Exascale?

Pete Beckman, Argonne National Laboratory

Director, Exascale Technology and Computing Institute
Co-Director, Northwestern University – Argonne Institute of Science and Engineering
Data from Peter Kogge, Notre Dame
What’s the Power Problem?

- *Mira*: Blue Gene/Q System
  - 20 times faster than BG/P *Intrepid* (10 PF)
  - ~4 times more power (~4 MW)
  - ~5X more power efficient than BG/P

- Repeat twice to reach Exascale?
  - 400 times faster than BG/Q *Mira* (4 EF)
  - ~16 times more power (~64 MW)
  - ~25X more power efficient than BG/Q
Japan: Japan’s Policy toward Exascale Computing

Yoshio KAWAGUCHI
Office for Promotion of Computing Science / MEXT
27 February, 2014

Japan Exascale System Development

Outline:
- Double-digits (higher) performance by 2020
- Push state of the art in power efficiency, scalability & reliability
- Enable unprecedented application capability
- AICS RIKEN in charge of exascale systems development
- Total project cost ca. JPY140 billion with about JPY 110 billion from the government’s budget (JPY 1.2 billion for 2014)

Schedule:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
<td>Basic Design</td>
<td>Trial Production</td>
<td>Detailed Design</td>
<td>Manufacturing</td>
<td>Installation</td>
<td>Operation</td>
<td></td>
</tr>
<tr>
<td>Application</td>
<td>Development and utilization of Application for Exascale Computing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| 1 | China | Tianhe-2 |
| 2 | USA   | Titan    |
| 3 | USA   | Sequoia  |
| 4 | Japan | K Computer |
| 5 | USA   | Mira     |
| 6 | Switzerland | Piz Daint |
| 7 | USA   | Stampede |
| 8 | Germany | JUQUEEN |
| 9 | USA   | Vulcan   |
| 10 | USA   | ???     |
US-Japan Agreement

Yoshio Kawaguchi (MEXT, Japan) and William Harrod (DOE, USA)

Yoshio Kawaguchi
Director
Office for the Promotion of Computing Science,
Research Promotion Bureau
Ministry of Education, Culture, Sports, Science and Technology (MEXT)

Shinya TAHATA
Director for Information Science and Technology, Information Division,
Research Promotion Bureau
Ministry of Education, Culture, Sports, Science and Technology (MEXT)
Europe:

**High Performance Computing in Horizon 2020**

Big Data and Extreme Scale Computing Workshop  
February 26-28, 2014 – Fukuoka Japan

*Excellence in Science*  
DG CONNECT  
European Commission

Jean-Yves Berthou ANR  
on behalf of the European Commission Leonardo Flores, Panagiotis Tsarchopoulos, Aniyan Varghese

• **HPC PPP starting 1st January 2014: 700 m€ for the period 2014-2020** (€143,4 million in Calls in 2014-2015)

<table>
<thead>
<tr>
<th>An Integrated HPC approach</th>
<th>Excellent Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ HPC strategy combining three elements:</td>
<td></td>
</tr>
<tr>
<td>(a) Computer Science: towards exascale HPC; A special FET initiative focussing on the next generations of exascale computing as a key horizontal enabler for advanced modelling, simulation and big-data applications [HPC in Future and Emerging Technologies (FET)]</td>
<td></td>
</tr>
<tr>
<td>(b) providing access to the best supercomputing facilities and services for both industry and academia; PRACE - world-class HPC infrastructure for the best research [HPC in e-infrastructures]</td>
<td></td>
</tr>
<tr>
<td>(c) achieving excellence in HPC applications; Centres of Excellence for scientific/industrial HPC applications in (new) domains that are most important for Europe [HPC in e-infrastructures]</td>
<td></td>
</tr>
<tr>
<td>• complemented with training, education and skills development in HPC</td>
<td></td>
</tr>
<tr>
<td>(a) and (c) will be implemented in the context of the HPC Public-Private Partnership</td>
<td></td>
</tr>
</tbody>
</table>
China: (currently with fastest machine)
Chinese Tianhe-2

32K Intel Ivy Bridge Xeon Sockets
48K Intel Phi Sockets
~3M cores
~55 PF/sec peak
~30 PF/sec Linpack
~24 MW power with cooling
History of Supercomputer Performance

Top 500 List


1 E flop/s
100 Pflop/s
10 P flop/s
1 Pflop/s
100 Tflop/s
10 T flop/s
1 Tflop/s
100 G flop/s
10 G flop/s
1 G flop/s
100 M flop/s

