An Overview of High Performance Computing and Trends

Jack Dongarra
Computer Science Department
University of Tennessee

Outline for the Next 3 Days

- Overview of High Performance Computing
- Impact of HPC on Linear Algebra Algorithms and Software
- Grid Computing
Innovative Computing Laboratory

- Numerical Linear Algebra
- Heterogeneous Distributed Computing
- Software Repositories
- Performance Evaluation

Software and ideas have found their way into many areas of Computational Science.

Around 40 people: At the moment...

16 Researchers: Research Assoc/Post-Doc/Research Prof
15 Students: Graduate and Undergraduate
8 Support staff: Secretary, Systems, Artist
1 Long term visitors (Japan)

Responsible for about $4M/years in research funding from NSF, DOE, DOD, etc.
Computational Science

- High Performance Computing offers a new way to do science:
  - Experiment - Theory - Computation
- Computation used to approximate physical systems - Advantages include:
  - Playing with simulation parameters to study emergent trends
  - Possible replay of a particular simulation event
  - Study systems where no exact theories exist

Why Turn to Simulation?

- When the problem is too:
  - Complex
  - Large / small
  - Expensive
  - Dangerous
- ... to do any other way.
- Climate / Weather Modeling
- Data intensive problems (data-mining, oil reservoir simulation)
- Problems with large length and time scales (cosmology)
Technology Trends:
Microprocessor Capacity

2X transistors/Chip Every 1.5 years
Called “Moore’s Law”

Microprocessors have become smaller, denser, and more powerful.
Not just processors, bandwidth, storage, etc

Gordon Moore (co-founder of Intel) predicted in 1965 that the transistor density of semiconductor chips would double roughly every 18 months.
H. Meuer, H. Simon, E. Strohmaier, & JD

- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP
  \[ Ax=b, \text{ dense problem} \]
- Updated twice a year
  SC‘xy in the States in November
  Meeting in Mannheim, Germany in June
- All data available from www.top500.org

---

**Fastest Computer Over Time**

<table>
<thead>
<tr>
<th>Year</th>
<th>GFlop/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>0</td>
</tr>
<tr>
<td>1992</td>
<td>10</td>
</tr>
<tr>
<td>1994</td>
<td>20</td>
</tr>
<tr>
<td>1996</td>
<td>30</td>
</tr>
<tr>
<td>1998</td>
<td>40</td>
</tr>
<tr>
<td>2000</td>
<td>50</td>
</tr>
</tbody>
</table>

In 1980 a computation that took 1 full year to complete can now be done in ~ 10 hours!
In 1980 a computation that took 1 full year to complete can now be done in ~ 16 minutes!

In 1980 a computation that took 1 full year to complete can today be done in ~ 27 seconds!
Livermore National Laboratory – IBM Blue Pacific and White SMP Superclusters

4TF Blue Pacific SST running
- 3 x 480 4-way SMP nodes
- 3.9 TF peak performance
- 2.6 TB memory
- 2.5 Tb/s bisectional bandwidth
- 62 TB disk
- 6.4 GB/s delivered I/O bandwidth

10TF ASCI White
- 512 Nighthawk 16-way SMP nodes
- 12. TF peak performance
- 4.0 TB memory
- 159 TB disk
- 2x I/O size and delivered bw over SST
- 2.5x external network improvement
- Sufficient swap for GANG scheduling

Fastest Computer Over Time

Japanese Earth Simulator
NEC 5104

In 1980 a computation that took 1 full year to complete can today be done in ~ 5.4 seconds!
TOP500 list - Data shown

- Manufacturer
- Computer Type
- Installation Site
- Location
- Year
- Customer Segment
- # Processors
- R\textsubscript{max}
- R\textsubscript{peak}
- N\textsubscript{max}
- N\textsubscript{1/2}
- N\textsubscript{world}

- Manufacturer or vendor
- indicated by manufacturer or vendor
- Customer
- Location and country
- Year of installation/last major update
- Academic, Research, Industry, Vendor, Class.
- Number of processors
- Maximal LINPACK performance achieved
- Theoretical peak performance
- Problemsize for achieving R\textsubscript{max}
- Problemsize for achieving half of R\textsubscript{max}
- Position within the TOP500 ranking
### TOP10

<table>
<thead>
<tr>
<th>Rank</th>
<th>Manufacturer</th>
<th>Computer</th>
<th>( R_{\text{max}} ) [TF/s]</th>
<th>Installation Site</th>
<th>Country</th>
<th>Year</th>
<th>Area of Installation</th>
<th># Proc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NEC</td>
<td>Earth-Simulator</td>
<td>55.86</td>
<td>Earth Simulator Center</td>
<td>Japan</td>
<td>2002</td>
<td>Research</td>
<td>5120</td>
</tr>
<tr>
<td>2</td>
<td>IBM</td>
<td>ASCI White SP Power3</td>
<td>7.23</td>
<td>Lawrence Livermore National Laboratory</td>
<td>USA</td>
<td>2000</td>
<td>Research</td>
<td>8192</td>
</tr>
<tr>
<td>3</td>
<td>HP</td>
<td>AlphaServer SC ES45 1 GHz</td>
<td>4.46</td>
<td>Pittsburgh Supercomputing Center</td>
<td>USA</td>
<td>2001</td>
<td>Academic</td>
<td>3016</td>
</tr>
<tr>
<td>4</td>
<td>HP</td>
<td>AlphaServer SC ES45 1 GHz</td>
<td>3.98</td>
<td>Commissariat a l’Énergie Atomique (CEA)</td>
<td>France</td>
<td>2001</td>
<td>Research</td>
<td>2560</td>
</tr>
<tr>
<td>5</td>
<td>IBM</td>
<td>SP Power3 375 MHz</td>
<td>3.05</td>
<td>NERSC/LBNL</td>
<td>USA</td>
<td>2001</td>
<td>Research</td>
<td>3328</td>
</tr>
<tr>
<td>6</td>
<td>HP</td>
<td>AlphaServer SC ES45 1 GHz</td>
<td>2.92</td>
<td>Los Alamos National Laboratory</td>
<td>USA</td>
<td>2002</td>
<td>Research</td>
<td>2048</td>
</tr>
<tr>
<td>7</td>
<td>Intel</td>
<td>ASCI Red</td>
<td>2.38</td>
<td>Sandia National Laboratory</td>
<td>USA</td>
<td>1999</td>
<td>Research</td>
<td>9632</td>
</tr>
<tr>
<td>8</td>
<td>IBM</td>
<td>gSeries 690 1.3 GHz</td>
<td>2.31</td>
<td>Oak Ridge National Laboratory</td>
<td>USA</td>
<td>2002</td>
<td>Research</td>
<td>864</td>
</tr>
<tr>
<td>9</td>
<td>IBM</td>
<td>ASCI Blue Pacific SST, IBM SP 604e</td>
<td>2.14</td>
<td>Lawrence Livermore National Laboratory</td>
<td>USA</td>
<td>1999</td>
<td>Research</td>
<td>5808</td>
</tr>
<tr>
<td>10</td>
<td>IBM</td>
<td>gSeries 690 1.3 GHz</td>
<td>2.00</td>
<td>IBM/US Army Research Lab (ARL)</td>
<td>USA</td>
<td>2002</td>
<td>Vendor</td>
<td>768</td>
</tr>
</tbody>
</table>

