AMES LABORATORY

InfiniBand Performance Review

Troy R. Benjegerdes
<troy@scl.ameslab.gov>
Scalable Computing Laboratory

www.scl.ameslab.gov

Motivation

InfiniBand survived the dot-com bust, hardware is available at competitive prices, now how well does it actually work?

Point-to-point works quite well. Some applications work well.

What doesn't work so well.. Deployment, Packaging, Portability

Why does an OS-bypass technology require so much software?

Introduction

- InfiniBand (aka 'IB') works well with vendor-provided software stacks
- Existing Mellanox-based hardware has excellent absolute performance and good price/performance characteristics
- Open Source stacks are still evolving
- Vendor stacks have a lot of 'baggage'

History

- Ottawa Linux Symposium 2001
- InfiniBand HCA's and switch from Mellanox
- Meanwhile, IBM and Intel canceled HCA plans
  - Intel did use IB physical layer signaling in PCI Express (2.5 ghz)
  - IBM went for 12x
  - Intel open-sourced the software stack
    - Intel Sourceforge InfiniBand project (IBAL)

IBAL (Intel Verbs)

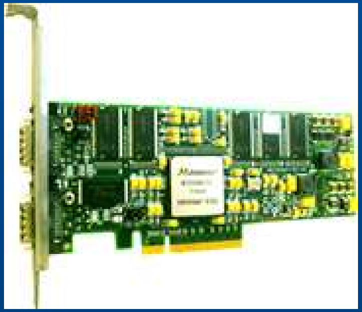
- Open source access layer, and ULP's (aka 'Upper Layer Protocols')
  - IBAL (access layer)
  - SDP (Sockets Direct)
  - OpenSM (Subnet Manager)
  - IPoIB (IP packets over InfiniBand)
  - kDAPL
  - uDAPL

IBAL: Where's the driver?

- Got a design, got working code, no hardware
- Mellanox driver was added, but NOT released as open source (Industry politics).
- Several amusing discussions on linux-kernel on (lack of) code quality
- Too big, too portable for Linux?

Available Hardware

- Mellanox
  - 8 port switch chips (4x IB)
  - 24 port switch (4x IB), 8 port 12x IB
  - PCI-X and PCI-Express 4x HCA's



- Agilent
  - 8 port switch chip (4x)
- Fujitsu
  - PCI-X 4x HCA

OEM's/Resellers

- InfiniCon (www.infinicon.com)
  - Mellanox-based HCA's
  - Mellanox and Agilent-based switches
  - I/O (Ethernet and Fibre Channel adapters)
- Voltaire (www.voltaire.com)
  - Mellanox HCA's and Switches
  - I/O (Ethernet and FC)
- Topspin (www.topspin.com)
  - Mellanox HCA's and Switches
  - I/O (Ethernet and FC)

SuperComputing 2002: InfiniBand strikes back

- New firmware and Mellanox drivers (THCA.. Tavor HCA) pushed peak bandwidth to over 6 Gigabits per second
- Several vendors were on the floor with hardware ready to ship early 2003
- D.K. Panda's group at Ohio State released the first MVAPICH release
- No shipping 10GigE solutions

My stack is better than yours

- Each vendor had a different software stack
  - Proprietary value-add
  - Market differentiation
  - blah blah $MARKETING blah blah
- Meanwhile, two 10 GigE drivers ended up in linux-2.6.6

Oh wait, customers really do like OSS

- April 2004, more source that I knew what to do with
  - Topspin, InfiniCon, Voltaire, Mellanox, and DivergeNet all announced, *and released* portions of their code.
- Shortly thereafter, OpenIB.org was formed
- Early Mozilla OSS days. It's there but.. errr, do you really want to do that?

OpenIB.org

- http://openib.org
- Hosted by Sandia California
  - Expected large IB End-user
- Collaboration of InfiniBand vendors, end-users (DOE, and others), OEMS
- Goal is to provide high-performance IB support for Linux
  - OEM's are interested in other OS'es

OpenIB challenges

- Put the code on a diet
- Maintain industry consensus
- Application developers like to bypass the OS but want the OS to clean up the mess
- Kernel developers don't like apps bypassing them

