Blue Gene/Q – Hardware
Blue Gene design goals

- System-on-Chip (SoC) design
  - Processor comprises both processing cores and network
- Optimal performance / watt ratio
- Small footprint
- Transparent high-speed reliable network
- Easy programming based on standard message passing interface (MPI)
- Extreme scalability (> 1,5Mi cores)
- High reliability
**Blue Gene/Q design**

3. **Compute Card (“Node”):**
   - One BQC Module (1x1x1x1x1),
   - 16 GB DDR3 Memory

4. **Node Card (“Node Board”):**
   - 32 Compute Cards (2x2x2x2x2),
   - Optical Modules, BQL Link Chips, Torus

5a. **Midplane:**
    - 16 Node Cards (4x4x4x4x2)

5b. **I/O drawer (1, 2 or 4 per rack):**
   - *8 I/O cards @ 16 GB,
   - *8 PCIe gen2 x8 slots (IB, 10GbE)

6. **Rack: 2 Midplanes (4x4x4x8x2)**
   - 1, 2 or 4 I/O Drawer

7. **System:**
   - e.g. 8 racks (8x8x8x8x2) = 1.7 PF/s
   - e.g. 28 racks (8x28x8x8x2) = 5.9 PF/s
   - e.g. 96 racks (16x12x16x16x2) = 20 PF/s

Source: IBM
Blue Gene/Q node card

Fiber-Optic Ribbons (36X, 12 Fibers each)

Compute Card with One Node (32X)

Water Hoses

48-Fiber Connectors

Redundant, Hot-Pluggable Power-Supply Assemblies

Source: IBM
Blue Gene/Q compute card

Source: Top500.org
Blue Gene/Q: Chip tomography

16+1+1 processing units

Two memory controller

L2 cache + crossbar switch

On-chip network

Source: IBM
Blue Gene/Q chip architecture

- 16+1 core SMP @ 1.6 GHz
  - Each core 4-way hardware threaded
  - 2-way concurrent issue
- Transactional memory and thread level speculation
- Quad floating point unit on each core
  - 204.8 GF peak node
- 563 GB/s bisection bandwidth to shared L2
- 32 MB shared L2 cache
- 42.6 GB/s DDR3 bandwidth (1.333 GHz DDR3)
  - (2 channels each with chip kill protection)
- 10 intra-rack inter-processor links each at 2.0GB/s (5D-Torus)
- one I/O link at 2.0 GB/s
- 16 GB memory/node
- ~60 watts max chip power
Blue Gene/Q: A2 processor core

- Simple core
  - designed for excellent power efficiency and small footprint
- Embedded 64bit PowerPC compliant
- 4 SMT threads typically get a high level of utilization on shared resources
  - Full register set for every thread
- 1.6 GHz @ 0.74V
- AXU port allows unique BG/Q floating point unit
- One AXU (FPU) and one other instruction issue per cycle
- In-order execution
Blue Gene/Q: Multithreading

- Four threads issuing to two pipelines
  - Impact of memory access latency reduced
- Issue
  - Up to two instructions issued per cycle
    - One Integer/Load/Store/Control instruction issue per cycle
    - One FPU instruction issue per cycle
  - At most one instruction issued per thread
- Flush
  - Pipeline is not stalled on conflict
  - Instead,
    - Instructions of conflicting thread are invalidated
    - Thread is restarted at conflicting instruction
  - Guarantees progress of other threads
Quad floating Point eXtension unit (QPX)

- 4 double precision pipelines (64bit):
  - scalar FPU
  - 4-wide FPU SIMD
  - 2-wide complex arithmetic SIMD
- 32 x 4 x 256 bit registers
- Instruction extensions to PowerISA
- 8 concurrent floating point ops (FMA) + load + store
- Permute instructions to reorganize vector data
- Supports a multitude of data alignments
- Peak performance 4FMA / cycle
  - 12.8 GFlops @ 1.6 GHz
Memory hierarchy

- **L2 cache**
  - 32 MBytes
  - Organised in 16 slices
  - 16-way associative

- **External memory**
  - 16 GBytes DDR3
  - 2 memory controllers

- **L1 cache**
  - 16 kB
  - 8-way associative
Blue Gene/Q: Advanced processor features

L1 pre-fetching engines
- Stream pre-fetching
  - 1 engine per core
- “Perfect” (list) pre-fetching
  - 4 engines per core
  - Hardware memorizes access sequence

