Lattice QCD with Commodity Hardware and Software

D.J. Holmgren , P.B. Mackenzie, D.L. Petravick,
R.D. Rechenmacher and J.N. Simone

Fermilab

CHEP2000, February 2000

Outline

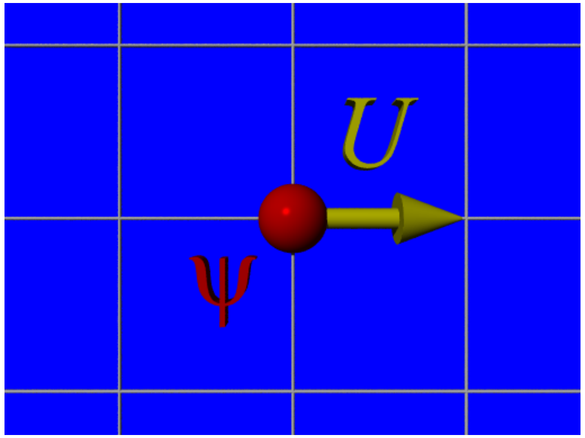
- Introduction
- Lattice QCD - Hardware and Software
- Early Results
- Future Plans

Introduction

- Lattice Quantum Chromodynamics solves the theory of quarks and gluons nonperturbatively by discretizing it on a lattice - the only way of obtaining many crucial QCD results
- New experiments are aiming for precision tests of the Standard Model
  - For example, Fermilab's CDF expects to measure x_s ($\overline{B}_s^0$-B_s^0 mixing) at the percent level.
- To connect this to CP violation requires a similarly precise QCD matrix element.
- This will require a 10^2 to 10^3-fold increase in lattice QCD computations.

Lattice QCD

- Lattice techniques are used to solve QCD from first principles.
- Problems are solved on a finite discrete periodic grid of space time points.
- Quark fields are represented by spin $\times$ color vectors, Ψ, at sites of the lattice.
- Gluon fields are represented by $SU(3)$ matrices, U, on all lattice links.



Existing Hardware - ACPMAPS

- Commissioned 1991. 356 dual-i860 boards, 32 MB/CPU.
- 36 crates containing 16-way crossbar-switched backplanes, connected via serial links into 3X3X3X2 hypercube
- Distributed memory architecture, 50 Gflop/sec peak, 10-20 Gflop/sec sustained
- Near end of designed lifetime

Future QCD Machines - Commodity Hardware/Software?

- Purpose built machines (for example, ACPMAPS):
  - Best “bang for the buck”, at least in first years of operation
  - But - typically no upgrade path without building a new machine
- Proposed commodity approach:
  - Use clustered commodity hardware: PCI bus, best price/performance CPU
  - Use “open” software (Linux and the like)
  - Use high performance parts (eg low latency networking) as necessary
  - Upgrade parts of the cluster each year, picking best price/performance
  - Linux + PCI allow changing CPU at any upgrade

PCQCD - Prototype Commodity QCD Cluster

- Built September,1999
- Eight dual 500-MHz Pentium III Nodes
- 64-bit Myrinet NIC's
- 8-port Myrinet Switch
- Fast ethernet NIC's and switch

Early Results

- Performance measurements were performed on several machine types:
  - PCQCD Prototype (Dual Pentium III cluster)
  - Quad Pentium III Xeon System
  - AMD Athlon (K7) System
  - Quad Alpha (21264) System - at www.testdrive.compaq.com - Linux + gcc
  - Alpha (21264) Cluster (1 CPU/node) - at www.testdrive.compaq.com - Tru64 + native CC
- Cluster calculations used MPI (MPICH) as message passing API

Benchmark Software Description

- Solving for quark propagators represents over 90% of the computational effort in the calculations we envision.
- The quark propagator, Ψ, is the solution of the sparse-matrix equation

