Igor Kolesnik & Mark Robinson
System Architecture
Project 1 Lecture Notes
21 October 2014

General History

The predecessor to the Blue Gene supercomputers was the QCDSP(Quantum Chromodynamics on Digital Signal Processors). The QCDSP was worked on from 1982-1998. It used thousands of cheap, off-the-shelf processors. It had a low cost and power consumption. The processors were connected via a Torus network.

Next was the QCDOC which was a Quantum Chromodynamics Digital On a Chip. It was an evolution of the QCDSP that used PPC 440 cores and a 6D Torus network.

At the end of 1999 IBM announced a $100 million research initiative to build a massively parallel computer to be used to study sciences such as protein folding. IBM wanted a way to study protein folding while exploring the possibilities of a massively parallel machine at a reasonable cost. In 2004 the first Blue Gene/L was completed with 16 racks each holding 1024 nodes. This computer had a performance of 70.72 TFLOPS. It held the first position in the TOP500 for 3.5 years. From 2004-2007 the Blue Gene/L installation at LLNL expanded to have 104 racks and had a performance of 478 TFLOPS. In 2005 the Blue Gene/L won the Gordon Bell Prize.

In 2007 IBM unveiled the Blue Gene/P, an evolution of the BG/L. The BG/P had a performance of 13.6 GFLOPS. It was able to achieve this performance at only 371 MFLOPS/W putting it near the top of the GREEN500 for 2 years.

The next evolution came in 2011 with the Blue Gene/Q with a peak performance of 20 Petaflops. The initial 4 rack system, similar to the original BG/L, performed at 677.1 TFLOPS. By 2012 as the BG/Q installations expanded it took 1st place in the TOP500, GRAPH500, and the GREEN500.

It is named Blue Gene because it was intended for the study of protein folding and gene development.

Goals of the Blue Gene:

- Achieve low power consumption by using less powerful processors.
- Achieve high performance via a large number of nodes (scalable in increments of up to at least 65,536).
- Low latency between nodes and stacks by using a torus network.
- System-on-a-chip design by having a lightweight OS per node for minimum system overhead.

- High Usability by using accessible languages such as POSIX and openMP, this made programming for BG/L easier as well as giving it a broader application front.
- Study protein folding and the possibilities of supercomputing.

Challenges

- Finding the right core that was both power efficient and decently powerful.
- How to get high speed communication between so many nodes. They solved this by using torus networks between the nodes and between the stacks.
- High bandwidth to/from cache and to external memory.

Differences from regular computers

BG/L leans towards MPI Co-Processor	BG/L leans away from General Purpose Computer
Space-shared nodes.	Time-shared nodes.
Use only real memory.	Virtual memory to disk.
No asynchronous OS activities.	OS services.
Distributed memory.	Shared memory.
No internal state between applications. [Helps performance and functional reproducibility.]	Built-in filesystem.
Reguires General Purpose Computer as Host.	

What Blue Gene computers have been used for:

- Weather Forecasts
- Disaster Management
- Precision Agriculture
- Protein Folding
- Chess
- Simulate part of the human brain
- Refining Medical Drugs

Blue Gene/L

Node Specification:

Compute Cards:	16/Node
CPUs:	64 PPC 440/Node (32-bit instruction set)
Rmax:	5.6 GF/s

Machine Specification:

# of cores:	212,992
# of nodes:	106,496
# of racks:	104
Rmax:	478 TF (A system's Rmax score describes its maximal achieved performance)
Rpeak:	596 TF (The Rpeak score describes its theoretical peak performance)
Power:	2,329.00 kW
MF/W:	210 MF/W

Summary:
Each Blue Gene/L node was attached to three parallel communications networks: a 3D toroidal network for peer-to-peer communication between compute nodes, a collective network for collective communication (broadcasts and reduce operations), and a global interrupt network for fast barriers. The I/O nodes, which run the Linux operating system, provided communication to storage and external hosts via an Ethernet network. The I/O nodes handled filesystem operations on behalf of the compute nodes. Finally, a separate and private Ethernet network provided access to any node for configuration, booting and diagnostics. To allow multiple programs to run concurrently, a Blue Gene/L system could be partitioned into electronically isolated sets of nodes. The number of nodes in a partition had to be a positive integer power of 2, with at least 25 - 32 nodes. To run a program on Blue Gene/L, a partition of the computer was first to be reserved. The program was then loaded and run on all the nodes within the partition, and no other program could access nodes within the partition while it was in use. Upon completion, the partition nodes were released for future programs to use.

