DOE Extreme Scale Computing

William Harrod
ASCR
October 19, 2011

Stratford, TX 1935
Outline

- View from Germantown
- Exascale Technical Challenges
- Reports to Congress
- Summary
The Scientific Challenges:
- Deliver next-generation scientific applications using today’s petascale computers.
- Discover, develop and deploy tomorrow’s Exascale computing and networking capabilities.
- Develop, in partnership with U.S. industry, next generation computing hardware and tools for science.
- Discover new applied mathematics and computer science for the ultra-low power, multicore-computing future.
- Provide technological innovations for U.S. leadership in Information Technology to advance competitiveness.

FY 2013 Highlights:
- Investments in U.S. industry to address critical challenges in hardware and technologies on the path to Exascale.
- Co-design centers to deliver next generation scientific applications by coupling application development with formulation of computer hardware architectures and system software.
- Operation of a 10 petaflop low-power IBM Blue Gene/Q at the Argonne Leadership Computing Facility and a hybrid, multi-core prototype computer at the Oak Ridge Leadership Computing Facility.
- ESnet puts 100Gbps technologies into production.
FY12 is 3rd year of this University & DOE Laboratory funding opportunity:

• Posted on July 19, 2011
• Pre-applications due Sep 1
• Encourage applications by Oct 3
• Full applications due Nov 29

ASCR base research programs:

• Applied Mathematics – 4 topics areas
• Computer Science – topic areas are focused on 3 key software challenges for Exascale platforms

ASCR Panel Reviews: Jan 24-25, 2012

• In Rockville/Bethesda, MD

For rules and requirements, see
http://science.energy.gov/early-career
The key finding of the Panel is that there are compelling needs for Exascale computing capability to support the DOE’s missions in energy, national security, fundamental sciences, and the environment. The DOE has the necessary assets to initiate a program that would accelerate the development of such capability to meet its own needs and by so doing benefit other national interests. Failure to initiate an Exascale program could lead to a loss of U. S. competitiveness in several critical technologies.”

Trivelpiece Panel Report, January, 2010
Senate Appropriations Committee

ENERGY AND WATER DEVELOPMENT APPROPRIATIONS BILL, 2012

- The Committee supports the Department’s initiative to develop Exascale computing—1,000 times more powerful than today’s most powerful computer. The Committee recommends $126,000,000 (FENCED) to support this initiative, which includes $90,000,000 for the Office of Science and $36,000,000 for the National Nuclear Security Administration.

- The Committee understands that the path to Exascale computing will be extremely challenging and will require significant research and development breakthroughs.

- The Committee understands that the Department (of Energy) will have the lead Government role in computing research and development. The Department’s role in developing more advanced computing platforms is even more important with the elimination of the DARPA High Performance Computing program.

NOT PASSED BY CONGRESS AT THIS TIME
Exascale Co-Design Centers

<table>
<thead>
<tr>
<th>Exascale Co-Design Centers</th>
<th>ExMatEx (Germann)</th>
<th>CESAR (Rosner)</th>
<th>EXACT (Chen)</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Labs</td>
<td>LANL</td>
<td>ANL</td>
<td>SNL</td>
</tr>
<tr>
<td></td>
<td>LLNL</td>
<td>PNNL</td>
<td>LBNL</td>
</tr>
<tr>
<td></td>
<td>SNL</td>
<td>LANL</td>
<td>LANL</td>
</tr>
<tr>
<td></td>
<td>ORNL</td>
<td>ORNL</td>
<td>ORNL</td>
</tr>
<tr>
<td></td>
<td>LLNL</td>
<td>LLNL</td>
<td>NREL</td>
</tr>
<tr>
<td>University &amp; Industry Partners</td>
<td>Stanford</td>
<td>Studsvik</td>
<td>Stanford</td>
</tr>
<tr>
<td></td>
<td>CalTech</td>
<td>TAMU</td>
<td>GA Tech</td>
</tr>
<tr>
<td></td>
<td>Rice</td>
<td>Rutgers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U Chicago</td>
<td>UT Austin</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IBM</td>
<td>Utah</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TerraPower</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>General Atomic</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Areva</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
FY 2010 Basic Research funded in Exascale related topic areas:
- **Applied Math**: Uncertainty Quantification
- **Computer Science**:
  - Advanced Architectures
  - X-Stack
  - Scientific Data Management and Analysis

FY2011:
- **Computational Partnerships**: 3 Co-Design Centers Funded
- Exascale MOU with NNSA signed
- Request for Information (released by ANL): critical and platform technologies

FY2012:
- Exascale Plan to Congress

On-going
- Exascale Coordination meetings with other Federal Agencies
Challenges to Achieving Exascale Computing

- It has been estimated that an Exascale computer will consist of one billion cores.

