Current State of HPC in the United States

Horst Simon
Lawrence Berkeley National Laboratory
and UC Berkeley

Simulation Summit, Washington DC
October 12, 2010

Against the backdrop of the Japanese Earth Simulator (2002) the NRC report confirmed the importance of HPC for
- leadership in scientific research
- economic competitiveness
- national security

Implementation of the report recommendations and the HECRTF plan led to the state of HPC in the US today
Key Message: State of HPC in the US in 2010

• In 2010 the US is the undisputed world leader in HPC hardware, software, and applications

• However, transition from petascale to exascale will be characterized by significant and dramatic changes in HPC.

• This transition will be highly disruptive, and will create opportunities for other players to emerge and to lead.
Overview

• The Good: current status, hardware/systems, software, applications
• The Bad: missed opportunities
• The Ugly: future challenges
Performance Development

- 1 Gflop/s
- 1 Tflop/s
- 100 Mflop/s
- 100 Gflop/s
- 100 Tflop/s
- 10 Gflop/s
- 10 Tflop/s
- 1 Pflop/s
- 100 Pflop/s
- 1.75 PFlop/s
- 24.67 TFlop/s
- 32.4 PFlop/s

SUM
N=1
N=500

400 MFlop/s
59.7 GFlop/s
1.17 TFlop/s

Years:
- 1994
- 1996
- 1998
- 2000
- 2002
- 2004
- 2006
- 2008
- 2010
Vendors / System Share

- IBM: 198 (40%)
- HP: 185 (37%)
- Others: 50 (10%)
- Cray Inc.: 21 (4%)
- SGI: 17 (3%)
- Dell: 17 (3%)
- Sun: 12 (2%)

Jaguar: World’s most powerful computer since 2009

<table>
<thead>
<tr>
<th>Feature</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>Processors</td>
<td>224K</td>
</tr>
<tr>
<td>Power</td>
<td>6.95 MW</td>
</tr>
</tbody>
</table>

Source: T. Zacharia, ORNL
DOE provides extreme scale computing today: 15 years of world leadership

Top 500 list, November 2009

<table>
<thead>
<tr>
<th>Machine</th>
<th>Place</th>
<th>Speed (max)</th>
<th>On list Since</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaguar</td>
<td>ORNL</td>
<td>1.75 PF</td>
<td>2009 (1)</td>
</tr>
<tr>
<td>Roadrunner</td>
<td>LANL</td>
<td>1.04 PF</td>
<td>2009 (3)</td>
</tr>
<tr>
<td>Dawn</td>
<td>LLNL</td>
<td>0.478 PF</td>
<td>2007 (8)</td>
</tr>
<tr>
<td>BG/P</td>
<td>ANL</td>
<td>0.458 PF</td>
<td>2007 (9)</td>
</tr>
<tr>
<td>NERSC</td>
<td>LBL</td>
<td>0.266 PF</td>
<td>2008 (15)</td>
</tr>
<tr>
<td>Red Storm</td>
<td>SNL</td>
<td>0.204 PF</td>
<td>2009 (10)</td>
</tr>
</tbody>
</table>

INCITE: 2.5x oversubscribed

ASC and ASCR provide much more than machines:
- Applications (Computational Science)
- Algorithms (Applied Mathematics)
- Systems (Computer Science)
- Integration (SciDAC, Campaigns)

Source: DOE - SC
# Delivering the Software Foundation

## Software Developed under ASCR Funding

### Programming Models
- Active Harmony
- ARMCI
- ATLAS
- Berkeley UPC Compiler
- Charm++
- Fountain
- FT-MPI
- Global Arrays
- Kepler
- MVAPICH
- OPEN-MPI
- OpenUH
- PVM

### Development/Performance Tools
- BABEL
- Berkeley Lab Checkpoint Restart (BLCR)
- Dyninst API
- Fast Bit
- Goanna
- HPCToolkit
- Jumpshot
- KOJAK
- MPI
- MRNet
- Net PIPE
- OpenAnalysis
- PAPI
- ROSE
- ScalaTrace
- STAT
- TAO
- TAU
- Hpcviewer

### Math Libraries
- ACTS COLLECTION
- ADIC
- HYPRE
- ITAPS Software Suite
- LAPACK
- Mesquite
- MPICH2
- OpenAD
- OPT++
- PETSc
- ROMIO
- ScaLAPACK
- Sparskit-CCA
- Trilinos

