Massively Multithreaded System Packaging Requirements

Dr. Robert Guernsey, Distinguished Engineer
IBM Next Generation Computing Systems & Technology

10th European System Packaging Workshop
January 30, 2007
A Disruptive Event

Flying around Mt. Rainer

Terrain Rendering Demonstration
Lambroghini Rendering
Cell Processor

- ~250M transistors
- ~235mm²
- Top frequency >4GHz
- 9 cores, 10 threads
- > 256 GFlops (SP) @4GHz
- > 26 GFlops (DP) @4GHz
- Up to 25.6GB/s memory B/W
- Up to 75 GB/s I/O B/W
- Large design investment (time & money)
Massively Multithreaded System Packaging Requirements. ESPW10 1/30/07

**IBM’s Objective: Putting Cell BE to “Work” …**

**Cell BE Broadband Engine & Game Consoles**
- Non-homogeneous coherent multi-Processor
  - Dual-threaded PPC control processor
  - 8 independent SIMD/Vector Accelerators
- ~250M transistors; ~235mm²
- Performance - Top frequency >4GHz
  - > 256 GFLOPs (Single Precision)
  - Up to 25.6GB/s memory B/W
  - Up to 75 GB/s I/O B/W

**Functions Which Benefit from Cell BE**
- Computationally Intense
  - Signal processing, DSP
  - FFT
  - Matrix, Vector mathematics
  - Parallel processing
  - Audio re-sampling
  - Noise generation
  - Pattern matching
  - Security encryption / decryption
  - Encoding / decoding
  - Curve and surface evaluation
  - Game Physics / Physics simulation
- Graphically/Visually Intense
  - Image processing
  - Transform-light
  - Surface subdivision
- Highly Interactive (Real-Time)
  - Game Physics / Physics simulation
  - Voice/Language parsing
  - TCP/IP offload
  - Real time processing
  - Video compression / decompression

**Potential Commercial Applications Areas**
- Medical imaging / Visualization
- Drug discovery
- Petroleum reservoir modeling
- Computational chemistry
- Seismic analysis
- Climate modeling
- Financial modeling/analysis
- Avionics & Targeting systems
- Air traffic control systems
- Radar/Sonar systems
- Training simulation
- Digital Video Surveillance
- Secure communications
- LAN/MAN Routers
- Network processing
- XML and SSL acceleration
- Voice and pattern recognition
- Video conferencing
- Data mining and analysis
- Media server & distribution
- On-Line Game Server

*IBM Focus Areas

**Develop & Deliver a Common, Modular Platform Which Addresses the Computationally Intense, Graphically Intense and Highly Interactive Needs of Emerging, Commercial Workloads**
Announced Sept 12, 2006

- **Cell BE Processor Blade (~600GFLOPS peak)**
  - Dual 3.2GHz Cell BE Processor Configuration
  - 1GB XDRAM (512MB per processor)
  - Blade-mounted 40GB IDE HDD
  - Dual Gigabit Ethernet (GbE) controllers
  - Double-wide blade (uses 2 BladeCenter slots)
  - Infiniband (IB) Option:
    - Qty 0-2 IB 4x Host Channel Adapters
  - 1 yr warranty (upgrades available for purchase)

- **BC Chassis Configuration (~3.5TFLOPS peak)**
  - Standard IBM BladeCenter One
  - Max. 7 Blades per chassis (QS20 - 2 slots each)
  - 2 Gigabit Ethernet switches
  - External IB switches required for IB option

Note: Intermixing Cell Blades with other blades in same chassis is not supported. BladeCenter-H not supported.
### November 2005: TOP10 of the TOP500 list

