QCDoc

A massively parallel computer optimized for lattice QCD

Lattice QCD NSF Presentation

October 10, 2003

Norman H. Christ
Columbia University
MOTIVATION

● QCD is a simple, complete, fundamental theory. For many important problems the only significant errors are numerical: **Natural target for Teraflops computing.**

● Space-time homogeneity supports easy parallelization and a mesh network.

● **System-on-a-chip technology permits a highly scalable and cost-effective design:**
  – Entire node (including interconnect logic) on a single chip.
  – The only extra components:
    * Serial nearest-neighbor connections.
    * Commercial Ethernet tree for booting, diagnostics and I/O.
  – Low power, compact design.

● **High-performance, cost-effective design, related to IBM’s BlueGene/L.**
CONTROLLED NUMERICAL ERRORS

Errors fall into five categories:

- **Perturbation expansion**: \( \left( \frac{g^2(\mu)}{4\pi} \right)^n |_{\mu=1/a} \)
  
  - Effort \( \sim e^{-c/\text{Error}} \)
  
  - Avoided by *non-perturbative normalization.*
  
  - Reduced to Effort \( \sim 1/\text{Error} \) by *step-scaling.*

- **Finite lattice spacing**: \( a^n \)
  
  - Effort \( \sim (1/\text{Error})^{7/n} \)
  
  - Power \( n \) increased by *Symanzik improvement.*

- **Light quark mass**: \( m/\Lambda_{QCD} \)
  
  - Two sources of error:
    
    * Explicit mass: Effort \( \sim (1/\text{Error})^2 \)
    
    * Chiral asymmetry: Effort \( \sim (1/\text{Error})^{3.5-7} \)
  
  - Second source removed by new *chiral formulations.*

- **Finite volume**: Effort \( \sim (1/\ln(\text{Error}))^4 \)

- **Limited statistics**: Effort \( \sim (1/\text{Error})^2 \)

Substantial computer resources are needed to employ these method and to develop new ones. The resulting benefits are far greater than simply increased Monte Carlo statistics.
HISTORY

Specially targeted parallel computers have been constructed to optimize lattice QCD calculations since the early eighties:

- **Cosmic Cube, CalTech (1984).**
- **Columbia machines:**
  - 16-node, 256 Mflops (1985)
  - 64-node, 1 Gflops (1987)
  - 256-node, 16 Gflops (1989)
  - 20,000-node, 1 Tflops (1998), QCDSP
- **APE, Rome/DESY (1987-present)**
- **CP-PACs, Tsukuba (1990,1997)**
- **GF11, IBM (1992)**
COLLABORATION

Columbia (DOE):  Norman Christ
                 Saul Cohen
                 Calin Cristian
                 Zhihua Dong
                 Changhoan Kim
                 Ludmilia Levkova
                 Xiaodong Liao
                 Guofeng Liu
                 Robert Mawhinney
                 Azusa Yamaguchi

BNL (SciDAC):    Robert Bennett
                 Chulwoo Jung
                 Konstantin Petrov
                 David Stampf

UKQCD (PPARC):  Peter Boyle
                 Michael Clark
                 Balint Joo

RBRC (RIKEN):   Shigemi Ohta (KEK)
                 Tilo Wettig (Yale)

IBM:            Dong Chen
                Alan Gara

Design groups:
               Yorktown Heights, NY;
               Rochester, MN; Raleigh, NC
DESIGN

• IBM-fabricated, single-chip node.
  [50 million transistors, 4-5 Watt, 1.3cm×1.3cm die]

• PowerPC 32-bit processor
  – 1 Gflops, 64-bit IEEE FPU.
  – Memory management.
  – GNU and XLC compilers.

• 4 Mbyte on-chip memory and up to 2.0 Gbyte/node on DIMM card.

• 6-dim communications network:
  – Efficient for small packet sizes, ≈ 200ns latency.
  – Global sum/broadcast functionality.
  – Minimal processor overhead.
  – Lower dimensional machine partitions.

• 100 Mbit/sec, Fast Ethernet
  – JTAG/Ethernet boot hardware.
  – Host-node OS communication.
  – Disk I/O.
  – RISCWatch debugger.

• ≈ 5 Watt, 15 in³ per node.
Complete Processor Node on a Single QCDQ Chip

- 4 MBytes of Embedded DRAM
- 8 Gbyte/sec Memory/Processor Bandwidth
- 1 Gflops Double Precision RISC Processor
- 2.6 GByte/sec Interface to External Memory
- 2.6 GByte/sec EDRAM/SDRAM DMA
- 24 Link DMA Communication Control
- 24 Off-Node Links 12 Gbit/sec Bandwidth
- Bootable Ethernet Interface
- 100 Mbit/sec Fast Ethernet
STATUS

• ASIC Design
  – RTL design complete 6/30/02.
  – Floorplanning netlist accepted 8/16/02.
  – Preliminary netlist accepted 9/6/02.
  – Final netlist accepted 11/12/02.
  – Final netlist revised 1/16/03.
  – Extensive testing:
    ∗ Real code on full design simulation 100 applications of Dirac operator: No bugs! (?)
    ∗ Two-node simulation:
      A sends Ethernet boot packets to B.
      B receives and executes boot code.
    ∗ Testbenches for faster simulation of subcomponents.
    ∗ Formal verification passed.
    ∗ Gate-level simulations with realistic timing passed.
    ∗ ASIC tape-out 4/8/03.
    ∗ Prototypes shipped 6/04/03.
STATUS

- Board and Cabinet Design
  - Daughter board designed and built.
– 64-node Mother board designed/built.
• Single Motherboard Cabinet.
- Serial communications simulated and tested.

