

Comparing Clusters and Supercomputers for Lattice QCD

S. Gottlieb, Indiana University

ABSTRACT

Since the development of the Beowulf project to build a parallel computer from commodity PC components there have been many such clusters built. The MILC QCD code has been run on a variety of clusters and supercomputers. Key design features are identified, and the cost effectiveness of clusters and supercomputers are compared. New opportunities in processors, motherboards and networking will also be discussed.

[Introduction to Beowulf](#)

[Keys to Performance](#)

[Benchmarks](#)

[Cost Effectiveness](#)

[New Opportunities](#)

Introduction to Beowulf

A web site for the Beowulf project may be found at <http://www.beowulf.org>.

The Beowulf Project was started at [CESDIS](#), which is operated for NASA by [USRA](#) in early 1994. In the summer of 1994 the first Beowulf 16 node cluster was constructed for the Earth and Space Sciences project, ([ESS](#)), at the Goddard Space Flight Center ([GSFC](#)). The project quickly spread to other NASA sites, other R&D labs and to universities around the world. The project's scope and the number of Beowulf installations have grown over the years.

There are about 100 clusters listed on the Beowulf home page. Within the MILC collaboration, we have access to five clusters at our universities.

[Gandhi was one of the first supporters of Linux!](#)

What are some advantages of clusters?

- Commodity hardware (Intel, AMD, Alpha(?), FastEthernet, Myrinet(?))
- Commodity software (Linux, MPICH)
- Programmability, Flexibility
- Community of users and developers
- Short design time; can take advantage of many developments

What are some disadvantages of clusters?

- **Nobody to complain to, i.e., no vendor (but that may change)**
- **Have to rely on design of others, i.e., what is currently available**

[CANDYCANE \(next slide\)](#)

[Back to Outline](#)

**It is difficult but not
impossible to conduct
strictly honest business....**

**What is true is that honesty
is incompatible with the
amassing of a large fortune.**

Mohandas K. Gandhi, Non-Violence in Peace and War

CANDYCANE



CPU
AND
NETWORK
DO
YOUR
CALCULATION
AND
NOTHING
ELSE

\$50K funding to physics dept. to build 32-node cluster

Sept. '98 build and test 4 node cluster

Oct. '98 put out bids for components

Nov. '98 last component arrives
Wednesday and Friday of Thanksgiving Break
Build 34 nodes and start cluster running.

Actual cost \$693/node PII350,
4.3GB, 64MB Ram
Fast ethernet
HP Procurve switch \$2,000
\$25,000

Today <\$400/node ~\$15K
1.2 - 1.5 GF 10-12.5 \$/MF

[Keys to Performance \(next slide\)](#)

[Back to Outline](#)

Keys to Performance

- **Single node floating point speed.**

quality of CPU

cache performance, size

compiler

memory bandwidth

- **Message passing performance**

latency

peak bandwidth

processor overhead

messaging software

[Single Node Performance \(next slide\)](#)