### System Software
- Cluster Command & Control
- High-Availability OSCAR HA-OSCAR
- LWK-Sandia
- PVFS
- ZeptoOS

### Collaboration
- enote

### Visualization/Data Analytics
- BeSTMan
- Parallel netCDF
- Virtual Data Tool Kit

### Miscellaneous
- Libmonitor

---

**Source:** DOE - SC
International Exascale Software Project (IESP)

- Worldwide collaboration to develop system software for exascale systems
- Based on recognition of difficulty and cost of software development for unique high end systems
- Initiated by DOE-SC and NSF
- US leadership clear
SciDAC - First Federal Program to Implement CSE

- SciDAC (Scientific Discovery through Advanced Computing) program created in 2001, re-competed in 2006
  - about $50M annual funding
  - foundation for advances in computational science
  - a decade ahead

Global Climate

Biology  Nanoscience  Combustion  Astrophysics
Leadership Computing: Scientific Progress at the Petascale

**Turbulence**
Understanding the statistical geometry of turbulent dispersion of pollutants in the environment.

**Nuclear Energy**
High-fidelity predictive simulation tools for the design of next-generation nuclear reactors to safely increase operating margins.

**Energy Storage**
Understanding the storage and flow of energy in next-generation nanostructured carbon tube supercapacitors

**Fusion Energy**
Substantial progress in the understanding of anomalous electron energy loss in the National Spherical Torus Experiment (NSTX).

**Biofuels**
A comprehensive simulation model of lignocellulosic biomass to understand the bottleneck to sustainable and economical ethanol production.

**Nano Science**
Understanding the atomic and electronic properties of nanostructures in next-generation photovoltaic solar cell materials.

All known sustained petascale science applications to date have been run on DOE systems.
• Initiated in 2004
• Provides Office of Science leadership computing resources to a small number of computationally intensive research projects of large scale, that can make high-impact scientific advances through the use of a large allocation of computer time and data storage
• Open to national and international researchers, including industry
• No requirement of DOE Office of Science funding
• Peer-reviewed
• 1.6 billion hours awarded in 2010
# Investments in leadership computing are enabling translation of science to industrial solutions

| High-efficiency thermoelectric materials enabling substantial increases in fuel efficiency | • Atomistic determination of PbTe-AgSbTe₂ nanocomposites and growth mechanism explains low thermal conductivity  
• DFT predictions of Ag atom interstitial position confirmed by high-resolution TEM  
• GM: Using improved insight to develop new material |
|---|---|
| Retrofit parts for improved fuel efficiency and CO₂ emissions for Class 8 long haul trucks | • BMI: Simulations enable design of retrofit parts, reducing fuel consumption by up to 3,700 gal and CO₂ by up to 41 tons per truck per year  
• 10–17% improvement in fuel efficiency exceeds regulatory requirement of 5% for trucks operating in California |
| Development and correlation of computational tools for transport airplanes | • Boeing: Reduced validation time to transition newer technology (CFD) from research to airplane design and development  
• Demonstrated and improved correlations between CFD and wind tunnel test data |

Source: T. Zacharia, ORNL
DDES, IDDES Simulation of Landing Gear

• Discretionary, PI Philipe Spalart (Boeing)
  – Paper In BANC (Conference on Benchmark Problems for Airframe Noise Computation)
    • Tandem cylinders benchmark problem represents landing gear
      – 9M Intrepid Core-Hours
        • 60 million grid points
        • Overlapping grids
        • 8 racks on Intrepid

Source: Argonne Leadership Computing Facility
Overview

• The Good: current status, hardware/systems, software, applications

• The Bad: distraction and missed opportunity

• The Ugly: future challenges
Cloud Computing is a Data Center Operations Model! Not a technology!

• Premise: it is more cost- and energy efficient to run one large datacenter than many smaller ones
• How do you enable one large datacenter to appear to be a set of private resources (how to outsource IT)
  – IBM and HP on-demand centers: Image nodes on-demand and long-term contracts
  – Amazon EC2: Using VMs to do rapid on-the-fly provisioning of machines with short-term (hourly) cost model
  – Both cases: technology enables the business model
Productivity and Use of HPC in the Automotive Business

Good correlation between Rev/Empl. and Gflops/Model for US and European companies. Japanese companies are a kind of exception as profitability comes primarily from efficient manufacturing processes. Nissan plant is rated number #1 in the productivity index list (Automotive News, Dec 16, 2002).