Blue Gene/L compute nodes used a minimal operating system supporting a single user program. Only a subset of POSIX calls was supported, and only one process could run at a time per node in co-processor mode, or one process per CPU in virtual mode. Programmers needed to implement green threads in order to simulate local concurrency. Application development was usually performed in C, C++, or Fortran using MPI for communication. However, some scripting languages such as Ruby and Python have been ported to the compute nodes.

source:http://en.wikipedia.org/wiki/Blue_Gene

Blue Gene/P
Node Specification:

Compute Cards:	32/Node
CPUs:	128 PPC 450/Node (32-bit instruction set)
Rmax:	13.6 GF/s

Machine Specification:

# of cores:	294,912
# of nodes:	2,304
# of racks:	72

Rmax: 825.5 TF (A system's Rmax score describes its maximal achieved performance)
Rpeak: 10,002.7 TF (The Rpeak score describes its theoretical peak performance)
Power: 2,268.00 kW
MF/W: 357 MF/W
Bandwidth: 5.1 GB/s per node (Global Collectivity)
Latency (MPI): 5 microsecond

Summary:
The Blue Gene/P supercomputer is a scalable, distributed-memory system consisting of up to 262,144 nodes. Each node is built around a single compute ASIC with 2GB or 4GB of external DDR2 DRAM. The compute ASIC is a highly integrated system-on-a-chip chip multiprocessor. It contains four PPC 450 embedded processor cores, each with private, highly-associative, 32KB first-level instruction and data caches. Each core is coupled to a dual-pipeline SIMD floating-point unit and to a small, private, second-level cache whose principal responsibility is to prefetch streams of data. In addition, the chip integrates an 8MB, shared third-level cache, two memory controllers, five network controllers, and a performance monitor.

The PPC 450 microprocessor is a high-performance, out-of-order industry-standard PPC processor originally targeted at high-end embedded systems. The processor supports 2-way superscalar instruction execution with a seven stage pipelined microarchitecture. The processor cores include 32KB first-level instruction and data caches organized as 16 associative sets with 64 ways per set.

A dual-pipeline, SIMD floating point unit is attached to each processor core. The floating point unit can execute two fused multiply-add instructions per cycle for a peak floating point performance of 13.6 GFLOPS/node. The floating point unit pairs two floating-point register files and two execution pipes. The primary and secondary register files are independently addressable, but they can be jointly accessed by SIMD instructions. SIMD execution exploits the data-level parallelism often present in high-performance computing workloads to reduce the number of instructions that must be executed, while increasing the number of operations completed.

Likes it predecessor, Blue Gene/P provides five dedicated communication networks: the torus network, the collective network, the barrier network, 10Gb/s Ethernet, and IEEE 1149.1 (JTAG). The network interfaces are integrated on the same chip as the processing units. The main network is the torus, which provides high performance data communication to nearest neighbor nodes in a 3D configuration with low latency and high throughput. The collective network supports efficient collective operations, such as broadcast and reduction, and serves as the I/O interconnect.

source:https://www.cs.utah.edu/hpca08/papers/Industrial_1_Salapura.pdf

Blue Gene/Q

Node Specification:

Compute Cards:	32/Node
CPUs:	32 PPC A2/Node (64-bit instruction set)
Cores:	576/Node
Rmax:	204 GF/Node

Machine Specification:

# of cores:	1,572,864
# of nodes:	98,304
# of racks:	96
Rmax:	17,173.2 TF (A system's Rmax score describes its maximal achieved performance)
Rpeak:	20,132.7 TF (The Rpeak score describes its theoretical peak performance)
Power:	7,890.00 kW
MF/W:	2,069 MF/W

