High Performance Computing, Computational Grid, and Numerical Libraries

Jack Dongarra  
Innovative Computing Lab  
University of Tennessee  
and  
Computer Science and Math Div  
Oak Ridge National Lab  
http://www.cs.utk.edu/~dongarra/

Technology Trends: Microprocessor Capacity

Gordon Moore (co-founder of Intel) predicted in 1965 that the transistor density of semiconductor chips would double roughly every 18 months.

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.  
2X memory and processor speed and ½ size, cost, & power 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
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!
Fastest Computer Over Time

In 1980 a computation that took 1 full year to complete can today be done in ~ 27 seconds!

In 1980 a computation that took 1 full year to complete can today be done in ~ 5.4 seconds!
Machines at the Top of the List

<table>
<thead>
<tr>
<th>Year</th>
<th>Computer</th>
<th>Measured Gflop/s</th>
<th>Factor Δ from Previous Year</th>
<th>Theoretical Peak Gflop/s</th>
<th>Factor Δ from Previous Year</th>
<th>Number of Processors</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>Earth Simulator Computer, NEC</td>
<td>35860</td>
<td>5.0</td>
<td>40960</td>
<td>3.7</td>
<td>5120</td>
<td>88%</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>65%</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>44%</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>74%</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>55%</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>73%</td>
</tr>
<tr>
<td>1996</td>
<td>Hitachi CP-PACS</td>
<td>368.2</td>
<td>1.3</td>
<td>614</td>
<td>1.8</td>
<td>2048</td>
<td>60%</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>83%</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>83%</td>
</tr>
<tr>
<td>1993</td>
<td>Fujitsu NWT</td>
<td>124.5</td>
<td>236</td>
<td>140</td>
<td>53%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A Tour de Force in Engineering

- **Homogeneous, Centralized, Proprietary, Expensive!**
- **Target Application:** CFD-Weather, Climate, Earthquakes
- **640 NEC SX/6 Nodes (mod)**
  - 5120 CPUs which have vector ops
  - Each CPU 8 Gflop/s Peak
- **40 TFlop/s (peak)**
- **$250-$500 million for things in building**
- **Footprint of 4 tennis courts**
- **7 MWatts**
  - Say 10 cent/KWhr - $16.8K/day = $6M/year!
- **Expect to be on top of Top500 until 60-100 TFlop ASCI machine arrives**

- **For the Top500 (November 2002)**
  - Performance of ESC
    - ≈ Σ Next Top 7 Computers
  - Σ of DOE computers (DP&OS) = 49 TFlop/s
### 20th List: The 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>35.86</td>
<td>Earth Simulator Center</td>
<td>Japan</td>
<td>2002</td>
<td>Research</td>
<td>5120</td>
</tr>
<tr>
<td>2</td>
<td>HP</td>
<td>ASCI Q, AlphaServer SC</td>
<td>7.73</td>
<td>Los Alamos National Laboratory</td>
<td>USA</td>
<td>2002</td>
<td>Research</td>
<td>4096</td>
</tr>
<tr>
<td>2</td>
<td>HP</td>
<td>ASCI Q, AlphaServer SC</td>
<td>7.73</td>
<td>Los Alamos National Laboratory</td>
<td>USA</td>
<td>2002</td>
<td>Research</td>
<td>4096</td>
</tr>
<tr>
<td>4</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>5</td>
<td>Linux NetworX</td>
<td>MCR Cluster</td>
<td>5.69</td>
<td>Lawrence Livermore National Laboratory</td>
<td>USA</td>
<td>2002</td>
<td>Research</td>
<td>8192</td>
</tr>
<tr>
<td>6</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>7</td>
<td>HP</td>
<td>AlphaServer SC ES45 1 GHz</td>
<td>3.98</td>
<td>Commissariat a l’Energie Atomique (CEA)</td>
<td>France</td>
<td>2001</td>
<td>Research</td>
<td>2560</td>
</tr>
<tr>
<td>8</td>
<td>HPTi</td>
<td>Xeon Cluster - Myrinet2000</td>
<td>3.34</td>
<td>Forecast Systems Laboratory - NOAA</td>
<td>USA</td>
<td>2002</td>
<td>Research</td>
<td>1536</td>
</tr>
<tr>
<td>9</td>
<td>IBM</td>
<td>pSeries 690 Turbo</td>
<td>3.16</td>
<td>HPCx</td>
<td>UK</td>
<td>2002</td>
<td>Academic</td>
<td>1280</td>
</tr>
<tr>
<td>10</td>
<td>IBM</td>
<td>pSeries 690 Turbo</td>
<td>3.16</td>
<td>NCAR (National Center for Atmospheric Research)</td>
<td>USA</td>
<td>2002</td>
<td>Research</td>
<td>1216</td>
</tr>
</tbody>
</table>

