CS 594 - 005
Scientific Computing for Engineers
HPC Start to Finish

Jack Dongarra
University of Tennessee

Simulation: The Third Pillar of Science

◆ Traditional scientific and engineering paradigm:
  1) Do theory or paper design.
  2) Perform experiments or build system.

◆ Limitations:
  Ø Too difficult -- build large wind tunnels.
  Ø Too expensive -- build a throw-away passenger jet.
  Ø Too slow -- wait for climate or galactic evolution.
  Ø Too dangerous -- weapons, drug design, climate experimentation.

◆ Computational science paradigm:
  3) Use high performance computer systems to simulate the phenomenon
     » Base on known physical laws and efficient numerical methods.
Computational Science

Applications to Energy

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

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

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

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

Smart Truck
Aerodynamic forces account for ~53% of long haul truck fuel use. ORNL's Jaguar predicted 12% drag reduction and yielded EPA-certified 6.9% increase in fuel efficiency.

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

Source: Steven E. Koonin, DOE
Computational Science Fuses Three Distinct Elements:

- Application Science/Engineering
- Computer Science
- Mathematics

Interconnected with:
- Compute-intensive applications
- Data-intensive applications
Computational Science As An Emerging Academic Pursuit

- **Many Programs in Computational Science**
  - College for Computing
    - Georgia Tech; NJIT; CMU; ...
  - Degrees
    - Rice, Utah, UCSB; ...
  - Minor
    - Penn State, U Wisc, SUNY Brockport
  - Certificate
    - Old Dominion, U of Georgia, Boston U, ...
  - Concentration
    - Cornell, Northeastern, Colorado State, ...
  - Courses
At the University of Tennessee

- A few years ago there was a discussion to create a program in Computational Science.
- This program evolved out of a set of meetings and discussions with faculty, students, and administration.
- Modeled on a similar minor degree program in the Statistics Department on campus.
- Appeared in the 2007-2008 Graduate Catalog.
Graduate Minor in Computational Science

- Students in one of the three general areas of Computational Science;
  - Applied Mathematics,
  - Computer related, or
  - a Domain Science

will become exposed to and better versed in the other two areas that are currently outside their “home” area.

- A pool of courses which deals with each of the three main areas has been put together by participating department for students to select from.

- Interdisciplinary Graduate Minor in Computational Science (IGMCS)
IGMCS: Requirements

- The Minor requires a combination of course work from three disciplines - Computer related, Mathematics/Stat, and a participating Science/Engineering domain (e.g., Chemical Engineering, Chemistry, Physics).

- At the Masters level a minor in Computational Science will require 9 hours (3 courses) from the pools.
  - At least 6 hours (2 courses) must be taken outside the student’s home area.
  - Students must take at least 3 hours (1 course) from each of the 2 non-home areas

- At the Doctoral level a minor in computation science will require 15 hours (5 courses) from the pools.
  - At least 9 hours (3 courses) must be taken outside the student’s home area.
  - Students must take at least 3 hours (1 course) from each of the 2 non-home areas
IGMCS Process for Students

1. A student, with guidance from their faculty advisor, lays out a program of courses
2. Next, discussion with department’s IGMCS liaison
3. Form generated with courses to be taken
4. Form is submitted for approval by the IGMCS Program Committee
# IGMCS Participating Departments

