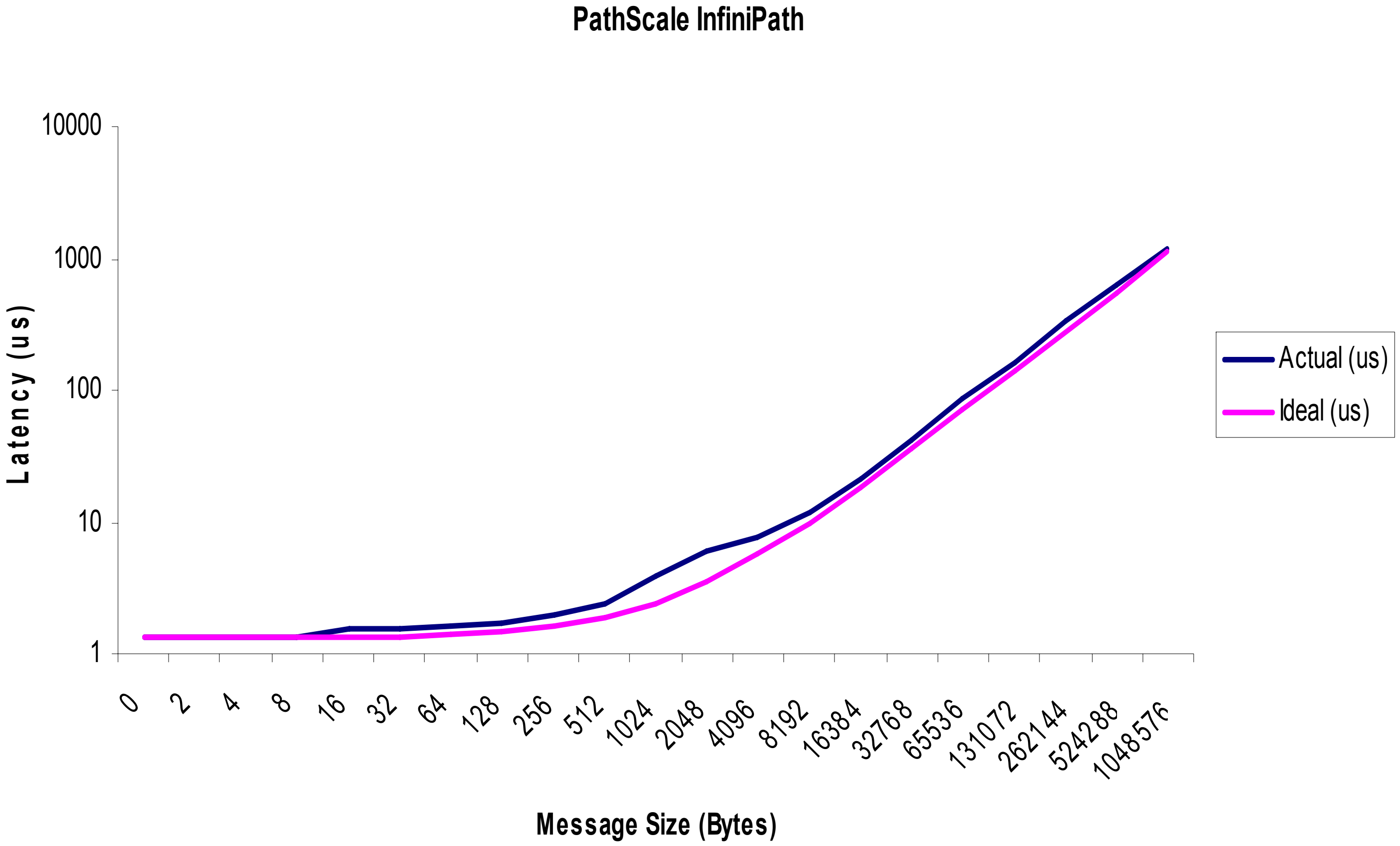
n ½ ≈ 90 bytes

Source: PathScale, 20may2005

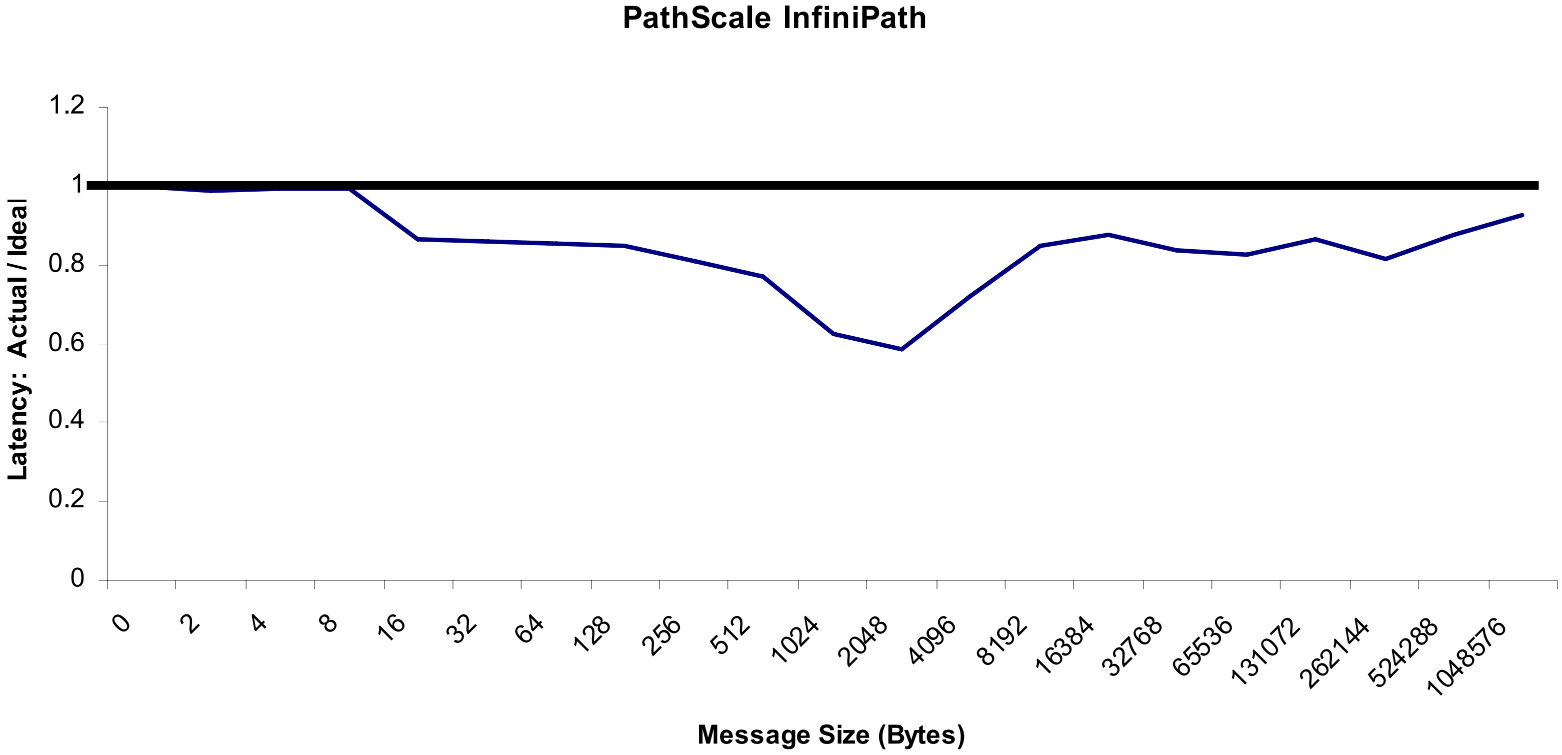
Log/Log Charts Are Misleading

Ideal(size) = latency(size=0) + size/bandwidth

Can you tell how much slower we are than ideal? 20%? 50%? 100%?

