Lattice Gauge Computing

Summary

Don Holmgren
Fermilab
Outline

- Communities / Collaborations
- Existing Facilities
- LQCD on Purpose-Built Machines
- LQCD on Clusters
Lattice Gauge Computing Requirements

- Very large floating point requirements:
  - Need 100’s to 1000’s of Gflop/sec-years for results
  - Primary work is inverting large 4-D sparse matrices
  - Sustained performance is often low (20% of peak)
  - Multiprocessors are essential – clusters or massively parallel

- Inter-processor communications requirements:
  - High bandwidth (100’s Mbyte/sec)
  - Low latency - O(μsec)
Communities / Collaborations

LATFOR (Lattice Forum) (K. Jansen)
- Forum of German lattice physicists, plus groups in Austria (Graz, Wien) and Switzerland (Bern)
- Universities: Berlin, Bielefeld, Darmstadt, Leipzig, Münster, Regensburg, Wuppertal
- Research Labs: DESY, GSI, NIC

Efforts:
- Coordinate physics program
- Share software
- Share configurations/propagators (ILDG = International Lattice Data Grid)
Communities / Collaborations

LATFOR - Requirements

- Typical application profile:
  - Lattice Size: $32^3 \times 64$
  - Memory: 100 Gbyte
  - I/O request: 0.1 Gbyte/Gflop
  - Runtime: 15 Teraflops-years

- Need 12.5 Teraflops sustained
Communities / Collaborations

- SciDAC Lattice Gauge Computing
  - Majority of US community
  - Labs: Brookhaven, Fermilab, Jefferson Lab
  - Universities: Indiana, Utah, UCSB, UCSD, MIT, BU, ASU, Illinois, OSU, Columbia…
  - [http://www.lqcd.org/](http://www.lqcd.org/)

- Efforts
  - QCDOC (Columbia/BNL)
  - Cluster Prototypes (Fermilab, JLAB)
  - Software (QMP, QLA, QIO, QDP) to allow applications to run on either type of hardware (Chulwoo Jung)
  - O(10 Tflops) in 2004 (QCDOC), 2005/6 (clusters)
Facilities

SciDAC:

- Jefferson Lab
  - Now: 128 nodes / 128 processors (2.0 GHz Xeon), Myrinet
  - Soon: 256 nodes / 256? Processors (Xeon), GigE mesh

- Fermilab
  - Now: 176 nodes / 352 processors (2.0/2.4 GHz Xeon), Myrinet
  - Autumn: O(256) processor expansion (Xeon? Itanium2?)
  - Myrinet? GigE?
Facilities

- DESY Clusters:
  - DESY Hamburg
    - 16 Dual 1.7 GHz Xeon
    - 16 Dual 2.0 GHz Xeon
    - Myrinet
  - DESY Zeuthen
    - 16 Dual 1.7 GHz Xeon
    - Myrinet
# APEmille installations

<table>
<thead>
<tr>
<th>Location</th>
<th>GF</th>
<th>Crates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bielefeld</td>
<td>130</td>
<td>2</td>
</tr>
<tr>
<td>Zeuthen</td>
<td>520</td>
<td>8</td>
</tr>
<tr>
<td>Milan</td>
<td>130</td>
<td>2</td>
</tr>
<tr>
<td>Bari</td>
<td>65</td>
<td>1</td>
</tr>
<tr>
<td>Trento</td>
<td>65</td>
<td>1</td>
</tr>
<tr>
<td>Pisa</td>
<td>325</td>
<td>5</td>
</tr>
<tr>
<td>Rome 1</td>
<td>520</td>
<td>8</td>
</tr>
<tr>
<td>Rome 2</td>
<td>130</td>
<td>2</td>
</tr>
<tr>
<td>Orsay</td>
<td>16</td>
<td>1/4</td>
</tr>
<tr>
<td>Swansea</td>
<td>65</td>
<td>1</td>
</tr>
</tbody>
</table>

**Gr. Total** ~1966 GF
LQCD On Purpose-Built Hardware

QCDOC: (T. Wettig)

- “QCD on a Chip” ASIC (0.18µ):
  - Partner with IBM
  - Based on PPC940 + 64-bit FPU, 1 Gflops peak
  - 4 MB EDRAM + controller for external DDR
  - 6 bidirectional LVDS channels, each 2 X 500 Mbps
  - Ethernet
  - ~ 5 Watts
  - $1/MF assuming 50% of peak

Packaging:
  - 2 ASICs/daughterboard, 32 daughterboards per motherboard
  - 8 motherboards per backplane (512 processors)
QCDOC

Schedule:
- Funded: 10 Tflops each in late 2003 for Edinburgh, RIKEN-BNL
  5 Tflops for Columbia (SciDAC) [pending]
- 10 – 20 Tflops for US @ BNL in 2004
- ASIC design is done, 1st chips in May
Performance: (P. Boyle)

- Single node kernels, based on simulations:
  - SU3 x SU3  800 Mflops/node
  - SU3 x 2spinor  780 Mflops/node

- Multinode:
  - Example – Wilson $2^4$, 470 Mflops/node, 22µsec/iteration, each iteration has 16 distinct communications
Fast global sums via network passthru:

**Estimated Scalability**

Based on cycle accurate simulation. *Wilson* CG performance on a $32^3 \times 64$ Lattice. Clover and Domain Wall Fermions even more scalable.

