Application Performance on Commodity-class and High-end Computers

Martyn F. Guest
CCLRC Daresbury Laboratory
m.f.guest@daresbury.ac.uk

Acknowledgements:
P. Sherwood, H.J. van Dam, I.J. Bush, DL
J. Nieplocha, E. Apra, PNNL

http://www.cse.cclrc.ac.uk/disco/hw-perf.shtml
Outline

- **Background - Distributed computing at DL**
- **Commodity-based and High-end Systems**
  - Single-node performance & the Interconnect bottleneck
  - Prototype Commodity Systems: CS1 - CS20
  - High-end Systems: IBM SP/p690+, SGI Altix 3700, plus HP/Compaq Alpha Server SC and SGI Origin 3800
  - Performance Metrics
- **Application performance**
  - Distributed vs. Replicated Data
  - Molecular Simulation (DLPOLY, DLMULTI and CHARMM)
  - Electronic Structure - Global Arrays (GAs) and Linear Algebra (PeIGS)
    - NWChem and GAMESS-UK
  - Materials Simulation (CPMD) and Computational Engineering (ANGUS, SBLI)
- **Summary**
### Capability and Capacity Computing - Cost and Performance

<table>
<thead>
<tr>
<th>Specification</th>
<th>Usage</th>
<th>Cost Units</th>
<th>CPU</th>
<th>Memory</th>
<th>I/O</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MPP/ASCI</strong></td>
<td><strong>HPC community</strong></td>
<td>10,000</td>
<td>2,000</td>
<td>3,000</td>
<td>200-300</td>
</tr>
<tr>
<td>1600 CPUs, IBM p690+ (1.5 TB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SMP</strong></td>
<td><strong>Capability</strong></td>
<td>100</td>
<td>20</td>
<td>25</td>
<td>20-30</td>
</tr>
<tr>
<td>16-processor SGI Altix 350, Itanium 2 (10GB RAM)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PC</strong></td>
<td><strong>Capacity</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pentium-4 Xeon / 3.4GHz (512 MByte, 30 GB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Commodity Systems (1.5 x N ??)**

3 year continual access to 32 CPUs:
High-end (£0.5 / CPU hour) : £420,000
in-house commodity system : £50,000
PNNL’s HPCS2 - 1960 Processor HP Supercluster

  - 1,960 Intel® Madison CPUs
  - QSNet2/Elan4 interconnect from Quadrics

- Timetable:
  - Prototype system (2002): 32 dual-CPU nodes with prototype McKinley processors
    - 256 peak GF, 256GB of total memory.
    - HP zx1 chipset (four independent PCI-X busses, one at 1GB/sec) - 12.8GBs memory bandwidth: Quadrics QSNet1/Elan3 interconnect
  - Phase 1 System (2002)
    - 128 Longs Peak nodes with 256 McKinley CPUs (1 TF, 572GB memory).
    - 1.0-GHz McKinley processors (3MB of on-chip L2 cache (86 % peak DGEMM).
  - Phase 2a System (2003).
    - 980 nodes with Madison/1.5 GHz CPUs.
    - 1960 IA-64 CPUs (6.9TB of memory), 53TB of global + 242 TB node storage).
    - QSNet2/Elan4 interconnect from Quadrics (1GB/sec PCI X2 bus - late 2003).
DisCo: Technical Progress in 2003-4

Hardware and Software Evaluation:

- **CPU**
  - IA32, x-86 and IA64 systems -
    - Intel Pentium 4 and Xeon Systems (3.06 GHz),
      AMD Opteron 246 & 248 (2.0 & 2.2 GHz)
  - Itanium2 (Intel Tiger 1, 1.2 and 1.5 GHz;
    HP systems, 900 MHz, 1 and 1.5 GHz;
    SGI Altix 3700 - 1.3 (3MB L3) & 1.5 GHz (6MB L3)

- **Networks**
  - Gigabit Ethernet options, cards, switches, channel-bonding, ...
  - SCI, Infiniband and Myrinet (P4/2400, P4/2666,
    P4/2800, Opteron 246 & 248 Clusters:
    OCF, Streamline, ClusterVision & Workstations UK), Quadrics

- **System Software**
  - message passing S/W (LAM MPI, LAM MPI-VIA, MPICH, VMI, SCAMPI), libraries
    (ATLAS, NASA, MKL, ACML, ScaLAPACK), compilers (Absoft, PGI, Intel’s ifc and
efc, Pathscale, GNU/g77), tools (GA tools, PNNL)
  - resource management software (PBS, TORQUE, GridEngine, LSF etc.)
Working with the Community

Chemistry Department, University of Bristol (Gabriel Balint Kurti)

- Provided initial benchmarking Suite incorporated in the procurement process.
- Exercised initial machine as part of acceptance tests (early user access), working with Bristol support staff (Ian Stewart)
  - Installed and validated wide variety of application codes as part of this process (NWChem, GAMESS-UK, DL_POLY, CHARMM, CPMD etc.)
  - Revealed shortcomings in initial Myrinet software
- Presented findings in visit to the University post-acceptance
- Provided application-code base to User Community (subject to appropriate licencing etc.)
- Potential “applications-driven” model for future support activities

Biochemistry Department, University of Essex (Chris Reynolds)
Working with the Community II.

Institute of Pharmaceutical Innovation, University of Bradford

- Exercised initial machine during acceptance tests (early user access), working with Bradford support staff (Victoria Pennington)
  - Installed and validated wide variety of application codes as part of this process (NWChem, GAMESS-UK, DL_POLY, CHARMM, CPMD etc.)
  - Revealed shortcomings in initial Myrinet / Redhat environment
- Presented findings in visit to the University post-acceptance
- Provided application-code base to User Community (subject to appropriate licencing etc.)
# Commodity Systems (CSx)

## Prototype / Evaluation Hardware

<table>
<thead>
<tr>
<th>Systems</th>
<th>Location</th>
<th>CPUs</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS1</td>
<td>Daresbury</td>
<td>32</td>
<td>PentiumIII / 450 MHz + FE (EPSRC)</td>
</tr>
<tr>
<td>CS2</td>
<td>Daresbury</td>
<td>64</td>
<td>24 X dual UP2000/EV67-667, QSNet Alpha/LINUX cluster, 8 X dual CS20/EV67-833 (“loki”)</td>
</tr>
<tr>
<td>CS3</td>
<td>RAL</td>
<td>16</td>
<td>Athlon K7 850MHz + myrinet</td>
</tr>
<tr>
<td>CS4</td>
<td>Sara</td>
<td>32</td>
<td>Athlon K7 1.2 GHz + FE</td>
</tr>
<tr>
<td>CS6</td>
<td>CLiC</td>
<td>528</td>
<td>PentiumIII / 800 MHz; fast ethernet (Chemnitzer Cluster)</td>
</tr>
<tr>
<td>CS7</td>
<td>Daresbury</td>
<td>64</td>
<td>AMD K7/1000 MP + SCALI/SCI (“ukcp”)</td>
</tr>
<tr>
<td>CS8</td>
<td>NCSA</td>
<td>320</td>
<td>160 dual IBM Itanium/800 + Myrinet 2k (“titan”)</td>
</tr>
<tr>
<td>CS9</td>
<td>Bristol</td>
<td>96</td>
<td>Pentium4 Xeon/2000 + Myrinet 2k (“dirac”)</td>
</tr>
</tbody>
</table>

## Prototype Systems

<table>
<thead>
<tr>
<th>Systems</th>
<th>Location</th>
<th>CPUs</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS0</td>
<td>Daresbury</td>
<td>10</td>
<td>10 CPUS, Pentium II/266</td>
</tr>
<tr>
<td>CS5</td>
<td>Daresbury</td>
<td>16</td>
<td>8 X dual Pentium III/933, SCALI</td>
</tr>
</tbody>
</table>
## Commodity Systems (CSx) II.

<table>
<thead>
<tr>
<th>Systems</th>
<th>Location</th>
<th>CPUs</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS10</td>
<td>Hull</td>
<td>64</td>
<td>Pentium4 Xeon/2667 + Myrinet 2k (&quot;eagle&quot;), Streamline/SCORE</td>
</tr>
<tr>
<td>CS11</td>
<td>Workstations</td>
<td>32</td>
<td>Pentium4 Xeon/2667 + GbitEther, ScaMPI</td>
</tr>
<tr>
<td>CS12</td>
<td>Essex</td>
<td>48</td>
<td>Pentium4 Xeon/2400 + GbitEther (&quot;sstream1&quot;), Streamline/SCORE</td>
</tr>
<tr>
<td>CS13</td>
<td>White Rose, Leeds</td>
<td>256</td>
<td>Pentium4 Xeon/2200-2400 + M2k (&quot;snowdon&quot;), Streamline/SCORE</td>
</tr>
<tr>
<td>CS14</td>
<td>NCSA</td>
<td>1024</td>
<td>Pentium III Xeon/1000 + M2k (&quot;platinum&quot;)</td>
</tr>
<tr>
<td>CS15</td>
<td>SDSC</td>
<td>128</td>
<td>Pentium III Xeon/800 + M2k (&quot;meteor&quot;)</td>
</tr>
<tr>
<td>CS16</td>
<td>SDSC</td>
<td>256</td>
<td>dual-Itanium2/1.3 GHz + M2k (&quot;Teragrid&quot;)</td>
</tr>
<tr>
<td>CS17</td>
<td>Daresbury</td>
<td>32</td>
<td>Pentium4 Xeon/2667 + GbitEther (&quot;ccp1&quot;), Streamline/SCORE</td>
</tr>
<tr>
<td>CS18</td>
<td>Bradford</td>
<td>78</td>
<td>Pentium4 Xeon/2800 + M2k/GbitE (&quot;grendel&quot;)</td>
</tr>
<tr>
<td>CS19</td>
<td>Daresbury</td>
<td>64</td>
<td>dual-Opteron/246 2.0 GHz nodes + Infiniband, Gbit and SCI (&quot;scaliwag&quot;)</td>
</tr>
<tr>
<td>CS20</td>
<td>RAL</td>
<td>256</td>
<td>dual-Opteron/248 2.2 GHz nodes + Myrinet (&quot;scarf&quot;)</td>
</tr>
</tbody>
</table>
High-End Systems Evaluated

