Oklahoma Supercomputing Symposium

Kent Winchell
Deep Computing CTO

October 2010
http://software-carpentry.org/

• Computers are as important to modern science as telescopes and test tubes. Unfortunately, most scientists are never taught how to use them effectively: most scientists have to figure out for themselves how to build, validate, maintain, and share complex programs. This is as fair as teaching someone arithmetic and then expecting them to figure out calculus on their own, and about as likely to succeed.
Smarter Planet Segments

- Smarter Education
- Cloud Computing
- Smarter Financial Systems
- Smarter Telecommunications
- Smarter Oil Management
- Public Safety
- Smarter Buildings
- Smarter Healthcare Systems
- Smarter Cities
- Smarter Planet
- Smarter Energy Grids
- Smarter Traffic Management
- Smarter Food Systems
- Products
- Retail
- Smarter Infrastructure
- Smarter Water Systems
- Smarter Information Management
- Government Services
High Performance Computing for a Smarter Planet (a partial list)

- Seismic Analysis
- Drug Discovery
- Actuarial Analysis
- Video On Demand
- Virtualization
- Earthquake Modeling
- Reservoir Analysis
- Protein Folding
- Asset Liability Management
- Network Optimization
- Data Management
- Climate Modeling
- Energy Conversion Systems
- Medical Imaging
- Portfolio Risk Analysis
- Gaming
- Cloud Computing
- Remediation

- Oil
- Healthcare
- Banking
- Telecom
- Infrastructure
- Environment
Impact on Industry

Weather Forecasting

Weather Research and Forecasting Models
IBM Power Systems, smaller versions of ASC Purple, used extensively throughout the industry

Large Scale Drop Impact Analysis of Mobile Phone

- ADVC, a commercial structural analysis code from Allied Engineering Corporation, Japan
  - An implicit based structural FEA system developed for parallel computation
- Drop Impact Analysis for a full assembly of a mobile phone is performed
  - Full model, including inner structure, virtually no simplifications
- Performance
  - 12 hours for 100 step simulation
  - 305 million degrees of freedom
  - 1.27 TF on 6192 Nodes of BG/L
  - 2.8% of Peak
  - 2.4 ms of real time
  - Drop height increased from 10 cm to 1.5 ml
- 2006 Gordon Bell Finalist

Real Time Options Trading

Current frontier with Code_Saturne

Calculation under way with 100 million cells
- PWR assembly mixing grid
- calculation on 4,000 to 8,000 procs
- major look due to mesh generation

Investment Banking Application
- Replace zoo of machines and Cisco with integrated Blue Gene/P
- Replace non-scalable communication infrastructure with System S
- Kittyhawk bridges between specialized Blue Gene/P hardware and legacy software

One minute peak Messages Per Second (MPS) rate

Aggregate view from the following systems/projections:
- Consolidated Tape System (CTS)
- Consolidated Quotation System (CQS)
- Options Pricing Reporting Authority (OPRA)
- NASDAQ Quotation Dissemination Service (NQDS)
BLUE WATERS
SUSTAINED PETASCALE COMPUTING

Blue Waters Update
Intense Computing at the Petascale and Beyond
Sustained Petascale computing will enable advances in a broad range of science and engineering disciplines:

- **Molecular Science**
- **Weather & Climate Forecasting**
- **Astrophysics**
- **Astronomy**
- **Earth Science**
- **Health**
- **Life Science**
- **Materials**
## Blue Waters Project Components

<table>
<thead>
<tr>
<th>Blue Waters Base System – Processors, Memory, Interconnect, On-line Storage, System Software, Programming Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Petascale Applications (Computing Resource Allocations)</td>
</tr>
<tr>
<td>Petascale Application Collaboration Team Support</td>
</tr>
<tr>
<td>Outstanding User and Production Support</td>
</tr>
<tr>
<td>WAN connections, Consulting, System Management, Security, Operations, ...</td>
</tr>
<tr>
<td>Value added Software – Collaborations</td>
</tr>
<tr>
<td>Value added hardware and software</td>
</tr>
<tr>
<td>Petascale Computing Facility</td>
</tr>
</tbody>
</table>