1.17 TFlop/s
59.7 GFlop/s
400 MFlop/s
N=1
N=500

SUM
162 PFlop/s
17.6 PFlop/s
76.5 TFlop/s

Courtesy of Jack Dongarra & Erich Strohmaier
History of Supercomputer Performance

Top 500 List

162 PFlop/s
17.6 PFlop/s
76.5 TFlop/s

my Laptop
my iPad


1 Eflop/s
100 Pflop/s
10 Pflop/s
1 Pflop/s
100 Tflop/s
10 Tflop/s
1 Tflop/s
100 Gflop/s
10 Gflop/s
1 Gflop/s
100 Mflop/s

100
75
56
48
46
42
10
70
63
65
60
632
595
507
561
50

my iPad
my Laptop

10
10

Pete Beckman
Argonne National Laboratory
To Reach Exascale: Will There be a Revolution?
The Software in the Middle: The Forces of Change

- Memory
- Threads
- Messaging
- Resilience
- Power
Memory: Technology Summary from Rob Schreiber

New memory on the horizon

- Spin-Torque-Transfer RAM (STTRAM)
  - Grandis (54nm, acquired by Samsung)

- Phase-Change RAM (PCRAM)
  - Samsung (20nm, diode, up to 8Gb)
  - Micron and Nokia – in phones now

- Resistive RAM (ReRAM)
  - Panasonic (180nm process, 4-layer xpoint)
  - Unity Semi (64MB, acquired by Rambus)

Spin transfer torque (STTRAM)

PCRAM

- Shipping today
- MLC (limited by resistance drift)
- Slow, expensive writes
- Wearout issue

ReRAM

Samsung, HP-Hynix, Sandisk, Toshiba
32Gb test chip (Sandisk/Toshiba, 24 nm, ISSCC 2013)
Fast (tens of nsecs) for both read and write
Good data retention and reliability
3D – 2 to 4 layers
MLC possible
Memory will be local, and include NVRAM

- Helps reduce power
- Helps with resilience
- Helps with cost
- Helps with Big Data (capacity)
Great! Where is the Revolution?

Programming Model and OS Interface:

- Need dynamic run-time to move ‘pages’ between RAM and NVRAM
- Need programming model that can hint (or manage) write-once and read only & data movement
- Coherence islands
- Add persistence to programming model
- Programming model for PIM
Threads/Tasks: Managing Exploding Parallelism

- Dynamic parallelism and decomposition
  - Programmer cannot hand-pick granularity / resource mapping
    - (equal work != equal time)

Variability is the new norm:
- Power
- Resilience
- Intrahode Contention
Google (re-discovers) OS Noise & Contention

Component-Level Variability Amplified By Scale
A common technique for reducing latency in large-scale online services is to parallelize sub-operations across many different machines, where each sub-operation is co-located with its portion of a large dataset. Parallelization happens by fanning out a request from a root to a large number of leaf servers and merging responses via a request-distribution tree. These sub-operations must all complete within a strict deadline for the...

Living with Latency Variability
The careful engineering techniques in the preceding section are essential for building high-performance interactive services, but the scale and complexity of modern Web services make it infeasible to eliminate all latency variability. Even if such perfect behavior could...

Reducing Component Variability
Interactive response-time variability can be reduced by ensuring interactive requests are serviced in a timely manner...
Messaging (send/recv & put/get):

Where is the Revolution?

- We have done inter-node messaging for decades... what is new?
- Millions of outstanding threads (messages) per socket to hide latency
- Hardware Wake-On Threads (no polling)
  - BG/Q
  - X86 Mwait
- Network -> Cache/NVRAM Injection
- New OS/R Interfaces
Fault Predictions are Hard
What will be the Revolution?

What we do know:
- Driving down power increases faults
- Vendors have great market incentive to redesign for reliable hardware
- Our current HPC software is very fragile
- We should improve resilience
- Build solutions at multiple layers

“Over the past thirty years there have been several predictions of the eminent cessation of the rate of improvement in computer performance. Every such prediction was wrong. They were wrong because they hinged on unstated assumptions that were overturned by subsequent events.”
Exploring Power and Temp

CMOS-based thermal sensors available via MSR (cpuonline)
Thermal Effects in High Performance Systems

- Temperature variation
  - Across cores, chips, memory versus cores, across nodes
    - input data dependent
    - spatial and temporal variation
    - hard to predict at design time
    - dynamic architecture highly desired
Revolution Areas:

- Architecture
- Memory
- Threads
- Messaging
- Resilience
- Power
Questions?  (then time for some more fun topics)