Hardware test environment

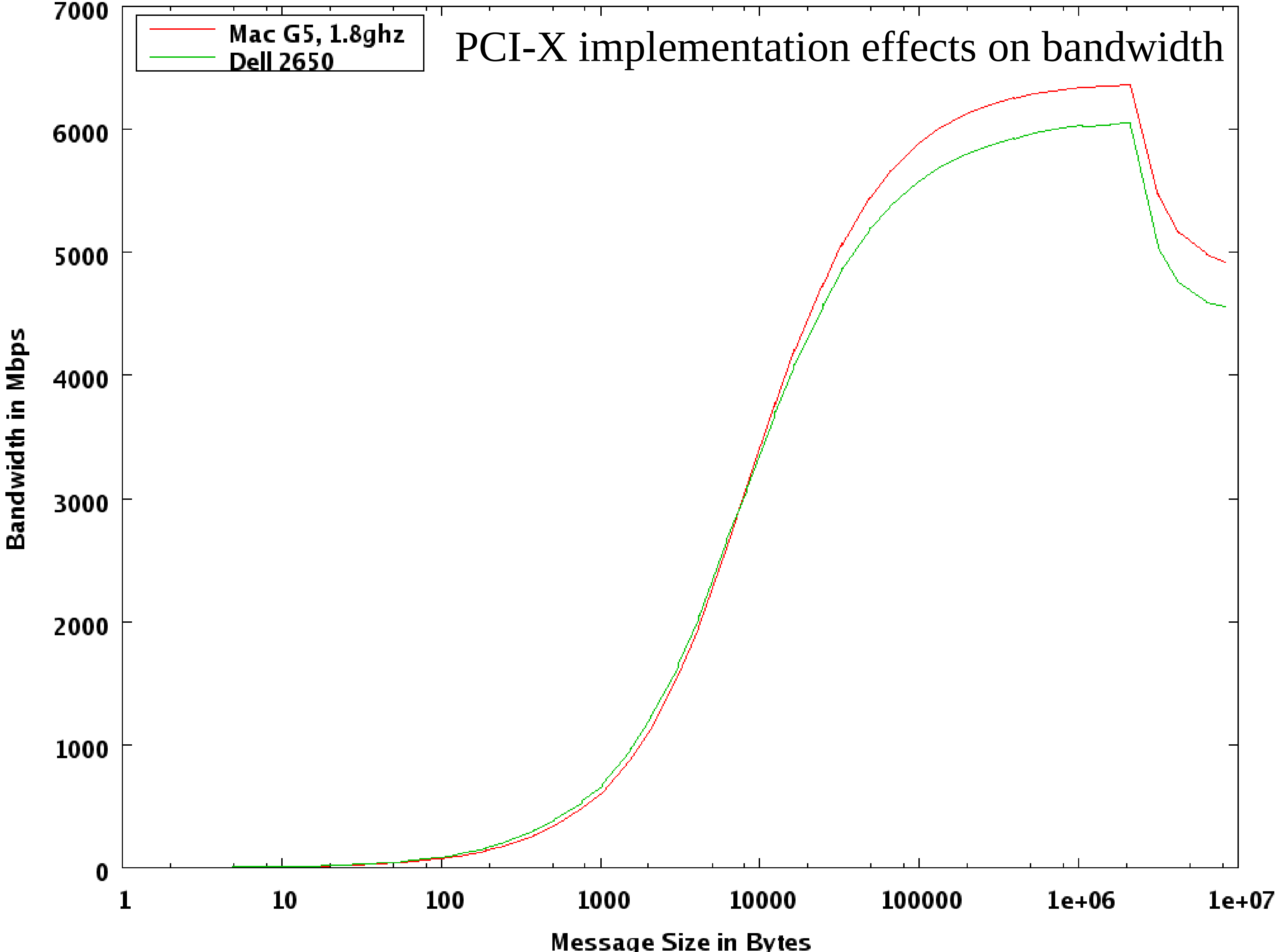
- Dell 2650 2.4 Ghz Xeon
  - Serverworks Grand Champion PCI-X
  - Chipset feature/bug
- Mac G5 (1.8 Ghz, and dual 2 Ghz)
  - AMD 8131 PCI-X bridge
- AMD Opteron 1.4 Ghz
  - RioWorks HDAMA motherboard
  - AMD 8131 PCI-X bridge

Software test environment

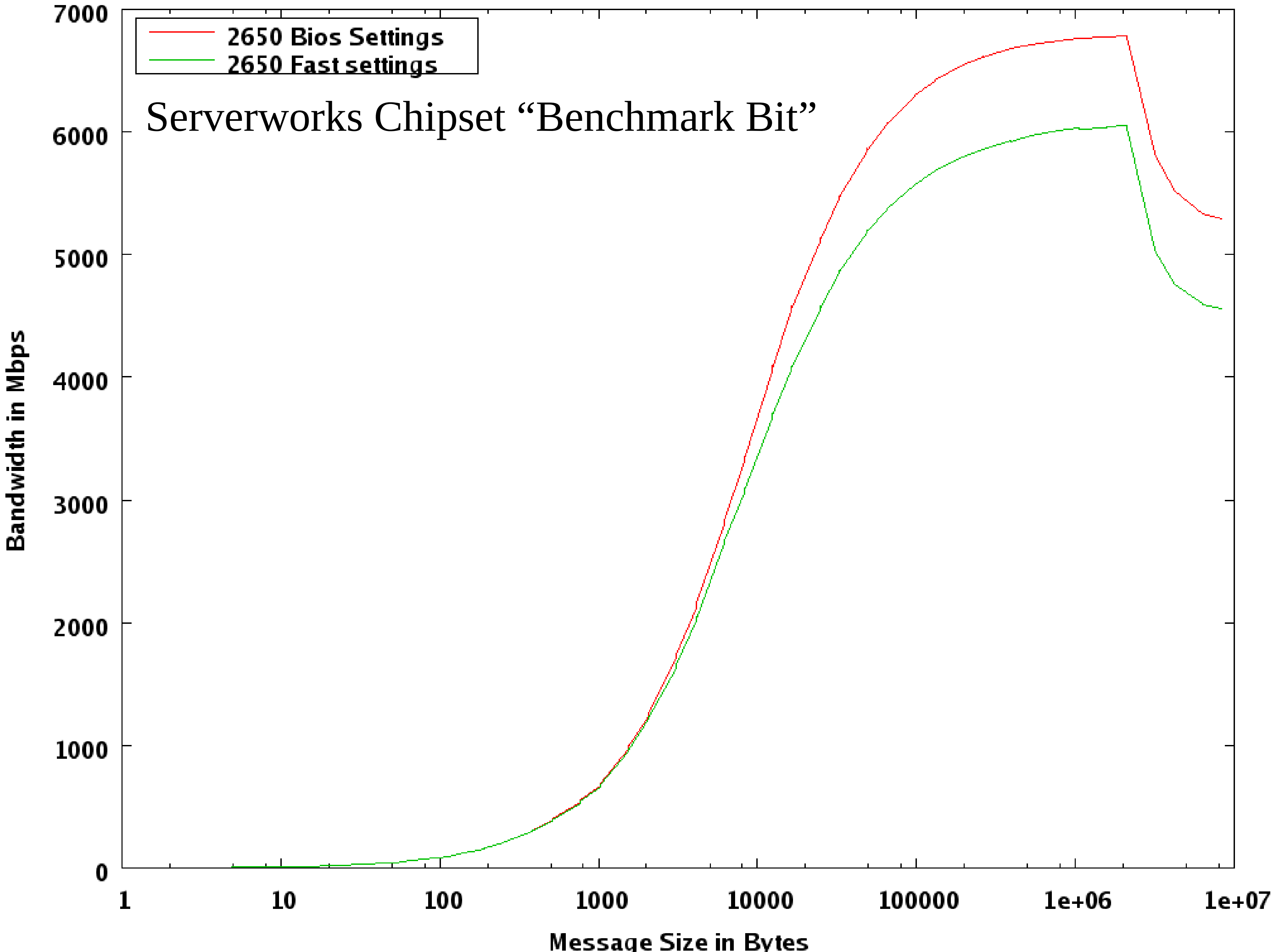
- Nov 2003 Results
  - Mellanox THCA
  - InfiniCon
- May 2004 Results
  - Mellanox THCA pre-release 3.2-rc9
  - Linux 2.6.5 kernel support
  - No PPC64
  - OpenIB.org appears usable, not enough time to test