### TOP500 - Performance

![TOP500 Performance Chart](chart.png)
“Moore’s Wall”
Horst Simon, NERSC

- Moore’s Law predicts exponential growth
  - Performance doubling every 18 months
  - Usually plotted on semi-log scale, appears as straight line
- Human experience has a hard time deal with log scale
  - In 1980 a computation that took 1 full year to complete can now be done in minutes!
  - We are sitting at the bend of an exponential curve
- From our perspective Moore’s Law appears as a “wall”
  - In a few years technology will again be completely different
  - Hard to predict what the future will be.

Performance Extrapolation
### Sun Systems on the Top500

<table>
<thead>
<tr>
<th>Rank</th>
<th>Manufacturer</th>
<th>Computer</th>
<th>Peak (GFLOPS)</th>
<th>Installation Site</th>
<th>Country</th>
<th>Year</th>
<th>Installation Type</th>
<th>Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Sun Microsystems</td>
<td>Sun Fire 4800</td>
<td>426.44</td>
<td>Service Provider</td>
<td>USA</td>
<td>2003</td>
<td>Industry</td>
<td>895</td>
</tr>
<tr>
<td>2.</td>
<td>Sun Microsystems</td>
<td>Sun Fire 4800</td>
<td>426.44</td>
<td>Service Provider</td>
<td>USA</td>
<td>2003</td>
<td>Industry</td>
<td>895</td>
</tr>
<tr>
<td>3.</td>
<td>Sun Microsystems</td>
<td>Sun Fire 4800</td>
<td>426.44</td>
<td>Service Provider</td>
<td>USA</td>
<td>2003</td>
<td>Industry</td>
<td>895</td>
</tr>
<tr>
<td>4.</td>
<td>Sun Microsystems</td>
<td>Sun Fire 4800</td>
<td>426.44</td>
<td>Service Provider</td>
<td>USA</td>
<td>2003</td>
<td>Industry</td>
<td>895</td>
</tr>
<tr>
<td>5.</td>
<td>Sun Microsystems</td>
<td>Sun Fire 4800</td>
<td>426.44</td>
<td>Service Provider</td>
<td>USA</td>
<td>2003</td>
<td>Industry</td>
<td>895</td>
</tr>
</tbody>
</table>

### French Top500 Computers

<table>
<thead>
<tr>
<th>Rank</th>
<th>Manufacturer</th>
<th>Computer</th>
<th>Peak (GFLOPS)</th>
<th>Installation Site</th>
<th>Country</th>
<th>Year</th>
<th>Installation Type</th>
<th>Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Sun Microsystems</td>
<td>Sun Fire 4800</td>
<td>426.44</td>
<td>Service Provider</td>
<td>USA</td>
<td>2003</td>
<td>Industry</td>
<td>895</td>
</tr>
<tr>
<td>2.</td>
<td>Sun Microsystems</td>
<td>Sun Fire 4800</td>
<td>426.44</td>
<td>Service Provider</td>
<td>USA</td>
<td>2003</td>
<td>Industry</td>
<td>895</td>
</tr>
<tr>
<td>3.</td>
<td>Sun Microsystems</td>
<td>Sun Fire 4800</td>
<td>426.44</td>
<td>Service Provider</td>
<td>USA</td>
<td>2003</td>
<td>Industry</td>
<td>895</td>
</tr>
<tr>
<td>4.</td>
<td>Sun Microsystems</td>
<td>Sun Fire 4800</td>
<td>426.44</td>
<td>Service Provider</td>
<td>USA</td>
<td>2003</td>
<td>Industry</td>
<td>895</td>
</tr>
<tr>
<td>5.</td>
<td>Sun Microsystems</td>
<td>Sun Fire 4800</td>
<td>426.44</td>
<td>Service Provider</td>
<td>USA</td>
<td>2003</td>
<td>Industry</td>
<td>895</td>
</tr>
</tbody>
</table>
### Excerpt from TOP500

