German National Research Center for
Information Technology
Research Institute for Computer Architecture

Experiences with a SMP PC Cluster

Antônio Augusto Fröhlich
guto@first.gmd.de

GI Meeting
February, 1998

GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

Outline

- The SNOW Project
- SNOW Hardware
- SNOW Software
- SNOW Performance
- Conclusions
- Alternatives

GMD FIRST

German National Research Center for Information Technology
Research Institute for Computer Architecture

The SNOW Project

- SNOW = Scalable Network of Workstations
- Operating Systems Group at GMD-FIRST
- Goals
  - Investigate and develop solutions to diminish the gap between expensive custom parallel machines and cluster of commodity workstations.

GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

The SNOW Project

- Motivation
  - Workstations performance is equivalent to MPP nodes.
  - Networks are getting closer to MPP interconnection systems.
- Challenge
  - Reach performance comparable to MPPs.
  - Software matter.

GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

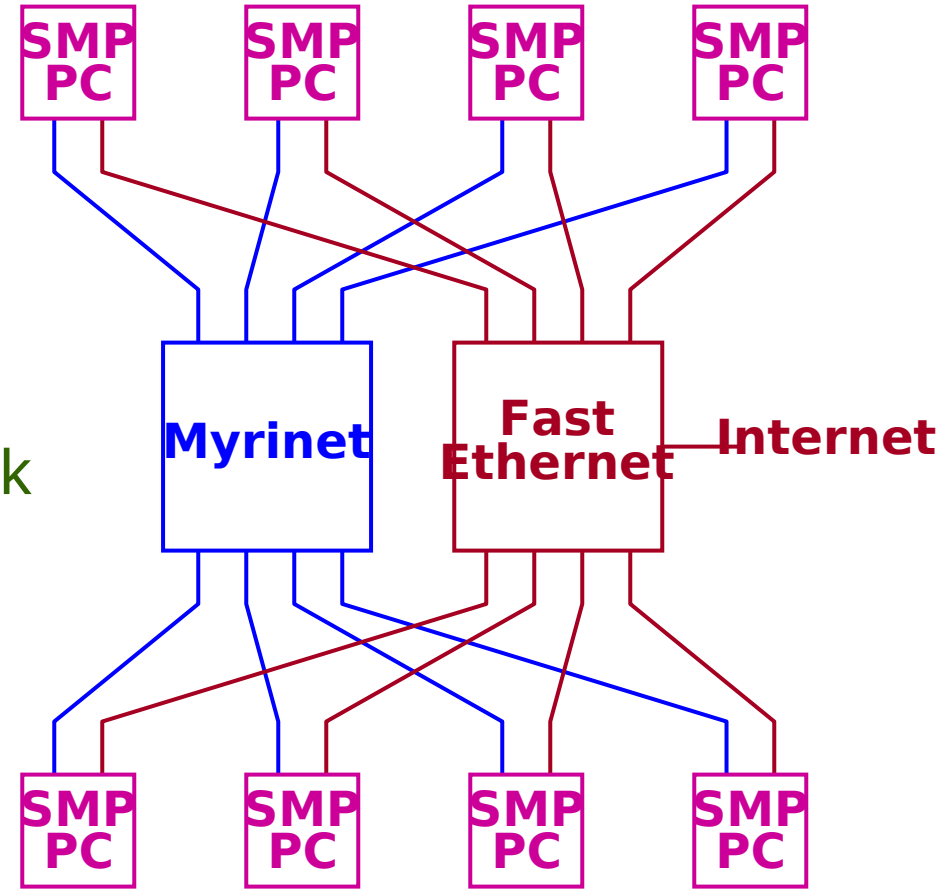
The SNOW Project

- Current stage
  - Evaluation of commodity SMP PC hardware and software to determine the lower edge of this technology.
  - How far from HPC systems in SNOW now?
  - Would careful software design fulfill the gap?

GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

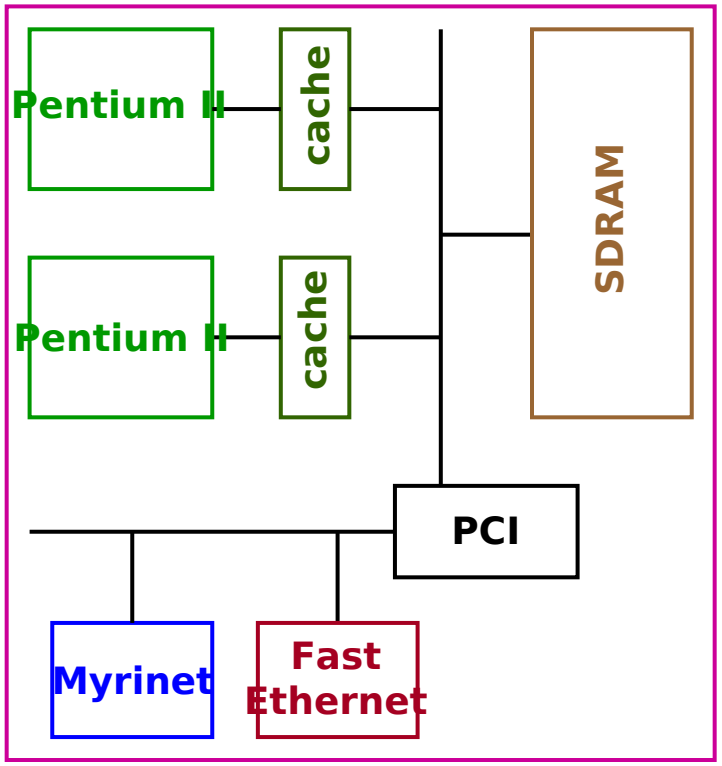
SNOW Hardware

- 8 SMP PCs
- Myrinet network
- Fast Ethernet network



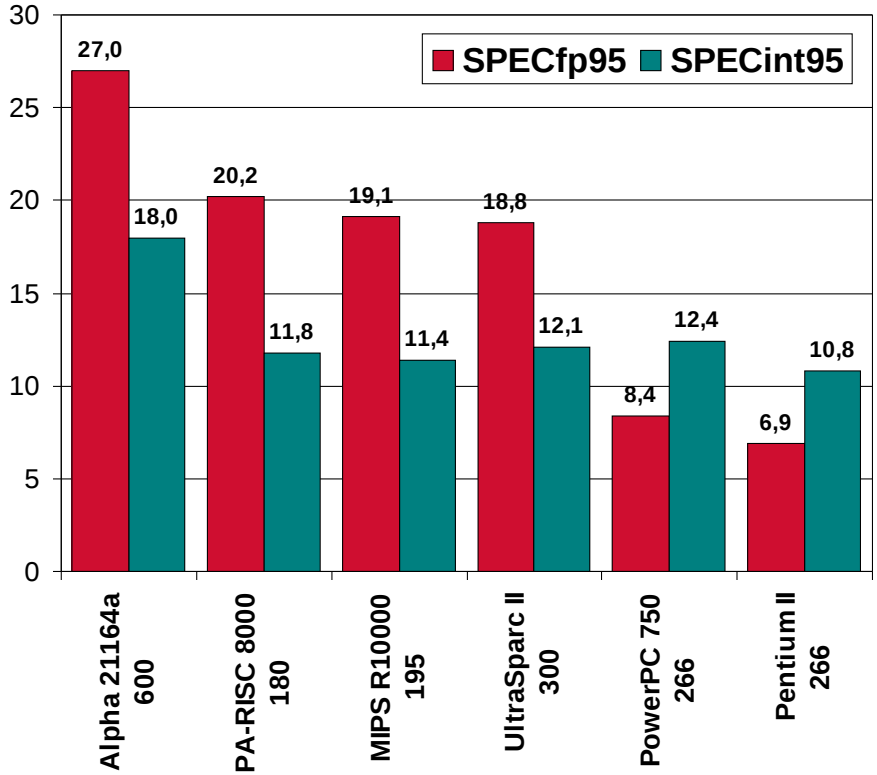
SNOW SMP PCs

- 2 x Pentium II 266 MHz
- 512 Kb L2 cache
- 128 Mbytes 10 ns SDRAM
- PCI bus
- Myrinet NIC
- Fast Ethernet NIC