MPI testing

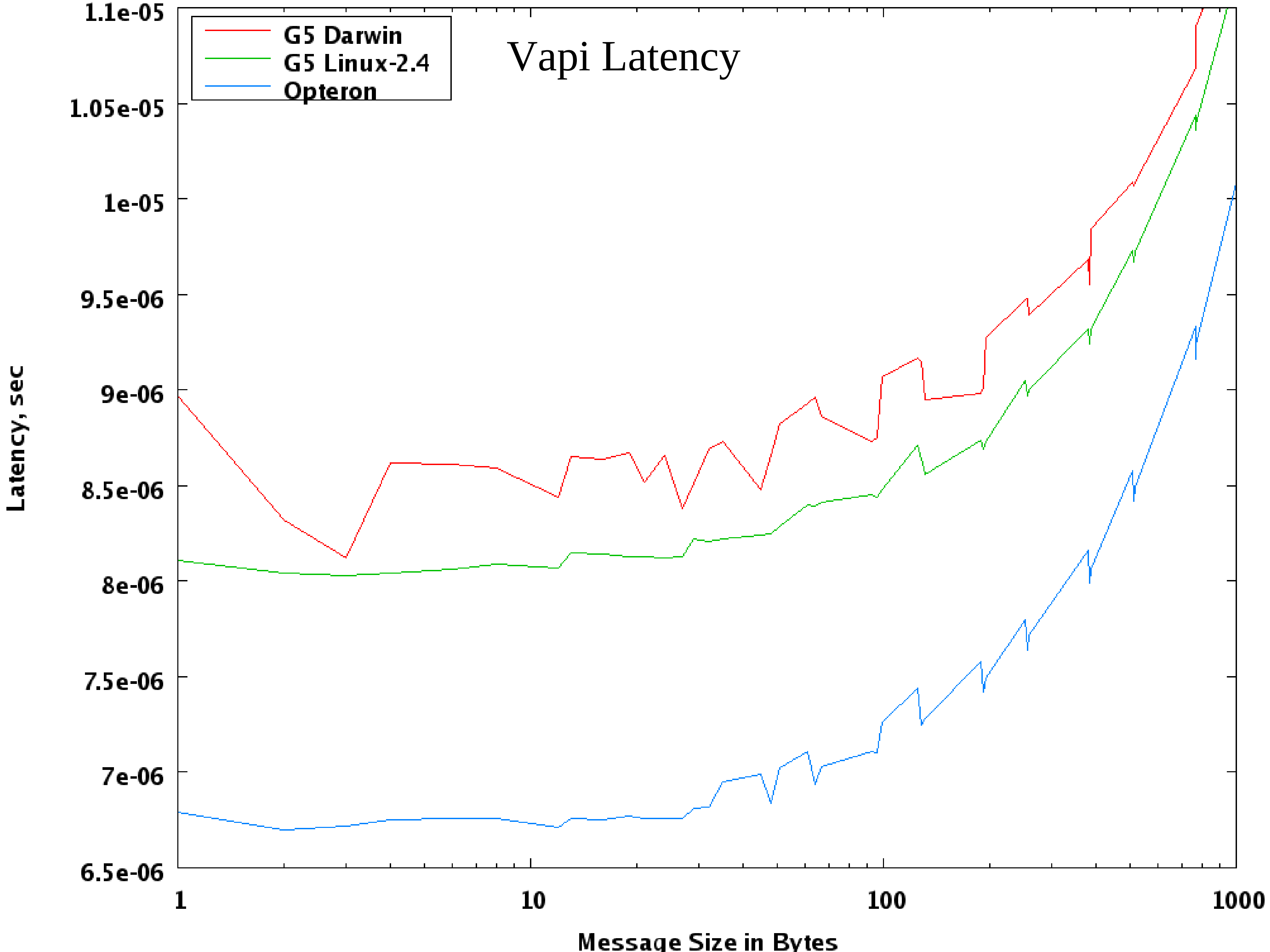
- Versions tested:
  - OSU-0.9.2 MVAPICH patches to MPI-1.2.5
  - LAM-MPI development version
    - April 28 checkout from Subversion repository
  - InfiniCon MPI from November 2003
    - Used MVAPICH as a base
    - different tuning