- Cray T3E/1200E (… historical …)
  - 816 processor system at Manchester (CSAR), 600 Mz Alpha EV56 CPU, 256 MB
- IBM pseries 690 and pseries 690+ (Daresbury)
  - *IBM p690* (8-way LPAR’d nodes, 1280 X 1.3 GHz CPUs with colony, HPCx)
  - *IBM p690+* (32-way nodes, 1600 X 1.7 GHz CPUs with HPS, HPCx- Phase2)
- Compaq AlphaServer SC
  - 4-way ES40/667 A21264A (APAC) and 833 MHz SMP nodes (2 GB RAM);
  - *TCS1 system at PSC* (750 4-way ES45 nodes - 3,000 EV68 CPUs - 4 GB memory per node, 8MB L2 cache), Quadrics interconnect (5 usec latency, 250 MB/sec B/W)
- SGI Origin 3800
  - SARA (1000 CPUs) - Numalink with MIPS R14k/500 CPUs
- SGI Altix 3700
  - *Linux Cluster - Numalink with Itanium 2 1.3 GHz CPUs, 3MB L3 cache*
    - CSAR (“newton” 512 CPUs) and SARA (“aster” - 416 CPUs - 7 nodes)
    - ORNL (“ram” 256 CPUs with Itanium 2 1.5 GHz CPUs, 6MB L3 cache)
Applications Performance Overview

- **Serial (SPEC, DL) & Communication Benchmarks**
- **Parallel Applications Performance**

1. Computational Chemistry:
   - Molecular Simulation &
   - Electronic Structure
2. Computational Materials Science
3. Atomic & Molecular Physics
4. Computational Engineering
5. Environmental Modelling

- Capacity-based group solution
- Issues of Cost effectiveness
- On e.g. 128-256 CPU cluster, modal job size is ~ 32 CPUs
- Increasing trend to hierarchical clusters - Gbit network with HEC core (with e.g. myrinet)

**Capability and Capacity Computing**

**Commodity vs. Proprietary Solutions**

Attempted to quantify delivered performance from the Commodity-based systems against MPP (CSAR Cray T3E/1200E) and ASCI-style SMP-node platforms (e.g. SGI Origin 3800) i.e.

Performance Metric (% 32-node Cray T3E)

\[
\frac{T_{32\text{-node} \text{ Cray T3E}}}{T_{32\text{-CPU} \text{ CS2 Pentium III/450 + FE}}} \times T_{32\text{-node} \text{ Cray T3E}} / \frac{T_{32\text{-node} \text{ Cray T3E}}}{T_{32\text{-node} \text{ Cray T3E}}}
\]

\[
\frac{T_{32\text{-node} \text{ Cray T3E}}}{T_{32\text{-node} \text{ Cray T3E}}}
\]

\[
\frac{T_{32\text{-node} \text{ Cray T3E}}}{T_{32\text{-node} \text{ Cray T3E}}}
\]

\[
\frac{T_{32\text{-node} \text{ Cray T3E}}}{T_{32\text{-node} \text{ Cray T3E}}}
\]

\[
\frac{T_{32\text{-node} \text{ Cray T3E}}}{T_{32\text{-node} \text{ Cray T3E}}}
\]

\[
\frac{T_{32\text{-node} \text{ Cray T3E}}}{T_{32\text{-node} \text{ Cray T3E}}}
\]

Performance Metrics: 2002

Performance Metric (% 32-node AlphaServer SC [PSC])

\[
\frac{T_{32\text{-CPU} \text{ AlphaServer SC ES45/1000}}}{T_{32\text{-CPU} \text{ CS9 Pentium 4 Xeon / 2000 + Myrinet 2k}}}
\]
Commodity Comparisons with High-end Systems

% of 32 CPUs of Cray T3E/1200E

Cluster CS6
Pentium3/800 + Fast Ethernet

CS6 - 98% of Cray T3E

2000
Commodity Comparisons with High-end Systems

% of 32 CPUs of Compaq AlphaServer SC ES45/1000

Cluster CS9
Pentium4/2000 Xeon + Myrinet

CS9 - 66% of ES45/1000

- ANGUS (288)
- ANGUS (144)
- CPMD
- CASTEP
- CHARMM
- DL_POLY_2 (Macromolecular)
- DL_POLY-2 (Ionic)
- GAMESS-UK + CHARMM
- NWChem (DFT Jfit)
- GAMESS-UK (HF Forces)
- GAMESS-UK (MP2 gradient)
- GAMESS-UK (DFT gradient)
- GAMESS-UK (DFT-Jfit)
- GAMESS-UK (DFT)
- GAMESS-UK (SCF)

2002
### SPEC CPU 2000 - SPECfp2000

Values relative to HP RX5670 Itanium2/1.5GHz

<table>
<thead>
<tr>
<th>System</th>
<th>SPEC CPU 2000</th>
<th>SPECfp2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMD A7N8X XP3200+/2.2 GHz</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>AMD MSI K8N Athlon64/2.6GHz</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>Intel D925XCV (3.6 GHz P4)</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>HP RX5670 Itanium2/1500</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>HP RX5670 Itanium2/1300</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>HP AlphaServer ES45 68/1250</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>HP AlphaServer GS1280 7/1300</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>Sun Blade 2000 / 1200 MHz</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>Sun Java Wrkst. (Opteron/2.4 GHz)</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>Fujitsu PP900 GP64/1890</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>SGI Origin 3200/R14k-600</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>SGI Altix 3000 Itanium 2/1300</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>SGI Altix 3000 Itanium 2/1500</td>
<td>102</td>
<td></td>
</tr>
<tr>
<td>HP 9000/C3750-875</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>IBM eServer p5 570/1900</td>
<td>102</td>
<td></td>
</tr>
<tr>
<td>IBM p-series 690/1700</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>IBM p-series 690/1300</td>
<td>81</td>
<td></td>
</tr>
</tbody>
</table>

2161 : HP RX4640 Madison/1500
2148 : SGI Altix Madison/1500
2702 : IBM p5 570 / 1.9 GHz

*Note: Values are relative to HP RX5670 Itanium2/1.5GHz.*
The GAMESS-UK Serial Benchmark

Performance relative to the HP RX5670 Itanium2/1.5GHz

<table>
<thead>
<tr>
<th>System</th>
<th>Performance Relative</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMD MP2400+ / 2000</td>
<td>53</td>
</tr>
<tr>
<td>Pentium 4 Xeon / 3066</td>
<td>78</td>
</tr>
<tr>
<td>SGI Altix3700 Itanium2/1500</td>
<td>79</td>
</tr>
<tr>
<td>HP RX4640 Itanium2/1300-H</td>
<td>85</td>
</tr>
<tr>
<td>HP RX2600 Itanium2/1500-L</td>
<td>94</td>
</tr>
<tr>
<td>HP RX5670 Itanium2/1500-H</td>
<td>100</td>
</tr>
<tr>
<td>Intel Tiger Itanium2/1500</td>
<td>93</td>
</tr>
<tr>
<td>AMD Opteron 848/ 2200</td>
<td>98</td>
</tr>
<tr>
<td>AMD Opteron 244/ 1800</td>
<td>83</td>
</tr>
<tr>
<td>Compaq Alpha ES45/1250</td>
<td>58</td>
</tr>
<tr>
<td>Compaq Marvel EV7 /1000</td>
<td>53</td>
</tr>
<tr>
<td>SUN Blade 2000 / 1056 Cu</td>
<td>40</td>
</tr>
<tr>
<td>SUN FireV880 / 900 Cu</td>
<td>33</td>
</tr>
<tr>
<td>SGI Origin3800/R14k-600</td>
<td>29</td>
</tr>
<tr>
<td>HP PA-9000/RP7410-875</td>
<td>57</td>
</tr>
<tr>
<td>IBM p-series 690/1700</td>
<td>73</td>
</tr>
<tr>
<td>IBM p-series 690/1300</td>
<td>100</td>
</tr>
</tbody>
</table>

2.3 minutes
Single CPU performance - A Case Study

- Scalability of Terascale applications is only part of the story …
- Absolute performance also depends on single cpu performance
- % of peak is seen as an important measure
- Comparison with other systems e.g. vector machines

Ran representative test cases on small numbers of IBM p690 processors for applications and some important kernels
Used IBM’s hpmlib to measure Mflop/s
Other hpmlib counters can help to understand performance e.g. memory bandwidth, cache miss rates, FMA count, computational intensity etc.
# Interconnects and Networking

- Ethernet, Gbit etc.
- Myrinet, Quadrics and Dolphin SCI
  - 64 bit PCI implementations, with faster PCI-X options near market
- Infiniband PCI-X interconnect

Key to any interconnect is the performance of the library implementation.

- MPICH (ANL)
- ScaMPI library from Scali
- PMB MPI Benchmarks (Pallas)

## Performance of current interconnects.

<table>
<thead>
<tr>
<th>Interconnect</th>
<th>Latency (µs)</th>
<th>Bandwidth (MB/s)</th>
<th>Switch size</th>
<th>Message size (kB)</th>
<th>1MB transmit (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadrics Elan3</td>
<td>5</td>
<td>325 (680)</td>
<td>128-2000</td>
<td>1.6</td>
<td>3.08</td>
</tr>
<tr>
<td>Dolphin SCI</td>
<td>5 (1.5 intranode)</td>
<td>326</td>
<td>N/A</td>
<td>1.6</td>
<td>3.08</td>
</tr>
<tr>
<td>Myrinet</td>
<td>~8 (6.3)</td>
<td>243 (500)</td>
<td>8-128</td>
<td>2</td>
<td>4.12</td>
</tr>
<tr>
<td>Mellanox Infiniband</td>
<td>7-10 (5.5)</td>
<td>800 (830)</td>
<td>8-96</td>
<td>5.6-8</td>
<td>1.26</td>
</tr>
<tr>
<td>Gbit. Ethernet</td>
<td>30-100</td>
<td>125</td>
<td>64</td>
<td>3.7-12.5</td>
<td>8.1</td>
</tr>
</tbody>
</table>
Communication Benchmarks
PMB: Pallas MPI Benchmark Suite (V2.2) and B_EFF

High-end Systems
- Cray T3E/1200E
- SGI Origin3800/R14k-500 (Teras)
- IBM SP/WH2-375 (“trailblazer” switch and SP/NH2-375 (single-plane colony)
- IBM p690 (8-way LPAR, HPCx - colony “Phase 1”)
- IBM p690+ (32-way SMP, HPCx - HPS “Phase2”)
- Compaq AlphaServer SC ES45/1GHz - TCS1
- SGI Altix3700/Itanium2-1.3GHz (Newton) and 1.5 GHz (Ram)