### Response to the Earth Simulator: IBM Blue Gene/L and ASCI Purple

- **Announced 11/19/02**
  - One of 2 machines for LLNL
  - 360 TFlop/s
  - 130,000 proc
  - Linux
  - FY 2005

- Plus ASCI Purple
  - IBM Power 5 based
  - 12K proc, 100 TFlop/s
DOE ASCI
Red Storm Sandia National Lab

- 10,368 compute processors, 108 cabinets
  - AMD Opteron @ 2.0 GHz
  - Cray integrator and providing the interconnect
- Fully connected high performance 3-D mesh interconnect.
  - Topology - 27 X 16 X 24
- Peak of ~ 40 TF
  - Expected MP-Linpack >20 TF
- Aggregate system memory bandwidth - ~55 TB/s
- MPI Latency - 2 ms neighbor, 5 ms across machine
- Bi-Section bandwidth ~2.3 TB/s
- Link bandwidth ~3.0 GB/s in each direction

2004 in operation

TOP500 - Performance
Performance Extrapolation

Performance Extrapolation

Earth Simulator

TFlop/s

To enter the list

PFlop/s

computer

My Laptop

ASCI II

12,544 proc

130,000 proc

N=500

N=1

100 MFlop/s

100 TFlop/s

1 PFlop/s

10 GFlop/s

100 GFlop/s

10 GFlop/s

10 TFlop/s

100 TFlop/s

100 MFlop/s

10 PFlop/s

10 PFlop/s

Jun-93

Jun-94

Jun-95

Jun-96

Jun-97

Jun-98

Jun-99

Jun-00

Jun-01

Jun-02

Jun-03

Jun-04

Jun-05

Jun-06

Jun-07

Jun-08

Jun-09

Jun-10
A total of 56 Intel based and 8 AMD based PC clusters are in the TOP500.  
- 31 of these Intel based cluster are IBM Netfinity systems delivered by IBM.

A substantial part of these are installed at industrial customers especially in the oil-industry.  
- Including 5 Sun and 5 Alpha based clusters and 21 HP AlphaServer.

15 of these clusters are labeled as 'Self-Made'.
Processor Breakdown for the 93 Clusters

- Pentium III: 28, 30%
- Pentium 4: 24, 26%
- Itanium: 4, 4%
- Sparc: 4, 4%
- AMD: 8, 9%
- Alpha: 25, 27%

Linux: Plotting The Future

- First Cluster on Top500 List
- Berkeley NOW (Solars)

Legend:
- % Linux Machines
- % Aggregate Performance
- Moore's Law (18mo)
Linux: Plotting The Future

How Long Until Total World Domination?

Predicting Future Market Share
How Large Can Linux Clusters Get?

Linux Cluster Sizes: Plotting The Future
Observations

- The adoption rate of Linux HPC is phenomenal!
  - Linux in the Top500 is doubling every 12 months
  - Linux adoption is not driven by bottom feeders
    - Adoption is actually faster at the ultra-scale!
- The CPU counts for the largest Linux clusters are currently doubling every year
- Prediction: by 2005, we will have a 10,000 CPU Linux cluster
- Prediction: by 2005, most top-performing supercomputers will be running Linux
- Adoption rate driven largely by economics and human factors

Distributed and Parallel Systems

- Distributed systems heterogeneous
  - SETI@home
  - Grid based Computing
  - Google
  - Special Interconnect
  - Parallel Dist Mem
  - ASCI TlOps
  - Massively parallel systems homogeneous

- 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
  - ~ 500,000 machines
  - Averaging 55 Tflop/s

- 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
- Earth Simulator
  - 5000 processors
  - Averaging 35 Tflop/s
SETI@home: Global Distributed Computing

♦ Running on 500,000 PCs, ~1300 CPU Years per Day
  ➢ 1.3M CPU Years so far
♦ Sophisticated Data & Signal Processing Analysis
♦ Distributes Datasets from Arecibo Radio Telescope

Use thousands of Internet-connected PCs to help in the search for extraterrestrial intelligence.
♦ When their computer is idle or being wasted this software will download ~ half a MB chunk of data for analysis. Performs about 3 Tflops for each client in 15 hours.
♦ The results of this analysis are sent back to the SETI team, combined with thousands of other participants.