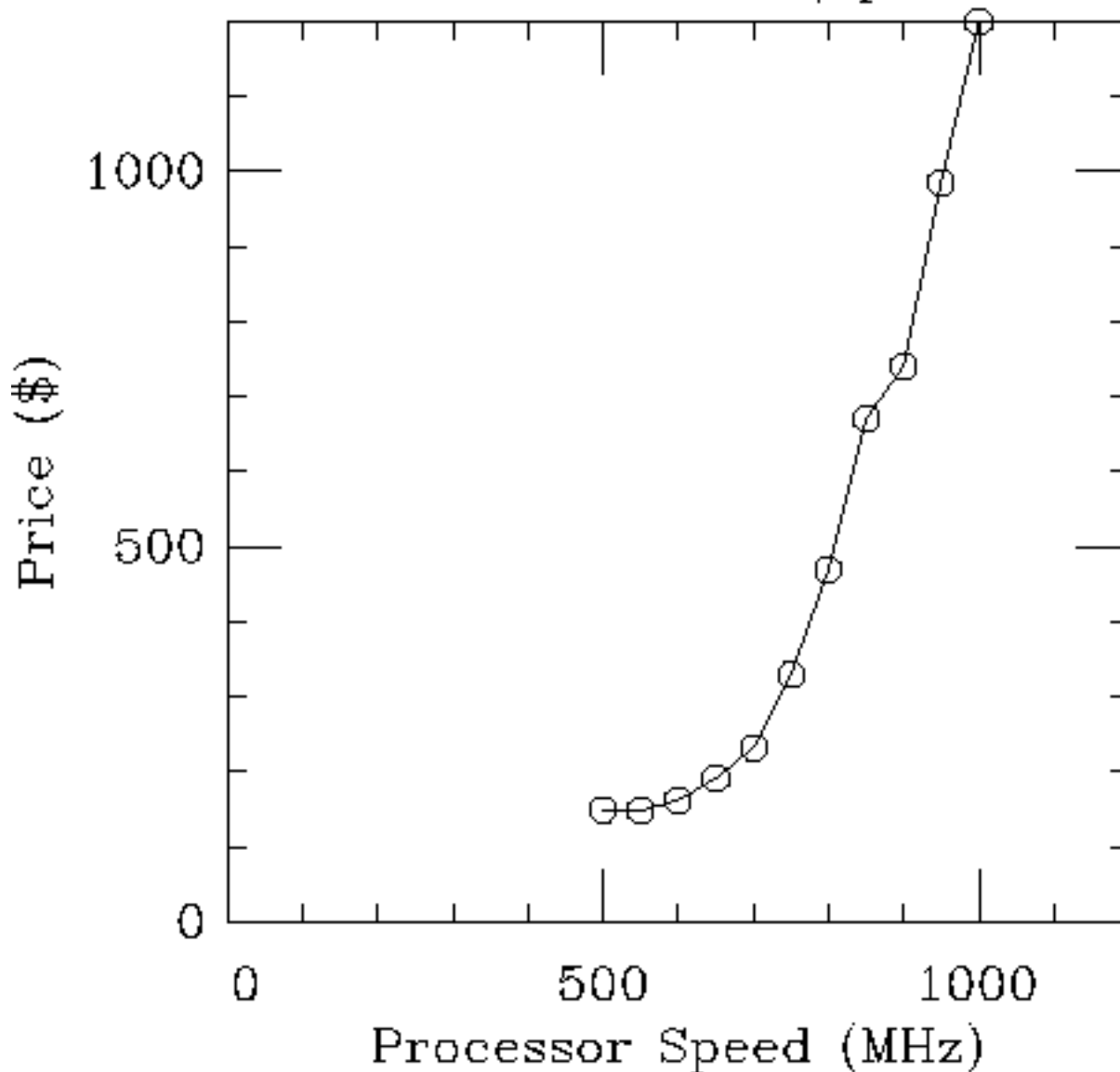
[Back to Outline](#)

Single Node Performance

It is easy to waste a lot of money on poor system design. To illustrate this, we consider the variety of AMD Athlon processors available and their costs. The same considerations apply to Intel or Alpha processors. Component prices vary a great deal during their lifetime, so we give a date for the graphs that depend upon price.