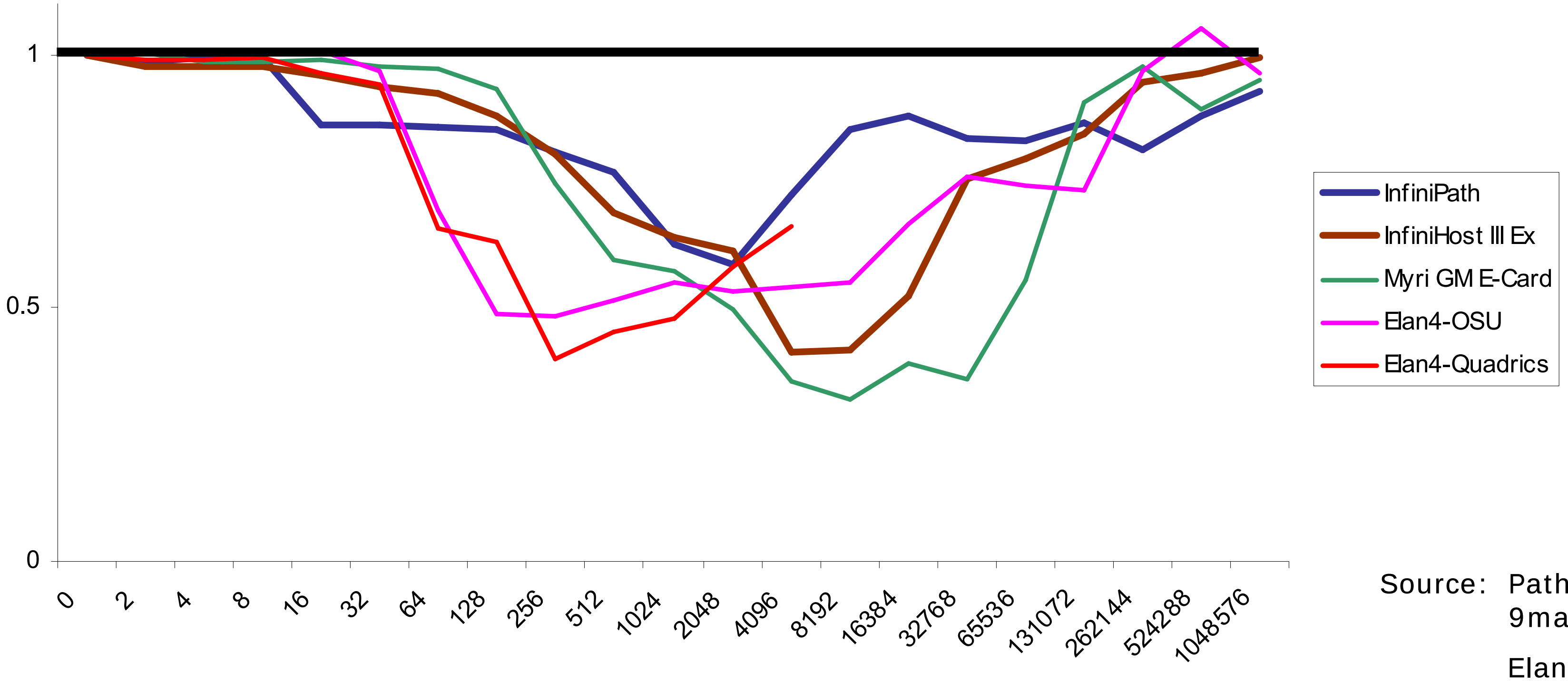


Linear scale makes it obvious



InfiniPath performance is closest to ideal

If you wanted to have a 3rd number in addition to "latency" and "bandwidth", the worst fraction of ideal would be a good one. It would range from 0.3 to 0.6 for various interconnects.



