Early Experience on the Blue Gene/Q Supercomputing System

Vitali Morozov

and Performance Engineering Team

Argonne Leadership Computing Facility
Argonne National Laboratory, Argonne, IL

Mira Performance Boot Camp

Argonne, May 21, 2013

Outline

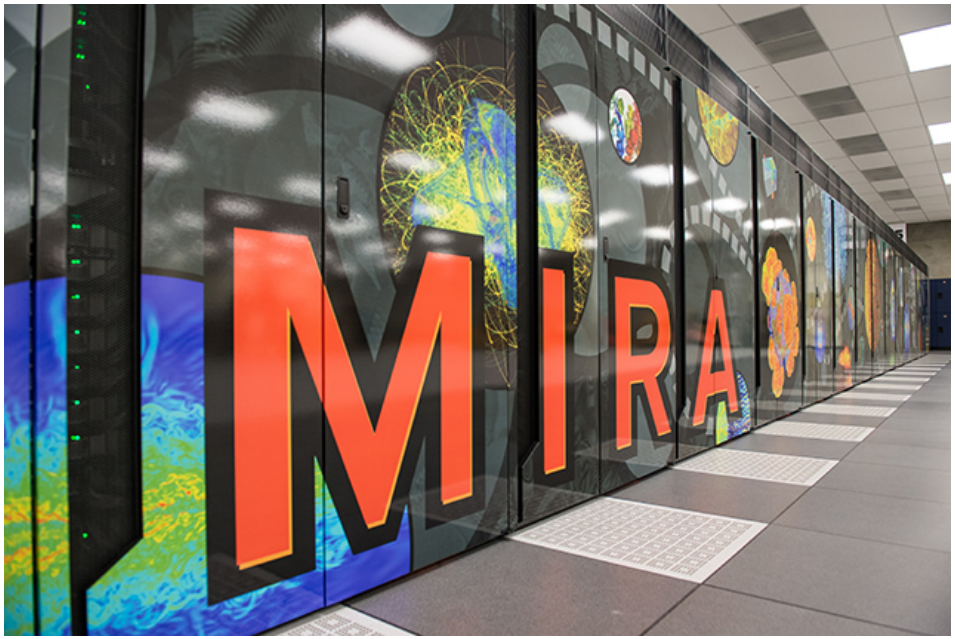
- Introduction
  - Argonne Leadership Computing Facility
  - From BG/P to BG/Q
- BG/Q System Characteristics
  - Memory Subsystem
  - Computation Capabilities
  - Interconnect and Communication Properties
  - Power Consumption and Efficiency
- BG/Q Specific Features
  - QPX - 4-way SIMD Vector Unit
  - Prefetching Unit and List Prefetching
- Application Studies
  - GTC, GFMC, DNS3D

ALCF @ Argonne

Production Systems:

Mira – *BG/Q system*

- 49,152 nodes / 786,432 cores
- 786 TB of memory
- Peak flop rate: 10 PF
- Linpack flop rate: 8.1 PF



Intrepid – *BG/P system*

- 40,960 nodes / *163,840* cores
- 82 TB of memory
- Peak flop rate: 557 TF
- Linpack flop rate: 459 TF

Storage

- Scratch: 28.8 PB raw capacity, 240 GB/s bw (GPFS); Home: 1.8 PB raw capacity

From BG/P to BG/Q: Another Step Forward

Design Parameters	BG/P	BG/Q	Improvement
Cores / Node	4	16	4x
Clock Speed (GHz)	0.85	1.6	1.9x
Flop / Clock / Core	4	8	2x
RAM / core (GB)	0.5	1	2x
Flops / Node (GF)	13.6	204.8	**15x**
Mem. BW/Node (GB/sec)	13.6	42.6	3x
1 Hop Latency (us)	[illegible]	[illegible]	--
Nodes / Rack	[illegible]	[illegible]	--
Concurrency / Rack	4,096	65,536	16x
Network Interconnect	3D torus	5D torus	Smaller diameter
GFlops/Watt	0.77	2.10	3x
Cooling	Air	Water	~30% savings/W
Total Racks at ALCF	40	48	556 TF to 10 PF

