Survey of high performance networks for lattice QCD

James N. Simone

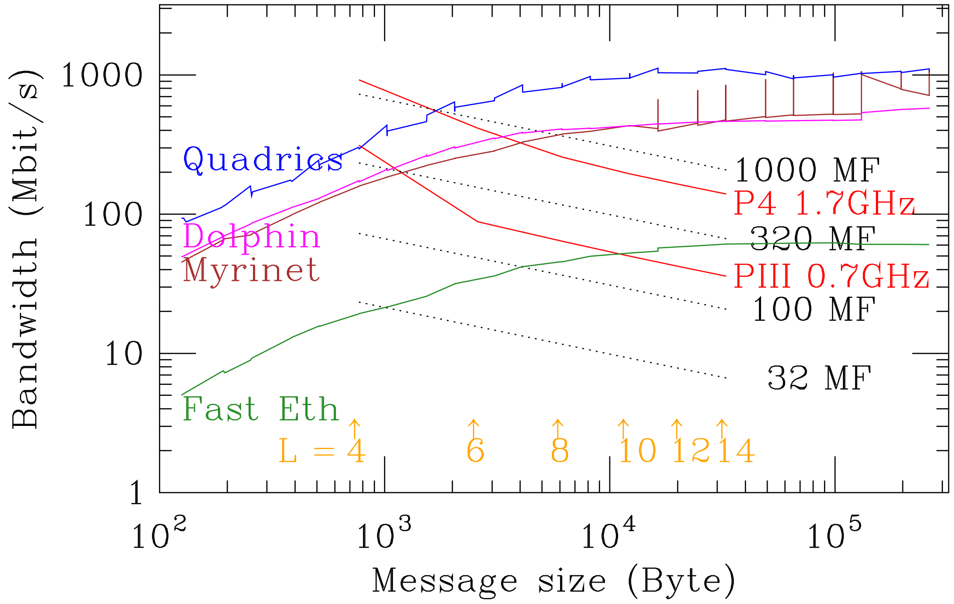
simone@fnal.gov

Fermi National Accelerator Laboratory

SciDAC site visit 3 July 2002

The need for communications performance

[S. Gottlieb]



In the application of the $\not{D}$ operator to a Kogut-Susskind quark field, to fully overlap computation with nearest-neighbor communications,

$$B \geq \frac{24 \text{ Bytes}}{66 \text{ Flops}} \frac{F}{L}$$

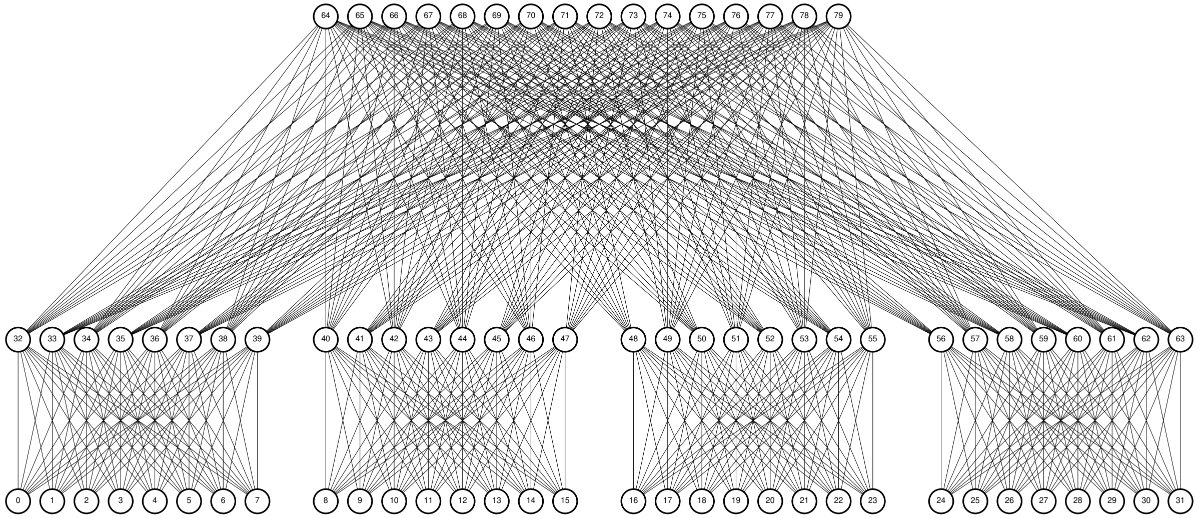
where B is the average bandwidth in MB/s, F is the sustained floating point rate in MFlops/s and L is the length of a sub-grid in sites.