<table>
<thead>
<tr>
<th>Year</th>
<th>Computer</th>
<th>Measured Tflop/s</th>
<th>Theoretical Peak Tflop/s</th>
<th>Number of Processors</th>
<th>% peak</th>
<th>Flops/Hz</th>
<th>Processor MHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>IBM BlueGene/L, LLNL</td>
<td>280.6</td>
<td>367.00</td>
<td>131072</td>
<td>76</td>
<td>4</td>
<td>700</td>
</tr>
<tr>
<td>2005</td>
<td>IBM BlueGene/L, IBM Yorktown</td>
<td>91.29</td>
<td>114.688</td>
<td>40960</td>
<td>80</td>
<td>4</td>
<td>700</td>
</tr>
<tr>
<td>2005</td>
<td>IBM ASC Purple</td>
<td>63.39</td>
<td>77.824</td>
<td>10240</td>
<td>81</td>
<td>4</td>
<td>1900</td>
</tr>
<tr>
<td>2004</td>
<td>Columbia, SGI Altix 3700, Itanium 2 for NASA</td>
<td>51.87</td>
<td>60.96</td>
<td>10160</td>
<td>84</td>
<td>4</td>
<td>1500</td>
</tr>
<tr>
<td>2005</td>
<td>Dell PowerEdge 1850 Thunderbird, Sandia</td>
<td>38.27</td>
<td>64.512</td>
<td>8000</td>
<td>59</td>
<td>2</td>
<td>3600</td>
</tr>
<tr>
<td>2005</td>
<td>Cray Redstorm, Sandia</td>
<td>36.19</td>
<td>43.52</td>
<td>10880</td>
<td>83</td>
<td>2</td>
<td>2000</td>
</tr>
<tr>
<td>2002</td>
<td>NEC Earth Simulator</td>
<td>35.86</td>
<td>40.96</td>
<td>5120</td>
<td>87.5</td>
<td>16</td>
<td>500</td>
</tr>
<tr>
<td>2005</td>
<td>Marc Nostrum, IBM JS20 PowerPC970</td>
<td>27.91</td>
<td>42.144</td>
<td>4800</td>
<td>66</td>
<td>4</td>
<td>2200</td>
</tr>
<tr>
<td>2005</td>
<td>IBM BlueGene/L, ASTRON</td>
<td>27.45</td>
<td>34.4064</td>
<td>12288</td>
<td>80</td>
<td>4</td>
<td>700</td>
</tr>
<tr>
<td>2005</td>
<td>Cray XT3, ORNL</td>
<td>20.527</td>
<td>24.96</td>
<td>5200</td>
<td>82</td>
<td>2</td>
<td>2400</td>
</tr>
</tbody>
</table>

BlueGene becomes the clear leader in the TOP500 list.
<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>Computer</th>
<th>Processors</th>
<th>Year</th>
<th>R_{max}</th>
<th>R_{peak}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DOE/NNSA/LLNL United States</td>
<td>BlueGene/L - eServer Blue Gene Solution IBM</td>
<td>131072</td>
<td>2005</td>
<td>280600</td>
<td>367000</td>
</tr>
<tr>
<td>2</td>
<td>NNSA/Sandia National Laboratories United States</td>
<td>Red Storm - Sandia/ Cray Red Storm, Opteron 2.4 GHz dual core Cray Inc.</td>
<td>26544</td>
<td>2006</td>
<td>101400</td>
<td>127411</td>
</tr>
<tr>
<td>3</td>
<td>IBM Thomas J. Watson Research Center United States</td>
<td>BGW - eServer Blue Gene Solution IBM</td>
<td>40960</td>
<td>2005</td>
<td>91290</td>
<td>114688</td>
</tr>
<tr>
<td>4</td>
<td>DOE/NNSA/LLNL United States</td>
<td>ASC Purple - eServer pSeries p5 575 1.9 GHz IBM</td>
<td>12208</td>
<td>2006</td>
<td>75760</td>
<td>92781</td>
</tr>
<tr>
<td>5</td>
<td>Barcelona Supercomputing Center Spain</td>
<td>MareNostrum - BladeCenter JS21 Cluster PPC 970, 2.3 GHz, Myrinet IBM</td>
<td>10240</td>
<td>2006</td>
<td>62630</td>
<td>94208</td>
</tr>
<tr>
<td>6</td>
<td>NNSA/Sandia National Laboratories United States</td>
<td>Thunderbird - PowerEdge 1850, 3.6 GHz, Infiniband Dell</td>
<td>9024</td>
<td>2006</td>
<td>53000</td>
<td>64972.8</td>
</tr>
<tr>
<td>7</td>
<td>Commissariat a l'Energie Atomique (CEA) France</td>
<td>Tera-10 - NovaScale 5160, Itanium2 1.6 GHz Quadrics Bull SA</td>
<td>9968</td>
<td>2006</td>
<td>52840</td>
<td>63795.2</td>
</tr>
<tr>
<td>8</td>
<td>NASA/Ames Research Center NAS United States</td>
<td>Columbia - SGI Altix 1.5 GHz, Voltaire Infiniband SGI</td>
<td>10160</td>
<td>2004</td>
<td>51870</td>
<td>60960</td>
</tr>
</tbody>
</table>

Source: www.top500.org
**Dual Node Compute Card**

- Heatsinks designed for 15W (measuring ~13W @1.6V)
- 9 x 256Mb DRAM; 16B interface
- Metral 4000 connector
- 54 mm (2.125”)
- 206 mm (8.125”) wide, 14 layers
BlueGene/L System Buildup