Source: PathScale InfiniPath, 9may2005

Elan4-Quadrics – Petrini IEEE Micro, 2005

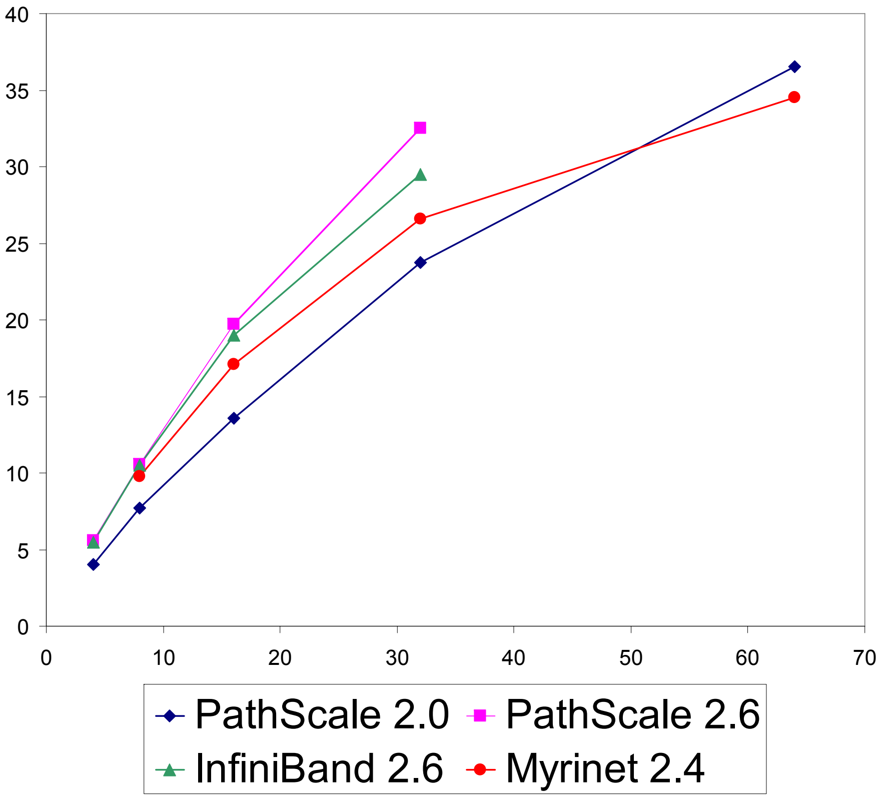
All others, Ohio State, 4sep2004.

Real Application: LS-DYNA (crash code)

neon_refined test case from topcrunch.org

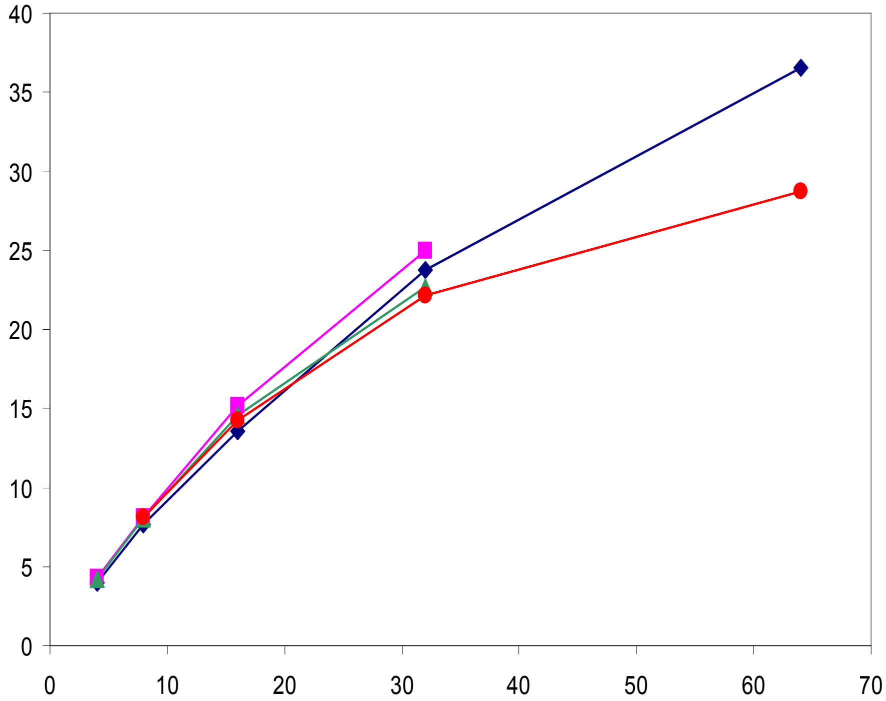
Comparison data from topcrunch.org, September 18, 2005

Performance



(all systems are AMD Opterons)