Survey of technologies

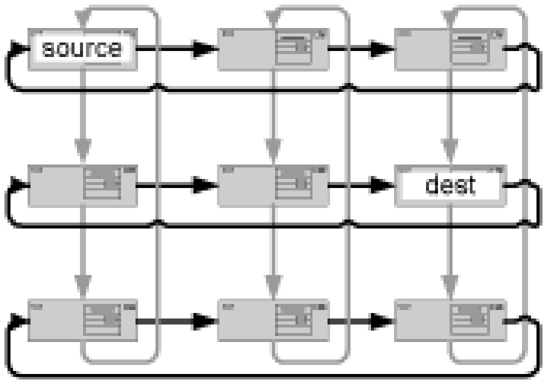
technology	topology	bw/node [Gb/s]	latency [μsec]	list cost/node
Gig. Ethernet	n-dim torus point-to-point	$2n \times 1.0$	12/link	$2n \times$ \$40-1,100
Myrinet	"fat tree" switch	$2.0 + 2.0$	7	\$1,830
Dolphin SCI	3-dim torus distrib. sw.	3×2.0	3.5	\$2,100
Quadrics	"fat tree" switch	$2.7 + 2.7$	5	\$3,500-4,000
InfiniBand	switch?	$2 + 2$ ($1\times$)	low/link	\$\$ development kit
		$8 + 8$ ($4\times$)		future?
		$24 + 24$ ($12\times$)		future?
Custom FPGA	6-dim torus	6×2.5	< 8	$\sim$ \$800 +development

Fabrics

256-port Clos switch



2-dim torroidal 3×3 mesh

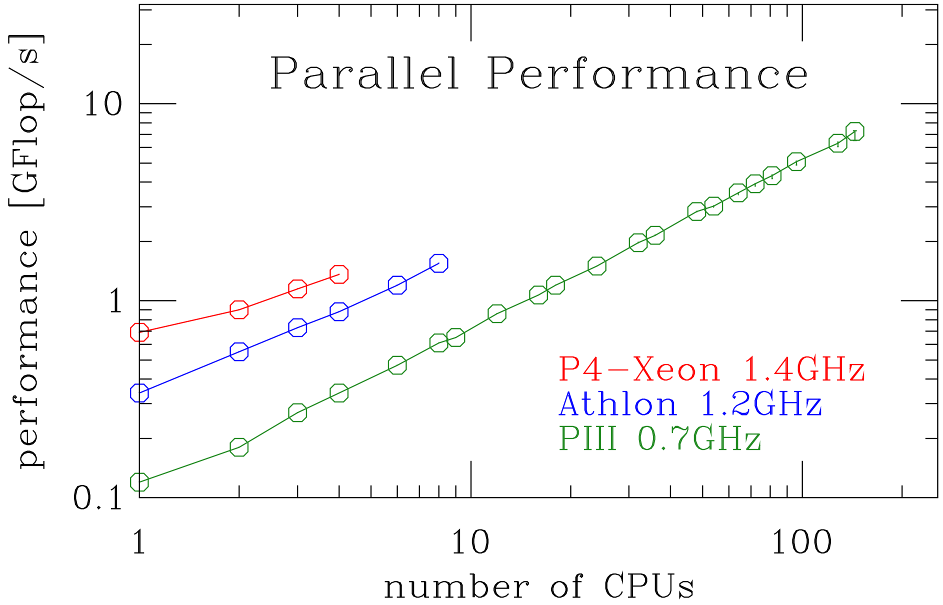


Bottleneck: NIC to memory

Current network cards connect with PCI interfaces to the compute nodes. PCI bus arbitration adds $1\text{-}2\ \mu$secs of latency per communication operation. Obtainable bandwidth depends on quality of PCI implementation in the CPU bridge chipset.

technology	theor. bw [MB/s]	best observed with Myrinet DMA
PCI 32-bit, 33 MHz	132	128
PCI 64-bit, 66 MHz	528	455(r) 512(w)
PCI-X 64-bit, 100 MHz	800	
PCI-X 64-bit, 133 MHz	1064	
PCI-express (3GIO)	2-64 pins $\times$ 2.0 Gb/s/pin	

Myrinet



- Myrinet 2000 reasonable match to current commodity computer systems
- will dual P4 systems need multiple NICs? switch costs double!
- future multi-port cards mix of Myrinet, InfiniBand, GigE and 3GIO
- future performance increases: 2.5 Gb/s ports, faster NIC processors, PCI-X and enhanced software (MPI support on NIC, GM, VI and sockets)
- Myricom trend: best performance @ const. price point

GigE mesh

Achieves $\$1/$MFlop (Wilson) $\$1.5/$MFlop (Kogut-Susskind) with MILC code on a 8×8 mesh of 1.7 GHz P4 machines [Z. Fodor, *et al.*].

Fermilab, MIT and Jlab are holding weekly phone meetings to coordinate GigE networking investigations.

