Early Evaluation of IBM BlueGene/P

Tyler Allen
Clemson University

September 28, 2015

Outline

Today we will:

- Discuss IBM's BlueGene/P Architecture
- Discuss the specifications of the hardware
- Examine benchmarks from a BlueGene/P implementation

Early Evaluation

- Goal: Evaluate performance of architecture
- Several metrics:
  - Hardware and Topology Specifications
  - Microbenchmarks and Kernels
  - Target Applications

BlueGene/P Architecture

- Developed by IBM
- Second Generation Architecture
- Successor to BlueGene/L
- Early Evaluation of IBM BlueGene/P by S. Alam et al. in 2008
- Authors from Oak Ridge National Laboratory

System on Chip

- Four PowerPC 450 cores at 850 MHz
- 3.4 GFlop/s per core
- On-chip routing and network protocols
- Low power consumption: 1.8 watts per GFlop/s

Nodes

- 2GB shared RAM per node, soldered
- 13.6 GFlop/s per node
- Connections to 5 networks
- 32K L1 Cache
- 14 Stream Prefetching L2
- 8MB Shared L3

Racks

- 4096 cores per rack
- Standard cooling
- Cores per Rack far exceeds prior architectures
- Possible because of low power consumption
- 72 Racks give 1 PFLOP/s

Racks

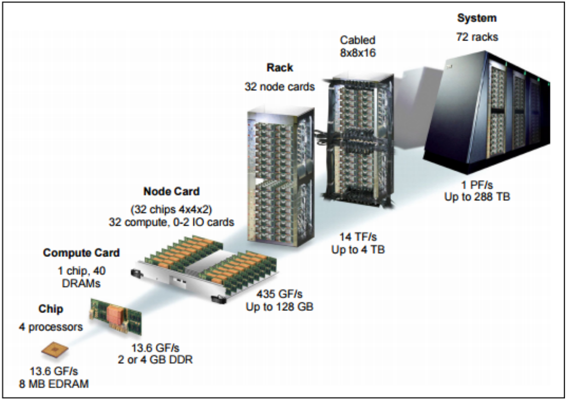


Figure: Image from IBM Redbook (2)

Topology

- 3-D Torus Topology
- Global Collective Tree (Global Broadcasts)
- 10 Gigabit Ethernet (IO)
  - IO requests route through Global Collective Tree
- Global interrupt network
- JTAG control network

Comparison to BlueGene/L

Feature	BlueGene/L	BlueGene/P
Node		
Cores per node	2	4
Core Clock Speed (MHz)	700	850
Cache Coherence	Software	Hardware
L1 Cache / Private per core	32K	32K
L2 Cache / Private per core	14 stream prefetching	14 stream prefetching
L3 Cache / Shared	4 MB Shared	8 MB Shared
Memory per Node (GB)	0.5 - 1	2
Main Memory Bandwidth (GB/s)	5.6	13.6
Peak Performance (GFlop/s per node)	5.6	13.6
Interconnects		
Torus Injection Bandwidth (GB/s)	2.1	5.1
Tree Bandwidth (MB/s)	700	1700

Node Hardware

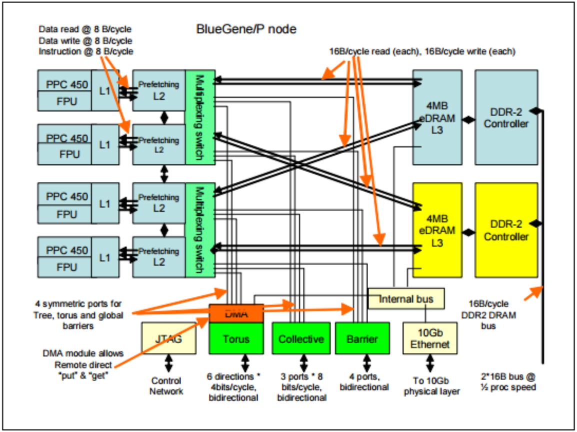


Figure: Image from IBM Redbook (2)

Node Modes

- Nodes can be used in different modes:
  - Symmetric Multiprocessor mode - 1 MPI Task, 4 threads each
  - Dual Node mode - 2 MPI Tasks, 2 threads each
  - Virtual Node mode - 4 MPI Tasks, 1 thread each

Benchmarks Overview

- Several types of benchmarks
  - HPCC Challenge Kernels - single/parallel performance
  - Halo Benchmark - Network and Communication
  - MPI Collective - MPI performance
  - Target Scientific Applications - Architecture Purpose

Benchmarks

We will look at an implementation of BG/P at Argonne National lab, Intrepid.