<table>
<thead>
<tr>
<th>Rank</th>
<th>Manufacturer</th>
<th>Computer</th>
<th>Rmax [GF/s]</th>
<th>Installation Site</th>
<th>Country</th>
<th>Area</th>
<th># Proc</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>IBM</td>
<td>SP Power3</td>
<td>795</td>
<td>Charles Schwab</td>
<td>USA</td>
<td>Finance</td>
<td>768</td>
</tr>
<tr>
<td>66</td>
<td>IBM</td>
<td>SP Power3</td>
<td>594</td>
<td>Sprint PCS</td>
<td>USA</td>
<td>Telecom</td>
<td>320</td>
</tr>
<tr>
<td>67</td>
<td>IBM</td>
<td>SP Power4</td>
<td>555</td>
<td>EDS General Motors</td>
<td>USA</td>
<td>Automotive</td>
<td>224</td>
</tr>
<tr>
<td>73</td>
<td>IBM</td>
<td>SP Power3</td>
<td>546</td>
<td>State Farm</td>
<td>USA</td>
<td>Database</td>
<td>520</td>
</tr>
<tr>
<td>125</td>
<td>IBM</td>
<td>Netfinity P5</td>
<td>366</td>
<td>WesternGeco</td>
<td>UK</td>
<td>Geophysics</td>
<td>1280</td>
</tr>
<tr>
<td>127</td>
<td>HP</td>
<td>SuperDome HyperPlex</td>
<td>361</td>
<td>Centrica Plc</td>
<td>UK</td>
<td>Energy</td>
<td>196</td>
</tr>
</tbody>
</table>

### Customer Types - Performance

![Customer Types - Performance Graph](image-url)
Producers

Producers - Performance
Processor Type

Chip Technology
Cumulative Performance
June 2002

222 TF/s

58 = Rank of ½ cumulative performance
To Run Benchmark for TOP500

**HPL**: High Performance Linpack
Antoine Petitet and Clint Whaley, ICL, UTK

- [icl.cs.utk.edu/hpl](http://icl.cs.utk.edu/hpl)
- Needs only
  - MPI
  - BLAS or VSIPL
- Highly scalable and efficient for the whole range of system sizes we see

---

1976: The Supercomputing “Island”

**Today:** A Continuum
Petaflop Computers Within the Next Decade

- **Five basis design points:**
  - **Conventional technologies**
    - 4.8 GHz processor, 8000 nodes, each w/16 processors
  - **Processing-in-memory (PIM) designs**
    - Reduce memory access bottleneck
  - **Superconducting processor technologies**
    - Digital superconductor technology, Rapid Single-Flux-Quantum (RSFQ) logic & hybrid technology multi-threaded (HTMT)
  - **Special-purpose hardware designs**
    - Specific applications e.g. GRAPE Project in Japan for gravitational force computations
  - **Schemes utilizing the aggregate computing power of processors distributed on the web**
    - [SETI@home](http://seti@home) ~26 Tflop/s

SETI@home: Global Distributed Computing

- Running on 500,000 PCs, ~1000 CPU Years per Day
  - 485,821 CPU Years so far
- Sophisticated Data & Signal Processing Analysis
- Distributions Datasets from Arecibo Radio Telescope
SETI@home

- Use thousands of Internet-connected PCs to help in the search for extraterrestrial intelligence.
- Uses data collected with the Arecibo Radio Telescope, in Puerto Rico
- When their computer is idle or being wasted this software will download a 300 kilobyte chunk of data for analysis.
- The results of this analysis are sent back to the SETI team, combined with thousands of other participants.

- Largest distributed computation project in existence
  - ~ 400,000 machines
  - Averaging 27 Tflop/s
- Today many companies trying this for profit.

Grid Computing - from ET to Anthrax
Petaflops ($10^{15}$ flop/s) Computer Today?

2 GHz processor ($O(10^9)$ ops/s)
- 1/2 Million PCs $O(10^6)$
- ~$2K each, $O(10^3)$ → $1B$
- 100 Mwatts
- 5 acres
- 500,000 Windows licenses!!
- PC failure every second

---

High-Performance Computing
Directions: Beowulf-class PC Clusters

**Definition:**
- **COTS PC Nodes**
  - Pentium, AMD, Alpha, PowerPC, SMP
- **COTS LAN/SAN Interconnect**
  - Ethernet, Myrinet, Giganet, ATM
- **Open Source Unix**
  - Linux, BSD
- **Message Passing Computing**
  - MPI, PVM

**Advantages:**
- Best price-performance
- Low entry-level cost
- Just-in-place configuration
- Vendor invulnerable
- Scalable
- Rapid technology tracking

*Enabled by* PC hardware, networks and operating system achieving capabilities of scientific workstations at a fraction of the cost and availability of industry standard message passing libraries. However, much more of a contact sport.
Excerpt from TOP500

<table>
<thead>
<tr>
<th>Rank</th>
<th>Manufacturer</th>
<th>Computer</th>
<th>Rmax [GF/s]</th>
<th>Installation Site</th>
<th>Country</th>
<th># Proc</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>Self-made</td>
<td>Cplane/Ross</td>
<td>707</td>
<td>Sandia National Lab</td>
<td>USA</td>
<td>1369</td>
</tr>
<tr>
<td>34</td>
<td>IBM</td>
<td>Titan Cluster</td>
<td>594</td>
<td>NCSA</td>
<td>USA</td>
<td>320</td>
</tr>
<tr>
<td>39</td>
<td>NEC</td>
<td>Magi Cluster</td>
<td>654</td>
<td>CBRC – Tsukuba Advanced Computing Center</td>
<td>Japan</td>
<td>1024</td>
</tr>
<tr>
<td>40</td>
<td>Self-made</td>
<td>SCoreiII</td>
<td>618</td>
<td>Real World Computing, Tsukuba</td>
<td>Japan</td>
<td>1024</td>
</tr>
<tr>
<td>41</td>
<td>IBM</td>
<td>Netfinity Cluster</td>
<td>594</td>
<td>NCSA</td>
<td>USA</td>
<td>1024</td>
</tr>
<tr>
<td>320</td>
<td>Dell</td>
<td>PowerEdge Cluster</td>
<td>121</td>
<td>Cornell Theory Center</td>
<td>USA</td>
<td>252</td>
</tr>
</tbody>
</table>