Pentium II

- Intel P6 family
  - Branch prediction
  - Data flow analysis
  - Speculative execution
  - 14 stages pipeline
  - 2 x IU
  - 2 x FPU
- Max 2-way SMP (Slot1)
- **ix86** macro-architecture
  - CISC
  - 6 registers



GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

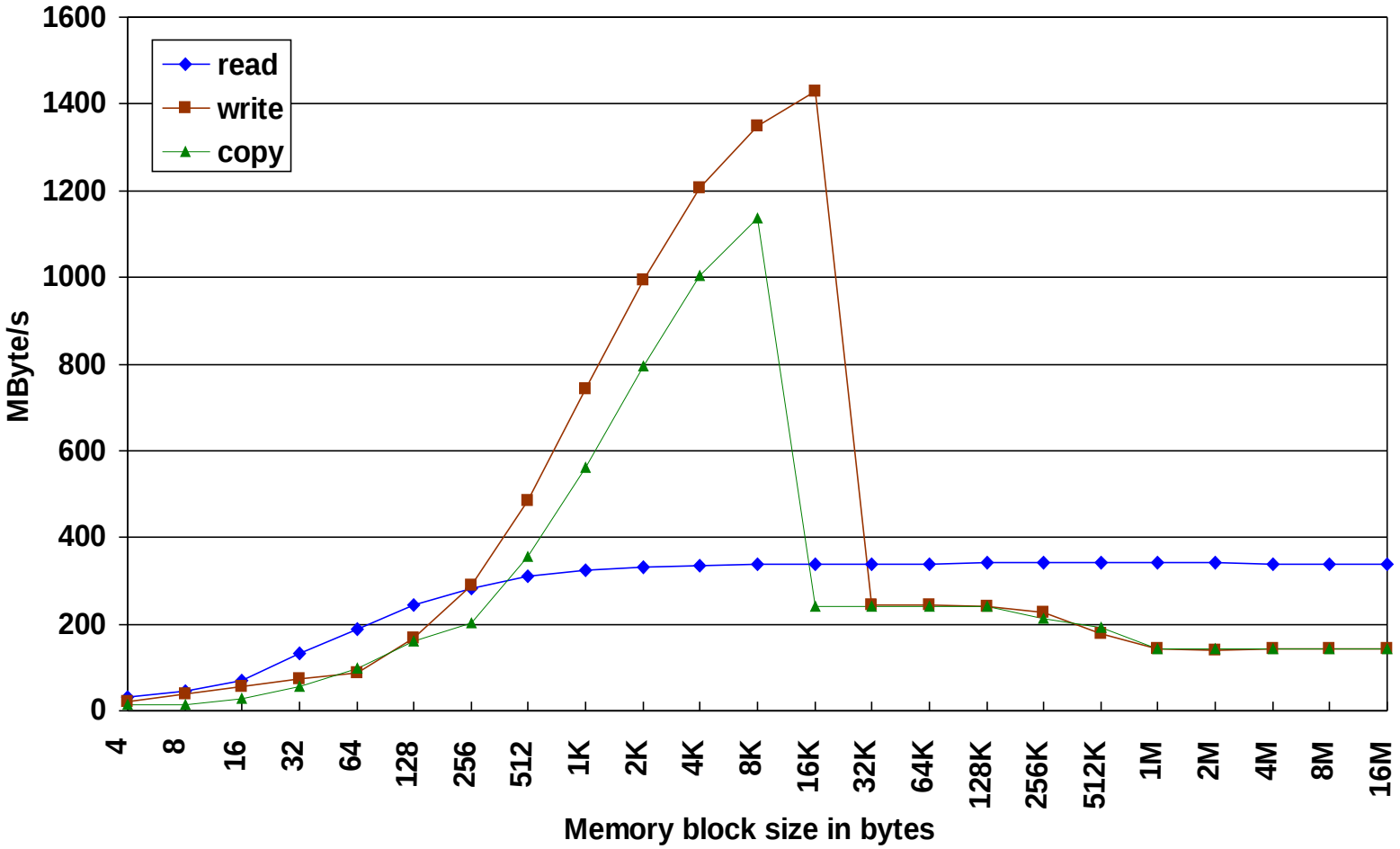
Memory

- L1 cache
  - 16 KB data
  - 16 KB code
  - 4256 MB/s
- L2 cache
  - 512 KB
  - 1064 MB/s
- SDRAM
  - 128 MB / 10 ns
  - 64 bits / 66 MHz
  - 422.4 MB/s

- Useful bandwidth
  - Read => 340 MB/s
  - Write => 143 MB/s
  - Copy => 140 MB/s
  - Stream => 145 MB/s
- SMP splits the bandwidth

GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

PC Memory Bandwidth (cache hit)



PCI Bus

- PCI bus
  - 32 bits
  - 33.33 MHz
- Bandwidth
  - DMA => 127 MB/s
  - PIO => 44.4 MB/s

- Considerations
  - DMA
    - Contiguous allocation
    - Locked pages
    - Physical address
    - Copy to buffer?
  - PIO
    - Logical address

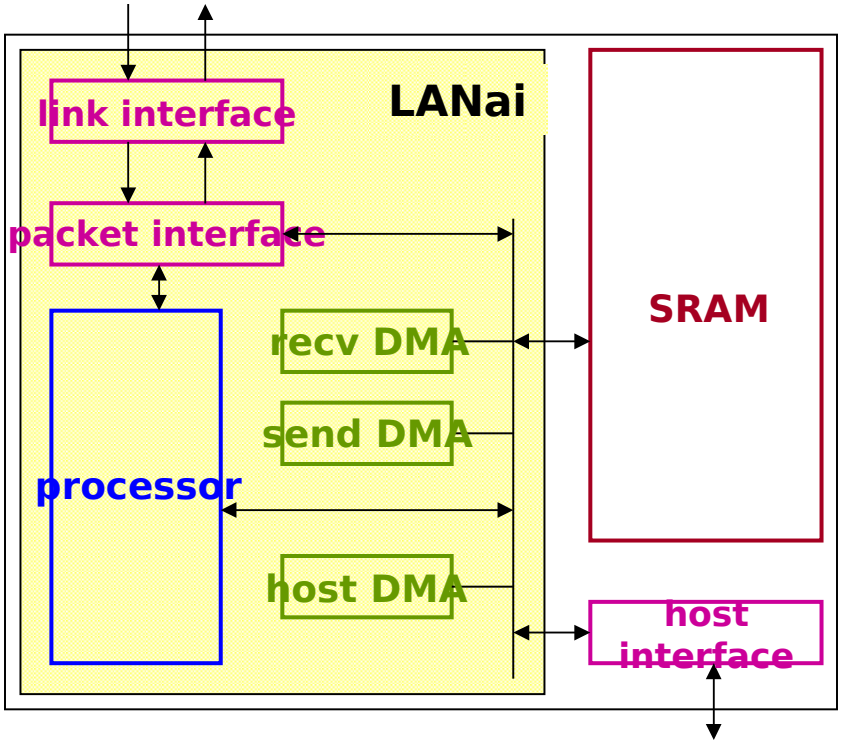
GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

SNOW Networks

- Fast Ethernet (100 Mbits/s full-duplex)
  - TCP/IP
  - File service (NFS)
  - Internet
- Myrinet (1.28 Gbits/s full-duplex)
  - HP applications
  - Free protocols and abstractions