BG/Q Node is richer

BG/Q Measured Memory Characteristics

In-house	Latency (ns)		Bandwidth (GB/s)	
	LLC	DRAM	Peak	Measured
BG/P	62	162	13.6	10.5
BG/Q	[illegible]	[illegible]	[illegible]	[illegible]

Lesson Learned:

1. Node memory bandwidth is limited.

2. With high level of concurrency, it can quickly be exhausted.

3. Approximately 4 cores can consume entire bandwidth.

STREA					
					ls
Copy					
Scale					
Add	3.7	3.3	8.9	8.7	28.4
Triad	3.7	3.3	8.9	8.5	28.5
Store	6.0	7.3	11.3	10.5	16.0

Single Node Computation / Scaling

dgemm	BG/P		BG/Q		Single core, N = 4096
1 core	1 thread	1 thread	2 threads	3 threads	4 threads
GFlops	2.68	5.18	7.28	8.97	10.87
% Peak	78.8	40.5	56.9	70.1	84.9

Lesson Learned:

1. Multiple threads on a core help increasing instruction throughput
2. Multiple threads on a core help hiding memory latency

dgen					4096
1 no					es
GFlo					.4
% Pe					3

Chargei	BG/Q Open MP Scaling, 1 thread per core			
(GTC)	1 core	2 cores	3 cores	4 cores
GFlops	0.315	0.628	0.941	1.249
Speedup	1	1.99	2.99	3.97

From BG/P to BG/Q: Another Step Forward

Design Parameters	BG/P	BG/Q	Improvement
Cores / Node	4	16	4x
Clock Speed (GHz)	0.85	1.6	1.9x
Flop / Clock / Core	4	8	2x
RAM / core (GB)	0.5	1	2x
Flops / Node			15x
Mem. BW/No			3x
1 Hop Latency (us)	2.6	2.2	--
Nodes / Rack	1,024	1,024	--
Concurrency / Rack	4,096	65,536	**16x**
Network Interconnect	3D torus	5D torus	Smaller diameter
GFlops/Watt	0.77	2.10	3x
Cooling	Air	Water	~30% savings/W
Total Racks at ALCF	40	48	556 TF to 10 PF

BG/Q Interconnect is faster

Communication Characteristics

P2P	Intra-Node		Single Link	
	BG/P	BG/Q	BG/P	BG/Q
0 B	2			
4 KB	6			
64 KB	32			

P2P	512 nodes	8 racks	32 racks	48 racks
0 B				3.66
4 KB				9.54
64 KB				49.75

Lesson Learned:

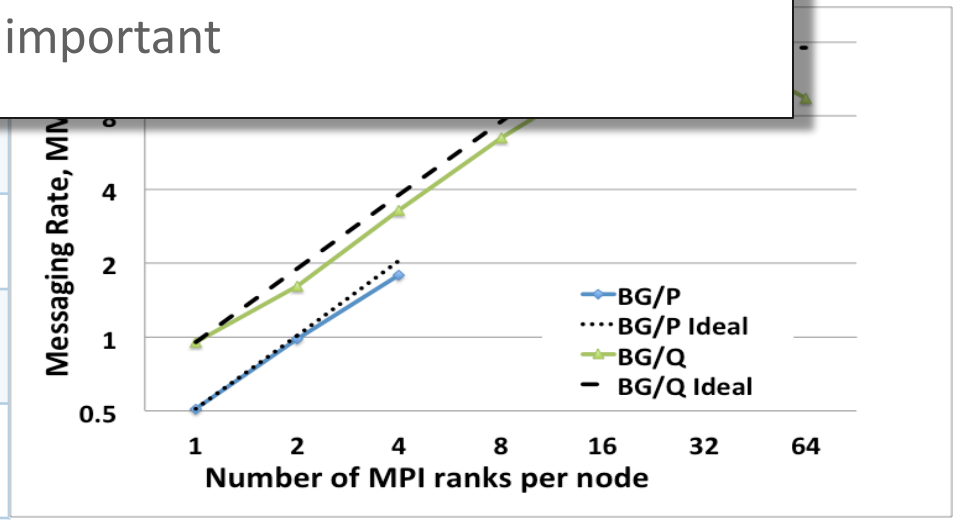
1. Interconnect continues to be fast and scalable