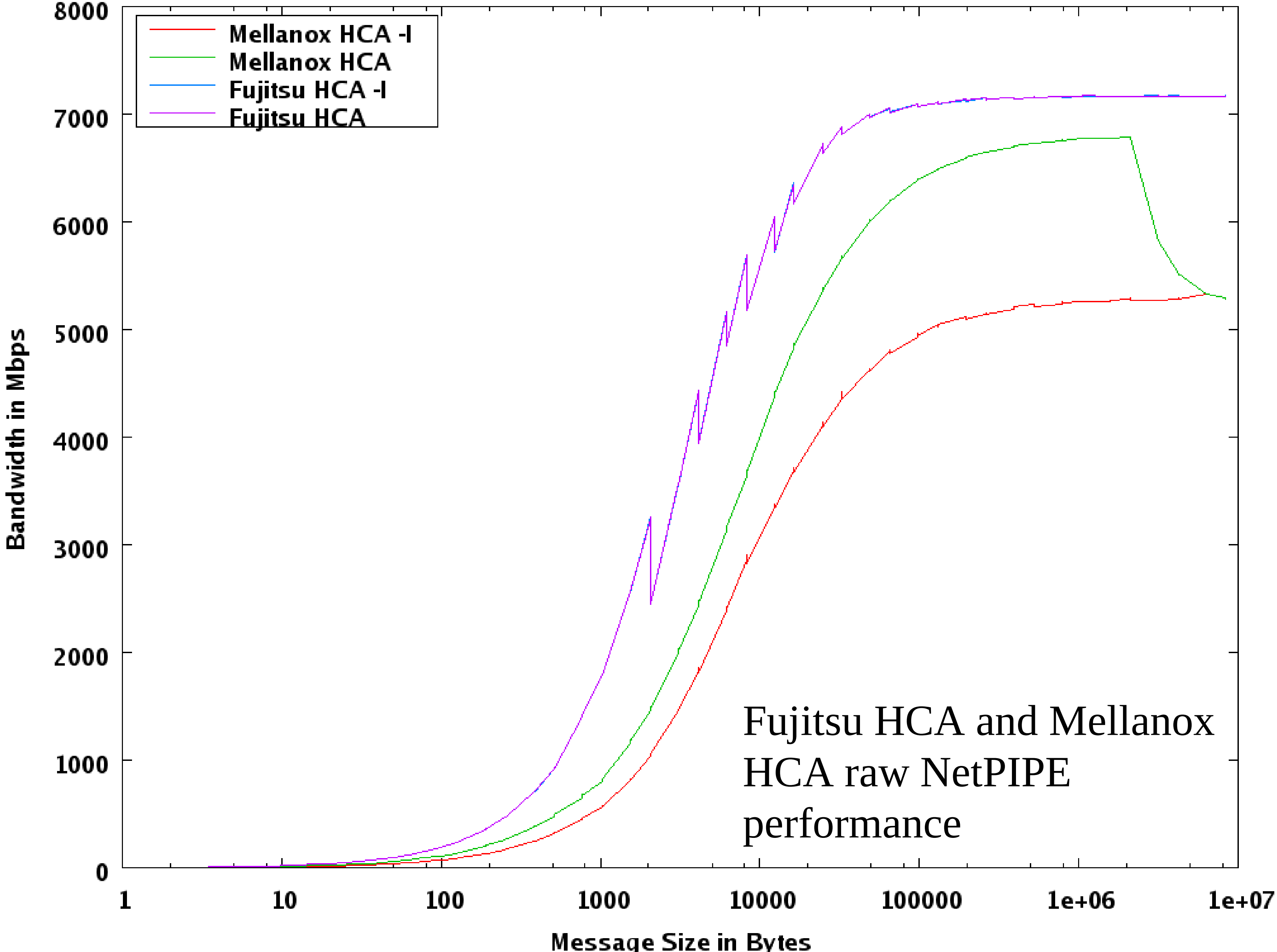


PCI-X implementation effects on bandwidth

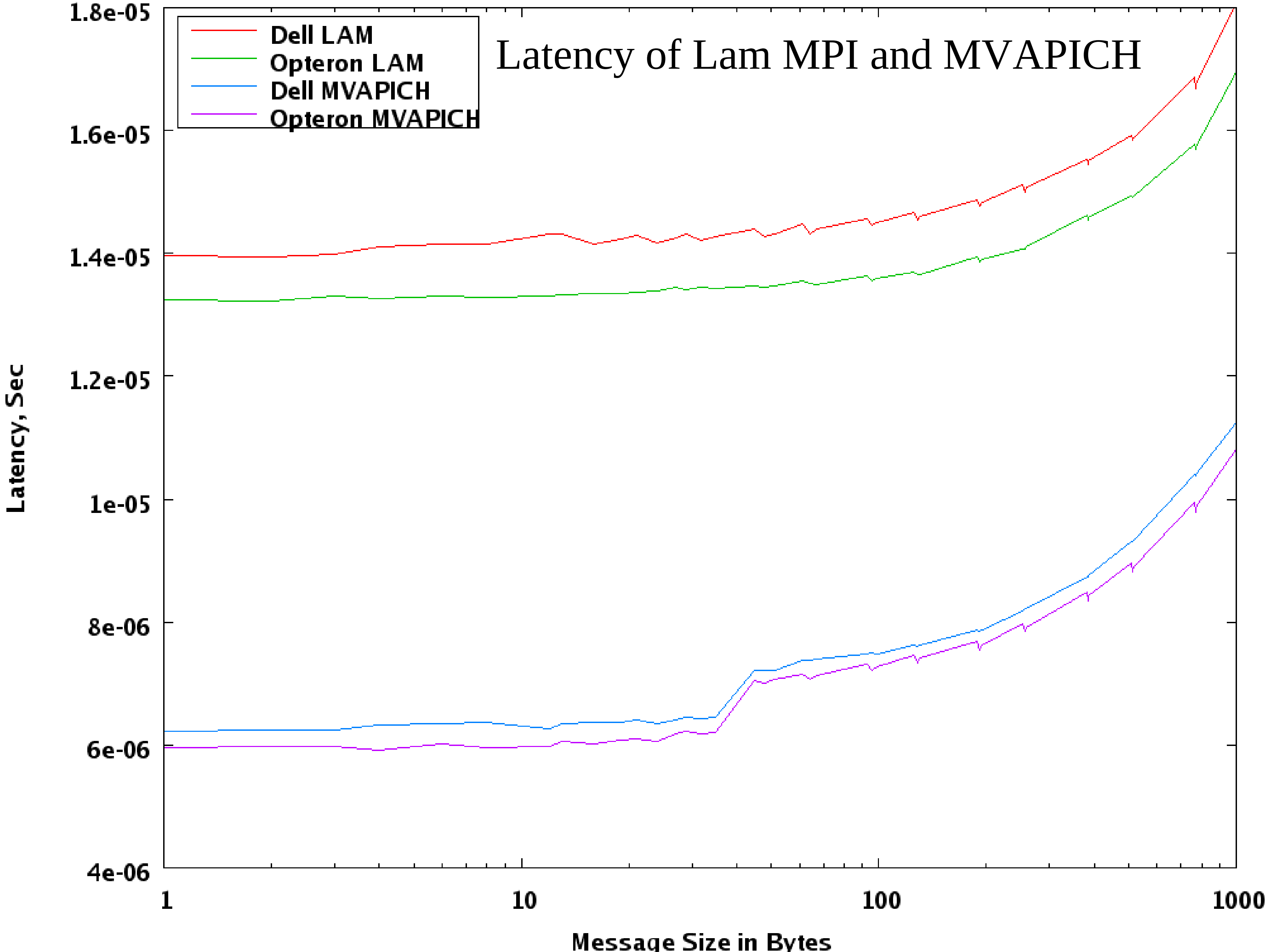


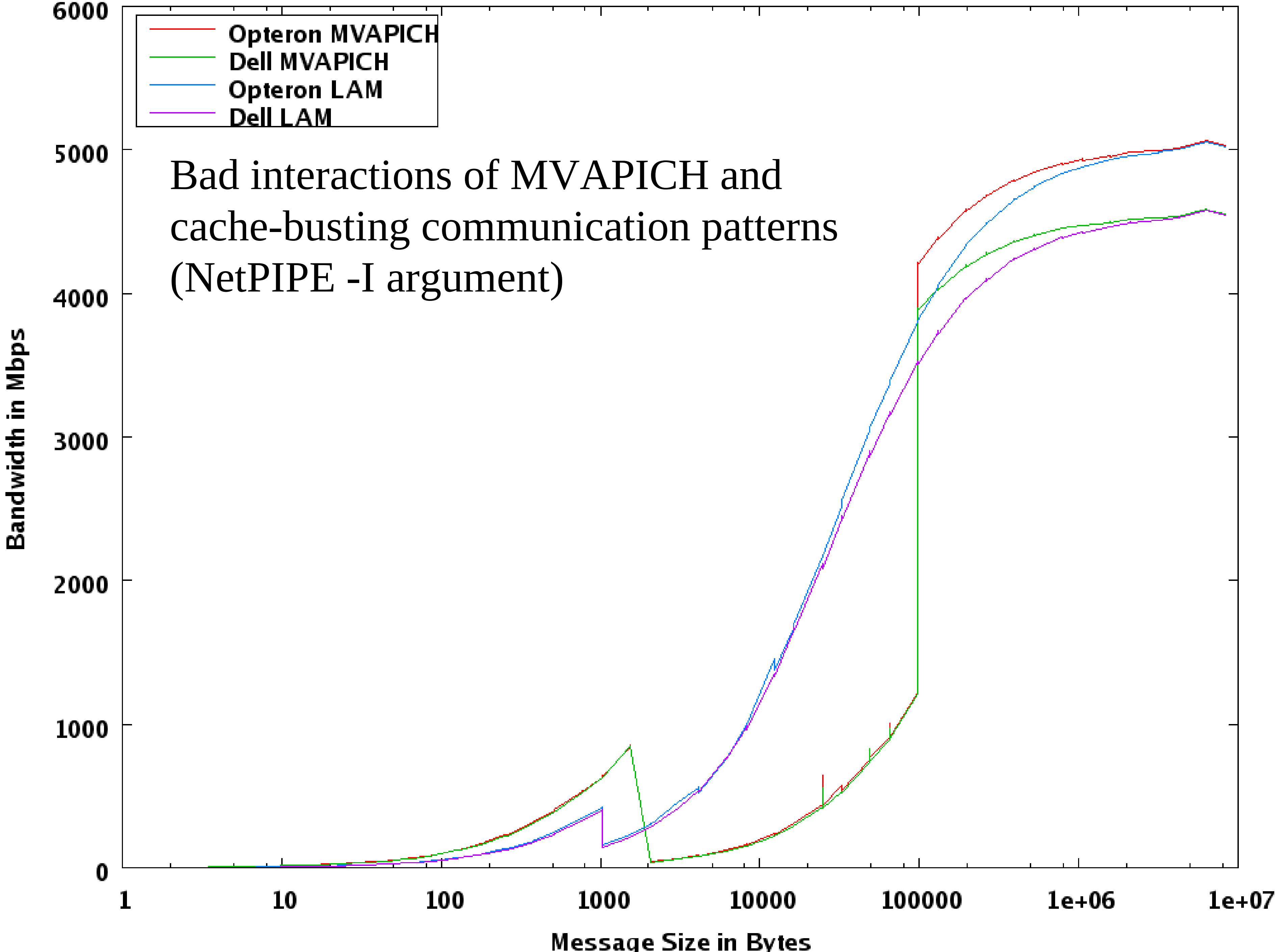