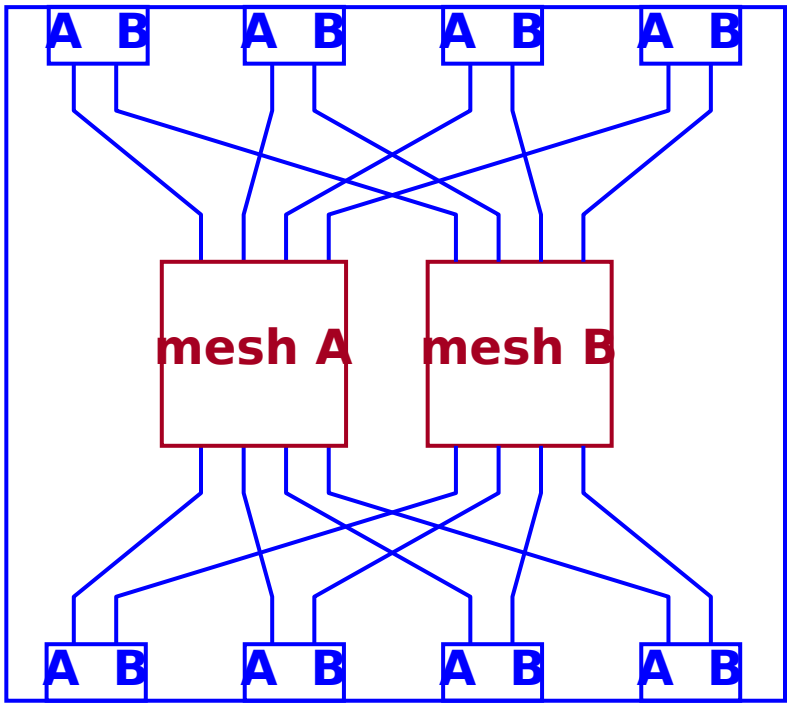
Myrinet NIC

- Myrinet NIC
  - LANai 4.1
    - 32 bits / 37.5 MHz RISC processor
    - 3 x DMA engine
    - link / packet int.
  - 1 Mbyte SRAM
  - PCI host interface
- 160 MB/s full-duplex
- CRC + flow control
- Very flexible



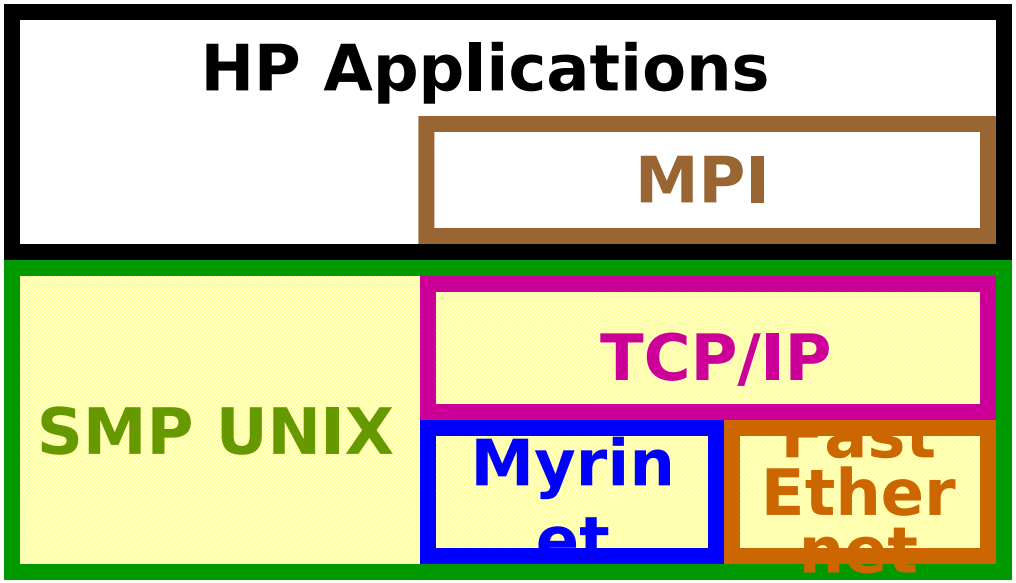
Myrinet Switch

- 8 x Ports
- 2 x crossbars
  - 8 x 8 mesh
  - 0.5 µs per hop latency
  - Worm-hole routing
- Flow control
- Arbitrary length packages
- Free topology



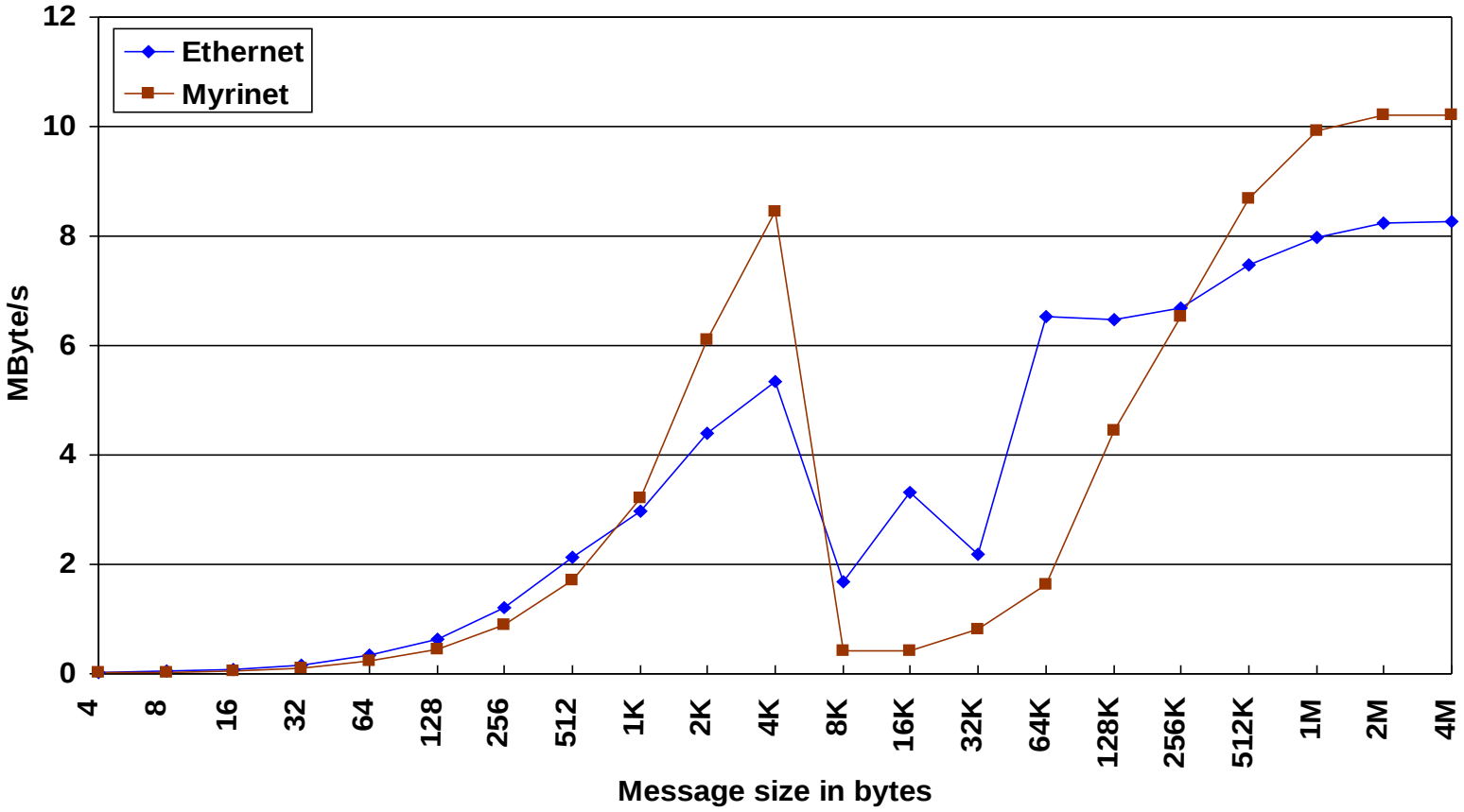
GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