- Ford, GM
- DCX, BMW, VW, Renault, PSA, Fiat, Saab
- Toyota, Honda

Source: C. Tanasescu, SGI
Barriers to Entry

• “Council on Competitiveness” report (2008) noted that there are three major barriers staling HPC adaption:
  – Lack of application software
  – Lack of sufficient talent
  – Cost

• These were the same constraints as noted in their 2004 report

• InterSect360 has a similar perspective:
  – “You could give companies free HW and SW and it would not solve these problems:
    • Political will to change a workflow and have confidence in simulation to supplement physical testing
    • Expertise and knowledge for using scalable systems, and
    • Creating digital models” (Addison Snell, InterSect360)

after Stephen Wheat, Intel
The Missing Middle

National Productivity Opportunity

“Missing Middle”

World Class/Leadership Computing

High-End HPC Users

Leading-Edge HPC Users

Source: Council on Competitiveness

© 2007 by the Council on Competitiveness and the University of Southern California

NUMBER OF USERS, APPLICATIONS

NUMBER OF PROCESSORS, MEMORY SIZE, JOB COMPLEXITY

Adapted from OSC Graphics
Overview

• The Good: current status, hardware/systems, software, applications

• The Bad: missed opportunities

• The Ugly: future challenges
## 35th List: The TOP10

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Manufacturer</th>
<th>Computer</th>
<th>Country</th>
<th>Cores</th>
<th>Rmax [Tflops]</th>
<th>Power [MW]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Oak Ridge National Laboratory</td>
<td>Cray</td>
<td>Jaguar Cray XT5 HC 2.6 GHz</td>
<td>USA</td>
<td>224,162</td>
<td>1,759</td>
<td>6.95</td>
</tr>
<tr>
<td>2</td>
<td>National Supercomputing Centre in Shenzhen</td>
<td>Dawning</td>
<td>Nebulae TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU</td>
<td>China</td>
<td>120,640</td>
<td>1,271</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>DOE/NNSA/LANL</td>
<td>IBM</td>
<td>Roadrunner BladeCenter QS22/LS21</td>
<td>USA</td>
<td>122,400</td>
<td>1,042</td>
<td>2.34</td>
</tr>
<tr>
<td>4</td>
<td>University of Tennessee</td>
<td>Cray</td>
<td>Kraken Cray XT5 HC 2.36GHz</td>
<td>USA</td>
<td>98,928</td>
<td>831.7</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Forschungszentrum Juelich (FZJ)</td>
<td>IBM</td>
<td>Jugene Blue Gene/P Solution</td>
<td>Germany</td>
<td>294,912</td>
<td>825.5</td>
<td>2.26</td>
</tr>
<tr>
<td>6</td>
<td>NASA/Ames Research Center/NAS</td>
<td>SGI</td>
<td>Pleiades SGI Altix ICE 8200EX</td>
<td>USA</td>
<td>56,320</td>
<td>544.3</td>
<td>3.1</td>
</tr>
<tr>
<td>7</td>
<td>National SuperComputer Center</td>
<td>NUDT</td>
<td>Tianhe-1 NUDT TH-1 Cluster, Xeon, ATI Radeon, Infiniband</td>
<td>China</td>
<td>71,680</td>
<td>563.1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>DOE/NNSA/LLNL</td>
<td>IBM</td>
<td>BlueGene/L eServer Blue Gene Solution</td>
<td>USA</td>
<td>212,992</td>
<td>478.2</td>
<td>2.32</td>
</tr>
<tr>
<td>9</td>
<td>Argonne National Laboratory</td>
<td>IBM</td>
<td>Intrepid Blue Gene/P Solution</td>
<td>USA</td>
<td>163,840</td>
<td>458.6</td>
<td>1.26</td>
</tr>
<tr>
<td>10</td>
<td>Sandia/NREL</td>
<td>Sun</td>
<td>Red Sky SunBlade x6275</td>
<td>USA</td>
<td>42,440</td>
<td>433.5</td>
<td></td>
</tr>
</tbody>
</table>
Growth of Chinese Investment in HPC

Sum of the performance of all systems in a country on the TOP500 list measured in LINPACK Rmax

Earth Simulator effect

Source: TOP500 list
www.top500.org
Growth of Chinese Investment in HPC

• China has made consistent investments over the last 10 years which led to consistent growth of its position in HPC
  ▪ Different from the Japanese Earth Simulator in 2002, which was a “Black Swan”, a one time occurrence
• China is now (6/1/2010) the #2 country on the list, ahead of Japan (since Nov. 2009) in terms of installed base of supercomputers
• China installed two significant TOP10 systems in the last year:
  ▪ Nebulae, which is located at the newly build National Supercomputing Centre in Shenzhen, China, achieved 1.271 PFlop/s running the Linpack benchmark, which puts it in the No. 2 spot on the TOP500 behind ORNL’s Jaguar and ahead of LANL’s Roadrunner. In part due to its NVidia GPU accelerators, Nebulae reports an impressive theoretical peak capability of almost 3 petaflop/s – the highest ever on the TOP500.
  ▪ Tianhe-1 (meaning River in Sky), installed at the National Super Computer Center in Tianjin, China is a second Chinese system in the TOP10 and ranked at No. 7. Tianhe-1 and Nebulae are both hybrid designs with Intel Xeon processors and AMD or NVidia GPUs used as accelerators. Each node of Tianhe-1 consists of two AMD GPUs attached to two Intel Xeon processors.