♦ Largest distributed computation project in existence
  ➢ Averaging 55 Tflop/s
♦ Today a number of companies trying this for profit.
The project employs computational chemistry to analyze chemical interactions between a library of 35 million potential drug molecules and several protein targets on the smallpox virus in the search for an effective anti-viral drug to treat smallpox post-infection.

Google query attributes

- 150M queries/day (2000/second)
- 100 countries
- 3B documents in the index

Data centers

- 15,000 Linux systems in 6 data centers
- 15 TFlop/s and 1000 TB total capability
- 40-80 1U/2U servers/cabinet
- 100 MB Ethernet switches/cabinet with gigabit Ethernet uplink
- growth from 4,000 systems (June 2000)
- 18M queries then

Performance and operation

- simple reissue of failed commands to new servers
- no performance debugging

Source: Monika Henzinger, Google
Today there is a complex interplay and increasing interdependence among the sciences. Many science and engineering problems require widely dispersed resources be operated as systems. What we do as collaborative infrastructure developers will have profound influence on the future of science. Networking, distributed computing, and parallel computation research have matured to make it possible for distributed systems to support high-performance applications, but...

- Resources are dispersed
- Connectivity is variable
- Dedicated access may not be possible

**Today: Collaboration**

### The Grid

**Problem Solving Environments**
Scientists and engineers using computation to accomplish lab missions

**Hardware**
Heterogeneous collection of high-performance computer hardware and software resources

**Software**
Software applications and components for computational problems

**Networking**
The hardware and software that permits communication among distributed users and computer resources

**Mass Storage**
A collection of devices and software that allow temporary and long-term archival storage of information

**Intelligent Interface**
A knowledge-based environment that offers user guidance on complex computing tasks

**Middleware**
Software tools that enable interaction among users, applications, and system resources

**Grid Operating System**
The software that coordinates the interplay of computers, networking, and software
The Knoxville Campus has two DS-3 commodity Internet connections and one DS-3 Internet2/Abilene connection. UT participates in several national networking initiatives including Internet2 (I2), Abilene, the federal Next Generation Internet (NGI) initiative, Southern Universities Research Association (SURA) Regional Information Infrastructure (RII), and Southern Crossroads (SoX).

The UT campus consists of a meshed ATM OC-12 being migrated over to switched Gigabit by early 2002.
Grids vs. Capability Computing

- **Not an “either/or” question**
  - Each addresses different needs
  - Both are part of an integrated solution

- **Grid strengths**
  - Coupling necessarily distributed resources
    - instruments, software, hardware, archives, and people
  - Eliminating time and space barriers
    - remote resource access and capacity computing
  - Grids are not a cheap substitute for capability HPC

- **Capability computing strengths**
  - Supporting foundational computations
    - terascale and petascale “nation scale” problems
  - Engaging tightly coupled teams and computations

Futures for Numerical Algorithms and Software

- **Numerical software will be adaptive, exploratory, and intelligent**
- **Determinism in numerical computing will be gone.**
  - After all, its not reasonable to ask for exactness in numerical computations.
  - Auditability of the computation, reproducibility at a cost
- **Fault Tolerance**
  - Google claims 15K nodes, what do they do when one goes down?
  - We must do better than “restart ALL nodes from last chkpt”
- **Importance of floating point arithmetic will be undiminished.**
  - 16, 32, 64, 128 bits and beyond.
- **Reproducibility, fault tolerance, and auditability**
- **Adaptivity** is a key so applications can effectively use the resources.
Collaborators / Support

- **TOP500**
  - H. Mauer, Mannheim U
  - H. Simon, NERSC
  - E. Strohmaier, NERSC

Thanks

Next Generation Software

SciDAC Scientific Discovery through Advanced Computing