Performance Numbers on RISC Processors

<table>
<thead>
<tr>
<th>Processor</th>
<th>Cycle Time</th>
<th>Linpack n=100</th>
<th>Linpack n=1000</th>
<th>Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel P4</td>
<td>2540</td>
<td>1190 (23%)</td>
<td>2355 (46%)</td>
<td>5080</td>
</tr>
<tr>
<td>Intel/HP Itanium 2</td>
<td>1000</td>
<td>1102 (27%)</td>
<td>3534 (88%)</td>
<td>4000</td>
</tr>
<tr>
<td>Compaq Alpha</td>
<td>1000</td>
<td>824 (41%)</td>
<td>1542 (77%)</td>
<td>2000</td>
</tr>
<tr>
<td>AMD Athlon</td>
<td>1200</td>
<td>558 (23%)</td>
<td>998 (42%)</td>
<td>2400</td>
</tr>
<tr>
<td>HP PA</td>
<td>550</td>
<td>468 (21%)</td>
<td>1583 (71%)</td>
<td>2200</td>
</tr>
<tr>
<td>IBM Power 3</td>
<td>375</td>
<td>424 (28%)</td>
<td>1208 (80%)</td>
<td>1500</td>
</tr>
<tr>
<td>Intel P3</td>
<td>933</td>
<td>234 (25%)</td>
<td>514 (55%)</td>
<td>933</td>
</tr>
<tr>
<td>PowerPC G4</td>
<td>533</td>
<td>231 (22%)</td>
<td>478 (45%)</td>
<td>1066</td>
</tr>
<tr>
<td>SUN Ultra 80</td>
<td>450</td>
<td>208 (23%)</td>
<td>607 (67%)</td>
<td>900</td>
</tr>
<tr>
<td>SGI Origin 2K</td>
<td>300</td>
<td>173 (29%)</td>
<td>553 (92%)</td>
<td>600</td>
</tr>
<tr>
<td>Cray T90</td>
<td>454</td>
<td>705 (39%)</td>
<td>1603 (89%)</td>
<td>1800</td>
</tr>
<tr>
<td>Cray C90</td>
<td>238</td>
<td>387 (41%)</td>
<td>902 (95%)</td>
<td>952</td>
</tr>
<tr>
<td>Cray Y-MP</td>
<td>166</td>
<td>161 (48%)</td>
<td>324 (97%)</td>
<td>333</td>
</tr>
<tr>
<td>Cray X-MP</td>
<td>118</td>
<td>121 (51%)</td>
<td>218 (93%)</td>
<td>235</td>
</tr>
<tr>
<td>Cray J-90</td>
<td>100</td>
<td>106 (53%)</td>
<td>190 (95%)</td>
<td>200</td>
</tr>
<tr>
<td>Cray 1</td>
<td>80</td>
<td>27 (17%)</td>
<td>110 (69%)</td>
<td>160</td>
</tr>
</tbody>
</table>
Pentium 4 - SSE2
Today’s “Sweet Spot” in Price/Performance

- 2.53 GHz, 400 MHz system bus, 16K L1 &
256K L2 Cache, theoretical peak of 2.53
Gflop/s, high power consumption
- Streaming SIMD Extensions 2 (SSE2)
  - which consists of 144 new instructions
  - includes SIMD IEEE double precision floating point
    - Peak for 64 bit floating point 2X (5.06 Gflop/s)
    - Peak for 32 bit floating point 4X (10.12 Gflop/s)
  - SIMD 128-bit integer
  - new cache and memory management instructions.
  - Intel’s compiler supports these instructions today
  - ATLAS was trained to probe and detect SSE2

---

Table 1: Performance in Solving a System of Linear Equations

<table>
<thead>
<tr>
<th>Computer</th>
<th>“LINPACK Benchmark”</th>
<th>“TFP” Best Effort</th>
<th>“Theoretical Peak” Mflop/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel Pentium 4 (2.53 GHz)</td>
<td>cif -O3 -XW -ipo -ip -align 1190</td>
<td>2535</td>
<td>5060</td>
</tr>
<tr>
<td>NEC SX-6/8 (1proc. 2.0 ms)</td>
<td>41520</td>
<td>64000</td>
<td></td>
</tr>
<tr>
<td>NEC SX-6/4 (4proc. 2.0 ms)</td>
<td>23680</td>
<td>32000</td>
<td></td>
</tr>
<tr>
<td>NEC SX-6/2 (2proc. 2.0 ms)</td>
<td>13350</td>
<td>16000</td>
<td></td>
</tr>
<tr>
<td>NEC SX-6/1 (1proc. 2.0 ms)</td>
<td>R12.1 -pi -Wf -prob.ave 1161</td>
<td>7575</td>
<td>8000</td>
</tr>
<tr>
<td>Fujitsu VPP5000/1 (1proc. 3.33ms)</td>
<td>ft -Wv -r128 -Of -KA32 1156</td>
<td>8784</td>
<td>9000</td>
</tr>
<tr>
<td>Cray T32 (32 proc. 2.2 ms)</td>
<td>29300</td>
<td>57000</td>
<td></td>
</tr>
<tr>
<td>Cray T32 (38 proc. 2.2 ms)</td>
<td>28340</td>
<td>50400</td>
<td></td>
</tr>
<tr>
<td>Cray T94 (24 proc. 2.2 ms)</td>
<td>26710</td>
<td>43200</td>
<td></td>
</tr>
<tr>
<td>Cray T94 (16 proc. 2.2 ms)</td>
<td>19980</td>
<td>28800</td>
<td></td>
</tr>
<tr>
<td>Cray T94 (8 proc. 2.2 ms)</td>
<td>10880</td>
<td>14400</td>
<td></td>
</tr>
<tr>
<td>Cray T94 (4 proc. 2.2 ms)</td>
<td>f90 -O2,inline2 1129</td>
<td>5736</td>
<td>7200</td>
</tr>
<tr>
<td>HP RX3670 Itanium 2/4 proc 1GHz</td>
<td>11420</td>
<td>16000</td>
<td></td>
</tr>
<tr>
<td>HP RX3670 Itanium 2/2 proc 1GHz</td>
<td>6284</td>
<td>12000</td>
<td></td>
</tr>
<tr>
<td>HP RX3670 Itanium 2/1 proc 1GHz</td>
<td>f90 +DSNcheck+O3 +Online,budget=100000 +Oso,ptrsto,globals 1182</td>
<td>3534</td>
<td>4000</td>
</tr>
<tr>
<td>HP RX2600 Itanium 2/2 proc 1GHz</td>
<td>6231</td>
<td>8000</td>
<td></td>
</tr>
<tr>
<td>HP RX2600 Itanium 2/1 proc 1GHz</td>
<td>f90 +DSNcheck+O3 +Online,budget=100000 +Oso,ptrsto,globals 1182</td>
<td>3528</td>
<td>4000</td>
</tr>
</tbody>
</table>
Notes on the Earth Simulator