2. Multiple processes must be used to saturate its capabilities

3. Partitioning continues to be important

Bi-sec			
0.5 racks	0.89	0.91	0.84
8 racks	3.55	3.64	3.36
32 racks	141.8	145.5	134.3
48 racks	212.4	218.3	201.4

BG/Q Bi-section bandwidth, TB/s



BG/Q Messaging rate, MMPS

From BG/P to BG/Q: Another Step Forward

Design Parameters	BG/P	BG/Q	Improvement
Cores / Node	4	16	4x
Clock Speed (GHz)	0.85	1.6	1.9x
Flop / Clock / Core	4	8	2x
RAM / core (GB)	0.5	1	2x
Flops / Node (GF)			15x
Mem. BW/Node (			3x
1 Hop Latency (us)	2.6	2.2	--
Nodes / Rack	1,024	1,024	--
Concurrency / Rack	4,096	65,536	16x
Network Interconnect	3D torus	5D torus	Smaller diameter
GFlops/Watt	0.77	2.10	**3x**
Cooling	Air	Water	~30% savings/W
Total Racks at ALCF	40	48	556 TF to 10 PF

BG/Q is Power Efficient

Power Efficiency and Characteristics

Idle node power: 50 W

	Time	Power	Energy	GFlops	%	Efficiency
TPC	sec	Watts	kJ	Node	Peak	GF/Watt
1	317.16	59.43	18.8	54.3	26.3	0.91
2	165.99	65.81	10.9	103.0	51.2	1.56
4	130.50	70.94	9.25	132.3	65.0	1.86

BG/Q single node power usage: **dgemm**, N=1024, 1, 2, and 4 threads per core (TPC)

STREAM	GB/s	Power	Efficiency
		Watts	GB/s/Watt
Copy	27.1	69.68	0.39
Scale	27.5	71.25	0.38
Add	28.4	68.43	0.42
Triad	28.5	68.32	0.42
Store	16.1	66.03	0.24

STREAM, single node, 1 thread per core

BG/Q Features: QPX Quad SIMD FPU

```
#include <math.h>
int exp_c( int n, float *r2, float m_d, float m_h, float *res0, float *res1 ) {
    for ( int i = 0; i < n; i++ )  {
        res0[i] = exp( -m_d * r2[i] );
        res1[i] = exp( -m_h * r2[i] );
    }
    return 0;
}
```

Regular C:

```
#include <massv.h>
int exp_v( int n, float *r2, float m_d, float m_h, float *res0, float *res1 ) {
    int I;  float p0[n], p1[n];

    for ( i = 0; i < n; i++ )  {
        p0[i] = -m_d * r2[i];
        p1[i] = -m_h * r2[i];
    }
```

C with MASSV:

of code
f peak
dup

Lesson Learned:

Find opportunities to use QPX extension

```
vector4doub
vector4doub

for ( i = 0, j =
        __dcbt

        s0  = v

        x0  = v
        f0 = ve
        f1 = ve
        f0 = vec_madd( x0, f1, a2 ); … m0 = vec_madd( x7, m1, a2 );
        f1 = vec_madd( x0, f0, a1 ); … m1 = vec_madd( x7, m0, a1 );
        f0 = vec_madd( x0, f1, a0 ); … m0 = vec_madd( x7, m1, a0 );
        f1 = vec_mul( f0, f0 ); … m1 = vec_mul( m0, m0 );
        f0 = vec_mul( f1, f1 ); … m0 = vec_mul( m1, m1 );
        f1 = vec_re( f0 ); … m1 = vec_re( m0 );
        f0 = vec_nmsub( f0, f1, a0 ); … m0 = ve
        f0 = vec_madd( f1, f0, f1 ); … m0 = vec_

        vec_st( f0, j   , res0 ); … vec_st( m0, j+4
    }
```