Performance scaled by CPU clock



Note: LSTC is not yet supporting InfiniPath

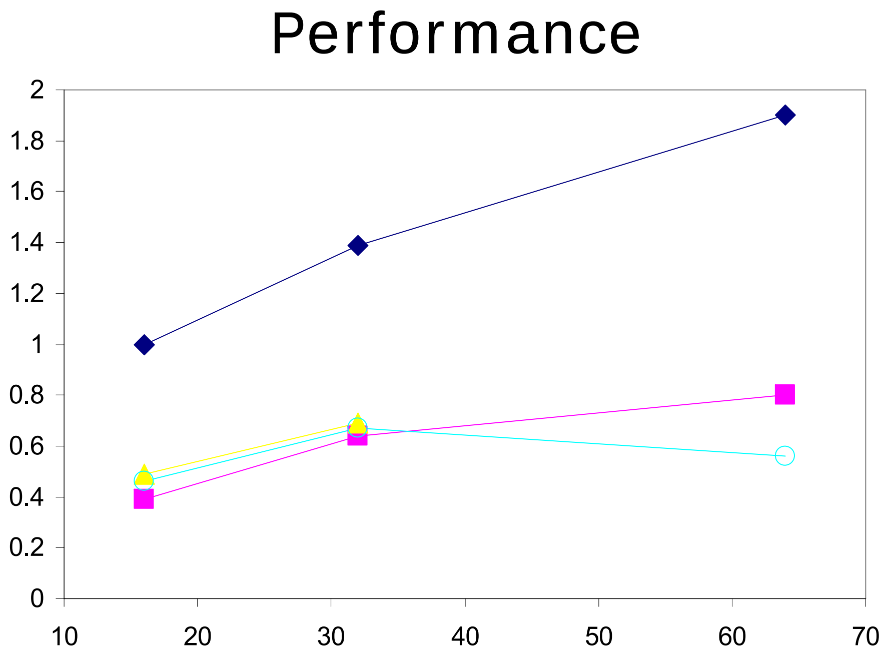
Real Applications

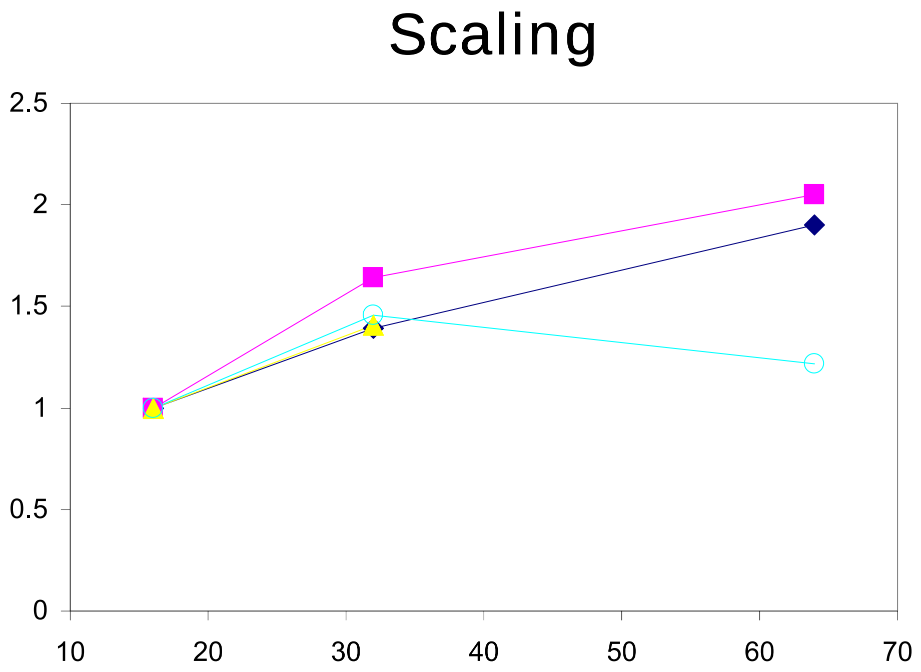
- Hard to find useful results
  - Competing vendors don’t publish many results other than hero numbers
  - Many results on the web use obsolete cpus
- So, we do “apples to cran-apples” comparison
  - Compare scaling, not absolute performance
  - Note that a faster cpu should make scaling worse
  - Reminder: InfiniPath benefit shows up when applications are not scaling perfectly
- We’d love to see more published results with modern cpus

Real Application: CHARMM

Charmm is a quantum chemistry app which is well known to be hard to scale.