- Intrepid is IBM standard configuration.
- Intrepid has 40 racks.
- We will compare Intrepid to the Cray XT system at ORNL.

Cray XT Comparison

Feature	BlueGene/L	BlueGene/P	Cray XT3	Cray XT4/DC	Cray XT4/QC
Node					
Cores per node	2	4	2	2	4
Core Clock Speed (MHz)	700	850	2600	2600	2100
Cache Coherence	Software	Hardware	Hardware	Hardware	Hardware
L1 Cache / Private per core	32K	32K	64K	64K	64K
L2 Cache / Private per core	14 stream prefetching	14 stream prefetching	1M	1M	512K
L3 Cache / Shared	4 MB Shared	8 MB Shared	n/a	n/a	2 MB Shared
Memory per Node (GB)	0.5 - 1	2	4	4	8
Main Memory Bandwidth (GB/s)	5.6	13.6	6.4	10.6	12.8/10.6
Peak Performance (GFlop/s per node)	5.6	13.6	10.4	10.4	16.8
Interconnects					
Torus Injection Bandwidth (GB/s)	2.1	5.1	6.4	6.4	6.4
Tree Bandwidth (MB/s)	700	1700	n/a	n/a	n/a

HPCC Performance Table

Test		Metric	BGP	XT
		N	468000	937000
		NB	144	168
DGEMM	Star	Avg GFLOP/s	2.4115	7.6004
	Single	GFLOP/s	2.4111	7.6759
STREAM	StarCopy	Avg GB/s	2.0789	2.3672
	StarScale	Avg GB/s	2.4038	1.3260
	StarAdd	Avg GB/s	2.3515	1.5252
	StarTriad	Avg GB/s	2.2441	1.5352
	SingleCopy	GB/s	3.8278	6.5867
	SingleScale	GB/s	3.6152	4.0444
	SingleAdd	GB/s	3.7085	4.4941
	SingleTriad	GB/s	3.7085	4.4742
RA OPT2	Star	Avg GUP/s	0.0042	0.0067
	Single	GUP/s	0.0067	0.0103
FFT	Star	Avg GFLOP/s	0.2729	0.4126
	Single	GFLOP/s	0.2885	0.5870
Comm	Ping-Pong Latency	Min, us	2.7490	6.6310
		Avg, us	3.5346	8.3476
		Max, us	4.1141	9.9391
	Ping-Pong Bandwidth	Min, GB/s	0.3745	1.2996
		Avg, GB/s	0.3852	1.6057
		Max, GB/s	0.3858	1.6792
	NOR Latency	us	5.4215	10.8004
	NOR Bandwidth	GB/s	0.1823	0.4093
	ROR Latency	us	6.0855	32.9361
	ROR Bandwidth	GB/s	0.0212	0.0603

Figure: Results of Single Processor and Embarrassingly Parallel Tests, as well as communication tests.

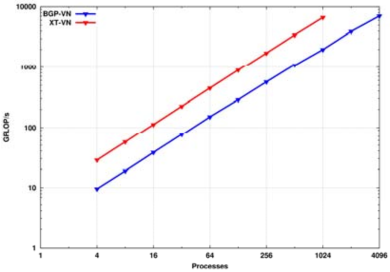
HPCC Performance

Comm	Ping-Pong Latency	Min, us	2.7490	6.6310
		Avg, us	3.5346	8.3476
		Max, us	4.1141	9.9391
	Ping-Pong Bandwidth	Min, GB/s	0.3745	1.2996
		Avg, GB/s	0.3852	1.6057
		Max, GB/s	0.3858	1.6792
	NOR Latency	us	5.4215	10.8004
	NOR Bandwidth	GB/s	0.1823	0.4093
	ROR Latency	us	6.0855	32.9361
	ROR Bandwidth	GB/s	0.0212	0.0603

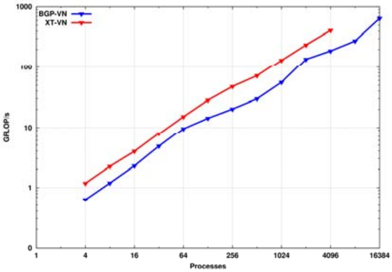
Single Process, Embarassingly Parallel, Communication

- BG/P does not perform as well on single process/embarassingly parallel tests. Why?
- BG/P has a lower clock rate than Cray XT
- Cray XT also has more memory
- Cray XT also has larger problem size to compensate
- BG/P is strong when latency is low

Parallel Tests



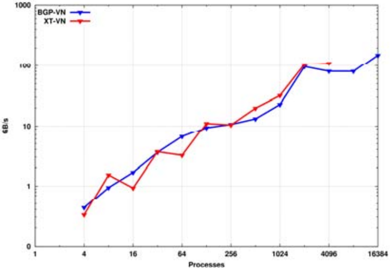
(a) HPL performance.