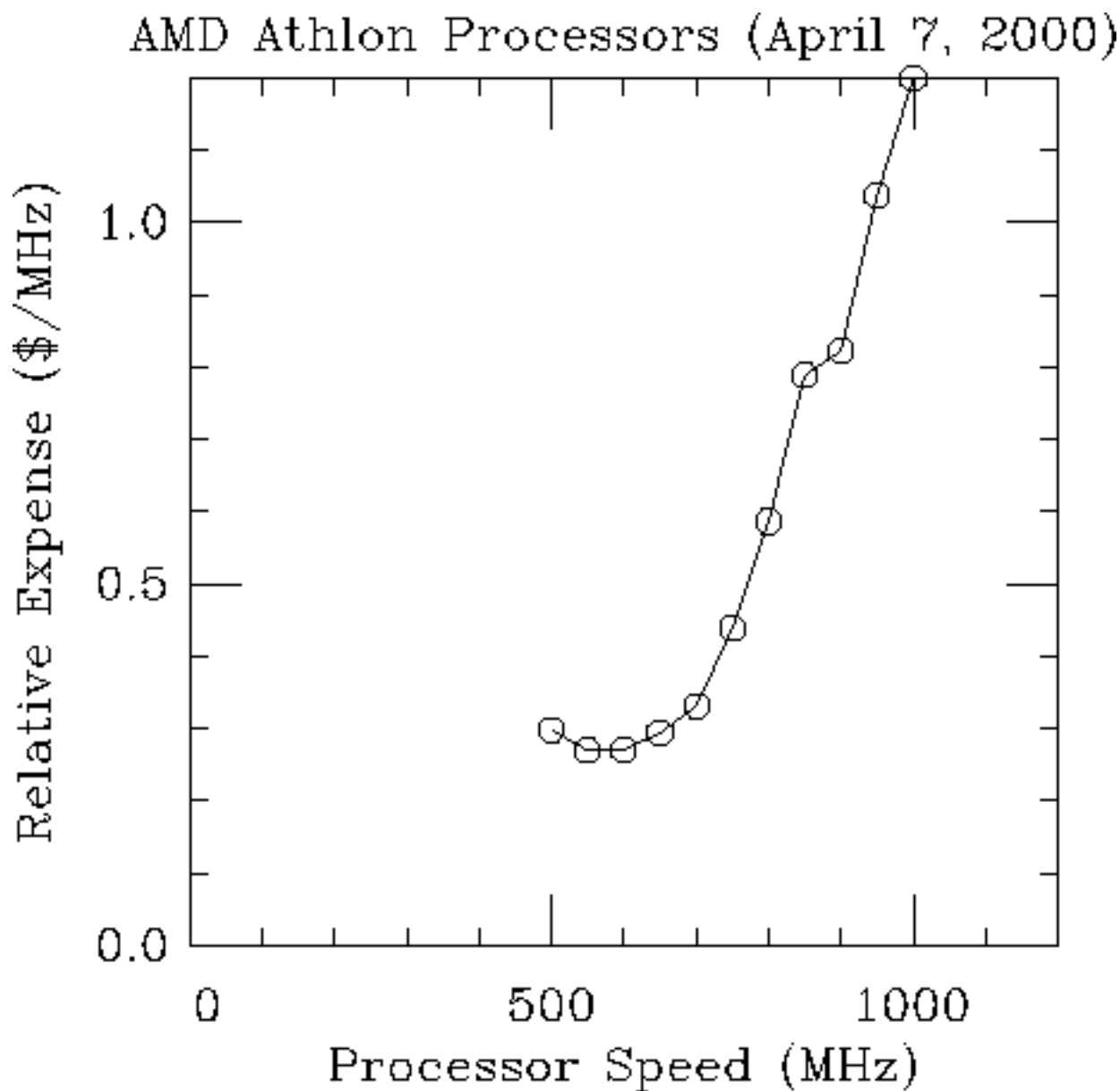
Processor price is a rapidly increasing function of speed.

AMD Athlon Processors (April 7, 2000)



Dividing by the speed of the chip, we still see that the

relative expense rises rapidly for the faster chips. In this case, there was an apparent sweet spot at 600 MHz. The faster chips have a higher price-performance ratio. Depending upon the costs of the other components of the system, the entire system may have a higher (undesirable) or lower (desirable) price-performance ratio.



For our QCD codes, access to memory is quite important. We demonstrate with the benchmarks below that performance does not increase in proportion to the speed of the chip. This is because memory speed is fixed when we compare 500 MHz and 600 MHz Athlons.

Results for 500 MHz Athlon

```

-----
L= 4  nodes= 1  230.80 +/- 5.10 MF/node
L= 6  nodes= 1  128.51 +/- 0.70 MF/node
L= 8  nodes= 1   97.20 +/- 0.30 MF/node
L= 10 nodes= 1   91.85 +/- 0.11 MF/node
L= 12 nodes= 1   89.91 +/- 0.06 MF/node
L= 14 nodes= 1   88.53 +/- 0.07 MF/node

```

Results for 600 MHz Athlon

```

-----
L= 4  nodes= 1  275.71 +/- 6.50 MF/node
L= 6  nodes= 1  135.16 +/- 0.67 MF/node
L= 8  nodes= 1  102.16 +/- 0.09 MF/node
L= 10 nodes= 1   96.50 +/- 0.09 MF/node
L= 12 nodes= 1   94.51 +/- 0.02 MF/node
L= 14 nodes= 1   93.43 +/- 0.15 MF/node

```

When comparing these two tables, we see that for $L = 4$, for which the problem fits in cache, there is a 19.5% speedup on the faster processor. But for all the larger problems, the speedup is only 5%. We expect that for even faster processors, the memory access will become an even greater issue and performance increases will be marginal.

Since memory access is so crucial, I have purchased a Pentium III 533B chip that uses PC133 memory. In theory, it should provide about 33% better performance than a similar chip with PC100 memory. I have tried three different motherboards using different support chips and the results are disappointing. The Gigabyte GA6VXE+ motherboard uses a VIA chipset, the Supermicro PIISED uses the Intel 810e chipset and I also tried an Intel CC820 motherboard using the Intel 820 chipset. The results are not particularly better than a PII 350 or 450 MHz chip using a BX motherboard.