SNOW Software

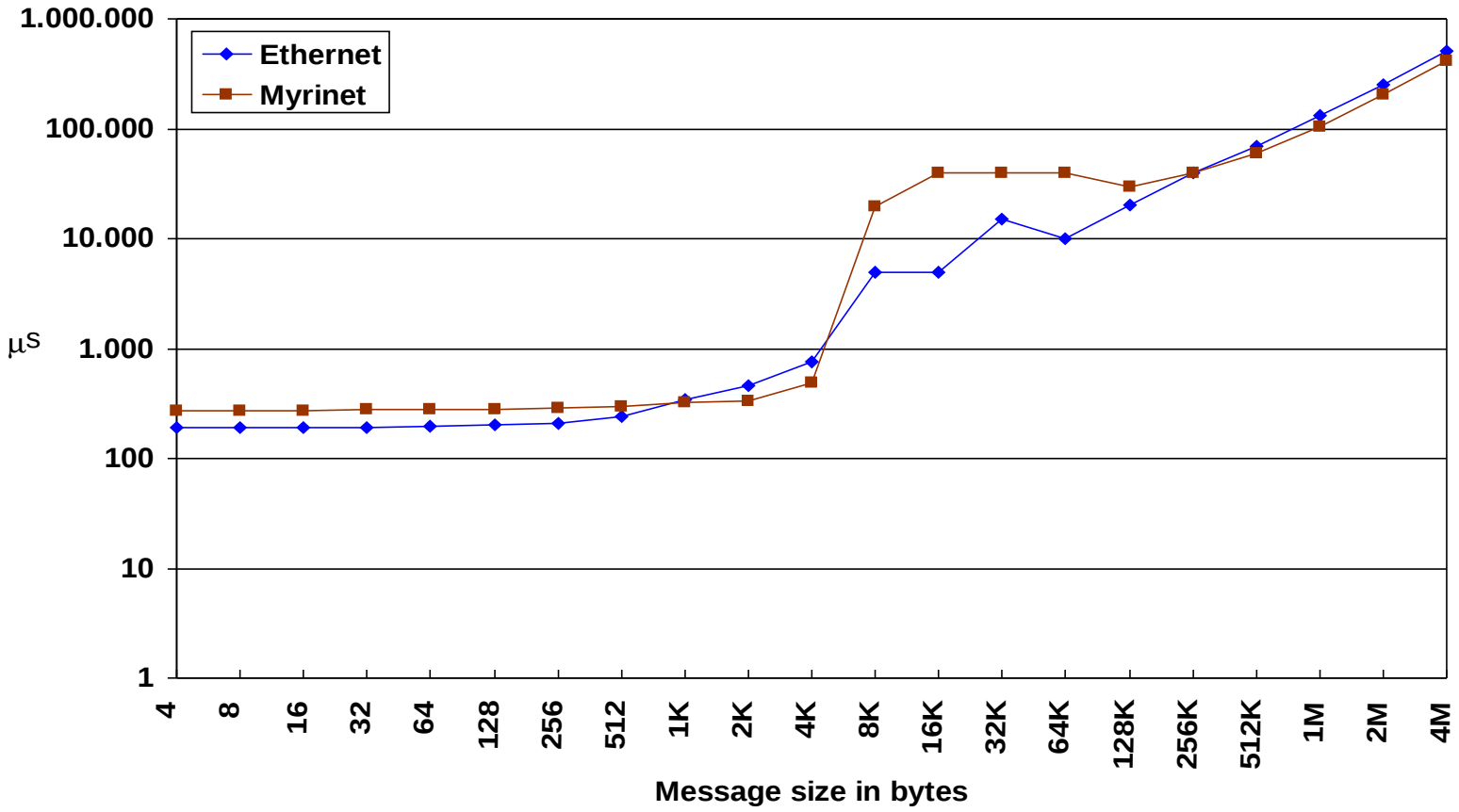


GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

MPI Bandwidth

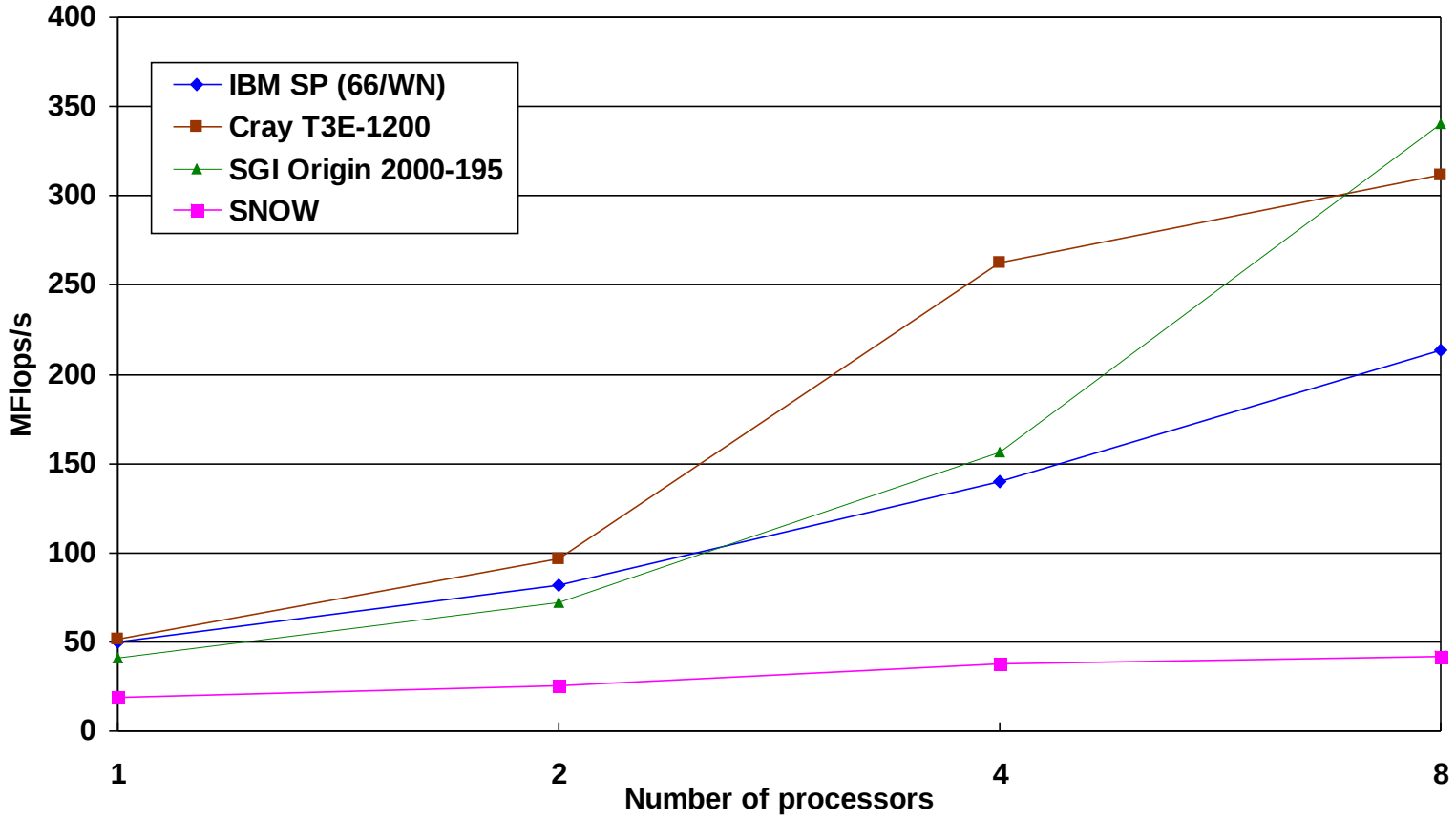


MPI One-way Latency



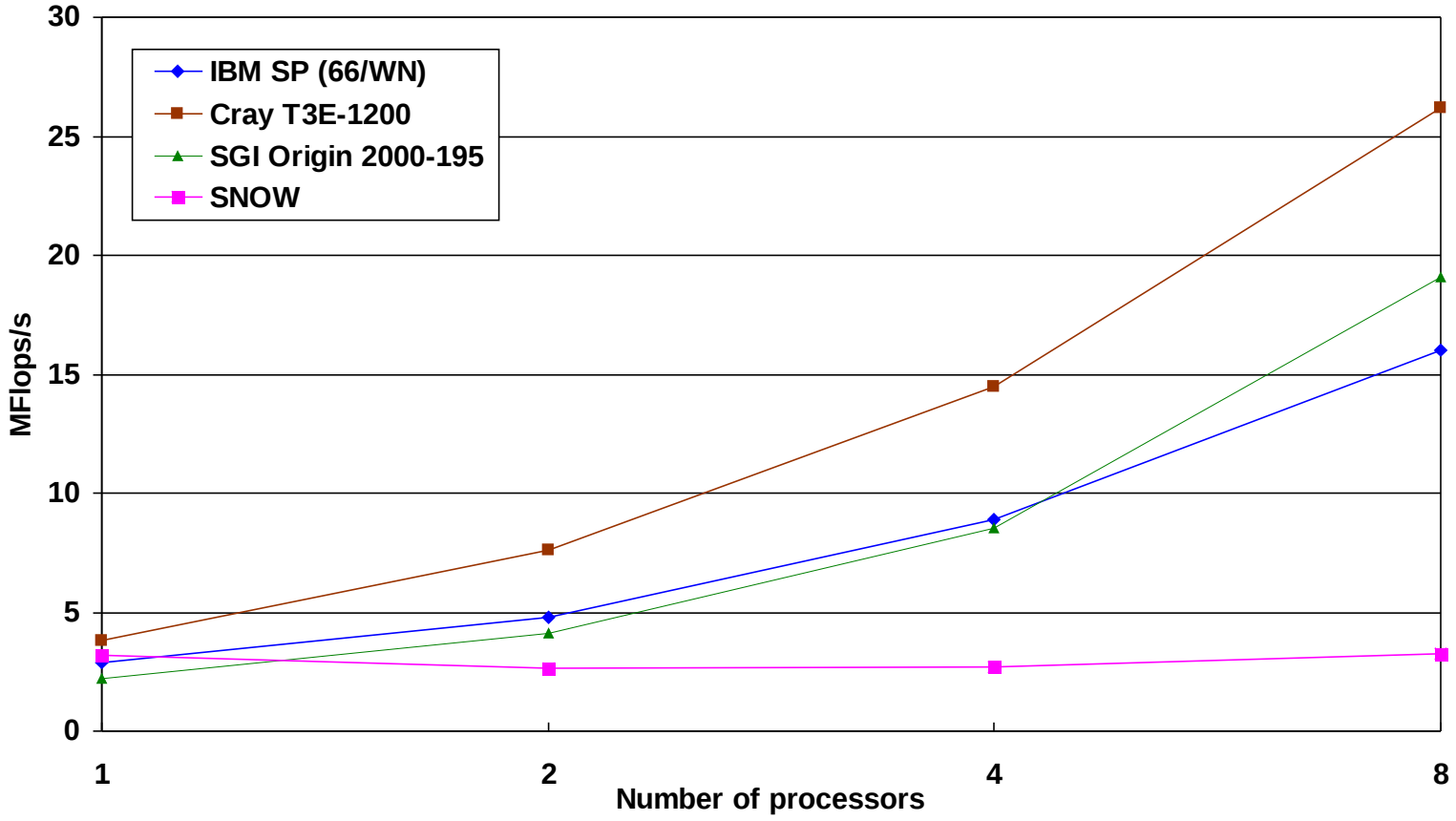
GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