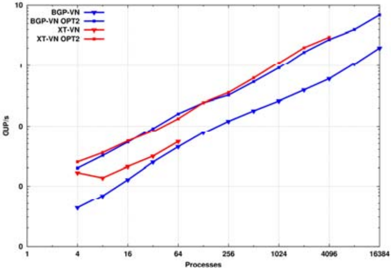
(b) FFT performance.

Figure: Similar scaling. XT has higher clock rate, larger problem size.

Parallel Tests



(c) PTRANS performance.



(d) RandomAccess performance.

Figure: Similar scaling. XT has higher clock rate, larger problem size.

Communication

- Communication benchmarks determine how different network parameters effect performance.
- Three parameters for Halo Benchmarks:
    - MPI Protocols
    - Process Mapping
    - Grid Selection
    - Identifies communication strengths/weaknesses

Halo Benchmark: MPI Protocols

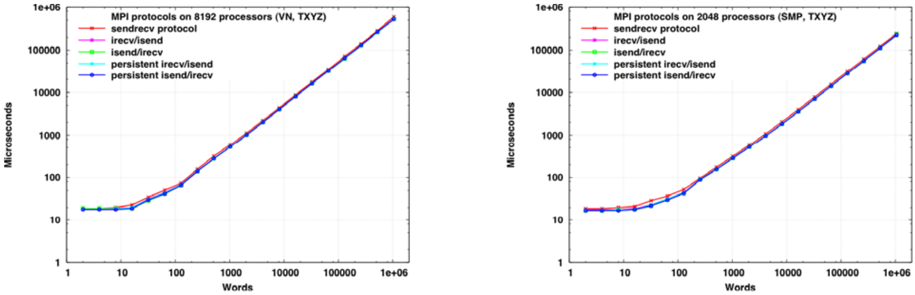
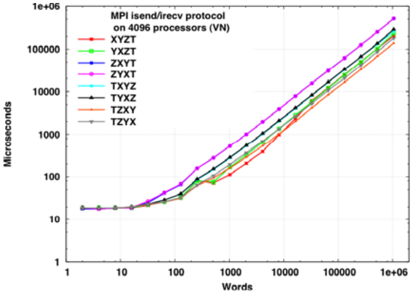


Figure: MPI Protocol is largely unimportant.

Halo Benchmark: Process Mappings



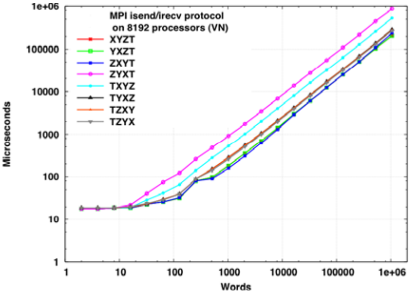
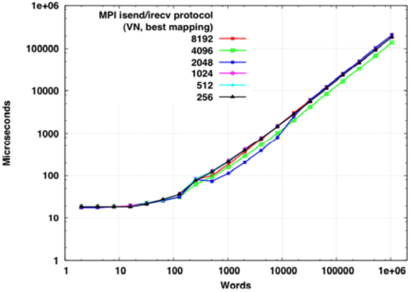


Figure: Mapping is unimportant for low volumes, but is important as size grows.

Halo Benchmark: Grid Size



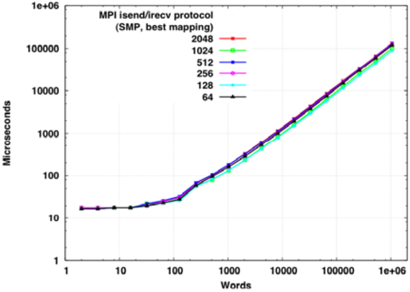
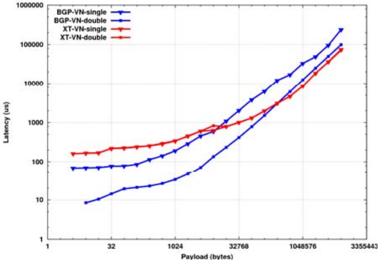


Figure: Performance is largely unrelated to processor grid size.

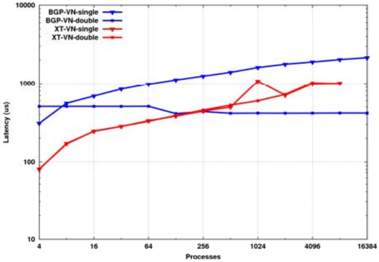
MPI Collective

- MPI Collective emphasises latency as problem sizes grow.
- Tests performance with MPI_SUM operation.
    - MPI_SUM reduces float values to a single value.
    - Modified version included to add double precision.
- B_CAST, a broadcast function, is also tested.