C with vector intrinsics:
150 lines of code
15-55% of peak
3x to 15x speedup

```
h;

....
qvfmadd 23,24,23,4  # f6 = a1 + x0 * f6;
....
qvfmadd 23,24,23,3  # f6 = a0 + x0 * f6;
qvfmul  11,11,11    # f0 = f0 * f0;
....
qvfmul  23,23,23    # f6 = f6 * f6;
qvfmul  11,11,11    # f0 = f0 * f0;
....
qvfmul  23,23,23    # f6 = f6 * f6;
qvfre   12,11       # f1 = __fre( f0 );
....
qvfre   24,23       # f7 = __fre( f6 );
bc      16, 0, LOOP0
```

Assembly:
245 lines of code
55-59% of peak
10x to 15x speedup

BG/Q Features: List Prefetch

Repeatable pattern for random memory access

```
for ( i = 0; i < ndim; i++ ) buffer[i] = (double)i;

for ( i = 0; i < nsamples; i++ ) index[i] = ndim * RAND[i];

for ( i = 0; i < nsamples; i++ )

    isum = isum + buffer[index[i]];
```

Original code

```
for ( i = 0; i < ndim; i++ ) buffer[i] = (double)i;

for ( i = 0; i < nsamples; i++ ) index[i] = ndim * RAND[i];

L1P_PatternConfigure( nsamples );
record = 1; // record = 0 for replay

L1P_PatternStart( record );

for ( i = 0; i < nsampl es; i++ )

    isum = isum + buffer[index[i]];

L1P_PatternStop();
```

Record/replay

BG/Q Features: List Prefetch Performance

	Original	Record	Replay	OMP	OMP/Rec	OMP/Play
Speedup	1.00	**-2%**	3.00	3.74	-12%	**8.67**
Inst rate	2.13%	2.12%	6.38%	6.90%	11.6%	18.3%
L1 cach						
L1p bu						
L2 cach						
DDR						
Memory bandwidth, Bytes/cycle. Peak approx. 18.3 Bytes/cycle						
DDR BW	1.20	3.93	3.93	4.79	5.36	11.46

Lesson Learned:

1. List prefetch is a new effective feature that helps hiding memory latency
2. Find the right balance between the level of concurrency, memory bandwidth and core resources contention

Argonne Science Applications

Apps	Q/P Ratio	Comments
DNS3D	13.5	64 RPN, MPI, turbulence, 3D FFT, All-to-all
GFMC	12.5	8 RPN, MPI-OMP, Nuclear Green function MC
HIRAM	11.9	8 RPN, MPI-OMP, Hi-res atmosphere
GTC	10.7	16 RPN, MPI-OMP, Gyro kinetics, particles
NAMD	9.7	16 RPN, Charm++ OMP, 3D FFT, Classical MD
FLASH	8.9	16 RPN, MPI-OMP, multi-physics, Hydro, AMR
LS3DF	8.1	64 RPN, MPI, ESSL OMP, DFT, divide-and-conq.
GPAW	7.6	32 RPN, MPI, ESSL OMP, DFT, PAW, real space
NEK	7.3	32 RPN, MPI, Spectral element CFD
MILC	6.1	64 RPN, MPI, Lattice QCD

Argonne Science Applications

Apps	Q/P Ratio	Comments
DNS3D	13.5	64 RPN, MPI, turbulence, 3D FFT, All-to-all
GFMC	12.5	8 RPN, MPI-OMP, Nuclear Green function MC
HIRAM	11.9	**Bi-section bandwidth** here
GTC	10.7	16 RPN, MPI-OMP, Gyro kinetics, particles
NAMD	9.7	16 RPN, Charm++ OMP, 3D FFT, Classical MD
FLASH	8.9	16 RPN, MPI-OMP, multi-physics, Hydro, AMR
LS3DF	8.1	64 RPN, MPI, ESSL OMP, DFT, divide-and-conq.
GPAW	7.6	32 RPN, MPI, ESSL OMP, DFT, PAW, real space
NEK	7.3	32 RPN, MPI, Spectral element CFD
MILC	6.1	64 RPN, MPI, Lattice QCD