Dark blue: 2.0 Ghz Opteron + InfiniPath
Pink: 2.0 Ghz Pentium4 + Myrinet
Yellow: 2.66 Ghz Pentium4 + Myrinet
Cyan: 1.0 Ghz AlphaServerSC + Quadrics Elan3





Data: http://www.cfs.dl.ac.uk/benchmarks/commodity/sld037.htm

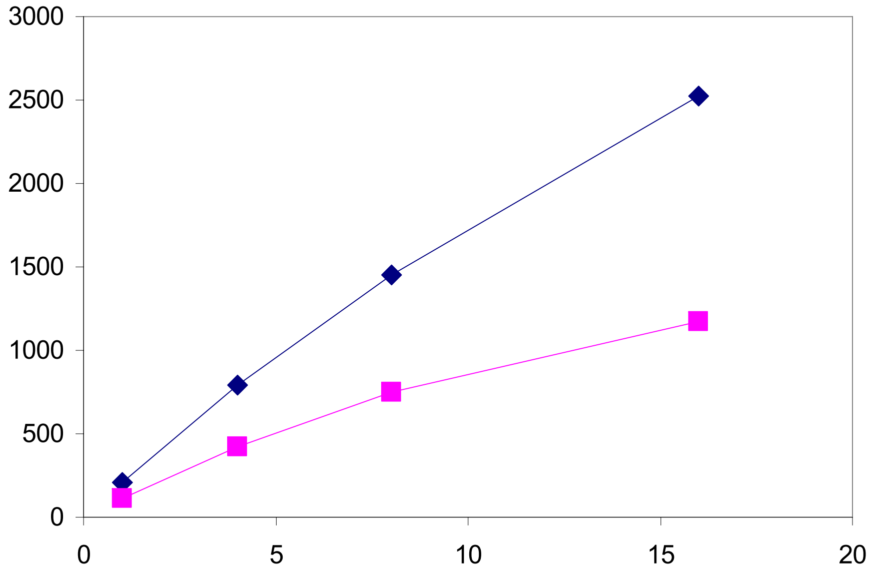
Real Application: Amber8

Amber8 is another chemistry app. These are "sander" benchmarks:

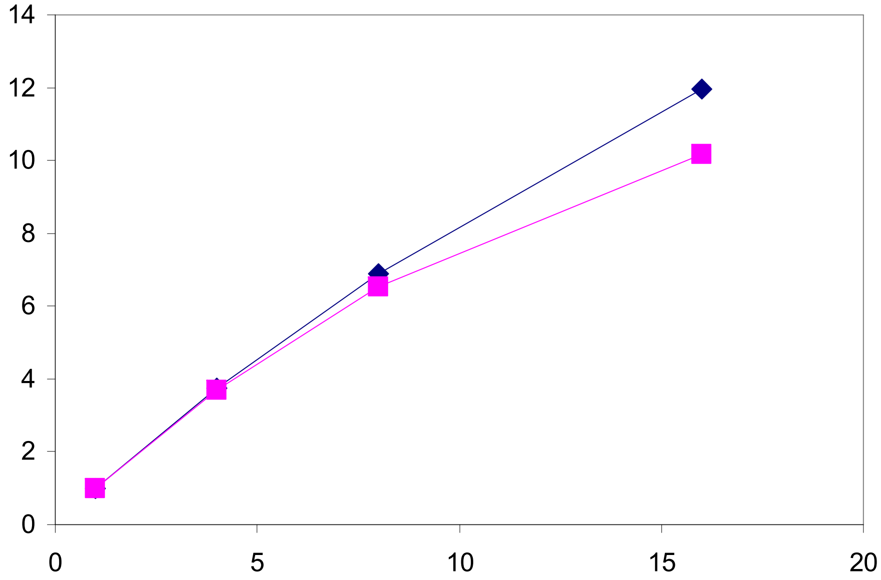
Blue: 2.6 Ghz Opteron + InfiniPath
Pink: 1.4 Ghz Opteron + Myrinet

Higher serial performance means scaling is harder!