Multi-versioning L2 cache
- Cache can track state changes caused by speculative threads
- At the end of speculative code: invalidate or commit and/or react with software notification
- Use: transactional memory, thread-level speculation
## Blue Gene/Q ↔ Blue Gene/P

<table>
<thead>
<tr>
<th>Property</th>
<th>Blue Gene/Q</th>
<th>Blue Gene/P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Node Properties</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Node Processors</td>
<td>16*4 PowerPC® A2</td>
<td>4* 450 PowerPC®</td>
</tr>
<tr>
<td>Processor Frequency</td>
<td>1.6GHz</td>
<td>0.85GHz</td>
</tr>
<tr>
<td>Coherency</td>
<td>SMP</td>
<td>SMP</td>
</tr>
<tr>
<td>Cache size (shared)</td>
<td>32MB</td>
<td>8MB</td>
</tr>
<tr>
<td>Main Store</td>
<td>16GB</td>
<td>2GB</td>
</tr>
<tr>
<td>Main Store Bandwidth</td>
<td>42.6GB/s</td>
<td>13.6 GB/s</td>
</tr>
<tr>
<td>Peak Performance</td>
<td>204.8GF/node</td>
<td>13.9 GF/node</td>
</tr>
<tr>
<td><strong>Torus Network</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topology</td>
<td>5D</td>
<td>3D</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>10<em>2</em>2GB/s=40GB/s</td>
<td>6<em>2</em>425MB/s=5.1GB/s</td>
</tr>
<tr>
<td>Hardware Latency</td>
<td>40ns (32B packet)</td>
<td>100ns (32B packet)</td>
</tr>
<tr>
<td>Nearest Neighbour</td>
<td>300?ns (512B packet)</td>
<td>800ns (256B packet)</td>
</tr>
<tr>
<td>Hardware Latency</td>
<td>2.6µs (31 hops)</td>
<td>3.2µs (64 hops)</td>
</tr>
<tr>
<td>Worst Case</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Tree Network</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bandwidth</td>
<td>6.5µs</td>
<td>2*0.85GB/s=1.7GB/s</td>
</tr>
<tr>
<td>Hardware Latency</td>
<td></td>
<td>3.5µs</td>
</tr>
<tr>
<td>worst case</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>System Properties</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area</td>
<td>~20m²</td>
<td>160m²</td>
</tr>
<tr>
<td>Peak Performance</td>
<td>1.7PF</td>
<td>1.008PF</td>
</tr>
<tr>
<td>Total Power</td>
<td>~400kW</td>
<td>~2.3MW</td>
</tr>
</tbody>
</table>
Execution Modes in BG/Q

- **64 MPI tasks**
  - 1 Thread / Task
  - P0 T0
  - P1 T0
  - Px T0
  - P63 T0

- **2,4,8,16,32 MPI tasks**
  - 32,16,8,4,2 Threads
  - P0 T0
  - Py T0
  - Tx
  - Tx

- **1 MPI task**
  - 1-64 Threads / Task
  - P0 T0
  - T1 T0
  - Tx
  - T63
Blue Gene/Q chip: the 17th Core

RAS Event handling and interrupt off-load
  - Reduce O/S noise and jitter
  - Core-to-Core interrupts when necessary

CIO Client Interface
  - Asynchronous I/O completion hand-off
  - Responsive CIO application control client

Application Agents: privileged application processing
  - Messaging assist, e.g., MPI pacing thread
  - Performance and trace helpers
Blue Gene/Q: New Network architecture

- 11 bi-directional chip-to-chip links
  - 2 GB/s bandwidth, about 40 ns latency
- 5-dimensional torus topology
  - Dimension E limited to length 2
- Why $d$-dimensional torus with large $d'$?
  - High bi-section bandwidth
  - Flexible partitioning in lower dimensions
- Deterministic/dynamic routing support
- Collective and barrier networks embedded in 5-D torus network
  - Floating point addition support in collective network
  - 11th port for auto-routing to IO fabric

Source: IBM
## Blue Gene/Q: System configurations

- **BG/Q Nodes form a 5D torus**
  - Nodecards: 2x2x2x2x2
  - Midplanes: 4x4x4x4x4x2
    - 4D are cabled to other midplanes
  - 5th dimension: extent 2 (stays within nodecard)
  - 6th dimension is cpu # within the node
  - Dim. labels: ABCDE T

- **Different floor shapes (Rows x Cols) for a given number of racks may correspond to the same, or to different torus shapes**