Argonne Science Applications

Apps	Q/P Ratio	Comments
DNS3D	13.5	64 RPN, MPI, turbulence, 3D FFT, All-to-all
GFMC	12.5	8 RPN, MPI-OMP, Nuclear Green function MC
HIRAM	11.9	8 RPN, MPI-OMP, Hi-res atmosphere
GTC	10.7	16 RPN, MPI-OMP, Gyro kinetics, particles
NAMD	9.7	16 RPN, Charm++ OMP, 3D FFT, Classical MD
FLASH	8	ydro, AMR
LS3DF	8.1	64 RPN, MPI, ESSL OMP, DFT, divide-and-conq.
GPAW	7.6	32 RPN, MPI, ESSL OMP, DFT, PAW, real space
NEK	7.3	32 RPN, MPI, Spectral element CFD
MILC	6.1	64 RPN, MPI, Lattice QCD

Hybrid MPI-OpenMP codes

Argonne Science Applications

Apps	Q/P Ratio	Comments
DNS3D	13.5	64 RPN, MPI, turbulence, 3D FFT, All-to-all
GFMC	12.5	8 RPN, MPI-OMP, Nuclear Green function MC
HIRAM	11.9	8 RPN, MPI-OMP, Hi-res atmosphere
GTC	10.7	16 RPN, MPI-OMP, Gyro kinetics, particles
NAMD	9.7	16 RPN, Charm++ OMP, 3D FFT, Classical MD
FLASH	8.9	16 RPN, MPI-OMP, multi-physics, Hydro, AMR
LS3DF	8.1	64 RPN, MPI, ESSL, OMP, DFT, divide and-conq.
GPAW		al space
NEK	7.3	32 RPN, MPI, Spectral element CFD
MILC	6.1	64 RPN, MPI, Lattice QCD

Experimental OpenMP capabilities

Argonne Science Applications

Apps	Q/P Ratio	Comments
DNS3D	13.5	64 RPN, MPI, turbulence, 3D FFT, All-to-all
GFMC	12.5	8 RPN, MPI-OMP, Nuclear Green function MC
HIRAM	11.9	8 RPN, MPI-OMP, Hi-res atmosphere
GTC	10.7	16 RPN, MPI-OMP, Gyro kinetics, particles
NAMD	[illegible]	[illegible] ical MD
FLASH	[illegible]	[illegible] dro, AMR
LS3DF	8.1	64 RPN, MPI, ESSL OMP, DFT, divide-and-conq.
GPAW	7.6	32 RPN, MPI, ESSL OMP, DFT, PAW, real space
NEK	7.3	32 RPN, MPI, Spectral element CFD
MILC	6.1	64 RPN, MPI, Lattice QCD

MPI-only codes, threading libraries

Argonne Science Applications

Apps	Q/P Ratio	Comments
DNS3D	13.5	64 RPN, MPI, turbulence, 3D FFT, All-to-all
GFMC	12.5	8 RPN, MPI-OMP, Nuclear Green function MC
HIRAM	11.9	8 RPN, MPI-OMP, Hi-res atmosphere
GTC	10.7	16 RPN, MPI-OMP, Gyro kinetics, particles
NAMD	9.7	16 RPN, Charm++ OMP, 3D FFT, Classical MD
FLASH	8.9	16 RPN, MPI-OMP, multi-physics, Hydro, AMR
LS3DF	8.1	[illegible]P, DFT, divide-and-conq.
GPAW	7.6	[illegible]P, DFT, PAW, real space
NEK	7.3	32 RPN, MPI, Spectral element CFD
MILC	6.1	64 RPN, MPI, Lattice QCD