NAS PB 2.3 Conjugate Gradient

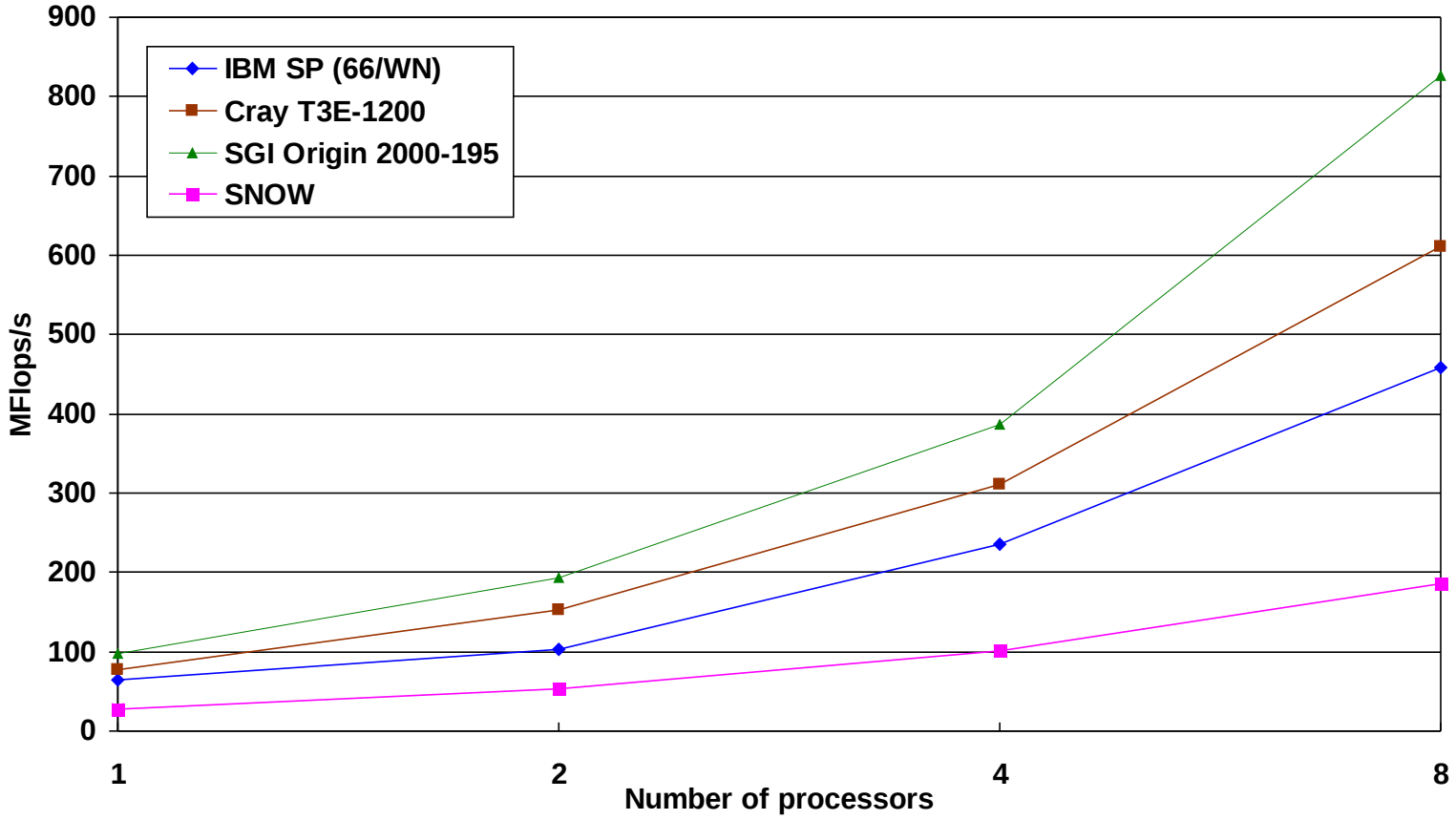


GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

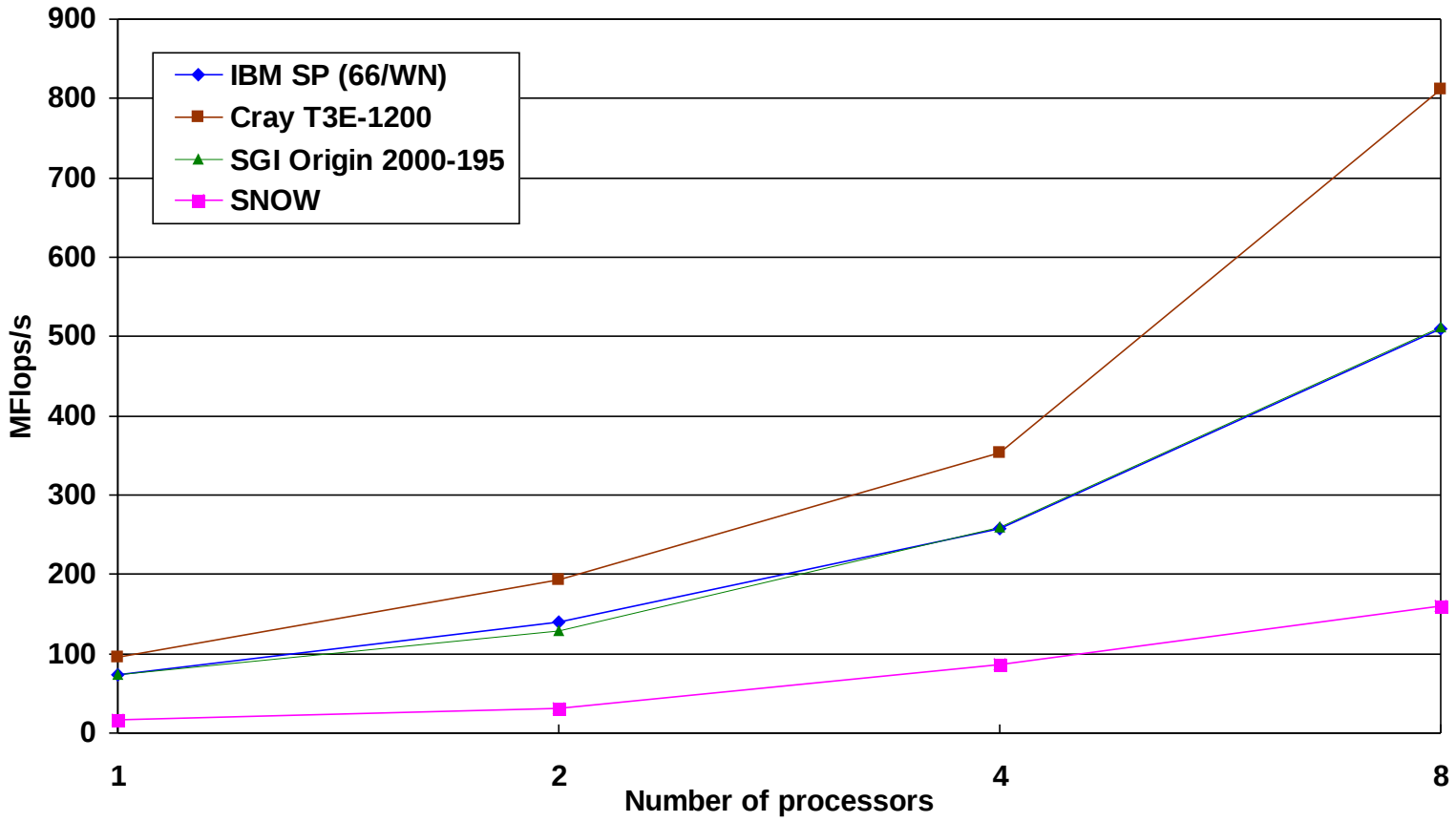
NAS PB 2.3 Integer Sort



NAS PB 2.3 Laplace Solver



NAS PB 2.3
3D Multigrid



Hardware Conclusions

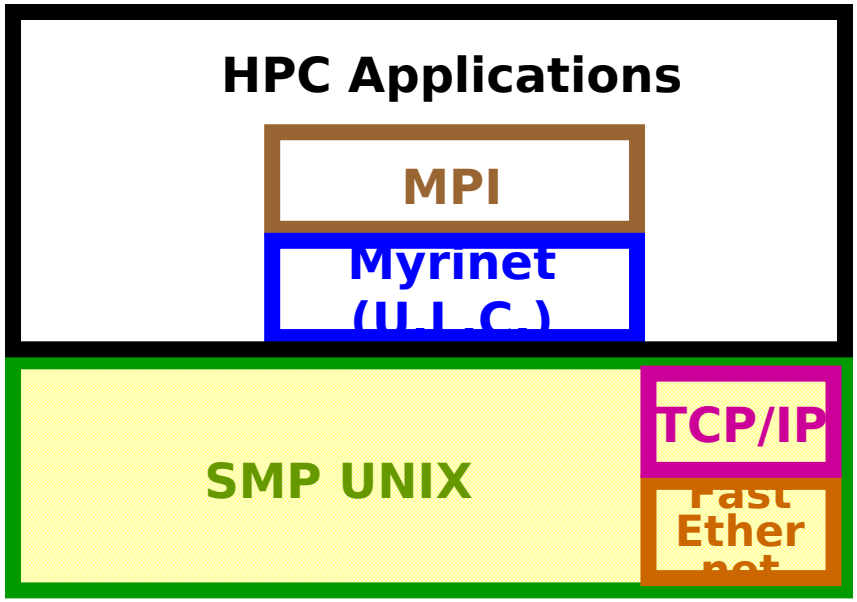
- SUN and SGI opened the way
- PCs
  - L2 cache at CPU speed or broader path
  - 100 MHz memory bus => 640 MB/s
  - 64 bits 66 MHz PCI bus => 528 MB/s
  - A matter of months (days)!