Serverworks Chipset “Benchmark Bit”


G5 Darwin
G5 Linux-2.4
Opteron
Vapi Latency
1.1e-05
1.05e-05
1e-05
9.5e-06
9e-06
8.5e-06
8e-06
7.5e-06
7e-06
6.5e-06
Latency, sec
1
10
100
1000
Message Size in Bytes



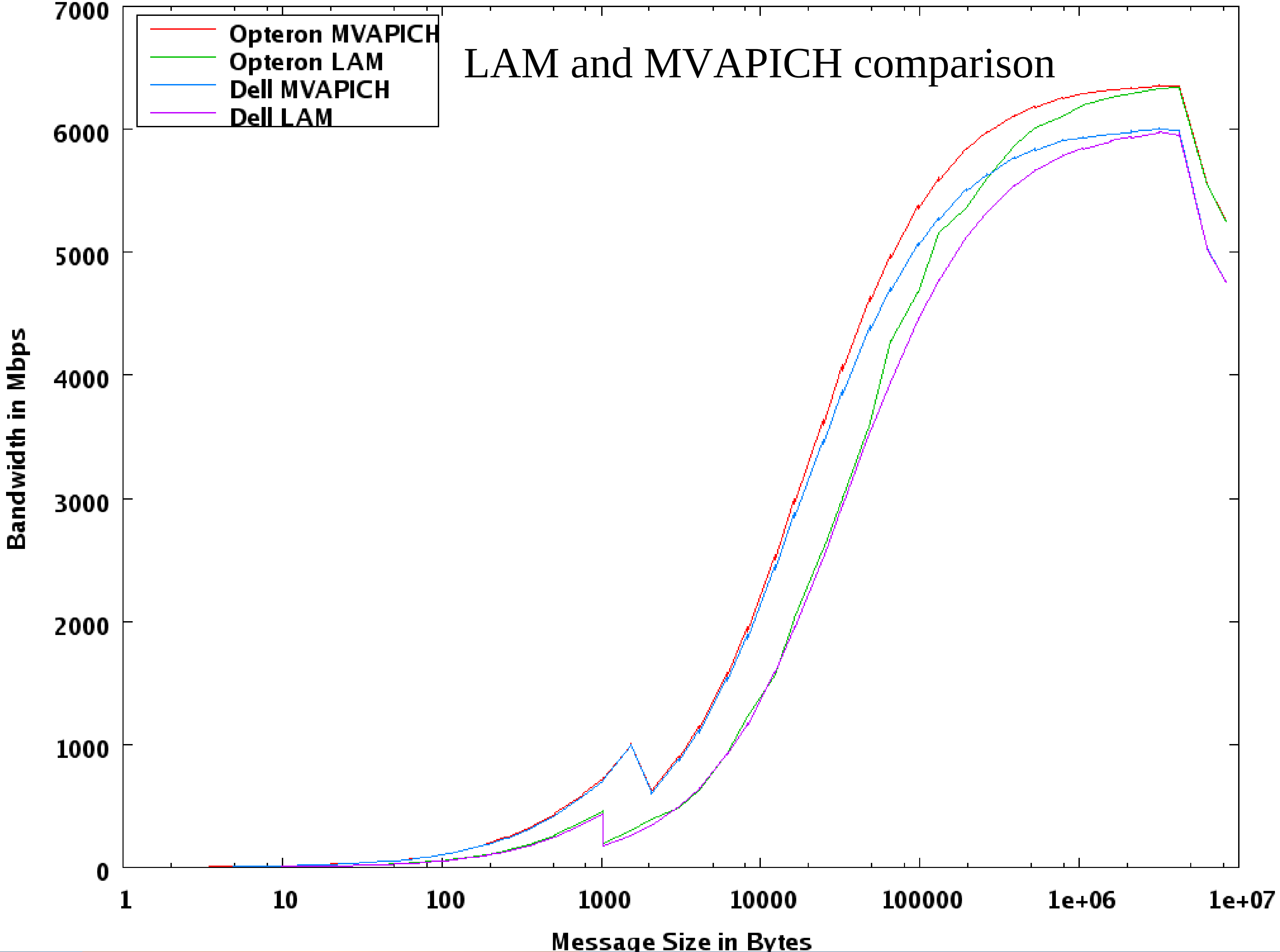
Fujitsu HCA and Mellanox HCA raw NetPIPE performance