**Great Lakes Consortium for Petascale Computing**

**Value added**

| Hardware and software |

**Petascale Education, Industry and Outreach**

**IBM Summit – Poughkeepsie, NY – 2/1782010**
Focus on Sustained Performance

• **Blue Water’s and NSF are focusing on sustained performance in a way few have been before.**
  - *Sustained* is the computer’s performance on a broad range of applications that scientists and engineers use every day.
    - Time to solution is the metric – not Ops/s
    - Determined with real applications that include time to read data and write the results
  - NSF’s call emphasized sustained performance, demonstrated on a collection of application benchmarks (application + problem set)
    - Not just simplistic metrics (e.g. HP Linpack)
    - Applications include both Petascale applications (effectively use the full machine, solving scalability problems for both compute and I/O) and applications that use a fraction of the system
    - Metric is the time to solution
  - Blue Waters project focus is on delivering sustained PetaFLOPS performance to all applications
    - Develop tools, techniques, samples, that exploit all parts of the system
    - Explore new tools, programming models, and libraries to help applications get the most from the system
**PACTs = Required Benchmarks**

- **Petascale Application Collaboration Team**
  - Formed around each required benchmark in NSF solicitation for sustained petascale system

- **Three petascale applications/problem sizes**
  - Lattice-Gauge QCD (MILC)
  - Molecular Dynamics (NAMD)
  - Turbulence (DNS3D)

- **Ultimate Milestone**
  - Time-to-solution target (or 1 PFLOP sustained) for specified problem (size, time, physics, method)

- **Three non-petascale applications/problem sizes**
  - Lattice-Gauge QCD (MILC)
  - Materials Science (PARATEC)
  - Climate modeling (WRF)
PRAC Program

- Petascale Computing Resource Allocations
  - NSF to allocate Blue Waters time primarily through PRACs
  - Selected by NSF based on
    - Need for sustained petascale platform to carry out ground-breaking research
    - Likely to be ready to use Blue Waters effectively in 2011
- PRAC awardees receive travel funds and “provisional time”
- Awardees (total ~ 36 before Blue Waters operation)
  - Announcement of first round to be completed soon (~6 more expected)
  - Will accept applications on a continuing basis in future
- Blue Waters application and consulting staff will support awardees in preparing codes
Service Balance

- Job and resource scheduling to enable jobs to run for long blocks of time on large numbers of processors (as “determined by user requirements”).
- It was expected that 20-50 percent of the system would be used by a single application most of the time
  - 50 percent or more of the system may be used for shorter periods.
- Scientific utilization was expected to be 95 percent of the available time used for “petascale science.”
File System is GPFS

- IBM is implementing scaling changes in GPFS for the HPCS/DARPA project.
- Blue Waters will implement those changes in a persistent manner.
- GPFS configured to accommodate other local systems in a single namespace.
- Performance requirements are appropriately scaled to BW characteristics.

Archive is HPSS

- HPSS Hardware consists of three tape robots and appropriate numbers of tape drives.
  - Expect to expand this thru the lifetime of BW.
- HPSS integrated with BW.
  - GPFS-HPSS Interface.
  - Import-Export Portal.
    - Traditional HPSS commands.
- NCSA is contributing RAIT implementation to the HPSS community as part of BW.