Software Conclusions

- UNIX is sometimes of big inconvenience
- Very high communication overhead
  - System calls
  - TCP/IP
  - Secure multiplexing
- Inadequate memory and I/O management
  - Paging (physical or virtual)
- Poor SMP support
  - Lack of control

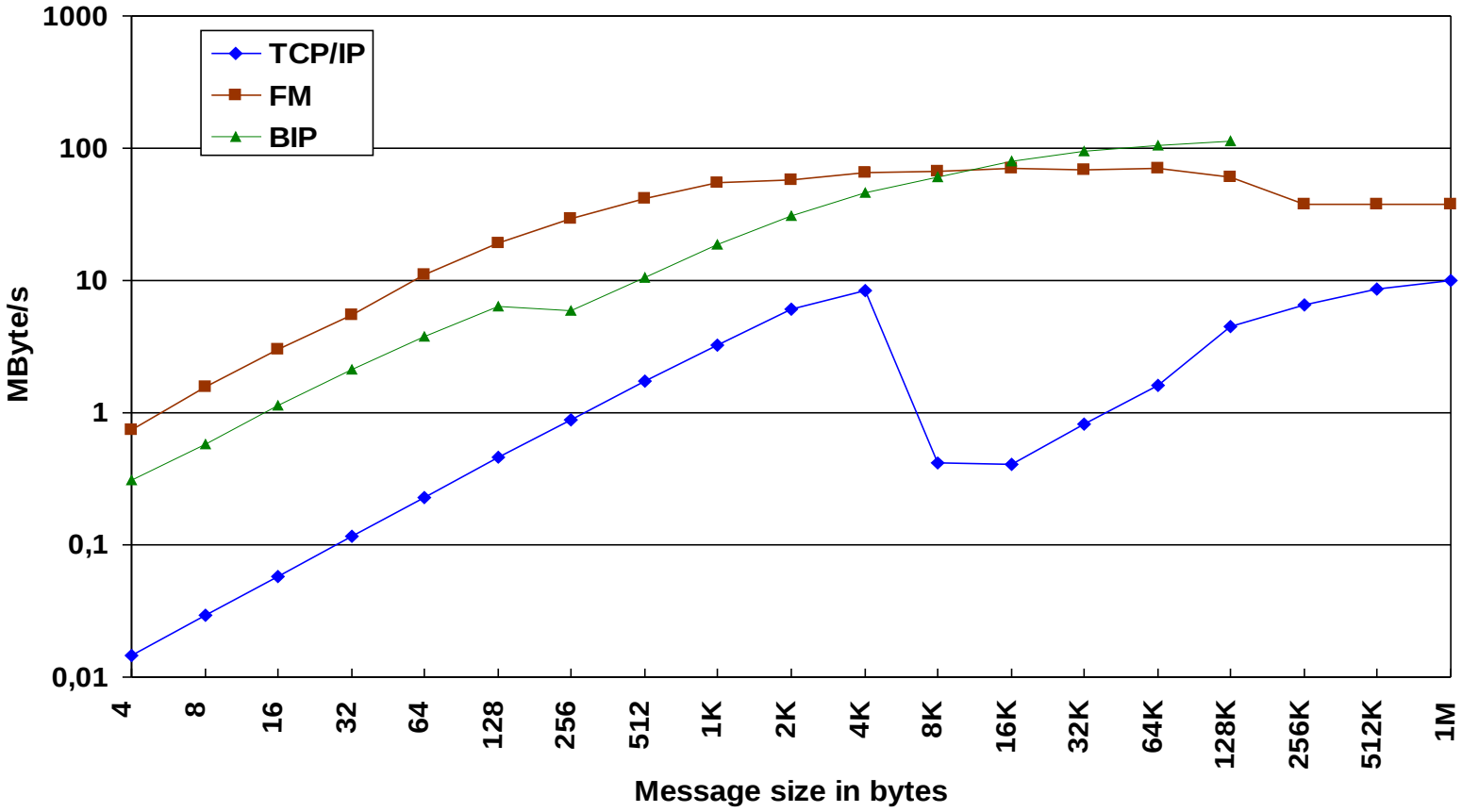
User Level Communication

- OS initializes the NIC and exports it to the applications
- Communication bypasses the OS
  - Lower overhead
  - More flexibility
- Multiplexing is a problem