<table>
<thead>
<tr>
<th>Department</th>
<th>IGMCS Liaison</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biochem &amp; Cell and Mole Bio</td>
<td>Dr. Albrecht von Arnim</td>
<td><a href="mailto:vonarnim@utk.edu">vonarnim@utk.edu</a></td>
</tr>
<tr>
<td>Chemical Engineering</td>
<td>Dr. Steven Abel</td>
<td><a href="mailto:abel@utk.edu">abel@utk.edu</a></td>
</tr>
<tr>
<td>Chemistry</td>
<td>Dr. Robert Hinde</td>
<td><a href="mailto:rhinde@utk.edu">rhinde@utk.edu</a></td>
</tr>
<tr>
<td>Civil &amp; Envir. Eng.</td>
<td>Dr. Joshua Fu</td>
<td><a href="mailto:jsfu@utk.edu">jsfu@utk.edu</a></td>
</tr>
<tr>
<td>Earth &amp; Planetary Sci</td>
<td>Dr. Edmund Perfect</td>
<td><a href="mailto:eperfect@utk.edu">eperfect@utk.edu</a></td>
</tr>
<tr>
<td>Ecology &amp; Evol. Biology</td>
<td>Dr. Paul Armsworth</td>
<td><a href="mailto:p.armsworth@utk.edu">p.armsworth@utk.edu</a></td>
</tr>
<tr>
<td>EECS</td>
<td>Dr. Jack Dongarra</td>
<td><a href="mailto:dongarra@eecs.utk.edu">dongarra@eecs.utk.edu</a></td>
</tr>
<tr>
<td>Genome Sci &amp; Tech</td>
<td>Dr. Albrecht von Arnim</td>
<td><a href="mailto:vonarnim@utk.edu">vonarnim@utk.edu</a></td>
</tr>
<tr>
<td>Geography</td>
<td>Dr. Bruce Ralston</td>
<td><a href="mailto:bralston@utk.edu">bralston@utk.edu</a></td>
</tr>
<tr>
<td>Information Science</td>
<td>Dr. Devendra Potnis</td>
<td><a href="mailto:dpotnis@utk.edu">dpotnis@utk.edu</a></td>
</tr>
<tr>
<td>Material Science &amp; Eng</td>
<td>Dr. David Keffer</td>
<td><a href="mailto:dkeffer@utk.edu">dkeffer@utk.edu</a></td>
</tr>
<tr>
<td>Mathematics</td>
<td>Dr. Vasilios Alexiades</td>
<td><a href="mailto:alexiades@utk.edu">alexiades@utk.edu</a></td>
</tr>
<tr>
<td>Mech, Aero &amp; Biomed Eng</td>
<td>Dr. Kivanc Ekici</td>
<td><a href="mailto:ekici@utk.edu">ekici@utk.edu</a></td>
</tr>
<tr>
<td>Microbiology</td>
<td>Dr. Gary Sayler</td>
<td><a href="mailto:sayler@utk.edu">sayler@utk.edu</a></td>
</tr>
<tr>
<td>Nuclear Engineering</td>
<td>Dr. Ivan Maldonado</td>
<td><a href="mailto:ivan.maldonado@utk.edu">ivan.maldonado@utk.edu</a></td>
</tr>
<tr>
<td>Physics</td>
<td>Dr. Thomas Papenbrock</td>
<td><a href="mailto:tpapenbr@utk.edu">tpapenbr@utk.edu</a></td>
</tr>
<tr>
<td>Statistics</td>
<td>Dr. Hamparsum Bozdogan</td>
<td><a href="mailto:bozdogan@utk.edu">bozdogan@utk.edu</a></td>
</tr>
</tbody>
</table>
Students in Departments Not Participating in the IGMCS Program

- A student in such a situation can still participate.
  - Student and advisor should submit to the Chair of the IGMCS Program Committee the courses to be taken.
  - Requirement is still the same:
    - Minor requires a combination of course work from three disciplines - Computer Science related, Mathematics/Stat, and a participating Science/Engineering domain (e.g., Chemical Engineering, Chemistry, Physics).
- Student’s department should be encouraged to participate in the IGMCS program.
  - Easy to do, needs approved set of courses and a liaison
Internship

- This is optional but strongly encouraged.
- Students in the program can fulfill 3 hrs. of their requirement through an Internship with researchers outside the student’s major.
- The internship may be taken offsite, e.g. ORNL, or on campus by working with a faculty member in another department.
- Internships must have the approval of the IGMCS Program Committee.
COMPUTATIONAL SCIENCE: INHERENTLY INTERDISCIPLINARY
Why Turn to Simulation?

◆ When the problem is too . . .
  ➢ Complex
  ➢ Large / small
  ➢ Expensive
  ➢ Dangerous
◆ to do any other way.
Why Turn to Simulation?

- Climate / Weather Modeling
- Data intensive problems (data-mining, oil reservoir simulation)
- Problems with large length and time scales (cosmology)
Data Driven Science

• Scientific data sets are growing exponentially
  – Ability to generate data is exceeding our ability to store and analyze
  – Simulation systems and some observational devices grow in capability with Moore’s Law

• Petabyte (PB) data sets will soon be common:
  - Climate modeling: estimates of the next IPCC data is in 10s of petabytes
  - Genome: JGI alone will have .5 petabyte of data this year and double each year
  - Particle physics: LHC is projected to produce 16 petabytes of data per year
  - Astrophysics: LSST and others will produce 5 petabytes/year (via 3.2 Gigapixel camera)

• Create scientific communities with “Science Gateways” to data
Weather and Economic Loss

- **$10T U.S. economy**
  - 40% is adversely affected by weather and climate

- **$1M in loss to evacuate each mile of coastline**
  - we now over warn by 3X!
  - average over warning is 200 miles, or $200M per event

- **Improved forecasts**
  - lives saved and reduced cost

- **LEAD**
  - Linked Environments for Atmospheric Discovery
    - Oklahoma, Indiana, UCAR, Colorado State, Howard, Alabama, Millersville, NCSA, North Carolina

Source: Kelvin Droegemeier, Oklahoma
Look at the Fastest Computers

- Strategic importance of supercomputing
  - Essential for scientific discovery
  - Critical for national security
  - Fundamental contributor to the economy and competitiveness through use in engineering and manufacturing

- Supercomputers are the tool for solving the most challenging problems through simulations
Units of Measure

- High Performance Computing (HPC) units are:
  - Flop: floating point operation, usually double precision unless noted
  - Flop/s: floating point operations per second
  - Bytes: size of data (a double precision floating point number is 8)

- Typical sizes are millions, billions, trillions...