Source: TOP500 list
www.top500.org
Growth of Chinese Investment in HPC

- Dawning in collaboration with ICT/CAS (Inst. of Computing Technology at the Chinese Academy of Science) is also pursuing its own line of processors based on the Godson-3 (commercially called Loongson) design. Dawning is expected to deliver a Petaflops system based on this processor later in 2010.
- Godson-3 is based on the MIPS core and emulates x86. It will dissipate only 5-10 W at 1 GHz.
- China is thus the only country outside the US that pursues its own processor technology to be used in HPC.


Figure 4. A massively parallel implementation of the Godson-3 could populate the on-chip mesh network with 16 or more quad-core clusters. This figure is based on a slide shown at the Hot Chips Symposium by one of the Godson-3 designers, so it’s not a far-fetched speculation. An implementation this large and powerful is probably intended for a future Chinese supercomputer.
Japan: Next Generation Supercomputer Project
NGSC

Brief history
• 2004-5: Blueprint Research
• 2005: NGSC as one of national key technologies
• 2006: Riken started the project
  Target: 10 Petaflops computing
  the sustained speed is over 1 petaflops
  (aimed at No.1 of HPC in the world SC record at that time)
  Major apps: nano, bio science & technology, new energy
  + for climate change, education, industries, • • •
  multi-physics, multi-disciplinary simulation

Budget: approx. 1.5 B$ for 2006-2012
Site: Kobe, Japan.

Source: H. Nakamura, RIST
KOBE site under construction

Image of completion

Construction

Machine room space

Source: H. Nakamura, RIST
Prototypes for Petaflop/s systems in 2009/2010

- IBM BlueGene/P (FZJ) 01-2008
- IBM Power6 (SARA) 07-2008
- Cray XT5 (CSC) 11-2008
- NEC SX9, vector part (HLRS) 02-2009
- Intel Nehalem/Xeon (CEA/FZJ) 06-2009
- IBM Cell/Power (BSC) 12-2008
PRACE – A Partnership with a Vision

- Provide **world-class** HPC systems for word-class science
- Support Europe in attaining **global leadership** in public and private research and development

... and a Mission

- Create a world-leading persistent high-end HPC infrastructure
  - Deploy 3 – 5 systems of the highest performance level (tier-0)
  - First European Petaflop/s system deployed in Jülich in 2009
  - Ensure a diversity of architectures to meet the needs of European user communities
  - Provide support and training
Why 2010 is different

• There always has been international competition in HPC

• 2010 is different from the past:
  – significant technology disruption
  – there is (not yet) a comprehensive plan to address this transition
  – different climate of global economic competitiveness
  – opportunity for others to leapfrog
Traditional Sources of Performance Improvement are Flat-Lining (2004)

- **New Constraints**
  - 15 years of *exponential* clock rate growth has ended

- **Moore’s Law reinterpreted:**
  - How do we use all of those transistors to keep performance increasing at historical rates?
  - Industry Response: #cores per chip doubles every 18 months *instead* of clock frequency!

*Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith*
Concurrent Levels

# processors

Minimum: 1
Average: 10
Maximum: 1,000

Jun-93 to Jun-15

Jack's Notebook
Multicore comes in a wide variety

- Multiple parallel general-purpose processors (GPPs)
- Multiple application-specific processors (ASPs)

---

**Intel 4004 (1971):**
- 4-bit processor,
- 2312 transistors,
- ~100 KIPS,
- 10 micron PMOS,
- 11 mm² chip

---

**Sun Niagara**
- 8 GPP cores (32 threads)

---

**Intel Network Processor**
- 1 GPP Core
- 16 ASPs (128 threads)

---

**IBM Cell**
- 1 GPP (2 threads)
- 8 ASPs

---

**Cisco CRS-1**
- 188 Tensilica GPPs

---

**Picochip DSP**
- 1 GPP core
- 248 ASPs

---

“The Processor is the new Transistor”

[Rowen]
What’s Next?