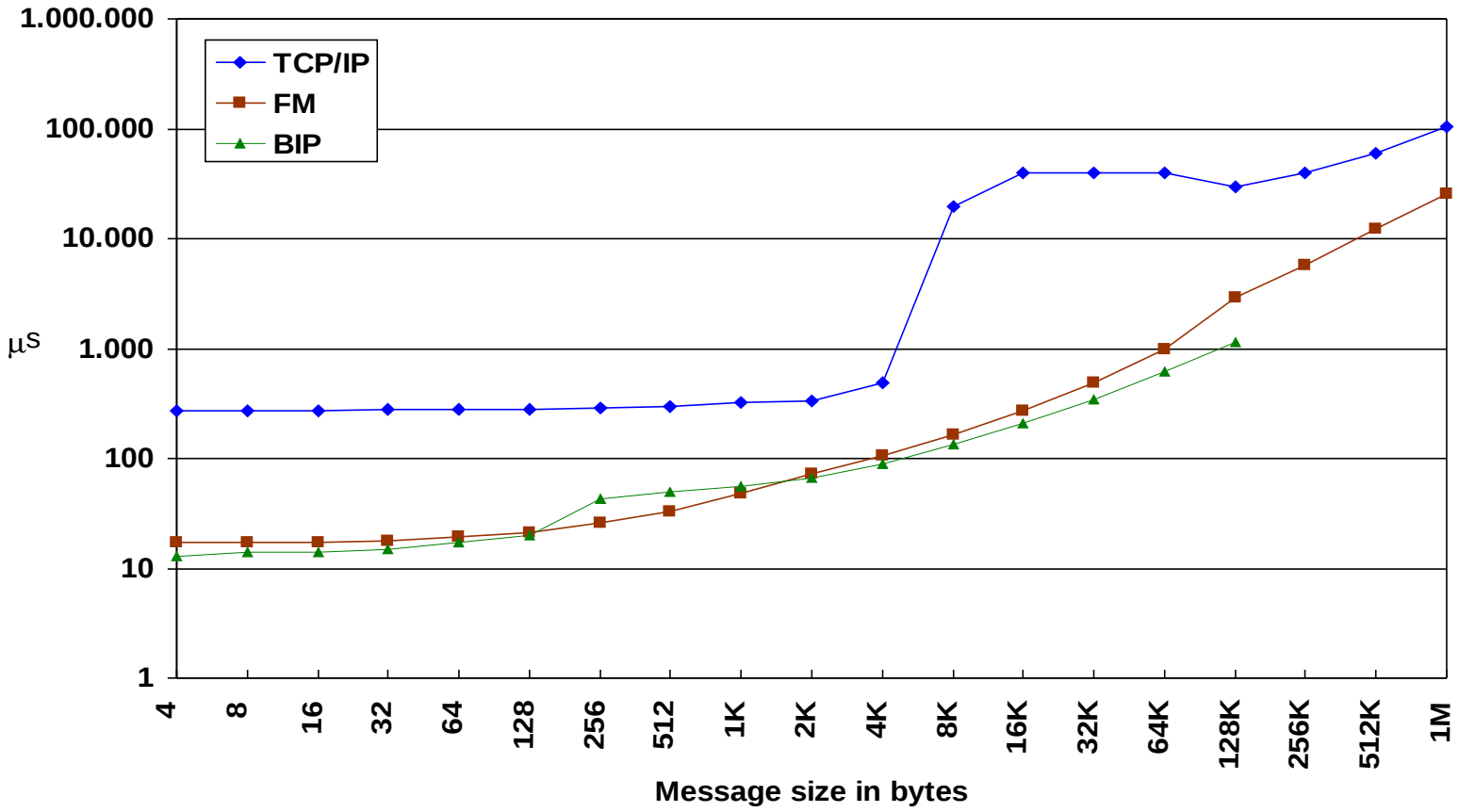
GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

MPI Bandwidth with ULC



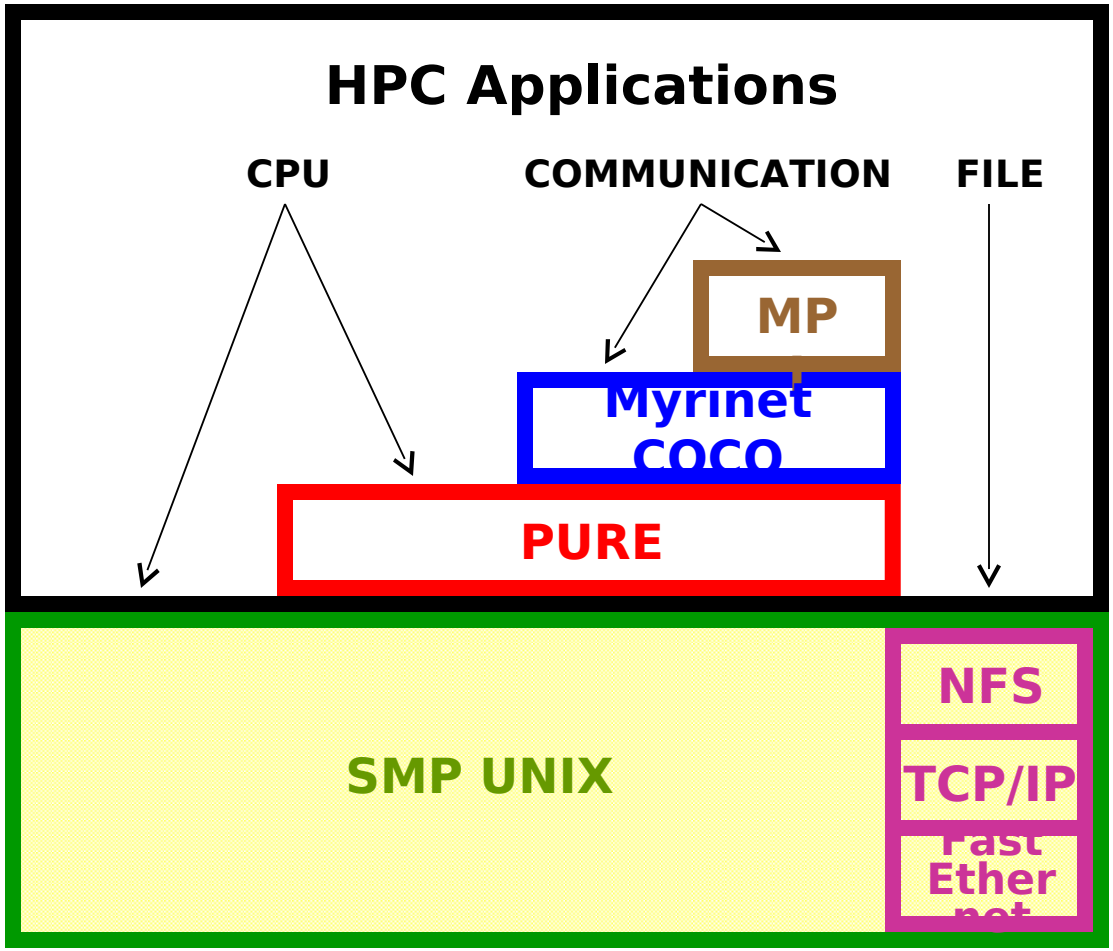
GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