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mega</td>
<td>Mflop/s</td>
<td>(10^6) flop/sec</td>
</tr>
<tr>
<td>Giga</td>
<td>Gflop/s</td>
<td>(10^9) flop/sec</td>
</tr>
<tr>
<td>Tera</td>
<td>Tflop/s</td>
<td>(10^{12}) flop/sec</td>
</tr>
<tr>
<td>Peta</td>
<td>Pflop/s</td>
<td>(10^{15}) flop/sec</td>
</tr>
<tr>
<td>Exa</td>
<td>Eflop/s</td>
<td>(10^{18}) flop/sec</td>
</tr>
<tr>
<td>Zetta</td>
<td>Zflop/s</td>
<td>(10^{21}) flop/sec</td>
</tr>
<tr>
<td>Yotta</td>
<td>Yflop/s</td>
<td>(10^{24}) flop/sec</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mbyte</td>
<td>(2^{20} = 1048576 \sim 10^6) bytes</td>
</tr>
<tr>
<td></td>
<td>Gbyte</td>
<td>(2^{30} \sim 10^9) bytes</td>
</tr>
<tr>
<td></td>
<td>Tbyte</td>
<td>(2^{40} \sim 10^{12}) bytes</td>
</tr>
<tr>
<td></td>
<td>Pbyte</td>
<td>(2^{50} \sim 10^{15}) bytes</td>
</tr>
<tr>
<td></td>
<td>Ebyte</td>
<td>(2^{60} \sim 10^{18}) bytes</td>
</tr>
<tr>
<td></td>
<td>Zbyte</td>
<td>(2^{70} \sim 10^{21}) bytes</td>
</tr>
<tr>
<td></td>
<td>Ybyte</td>
<td>(2^{80} \sim 10^{24}) bytes</td>
</tr>
</tbody>
</table>

- Current fastest (public) machine \(\sim 33.9\) Pflop/s
  - Up-to-date list at www.top500.org
In the past decade, the world has experienced one of the most exciting periods in computer development. Microprocessors have become smaller, denser, and more powerful. The result is that microprocessor-based supercomputing is rapidly becoming the technology of preference in attacking some of the most important problems of science and engineering.
2X transistors/Chip Every 1.5 years
Called “Moore’s Law”
Moore’s Secret Sauce: Dennard Scaling

Moore’s Law put lots more transistors on a chip…but it’s Dennard’s Law that made them useful

Dennard Scaling:
• Decrease feature size by a factor of \( \lambda \) and decrease voltage by a factor of \( \lambda \); then
  • # transistors increase by \( \lambda^2 \)
  • Clock speed increases by \( \lambda \)
  • Energy consumption does not change

Dennard observed that voltage and current should be proportional to the linear dimensions of a transistor

2x transistor count
40% faster
50% more efficient

Design of Ion-Implanted MOSFET’s with Very Small Physical Dimensions

ROBERT H. DENNARD, MEMBER, IEEE, FRITZ H. GAENSSLEN, KWA-NIEN YU, MEMBER, IEEE, V. LEO RIDEOUT, MEMBER, IEEE, ERNEST BASSOUS, AND ANDRÉ H. LEBLANC, MEMBER, IEEE

Abstract—This paper considers the design, fabrication, and characterization of ion-implanted MOSFET’s, which have been used for digital integrated circuits using dimensions in the order of \( \lambda \). Scaling relationships are presented which show how a conventional MOSFET can be reduced in size. An improved small device structure is presented that uses ion implantation to provide a shallow source and drain region and a continuous substrate doping profile. One-dimensional models are used to predict the substrate doping profile and the corresponding threshold voltage versus source voltage characteristics. A two-dimensional current transport model is used to calculate the subthreshold and cutoff characteristics for different device parameter combinations. Polysilicon-gate MOSFET’s with channel lengths as short as 6.5 \( \mu \) are fabricated, and the device characteristics measured and compared with predicted values. The performance improvement expected from using these very small devices in highly miniaturized integrated circuits is proposed.

Manuscript received May 30, 1974; revised July 5, 1974.

The authors are with the IBM T. J. Watson Research Center, Yorktown Heights, N.Y. 10598.

List of Symbols

- Inverse semilogarithmic slope of subthreshold characteristics.
- Width of idealized step function profile for channel implant.
- Work function difference between gate and substrate.
- Effective channel mobility.
- Substrate acceptor concentration.
- Drain current.
- Boltzmann’s constant.
- Unitless scaling constant.
- MOSFET channel length.
- Effective surface mobility.
- Intrinsic carrier concentration.
- Substrate acceptor concentration.
- Bond bending in silicon at the onset of strong inversion for non-substrate voltage.

[Denard, Gaensslen, Yu, Rideout, Bassous, Leblanc, IEEE JSSC, 1974]
Unfortunately Dennard Scaling is Over: What is the Catch?

Breakdown is the result of small feature sizes, current leakage poses greater challenges, and also causes the chip to heat up.

