LINPACK Benchmarking and Beyond

Jack Dongarra

University of Tennessee
Oak Ridge National Laboratory
H. Meuer, H. Simon, E. Strohmaier, & JD

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

- 59.7 GFlop/s
- 1.17 TFlop/s
- 100 P flop/s
- 24.7 TFlop/s
- 32.4 PFlop/s
- 1.76 PFlop/s
- 6-8 years
- N=500
- My Laptop
- SUM
- N=1
- 400 MFlop/s
- 1 Gflop/s
- 100 Gflop/s
- 100 Mflop/s
Processors Used in the Top500 Systems

- Intel 81%
- AMD 10%
- IBM 8%
Today’s Multicores
99% of Top500 Systems Are Based on Multicore

Sun Niagara2 (8 cores)
IBM Power 7 (8 cores)
Fujitsu Venus (8 cores)
IBM Cell (9 cores)
AMD Istambul (6 cores)
IBM BG/P (4 cores)
Intel Xeon (8 cores)
Intel Polaris [experimental] (80 cores)

Of the Top500, 499 are multicore.
Performance of Countries

Total Performance [Tflop/s]

US
Performance of Countries

[Graph showing the total performance in Tflops/s for the US and EU from 2000 to 2010. The graph indicates a steady increase in performance for both countries, with the US surpassing the EU in 2008 and maintaining a lead through 2010.]
Performance of Countries

Total Performance [Tflop/s]
Performance of Countries

Total Performance [Tflop/s]