Latency of Lam MPI and MVAPICH
Dell LAM
Opteron LAM
Dell MVAPICH
Opteron MVAPICH
Latency, Sec
1.8e-05
1.6e-05
1.4e-05
1.2e-05
1e-05
8e-06
6e-06
4e-06
1
10
100
1000
Message Size in Bytes

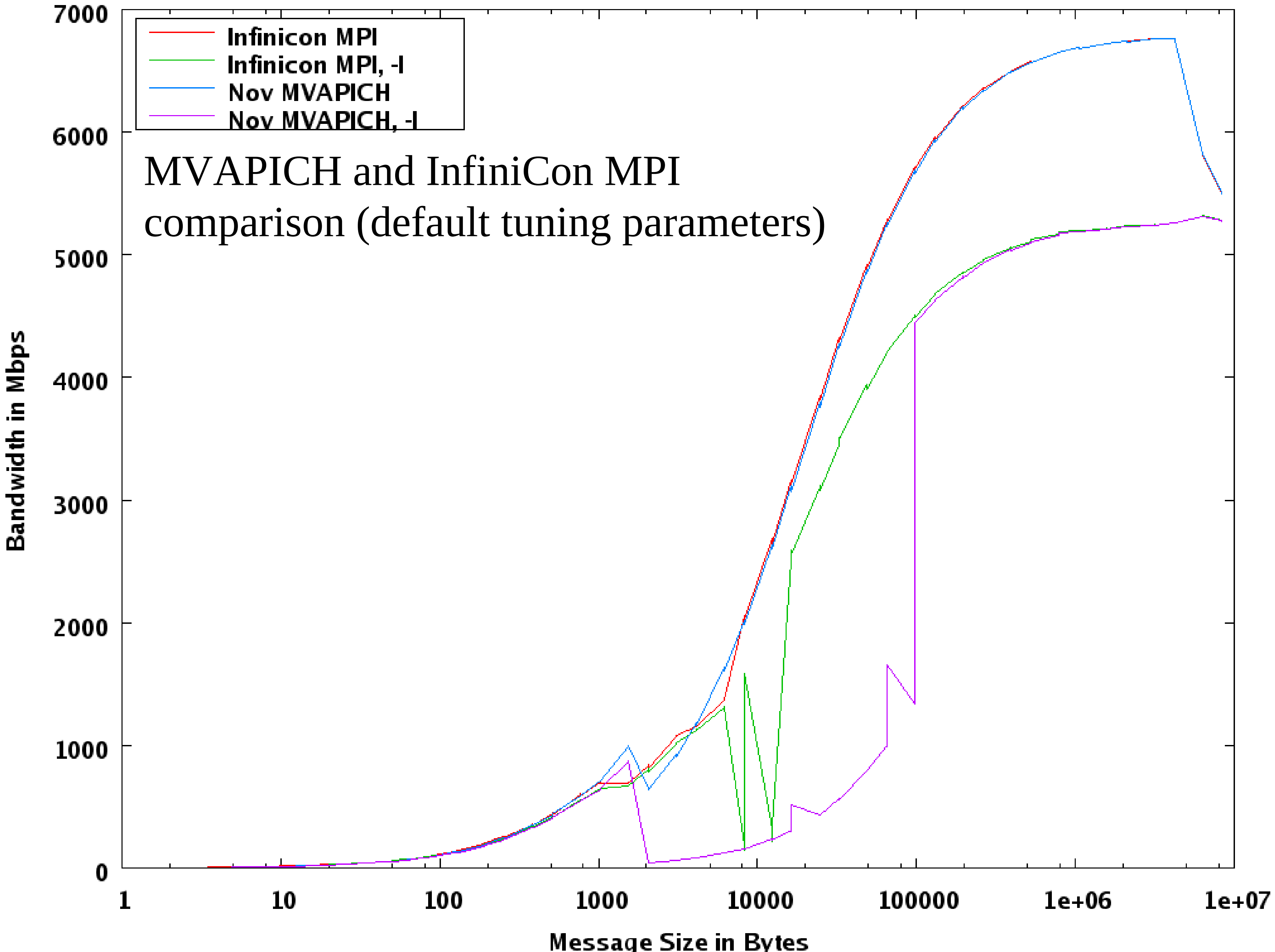


Bad interactions of MVAPICH and cache-busting communication patterns (NetPIPE -I argument)