Powering the transistors without melting the chip.
The primary reason cited for the breakdown is that at small sizes, current leakage poses greater challenges, and also causes the chip to heat up, which creates a threat of thermal runaway and therefore further increases energy costs.
Moore’s Law is Alive and Well

Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanović
Slide from Kathy Yelick
Transistors into Cores
Today’s Multicores
99% of Top500 Systems Are Based on Multicore

- Sun Niagara2 (8 cores)
- Intel Xeon Phi (60 cores)
- Fujitsu Venus (16 cores)
- Nvidia Kepler (2688 Cuda cores)
- IBM Power 7 (8 cores)
- AMD Interlagos (16 cores)
- IBM BG/Q (18 cores)
- Intel Westmere (10 cores)
But Clock Frequency Scaling Replaced by Scaling Cores / Chip

15 Years of exponential growth ~2x year has ended

Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanović
Slide from Kathy Yelick
Performance Has Also Slowed, Along with Power

Power is the root cause of all this

A hardware issue just became a software problem

Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanović

Slide from Kathy Yelick
# Power Cost of Frequency

- **Power** $\propto \text{Voltage}^2 \times \text{Frequency}$ \hspace{1cm} ($V^2F$)
- **Frequency** $\propto \text{Voltage}$
- **Power** $\propto \text{Frequency}^3$

<table>
<thead>
<tr>
<th></th>
<th>Cores</th>
<th>V</th>
<th>Freq</th>
<th>Perf</th>
<th>Power</th>
<th>PE (Bops/watt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superscalar</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>&quot;New&quot; Superscalar</td>
<td>1X</td>
<td>1.5X</td>
<td>1.5X</td>
<td>1.5X</td>
<td>3.3X</td>
<td>0.45X</td>
</tr>
</tbody>
</table>
## Power Cost of Frequency

- **Power** $\propto$ **Voltage$^2$ x Frequency** \((V^2F)\)

- **Frequency** $\propto$ **Voltage**

- **Power** $\propto$ **Frequency$^3$$\)**

<table>
<thead>
<tr>
<th></th>
<th>Cores</th>
<th>V</th>
<th>Freq</th>
<th>Perf</th>
<th>Power</th>
<th>PE (Bops/watt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superscalar</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>&quot;New&quot; Superscalar</td>
<td>1X</td>
<td>1.5X</td>
<td>1.5X</td>
<td>1.5X</td>
<td>3.3X</td>
<td>0.45X</td>
</tr>
<tr>
<td>Multicore</td>
<td>2X</td>
<td>0.75X</td>
<td>0.75X</td>
<td>1.5X</td>
<td>0.8X</td>
<td>1.88X</td>
</tr>
</tbody>
</table>

(Bigger number is better)

50% more performance with 20% less power

Preferable to use multiple slower devices, than one superfast device
Look at the Fastest Computers

- **Strategic importance of supercomputing**
  - Essential for scientific discovery
  - Critical for national security
  - Fundamental contributor to the economy and competitiveness through use in engineering and manufacturing

- **Supercomputers are the tool for solving the most challenging problems through simulations**
Example of typical parallel machine
Example of typical parallel machine
Example of typical parallel machine

Shared memory programming between processes on a board and a combination of shared memory and distributed memory programming between nodes and cabinets
Example of typical parallel machine

Combination of shared memory and distributed memory programming
What do you mean by performance?

◆ What is a xflop/s?
  ➢ xflop/s is a rate of execution, some number of floating point operations per second.
    » Whenever this term is used it will refer to 64 bit floating point operations and the operations will be either addition or multiplication.

◆ What is the theoretical peak performance?
  ➢ The theoretical peak is based not on an actual performance from a benchmark run, but on a paper computation to determine the theoretical peak rate of execution of floating point operations for the machine.
  ➢ The theoretical peak performance is determined by counting the number of floating-point additions and multiplications (in full precision) that can be completed during a period of time, usually the cycle time of the machine.
  ➢ For example, an Intel Xeon 5570 quad core at 2.93 GHz can complete 4 floating point operations per cycle or a theoretical peak performance of 11.72 GFlop/s per core or 46.88 Gflop/s for the socket.
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 Germany in June
- All data available from www.top500.org
Performance Development of HPC over the Last 24 Years from the Top500

- 59.7 GFlop/s
- 400 MFlop/s
- 1.17 TFlop/s
- 1.17 TFlop/s
- SUM
- N=1
- N=500
- 59.7 GFlop/s
- 400 MFlop/s
- My Laptop 70 Gflop/s
- My iPhone 4 Gflop/s
- 420 PFlop/s
- 33.9 PFlop/s
- 206 TFlop/s

State of Supercomputing in 2016

• Pflops (> $10^{15}$ Flop/s) computing fully established with 81 systems.

• Three technology architecture possibilities or “swim lanes” are thriving.
  • Commodity (e.g. Intel)
  • Commodity + accelerator (e.g. GPUs) (104 systems)
  • Special purpose lightweight cores (e.g. IBM BG, ARM, Intel’s Knights Landing)

• Interest in supercomputing is now worldwide, and growing in many new markets (around 50% of Top500 computers are used in industry).

• Exascale ($10^{18}$ Flop/s) projects exist in many countries and regions.