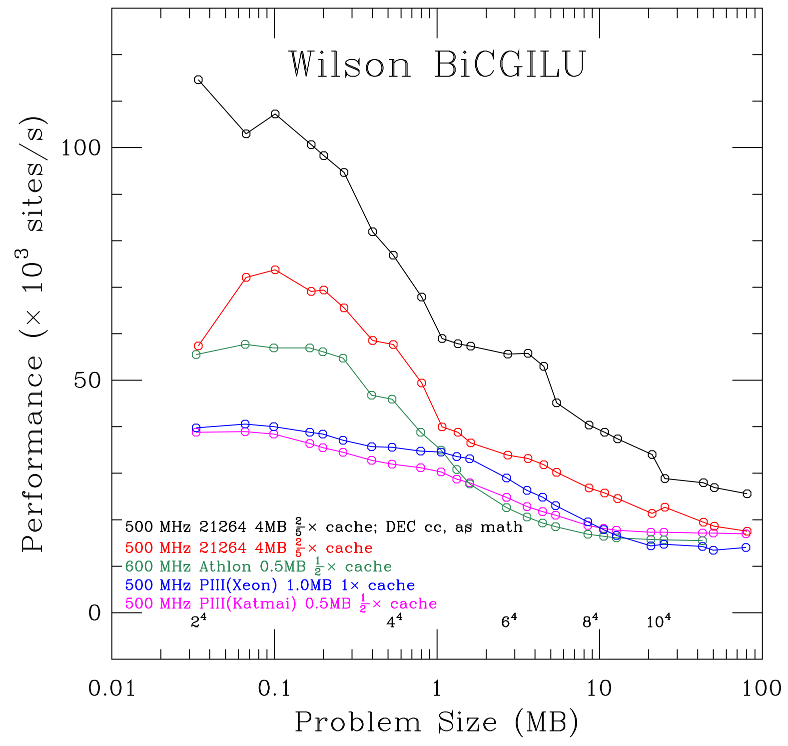
$$M\Psi = s$$

for a specified source, s.

- The MILC collaboration's Wilson quark solver is a state-of-the-art parallel production program, hence it's a good predictor of QCD performance.
- The MILC solver is based upon the Biconjugate Gradient algorithm with an incomplete LU decomposition preconditioner.

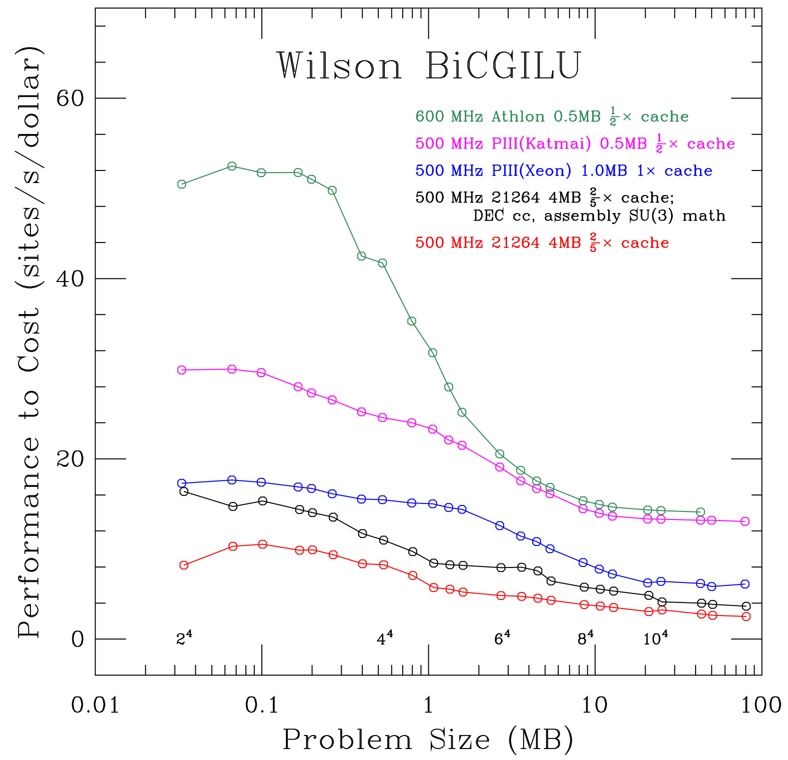
Single CPU Performance

- Achieved performance is a strong function of the problem (lattice) size. Some speculations:
  - Tiny lattices: Clock speed dominates.
  - Small lattices: Secondary cache dominates.
  - Large lattices (typical in QCD): Main memory bandwidth dominates.