Commodity-based Systems
- CS2 Alpha Linux Cluster dual UP2000/667
- CS8 Itanium/800 + Myrinet 2k (NCSA)
- CS9 dual P4/2000 Xeon + Myrinet 2k
- CS10 P4/2666 + Myrinet 2k (Streamline, Hull U.)
- CS12 P4/2400 + GbitEther (Streamline, Essex U.)
- CS13 P4/2200-2400 + Myrinet 2k (Streamline, Leeds U.)
- CS14 PIII/1000 + Myrinet (VMI, NCSA)
- CS16 Itanium2 1300 + Myrinet2k (SDSC)
- CS18 P4/2800 + Myrinet 2k, GBitE (Clustervision, Bradford U.)
- CS19 Opteron 246/2000 + Infiniband, SCI, GbitEther (OCF, Daresbury).
- CS20 Opteron 248/2200 + Myrinet, (Streamline, RAL).
PingPong Performance

Commodity Systems

PMB Benchmark (V2.2, Pallas)

Bandwidth (MByte/sec)

Message Length (Bytes)

- Cray T3E/1200E
- CS18 Pentium 4/2800 + Myrinet 2k
- CS15 Itanium2/1300 + Myrinet 2k
- CS10 Pentium 4/2666 + Myrinet 2k
- CS14 Platinum/PIII 1000 + Myrinet
- CS19 Opteron 246/2.0 + SCI
- CS19 Opteron 246/2.0 + IB
- CS12 Pentium 4/2400 + GbitEther
- CS18 Pentium 4/2800 + GbitEther
Interconnect Benchmark - EFF_BW

- SGI Origin 3800/R14k-500
- SGI Altix 3700/Itanium2 1300
- AlphaServer SC ES45/1000 (1CPU)
- AlphaServer SC ES45/1000 (4CPU)
- IBM p690 (16 - 16X1)
- CS19 Opteron 246/2000 + Infiniband (1 CPU)
- Cray T3E/1200E
- CS2 QSNet AlphaEV67 (1 CPU)
- CS2 QSNet AlphaEV67 (2 CPU)
- CS10 P4/2666 Xeon - Myrinet 2k (1 CPU)
- CS10 P4/2666 Xeon - Myrinet 2k (2 CPU)
- CS14 PIII/1000 Xeon - Myrinet/VMI (1 CPU)
- CS14 PIII/1000 Xeon - Myrinet/VMI (2 CPU)
- CS19 Opteron 246/2000 + SCI (1 CPU)
- CS18 P4/2400 Xeon - GbitEther (1 CPU)
- CS12 P4/2400 Xeon - GbitEther (1 CPU)
- CS11 P4/2400 Xeon - GbitEther (1 CPU)
- CS11 P4/2400 Xeon - GbitEther (2 CPU)
- CS6 PIII/800 + FastEther (LAM)

Interconnects:
- QSNet
- Infiniband
- Myrinet 2k
- Myrinet
- SCALI
- Gbit Ethernet
- Fast Ethernet

Throughput in MBytes/sec.
Collective Operations
(Time - usec) - as function of no. of CPUs

Communications among groups of processors in the cluster e.g.
- Distribute data to all nodes in the cluster (scatter, all-to-all)
- Collect data from all nodes (gather)
- Collect information from all nodes to determine an overall condition e.g. minima or maxima (reduce).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b,c,d</td>
<td>e,f,g,h</td>
<td>i,j,k,l</td>
<td>m,n,o,p</td>
<td>MPI_Alltoall</td>
<td>a,e,i,m</td>
<td>b,f,j,n</td>
<td>c,g,k,o</td>
<td>d,h,l,p</td>
</tr>
</tbody>
</table>

Alltoall

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>c</td>
<td>MPI_Allgather</td>
<td>a,b,c</td>
<td>a,b,c</td>
<td>a,b,c</td>
</tr>
</tbody>
</table>

Allgather
MPI_allreduce Performance

- Cray T3E/1200E
- CS18 Pentium 4/2800 + Myrinet 2k
- CS15 Itanium2/1300 + Myrinet 2k
- CS10 Pentium 4/2666 + Myrinet 2k
- CS14 Platinum/PIII 1000 + Myrinet
- CS19 Opteron 246/2.0 + SCI
- CS19 Opteron 246/2.0 + IB
- CS12 Pentium 4/2400 + GbitEther
- CS18 Pentium 4/2800 + GbitEther

Commodity Systems

Message Length (Bytes)

Measured Time (us)

PMB Benchmark (Pallas)

16 CPUs
64-CPU Relative Performance for Allreduce

- IBM p690+ // HPS - SP7
- IBM p690+ // HPS (Jun’04)
- SGI Origin3800/R14k-500
- Compaq AlphaServer SC ES45/1000
- SGI Altix 3700/Itanium2 1500
- SGI Altix 3700/Itanium2 1300 (PAM)
- CS19 Opteron 246/2.0 + SCI
- CS19 Opteron 246/2.0 + IB
- CS20 Opteron 248/2.2 + Myrinet

Message Length (Bytes)
128-CPU Relative Performance for Allreduce

Message Size (Bytes)

- IBM p690+ // HPS - SP7
- SGI Origin3800/R14k-500
- Compaq AlphaServer SC ES45/1000
- SGI Altix 3700/Itanium2 1300 (PAM)
- CS20 Opteron 248/2.2 + Myrinet
64-CPU Relative Performance for Allgather

- IBM p690+ // HPS - SP7
- SGI Altix 3700/Itanium2 1500
- SGI Altix 3700/Itanium2 1300 (PAM)
- SGI Origin3800/R14k-500
- Compaq AlphaServer SC ES45/1000
- CS19 Opteron 246/2.0 + SCI
- IBM p690+ // HPS (Jun'04)
- CS20 Opteron 248/2.2 + Myrinet

Performance (Relative to IBM p690+ HPS, SP7)

Message Length (Bytes)

IBM p690+, Service Pack 7
128-CPU Relative Performance for Allgather

- IBM p690+ // HPS - SP7
- SGI Altix 3700/Itanium2 1300 (PAM)
- SGI Origin3800/R14k-500
- Compaq AlphaServer SC ES45/1000
- CS20 Opteron 248/2.2 + Myrinet

Message Size (Bytes)
Application Codes

Application-driven performance comparisons between Commodity-based systems and current High-end platforms:

- **Computational Chemistry**
  - **Molecular Simulation**
    - DL_POLY, DLMULTI - parallel MD codes with many applications
    - CHARMM - macromolecular MD and energy minimisation
  - **Molecular Electronic Structure**
    - GAMESS-UK and NWChem - Ab initio Electronic structure codes
- **Materials**
  - CPMD, CASTEP - Car-Parrinello plane wave codes
- **Engineering**
  - ANGUS, SBLI - regular-grid domain decomposition engineering codes

**Performance Metric (% 32-node High-end system)**
Performance Metrics: 2004

Attempt to quantify delivered performance from current Commodity-based systems against high-end platforms: SGI Altix 3700/Itanium2 1.3 GHz (CSAR service) and IBM p690+ (HPCx service) i.e.

Performance Metric (% 32-node SGI Altix 3700/Itanium2 1.3 GHz)

\[
\frac{T_{32-CPU} \text{ SGI Altix 3700}}{T_{32-CPU} \text{ CS20 Opteron 248 / 2800 + Myrinet 2k}} \quad \text{and} \quad \frac{T_{32-CPU} \text{ SGI Altix 3700}}{T_{32-CPU} \text{ CS18 Pentium 4 Xeon / 2800 + Gbit Ethernet}}
\]
Molecular Simulation

Molecular Dynamics Codes:
DL_POLY, DLMULTI and CHARMM

- **DL_POLY**
  - Developed as CCP5 parallel MD code by W. Smith and T.R. Forester
    - UK CCP5 + International user community
    - DLPOLY_2 (replicated data) and DLPOLY_3 (distributed data)
  - 830 licences issued since 1994.
  - Areas of application:
    - liquids, solutions, spectroscopy, ionic solids, molecular crystals, polymers, glasses, membranes, proteins, metals, solid and liquid interfaces, catalysis, clathrates, liquid crystals, biopolymers, polymer electrolytes.
DL_POLY: A Parallel MD Package

- Parallel molecular dynamics code by W. Smith and T.R. Forester
- Wide selection of Periodic Boundaries and Integration Algorithms
- Wide Range of Target systems
  - Atomic systems & mixtures (Ne, Ar, etc.), ionic melts & crystals (NaCl, KCl etc.)
  - Polarisable ionics (ZSM-5, MgO etc.)
  - Molecular liquids, solids (CCl₄, Bz etc.) and ionics (KNO₃, NH₄Cl, H₂O etc.)
  - Synthetic polymers ([PhCHCH₂]ₙ etc.), Bio-polymers and macromolecules
  - Polymer electrolytes, Membranes, Aqueous solutions, Metals
- Migration from Replicated (V2) to Distributed data (V3)
  - Distribute atoms, forces across the nodes
    - More memory efficient to address larger cases (10⁵-10⁷)
  - Shake and short-ranges forces require only neighbour communications
    - Communications scale linearly with number of nodes
  - Coulombic energy remains global
    - Strategy depends on problem & machine characteristics
    - Adopt particle Mesh Ewald scheme
**DL_POLY Parallel Benchmarks (Cray T3E/1200)**

V2: Replicated Data

4. NaCl; Ewald, 27,000 ions
5. NaK-disilicate glass; 8,640 atoms, MTS+ Ewald
8. MgO microcrystal: 5,416 atoms

9. Model membrane/Valinomycin (MTS, 18,886)
7. Gramicidin in water (SHAKE, 12,390)
6. K/valinomycin in water (SHAKE, AMBER, 3,838)
1. Metallic Al (19,652 atoms, Sutton Chen)
3. Transferrin in Water (neutral groups + SHAKE, 27,593)
2. Peptide in water (neutral groups + SHAKE, 3993).

