The Sequoia story

Presented to:
HPC User Forum

Terri Quinn, Principal Deputy Department Head, Integrated Computing and Communications

September 18, 2012
The Dearborn Inn
Dearborn, MI
Sequoia is a story about a partnership that created a unique tool for scientific simulation.
BlueGene started with a recognition that a disruption in HPC technology was needed

- Total Cost of Ownership
- Facility demands
- Computing Power (UQ)
- Few or no App code changes

and yet

Spawned a decade of a partnership with IBM and ANL to create the Blue Gene line of computers, sharing the costs and the risks
Blue Gene Roadmap

Goals:
- Three orders of magnitude performance in 10 years
- Push state of the art in Power efficiency, scalability, & reliability
- Enable unprecedented application capability
- Exploit new technologies: PCM, photonics, 3DP

Performance

Blue Gene / L
PPC 440 @700MHz
596+ TF

Blue Gene / P
PPC 450 @850MHz
1+ PF

Blue Gene / Q
In progress
20+ PF

Goals:
- Lay the ground work for ExaFlop & usability
- Address many of the power efficiency, reliability and technology challenges

2004  2008  2012  2016  2020
**BGQ System Highlights**

- **Highly scalable homogeneous system**
  - Each rack of BG/Q has 1,024 16-core compute nodes (209TF/sec peak)
  - Can scale to 512 racks achieving 100 PF/sec peak

- **Open source and standards-based programming environment**
  - Full-featured Linux on service, front end, and I/O nodes
  - Lightweight Compute Node Kernel (CNK)
  - Standards based program env. (e.g., OpenMP3.0, GNU tools)

- **Efficient operational characteristics**
  - 5D torus for tremendous bisection bandwidth
  - Speculative execution, transactional memory, fast thread handoff
  - Partitionable application isolation with reproducible runtime results

**Most power and space efficient supercomputer today**
Blue Gene/Q Compute Card

- Basic Field Replacement Unit of a BlueGene/Q system
- Compute Card has 1 BQC chip + 72 SDRAMs (16GB DDR3)
- 204.8 GLOPs Peak
- 16 + 1 cores SMP
- Cores 1.6 GHz
- 4-way hw threads
- Quad SIMD FPU
- 32 M shared L2 cache
Blue Gene/Q Node Card assembly

- Power efficient processor chips allow dense packaging
- High bandwidth / low latency electrical interconnect on-board
- Compute Node Card assembly is water-cooled (18-25°C – above dew point)
- Redundant power supplies
IBM

Inter-Processor Communication

- Integrated 5D torus
  - Hardware assisted collective and barrier
  - FP addition support in network
  - Virtual Cut Through
  - RDMA direct from application
  - Wrapped

- 2 GB/s bandwidth on all 10 links (4 GB/s bidi)

- 5D nearest neighbor exchange measured at ~1.75 GB/s per link

Network Performance
- All-to-all: 97% of peak
- Bisection: > 93% of peak
- Nearest-neighbor: 98% of peak
- Collective: FP reductions at 94.6% of peak

- Hardware latency
  - Nearest: 80ns
  - Farthest: 3us (96-rack 20PF system)
The 20PF Sequoia system will develop key technologies for predictive simulation

**Sequoia at glance**
- 20 PF/s peak
- Memory 1.5 PB, 4 PB/s BW
- 1.5M Cores, 6M threads
- 50 PB Disk
- 9.6MW max power; 4,000 ft²
- Third-generation IBM BlueGene

**Mission requirements**
- Run 24 simultaneous Purple-class Integrated Design Code calculations while also running....
- Weapons science at 4 PF sustained
## It’s a long road to a production 20PF machine

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Scalable Application Preparation Project</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dawn(BG/P) Early Demonstration System in Production</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statement of Work Finalized</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Sequoia Facilities Preparation**

- Sequoia File System Delivery
- Sequoia Delivery
  - File System Integration
  - Sequoia Integration

**Sequoia Acceptance**

- Production Sequoia
Movie interlude
Facilities work was a grand challenge

- Custom designed receptacles
  - Reduced under-floor congestion
  - All 4 feet of subfloor are packed

- Mechanical room requirements:
  91% liquid cooled and 9% air cooled
  - New tertiary CHW loop
  - GPM/rack = 25 to 30

- Electrical requirements:
  100 kW/rack = 9.6MW

- Cooling monitoring
  - Utility grade monitoring and control system

- Physical requirements
  - 96 racks in 4,000 ft²
  - 4,500 lbs/racks = 210 tons total
Extrem scale HPC pushes the boundaries of the facility electrical, mechanical and structural infrastructure.
Extreme scale HPC pushes the boundaries of the facility electrical, mechanical and structural infrastructure
HPC facility strategy plan

- Focus on sustainability and energy efficiency through flexibility
- Scale footprint with the computational technology
- Deploy innovative HPC facility design methodologies
Scalable building concept

- Design the facility with the technology
- Make it flexible
  - Scale power
  - Scale square footage and structure
  - Scale cooling solutions
- Minimize the use of cooling towers and chillers
- More cost effective than containerized solutions
- One modular unit
  - 3 to 6MW in 6,000 SF
How will we fare if we do not invest in technology for the future? Poorly…..

We may have to buy 9x to 10x computing HW to get 2x to 3x performance

- Many DOE missions demand higher fidelity multi-physics simulations and more capable HPC systems
- However ….. trends in computing HW lead to degraded performance and far higher energy costs

Ratio of memory bandwidth & capacity to computing is shrinking

Peak performance = what we buy 9x to 10x more hardware
Effective performance = what we get is 2x to 3x app performance
Operating costs are expected to increase by 2x to 3x due to system power

An $150M electric bill for a system that costs $180M!
# FastForward in a nutshell

| Who       | 2 DOE Orgs (Science/NNSA)  
|           | 7 National Labs  
|           | 5 (now 4) US companies |
| What      | Fund $62.5M of R&D for processors, memory, and storage technologies for a broad market |
| When      | February 2012 to June 30, 2014 |
| Why       | Influence critical HPC technologies |
| Results   | Contracts were awarded by June 29, 2012 |
| Next step | Set up DOE/FF awardee 2-year collaborations |
FastForward is an offensive maneuver to tackle the problem early

- High-value R&D promising to:
  - increase performance of DOE simulations
  - decrease energy usage
  - benefit the broad market
  - be available in large-scale DOE systems in 5 to 10 years

### FastForward Awardees

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Value</th>
<th>Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMD Advanced Research LLC</td>
<td>$12,600,000</td>
<td>Processor/Memory R&amp;D</td>
</tr>
<tr>
<td>IBM Corporation</td>
<td>$10,476,714</td>
<td>Memory R&amp;D</td>
</tr>
<tr>
<td>Intel Federal LLC</td>
<td>$18,963,437</td>
<td>Processor/Memory R&amp;D</td>
</tr>
<tr>
<td>Nvidia Corp.</td>
<td>$12,398,893</td>
<td>Processor R&amp;D</td>
</tr>
<tr>
<td>Whamcloud Inc. (Now Intel Federal LLC)</td>
<td>$7,996,053</td>
<td>Storage and I/O R&amp;D</td>
</tr>
<tr>
<td><strong>Total Subcontract Value</strong></td>
<td><strong>$62,435,097</strong></td>
<td></td>
</tr>
</tbody>
</table>