Jack Dongarra
Computer Science Department
University of Tennessee
Earth Simulator

- Based on the NEC SX architecture, 640 nodes, each node with 8 vector processors (8 Gflop/s peak per processor), 2 ns cycle time, 16GB shared memory.
  - Total of 5104 total processors, 40 TFlop/s peak, and 10 TB memory.
- It has a single stage crossbar (1800 miles of cable) 83,000 copper cables, 16 GB/s cross section bandwidth.
- 700 TB disk space
- 1.6 PB mass store
- Area of computer = 4 tennis courts, 3 floors
Earth Simulator in a Nutshell

Specifications

<table>
<thead>
<tr>
<th></th>
<th>Total number of processors</th>
<th>5,120</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak performance / processor</td>
<td>8 GFlops</td>
<td></td>
</tr>
<tr>
<td>Peak performance / node</td>
<td>64 GFlops</td>
<td></td>
</tr>
<tr>
<td>Shared memory</td>
<td>16 GB</td>
<td></td>
</tr>
</tbody>
</table>

Total number of nodes: 640
Total peak performance: 40 TFlops
Total main memory: 10 TB

Outline of the Earth Simulator Computer

- **Architecture**: A MIMD-type, distributed memory, parallel system consisting of computing nodes in which vector-type multi-processors are tightly connected by sharing main memory.

- **Total number of processor nodes**: 640
- **Number of PE’s for each node**: 8
- **Total number of PE’s**: 5120
- **Peak performance of each PE**: 8 GFLOPS
- **Peak performance of each node**: 64 GFLOPS

- **Main memory**: 10 TB (total).
  - Shared memory / node: 16 GB

- **Interconnection network**: Single-Stage Crossbar Network

- **Performance**: Assuming the efficiency 12.5%,
  the peak performance 40 TFLOPS
  (the effective performance for an atmospheric circulation model is more than 5 TFLOPS).
**R&D results**

### Comparison of vector processors

**SX4**
- 8 Gflops (2 Gflop/s x 4)
- Clock: 125MHz
- LSI: 0.35μm CMOS
- 37 x 4 = 148 LSIs

**SX5**
- 8 Gflop/s
- Clock: 250MHz
- LSI: 0.25μm CMOS
- 32 LSIs

**Earth Simulator**
- 8 Gflop/s
- Clock: 500MHz/1GHz
- LSI: 0.15μm CMOS
- 1 chip processor

---

### Comparison of cabinets for 1 node

**Present distributed-memory supercomputer (SX-4) 1 node**
- Peak Performance: 64 Gflops
- Main Memory: 16GB
- Electric Power: 90KVA

**Earth Simulator 1 node**
- Peak Performance: 64 Gflops
- Main Memory: 16GB
- Electric Power: 8KVA

---

*Earth Simulator Research and Development Center*
**R&D results**

**Connection between processor nodes (crossbar network)**

- **Total number of cables**: 640 x 130 = 83,200
- **Total length of cables**: 2,900 m
- **Total weight of cables**: 220 t

---

**R&D issues on Hardware Technologies**

1. **LSI Technology**
   - Enhancement of clock cycle: 150MHz → 500MHz (partly 1GHz)
   - Development of high density LSI
     - 0.15µm CMOS + Cu interconnection (8 layers)
     - 1.50-2.0 million transistors/cm² → 10 million transistors/cm²
   - Enlargement of chip size: (about 2cm x 2cm)
     - High performance one-chip vector processor: OCVP-ES

2. **Packaging Technology**
   - Build-up PCB (110mm x 115mm)
     - Line width / Spacing: 25µm / 25µm
     - 6 core layers + 4 build-up layers on both surfaces
   - Number of pins/chip: <1000 (present) → 4000 - 5000

3. **Cooling Technology**
   - Air cooling using heat pipe technology (Max. 170W per chip)

4. **Board to Board Interconnection Technology**
   - Interface connector: 0.5mm pitch surface mount
   - Interface cable: 0.6mm diameter coaxial cable and 3.8ns/m delay time

5. **PN-IN Interconnection Technology**
   - 40m transmission distance with fine tuned equalizer circuit
Bird’s-eye View of the Earth Simulator System

- Disks
- Cartridge Tape Library System
- Processor Node (PN) Cabinets
- Interconnection Network (IN) Cabinets
- Air Conditioning System
- Power Supply System
- Double Floor for IN Cables

Cross-sectional View of the Earth Simulator Building

- Lightning protection system
- Air-conditioning return duct
- Double floor for IN cables and air-conditioning
- Air-conditioning system
- Power supply system
- Seismic isolation system
New Earth Simulator Facilities

- Power plant
- Building for operation and research
- Building for computer system