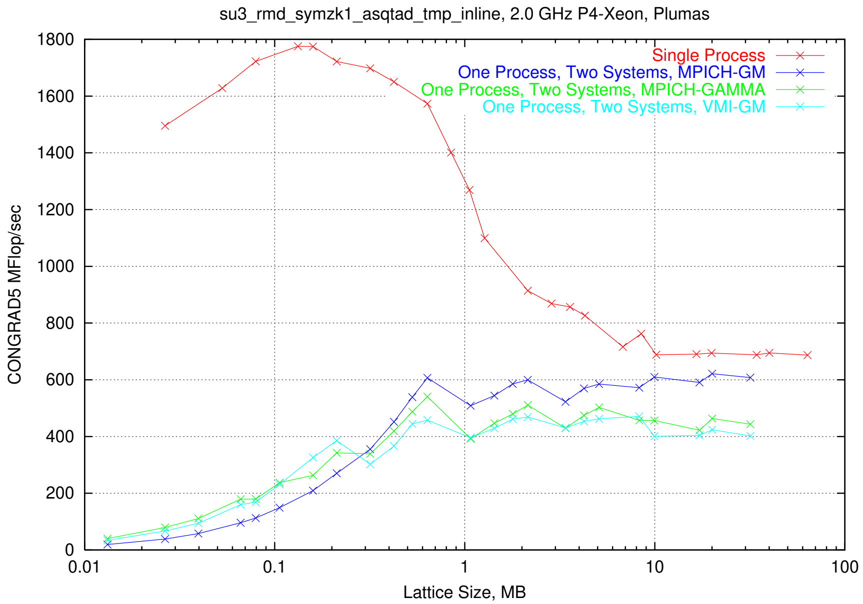
Advantages:

- scalable performance
- inexpensive: GigE NICs approaching commodity pricing (\$40-$1,100$)
- no GigE switches: expensive/port, store-and-forward routing (Avici ?)
- low latency for nearest-neighbor communications with "os-bypass" drivers *e.g.* Gamma (Genoa), EMP (OSU), mvia (NERSC)

Potential disadvantages

- routing of non nearest-neighbor messages (*e.g.* FFT) requires CPU cycles
- low dim. meshes require physical re-wiring to repartition nodes

GigE link tests for MILC code



- GigE test: GAMMA (Genoa Active Messages) drivers and MPICH
- Netgear GA622T NICs ($58 ea)
- zero-length message latency 12 μsecs
- asymptotic bandwidth a modest 50 MB/s (NETpipe)
- Fermilab in contact with GAMMA group to investigate performance of GA622T

Custom hardware

Companies such as Lattice and Xilinx provide FPGA hardware and the intellectual property (IP) building blocks needed to for high-performance data routing, translation and aggregation. Uses in voice and data products.

Features summary Xilinx Virtex-II Pro

- $3{,}168$ to $50{,}832$ logic cells
- 0-4 PowerPC 405 cores, 300 MHz, no FPU
- DDR memory controller (IP solution)
- 4-16 full duplex serial links, each up to 2.5 Gb/s (3.125 Gbaud)
- multi-protocols: InfiniBand, 1 GigE, 10 GigE, 3GIO
- aggregate bandwidth up to 40 Gb/s
- PCI-X 64-bit 133 MHz, support for DMA (IP solution)

Custom Network Interfaces

Advantages

- up to 16 multi Gigabit per second ports per NIC
- routing of messages by the NICS (low system overhead)
- *e.g.* 6-dim torus topology: allow partitioning multiple 4-dim jobs
- hardware design can be well matched to SciDAC QCDMP API, *e.g.* simultaneous gather/scatter DMA operations
- PCI-X now, 3GIO future
- share development and intellectual property costs with experiments?

Disadvantages

- what are the (true) development costs?
- how long before industry overtakes a custom NIC in price/performance?

Custom Prototypes

The Fermilab Electronic Systems Engineering Group is prototyping the Xilinx Virtex-II and Lattice Semiconductor chips for future high-speed DAQ system applications (CKM, BTeV).

ESE will build an additional four Virtex-II prototype cards for lattice QCD investigations: latency, bandwidth, drivers, protocols

Features

- four 2.0 Gb$/$s serial ports
- Cu (Optical) cabling
- 32-bit 33 MHz PCI bus supporting initiator/target block memory moves (DMA)
- rudimentary DMA engine

Summary

Fermilab continues to investigate networking fabrics for lattice QCD. We seek the best balance between communications performance and floating point performance on present and future commodity systems while seeking to reduce the total cluster cost.

Dolphin SCI, Quadrics, InfiniBand

Dolphin

- Fermilab in contact with Dolphin to benchmark $4 \times 5 \times 6$ torus of dual AMD 760MPX 2000+ systems.

Quadrics

- MILC achieves MFlop$/$s on Pittsburgh Quadrics cluster.
- excellent RDMA (remote get, put) performance
- costly at list price

InfiniBand

- only development hardware available now
- current switches have at most 32 ports
- diminished hopes for InfiniBand integration into Intel CPU bridge chip sets