MPI-only codes

Argonne Science Applications

Apps	Q/P Ratio	Comments
DNS3D	13.5	64 RPN, MPI, turbulence, 3D FFT, All-to-all
GFMC	12.5	8 RPN, MPI-OMP, Nuclear Green function MC
HIR		
GT		
NAM		
FLA		
LS3		
GPAW	7.6	32 RPN, MPI, ESSL OMP, DFT, PAW, real space
NEK	7.3	32 RPN, MPI, Spectral element CFD
MILC	6.1	64 RPN, MPI, Lattice QCD

Lesson Learned:

1. MPI only codes may run "AS IS"

2. Threading is key to high performance

3. BG/Q provides many opportunities for tuning

GTC: BG/Q Performance, 2048 cores

Random memory access, threads help to hide memory latency

Threads per core =>	1	2	3	4
GFlops/node	3.81	5.76	6.80	7.10
Relative Speedup	**1.00**	**1.51**	**1.78**	**1.87**
Instructions/cycle/core	0.30	0.47	0.57	0.61
% Max issue rate	30.3	33.8	40.9	44.1
Relative total instructions	**1.00**	**1.04**	**1.06**	**1.08**
Relative FPU instructions	0.42	0.40	0.39	0.38
Load source	% loads that	hit in		
L1D cache	94.3	94.1	93.8	93.4
L1P Buffer	1.95	1.71	1.55	1.42
L2 Cache	3.12	3.26	3.58	3.95
DDR	0.67	0.89	1.09	1.24
DDR bandwidth/node	Bytes/cycle	measured limit 18		
Loads	1.34	2.32	3.03	3.51
Stores	0.68	1.14	1.46	1.68
Total	**2.02**	**3.46**	**4.49**	**5.19**

GFMC: BG/Q Performance, 16K cores, 8 RPN

Sparse matrix vector multiply, threads help to fill FP pipeline

Threads per core =>	1	2	4
Gflops/node	8.29	13.2	12.7
Relative Speedup	1.00	**1.58**	**1.50**
Instructions/cycle/core	0.25	0.40	0.40
% Max issue rate	24.8	23.9	24.2
Relative total instructions	1.00	1.02	1.08
Relative FPU instructions	0.40	0.40	0.39
Load source	% loads that	hit in	
L1D cache	**82.5**	**83.4**	**82.2**
L1P Buffer	**14.3**	**13.2**	**12.7**
L2 Cache	2.01	1.97	3.06
DDR	1.11	1.29	2.05
DDR bandwidth/node	Bytes/cycle	measured limit 18	
Loads	3.94	6.93	9.08
Stores	1.96	4.36	5.67
Total	5.90	10.4	**14.8**

DNS3D: BG/Q Performance, 1024 cores

MPI-only, all-to-all. Tasks help increasing the instruction throughput

Ranks per core =>	1	2	4
Gflops/node	3.59	5.08	5.87
Relative Speedup	**1.00**	**1.39**	**1.59**
Instructions/cycle/core	**0.34**	**0.56**	**0.67**
% Max issue rate	33.6	41.9	51.1
Relative total instructions	1.00	1.19	1.25
Relative FPU instructions	0.42	0.36	0.34
Load source	% loads that	hit in	
L1D cache	93.1	93.2	88.2
L1P Buffer	3.57	3.14	2.67
L2 Cache	1.34	2.04	7.56
DDR	1.98	1.66	1.57
DDR bandwidth/node	Bytes/cycle	measured limit 18	
Loads	2.92	4.20	4.77
Stores	2.40	3.38	3.93
Total	5.32	7.57	8.70

Lessons Learned

- Applications that run on BG/P should run on BG/Q **"AS-IS"**
- Key BG/Q properties for memory latency hiding are
  - Single-core **multithreading**
  - Prefetch unit and different prefetch strategies
- Balancing instruction throughput, latency hiding, memory pressure, and concurrency level is a key
  - Contention for L1/L1p resources
  - **Scaling** without extra work
  - Coarse-grain MPI versus fine-grain OMP
  - **Memory bandwidth** can quickly be exhausted
- BG/Q continues delivering fast interconnect
  - High-concurrency messaging is essential
  - Low-diameter 5D network reduces latency on scale
- **QPX** vector unit helps, if application can use it