Performance



Scaling



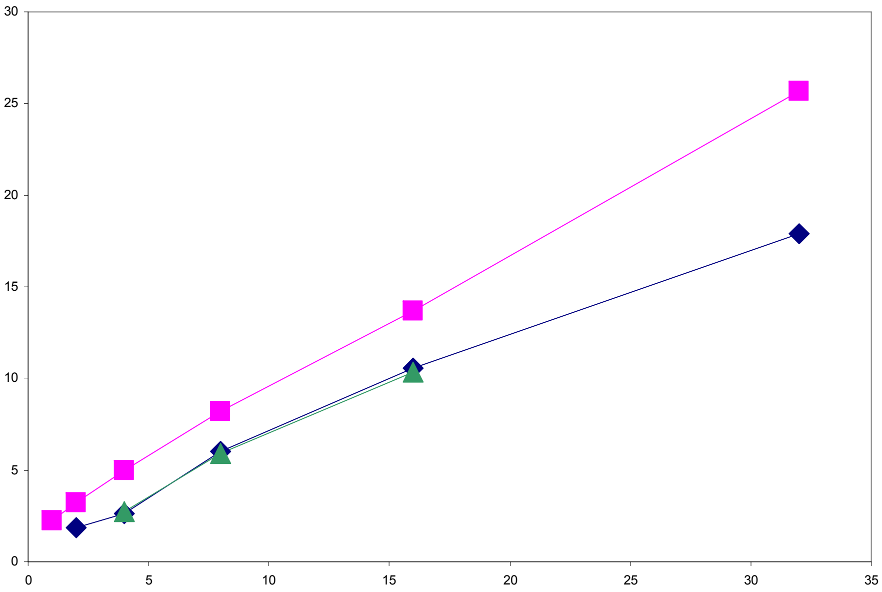
http://amber.scripps.edu/amber8.bench1.html

Real Application: MILC

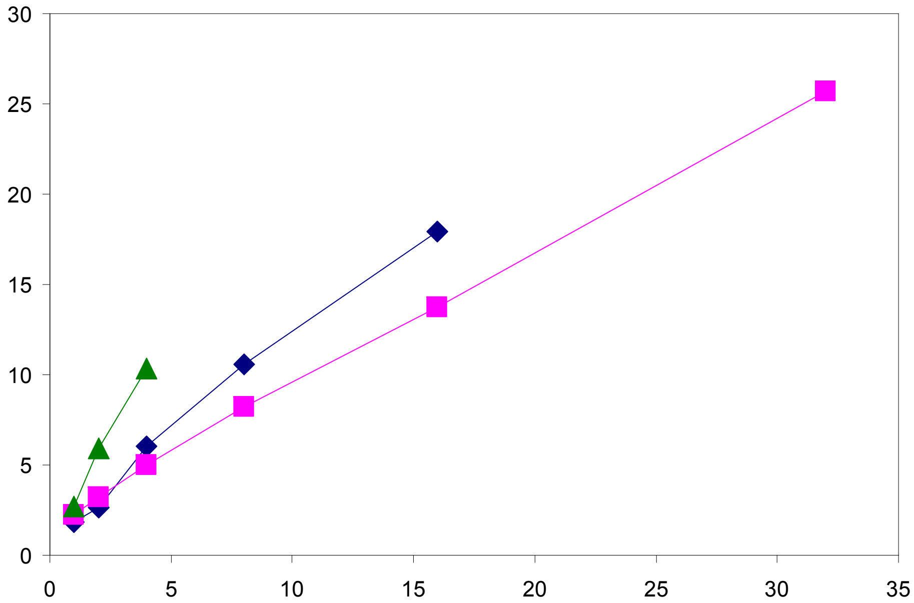
MILC is a quantum chromo-dynamics program. It is known to not scale to 2-cpu nodes on Pentium, hence the comparison (pink) is a single-cpu 3.6 Ghz Pentium4 + Mellanox cluster. Our lines are (dark blue) 2-cpu 2.0 Ghz Opteron + InfiniPath and (green) 2-socket dual-core 2.2 Ghz Opteron + InfiniPath

Note that the P4 has hand-tuned assembly.

Performance by cpu count



Performance by node count



http://physics.indiana.edu/~sg/milc/benchmark.html