---

![Speed-up Graph](Image)
DL_POLY V2: Bench 4 - Commodity-based Systems

Performance

- NaCl; 27,000 ions, Ewald, 75 time steps, Cutoff=24Å

Graph showing performance with different systems and configurations. The chart includes
- CS12 P4/2400 + GbitE
- CS10 P4/2666 + Myrinet
- CS18 P4/2800 + GbitE
- CS18 P4/2800 + Myrinet
- CS19 Opteron246/2.0 + IB
- CS19 Opteron246/2.0 + SCI
- CS20 Opteron246/2.2 + Myrinet
- CS16 Itanium2/1300 + Myrinet
- SGI Altix 3700/Itanium2 1300

The graph compares performance across different number of CPUs (16 and 32) with T3E128 = 2.13 and 30%, 68%.
DL_POLY V2: Bench 5 - Commodity-based Systems

Performance

NaK-disilicate glass;
8,640 atoms, MTS + Ewald: 270 time steps

Number of CPUs

T3E₁₂₈ = 3.1
**DL_POLY V2: Bench 7 - Commodity-based Systems**

**Performance**

- **CS12 P4/2400 + GbitE**
- **CS10 P4/2666 + Myrinet**
- **CS18 P4/2800 + GbitE**
- **CS18 P4/2800 + Myrinet**
- **CS19 Opteron246/2.0 + SCI**
- **CS19 Opteron246/2.0 + IB**
- **CS20 Opteron246/2.2 + Myrinet**
- **SGI Altix 3700/Itanium2 1300**

**Gramicidin in water; rigid bonds and SHAKE, 12,390 atoms, 500 time steps**

**T3E_{128} = 2.4**
Migration from Replicated to Distributed data:

**DL_POLY-3 : Domain Decomposition**

- Distribute atoms, forces across the nodes
  - More memory efficient, can address much larger cases ($10^5$-$10^7$)
- Shake and short-ranges forces require only neighbour communication
  - Communications scale linearly with number of nodes
- Coulombic energy remains global
  - Strategy depends on problem and machine characteristics
  - Adopt Smooth Particle Mesh Ewald scheme
    - Includes Fourier transform smoothed charge density (reciprocal space grid typically 64x64x64 - 128x128x128)

**W. Smith and T.R. Forester**

http://www.cse.clrc.ac.uk/msi/software/DL_POLY
Conventional routines (e.g. fftw) assume plane or column distributions. A global transpose of the data is required to complete the 3D FFT and additional costs are incurred re-organising the data from the natural block domain decomposition. An alternative FFT algorithm has been designed to reduce communication costs. The 3D FFT are performed as a series of 1D FFTs, each involving communications only between blocks in a given column. More data is transferred, but in far fewer messages. Rather than all-to-all, the communications are column-wise only.
# DL_POLY3 Coulomb Energy Evaluation

## Commodity-based Systems

<table>
<thead>
<tr>
<th>Performance</th>
<th>Number of CPUs</th>
<th>DL_POLY3 Coulomb Energy Evaluation</th>
<th>NaCl Simulation:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16</td>
<td>216,000 ions, 200 time steps, Cutoff=12Å</td>
<td></td>
</tr>
</tbody>
</table>

### Commodity-based Systems

<table>
<thead>
<tr>
<th>System</th>
<th>Type</th>
<th>NaCl Simulation</th>
<th>DL_POLY3 Coulomb Energy Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS12 Pentium 4/2400 + Gbit Ethernet</td>
<td>Commodity based</td>
<td>216,000 ions, 200 time steps, Cutoff=12Å</td>
<td>49%</td>
</tr>
<tr>
<td>CS10 Pentium 4/2666 + Myrinet 2k</td>
<td>Commodity based</td>
<td>216,000 ions, 200 time steps, Cutoff=12Å</td>
<td>93%</td>
</tr>
<tr>
<td>CS18 P4/2800 + GbitE</td>
<td>Commodity based</td>
<td>216,000 ions, 200 time steps, Cutoff=12Å</td>
<td></td>
</tr>
<tr>
<td>CS18 P4/2800 + Myrinet</td>
<td>Commodity based</td>
<td>216,000 ions, 200 time steps, Cutoff=12Å</td>
<td></td>
</tr>
<tr>
<td>CS19 Opteron246/2.0 + SCI</td>
<td>Commodity based</td>
<td>216,000 ions, 200 time steps, Cutoff=12Å</td>
<td></td>
</tr>
<tr>
<td>CS19 Opteron246/2.0 + IB</td>
<td>Commodity based</td>
<td>216,000 ions, 200 time steps, Cutoff=12Å</td>
<td></td>
</tr>
<tr>
<td>CS20 Opteron248/2.2 + Myrinet</td>
<td>Commodity based</td>
<td>216,000 ions, 200 time steps, Cutoff=12Å</td>
<td></td>
</tr>
<tr>
<td>SGI Altix 3700/Itanium2 1300</td>
<td>Commodity based</td>
<td>216,000 ions, 200 time steps, Cutoff=12Å</td>
<td></td>
</tr>
</tbody>
</table>

### Graphical Representation

- **X-axis**: Number of CPUs (16, 32)
- **Y-axis**: Performance
- **Legend**:
  - CS12 Pentium 4/2400 + Gbit Ethernet
  - CS10 Pentium 4/2666 + Myrinet 2k
  - CS18 P4/2800 + GbitE
  - CS18 P4/2800 + Myrinet
  - CS19 Opteron246/2.0 + SCI
  - CS19 Opteron246/2.0 + IB
  - CS20 Opteron248/2.2 + Myrinet
  - SGI Altix 3700/Itanium2 1300

### Performance Metrics

- **DL_POLY3 Coulomb Energy Evaluation**
  - 49%
  - 93%

### NaCl Simulation

- 216,000 ions, 200 time steps, Cutoff=12Å
DL_POLY3 Macromolecular Simulations

Performance

High-end Systems

Gramicidin in water; rigid bonds + SHAKE: 792,960 ions, 50 time steps

Relies on alternative FFT algorithm designed to reduce communication costs.

DL_POLY3 Macromolecular Simulations

Number of CPUs

<table>
<thead>
<tr>
<th>Number of CPUs</th>
<th>SGI Origin 3800/R14k-500</th>
<th>IBM p690</th>
<th>AlphaServer SC ES45/1000</th>
<th>SGI Altix 3700/Itanium2 1300</th>
<th>SGI Altix 3700/Itanium2 1500</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>6</td>
<td>11</td>
<td>12</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>64</td>
<td>10</td>
<td>11</td>
<td>19</td>
<td>18</td>
<td>14</td>
</tr>
<tr>
<td>128</td>
<td>14</td>
<td>18</td>
<td>30</td>
<td>37</td>
<td>30</td>
</tr>
<tr>
<td>256</td>
<td>29</td>
<td>19</td>
<td>41</td>
<td>41</td>
<td>41</td>
</tr>
</tbody>
</table>
DL_POLY3 Macromolecular Simulations

**Performance**

- **Commodity-based Systems**

<table>
<thead>
<tr>
<th>System</th>
<th>Number of CPUs</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS12 Pentium 4/2400 + Gbit Ethernet</td>
<td>16</td>
<td>2.7%</td>
</tr>
<tr>
<td>CS10 Pentium 4/2666 + Myrinet 2k</td>
<td>16</td>
<td>2.8%</td>
</tr>
<tr>
<td>CS18 P4/2800 + GbitE</td>
<td>32</td>
<td>4.3%</td>
</tr>
<tr>
<td>CS18 P4/2800 + Myrinet</td>
<td>32</td>
<td>4.7%</td>
</tr>
<tr>
<td>CS19 Opteron246/2.0 + SCI</td>
<td>32</td>
<td>5.3%</td>
</tr>
<tr>
<td>CS19 Opteron246/2.0 + IB</td>
<td>32</td>
<td>4.9%</td>
</tr>
<tr>
<td>CS20 Opteron248/2.2 + Myrinet</td>
<td>32</td>
<td>5.0%</td>
</tr>
<tr>
<td>SGI Altix 3700/Itanium2 1300</td>
<td>32</td>
<td>7.1%</td>
</tr>
</tbody>
</table>

**Gramicidin in water; rigid bonds + SHAKE:** 792,960 ions, 50 time steps

- **52%, 88%**
DLMULTI

- **DL_MULTI** - A DL_POLY package to simulate rigid molecules with multipoles

- Allows molecular dynamics simulation of rigid molecules with the electrostatics described by a distributed multipole.

- Work has been carried out on implementing constant pressure simulations, involving extensive modifications of the existing DL_POLY code as well as addition of new code.

- The program has been tested on a variety of parallel architectures. Future work will initially focus on using the package to investigate the thermodynamic stability of polymorphs of organic crystals e.g. 5-and 6-azauracil
  - Electrostatics carried out by Ewald sums.
  - Direct space sum uses neighbour list. - No communication. Requires a large cutoff to avoid stability problems in the energy
  - Reciprocal space sum has to compute global sums – number of sums scales as \( l^{**2} \) where \( l \) is the multipole order.
DLMULTI: High-end and Commodity-based Systems

Performance

- CS12 P4/2400 + GbitE
- CS10 P4/2666 + Myrinet
- CS18 P4/2800 + Myrinet
- CS18 P4/2800 + GbitE
- CS19 Opteron246/2.0 + SCI
- CS19 Opteron246/2.0 + IB
- CS20 Opteron248/2.2 + Myrinet
- SGI Altix 3700/Itanium2 1300
- IBM p690 1.3 GHz
- AlphaServer SC ES45/1000

Bench 2: Relaxed Supercell of 57,024 atoms (5-azauracil), 2 time steps.
DLMULTI: Commodity-based Systems

Performance

Bench 2: Relaxed Supercell of 57,024 atoms (5-azauracil), 2 time steps.
CHARMM

• CHARMM (Chemistry at HARvard Macromolecular Mechanics) is a general purpose molecular mechanics, molecular dynamics and vibrational analysis package for modelling and simulation of the structure and behaviour of macromolecular systems (proteins, nucleic acids, lipids etc.)

• Supports energy minimisation and MD approaches using a classical parameterised force field.

• J. Comp. Chem. 4 (1983) 187-217

• Parallel Benchmark - MD Calculation of Carboxy Myoglobin (MbCO) with 3830 Water Molecules.

• QM/MM model for study of reacting species
  ■ incorporate the QM energy as part of the system into the force field
  ■ coupling between GAMESS-UK (QM) and CHARMM.
Parallel CHARMM Benchmark

Benchmark MD Calculation of Carboxy Myoglobin (MbCO) with 3830 Water Molecules: 14026 atoms, 1000 steps (1 ps), 12-14 Å shift.

Measured Time (seconds)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>104</td>
<td>114</td>
<td>83</td>
<td>72</td>
<td>64</td>
<td>61</td>
<td>54</td>
<td>42</td>
</tr>
<tr>
<td>32</td>
<td>89</td>
<td>64</td>
<td>59</td>
<td>61</td>
<td>54</td>
<td>51</td>
<td>46</td>
<td>37</td>
</tr>
<tr>
<td>64</td>
<td>64</td>
<td>51</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>46</td>
<td>25</td>
</tr>
</tbody>
</table>

T3E\textsubscript{16} = 665, T3E\textsubscript{128} = 106
Molecular Electronic Structure

*Ab initio* Electronic structure Codes:
NWChem and GAMESS-UK
CCP1: Molecular Electronic Structure

- Working Group (29 Members from 17 Universities)
  - Chaired by Prof P.J. Knowles (University of Birmingham)

- Software
  - GAMESS-UK, CRYSTAL and ChemShell

- Study Weekends and Workshops
  - QM/MM methods (St Andrews, 1995)
  - Quantum Chemistry on MPP Computers (Cambridge, 1995)
  - Quantum Mechanics of Large systems (Daresbury, 1996)
  - Ab Initio Molecular Dynamics (Daresbury, 1998)

- Flagship Projects

- DL Staff Support
  - M. F. Guest, P. Sherwood, J.M. Thomas and H.J.J. van Dam
Molecular Electronic Structure

- Replicated Data Strategy (RD)
  - Not efficient for large parallel computers due to both system size and parallel scalability limitations.
  - Problem size driven by memory available to one CPU, not the whole machine.
  - Communications overhead – results of parallel OP must be re-replicated so that each CPU has a current version i.e. expensive global communication - time taken grows with the number of CPUs.
  - RD does not fit well with many algorithms e.g. diagonalization - increasingly important stage of SCF/DFT calculations as the system size increases.