- A core part of a new operational concept.
  - Transparent data management for Users.
  - “Virtual file system” for very large data.
  - Improved productivity and schedule effectiveness.
  - Lighter-weight backup.
POWER Platforms

- Production ready, ultra reliable
- Market leader – sustained application performance
- Blades scaling to large memory SMP
- Rich s/w stack (from PERCS)
- Fast interconnect
- Very dense packaging
POWER7 Processor Chip

Core options: 8 (For HPC)

567mm² Technology:
- 45nm lithography, Cu, SOI, eDRAM

Transistors: 1.2B
- Equivalent function of 2.7B
- eDRAM efficiency

Eight processor cores
- 12 execution units per core
- 4 Way SMT per core
- 32 Threads per chip
- 256 KB L2 per core

32MB on chip eDRAM shared L3

Dual DDR3 Memory Controllers
- 100 GB/s Memory bandwidth per chip

Scalability up to 32 Sockets
- 360 GB/s SMP bandwidth/chip
- 20,000 coherent operations in flight

Advanced pre-fetching Data and Instruction

Binary Compatibility with POWER6
PERCS POWER7 Hierarchical Structure

- **POWER7 Chip**
  - 8 Cores

- **POWER7 QCM & Hub Chips**
  - QCM: 4 POWER7 Chips
    - 32 Core SMP Image
  - Hub Chip: One per QCM
    - Interconnect QCM, Nodes, and Super Nodes

- **POWER7 IH Node**
  - 2U Node
  - 8 QCMs
    - 256 Cores

- **POWER7 ‘Super Node’**
  - 4 Drawers / Nodes
    - 1024 Cores

- **Full System**
  - Up to 512 ‘Super Nodes’
  - 512K Cores
Planned POWER7 Compute Node

**Chip Performance:** \( \geq 224 \text{ GFLOPS} \)
- 8 Cores per Chip
- Core Freq: 3.7GHz+
- 4 Floating Point Units (FPU) per core
- 2 FLOPS/Cycle
- 8 cores x 3.7 GHz x 4 FPU/Core x 2 Flops/Cycle

**Node Performance:** \( \geq 7.6 \text{ TF w/ Integrated SMP Fabric} \)
- 256 Cores
- 32 Chips x \( \geq 237 \) GFLOPs

<table>
<thead>
<tr>
<th>POWER7 Compute Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
</tr>
<tr>
<td>Architecture</td>
</tr>
<tr>
<td>Cache</td>
</tr>
<tr>
<td>DDR3 Memory</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>PCI Expansion / Node</td>
</tr>
<tr>
<td>Storage Enclosure</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Ethernet / Node</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Cluster Attach</td>
</tr>
<tr>
<td>Power</td>
</tr>
<tr>
<td>Cooling</td>
</tr>
</tbody>
</table>
P7 IH System Hardware – Node Front View

- D-Link Optical Interface
  Connects to other Super Nodes
- PCIe Interconnect
- L-Link Optical Interface
  Connects 4 Nodes to form Super Node
- Hub Module (8x)
- Memory DIMM’s (64x)
- Memory DIMM’s (64x)
- D-Link Optical Interface
  Connects to other Super Nodes
- PCIe Interconnect
- D-Link Optical Interface
  Connects to other Super Nodes
- D-Link Optical Fiber
- P7 QCM (8x)
- 360VDC Input Connector
- Water Connection

Dimensions: 39”W x 72”D x 83”H
PERCS POWER7 Super Node Description

- 128 P7 Chips + 32 Hub Chips
- 32W O/S Image
- 8 X 32Ws Coherent SMP & I/O Size
- 4 Compute Node (256 Cores / Node)
- 8U in 30” Rack
- 32 SCMs + 32 QCMs
- 1024 Cores
- Up to 8 TB Memory
Advanced Application Development Workbench

Coding & Analysis Tools

Launch & Monitoring Tools

Performance Tuning Tools

Debugging Tools
Portfolio

- **POWER Platforms**
  - Production ready, ultra reliable
  - Market leader – *sustained* application performance
  - Large memory SMP
  - Rich s/w stack (from PERCS)
  - Fast interconnect
  - Very dense packaging

- **Blue Gene**
  - Production ready, ultra reliable
  - Ultra high scaling capability
  - Fast interconnect
  - Highly energy efficient
  - Very dense packaging
  - Strong PEAK $/Mflp price/performance

- **X86 Clusters**
  - Focused on “capacity”, scalability
  - High ISV coverage
  - Strong PEAK $/Mflp price/performance