US
EU
Japan
China
Countries / System Share
## 35rd List: The TOP10

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Computer</th>
<th>Country</th>
<th>Cores</th>
<th>Rmax [Pflops]</th>
<th>% of Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DOE / OS Oak Ridge Nat Lab</td>
<td>Jaguar / Cray Cray XT5 sixCore 2.6 GHz</td>
<td>USA</td>
<td>224,162</td>
<td>1.76</td>
<td>75</td>
</tr>
<tr>
<td>2</td>
<td>Nat. Supercomputer Center in Shenzhen</td>
<td>Nebulea / Dawning / TC3600 Blade, Intel X5650, Nvidia C2050 GPU</td>
<td>China</td>
<td>120,640</td>
<td>1.27</td>
<td>43</td>
</tr>
<tr>
<td>3</td>
<td>DOE / NNSA Los Alamos Nat Lab</td>
<td>Roadrunner / IBM BladeCenter QS22/LS21</td>
<td>USA</td>
<td>122,400</td>
<td>1.04</td>
<td>76</td>
</tr>
<tr>
<td>4</td>
<td>NSF / NICS / U of Tennessee</td>
<td>Kraken/ Cray Cray XT5 sixCore 2.6 GHz</td>
<td>USA</td>
<td>98,928</td>
<td>.831</td>
<td>81</td>
</tr>
<tr>
<td>5</td>
<td>Forschungszentrum Juelich (FZJ)</td>
<td>Jugene / IBM Blue Gene/P Solution</td>
<td>Germany</td>
<td>294,912</td>
<td>.825</td>
<td>82</td>
</tr>
<tr>
<td>6</td>
<td>NASA / Ames Research Center/NAS</td>
<td>Pleiades / SGI SGI Altix ICE 8200EX</td>
<td>USA</td>
<td>56,320</td>
<td>.544</td>
<td>82</td>
</tr>
<tr>
<td>7</td>
<td>National SC Center in Tianjin / NUDT</td>
<td>Tianhe-1 / NUDT TH-1 / IntelQC + AMD ATI Radeon 4870</td>
<td>China</td>
<td>71,680</td>
<td>.563</td>
<td>46</td>
</tr>
<tr>
<td>8</td>
<td>DOE / NNSA Lawrence Livermore NL</td>
<td>BlueGene/L IBM eServer Blue Gene Solution</td>
<td>USA</td>
<td>212,992</td>
<td>.478</td>
<td>80</td>
</tr>
<tr>
<td>9</td>
<td>DOE / OS Argonne Nat Lab</td>
<td>Intrepid / IBM Blue Gene/P Solution</td>
<td>USA</td>
<td>163,840</td>
<td>.458</td>
<td>82</td>
</tr>
<tr>
<td>10</td>
<td>DOE / NNSA Sandia Nat Lab</td>
<td>Red Sky / Sun / SunBlade 6275</td>
<td>USA</td>
<td>42,440</td>
<td>.433</td>
<td>87</td>
</tr>
<tr>
<td>Rank</td>
<td>Site</td>
<td>Computer</td>
<td>Country</td>
<td>Cores</td>
<td>Rmax [Pflops]</td>
<td>% of Peak</td>
</tr>
<tr>
<td>------</td>
<td>------------------------------------------</td>
<td>---------------------------------------</td>
<td>---------</td>
<td>--------</td>
<td>---------------</td>
<td>-----------</td>
</tr>
<tr>
<td>1</td>
<td>DOE / OS Oak Ridge Nat Lab</td>
<td>Jaguar / Cray Cray XT5 sixCore 2.6 GHz</td>
<td>USA</td>
<td>224,162</td>
<td>1.76</td>
<td>75</td>
</tr>
<tr>
<td>2</td>
<td>Nat. Supercomputer Center in Shenzhen</td>
<td>Nebulea / Dawning / TC3600 Blade, Intel X5650, Nvidia C2050 GPU</td>
<td>China</td>
<td>120,640</td>
<td>1.27</td>
<td>43</td>
</tr>
<tr>
<td>3</td>
<td>DOE / NNSA Los Alamos Nat Lab</td>
<td>Roadrunner / IBM BladeCenter QS22/LS21</td>
<td>USA</td>
<td>122,400</td>
<td>1.04</td>
<td>76</td>
</tr>
<tr>
<td>4</td>
<td>NSF / NICS / U of Tennessee</td>
<td>Kraken/ Cray Cray XT5 sixCore 2.6 GHz</td>
<td>USA</td>
<td>98,928</td>
<td>.831</td>
<td>81</td>
</tr>
<tr>
<td>5</td>
<td>Forschungszentrum Juelich (FZJ)</td>
<td>Jugene / IBM Blue Gene/P Solution</td>
<td>Germany</td>
<td>294,912</td>
<td>.825</td>
<td>82</td>
</tr>
<tr>
<td>6</td>
<td>NASA / Ames Research Center/NAS</td>
<td>Pleiades / SGI SGI Altix ICE 8200EX</td>
<td>USA</td>
<td>56,320</td>
<td>.544</td>
<td>82</td>
</tr>
<tr>
<td>7</td>
<td>National SC Center in Tianjin / NUDT</td>
<td>Tianhe-1 / NUDT TH-1 / IntelQC + AMD ATI Radeon 4870</td>
<td>China</td>
<td>71,680</td>
<td>.563</td>
<td>46</td>
</tr>
<tr>
<td>8</td>
<td>DOE / NNSA Lawrence Livermore NL</td>
<td>BlueGene/L IBM eServer Blue Gene Solution</td>
<td>USA</td>
<td>212,992</td>
<td>.478</td>
<td>80</td>
</tr>
<tr>
<td>9</td>
<td>DOE / OS Argonne Nat Lab</td>
<td>Intrepid / IBM Blue Gene/P Solution</td>
<td>USA</td>
<td>163,840</td>
<td>.458</td>
<td>82</td>
</tr>
<tr>
<td>10</td>
<td>DOE / NNSA Sandia Nat Lab</td>
<td>Red Sky / Sun / SunBlade 6275</td>
<td>USA</td>
<td>42,440</td>
<td>.433</td>
<td>87</td>
</tr>
</tbody>
</table>
Recently upgraded to a 2 Pflop/s system with more than 224K cores using AMD's 6 Core chip.

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak performance</td>
<td>2.332 PF</td>
</tr>
<tr>
<td>System memory</td>
<td>300 TB</td>
</tr>
<tr>
<td>Disk space</td>
<td>10 PB</td>
</tr>
<tr>
<td>Disk bandwidth</td>
<td>240+ GB/s</td>
</tr>
<tr>
<td>Interconnect bandwidth</td>
<td>374 TB/s</td>
</tr>
</tbody>
</table>
Nebulae

Hybrid system, commodity + GPUs

Theoretical peak 2.98 Pflop/s

Linpack Benchmark at 1.27 Pflop/s

4640 nodes, each node:
  2 Intel 6-core Xeon5650 + Nvidia Fermi C2050 GPU (each 14 cores)
  ➢ 120,640 cores
  ➢ Infiniband connected