- Distributed Data Strategy (DD) - NWChem
  - Each CPU holds only part of each of the (major) data structures, overcoming many of the drawbacks of replicated data;
    - Limit on system size is now that given by all memory on the computer,
    - global communications largely avoided as re-replication is not required,
    - many of the algorithms that can not be effectively addressed by a RD strategy fit more naturally within a distributed data one.

RD & DD Strategies

http://www.emsl.pnl.gov/docs/nwchem/nwchem.html
Pictorial representation of the iterative SCF process in (i) a sequential process, and (ii) a distributed data parallel process: **MOAO** represents the molecular orbitals, **P** the density matrix, and **F** the Fock or Hamiltonian matrix.
Global Arrays

Physically distributed data

- Single, shared data structure

- Shared-memory-like model
  - Fast local access
  - NUMA aware and easy to use
  - MIMD and data-parallel modes
  - Inter-operates with MPI, ...

- BLAS and linear algebra interface

- Ported to major parallel machines
  - IBM, Cray, SGI, clusters,...

- Originated in an HPCC project

- Used by 5 major chemistry codes, financial futures forecasting, astrophysics, computer graphics

Tools developed as part of the NWChem project at PNNL; R.J. Harrison, J. Nieplocha et al.
Global Array Benchmark 1. GET

I. GET

Singularities caused by "cold start" conditions: 1st reference to memory across partitions leads to a kernel trap to address memory protection.

LAPI-based implementation on the IBM p690

Service Packs 7 & 9 on IBM p690+

Remote 2-D Array Section (GET)

- GET (p690+ HPS SP9)
- GET (p690+ HPS SP7)
- GET (p690+ / HPS)
- GET (p690 / Colony)
- GET (SGI Altix 3700/1.3)
- GET (P4/Xeon 2800 + Myrinet)
- GET (SGI O3800/R14k-500)
- GET (HP Alphaserver ES45/1000)
II. PUT

Performance singularities for initial access are well known. Driver has to establish mapping of the newly referenced memory and takes a trap. Subsequent references to that memory are fast.
III. ACCUMULATE

Remote 2-D Array Section (ACCUMULATE)
High-End Computational Chemistry
The NWChem Software

- **Capabilities (Direct, Semi-direct and conventional):**
  - RHF, UHF, ROHF using up to 10,000 basis functions; analytic 1st and 2nd derivatives.
  - DFT with a wide variety of local and non-local XC potentials, using up to 10,000 basis functions; analytic 1st and 2nd derivatives.
  - CASSCF; analytic 1st and numerical 2nd derivatives.
  - Semi-direct and RI-based MP2 calculations for RHF and UHF wave functions using up to 3,000 basis functions; analytic 1st derivatives and numerical 2nd derivatives.
  - Coupled cluster, CCSD and CCSD(T) using up to 3,000 basis functions; numerical 1st and 2nd derivatives of the CC energy.
  - Classical molecular dynamics and free energy simulations with the forces obtainable from a variety of sources.
Case Studies - Zeolite Fragments

\[
\begin{align*}
\text{Si}_8\text{O}_7\text{H}_{18} & \quad 347/832 \\
\text{Si}_8\text{O}_{25}\text{H}_{18} & \quad 617/1444 \\
\text{Si}_{26}\text{O}_{37}\text{H}_{36} & \quad 1199/2818 \\
\text{Si}_{28}\text{O}_{67}\text{H}_{30} & \quad 1687/3928
\end{align*}
\]

- DFT Calculations with Coulomb Fitting
  
  Basis (Godbout et al.)
  
  DZVP - O, Si
  
  DZVP2 - H

  Fitting Basis:
  
  DGAUSS-A1 - O, Si
  
  DGAUSS-A2 - H

- NWChem uses auxiliary fitting basis for coulomb energy, with 3 centre 2 electron integrals held in core.
DFT Coulomb Fit - NWChem 4.6

Performance

\[
\text{Si}_{26}\text{O}_{37}\text{H}_{36} \quad 1199/2818
\]

- CS18 P4/2800 + GbitEther
- CS18 P4/2800 + Myrinet
- CS20 Opteron 248/2.2 + Myrinet
- AlphaServer SC ES45/1000
- SGI Altix 3700/1300
- IBM p690+ HPS (SP9)

Basis (Godbout et al.)
- DZVP - O, Si
- DZVP2 - H

Fitting Basis:
- DGAUSS-A1 - O, Si
- DGAUSS-A2 - H

Use auxiliary fitting basis for coulomb energy, with 3 centre 2 electron integrals held in core (%).

“memory 420 mb”
DFT Coulomb Fit - NWChem 4.6

Performance

Si$_{28}$O$_{67}$H$_{30}$ 1687/3928

Number of CPUs

32 64

3.0 4.0

0.0 2.0 4.0 6.0 8.0 10.0 12.0 14.0

0.0 2.5% 4.0 6.3 6.1 7.4 7.0 7.2 11.0 10.2 11.1 8.4

Basis (Godbout et al.)
- DZVP - O, Si
- DZVP2 - H

Fitting Basis:
- DGAUSS-A1 - O, Si
- DGAUSS-A2 - H

Use auxiliary fitting basis for coulomb energy, with 3 centre 2 electron integrals held in core (%).

“memory 420 mb”
Exploiting Global Memory: NWChem

**Zeolite ZSM-5**

DFT (LDA) - Fitted Coulomb Approach - use memory to avoid re-computation of the 3c 2e-integrals

- DZVP Basis (DZV_A2) and Dgauss A1_DFT Fitting basis:
  - AO basis: 3554
  - CD basis: 12713

- IBM p690+ HPS (SP9)

**Wall time (13 SCF iterations):**

32 CPUs = 5,499 seconds (0%)
128 CPUs = 1,434 seconds (100%)

- 3-centre 2e-integrals = 1.00 X 10^{12}
- Schwarz screening = 5.94 X 10^9
- % 3c 2e-ints. In core = 100%
Exploiting HPC: The PNNL Collaboration

- Long standing collaboration with HPCC group within EMSL
- Tools
  - Global arrays:
    - portable distributed data tool:
      - Physically distributed data
  - NWChem
    - Highly efficient and portable MPP computational chemistry package
    - Distributed Data - Scalable with respect to chemical system size as well as MPP hardware size
- Extensible Architecture
  - Object-oriented design
    - abstraction, data hiding, handles, APIs
  - Parallel programming model
    - non-uniform memory access, global arrays
  - Infrastructure
    - GA, Parallel I/O, RTDB, MA, …
- PeIGS:
  - parallel eigensolver,
  - guaranteed orthogonality of eigenvectors
- Wide range of parallel functionality essential for HPCx

Single, shared data structure
- Used by CCP1 groups (e.g. MOLPRO)
GAMESS-UK

GAMESS-UK is the general purpose ab initio molecular electronic structure program for performing SCF-, MCSCF- and DFT-gradient calculations, together with a variety of techniques for post Hartree Fock calculations.

- Derived from the original GAMESS code, obtained from Michel Dupuis in 1981 (then at the NRCC), and has been extensively modified and enhanced over the past two decades.
- This work has included contributions from numerous authors†, and has been conducted largely at the CCLRC Daresbury Laboratory, under the auspices of the UK's Collaborative Computational Project No. 1 (CCP1). Other major sources that have assisted in the on-going development and support of the program include various academic funding agencies in the Netherlands, and ICI plc.

Additional information on the code may be found from links at: http://www.cfs.dl.ac.uk

GAMESS-UK features 1.

- **Hartree Fock:**
  - Segmented/ GC + spherical harmonic basis sets
  - SCF-Energies and Gradients: conventional, in-core, direct
  - SCF-Frequencies: numerical and analytic 2nd derivatives
  - Restricted, unrestricted open shell SCF and GVB.

- **Density Functional Theory**
  - Energies + gradients, conventional and direct including Dunlap fit
  - B3LYP, BLYP, BP86, B97, HCTH, B97-1, FT97 & LDA functionals
  - Analytic 2nd derivatives

- **Electron Correlation:**
  - MP2 energies, gradients and frequencies, Multi-reference MP2, MP3 Energies
  - MCSCF and CASSCF Energies, gradients and numerical 2nd derivatives
  - MR-DCI Energies, properties and transition moments (semi-direct module)
  - CCSD and CCSD(T) Energies
  - RPA (direct) and MCLR excitation energies / oscillator strengths, RPA gradients
  - Full-CI Energies
  - Green’s functions calculations of IPs.
  - Valence bond (Turtle)
GAMESS-UK features 2.

- Molecular Properties:
  - Mulliken and Lowdin population analysis, Electrostatic Potential-Derived Charges
  - Distributed Multipole Analysis, Morokuma Analysis, Multipole Moments
  - Natural Bond Orbital (NBO) + Bader Analysis
  - IR and Raman Intensities, Polarizabilities & Hyperpolarizabilities
  - Solvation and Embedding Effects (DRF)
  - Relativistic Effects (ZORA)

- Pseudopotentials:
  - Local and non-local ECPs.