- **Power**
  - 1-2 nJ/operation today
  - 20 pJ/flop (or 50 GFLOPS/W system) required for ExaScale (decrease by 1000x)
  - Dominated by data movement and overhead

- **Programmability**
  - Writing an efficient parallel program is hard
  - Locality required for efficiency
  - System complexity greatly inhibits programmability

Exascale Goal 50 GFLOPS/W
35 YEARS OF MICROPROCESSOR TREND DATA

- Transistors (thousands)
- Single-thread Performance (SpecINT)
- Frequency (MHz)
- Typical Power (Watts)
- Number of Cores

Optimized Power Performance
(Data from Intel: Shekhar Borkar)

65nm CMOS, 50°C

Maximum Frequency (MHz)

Total Power (mW)

Supply Voltage (V)

Energy Efficiency (GOPS/Watt)

Active Leakage Power (mW)

320mV

=V_T

9.6X

320mV

Subthreshold Region

October 19, 2011
Data Path Insights

• The cost of data storage and access – *Dominates energy consumption*

• Computer design must keep or make data effectively local to processing

• Cost to access increasingly distant data – energy and latency – may appear to be “linear”
  - It is not, data access costs increases dramatically with distance to data
  - Interconnect, latency, and overhead costs grow disproportionately with distance to data

• Need to redesign the entire data access path:
  - Memory technologies – reduce access cost
    - Transfer energy: DDR3 DRAM 40 pJ/bit $\Rightarrow$ 1-3 pJ/bit
  - Packaging costs – reduce bit movement cost
Peak Exascale Machine Power

Goals
20 MW
20 pJ/Flop

980 MW
980 pJ/Flop

Shekhar Borkar, Intel
October 19, 2011
• ASCR is preparing for Extreme Scale computing and continuing to support current generation HPC systems.

• Applied Mathematics projects should have DOE relevance across a broad range of science domains:
  – SC offices (BES, BER, FES, ...)
  – Applied Offices (Nuclear Energy, Fossil Energy, Office of Electricity Delivery and Energy Reliability, ...)
  – NNSA
Anticipated Future Programs
(more to come)

- **X-Stack: Runtimes, Programming Models, Languages, Compilers, and Tools**
  - Minimize exposure of system complexity
  - Extreme concurrency
  - Heterogeneous system
  - Minimize data movement
  - Runtimes for efficiency and resiliency
  - Self-aware OS/runtime

- **Extreme Scale Solver Algorithms**
  - Fine grain parallelism
  - Data movement & locality
The Real Challenges

• Avoiding mediocre solutions
  – Evolving existing systems

• Developing a new software stack for Extreme Scale systems
  – Not treating it as an “after thought”
  – Failure to develop programmable systems

• Developing efficient mathematical algorithms that deliver high performance for Extreme Scale systems
  – Minimize communications
  – Fault tolerant computations
  – Asynchronous, fine grain parallelism

• New computers designs based on a new execution model
  – Must be based on 2020 COTS technology
  – Exotic technology is not an option

It is a Budget Problem!
Voltage scaling to reduce power and energy
   Explodes parallelism
   Cost of communication vs. computation—critical balance

It’s not about the FLOPS. It’s about data movement.
   Algorithms should be designed to perform more work per unit data movement.
   Programming systems should further optimize this data movement.
   Architectures should facilitate this by providing an exposed hierarchy and efficient communication.

System software to orchestrate all of the above
   Self aware operating system

Stratford, TX 2011
October 19, 2011
### Exascale Timeline

*(from ASCR Arch I Workshop)*

<table>
<thead>
<tr>
<th>Event Description</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Year 5</th>
<th>Year 6</th>
<th>Year 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exascale Programming Environment – v0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exascale Programming Environment – v0.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software Technology – OS, runtime, programming environment &amp; resiliency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exascale Co-Design</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System Build Phase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RFP issued</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System Build Contract Starts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First node</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Cabinet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accept Exascale System</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Notes:**

- **October 19, 2011**: Applied Math PI Meeting
- **October 2011**: Exascale Co-Design