➢ 500 MB/s peak per link and 8 GB/s
Commodity plus Accelerators

**Commodity**
- Intel Xeon
- 8 cores
- 3 GHz
- 8*4 ops/cycle
- 96 Gflop/s (DP)

**Accelerator (GPU)**
- Nvidia C2050 “Fermi”
- 448 “Cuda cores”
- 1.15 GHz
- 448 ops/cycle
- 515 Gflop/s (DP)

---

**Diagram:**
- X86 Host
- Host Memory
- Device Memory
- Interconnect
  - PCI Express
  - 512 MB/s to 32GB/s
  - 8 MW – 512 MW
- Thread Execution Control Unit
- Thread Processors
- Special Function Unit
- DMA
#3 LANL Roadrunner
A Petascale System in 2008

“Connected Unit” cluster
192 Opteron nodes
(180 w/ 2 dual-Cell blades connected w/ 4 PCIe x8
links)

17 clusters

≈ 13,000 Cell HPC chips
≈ 1.33 PetaFlop/s (from Cell)
≈ 7,000 dual-core Opterons
≈ 122,000 cores

2nd stage InfiniBand 4x DDR interconnect
(18 sets of 12 links to 8 switches)

2nd stage InfiniBand interconnect (8 switches)

Based on the 100 Gflop/s (DP) Cell chip

Hybrid Design (2 kinds of chips & 3 kinds of cores)
Programming required at 3 levels.

Dual Core Opteron Chip
Looking at the Gordon Bell Prize
(Recognize outstanding achievement in high-performance computing applications and encourage development of parallel processing)

- 1 GFlop/s; 1988; Cray Y-MP; 8 Processors
  - Static finite element analysis

- 1 TFlop/s; 1998; Cray T3E; 1024 Processors
  - Modeling of metallic magnet atoms, using a variation of the locally self-consistent multiple scattering method.

- 1 PFlop/s; 2008; Cray XT5; 1.5x10^5 Processors
  - Superconductive materials

- 1 EFlop/s; ~2018; ?; 1x10^7 Processors (10^9 threads)
## Potential System Architecture

<table>
<thead>
<tr>
<th></th>
<th>2009</th>
<th>2019</th>
<th>Difference Today &amp; 2019</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System peak</strong></td>
<td>2 Pflop/s</td>
<td>1 Eflop/s</td>
<td>O(1000)</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>6 MW</td>
<td>~20 MW</td>
<td></td>
</tr>
<tr>
<td><strong>System memory</strong></td>
<td>0.3 PB</td>
<td>32 - 64 PB [ .03 Bytes/Flop ]</td>
<td>O(100)</td>
</tr>
<tr>
<td><strong>Node performance</strong></td>
<td>125 GF</td>
<td>1,2 or 15TF</td>
<td>O(10) - O(100)</td>
</tr>
<tr>
<td><strong>Node memory BW</strong></td>
<td>25 GB/s</td>
<td>2 - 4TB/s [ .002 Bytes/Flop ]</td>
<td>O(100)</td>
</tr>
<tr>
<td><strong>Node concurrency</strong></td>
<td>12</td>
<td>O(1k) or 10k</td>
<td>O(100) - O(1000)</td>
</tr>
<tr>
<td><strong>Total Node Interconnect BW</strong></td>
<td>3.5 GB/s</td>
<td>200-400GB/s (1:4 or 1:8 from memory BW)</td>
<td>O(100)</td>
</tr>
<tr>
<td><strong>System size (nodes)</strong></td>
<td>18,700</td>
<td>O(100,000) or O(1M)</td>
<td>O(10) - O(100)</td>
</tr>
<tr>
<td><strong>Total concurrency</strong></td>
<td>225,000</td>
<td>O(billion) [O(10) to O(100) for latency hiding]</td>
<td>O(10,000)</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>15 PB</td>
<td>500-1000 PB (&gt;10x system memory is min)</td>
<td>O(10) - O(100)</td>
</tr>
<tr>
<td><strong>IO</strong></td>
<td>0.2 TB</td>
<td>60 TB/s (how long to drain the machine)</td>
<td>O(100)</td>
</tr>
<tr>
<td><strong>MTTI</strong></td>
<td>days</td>
<td>O(1 day)</td>
<td>- O(10)</td>
</tr>
</tbody>
</table>
Exascale \((10^{18} \text{ Flop/s})\) Systems: Two possible paths

- **Light weight processors (think BG/P)**
  - \(~1 \text{ GHz processor (}10^9\)\)
  - \(~1 \text{ Kilo cores/socket (}10^3\)\)
  - \(~1 \text{ Mega sockets/system (}10^6\)\)

- **Hybrid system (think GPU based)**
  - \(~1 \text{ GHz processor (}10^9\)\)
  - \(~10 \text{ Kilo FPUs/socket (}10^4\)\)
  - \(~100 \text{ Kilo sockets/system (}10^5\)\)
Factors that Necessitate Redesign of Our Software