MPI One-way Latency with ULC



GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

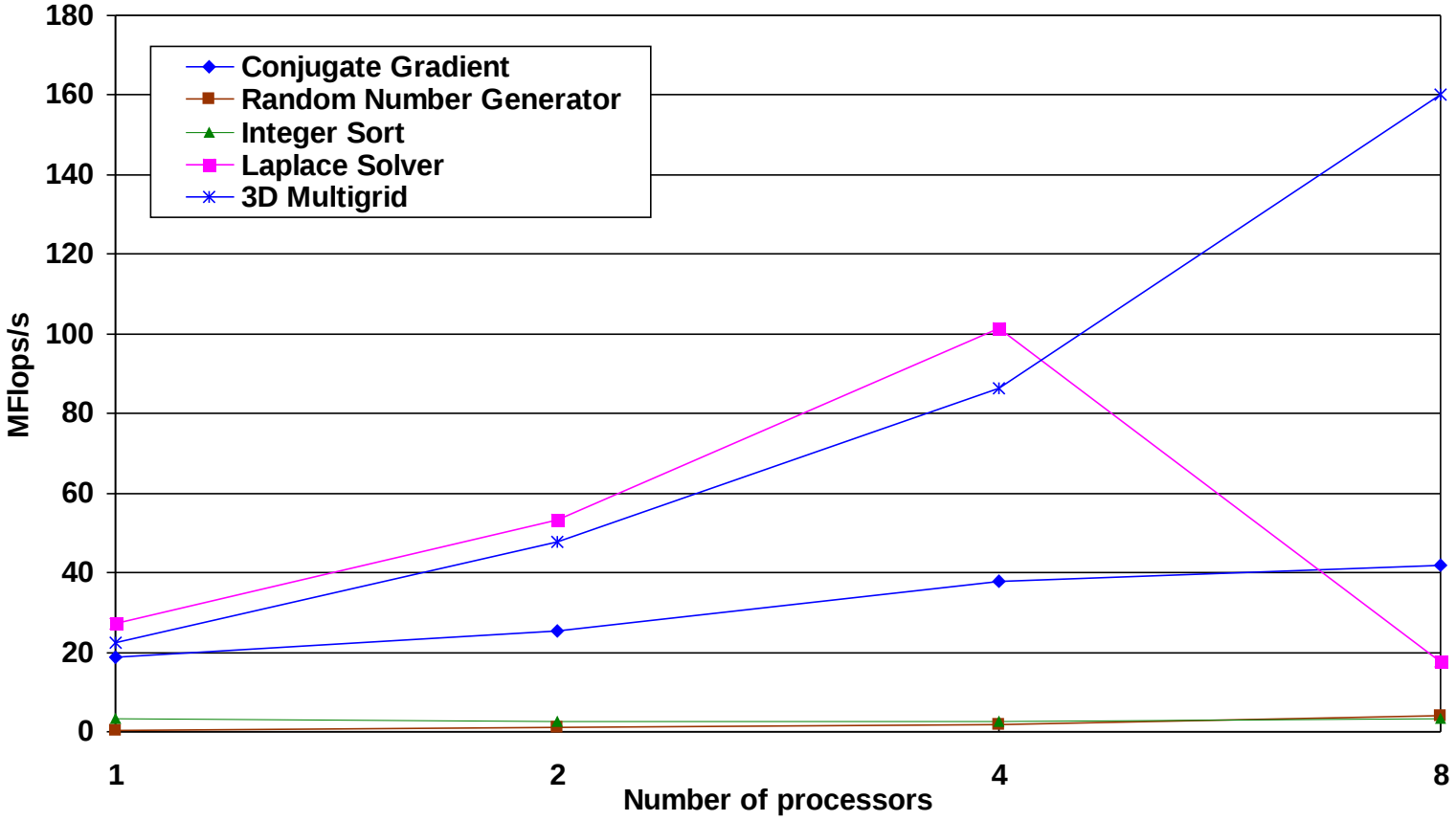
Further Steps



Berkeley NOW

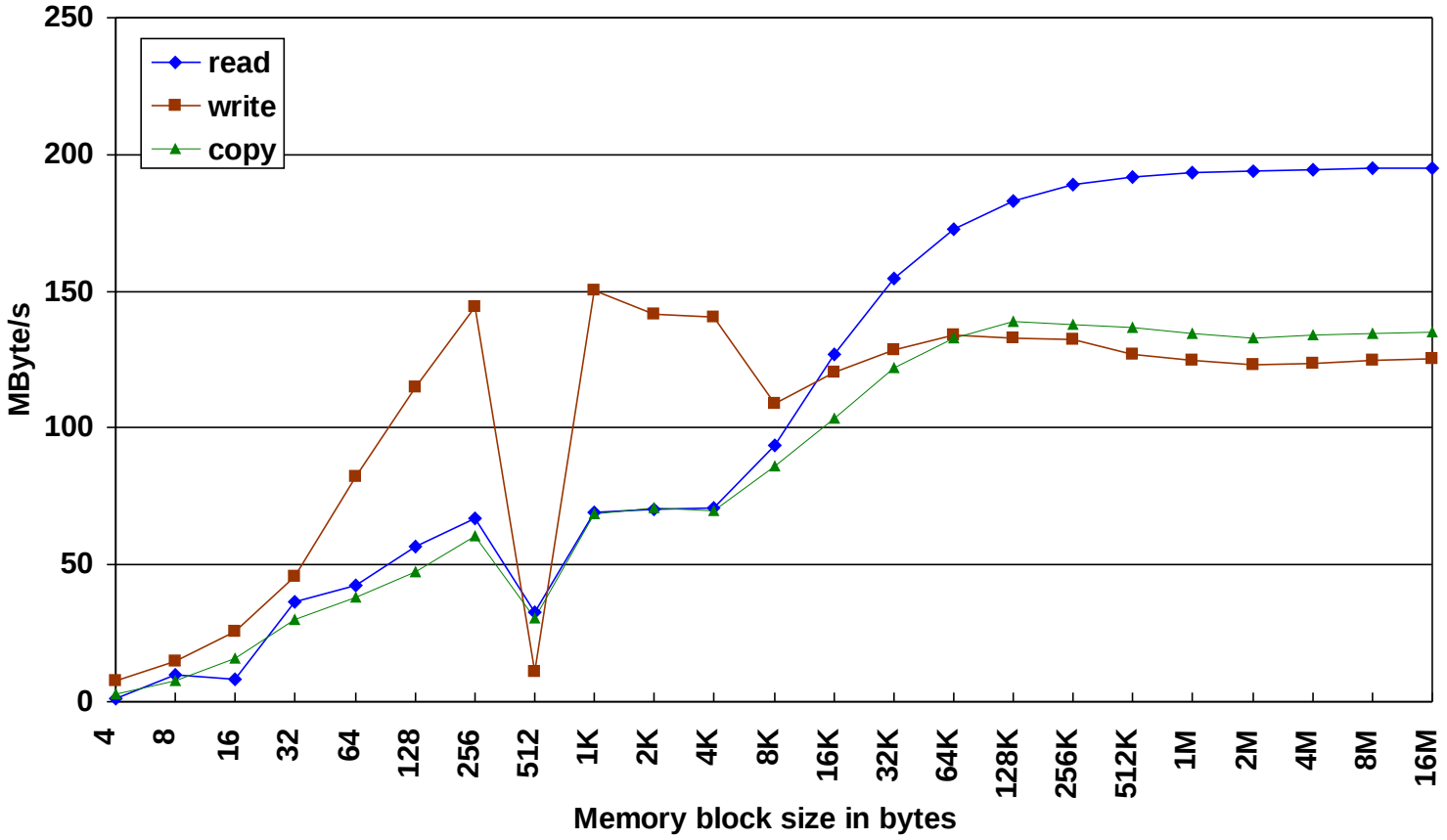
- Microprocessor performance improves 50% per year
  - MPPs come 1 to 2 year latter than WS
  - MPPs are 1.5 to 2.25 „slower“ than WS
- Costs drop 10% when volume doubles
  - MPPs can sell 40 x 200 nodes
  - WS can sell 100.000
  - WS are 1/3 cheaper than MPPs

NAS


Conjugate Gradient
Random Number Generator
Integer Sort
Laplace Solver
3D Multigrid
MFlops/s
180
160
140
120
100
80
60
40
20
0
1
2
4
8
Number of processors

GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

PC Memory Bandwidth (cache miss)



GMD FIRST
German National Research Center for Information Technology
Research Institute for Computer Architecture

NAS PB 2.3 Random Number Generator