Single CPU Performance Normalized by Cost

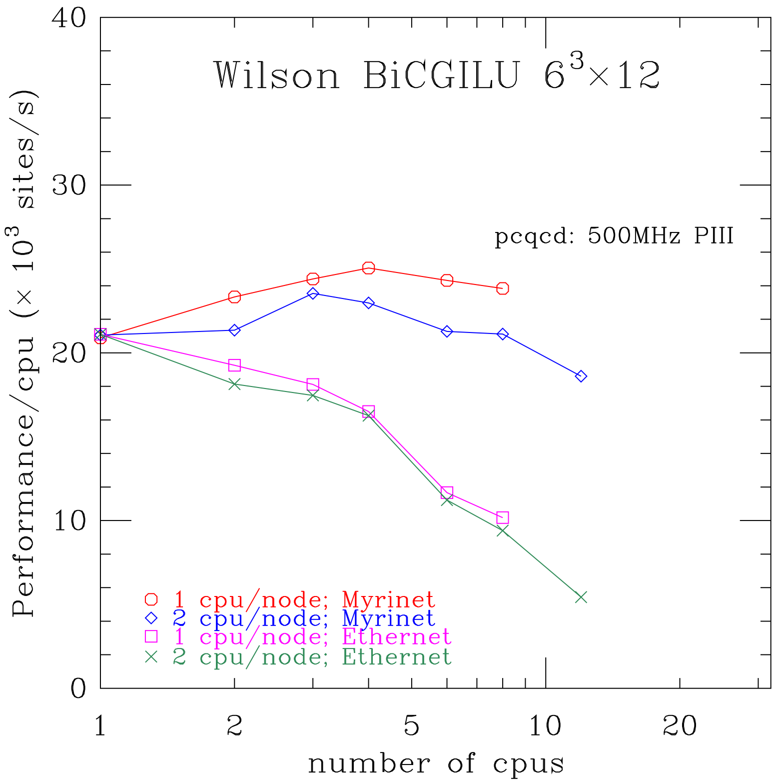
- Estimated single CPU system costs (no Myrinet):
  - Athlon: 600 MHz, $1100
  - Pentium III: 500 MHz, $1300
  - Xeon: 500 MHz, 1 MB L2, $2300
  - Alpha: 500 MHz, 4 MB L2, $7000



- Athlon, Pentium III clear winners for large lattices

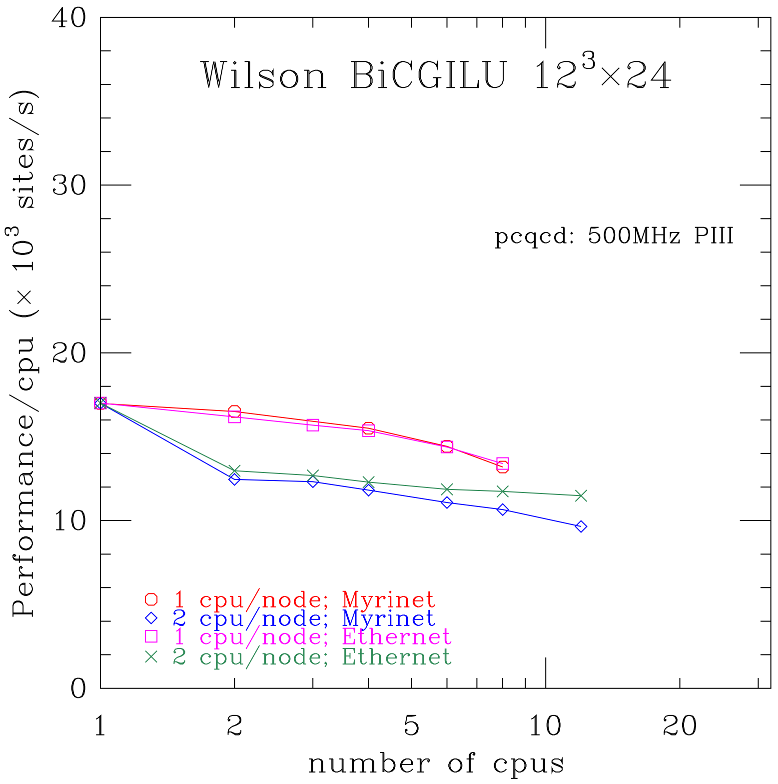
Parallel Performance: Pentium III

- Myrinet outperforms ethernet in all parallel configurations - better bandwidth and latency
- Positive slope on Myrinet probably a result of small problem size - more of the lattice fits into L2 as CPUs are added