• Intel processors largest share, 89% followed by AMD, 4%.
Countries Share

<table>
<thead>
<tr>
<th>Country</th>
<th>Absolute Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>201</td>
</tr>
<tr>
<td>China</td>
<td>109</td>
</tr>
<tr>
<td>Japan</td>
<td>38</td>
</tr>
<tr>
<td>UK</td>
<td>18</td>
</tr>
<tr>
<td>France</td>
<td>18</td>
</tr>
<tr>
<td>Germany</td>
<td>32</td>
</tr>
</tbody>
</table>

China nearly tripled the number of systems on the latest list, while the number of systems in the US has fallen to the lowest point since the TOP500 list was created.
Performance of Countries

Total Performance [Tflop/s]

US
Performance of Countries

Total Performance [Tflop/s]

- US
- EU

- 2000
- 2002
- 2004
- 2006
- 2008
- 2010
- 2012
- 2014
## November 2015: The TOP 10 Systems

<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 Power [MW]</th>
<th>MFlops/Watt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>National Super Computer Center in Guangzhou</td>
<td>Tianhe-2 NUDT, Xeon 12C + Intel Xeon Phi (57c) + Custom</td>
<td>China</td>
<td>3,120,000</td>
<td>33.9</td>
<td>62</td>
<td>17.8</td>
</tr>
<tr>
<td>2</td>
<td>DOE / OS Oak Ridge Nat Lab</td>
<td>Titan, Cray XK7, AMD (16C) + Nvidia Kepler GPU (14c) + Custom</td>
<td>USA</td>
<td>560,640</td>
<td>17.6</td>
<td>65</td>
<td>8.3</td>
</tr>
<tr>
<td>3</td>
<td>DOE / NNSA L Livermore Nat Lab</td>
<td>Sequoia, BlueGene/Q (16c) + custom</td>
<td>USA</td>
<td>1,572,864</td>
<td>17.2</td>
<td>85</td>
<td>7.9</td>
</tr>
<tr>
<td>4</td>
<td>RIKEN Advanced Inst for Comp Sci</td>
<td>K computer Fujitsu SPARC64 VIIIfx (8c) + Custom</td>
<td>Japan</td>
<td>705,024</td>
<td>10.5</td>
<td>93</td>
<td>12.7</td>
</tr>
<tr>
<td>5</td>
<td>DOE / OS Argonne Nat Lab</td>
<td>Mira, BlueGene/Q (16c) + Custom</td>
<td>USA</td>
<td>786,432</td>
<td>8.16</td>
<td>85</td>
<td>3.95</td>
</tr>
<tr>
<td>6</td>
<td>DOE / NNSA / Los Alamos &amp; Sandia</td>
<td>Trinity, Cray XC40, Xeon 16C + Custom</td>
<td>USA</td>
<td>301,056</td>
<td>8.10</td>
<td>80</td>
<td>2.3</td>
</tr>
<tr>
<td>7</td>
<td>Swiss CSCS</td>
<td>Piz Daint, Cray XC30, Xeon 8C + Nvidia Kepler (14c) + Custom</td>
<td>Switzerland</td>
<td>115,984</td>
<td>6.27</td>
<td>81</td>
<td>2.3</td>
</tr>
<tr>
<td>8</td>
<td>HLRS Stuttgart</td>
<td>Hazel Hen, Cray XC40, Xeon 12C + Custom</td>
<td>Germany</td>
<td>185,088</td>
<td>5.64</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>KAUST</td>
<td>Shaheen II, Cray XC40, Xeon 16C + Custom</td>
<td>Saudi Arabia</td>
<td>196,608</td>
<td>5.54</td>
<td>77</td>
<td>2.8</td>
</tr>
<tr>
<td>10</td>
<td>Texas Advanced Computing Center</td>
<td>Stampede, Dell Intel (8c) + Intel Xeon Phi (61c) + IB</td>
<td>USA</td>
<td>204,900</td>
<td>5.17</td>
<td>61</td>
<td>4.5</td>
</tr>
</tbody>
</table>

500 (368) Karlsruher MEGAWARE Intel Germany 10,800 .206 95
Comparison

1 Gflop/s  
70 Gflop/s  
33,000,000 Gflop/s

~factor of 500,000
China, 2013: **the 34 PetaFLOPS**

Developed in cooperation between NUDT and Inspur for National Supercomputer Center in Guangzhou

**Peak performance of 54.9 PFLOPS**

- 16,000 nodes contain 32,000 Xeon Ivy Bridge processors and 48,000 Xeon Phi accelerators totaling 3,120,000 cores
- 162 cabinets in 720m² footprint
- Total 1.404 PB memory (88GB per node)
- Each Xeon Phi board utilizes 57 cores for aggregate 1.003 TFLOPS at 1.1GHz clock
- Proprietary TH Express-2 interconnect (fat tree with thirteen 576-port switches)
- 12.4 PB parallel storage system
- 17.6MW power consumption under load; 24MW including (water) cooling
- 4096 SPARC V9 based Galaxy FT-1500 processors in front-end system
ORNL’s “Titan” Hybrid System: Cray XK7 with AMD Opteron and NVIDIA Tesla processors