- Steepness of the ascent from terascale to petascale to exascale
- Extreme parallelism and hybrid design
  - Preparing for million/billion way parallelism
- Tightening memory/bandwidth bottleneck
  - Limits on power/clock speed implication on multicore
  - Reducing communication will become much more intense
  - Memory per core changes, byte-to-flop ratio will change
- Necessary Fault Tolerance
  - MTTF will drop
  - Checkpoint/restart has limitations

Software infrastructure does not exist today
Moore’s Law reinterpreted

• Number of cores per chip will double every two years
• Clock speed will not increase (possibly decrease) because of Power

\[ \text{Power} \propto \text{Voltage}^2 \times \text{Frequency} \]
\[ \text{Voltage} \propto \text{Frequency} \]
\[ \text{Power} \propto \text{Frequency}^3 \]

• Need to deal with systems with millions of concurrent threads
• Need to deal with inter-chip parallelism as well as intra-chip parallelism
What’s Next?

Different Classes of Chips
- Home
- Games / Graphics
- Business
- Scientific

- All Large Core
- Mixed Large and Small Core
- Many Small Cores
- All Small Core
- Many Floating-Point Cores
- 3D Stacked Memory
- photonic NoC
- 3D memory layers
- multi-core processor layer

+ 3D Stacked Memory
Major Changes to Software

• Must rethink the design of our software
  ▪ Another disruptive technology
    • Similar to what happened with cluster computing and message passing
  ▪ Rethink and rewrite the applications, algorithms, and software

• Numerical libraries for example will change
  ▪ For example, both LAPACK and ScaLAPACK will undergo major changes to accommodate this
Five Important Features to Consider When Computing at Scale

1. Effective Use of Many-Core and Hybrid architectures
   - Break fork-join parallelism
   - Dynamic Data Driven Execution
   - Block Data Layout

2. Exploiting Mixed Precision in the Algorithms
   - Single Precision is 2X faster than Double Precision
   - With GP-GPUs 10x
   - Power saving issues

3. Self Adapting / Auto Tuning of Software
   - Too hard to do by hand

4. Fault Tolerant Algorithms
   - With 1,000,000’s of cores things will fail

5. Communication Reducing Algorithms
   - For dense computations from $O(n \log p)$ to $O(\log p)$ communications
   - Asynchronous iterations
   - GMRES k-step compute ($x, Ax, A^2x, \ldots A^kx$)
**LAPACK LU - Intel64 - 16 cores**

DGETRF - Intel64 Xeon quad-socket quad-core (16 cores) - th. peak 153.6 Gflop/s
LAPACK LU

• Fork-join, bulk synchronous processing
Parallel Tasks in LU

- Break into smaller tasks and remove dependencies
LU - Intel64 - 16 cores

DGETRF - Intel64 Xeon quad-socket quad-core (16 cores)
theoretical peak 153.6 Gflop/s

Matrix size vs Gflop/s for DGEMM, PLASMA, and LAPACK.
Performance - 24 cores, LU

LU Performance (double prec.)

AMD Istanbul, 2.8 GHz, 4 sockets (24 cores)

- PLASMA
- MKL 11.0
- LAPACK

Gflop/s vs. Size

0 2000 4000 6000 8000 10000 12000
Fault Tolerance

• Trends in HPC:
  ▪ High end systems with millions of processors.

• Increased probability of a system failure
  ▪ Most nodes today are robust, 3 year life.
  ▪ Mean Time to Failure is growing shorter as systems grow, devices shrink, parts are powered on and off...

• MPI widely accepted in scientific computing.
  ▪ Process faults not tolerated in MPI model.

Interesting studies:
• The computer failure data repository (CFDR) http://cfdr.usenix.org/
Classic approach for FT: Checkpoint-Restart

Typical “Balanced Architecture” for PetaScale Computers

<table>
<thead>
<tr>
<th>Systems</th>
<th>Perf.</th>
<th>Ckpt time</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoadRunner</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLNL BG/L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argonne BG/P</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total SGI Altix</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDRIS BG/P</td>
<td>100 TF</td>
<td>30 min.</td>
<td>IDRIS</td>
</tr>
</tbody>
</table>
• Hardware has changed dramatically while software ecosystem has remained stagnant
• Need to exploit new hardware trends (e.g., manycore, heterogeneity) that cannot be handled by existing software stack, memory per socket trends
• Emerging software technologies exist, but have not been fully integrated with system software, e.g., UPC, Cilk, CUDA, HPCS
• Community codes unprepared for sea change in architectures
• No global evaluation of key missing components
International Exascale Software Program