Wiring of interconnection network cables
Cables

R&D results
Total length of IN cables
Wiring of interconnection network cables

Processor Cabinets
Peak Performance
Earth Simulator Computer (ESC)

- \( \text{Rmax from LINPACK MPP Benchmark} \ \text{Ax=b, dense problem} \)
  - Linpack Benchmark = 35.6 TFlop/s
  - Problem of size \( n = 1,041,216 \); (8.7 TB of memory)
  - Half of peak \( \left( n_{\text{f}} \right) \) achieved at \( n_{\text{f}} = 265,408 \)
  - Benchmark took 5.8 hours to run.
  - Algorithm: LU w/partial pivoting
  - Software: for the most part Fortran using MPI

- For the Top500
  - \( \Sigma \) of all the DOE computers = 27.5 TFlop/s
  - Performance of ESC \( \sim \frac{1}{6} \Sigma \) (Top 500 Computers)
  - Performance of ESC \( > \Sigma \) (Top 12 Computers)
  - Performance of ESC \( > \Sigma \) (Top 15 Computers in the US)
  - Performance of ESC \( > \) All the DOE and DOD machines (37.2 TFlop/s)
  - Performance of ESC \( > > \) the 3 NSF Center's computers (7.9 TFlop/s)

SETI@home \( \sim 27 \) TFlop/s

<table>
<thead>
<tr>
<th>Year</th>
<th>Computer</th>
<th>Measured GFlop/s</th>
<th>Measured GFlop/s from Previous Year</th>
<th>Estimated Peak GFlop/s</th>
<th>Estimated Peak GFlop/s from Previous Year</th>
<th>Number of Processors</th>
<th>Size of Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>Earth Simulator Computer, NEC</td>
<td>35610</td>
<td>+ 4.9</td>
<td>40832</td>
<td>+ 3.7</td>
<td>5104</td>
<td>1041216</td>
</tr>
<tr>
<td>2001</td>
<td>ASCI White-Pacific, IBM SP Power 3</td>
<td>7226</td>
<td>1.5</td>
<td>11136</td>
<td>1.0</td>
<td>7424</td>
<td>518996</td>
</tr>
<tr>
<td>2000</td>
<td>ASCI White-Pacific, IBM SP Power 3</td>
<td>4938</td>
<td>2.1</td>
<td>11136</td>
<td>3.5</td>
<td>7424</td>
<td>434600</td>
</tr>
<tr>
<td>1999</td>
<td>ASCI Red Intel Pentium II Xeon core</td>
<td>2379</td>
<td>1.1</td>
<td>3207</td>
<td>0.8</td>
<td>9632</td>
<td>362880</td>
</tr>
<tr>
<td>1998</td>
<td>ASCI Blue-Pacific SST, IBM SP 604E</td>
<td>2144</td>
<td>1.6</td>
<td>3868</td>
<td>2.1</td>
<td>5808</td>
<td>431344</td>
</tr>
<tr>
<td>1997</td>
<td>Intel ASCI Option Red (200 MHz Pentium Pro)</td>
<td>1338</td>
<td>3.6</td>
<td>1830</td>
<td>3.0</td>
<td>9152</td>
<td>235000</td>
</tr>
<tr>
<td>1996</td>
<td>Hitachi CP-PACS</td>
<td>3682</td>
<td>1.3</td>
<td>614</td>
<td>1.8</td>
<td>2048</td>
<td>103680</td>
</tr>
<tr>
<td>1995</td>
<td>Intel Paragon XPS MP</td>
<td>281.1</td>
<td>1</td>
<td>338</td>
<td>1.0</td>
<td>6768</td>
<td>128600</td>
</tr>
<tr>
<td>1994</td>
<td>Intel Paragon XPS MP</td>
<td>281.1</td>
<td>2.3</td>
<td>338</td>
<td>1.4</td>
<td>6768</td>
<td>128600</td>
</tr>
<tr>
<td>1993</td>
<td>Fujitsu NWT</td>
<td>124.5</td>
<td>236</td>
<td></td>
<td></td>
<td>140</td>
<td>31920</td>
</tr>
</tbody>
</table>
# LINPACK Benchmark List