SYSTEM SPECIFICATIONS:
- Peak performance of 27 PF
  - 24.5 Pflop/s GPU + 2.6 Pflop/s AMD
- 18,688 Compute Nodes each with:
  - 16-Core AMD Opteron CPU
  - NVIDIA Tesla “K20x” GPU
  - 32 + 6 GB memory
- 512 Service and I/O nodes
- 200 Cabinets
- 710 TB total system memory
- Cray Gemini 3D Torus Interconnect
- 9 MW peak power

4,352 ft²
404 m²
Cray XK7 Compute Node

**XK7 Compute Node Characteristics**

- **AMD Opteron 6274 Interlagos**
  - 16 core processor
- **Tesla K20x @ 1311 GF**
- **Host Memory**
  - 32GB
  - 1600 MHz DDR3
- **Tesla K20x Memory**
  - 6GB GDDR5
- **Gemini High Speed Interconnect**

Slide courtesy of Cray, Inc.
Titan: Cray XK7 System

System:
- 200 Cabinets
- 18,688 Nodes
- 27 PF
- 710 TB

Cabinet:
- 24 Boards
- 96 Nodes
- 139 TF
- 3.6 TB

Board:
- 4 Compute Nodes
- 5.8 TF
- 152 GB

Compute Node:
- 1.45 TF
- 38 GB
USA, 2012: BlueGene strikes back

- Built by IBM for NNSA and installed at LLNL
- 20,123.7 TFLOPS peak performance
  - Blue Gene/Q architecture
  - 1,572,864 total PowerPC A2 cores
  - 98,304 nodes in 96 racks occupy 280m²
  - 1,572,864 GB DDR3 memory
  - 5-D torus interconnect
  - 768 I/O nodes
  - 7890kW power, or 2.07 GFLOPS/W
  - Achieves 16,324.8 TFLOPS in HPL (#1 in June 2012), about 14 PFLOPS in HACC (cosmology simulation), and 12 PFLOPS in Cardioid code (electrophysiology)
Linpack run with 705,024 cores at 10.51 Pflop/s (88,128 CPUs), 12.7 MW; 29.5 hours. Fujitsu to have a 100 Pflop/s system in 2014.
China

First Chinese Supercomputer to use a Chinese Processor

- Sunway BlueLight MPP
- ShenWei SW1600 processor, 16 core, 65 nm, fabbed in China
- 125 Gflop/s peak
- #14 with 139,364 cores, 0.796 Pflop/s & 1.07 Pflop/s Peak
- Power Efficiency 741 Mflops/W

Coming soon, Loongson (Godson) processor

- 8-core, 65nm Loongson 3B processor runs at 1.05 GHz, with a peak performance of 128 Gflop/s
Recent Developments

- US DOE planning to deploy $O(100)$ Pflop/s systems for 2017-2018 - $525M$ hardware
- Oak Ridge Lab and Lawrence Livermore Lab to receive IBM and Nvidia based systems
- Argonne Lab to receive Intel based system
  - After this Exaflops
- US Dept of Commerce is preventing some China groups from receiving Intel technology
  - Citing concerns about nuclear research being done with the systems; February 2015.
  - On the blockade list:
    - National SC Center Guangzhou, site of Tianhe-2
    - National SC Center Tianjin, site of Tianhe-1A
    - National University for Defense Technology
    - National SC Center Changsha
Yutong Lu from NUDT at the International Supercomputer Conference in Germany in July

### Status of Tianhe System

<table>
<thead>
<tr>
<th>System</th>
<th>Tianhe-1A</th>
<th>Tianhe-2</th>
<th>Tianhe-2A</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Peak (PF)</td>
<td>4.7</td>
<td>54.9</td>
<td>~100</td>
</tr>
<tr>
<td>Peak Power (MW)</td>
<td>4.04</td>
<td>17.8</td>
<td>~18</td>
</tr>
<tr>
<td>Total System Memory</td>
<td>262 TB</td>
<td>1.4 PB</td>
<td>~3 PB</td>
</tr>
<tr>
<td>Node Performance (TF)</td>
<td>0.655</td>
<td>3.431</td>
<td>~6</td>
</tr>
<tr>
<td>Node processors</td>
<td>Xeon X5670, Nvidia M2050</td>
<td>Xeon E5 2692, Xeon Phi</td>
<td>Xeon E5 2692, China Accelerator</td>
</tr>
<tr>
<td>System size (nodes)</td>
<td>7,168 nodes</td>
<td>16,000 nodes</td>
<td>~18,000</td>
</tr>
<tr>
<td>System Interconnect</td>
<td>TH Express-1</td>
<td>TH Express-2</td>
<td>TH Express-2+</td>
</tr>
<tr>
<td>File System</td>
<td>2 PB Lustre</td>
<td>12.4 PB H²FS+Lustre</td>
<td>~30 PB H²FS+TDM</td>
</tr>
</tbody>
</table>
China Accelerator

Matrix2000 GPDSP

- High Performance
  - 64bit Supported
  - ~2.4/4.8TFlops (DP/SP)
  - 1GHz, ~200W

- High Throughput
  - High-bandwidth Memory
  - 32~64GB
  - PCIe 3.0, 16x
Industrial Use of Supercomputers