MPI Collective Performance



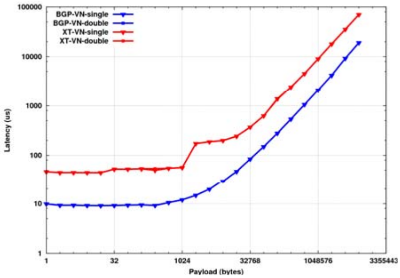
(a) IMB Allreduce operation latency versus message payload size.



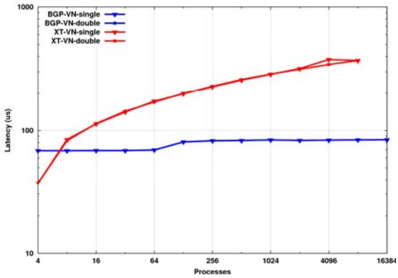
(b) IMB Allreduce operation latency versus process count.

Figure: BG/P scales better than XT with double precision values.

MPI Collective Performance



(c) IMB Bcast operation latency versus message payload size.



(d) IMB Bcast operation latency versus process count.

Figure: BG/P scales better than XT thanks to Global Collective Tree.

TOP500

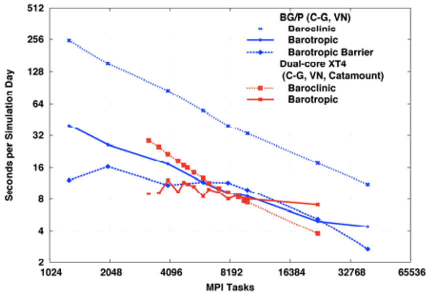
- With Linpack HPL Benchmark, BG/P ranked 74 on June 2008 TOP500.
- Ranked 5^{th} overall on on Green500.

Sample Applications

- Several Department of Energy applications were used as benchmarks.
- Applications represent target domain of BG/P
- Performance in these applications is higher priority than individual metrics.

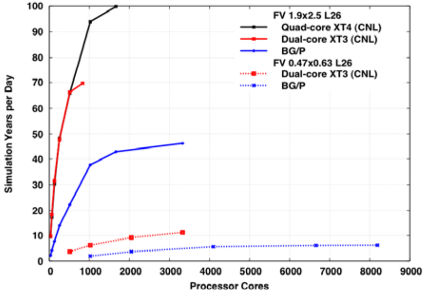
Sample Applications

- BG/P is competitive at tasks with high communication cost.
- One example is the Parallel Ocean Program with load imbalances.



Sample Applications cont.

- Some applications do not perform as well on BG/P
- The Community Atmospheric Model (CAM) is one example.
- Issues with CAM's performance are currently inherent to the program.



Power Consumption

	BG/P	XT/QC
Cores	8192	30976
Measured Aggregate Power / HPL (kW)	63	1580
Per core (W)	7.7	51.0
Measured Aggregate Power / Normal (kW)	60	1500
Per core (W)	7.3	48.4
Peak Flop/s (Tflops/s)	27.9	260.2
HPL Rmax	21.9	205.0
HPL Flop/s Power Ratio (Mflops/s per W)	347.6	129.7
POP SYD @ 8192 cores	3.6	12.5
Aggregate power required (kW)	60.0	396.7
Approximate Cores for POP SYD of 12	40000	7500
Aggregate power required (kW)	293.0	363.2

- Power consumption is very low on BP/G
- Power consumption/performance can be low for scientific computing.

Summary

- BG/P is an IBM Supercomputer Architecture
- BG/P shows good performance in low latency conditions.
- BG/P scales for most applications.
- BG/P also has low power consumption.
- For some tasks, BG/P may have much lower performance.
- BG/P has very low power usage compared to other clusters for most applications.

References

(1) S. Alam, R. Barrett, M. Bast, M.R. Fahey, J. Kuehn, C. McCurdy, J. Rogers, P. Roth, R. Sankaran, J.S. Vetter, P. Worley, and W. Yu, *Early evaluation of ibm bluegene/p*, High performance computing, networking, storage and analysis, 2008. sc 2008. international conference for, 2008Nov, pp. 1–12.

(2) C. Sosa and B. Knudson, *Ibm system blue gene solution: Blue gene/p application development*, 2009. `http://www.redbooks.ibm.com/redbooks/pdfs/sg247287.pdf`, Accessed: 9/26/2015.