::::::::::::

Gigabyte_GA6VXE+

::::::::::::

L= 4	nodes= 1	185.70	+/-	4.33	MF/node
L= 6	nodes= 1	106.01	+/-	0.15	MF/node
L= 8	nodes= 1	81.08	+/-	0.01	MF/node
L= 10	nodes= 1	75.78	+/-	0.01	MF/node
L= 12	nodes= 1	75.86	+/-	0.01	MF/node
L= 14	nodes= 1	73.37	+/-	0.00	MF/node

::::::::::::

Intel_CC820

::::::::::::

L= 4	nodes= 1	181.55	+/-	3.62	MF/node
L= 6	nodes= 1	97.57	+/-	0.09	MF/node
L= 8	nodes= 1	75.73	+/-	0.03	MF/node
L= 10	nodes= 1	71.61	+/-	0.07	MF/node
L= 12	nodes= 1	70.42	+/-	0.01	MF/node
L= 14	nodes= 1	70.15	+/-	0.00	MF/node

::::::::::::

Supermicro_PIII5ED

::::::::::::

L= 4	nodes= 1	174.06	+/-	3.01	MF/node
L= 6	nodes= 1	93.91	+/-	0.18	MF/node
L= 8	nodes= 1	72.89	+/-	0.02	MF/node
L= 10	nodes= 1	70.12	+/-	0.02	MF/node
L= 12	nodes= 1	69.28	+/-	0.00	MF/node
L= 14	nodes= 1	68.88	+/-	0.05	MF/node

::::::::::::

CANDYCANE PII350 MHz with gcc compiler

::::::::::::

L= 4	nodes= 1	113.63	+/-	0.97	MF/node
L= 6	nodes= 1	82.65	+/-	0.13	MF/node
L= 8	nodes= 1	71.89	+/-	0.02	MF/node
L= 10	nodes= 1	70.26	+/-	0.01	MF/node

L= 12 nodes= 1 69.60 +/- 0.00 MF/node
 L= 14 nodes= 1 69.68 +/- 0.00 MF/node

::::::::::::

Roadrunner PII450 MHz with PGI compiler

::::::::::::

L= 4 nodes= 1 141.95 +/- 1.64 MF/node
 L= 6 nodes= 1 98.85 +/- 0.17 MF/node
 L= 8 nodes= 1 81.97 +/- 0.05 MF/node
 L= 10 nodes= 1 78.78 +/- 0.01 MF/node
 L= 12 nodes= 1 78.06 +/- 0.01 MF/node
 L= 14 nodes= 1 77.84 +/- 0.01 MF/node

However, there are support chips from ServerWorks that do support PC133 memory. Here are results from Los Lobos, that uses Intel 733 MHz chips in IBM NetFinity servers.

::::::::::::

Los Lobos

::::::::::::

L= 4 nodes= 1 283.90 +/- 3.28 MF/node
 L= 6 nodes= 1 128.22 +/- 0.48 MF/node
 L= 8 nodes= 1 119.97 +/- 0.49 MF/node
 L= 10 nodes= 1 106.88 +/- 2.73 MF/node
 L= 12 nodes= 1 114.95 +/- 2.36 MF/node
 L= 14 nodes= 1 103.93 +/- 0.64 MF/node

Currently, one cannot get dual processor motherboards for the Athlon processor; however, the above results show that it would be a better choice for this code if single processor motherboards will be used. The motherboards with ServerWorks support chips are expensive.

[Graph of Performance Model \(next slide\)](#)

[Back to Outline](#)

A Simple Performance Model

A simple performance model of the Kogut-Susskind Conjugate Gradient algorithm gives this bandwidth requirement to overlap communication and floating point operations:

$$MB = 48 MF / (132 L) = 0.364 MF / L ,$$

where MB is the achieved bandwidth in Megabytes/s, MF is the achieved floating point speed in Megaflops/s and an L^4 portion of the grid is on each node.

A graph shows measured bandwidth for a ping-pong test for three types of hardware and the performance model for several processor speeds.