- **This list is not complete; other configs are possible...**
  - up to 16x16 = 256 racks

### System configurations table

<table>
<thead>
<tr>
<th>Racks</th>
<th>Rows</th>
<th>Col.</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>Torus size, in nodes</th>
<th>Torus size, in midplanes</th>
</tr>
</thead>
<tbody>
<tr>
<td>mid</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>512</td>
<td>8.192</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>12</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>6</td>
<td>12</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td>8</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>12</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>2</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>20</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>6</td>
<td>12</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>8</td>
<td>16</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td>20</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>24</td>
<td>6</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td>16</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>24</td>
<td>2</td>
<td>12</td>
<td>8</td>
<td>8</td>
<td>12</td>
<td>16</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>28</td>
<td>7</td>
<td>4</td>
<td>8</td>
<td>28</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>32</td>
<td>8</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>16</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>32</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>8</td>
<td>16</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>40</td>
<td>5</td>
<td>8</td>
<td>20</td>
<td>16</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>48</td>
<td>6</td>
<td>8</td>
<td>16</td>
<td>12</td>
<td>16</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>48</td>
<td>4</td>
<td>12</td>
<td>8</td>
<td>16</td>
<td>12</td>
<td>16</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>48</td>
<td>3</td>
<td>16</td>
<td>8</td>
<td>12</td>
<td>16</td>
<td>16</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>56</td>
<td>7</td>
<td>8</td>
<td>28</td>
<td>16</td>
<td>16</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>64</td>
<td>8</td>
<td>8</td>
<td>16</td>
<td>8</td>
<td>16</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>64</td>
<td>4</td>
<td>16</td>
<td>8</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>72</td>
<td>9</td>
<td>8</td>
<td>12</td>
<td>12</td>
<td>16</td>
<td>16</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>80</td>
<td>10</td>
<td>8</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>96</td>
<td>12</td>
<td>8</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>96</td>
<td>6</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Torus node to MPI mapping

- Block's physical network topology: 6 dimensional $A \times B \times C \times D \times E \times T$
  - local on each node; shared memory communication (fast); $T=0..63$
  - $A,B,C,D,E$ depend on size of block, e.g. $2 \times 2 \times 4 \times 4 \times 2$ for quarter midplane
- Processes need to be placed to minimize maximum load on network-links
  - take advantage of logical decomposition of work
  - take advantage of link in $E$ direction: double bandwidth available
  - e.g. domain decomposition: find best match of physical and logical dimensions
- Three options to define mapping:
  - By permutation of physical dimensions (--mapping DCTEBA)
  - Use MPI_Cart_Create() (to date: only reordering of dimensions)
  - By a file with a line for each process, specifying its physical position (--mapping <filename>)
Linktest: Blue Gene torus link bandwidth tester

- All-to-all ping-pong test
- Bandwidth distribution
  - *Intra-node communication*
  - *Communication via link A, B, C, D*
  - *Communication via link E*
Blue Gene/Q: I/O architecture

- **I/O Network to/from Compute rack**
  - 2 links (4GB/s in 4GB/s out) feed an I/O PCI-e port
  - Every node card has up to 4 ports (8 links)
  - Typical configurations
    - 8 ports (32GB/s/rack)
    - 16 ports (64 GB/s/rack)
    - 32 ports (128 GB/s/rack)
  - Extreme configuration 128 ports (512 GB/s/rack)

- **I/O Drawers**
  - 8 I/O nodes/drawer with 8 ports (16 links) to compute rack
  - 8 PCI-e gen2 x8 slots (32 GB/s aggregate)
  - 4 I/O drawers per compute rack
  - Optional installation of I/O drawers in external racks for extreme bandwidth configurations
JUQUEEN Configuration (01/2013)

28 Racks Blue Gene/Q
- 28672 Compute Nodes (16 cores, 16 GB memory)
- 458752 cores / 1.8M threads
- 5.88 PFlop/s peak performance
- 248 I/O nodes (10GigE) ← (1x32 + 27x8)
- 2.3 MW power consumption (10-80 kW per rack)

4 Frontend Nodes (user login) + 2 Service Nodes (system, database)
- IBM p7 740, 8 cores (3.55 GHz), 128 GB memory
- local storage device DS5020 (16 TB)
JuQueen Environment

BG/Q

Control-System

runjob

ssh

Fileserver JUST

GPFS

Service Node

Service Node

FrontEnd

FrontEnd

RAB

DB2

05.07.2012