Improve the world’s simulation and modeling capability by improving the coordination and development of the HPC software environment

Workshops:

Build an international plan for coordinating research for the next generation open source software for scientific high-performance computing

www.exascale.org
International Community Effort

- We believe this needs to be an international collaboration for various reasons including:
  - The scale of investment
  - The need for international input on requirements
  - US, Europeans, Asians, and others are working on their own software that should be part of a larger vision for HPC.
  - No global evaluation of key missing components
  - Hardware features are uncoordinated with software development
Where We Are Today:

- SC08 (Austin TX) meeting to generate interest
- Funding from DOE’s Office of Science & NSF Office of Cyberinfrastructure and sponsorship by Europeans and Asians
- US meeting (Santa Fe, NM) April 6-8, 2009
  - 65 people
- European meeting (Paris, France) June 28-29, 2009
  - 70 people
  - Outline Report
- Asian meeting (Tsukuba Japan) October 18-20, 2009
  - Draft roadmap
  - Refine Report
- SC09 (Portland OR) BOF to inform others
  - Public Comment
  - Draft Report presented
- European meeting (Oxford, UK) April 13-14, 2010
  - Refine and prioritize roadmap
  - Explore governance structure and management models for IESP
4.1 Systems Software
- 4.1.1 Operating systems
- 4.1.2 Runtime Systems
- 4.1.2 I/O systems
- 4.1.3 External Environments
- 4.1.4 Systems Management

4.2 Development Environments
- 4.2.1 Programming Models
- 4.2.2 Frameworks
- 4.2.3 Compilers
- 4.2.4 Numerical Libraries
- 4.2.5 Debugging tools

4.3 Applications
- 4.3.1 Application Element: Algorithms
- 4.3.2 Application Support: Data Analysis and Visualization
- 4.3.3 Application Support: Scientific Data Management

4.4 Crosscutting Dimensions
- 4.4.1 Resilience
- 4.4.2 Power Management
- 4.4.3 Performance Optimization
- 4.4.4 Programmability
• www.exascale.org
If you are wondering what’s beyond ExaFlops

Mega, Giga, Tera, Peta, Exa, Zetta ...

<table>
<thead>
<tr>
<th>Power</th>
<th>Symbol</th>
<th>Prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^3$</td>
<td>kilo</td>
<td></td>
</tr>
<tr>
<td>$10^6$</td>
<td>mega</td>
<td></td>
</tr>
<tr>
<td>$10^9$</td>
<td>giga</td>
<td></td>
</tr>
<tr>
<td>$10^{12}$</td>
<td>tera</td>
<td></td>
</tr>
<tr>
<td>$10^{15}$</td>
<td>peta</td>
<td></td>
</tr>
<tr>
<td>$10^{18}$</td>
<td>exa</td>
<td></td>
</tr>
<tr>
<td>$10^{21}$</td>
<td>zetta</td>
<td></td>
</tr>
<tr>
<td>$10^{24}$</td>
<td>yotta</td>
<td></td>
</tr>
<tr>
<td>$10^{27}$</td>
<td>xona</td>
<td></td>
</tr>
<tr>
<td>$10^{30}$</td>
<td>weka</td>
<td></td>
</tr>
<tr>
<td>$10^{33}$</td>
<td>vunda</td>
<td></td>
</tr>
<tr>
<td>$10^{36}$</td>
<td>uda</td>
<td></td>
</tr>
<tr>
<td>$10^{39}$</td>
<td>treda</td>
<td></td>
</tr>
<tr>
<td>$10^{42}$</td>
<td>sorta</td>
<td></td>
</tr>
<tr>
<td>$10^{45}$</td>
<td>rinta</td>
<td></td>
</tr>
<tr>
<td>$10^{48}$</td>
<td>quexa</td>
<td></td>
</tr>
<tr>
<td>$10^{51}$</td>
<td>pepta</td>
<td></td>
</tr>
<tr>
<td>$10^{54}$</td>
<td>ocha</td>
<td></td>
</tr>
<tr>
<td>$10^{57}$</td>
<td>nena</td>
<td></td>
</tr>
<tr>
<td>$10^{60}$</td>
<td>minga</td>
<td></td>
</tr>
<tr>
<td>$10^{63}$</td>
<td>luma</td>
<td></td>
</tr>
</tbody>
</table>