- Of the 500 Fastest Supercomputer
  - Worldwide, Industrial Use is ~ 50%

- Aerospace
- Automotive
- Biology
- CFD
- Database
- Defense
- Digital Content Creation
- Digital Media
- Electronics
- Energy
- Environment
- Finance
- Gaming
- Geophysics
- Image Proc./Rendering
- Information Processing Service
- Information Service
- Life Science
- Media
- Medicine
- Pharmaceutics
- Research
- Retail
- Semiconductor
- Telecomm
- Weather and Climate Research
- Weather Forecasting
Power is an Industry Wide Problem

Google facilities
- leveraging hydroelectric power
- old aluminum plants

“Hiding in Plain Sight, Google Seeks More Power”, by John Markoff, June 14, 2006

Microsoft and Yahoo are building big data centers upstream in Wenatchee and Quincy, Wash.
- To keep up with Google, which means they need cheap electricity and readily accessible data networking

Microsoft Quincy, Wash.
470,000 Sq Ft, 47MW!
Commercial Data Centers

- Facebook: 300,000 sq ft, 1.5 cents per kW hour, Prineville OR

- Microsoft: 700,000 sq ft in Chicago

- Apple: 500,000 sq ft in Rural NC, 4 cents kW/h
**COOLING:** High-efficiency water-based cooling systems—less energy-intensive than traditional chillers—circulate cold water through the containers to remove heat, eliminating the need for air-conditioned rooms.

**STRUCTURE:** A 24,000-square-meter facility houses 400 containers. Delivered by trucks, the containers attach to a spine infrastructure that feeds network connectivity, power, and water. The data center has no conventional raised floors.

**POWER:** Two power substations feed a total of 300 megawatts to the data center, with 200 MW used for computing equipment and 100 MW for cooling and electrical losses. Batteries and generators provide backup power.

**CONTAINER:** Each 67.5-cubic-meter container houses 2,500 servers, about 10 times as many as conventional data centers pack in the same space. Each container integrates computing, networking, power, and cooling systems.
Future Computer Systems

- Most likely be a hybrid design
- Think standard multicore chips and accelerator (GPUs)
- Today accelerators are attached over slow links
- Next generation more integrated
- Intel’s Xeon Phi
  - 244 “threads” 61 cores
- AMD’s Fusion
  - Multicore with embedded graphics ATI
- Nvidia’s Kepler with 2688 “Cuda cores”, 14 cores
  - Project Denver plans to develop an integrated chip using ARM architecture in future?
Commodity plus Accelerator Today

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

**Accelerator (GPU)**
- Nvidia K20X “Kepler”
- 2688 “Cuda cores”
- .732 GHz
- 2688*2/3 ops/cycle
- 1.31 Tflop/s (DP)

**Interconnect**
- PCI-X 16 lane
- 64 Gb/s (8 GB/s)
- 1 GW/s

**Host Memory**

**Device Memory**
- 6 GB

**Thread Execution Control Unit**

**192 Cuda cores/SM**
Performance Share of Accelerators

Fraction of Total TOP500 Performance

- 0%
- 5%
- 10%
- 15%
- 20%
- 25%
- 30%
- 35%
- 40%

Years:
- 2006
- 2007
- 2008
- 2009
- 2010
- 2011
- 2012
- 2013
- 2014
Multi- to Many-Core

- **Complex cores:** huge, complex, lots of internal concurrency latency hiding
- **Simple cores:** small, simpler core little internal concurrency latency-sensitive

- All Complex Cores  
  *e.g. Intel Xeon*

- Mixed Big & Small Cores

- All Small Cores  
  *e.g. Intel MIC*
Challenges of using Accelerators

- **High levels of parallelism**
  Many GPU/Accelerator cores, serial kernel execution
  [ e.g. 240 in the Nvidia Tesla; up to 512 in *Fermi* - to have concurrent kernel execution ]

- **Hybrid/heterogeneous architectures**
  Match algorithmic requirements to architectural strengths
  [ e.g. small, non-parallelizable tasks to run on CPU, large and parallelizable on GPU ]

- **Compute vs communication gap**
  Exponentially growing gap; persistent challenge
  [ Processor speed improves 59%, memory bandwidth 23%, latency 5.5% ]
  [ on all levels, e.g. a GPU Tesla C1070 (4 x C1060) has compute power of O(1,000) Gflop/s but GPUs communicate through the CPU using O(1) GB/s connection ]
Problem with Multicore

• As we put more processing power on the multicore chip, one of the problems is getting the data to the cores

• Next generation will be more integrated, 3D design with a photonic network
• Number of cores per chip doubles every 2 year, while clock speed decreases (not increases).
  • Need to deal with systems with millions of concurrent threads
    • Future generation will have billions of threads!
  • Need to be able to easily replace inter-chip parallelism with intro-chip parallelism
• Number of threads of execution doubles every 2 year
Power Level (kW)

- Top 25
  - 43 machines > 1 MW
  - 20 machines > 2 MW
Power Consumption

- TOP10: 5.04 x in 5 y
- TOP50: 3.13 x in 5 y
- TOP500: 3.25 x in 5 y
Power Efficiency