**Rack**
- 32 Node Cards

**Node Card**
- (32 chips 4x4x2)
- 16 compute, 0-2 IO cards

**Compute Card**
- 2 chips, 1x2x1

**Chip**
- 2 processors

**System**
- 64 Racks, 64x32x32

**System Performance**
- 2.8/5.6 TF/s
- 512 GB
- 90/180 GF/s
- 16 GB

**System Capacity**
- 180/360 TF/s
- 32 TB

**System Buildup**
- 2 processors
- 2 chips, 1x2x1
- 32 Node Cards
- (32 chips 4x4x2)
- 16 compute, 0-2 IO cards
- 2.8/5.6 TF/s
- 512 GB
- 90/180 GF/s
- 16 GB
- 180/360 TF/s
- 32 TB

**System Dimensions**
- Rack: 64 Racks, 64x32x32
- Node Card: 32 Node Cards
Blue Gene Interconnection Networks
*Optimized for Parallel Programming and Scalable Management*

**3 Dimensional Torus**
- Interconnects all compute nodes (65,536)
- Virtual cut-through hardware routing
- 1.4Gb/s on all 12 node links (2.1 GB/s per node)
- Communications backbone for computations
- 0.7/1.4 TB/s bisection bandwidth, 67TB/s total bandwidth

**Global Collective Network**
- One-to-all broadcast functionality
- Reduction operations functionality
- 2.8 Gb/s of bandwidth per link; Latency of tree traversal 2.5 µs
- ~23TB/s total binary tree bandwidth (64k machine)
- Interconnects all compute and I/O nodes (1024)

**Low Latency Global Barrier and Interrupt**
- Round trip latency 1.3 µs

**Control Network**
- Boot, monitoring and diagnostics

**Ethernet**
- Incorporated into every node ASIC
- Active in the I/O nodes (1:64)
- All external comm. (file I/O, control, user interaction, etc.)
BlueGene/L Takes Up Little Space
A Fundamental Disruption

Highest frequency can give lowest performance

- Highest frequency
- Lowest frequency

Number of Processor Cores

Relative Performance

Fixed chip size
Elements of System Performance

Customer Perceived System Performance \( \approx \frac{(N \times \text{Frequency})}{(\text{CPI} \times \text{PathLength})} \)

- \( N = \) number of computing elements or threads.
- Frequency (F) = megahertz
- CPI = cycles per instruction for a computing element.
- PathLength (PL) = software pathlength per function, measured in instructions

- *Frequency is no longer a useful "lever".*
- *Lower CPI takes too many circuits.*
- *Must increase N while maintaining or decreasing Pathlength*
System Performance Stack

IBM Systems & Technology Group

S/W Development Tools

Middleware

Operating System

Hypervisor

SMP System Structure

Fabric, switches, busses, memory system, protocols, ...

Compilers

Number of Computing Elements

Pathlength

Microprocessor Core

Microarchitecture, logic circuits, design methodology

Cache

Cache levels, granularity, latency, throughput

Interconnect

I/O

Package

Buses, Memory Bandwidth, Memory Size & Latency, Cooling

Semiconductors

Device, process, interconnect

Cycles / Instruction

Frequency
System Performance Stack: Applications / Workloads.

- **Application**
- **Middleware**
  - **S/W Development Tools**
  - **Operating System**
  - **Hypervisor**
- **SMP System Structure**
  - Fabric, switches, busses, memory system, protocols, ...
- **Compilers**
- **Microprocessor Core**
  - Microarchitecture, logic circuits, design methodology
- **Cache**
  - Cache levels, granularity, latency, throughput
- **Interconnect**
- **Package**
  - Buses, Memory Bandwidth, Memory Size & Latency, Cooling
- **Semiconductors**
  - Device, process, interconnect

**Pathlength**

**Cycles / Instruction**

**Frequency**

**Number of Computing Elements**
New workloads and environments emerge and today’s workloads will evolve.

Today’s Applications, Workloads, & Architectures
Near-Term Systems

Emerging Applications, Workloads, & Architectures
- Event Based Systems
- Real-Time Analytics
- Rich Media
- XML Data
- Secure Computing
- Sensor Infrastructure
- Next Gen UI
- SOA Infrastructure
- Scale Out Infrastructure

Next Generation Systems Era: Applications / Workloads.
**System Performance Stack: System Hardware Design.**