Source: Jack Dongarra, ISC 2008
Current Technology Roadmaps will Depart from Historical Gains

Power is the Leading Design Constraint

From Peter Kogge, DARPA Exascale Study
… and the power costs will still be staggering

$1M per megawatt per year! (with CHEAP power)

From Peter Kogge,
DARPA Exascale Study
A decadal DOE plan for providing exascale applications and technologies for DOE mission needs

Rick Stevens and Andy White, co-chairs
Pete Beckman, Ray Bair-ANL; Jim Hack, Jeff Nichols, Al Geist-ORNL; Horst Simon, Kathy Yelick, John Shalf-LBNL; Steve Ashby, Moe Khaleel-PNNL; Michel McCoy, Mark Seager, Brent Gorda-LLNL; John Morrison, Cheryl Wampler-LANL; James Peery, Sudip Dosanjh, Jim Ang-SNL; Jim Davenport, Tom Schlagel, BNL; Fred Johnson, Paul Messina, ex officio
Broad Community Support to Address the Challenge

- **Town Hall Meetings April-June 2007**
- **Scientific Grand Challenges Workshops Nov, 2008 – Oct, 2009**
  - Climate Science (11/08),
  - High Energy Physics (12/08),
  - Nuclear Physics (1/09),
  - Fusion Energy (3/09),
  - Nuclear Energy (5/09),
  - Biology (8/09),
  - Material Science and Chemistry (8/09),
  - National Security (10/09)
  - Cross-cutting technologies (2/10)
- **Exascale Steering Committee**
  - “Denver” vendor NDA visits 8/2009
  - SC09 vendor feedback meetings
  - Extreme Architecture and Technology Workshop 12/2009
- **International Exascale Software Project**
What are critical exascale technology investments?

- **System power** is a first class constraint on exascale system performance and effectiveness.
- **Memory** is an important component of meeting exascale power and applications goals.
- **Programming model.** Early investment in several efforts to decide in 2013 on exascale programming model, allowing exemplar applications effective access to 2015 system for both mission and science.
- **Investment in exascale processor design** to achieve an exascale-like system in 2015.
- **Operating System strategy for exascale** is critical for node performance at scale and for efficient support of new programming models and run time systems.
- **Reliability and resiliency are critical at this** scale and require applications neutral movement of the file system (for check pointing, in particular) closer to the running apps.
- **HPC co-design strategy and implementation** requires a set of hierarchical performance models and simulators as well as commitment from apps, software and architecture communities.
A Revolution is Underway

• Rapidly Changing Technology Landscape
  – Evolutionary change between nodes (*10x more explicit parallelism*)
  – Revolutionary change within node (*100x more parallelism*, with diminished memory capacity and bandwidth)
  – Multiple Technology Paths (GPU, manycore/embedded, x86/PowerX)

• The technology disruption will be pervasive (*not just exascale*)
  – *Assumptions that our current software infrastructure is built upon are no longer valid*
  – Applications, Algorithms, System Software *will all break*
  – As significant as migration from vector to MPP (early 90’s)

• Need a new approach to ensuring continued application performance improvements
  – This isn’t just about Exaflops – this is for all system scales
Technology Paths to Exascale

• Leading Technology Paths (Swim Lanes)
  – Multicore: *Maintain complex cores, and replicate* (x86 and Power7, Blue Waters, NGSC)
  – Manycore/Embedded: *Use many simpler, low power cores from embedded* (BlueGene,Dawning)
  – GPU/Accelerator: *Use highly specialized processors from gaming/graphics market space* (NVidia Fermi, Cell, Intel Knights Corner/Larrabee)

• Risks in Swim Lane selection
  – Select too soon: *Users cannot follow*
  – Select too late: *Fall behind performance curve*
  – Select incorrectly: *Subject users to multiple disruptive technology changes*
Navigating Technology Phase Transitions

- Franklin (N5): 19 TF Sustained, 101 TF Peak
- Hopper (N6): >1 PF Peak
- Franklin (N5) + QC: 36 TF Sustained, 352 TF Peak
- COTS/MPP + MPI
- COTS/MPP + MPI (+ OpenMP)
- NERSC-7: 10 PF Peak
- NERSC-8: 100 PF Peak
- NERSC-9: 1 EF Peak
- Exascale + ???

GPU CUDA/OpenCL
Or Manycore BG/Q, R

Top500

Peak Teraflop/s

Summary

- In 2010 HPC in the US is a strong position of world leadership
- There are missed opportunities, in particular with respect to a broader application of HPC in industry
- Major Challenges are ahead for HPC
  - Technology transitions
  - International competition
- We are at a critical juncture: the right decisions in 2011-2013 can assure US leadership in HPC for another generation
Shackleton’s Quote on Exascale

Ernest Shackleton’s 1907 ad in London’s Times, recruiting a crew to sail with him on his exploration of the South Pole