[Click here to view the graph in a PostScript viewer](#)

- Note that this is a log-log plot.
- The messages vary in size from 800 bytes to 30 KB for problem sizes of interest. The arrows near the bottom of the graph correspond to different L values.
- The green and blue curves come from measured performance on the Roadrunner supercluster at the [Albuquerque High Performance Computer Center](#). The Quadric curve comes from the Teracluster at LLNL. The measurement is done using the Netpipe program from the [Ames Scalable Computing Laboratory](#)
- The straight red lines come from the performance model presented and are plotted for matrix vector speeds of 50, 100, 200 and 400 MF. We need to run at a large enough value of L so that the measured

bandwidth is above the red line (for what ever speed our processor achieves for the corresponding value of L).

- **Pushing up the communication rate for small messages is important.**
- **It is especially nice when we don't need more expensive hardware. There is a huge price range among Quadrics, Myrinet and Fast Ethernet.**

[Latency \(next slide\)](#)

[Back to Outline](#)

Latency

With Netpipe it is easy to determine bandwidth curves as presented on the previous slide. It is also easy to get the latency for short messages and we summarize some results here.

To download Netpipe from the [Ames Scalable Computing Laboratory](http://www.scl.ameslab.gov) use this URL:

<ftp://ftp.scl.ameslab.gov/pub/netpipe/>. You will want the README and one of the gzip'ed tar files.

Latency with Fast Ethernet Hardware

CANDYCANE MPICH

151 microsec (Feb. 7, 1999)

166 microsec (Oct. 11, 1999)

Roadrunner MPICH

169-179 microsec (dual processors)

CANDYCANE MVIA

60 microsec

CANDYCANE GAMMA

30 microsec + 7 microsec (switch) + 4 microsec (MPI)
= 41 microsec

Latency with Myrinet Hardware

Roadrunner

31-34 microsec

Latency with Quadrics Hardware

LLNL TeraCluster

7 microsec

[Benchmarks \(next slide\)](#)

[Back to Outline](#)

Benchmarks

A web site for MILC benchmarks may be found at <http://physics.indiana.edu/~sg/milc/benchmark.html>. Additional graphs and explanations may be found there.

The simple performance model presented above can help us guess when the communication and floating point are in reasonable balance, but it is no substitute for real benchmarks.

- Key variables**
- **problem size**
 L^4/node
 - **# of CPU's or nodes**

All benchmarks are for Single precision Kogut-Susskind Conjugate Gradient. Use the back button on your browser to return to this page. These graphs are in PostScript. Other tables and axis are available from the [MILC benchmark site](#).

- [CANDYCANE](#)
- [Comparison of Fast Ethernet and Myrinet on Roadrunner with one CPU per node](#)
- [Comparison of Fast Ethernet and Myrinet on Roadrunner with two CPUs per node](#)
- [Comparison of MPICH and MVIA under Fast Ethernet on CANDYCANE](#)
- [Cray T3E 900](#)
- [IU IBM SP Winterhawk II \(4 way SMP\)](#)
- [Comparison of different IBM SP models](#)
- [SGI Origin \(195 MHz\)](#), note that 250 MHz is also available
- [Sun E10000](#)
- [LLNL Teracluster](#), with some assembly code and 2 EV67 667 MHz

processors per box

- [Comparison of several commercial machines for L=8](#)

[Price/Performance \(next slide\)](#)

[Back to Outline](#)

Price-Performance Ratios

Caveats

- Not so easy to get prices for supercomputers. Some quotes are old.
- Myrinet estimates were done some time ago.

		\$/MF
Intel Fast Ethernet (single processor)		10-13
	(dual processor)	worse
Myrinet	(single processor)	~27
	(dual processor)	~22

AMD may be better than Intel, but dual processing not yet available.

2/99	64 node SGI origin (250 MHz) million @ 113 MF/node	193
2/99	44 node Cray T3E \$1.9 million @ 90 MF/node	480
	256 node IBM SP Power 3 @ 173 MF/node	166 (91 discount)
	Compaq Alpha Server SC 64CPU \$2.7 million @ 280 MF/node	150

[**New Opportunities**](#)

[**Back to Outline**](#)

New Opportunities

One of the nice things about clusters is that they can rapidly incorporate advances in technology. When new processors come out, for example, there must be motherboards available soon after to support them.

There are several key area of technology:

- **Processors**
- **Memory systems**
- **Network hardware**
- **Network software**

I organized a session at the APS March meeting on PC Clusters, two well known cluster experts Rick Stevens and Thomas Sterling spoke about node technology and networking, respectively. Their talks, along with others are on the web.

<http://physics.indiana.edu/~sg/pccluster.html>

[Back to Outline](#)