  - Simulation:

  ![Simulation Graph]

  - Real 6 meter cable:

  ![Real Cable Graph]
RELIABILITY

- ECC on external DIMM and EDRAM.

- Automatic recovery from single-bit communications errors.

- Running check sum on both ends of each serial channel.

- Number of components similar to QCDS[1]: 1-2 failures/week on 10K node machine.

- Soft error rate estimated at < 1/week on 10K nodes (low-α lead in solder balls).
SOFTWARE STRATEGY

• Industry standard, “RISC” processor.
  – GNU C++ compiler.
  – Memory management.

• “Front-end” OS: qdaemon.
  – Normal Unix daemon.
  – Boots, partitions and manages “back-end” grid of QCDOC nodes via RPC.
  – Aggressively multi-threaded.
  – Runs on an SMP machine.
  – Controlled by a modified shell “qcsh” or a PBS queuing system.

• “Back-end” node OS: run-kernel
  – Non-preemptive.
  – Uses 440 MMU for protection, not translation.
  – 64-entry TLB covers entire QCDOC memory.
  – Services hardware and software interrupts.
  – Provides standard C run-time environment.
  – Contains extensive hardware diagnostics.

• Partitioning
  – 6-D torus divided into 6-D rectangular solids.
  – Lower dim. torii folded within these 6-D solids.
  – Implemented as an automatic algorithm.
PARTITIONING

Folding two machine axes into one periodic application axis.

Folding three pairs of machine axes into three periodic application axes.
PERFORMANCE†

<table>
<thead>
<tr>
<th>Operation</th>
<th>Local Vol.</th>
<th>Mflops/node</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Optimized assembly code:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SU3-SU3</td>
<td>-</td>
<td>800</td>
</tr>
<tr>
<td>SU3-2spinor</td>
<td>-</td>
<td>780</td>
</tr>
<tr>
<td>DAXPY</td>
<td>-</td>
<td>190</td>
</tr>
<tr>
<td>ZAXPY</td>
<td>-</td>
<td>450</td>
</tr>
<tr>
<td>DAXPY-Norm</td>
<td>-</td>
<td>350</td>
</tr>
<tr>
<td>Wilson $D_{eo}$</td>
<td>$2^4$</td>
<td>470</td>
</tr>
<tr>
<td>Wilson $D_{eo}$</td>
<td>$4^4$</td>
<td>535</td>
</tr>
<tr>
<td>Clover $D_{eo}$</td>
<td>$2^4$</td>
<td>560</td>
</tr>
<tr>
<td>Clover $D_{eo}$</td>
<td>$4^4$</td>
<td>590</td>
</tr>
<tr>
<td>Staggered $D_{eo}$</td>
<td>$2^4$</td>
<td>370</td>
</tr>
<tr>
<td>Staggered $D_{eo}$</td>
<td>$2^2 . 4^2$</td>
<td>430</td>
</tr>
<tr>
<td>Asqtad $D_{eo}$</td>
<td>$4^4$</td>
<td>440</td>
</tr>
</tbody>
</table>

| **Comparison with compiled code:** |            |             |
| CloverTerm/asm           | -          | 790         |
| CloverTerm/gcc, no dcbt  | -          | 150         |
| CloverTerm/xlc, no dcbt  | -          | 300         |

| **Generic (MILC) C code with QMP:** |            |             |
| Staggered $D_{eo}$         | $2^4$      | 171         |
| Staggered $D_{eo}$         | $4^4$      | 207         |
| Asqtad $D_{eo}$            | $4^4$      | 157         |
| Asqtad Force               | $2^4$      | 140         |
| Asqtad Force               | $4^4$      | 200         |

From gate-level simulation of ASIC at “nominal” 500MHz speed. Communication, cache flush overhead and 20 ns pin-pin wire delay included.

† [Boyle, Jung, Wettig http://arXiv.org/abs/hep-lat/0306023]
32K-node machine will deliver 15 Tflops at $1/Mflops.
OUTLOOK

• Phase I construction: $1.0 \times 3$ Tf, Jan-Feb 2004.

• Large-scale machines:
  (Performance numbers are sustained for LGT.)
  – RBRC: 5 Tf, $5M$, funded, July 2004.
  – UKQCD: 5 Tf, $5M$, funded, July 2004.

• Break-through discovery potential in High Energy and Nuclear Physics.