- Visualisation: tools include CCP1 GUI

- Hybrid QM/MM (ChemShell + CHARMM QM/MM)

- Semi-empirical: MNDO, AM1, and PM3 hamiltonians

- Parallel Capabilities:
  - MPP and SMP implementations (GA tools)
  - SCF/DFT energies, gradients, frequencies
  - MP2 energies and gradients
  - Direct RPA and Direct-CI
Parallel Implementation of GAMESS-UK

- Early implementation based on message passing
- Subsequent activities under HEC Facilities Agreement with support from European projects
  - IMMP (1994-1997, part of EUROPORT)
    - Partners: Guest, Sherwood (Daresbury) - GAMESS-UK, Baerends (Amsterdam) - ADF, Clark (Erlangen) - VAMP
    - Focus on MPP systems (e.g. T3E)
    - Mapping of disk files into global memory (uses GAs)
    - First MPP MP2 algorithm
      - GA storage of transformed integrals
  - QUASI (1998-2001)
    - Application of QM/MM methods in Industry
    - Led by Daresbury, Partners: Catlow (RI), Thiel (MPI), BASF, ICI, Hydro
    - Focus on commodity systems, cost-effective computing in industry
      - demonstrated using Linux alpha commodity cluster at Daresbury.
Parallel Implementation of GAMESS-UK

- Current approach between extremes of replicated and distributed data.
- Most data is replicated - when a parallel linear algebra OP is to be performed the data is copied into a global array, GA tools used to perform the operation, and the data copied back into a replicated object.
  - SCF and DFT
    - GA Tools for caching of I/O for restart and checkpoint files
    - Storage of 3-centre 2-e integrals in DFT Jfit
    - Linear Algebra (via Parallel Linear Algebra from NWChem Project – PeIGs) - DIIS/MMOs, Inversion of Fitting matrix)
  - MP2 gradients, SCF and DFT second derivatives
    - Distribution of <v_vo> and <v_vo> integrals via GAs
  - Direct RPA Excited States
    - Replicated data with parallelisation of direct integral evaluation
- This solves some of the problems associated with a RD strategy, but large memory requirements and the re-replication overheads remain.
- Efficiency dependent on optimised GA implementation.
Parallel Implementations of GAMESS-UK

- Extensive use of GA Tools and Parallel Linear Algebra from NWChem Project (EMSL)
- SCF and DFT energies and gradients
  - Replicated data, but …
    - GA Tools for caching of I/O for restart and checkpoint files
    - Storage of 3-centre 2-e integrals in DFT Jfit
    - Linear Algebra (via PeIGs, DIIS/MMOs, Inversion of 2c-2e matrix)
- SCF and DFT second derivatives
  - Distribution of \( <vvoo> \) and \( <vovo> \) integrals via GAs
- MP2 gradients
  - Distribution of \( <vvoo> \) and \( <vovo> \) integrals via GAs
- Direct-CI
  - Storing the most current \( C \)- and \( Z \)-vectors via GAs
- Direct RPA Excited States
  - Replicated data with parallelisation of direct integral evaluation
Parallel Linux Implementations of GAMESS-UK

- Issues arising from the use of Global Array (GA) Tools
- Efficiency dependent on optimised GA implementation
  - e.g. SCALI, GM vs. PM Myrinet port (CS7, CS12 and CS14)
- SHMMAX
  - The Linux kernel has traditionally fairly small limit for the shared memory segment size (SHMMAX). In kernels 2.2.x it is 32MB on Intel
  - There are two ways to increase this limit:
    - rebuild the kernel after changing SHMMAX in /usr/src/linux/include/asm-i386/shmparm.h, for example, setting SHMMAX as 0x8000000 (128MB)
    - a system admin can increase the limit without rebuilding the kernel, e.g.
      - echo "134217728" >/proc/sys/kernel/shmmmax
- More info on this subject:
  - and on the Linux SMP discussion list.

On “grendel” - /proc/sys/kernel/shmmmax : 33554432
Parallel Linear Algebra

Symmetric Eigensolver Routines

- ScaLAPACK
  - drivers for solving standard and generalized dense symmetric or dense Hermitian Eigenproblems.
  - PDSYEV (QR Method) (Scalapack 1.5)
  - PDSYEVX (Bisection & Inverse Iteration) (Scalapack 1.5)
  - PDSYEVD (D&C Method) (Scalapack 1.7)

- PeIGS (PNNL)
  - General symmetric and standard symmetric eigenproblems
    - PDSPEV (Bisection & Inverse Iteration)

- BFG (I. Bush)
  - Block Jacobi Method on full dense symmetric matrix (+ Hermitian)

- Plapack
  - QR method
  - MRRR ‘Multiple Relatively Robust Representations’
PeIGS 3.0 Parallel Performance

(Solution of real symmetric generalized and standard eigensystem problems)

Features (not always available elsewhere):
- Inverse iteration using Dhillon-Fann-Parlett’s parallel algorithm (fastest uniprocessor performance and good parallel scaling)
- Guaranteed orthonormal eigenvectors in the presence of large clusters of degenerate eigenvalues
- Packed Storage
- Smaller scratch space requirements
- PDSYEV (Scalapack 1.5) and PDSYEVD (Scalapack 1.7)

Consider performance on Fock matrices generated in typical Hartree Fock SCF calculations (1152 and 7194 basis functions).

Developed as part of the NWChem project at PNNL; R.J. Harrison, J. Nieplocha et al.
Eigensolver Performance - “Small” case
(IBM p690+)
Fock Matrix, N = 1152

- Peigs 2.1 (PDSPEV)
- Scalapack (PDSYEV)
- Scalapack (PDSYEV)
- BFG

Number of Processors vs. Time (secs)
PDSYEV - IBM p690+ and SGI Altix 3700

Fully symmetric Fock matrices from a CRYSTAL simulation, representing lithium fluoride with an F centre used to measure the performance of PDSYEV, the ScaLAPACK Parallel Divide and Conquer routine.

Comparing the p690 and SGI Altix, the Atix scales better to 128 CPUs but its performance falls off giving a relative advantage to the IBM p690 at 192 and 225 CPUs.

Here the p690+ has a notable performance lead.

Performance is shown for a matrix of dimension 7194.
Eigensolver Performance - “Largest” Case

(IBM p690+)

Fock Matrix, N = 20480

Total Time (seconds)

Number of Processors

MRRR
PDSYEV

0 350 700 1050 1400

32 64 128 256 512
GAMESS-UK ΔSCF Performance
IBM SP/p690, High-end and Commodity-based Systems

Performance

- CS17 P4/2667 + Gbit Ether
- CS18 P4/2800 + Gbit Ether
- CS18 P4/2800 + Myrinet
- CS20 Opteron248/2.2 + Myrinet
- Compaq AlphaServer SC/1000
- IBM p690
- SGI Altix 3700/Itanium2 1300
- SGI Altix 3700/Itanium2 1500
- IBM p690+ // HPS (SP9)

Impact of Serial Linear Algebra:
\[ T_{CS20-Opt248}(16) = 1306 \text{ [231]} \]
\[ T_{CS20-Opt248}(32) = 1270 \text{ [160]} \]

Cyclosporin:(3-21G Basis, 1000 GTOS)

PDSYEVD (D&C Method)
(Scalapack (1.7))
GAMESS-UK. DFT B3LYP Performance
IBM SP/p690+, High-end and Commodity-based Systems

Performance

Basis: 6-31G

Cyclosporin, 1000 GTOs
GAMESS-UK: DFT HCTH on Valinomycin. Impact of Coulomb Fitting

$T_{32} \text{ SGI Altix 3700/1500} = 237 \text{ secs}$.

$T_{32} \text{ SGI Altix 3700/1500} = 780 \text{ secs}$.

Basis: DZV_A2 (Dgauss)
A1_DFT Fit: 882/3012

Distributed Computing, DisCo
October 2004
GAMES-UK - GA-based Implementation

DFT HCTH Performance: IBM p690+ and High-end Systems

- SGI Origin 3800/R14k-500
- IBM p690
- AlphaServer ES45/1000
- SGI Altix 3700 Itanium2/1300
- IBM p690+ // HPS (SP9)

Valinomycin, 1620 GTOs (DZVP2_A2 (Dgauss))

\[ S_{p690+//HPS\ (SP9)} \times 256 = 152 \]
GAMESS-UK - GA-based Implementation

**Performance**

<table>
<thead>
<tr>
<th>System</th>
<th>Number of CPUs</th>
<th>Valinomycin, DFT-HCTH (1620 GTOs - DZVP2_A2, Dgauss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS18 P4/2800 + GbitE</td>
<td>32</td>
<td>2.5</td>
</tr>
<tr>
<td>CS20 Opteron 248/2.2 + Myrinet</td>
<td>32</td>
<td>2.4</td>
</tr>
<tr>
<td>SGI Origin 3800/R14k-500</td>
<td>32</td>
<td>2.9</td>
</tr>
<tr>
<td>IBM p690</td>
<td>64</td>
<td>6.1</td>
</tr>
<tr>
<td>AlphaServer ES45/1000</td>
<td>64</td>
<td>4.6</td>
</tr>
<tr>
<td>SGI Altix 3700 Itanium2/1300</td>
<td>64</td>
<td>5.3</td>
</tr>
<tr>
<td>IBM p690+ // HPS (SP9)</td>
<td>64</td>
<td>104%, 141%</td>
</tr>
</tbody>
</table>

**Commodity-based and High-end Systems**
DFT BLYP Gradient:
High-end and Commodity-based Systems

Performance

- CS17 P4/2667 + Gbit Ether
- CS18 P4/2800 + GbitEther
- CS20 Opteron248/2.2 + Myrinet
- SGI Origin3800/R14k-500
- IBM p690
- AlphaServer SC ES45/1000
- IBM p690+ // HPS (SP9)
- SGI Altix 3700/Itanium2 1300
- SGI Altix 3700/Itanium2 1500

Geometry optimisation of polymerisation catalyst
Cl(C₃H₅O).Pd[(P(CMe₃)₂)₂.C₆H₄]

Basis 3-21G* (446 GTOs): 10 energy + gradient evaluations

Distributed Computing, DisCo 79 October 2004
SCF and DFT Analytic 2nd Derivatives

- **SCF**
  - Solve DFT equations

- **Derivative Fock**
  - Compute 1st derivatives of DFT matrices

- **CPHF**
  - Solve systems of linear equations for 1st derivatives of DFT wavefunction

- **2nd Derivative**
  - Compute 2nd derivatives of DFT functionals

- **Hessian**
  - Sum partial derivatives

- Distribution of $<v\nu o>$ and $<v\nu o>$ integrals via GAs
- All DFT contributions calculated on the fly
- Extensive use of screening
- Evaluation in AO or MO basis available
- CPHF arrays replicated but all systems of equations can be solved separately to save memory
- DFT terms in CPHF equations can be calculated at lower accuracy
SCF Analytic 2nd Derivatives Performance
IBM SP/p690+, High-end and Commodity-based Systems

\((C_6H_4(CF_3))_2\): Basis 6-31G (196 GTO)

Terms from MO 2e-integrals in GA storage (CPHF & pert. Fock matrices); Calculation dominated by CPHF
DFT Analytic 2nd Derivatives Performance
Commodity-based Systems - HCTH functional

\((C_6H_4(CF_3))_2\): Basis 6-31G (196 GTO)