For Exascale need to be around 50 Gflops/W
A factor of 20 needed
The High Cost of Data Movement

• Flop/s or percentage of peak flop/s become much less relevant

Approximate power costs (in picoJoules)

<table>
<thead>
<tr>
<th></th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP FMADD flop</td>
<td>100 pJ</td>
</tr>
<tr>
<td>DP DRAM read</td>
<td>4800 pJ</td>
</tr>
<tr>
<td>Local Interconnect</td>
<td>7500 pJ</td>
</tr>
<tr>
<td>Cross System</td>
<td>9000 pJ</td>
</tr>
</tbody>
</table>

Source: John Shalf, LBNL

• Algorithms & Software: minimize data movement; perform more work per unit data movement.
nVIDIA's 28nm chips.
For a floating point operation costs 20pJ.
Getting the operands from local memory (situated 1mm away) consumes 26pJ.
If the operands need to be obtained from the other end of the die, it requires 1nJ
while if the operands need to be read from DRAM, the cost is 16nJ.
Energy Cost Challenge

• At ~$1M per MW energy costs are substantial
  - 10 Pflop/s in 2011 uses ~10 MWs
  - 1 Eflop/s in 2020 > 100 MWs

• DOE Target: 1 Eflop/s around 2020-2022 at 20 MWs
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)
Can I Build an Exascale Machine Today?

- Sure, #1 system is at ~33 Pflop/s
- Need 30 of them.
- 33 Pflop/s at 20 MW; need 600 MW
- 33 Pflop/s cost ~$400 M; need $12 B
- 33 Pflop/s has 3M cores; need 100 M cores
- 33 Pflop/s in 1 building; need 30 buildings

- Each of these pose great challenges.
The winning architecture for building exascale systems, heterogeneous or homogeneous, and why?

- **Multicore**: Maintain complex cores, and replicate (x86, SPARC, Power7) [#4 and 9]

- **Manycore/Embedded**: Use many simpler, low power cores from embedded (BlueGene, future ARM) [#3, 5, 7, and 8]

- **GPU/Coprocessor/Accelerator**: Use highly specialized processors from graphics market space (NVidia Fermi, Intel Xeon Phi, AMD) [# 1, 2, 6, and 10]
## Conventional Wisdom is Changing

<table>
<thead>
<tr>
<th>Old Conventional Wisdom</th>
<th>New Conventional Wisdom</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Peak clock frequency as primary limiter for performance improvement</td>
<td>• Power is primary design constraint for future HPC system design</td>
</tr>
<tr>
<td>• Cost: FLOPs are biggest cost for system: optimize for compute</td>
<td>• Cost: Data movement dominates optimize to minimize data movement</td>
</tr>
<tr>
<td>• Concurrency: Modest growth of parallelism by adding nodes</td>
<td>• Concurrency: Exponential growth of parallelism within chips</td>
</tr>
<tr>
<td>• Memory scaling: maintain byte per flop capacity and bandwidth</td>
<td>• Memory Scaling: Compute growing 2x faster than capacity or bandwidth</td>
</tr>
<tr>
<td>• Uniformity: Assume uniform system performance</td>
<td>• Heterogeneity: Architectural and performance non-uniformity increase</td>
</tr>
<tr>
<td>• Reliability: It’s the hardware’s problem</td>
<td>• Reliability: Cannot count on hardware protection alone</td>
</tr>
</tbody>
</table>
Evolution Over the Last 30 Years

- Initially, commodity PCs were decentralized systems.
- As chip manufacturing process shrank to less than a micron, they started to integrate features on-die:
  - 1989: FPU (Intel 80486DX)
  - 1999: SRAM (Intel Pentium III)
  - 2009: GPU (AMD Fusion)
  - 2016: DRAM on chip (3D stacking)
Future Systems May Be Composed of Different Kinds of Cores

- DRAM chips (cells)
- 3D DRAM (cells)

Latency: lower latency
Bandwidth: higher bandwidth

Address | Data
Memory controller

12/1
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
Critical Issues at Peta & Exascale for Algorithm and Software Design

- **Synchronization-reducing algorithms**
  - Break Fork-Join model

- **Communication-reducing algorithms**
  - Use methods which have lower bound on communication

- **Mixed precision methods**
  - 2x speed of ops and 2x speed for data movement

- **Autotuning**
  - Today’s machines are too complicated, build “smarts” into software to adapt to the hardware

- **Fault resilient algorithms**
  - Implement algorithms that can recover from failures

- **Reproducibility of results**
  - Today we can’t guarantee this. We understand the issues, but some of our “colleagues” have a hard time with this.
Conclusions

• For the last decade or more, the research investment strategy has been overwhelmingly biased in favor of hardware.
• This strategy needs to be rebalanced - barriers to progress are increasingly on the software side.
• Moreover, the return on investment is more favorable to software.
  ▪ Hardware has a half-life measured in years, while software has a half-life measured in decades.
• High Performance Ecosystem out of balance
  ▪ Hardware, OS, Compilers, Software, Algorithms, Applications
    • No Moore’s Law for software, algorithms and applications