LAM and MVAPICH comparison



MVAPICH and InfiniCon MPI comparison (default tuning parameters)

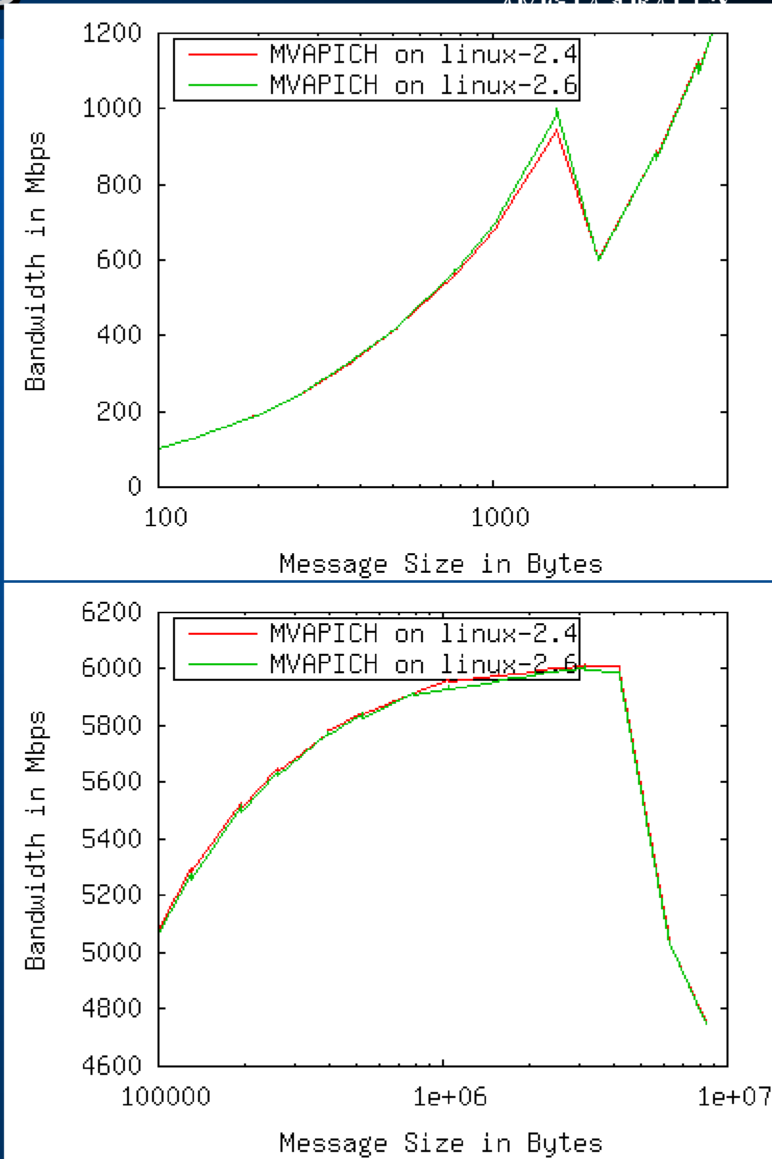


MPI results

- MVAPICH is more mature and tuned
  - RDMA-write and memory polling
    - lower latency than netpipe, which uses POLL_CQ and extra PCI-X cycles
- Tuneing != performance
  - NetPIPE cache-invalidate (-I) option
    - Worst case cache behavior on cpu cache, MPI eager buffers, Mellanox TPT-cache
  - Simpler is more robust

Linux kernel versions

- linux-2.4 vs linux-2.6.5, with Mellanox THCA-3.2-rc9
- Nothing noticeable
- driver bypasses most all OS features, even re-implementing pagetable walking



Conclusions

- Performance is good
  - Can IB compete with other custom cluster interconnects? Probably.
  - Can IB compete with Ethernet??
  - Good IB vendor support for commercial linux distributions (Suse and Red Hat)
- What's missing
  - Open source/Vendor neutral API (OpenIB is beginning to address this)

Commodity Interconnect?

- Hardware is mostly there
  - $750/port HCA cost, $300/port switch cost
  - Cray (XD1) uses 4x IB
  - IBM (pSeries) will use 12x IB as an internal system interconnect

Software story is lacking

- Compared to other commodity interconnects, IB software is
  - more complex (200K lines)
  - less robust
  - requires special knowledge/training
- Compared to other high-performance cluster interconnects...
  - similar complexity

Linux integration

- IB software needs a diet
  - OS-bypass hardware with half an OS worth of driver code
- No IB drivers in 2.6 kernel
- TWO 10 Gigabit drivers in-kernel

RDMA issues

- Interactions with MM subsystem
  - Kernel developers don't seem to fully understand RDMA
  - Hardware (read IB vendors) don't fully understand virtual memory voodoo
- Security
  - If you thought buffer overflows were a problem now...

Research areas

- Demand-paging for registered memory
    - very non-trivial
    - requires part of the kernel MM subsystem to run on the InfiniBand hardware
    - very race-prone
    - May be the best way to resolve application programmer issues

Thank you

- $$$’s
  - U.S. Department of Energy
  - NSF
- Hardware
  - InfiniCon
  - DivergeNet
  - Mellanox
- People
  - Brett Bode
  - Jeff Kirk (Mellanox)

Questions?

?