Summary:
The BG/Q processor is an 18-core CPU and only 16 cores are used to perform mathematical calculations.The 17th core is used for node control tasks such as offloading I/O operations which ”talk” to Linux running on the I/O node. (Note, the I/O nodes are separate from the compute nodes; so, Linux is not actually running on the 17th core.) The 18th core is a spare core which is used when there are corrupt cores on the chip. The corrupt core is swapped and software transparent. The processor uses PowerPC A2 cores, operating at a moderate clock frequency of 1.6 GHz and consuming a modest 55 watts at peak. The Blue Gene line has always been known for throughput and energy efficiency, and so emphasizes the A2 architecture. Despite the low power consumption, the chip delivers a very respectable 204 Gflops. This is due to a combination of features like the tight core count, support for up to four threads per core, and a quad floating-point unit.

The quad double-precision Floating Point Unit (FPU) (available on each core) has four pipelines which can be used to execute scalar floating point instructions, four SIMD instructions, or two complex arithmetic SIMD instructions. These instructions are extensions of the Power instruction set. The FPU has a six-stage pipeline and has permutation instructions to reorganize vector data on the fly; it can perform a maximum of eight concurrent floating point operations per clock cycle plus a load and a store. Every BG/Q processor has two DDR3 memory controllers, each interfacing with eight slices of the L2 cache to handle their cache misses (one controllers for each half of the 16 cores on the chip).

BG/Q peer-to-peer communication between compute nodes is performed over a 5-dimensional (5D) Torus network (note that BG/L and P feature a 3D Torus). Each node has 11 links and each link can transmit data at 2 GB/s and simultaneously receive at 2 GB/s for a total bandwidth of 44 GB/s. While 10 links connect the compute nodes, the 11th link provides connection to the I/O

nodes. By default a custom lightweight operating system called Compute Node Kernel (CNK) is loaded on the compute nodes while I/O nodes run Linux OS. The I/O architecture is significantly different from previous BG generations since it is separated from the compute nodes and moved to independent I/O racks.