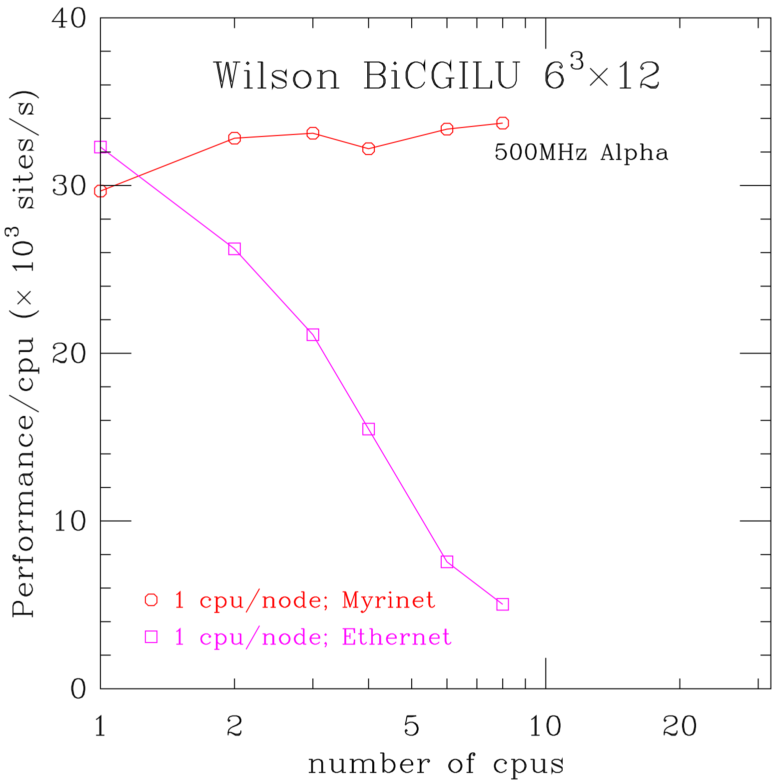
Parallel Performance: Pentium III

- Very large lattice, always exceeds L2 size
- Better performance when CPUs are distributed on more nodes (single process per node)
- No advantage for Myrinet over ethernet



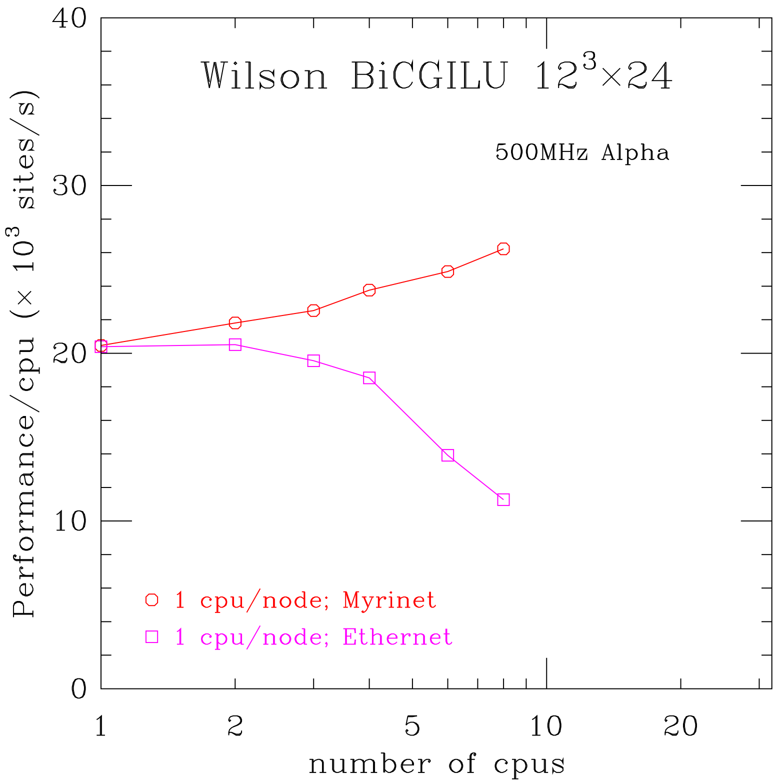
Parallel Performance: Alpha(21264)

- Thanks to Compaq for providing access to cluster.
- Similar results to Pentium III cluster - Myrinet beats ethernet in all configurations.



Parallel Performance: Alpha(21264)

- Large lattice will not fit into cache. But, data organization in MILC is designed to optimize cache reuse.
- Unlike Pentium III, Myrinet beats ethernet in all configurations



Future Work

- We've just started - lots of work to do to understand and optimize performance on 8-node Pentium III cluster.
- Conventional wisdom favors low latency (and expensive!) high performance networks. We will explore option of using commodity networks, carefully optimizing algorithms and selecting/evaluating ethernet switches.
- A full ACPMAPS replacement would require about 50 nodes. We will build and operate a production cluster of 32 to 64 nodes (depending upon budget).
- Physics demands will push HEP towards superclusters of 1000+ CPU's (teraflop-scale). We will explore and attempt to solve the many issues of building, administrating, and maintaining superclusters.