<table>
<thead>
<tr>
<th>Computer</th>
<th>Number of Processes</th>
<th>( R_{\text{max}} ) (Gflop/s)</th>
<th>( N_{\text{max}} ) order</th>
<th>( N_{1/3} ) order</th>
<th>( R_{\text{peak}} ) (Gflop/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earth Simulator, NEC processors****</td>
<td>esc</td>
<td>5104</td>
<td>35610</td>
<td>1041216</td>
<td>263408</td>
</tr>
<tr>
<td>ASCI White-Pacific, IBM SP Power 3 (335 MHz) lnl</td>
<td>8000</td>
<td>7226</td>
<td>518906</td>
<td>179000</td>
<td>12000</td>
</tr>
<tr>
<td>Compaq AlphaServer SC ES45/EV68 1GHz psc</td>
<td>3016</td>
<td>4463</td>
<td>280000</td>
<td>8500</td>
<td>6032</td>
</tr>
<tr>
<td>Compaq AlphaServer SC ES45/EV68 1GHz psc</td>
<td>3024</td>
<td>4059</td>
<td>525000</td>
<td>105000</td>
<td>6048</td>
</tr>
<tr>
<td>Compaq AlphaServer SC ES45/EV68 1GHz cea</td>
<td>2560</td>
<td>3980</td>
<td>360000</td>
<td>8500</td>
<td>5120</td>
</tr>
<tr>
<td>IBM SP Power3 208 nodes 375 MHz lbr</td>
<td>3228</td>
<td>3052</td>
<td>371712</td>
<td>4900</td>
<td></td>
</tr>
<tr>
<td>Compaq Alphaserver SC ES45/EV68 1GHz lbr</td>
<td>2048</td>
<td>2916</td>
<td>272000</td>
<td>4900</td>
<td></td>
</tr>
<tr>
<td>IBM SP Power3 158 nodes 375 MHz lbr</td>
<td>2528</td>
<td>2526</td>
<td>371712</td>
<td>102400</td>
<td>3792</td>
</tr>
<tr>
<td>ASCI Red Intel Pentium II Xeon core 333MHz srl</td>
<td>9632</td>
<td>2379.6</td>
<td>362880</td>
<td>75400</td>
<td>3207</td>
</tr>
<tr>
<td>ASCI Blue-Pacific SST, IBM SP 60E(332 MHz) lbr</td>
<td>5808</td>
<td>2144</td>
<td>431344</td>
<td>432344</td>
<td>3868</td>
</tr>
<tr>
<td>ASCI Red Intel Pentium II Xeon core 333MHz srl</td>
<td>9472</td>
<td>2121.3</td>
<td>251904</td>
<td>66000</td>
<td>3154</td>
</tr>
<tr>
<td>Compaq Alphaserver SC ES45/EV68 1GHz lbr</td>
<td>1520</td>
<td>2096</td>
<td>390000</td>
<td>71000</td>
<td>3040</td>
</tr>
<tr>
<td>IBM SP 112 nodes (375 MHz POWER3 High) ibm</td>
<td>1792</td>
<td>1791</td>
<td>275000</td>
<td>275000</td>
<td>2688</td>
</tr>
<tr>
<td>HITACHI SR8000/MPP/1152(450MHz) u tokyo</td>
<td>1152</td>
<td>1709.1</td>
<td>141000</td>
<td>16000</td>
<td>2074</td>
</tr>
<tr>
<td>HITACHI SR8000-F1/168(375MHz) leibniz</td>
<td>198</td>
<td>1653</td>
<td>160000</td>
<td>19560</td>
<td>2016</td>
</tr>
<tr>
<td>ASCI Red Intel Pentium II Xeon core 333MHz srl</td>
<td>6720</td>
<td>1633.3</td>
<td>306720</td>
<td>525000</td>
<td>2228</td>
</tr>
<tr>
<td>SGI ASCI Blue Mountain lnl</td>
<td>5040</td>
<td>1608</td>
<td>374400</td>
<td>138000</td>
<td>2520</td>
</tr>
<tr>
<td>IBM SP 328 nodes (375 MHz POWER3 Thin) nno</td>
<td>1312</td>
<td>1417</td>
<td>374000</td>
<td>374000</td>
<td>1968</td>
</tr>
<tr>
<td>Intel ASCII Option Red (200 MHz Pentium Pro) srl</td>
<td>9152</td>
<td>1338</td>
<td>235000</td>
<td>63000</td>
<td>1830</td>
</tr>
<tr>
<td>NEC SX-5/128M8(3.2ns) osaka</td>
<td>128</td>
<td>1192.0</td>
<td>129536</td>
<td>10240</td>
<td>1280</td>
</tr>
<tr>
<td>CRAY T3E-1200 (600 MHz) us government</td>
<td>1488</td>
<td>1127</td>
<td>148800</td>
<td>28272</td>
<td>1786</td>
</tr>
<tr>
<td>HITACHI SR8000-F1/112(375MHz) leibniz</td>
<td>112</td>
<td>1035.0</td>
<td>120000</td>
<td>15160</td>
<td>1344</td>
</tr>
</tbody>
</table>

---

# Performance of AFES Climate Code
Physics Model of AFES

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cumulus convection</td>
<td>Condensation, precipitation, convection</td>
</tr>
<tr>
<td></td>
<td>- Simplified Arakawa-Schubert (Arakawa and Schubert, 1974; Moorthi &amp; Suarez, 1992)</td>
</tr>
<tr>
<td></td>
<td>- Kuo scheme + shallow convection</td>
</tr>
<tr>
<td></td>
<td>- Manabe’s moist convection</td>
</tr>
<tr>
<td>Large-scale condensation</td>
<td>Other cloud processes and prediction of cloud water</td>
</tr>
<tr>
<td></td>
<td>(Le Treut &amp; Li, 1990)</td>
</tr>
<tr>
<td>Radiation</td>
<td>2-stream k-distribution scheme (Nakajima &amp; Tanaka, 1986)</td>
</tr>
<tr>
<td>Vertical diffusion</td>
<td>Transport of heat, momentum, and moisture in PBL</td>
</tr>
<tr>
<td></td>
<td>Level 2 turbulence scheme (Mellor &amp; Yamada, 1974, 1982)</td>
</tr>
<tr>
<td>Surface flux</td>
<td>Fluxes in surface boundary layer (Louis, 1979)</td>
</tr>
<tr>
<td></td>
<td>(Mellor et al., 1992)</td>
</tr>
<tr>
<td>Ground process</td>
<td>Multi-layer heat conduction, Hydrology (Manabe, 1979)</td>
</tr>
<tr>
<td></td>
<td>Ground moisture (Manabe et al., 1965)</td>
</tr>
<tr>
<td></td>
<td>Frozen soil process (Clapp &amp; Hornberger, 1978)</td>
</tr>
<tr>
<td></td>
<td>Bucket model (Kondo, 1993)</td>
</tr>
<tr>
<td>Ocean mixing layer</td>
<td>Ocean temperature (Wilson et al, 1987)</td>
</tr>
<tr>
<td></td>
<td>Sea ice</td>
</tr>
<tr>
<td>Gravity wave-induced drag</td>
<td>Orographic effect (McFarlane, 1987)</td>
</tr>
<tr>
<td>Others</td>
<td>Dry convection adjustment</td>
</tr>
</tbody>
</table>

Parallelization of AFES

- **MPI (Top-down approach)** --> among processor nodes
  - Domain decomposition w.r.t. latitude in grid space (S.Pole to N.Pole)
  - Decomposition w.r.t. wave number of Fourier transform in wave domain