<table>
<thead>
<tr>
<th><strong>Application</strong></th>
<th><strong>Middleware</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S/W Development Tools</strong></td>
<td><strong>Operating System</strong></td>
</tr>
<tr>
<td><strong>Middleware</strong></td>
<td><strong>Hypervisor</strong></td>
</tr>
<tr>
<td><strong>SMP System Structure</strong></td>
<td><strong>Compilers</strong></td>
</tr>
<tr>
<td>Fabric, switches, busses, memory system, protocols, ...</td>
<td><strong>Microprocessor Core</strong></td>
</tr>
<tr>
<td><strong>Microarchitecture, logic circuits, design methodology</strong></td>
<td><strong>Cache</strong></td>
</tr>
<tr>
<td><strong>Cache levels, granularity, latency, throughput</strong></td>
<td><strong>Interconnect</strong></td>
</tr>
<tr>
<td><strong>Package</strong></td>
<td><strong>I/O</strong></td>
</tr>
<tr>
<td>Buses, Memory Bandwidth, Memory Size &amp; Latency, Cooling</td>
<td><strong>Number of Computing Elements</strong></td>
</tr>
<tr>
<td><strong>Semiconductors</strong></td>
<td><strong>Pathlength</strong></td>
</tr>
<tr>
<td>Device, process, interconnect</td>
<td><strong>Cycles / Instruction</strong></td>
</tr>
<tr>
<td><strong>Frequency</strong></td>
<td></td>
</tr>
</tbody>
</table>
New Focus for System Hardware Design

- **Significantly higher focus on performance per watt.**
  New tradeoffs between system and single thread performance

- **Increased focus on performance density**

- **Emphasis on more threads of less performance.**
  See this in Niagara now.

- **System performance dominated by factors other than processors:**
  Memory, interconnect, and disk.

- **Software must help provide performance growth.**

- **Emphasis on system management and virtualization.**

- **Modularity and flexibility in hardware design.**
Threads per Socket
Estimates from IEEE Computer Elements Vail Workshop

- 2007: 64
  Niagara

- 2009: 100's

- 2012: ~1000
**On-Line Transaction Processing Example***

**Example Socket Performance**  (Rough sizings of "what-if" systems)

<table>
<thead>
<tr>
<th>Unconstrained</th>
<th>Constrained by</th>
<th>Memory Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 K TPM-C</td>
<td>90 K TPMC</td>
<td>32 GB</td>
</tr>
<tr>
<td>400</td>
<td>100</td>
<td>32</td>
</tr>
<tr>
<td>400</td>
<td>300</td>
<td>128</td>
</tr>
<tr>
<td>800</td>
<td>400</td>
<td>128</td>
</tr>
<tr>
<td>800</td>
<td>600</td>
<td>256</td>
</tr>
<tr>
<td>2100</td>
<td>700</td>
<td>256</td>
</tr>
</tbody>
</table>

Memory bandwidth must track memory size.

* SAP and ERP workloads are less demanding of memory.
Memory Bandwidth Must Track Memory Size

Desired Socket Memory Bandwidth (GB/s)

- 2006 / 2007: 25 - 50 GB/s
- 2008 / 2009: 100 - 200 GB/s

Must achieve at acceptable cost. That makes it hard!
Future System Vision: Integration and Modularity

- Increased reliance on modular components, with variety of components
  - Configurability emerging as possible replacement for traditional design
- Software enablement becoming ever more important in exploitation of modularity and configurability (compilers, operating systems, middleware, etc.)
  - **Must co-design HW and SW systems**
### System Performance Stack: Cooling

#### S/W Development Tools
- Application
- Middleware
- S/W Development

#### Middleware
- Operating System
- Hypervisor

#### SMP System Structure
Fabric, switches, busses, memory system, protocols, ...

#### Compilers
- SMP System Structure
  - Fabric, switches, busses, memory system, protocols, ...

#### Microprocessor Core
- Microarchitecture, logic circuits, design methodology

#### Cache
- Cache levels, granularity, latency, throughput

#### Interconnect

#### I/O

#### Package
Buses, Memory Bandwidth, Memory Size & Latency, Cooling

#### Semiconductors
Device, process, interconnect

---

**Pathlength**

**Cycles / Instruction**

**Frequency**

**Number of Computing Elements**
Technology Longer Term: Inboard Cooling

- **Integrate air-conditioner and computer in the frame.**
  - Air circulates internally with none leaving the frame.
  - Simplifies cooling engineering.
  - Greatly reduces acoustic noise in room.

- **Enables lower temperature operation.**
  - Can maintain dry environment
    - No condensation -- frost-free.
  - Quasar low temperature experiments this year near -20C.

Advanced Server Hardware Systems 80kW Frame Inboard Cooling Demonstration at Yorktown Labs.
Paul Coteus, Shawn Hall, John Karidis, Alphonso Lanzetta, Rick Rand, & Shurong tian.

HEX = Air to Water Heat Exchanger
Enterprise server performance per socket is poised to grow dramatically through the use of massive multi-threading.

This will bring significant packaging challenges in --

- Dense packaging of large memory.
- High socket memory bandwidth at affordable cost.
- Innovative cooling for dense large systems.

Enjoy!!!