source:http://icl.cs.utk.edu/news_pub/submissions/PAPI-for-BGQ_mccraw.pdf

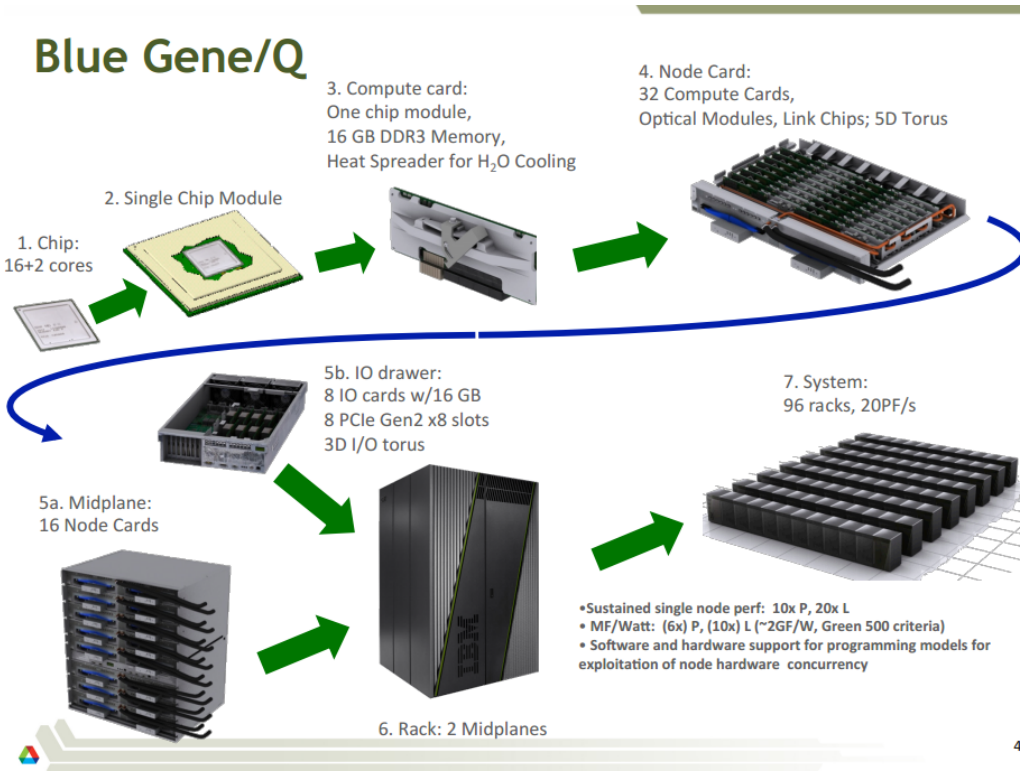

Figure 1
Source:
http://www.bgconsortium.org/sites/bgconsortium.drupalgardens.com/files/BGQ-arch.pdf

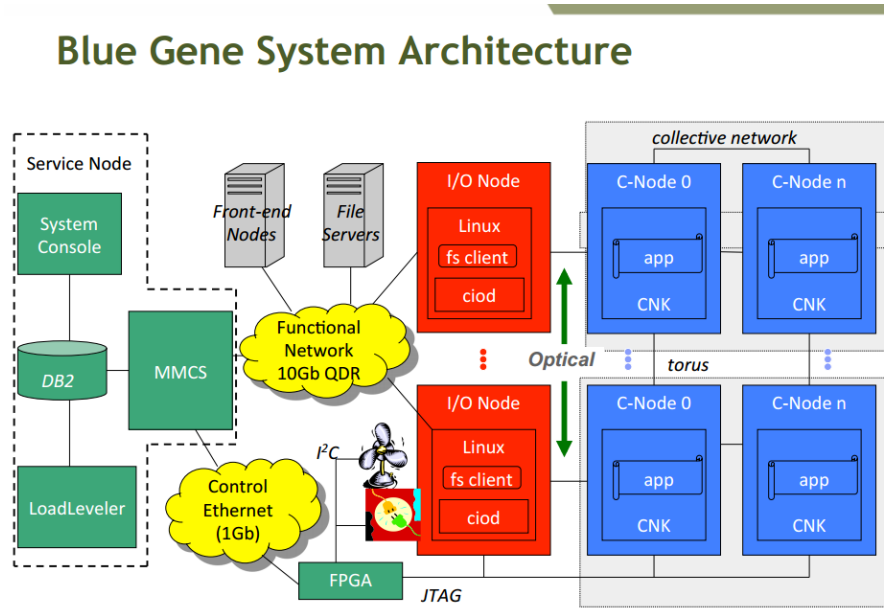

Figure 2
Source:
http://www.bgconsortium.org/sites/bgconsortium.drupalgardens.com/files/BGQ-arch.pdf

Sources:

http://www.slideshare.net/msramakrishna/documents-of-blue-genel

http://www.nersc.gov/assets/Uploads/IBM-Blue-Gene-Architecture.10-10-11.pdf

http://pic.dhe.ibm.com/infocenter/compbg/v121v141/index.jsp?topic=%2Fcom.ibm.xlf141.bg.doc%2Fgetstart%2Fbg_arch.html

https://www.alcf.anl.gov/files/IBM_BGQ_Architecture_0.pdf

http://www-zeuthen.desy.de/technisches_seminar/texte/bluegenel.pdf

http://www.princeton.edu/~achaney/tmve/wiki100k/docs/Blue_Gene.html

https://asc.llnl.gov/computing_resources/bluegenel/configuration.html

http://www.training.prace-ri.eu/uploads/tx_pracetmo/BG-Q-_Vezolle.pdf

http://www.bgconsortium.org/sites/bgconsortium.drupalgardens.com/files/BGQ-arch.pdf

https://computing.llnl.gov/tutorials/bgq/#Evolution