Terms from MO 2e-integrals in GA storage (CPHF & pert. Fock matrices);
Calculation dominated by CPHF:
MP2 Gradient Algorithms

Serial

- Conventional
  - integrals written to disk
  - read back, transformed, written out, resorted etc.
  - heavy I/O demands

- Direct/Semi-direct (Frisch, Head-Gordon & Pople, Hasse and Ahlrichs)
  - replace all/some I/O with batched integral recomputation

Parallel

- Poor I/O-to-compute performance of MPPs
  - direct approach

- Current MPPs have large global memories

- Store subset of MO integrals
  - reduce number of integral recomputations
  - increase communication overhead

- Subset includes VOVO, VVOO, VOOO,
  - VVVO-class too large to store
  - compute VVVO-terms in separate step
Performance of MP2 Gradient Module
High-end and Commodity-based Systems

Performance

Mn(CO)$_5$H - MP2 geometry optimisation
BASIS: TZVP + f (217 GTOs)

Latency sensitive - optimised GAs
Distributed Data Implementation of GAMESS-UK

Zeolite Y cluster

DZVP Basis (DZV_A2): 3975 GTOs
Hartree Fock (IBM p690+)

Elapsed Time (seconds)

<table>
<thead>
<tr>
<th>Number of CPUs</th>
<th>GA-based Implementation</th>
<th>GA-based Implementation (SP9)</th>
<th>MPI/ScaLAPACK</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>8000</td>
<td>6300</td>
<td>7000</td>
</tr>
<tr>
<td>128</td>
<td>4200</td>
<td>3100</td>
<td>3400</td>
</tr>
<tr>
<td>256</td>
<td>2100</td>
<td>1500</td>
<td>1700</td>
</tr>
<tr>
<td>512</td>
<td>1000</td>
<td>700</td>
<td>800</td>
</tr>
<tr>
<td>1024</td>
<td>500</td>
<td>350</td>
<td>400</td>
</tr>
</tbody>
</table>

Speed-up

<table>
<thead>
<tr>
<th>Number of CPUs</th>
<th>GA-based Implementation</th>
<th>GA-based Implementation (SP9)</th>
<th>MPI/ScaLAPACK</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>1200</td>
<td>900</td>
<td>1050</td>
</tr>
<tr>
<td>128</td>
<td>600</td>
<td>450</td>
<td>525</td>
</tr>
<tr>
<td>256</td>
<td>300</td>
<td>225</td>
<td>262.5</td>
</tr>
<tr>
<td>512</td>
<td>150</td>
<td>112.5</td>
<td>131.25</td>
</tr>
<tr>
<td>1024</td>
<td>75</td>
<td>56.25</td>
<td>65.625</td>
</tr>
</tbody>
</table>
Coupled QM/MM Calculations

- To extend application of *ab-initio* techniques to complex systems
  - Treat reacting core by Quantum Mechanics
    - High accuracy
    - High cost
  - Model environment by Molecular Mechanics
    - Parameterised force field
    - Classical electrostatics
  - QM-MM Junction
    - Link atoms (e.g. H)
    - mechanical coupling
    - polarisation of QM region
GAMESS-UK Interface with CHARMM

- Implemented in collaboration with Bernie Brooks, Eric Billings, (NIH, Bethesda Maryland)
- Functionality:
  - Similar to existing ab-initio interfaces; CHARMM side follows coupling to GAMESS(US) (Milan Hodoscek)
  - Support for Gaussian delocalised point charges implemented in GAMESS-UK, based on 2- and 3- centre integral and derivative integral drivers from the DFT module.
- Availability:
  - CHARMM-capable code incorporated into GAMESS-UK Version 6.2.
  - CHARMM (implemented in c26b2 onwards) requires independent licencing from Martin Karplus.
- Ported to a wide variety of systems including MPPs
  - Origin (Green), Alphaserver SC (PSC), Beowulfs ….
QM/MM Applications


Triosephosphate isomerase (TIM)

- Central reaction in glycolysis, catalytic interconversion of DHAP to GAP
- Demonstration case within QUASI (Partners UZH, and BASF)

QM region 35 atoms (DFT BLYP)
- include residues with possible proton donor/acceptor roles
- GAMESS-UK, MNDO, TURBOMOLE

MM region (4,180 atoms + 2 link)
- CHARMM force-field, implemented in CHARMM, DL_POLY

Measured Time (seconds)

<table>
<thead>
<tr>
<th></th>
<th>CS9 P4/2000 + Myrinet 2k</th>
<th>SGI Origin3800/R14k-500</th>
<th>AlphaServer SC ES45/1000</th>
<th>IBM SP/p690</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>1030</td>
<td>540</td>
<td>308</td>
<td>196</td>
</tr>
<tr>
<td>16</td>
<td>714</td>
<td>419</td>
<td>274</td>
<td>213</td>
</tr>
<tr>
<td>32</td>
<td>687</td>
<td>376</td>
<td>222</td>
<td>213</td>
</tr>
<tr>
<td>64</td>
<td>1487</td>
<td>778</td>
<td>428</td>
<td>231</td>
</tr>
</tbody>
</table>

\[ T_{128} \text{ (IBM SP/p690) = 134 secs} \]
Materials Simulation Codes

Plane Wave DFT Codes:
• CASTEP
• VASP
• CPMD

These codes have similar functionality, power and problems. CASTEP is the flagship code of UKCP.

Local Gaussian Basis Set Codes:
• CRYSTAL

This code presents a different set of problems when considering performance (matrix diagonalisation).

SIESTA and CONQUEST:

• O(n) scaling codes which will be extremely attractive to users.
• Both are currently development rather than production codes.
Materials Simulation. Plane Wave Methods: CASTEP, CPMD

Direct minimisation of the total energy (avoiding diagonalisation)

\[
\psi_j(\vec{r}) = \sum_{\vec{G}} C_{j,\vec{G}} e^{-i(\vec{k} + \vec{G}).\vec{r}} < E_{cut}^{\vec{k} + \vec{G}}\]

- Pseudopotentials must be used to keep the number of plane waves manageable
- Large number of basis functions $N \sim 10^6$ (especially for heavy atoms).

The plane wave expansion means that the bulk of the computation comprises large 3D Fast Fourier Transforms (FFTs) between real and momentum space.

- These are distributed across the processors in various ways.
- The actual FFT routines are optimized for the cache size of the processor.
CPMD - Car-Parrinello Molecular Dynamics

- Developed at IBM Zurich from the original Car-Parrinello Code in 1993;
- Developments at many other sites during the years (more than 150,000 lines of code); it has many unique features, e.g. path-integral MD, QM/MM interfaces, TD-DFT and LR calculations;
- Since 2001 distributed free to academic institutions (www.cpmd.org); more than 2500 licenses in the first year in more than 50 countries.

Optimization of Molecular Structure

Optimization of $E_{el}$ $\longrightarrow$ Forces on Ions $\longrightarrow$ Structure optimization or Molecular Dynamics

$\Psi_i (r) = \sum_j c_{ij} \Phi_j$

Localized basis set (e.g. gaussian functions)

Extended basis set (Plane Waves)

Direct Minimization (Orthogonalization)

Car-Parrinello

CCP1 Flagship Project
CPMD - Mixed Mode Programming

- Developed at IBM Zurich from the original Car-Parrinello Code in 1993;
- Since 2001 distributed free to academic institutions (www.cpmd.org); more than 2500 licenses in the first year in more than 50 countries.

Performance has been tested using up to 1280 processors of HPCx (IBM p690/colony).

- The maximum performance achieved was above 1 Tflop/s for a system consisting of 1000 atoms. (~20% of peak performance).

- 1000 atoms SiC supercell, 256x256x256 Mesh, Cutoff 60 Ry (2,209,586 plane waves)

- Mixed approach instrumental to obtain these results

- Larger SMPs / better switches will help
CPMD 3.7 - C_{120} Benchmark

**Singlet**

- BLYP functional, Kleinman pseudopotential, Wavefunction cutoff 35 Ry
- no. of plane waves ~360K,
- Real space mesh 154X96X96

### Performance

**Singlet**

- BLYP functional, Kleinman pseudopotential, Wavefunction cutoff 35 Ry
- no. of plane waves ~360K,
- Real space mesh 154X96X96
CPMD 3.7 - C_{120} Benchmark

Performance

- CS12 P4/2400 + GbitEther
- CS12 P4/2400 + GbitE (1CPU/node)
- CS10 P4/2666 + Myrinet
- CS10 P4/2666 + Myrinet (1CPU/node)
- CS19 Opteron246/2.0 + IB
- CS19 Opteron246/2.0 + SCI
- CS20 Opteron248/2.2 + Myrinet
- AlphaServer SC ES45 / 1000
- SGI Altix 3700/Itanium2 1300
- SGI Altix 3700/Itanium2 1500

Blended LYP functional, Kleinman pseudopotential, Wavefunction cutoff 35 Ry
no. of plane waves ~360K,
Real space mesh 154X96X96

Number of CPUs

16 32
Benchmark Example:

$\text{Si}_{512}$

Electronic Structure:
- LDA, Kleinman pseudopotential,
- Wavefunction cutoff 20 Ry
- no. of plane waves ~320K,
- Real space mesh 108X108X108

CPMD 3.7 - Si$_{512}$ Benchmark

Performance

- CS10 P4/2666 + Myrinet
- CS18 P4/2800 + Myrinet
- CS19 Opteron246/2.0 + SCI
- CS19 Opteron246/2.0 + IB
- CS20 Opteron248/2.2 + Myrinet
- AlphaServer SC ES45 / 1000
- IBM p690
- IBM p690+ // HPS
- SGI Altix 3700/Itanium2 1300
- SGI Altix 3700/Itanium2 1500

Benchmark Example:

Electronic Structure:
- LDA, Kleinman pseudopotential,
- Wavefunction cutoff 20 Ry
- no. of plane waves ~320K,
- Real space mesh 108X108X108

Performance

- CS10 P4/2666 + Myrinet
- CS18 P4/2800 + Myrinet
- CS19 Opteron246/2.0 + SCI
- CS19 Opteron246/2.0 + IB
- CS20 Opteron248/2.2 + Myrinet
- AlphaServer SC ES45 / 1000
- IBM p690
- IBM p690+ // HPS
- SGI Altix 3700/Itanium2 1300
- SGI Altix 3700/Itanium2 1500

Benchmark Example:

Electronic Structure:
- LDA, Kleinman pseudopotential,
- Wavefunction cutoff 20 Ry
- no. of plane waves ~320K,
- Real space mesh 108X108X108
Computational Engineering

Direct numerical simulation (DNS) Codes:
ANGUS and SBLI
**ANGUS: Combustion modelling (regular grid)**

High-end and **Commodity-based Systems**

<table>
<thead>
<tr>
<th>Performance</th>
<th>Conjugate Gradient + ILU</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 Iterations</td>
<td>Grid Size - $144^3$</td>
</tr>
</tbody>
</table>

- **CS12 P4/2400 + GbitEther**
- **CS9 P4/2000 Xeon+ Myrinet 2k**
- **CS10 P4/2666 Xeon+ Myrinet 2k**
- **CS18 P4/2800 + GbitE**
- **CS18 P4/2800 + Myrinet**
- **CS19 Opteron246/2.0 + IB**
- **CS19 Opteron246/2.0 + SCI**
- **CS20 Opteron248/2.2 + Myrinet**
- **SGI Altix 3700/Itanium2 1300**
- **SGI Altix 3700/Itanium2 1500**
- **IBM p690+ 1.7 GHz // HPS**

Direct numerical simulations (DNS) of turbulent pre-mixed combustion solving the augmented Navier-Stokes equations for fluid flow.