<table>
<thead>
<tr>
<th>Nodes</th>
<th>$M^\dagger M$</th>
<th>Gsum</th>
<th>Sust. T\text{flops}</th>
</tr>
</thead>
<tbody>
<tr>
<td>4096</td>
<td>2620\mu s</td>
<td>10\mu s</td>
<td>2.15</td>
</tr>
<tr>
<td>8192</td>
<td>1310\mu s</td>
<td>11.5\mu s</td>
<td>4.2</td>
</tr>
<tr>
<td>16384</td>
<td>680 \mu s</td>
<td>13\mu s</td>
<td>8.1</td>
</tr>
<tr>
<td>32768</td>
<td>340 \mu s</td>
<td>15\mu s</td>
<td>15.6</td>
</tr>
</tbody>
</table>
In 0.18µm ASIC, native implementation of “fused multiply-add”:

\[ a \times b + c \] (complex numbers)

256 registers, each holding a complex number (double precision)

200 MHz, ~5 watts, 1.6 Gflops/chip, $1/Mflop @ 50% of peak

6+1 channels of LVDS communications, each 200 Mbyte/sec
  - 3D torus (x, y, z)
  - Fast I/O link to UNIX host

Schedule:
  - 300 to 600 ASIC’s to be delivered in June `03
  - From these, assemble and test 256-processor crate
  - Mass production to start in Sept `03
The apeNEXT architecture (2)

- Two directions (Y,Z) on the backplane
- Direction X through front panel cables

System topologies:

- Processing Board: $4 \times 2 \times 2 \sim 26 \text{ GF}$
- subCrate (16 PB): $4 \times 8 \times 8 \sim 0.4 \text{ TF}$
- Crate (32 PB): $8 \times 8 \times 8 \sim 0.8 \text{ TF}$
- Large systems: $(8^*n) \times 8 \times 8$
LQCD on Clusters

- Performance (DESY: P. Wegner, Fermilab: D. Holmgren)
- Common themes:
  - SSE/SSE2 very important for optimizing performance
  - Large memory bandwidth increase in Pentium 4 critical to LQCD
  - Building clusters with Myrinet, investigating other networks
  - Pay attention to PCI bus performance, varies with chipset
Parallel (1-dim) Dirac Operator Benchmark (SSE), even-odd preconditioned, $2 \times 16^3$, XEON CPUs, single CPU performance

- **Myrinet2000**
  - i860: 90 MB/s
- **E7500**
  - 190 MB/s
## Maximal Efficiency of external I/O

<table>
<thead>
<tr>
<th>System</th>
<th>MFLOPs (without communication)</th>
<th>MFLOPS (with communication)</th>
<th>Maximal Bandwidth</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Myrinet (i860), SSE</td>
<td>579</td>
<td>307</td>
<td>90 + 90</td>
<td>0.53</td>
</tr>
<tr>
<td>Myrinet/GM (E7500), SSE</td>
<td>631</td>
<td>432</td>
<td>190 + 190</td>
<td>0.68</td>
</tr>
<tr>
<td>Myrinet/Parastation (E7500), SSE</td>
<td>675</td>
<td>446</td>
<td>181 + 181</td>
<td>0.66</td>
</tr>
<tr>
<td>Myrinet/Parastation (E7500), non-blocking, non-SSE</td>
<td>406</td>
<td>368</td>
<td>hidden</td>
<td>0.91</td>
</tr>
<tr>
<td>Gigabit, Ethernet, non-SSE</td>
<td>390</td>
<td>228</td>
<td>100 + 100</td>
<td>0.58</td>
</tr>
<tr>
<td>Infiniband non-SSE</td>
<td>370</td>
<td>297</td>
<td>210 + 210</td>
<td>0.80</td>
</tr>
</tbody>
</table>
Infiniband interconnect
LQCD on Clusters - Management

- See talks by A. Gellrich, A. Singh

Common themes:
- Linux, PBS, Myrinet, MPI, private networks
- Web based monitoring and alarms
- MRTG history plots (e.g. temperatures, fans)
- Some Myrinet hardware reliability issues