- **Microtasking (Bottom-up approach)** --> within node
  - Parallel decomposition of collapsed DO-loop to maximize the length of vector loop
  - Parallelism
    - Vertical direction for Legendre transform
    - Column-wise (2-dimensional) for physical process

- **Vectorization (Bottom-up approach)** --> with 1PE
  - Optimization of vector loop
  - Maximization of loop length with DO-loop collapse
**Optimization Strategies for AIRS Climate Model**

- High resolution (10km) resulting in increased cost concentration on vector-tailored dynamics part (>75%)
- MPI among nodes / Microtasking within node
- Domain decomposition that fully exploits parallel nodes (>99% parallelization ratio) with less communication
- Reduced load imbalance due to improved algorithms (e.g., Use of increasingly popular Kuo cloud physics model)
- Improved vector performance with DO-loop optimization
- Combined use of assembler coding for part of matrix operations

---

**Strategy for Performance Enhancement for the ES**

- **Minimization of serial sections**
  - Most dominant factors affecting the total performance of applications

- **Pursuit of reduced communication overhead**

- **Increase of vector performance**
  - Effective combination of vector and parallel processing efficiency
Effective Performance of the AFES Climate Code on the ES with the Kuo’s Cumulus Convection Scheme for a T1279L96 Resolution Model

**26.6 TFLOP/S sustained performance with the 640 full nodes (5120 CPUs/peak 40 TFLOP/S)**
Results from AFES

Precipitation (312km T42L24)
Precipitation (125 km T106L24)

Precipitation (20.8 km T639L24)
Precipitation (10.4 km T1279L24)

Specific Humidity at 850hPa (about 1500 m a.s.l.)
Cyclones around the Madagascar Islands

Seasonal Variation of Sea Surface Temperature

10 km resolution for oceans (previously 100km)
Distributed and Parallel Systems

Distributed systems heterogeneous

- Gather (unused) resources
- Steal cycles
- System SW manages resources
- System SW adds value
- 10% - 20% overhead is OK
- Resources drive applications
- Time to completion is not critical
- Time-shared
- SETI@home
  - ~ 400,000 machines
  - Averaging 27 Tflop/s

Massively parallel systems homogeneous

- Bounded set of resources
- Apps grow to consume all cycles
- Application manages resources
- System SW gets in the way
- 5% overhead is maximum
- Apps drive purchase of equipment
- Real-time constraints
- Space-shared
- ASCI White LLNL
  - 8000 processors
  - Averaging 7.2 Tflop/s

Performance

1976: The Supercomputing “Island”

Today: A Continuum

NUMBER OF MACHINES

PERFORMANCE
The Future of HPC

- Great excitement in the area of High Performance Computing
- The expense of being different is being replaced by the economics of being the same
- HPC needs to lose its "special purpose" tag
- Still has to bring about the promise of scalable general purpose computing ...
- … but it is dangerous to ignore this technology
- Final success when MPP technology is embedded in desktop computing
- Yesterday's HPC is today's mainframe is tomorrow's workstation

Highly Parallel Supercomputing: Where Are We?

- Performance:
  - Sustained performance has dramatically increased during the last year.
  - On most applications, sustained performance per dollar now exceeds that of conventional supercomputers. But...
  - Conventional systems are still faster on some applications.
- Languages and compilers:
  - Standardized, portable, high-level languages such as HPF, PVM and MPI are available. But ...
  - Initial HPF releases are not very efficient.
  - Message passing programming is tedious and hard to debug.
  - Programming difficulty remains a major obstacle to usage by mainstream scientist.
Highly Parallel Supercomputing: Where Are We?

- Operating systems:
  - Robustness and reliability are improving.
  - New system management tools improve system utilization. But...
  - Reliability still not as good as conventional systems.

- I/O subsystems:
  - New RAID disks, HiPPI interfaces, etc. provide substantially improved I/O performance. But...
  - I/O remains a bottleneck on some systems.

The Importance of Standards - Software

- Writing programs for MPP is hard ...
- But ... one-off efforts if written in a standard language
- Past lack of parallel programming standards ...
  - ... has restricted uptake of technology (to "enthusiasts")
  - ... reduced portability (over a range of current architectures and between future generations)
- Now standards exist: (MPI, OpenMP, PVM, & HPF), which ...
  - ... allows users & manufacturers to protect software investment
  - ... encourage growth of a “third party” parallel software industry & parallel versions of widely used codes
The Importance of Standards - Hardware

- Processors
  - commodity RISC processors
- Interconnects
  - high bandwidth, low latency communications protocol
  - no de-facto standard yet (ATM, Fibre Channel, HPPI, FDDI)
- Growing demand for total solution:
  - robust hardware + usable software
- HPC systems containing all the programming tools / environments / languages / libraries / applications packages found on desktops

Achieving TeraFlops

- In 1991 we had, 1 Gflop/s
- Today, 1000 fold increase
  - Architecture
    - exploiting parallelism
  - Processor, communication, memory
    - Moore’s Law
  - Algorithm improvements
    - block-partitioned algorithms
Future: Petaflops ( $10^{15}$ flops/s)

Today $\approx \sqrt{10^{15}}$ flops for our workstations

- A Pflop for 1 second $\approx$ a typical workstation computing for 1 year.
- From an algorithmic standpoint
  - concurrency
  - data locality
  - latency & sync
  - floating point accuracy
  - dynamic redistribution of workload
  - new language and constructs
  - role of numerical libraries
  - algorithm adaptation to hardware failure

A Petaflops Computer System

- 1 Pflop/s sustained computing
- Between 10,000 and 1,000,000 processors
- Between 10 TB and 1PB main memory
- Commensurate I/O bandwidth, mass store, etc.
- If built today, cost $40B and consume 1 TWatt.
- May be feasible and “affordable” by the year 2010