Discretisation of equations is performed using standard 2nd order central differences on a 3D-grid.

Pressure solver utilises either a conjugate gradient method with modified incomplete LU preconditioner or a multi-grid solver (both make extensive use of Level 1 BLAS) or fast Fourier transform.
ANGUS: Combustion modelling (regular grid)

High-end and Commodity-based Systems

Performance

Conjugate Gradient + ILU

100 Iterations

Grid Size - $144^3$

Number of CPUs

Conjugate Gradient + ILU

- 32 CPUs
- 64 CPUs

Grid Size

- 144^3

Performance

- CS10 P4/2666 Xeon+ Myrinet 2k
- CS19 Opteron246/2.0 + IB
- CS19 Opteron246/2.0 + SCI
- CS20 Opteron248/2.2 + Myrinet
- SGI Origin 3800/R14k-500
- IBM p690 1.3GHz // colony
- IBM p690+ 1.7 GHz // HPS
- AlphaServer SC ES45/1000
- SGI Altix 3700/Itanium2 1300
- SGI Altix 3700/Itanium2 1500

- Conjugate Gradient + ILU
- 100 Iterations
- Grid Size - $144^3$
ANGUS: Combustion modelling (regular grid)
Memory Bandwidth Effects: Pentium Xeon and Opteron Systems

Conjugate Gradient + ILU

<table>
<thead>
<tr>
<th>Grid Size - $144^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS18 P4 Xeon/2800 + Myrinet 2k</td>
</tr>
<tr>
<td>CS19 Opteron246/2.0 + SCI</td>
</tr>
</tbody>
</table>

Number of CPUs

<table>
<thead>
<tr>
<th>Number of CPUs</th>
<th>CS18 P4 Xeon/2800 + Myrinet 2k</th>
<th>CS19 Opteron246/2.0 + SCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td><img src="chart-CS18-32.png" alt="Chart" /></td>
<td><img src="chart-CS19-32.png" alt="Chart" /></td>
</tr>
<tr>
<td>16</td>
<td><img src="chart-CS18-16.png" alt="Chart" /></td>
<td><img src="chart-CS19-16.png" alt="Chart" /></td>
</tr>
<tr>
<td>8</td>
<td><img src="chart-CS18-8.png" alt="Chart" /></td>
<td><img src="chart-CS19-8.png" alt="Chart" /></td>
</tr>
</tbody>
</table>

Measured Time (seconds)

<table>
<thead>
<tr>
<th>Measured Time (seconds)</th>
<th>CS18 P4 Xeon/2800 + Myrinet 2k</th>
<th>CS19 Opteron246/2.0 + SCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>596</td>
<td><img src="chart-CS18-596.png" alt="Chart" /></td>
<td><img src="chart-CS19-596.png" alt="Chart" /></td>
</tr>
<tr>
<td>499</td>
<td><img src="chart-CS18-499.png" alt="Chart" /></td>
<td><img src="chart-CS19-499.png" alt="Chart" /></td>
</tr>
<tr>
<td>454</td>
<td><img src="chart-CS18-454.png" alt="Chart" /></td>
<td><img src="chart-CS19-454.png" alt="Chart" /></td>
</tr>
<tr>
<td>353</td>
<td><img src="chart-CS18-353.png" alt="Chart" /></td>
<td><img src="chart-CS19-353.png" alt="Chart" /></td>
</tr>
<tr>
<td>236</td>
<td><img src="chart-CS18-236.png" alt="Chart" /></td>
<td><img src="chart-CS19-236.png" alt="Chart" /></td>
</tr>
<tr>
<td>16</td>
<td><img src="chart-CS18-16.png" alt="Chart" /></td>
<td><img src="chart-CS19-16.png" alt="Chart" /></td>
</tr>
<tr>
<td>8</td>
<td><img src="chart-CS18-8.png" alt="Chart" /></td>
<td><img src="chart-CS19-8.png" alt="Chart" /></td>
</tr>
</tbody>
</table>

Distributed Computing, DisCo October 2004
ANGUS: Combustion modelling (regular grid)
High-end and Commodity-based Systems

Performance

Direct numerical simulations (DNS) of turbulent pre-mixed combustion solving the augmented Navier-Stokes equations for fluid flow.
Computational Engineering: UK Turbulence Consortium

- Focus on compute-intensive methods (Direct Numerical Simulation, Large Eddy Simulation, etc) for the simulation of turbulent flows.
- Shock boundary layer interaction modelling is critical for accurate aerodynamic design but is still poorly understood.
- The following results are from a compressible turbulent channel flow benchmark using Southampton’s DNS code.
  - 3D compressible Navier-Stokes equations.
  - General grid transformation for complex geometries.
  - High-order (4th or 6th) central difference scheme.
  - Reduced-order and stable boundary treatment.
  - Entropy splitting for Euler terms.
  - Shock capturing TVD scheme with artificial compression method.

http://www.afm.ses.soton.ac.uk/
Direct Numerical Simulation: $360^3$ benchmark

High-End Systems

Performance (Mgrid-points*iterations/sec) vs. Number of processors

- **SBLI code – PDNS3D**
- **UK Turbulence Consortium**

**High-End Systems**

- SGI Altix 3700/1500
- SGI Altix 3700/1300
- IBM p690+ xlf8115
- IBM p690 SP
- Cray T3E-1200E

**T3 (360 X 360 X 360)**
DNS: $120^3$ benchmark.

High-end and Commodity-based Systems

Performance

- **CS12 P4/2400 Xeon + GbitE**
- **CS12 P4/2400 Xeon + GbitE (1CPU/node)**
- **CS18 P4/2800 + Myrinet**
- **CS18 P4/2800 + Myrinet (1CPU/node)**
- **CS19 Opteron246/2.0 + SCI**
- **CS19 Opteron246/2.0 + SCI (1CPU/node)**
- **CS20 Opteron248/2.2 + Myrinet**
- **AlphaServer SC ES45/1000**
- **AlphaServer SC ES45/1000 (2CPUs/node)**
- **IBM p690 1.3GHz**
- **IBM p690 1.3GHz (4 CPUs/node)**
- **SGI Altix 3700/Itanium2 1300**

---

**Dramatic Impact of Memory Bandwidth on Xeon Systems**

36%, 82%
DNS: $120^3$ benchmark. High-end and Commodity-based Systems

Dramatic Impact of Memory Bandwidth on Xeon Systems

Number of processors

<table>
<thead>
<tr>
<th>Number of processors</th>
<th>CS10 P4/2666 Xeon + Myrinet</th>
<th>CS12 P4/2400 Xeon + GbitE</th>
<th>CS18 P4/2800 + Myrinet</th>
<th>CS19 Opteron246/2.0 + SCI</th>
<th>CS19 Opteron246/2.0 + IB</th>
<th>CS20 Opteron248/2.2 + Myrinet</th>
<th>AlphaServer SC ES45/1000</th>
<th>IBM p690 1.3GHz</th>
<th>SGI Altix 3700/Itanium2 1300</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>1.2</td>
<td>2.7</td>
<td>3.1</td>
<td>5.3</td>
<td>5.6</td>
<td>3.4</td>
<td>2.3</td>
<td>4.1</td>
<td>8.5</td>
</tr>
<tr>
<td>32</td>
<td>2.7</td>
<td>3.4</td>
<td>3.1</td>
<td>5.3</td>
<td>5.6</td>
<td>3.4</td>
<td>2.3</td>
<td>4.1</td>
<td>8.5</td>
</tr>
<tr>
<td>64</td>
<td>3.1</td>
<td>3.4</td>
<td>3.1</td>
<td>5.3</td>
<td>5.6</td>
<td>3.4</td>
<td>2.3</td>
<td>4.1</td>
<td>8.5</td>
</tr>
</tbody>
</table>

Performance

- CS10 P4/2666 Xeon + Myrinet
- CS12 P4/2400 Xeon + GbitE
- CS18 P4/2800 + Myrinet
- CS19 Opteron246/2.0 + SCI
- CS19 Opteron246/2.0 + IB
- CS20 Opteron248/2.2 + Myrinet
- AlphaServer SC ES45/1000
- IBM p690 1.3GHz
- SGI Altix 3700/Itanium2 1300

Dramatic Impact of Memory Bandwidth on Xeon Systems
Commodity Comparisons with High-end Systems

% of 32 CPUs of SGI Altix 3700 / 1300

Cluster CS18
Pentium4/2800 Xeon + Gbit Ethernet

CS18 - 49% of Altix
Commodity Comparisons with High-end Systems

% of 32 CPUs of SGI Altix 3700 / 1300

Cluster CS20
AMD Opteron 248/2.2 + Myrinet

CS20 - 100% of Altix

Distributed Computing, DisCo 106 October 2004
Summary

- Commodity-based and High-end Systems
  - Commodity Systems under evaluation; CS1 – CS20
  - High-end systems from SGI, IBM and HP/Compaq
  - Serial CPU and Communications Performance; Evaluation Metrics

- Application performance
  - Molecular Simulation (replicated & distributed data)
    - DL POLY, DLMULTI and CHARMM
  - Molecular Electronic Structure & Materials Simulation
    - NWChem, GAMESS-UK and CPMD
  - Computational Engineering
    - ANGUS, SBLI

- Notable Performance of Opteron clusters with Myrinet:
  - CS20 Opteron248/2200 cluster with Myrinet delivers on average 100% of SGI Altix 3700/1300.
  - Cost effectiveness of the Clusters (e.g., CS18) with Gbit Ethernet - delivers 49% of the SGI Altix 3700/1300.