From Laptops to Petaflops

Portable Productivity and Performance using Open-Source Development Tools

Jeff Hammond and Katherine Riley

(with plenty of help from our colleages)

Mira Out Of The Box

Hardware:

- 48 racks – 10 Petaflops peak
- 48k compute nodes / 768k cores
  - PowerPC A2 with 16 cores/node at 1.6 GHz
  - 4 way multi-threading
  - Quad FPU – 4 wide double precision SIMD
- 384 IO nodes
- 768 TB Memory – 16 GB/node
- 5D Torus network
- Login, Service, & Management Nodes

Software:

- Compute Node Kernel OS + IO Node Linux
- XL Fortran, C, C++ Compilers
- GNU Compilers + GNU Toolchain
- Messaging Stack – SPI, PAMI, MPI
- ESSL (BLAS/LAPACK)
- Control System – MMCS, CIOD, CDTI

What's not in the box

The BG/Q software stack is a good start but users want more, including:

- Performance Tools:
  - PAPI, TAU, HPCToolkit, mpiP, ….
- Debuggers:
  - TotalView, DDT
- Libraries:
  - FFTW, ScaLAPACK, ParMetis, P3DFFT ….
- Programming Model:
  - UPC, Charm++, Global Arrays, ...

A collaborative "Early Software" project was initiated at Argonne to make this software available when Mira comes online

Building the Blue Gene/Q scientific software ecosystem

IBM has provided OSS all the way down:

- MPI sits on top of PAMI sits on top of SPI, all of which are OSS.
- CNK, which is Linux-like, is OSS.
- Python, GCC, glibc, etc. all OSS (obviously, due to GPL for GNU codes).

What IBM does not provide is readily ported by learning from these.

For example, GASNet was ported by learning PAMI and MPI design and LLVM by learning from GCC.

We are working to enable a seamless transition between your laptop and Blue Gene/Q.

Why not just build an x86 supercomputer?

Listed below are the The Green500's Top 10 most energy-efficient supercomputers in the world as of June 2012.

Comma Separated Values (cs
Microsoft Excel Spreadsheet (

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	2100.88	DOE/NNSA/LLNL	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	41.10
2	2100.88	IBM Thomas J. Watson Research Center	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	41.10
3	2100.86	DOE/SC/Argonne National Laboratory	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	82.20
4	2100.86	DOE/SC/Argonne National Laboratory	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	82.20
5	2100.86	Rensselaer Polytechnic Institute	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	82.20
6	2100.86	University of Rochester	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	82.20
7	2100.86	IBM Thomas J. Watson Research Center	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	82.20
8	2099.56	University of Edinburgh	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	493.10
9	2099.50	Science and Technology Facilities Council - Daresbury Laboratory	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	575.30
10	2099.46	Forschungszentrum Juelich (FZJ)	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	657.50

* Performance data obtained from publicly available sources including TOP500

11	2099.39	CINECA	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	821.90
12	2099.14	High Energy Accelerator Research Organization /KEK	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	246.60
13	2099.14	EDF R&D	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	328.80
14	2099.14	IDRIS/GENCI	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	328.80
15	2099.14	Victorian Life Sciences Computation Initiative	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	328.80
16	2099.14	IBM - Rochester	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	164.40
17	2099.14	IBM - Rochester	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	164.40
18	2099.14	DOE/NNSA/LLNL	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	164.40
19	2069.04	DOE/SC/Argonne National Laboratory	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	3945.00
20	2069.04	DOE/NNSA/LLNL	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	7890.00
21	1380.67	Intel	Intel Cluster, Xeon E5-26[illegible] 8C 2.600GHz, Infiniband FDR, Intel MIC	72.90
22	1379.79	Nagasaki University	DEGIMA Cluster, Intel i5, ATI Radeon GPU, Infiniband QDR	47.00
23	1266.26	Barcelona Supercomputing Center	Bullx B505, Xeon E5649 6C 2.53GHz, Infiniband QDR, NVIDIA 2090	81.50
24	1151.91	Center for Computational Sciences, University of Tsukuba	Xtream-X GreenBlade [illegible]4, Xeon E5-2670 8C 2.600GHz, Infiniband QD[illegible] [illegible]VIDIA 2090	366.00
25	1050.26	Los Alamos National Laboratory	Xtreme-X , Xeon E5-2670 8C 2.600GHz, Infiniband QDR, NVIDIA 2090	226.80

26	1010.11	CEA/TGCC-GENCI	Bullx B505, Xeon E5640 2.67 GHz, Infiniband QDR	108.80
27	994.14	CSIRO	Nitro G16 3GPU, Xeon E[illegible]2650 8C 2.000GHz, Infiniband FDR, NVIDIA 20[illegible]0	110.30
28	958.35	GSIC Center, Tokyo Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows	1243.80
29	953.29	Virginia Tech	SuperServer 2026GT-TRF, Xeon E5645 6C 2.40GHz, Infiniband QDR, NVIDIA 2050	126.30
30	932.83	University of Chicago	iDataPlex DX360M4, Xeon E5-2670 8C 2.600GHz, Infiniband FDR	73.10
31	932.78	Centro Euro-Mediterraneo per i Cambiamenti Climatici	iDataPlex DX360M4, Xeon E5-2670 8C 2.600GHz, Infiniband FDR	159.50
32	932.71	Indian Institute of Technology Madras	iDataPlex DX360M4, Xeon E5-2670 8C 2.600GHz, Infiniband FDR	97.70
33	932.62	Automotive Company	iDataPlex DX360M4, Xeon E5-2670 8C 2.600GHz, Infiniband FDR	67.30
34	932.53	Durham University	iDataPlex DX360M4, Xeon E5-2670 8C 2.600GHz, Infiniband FDR	139.60
35	932.41	Science and Technology Facilities Council - Daresbury Laboratory	iDataPlex DX360M4, Xeon E5-2670 8C 2.600GHz, Infiniband FDR	170.20
36	932.19	Universidad de Cantabria - SSC	iDataPlex DX360M4, Xeon E5-2670 8C 2.600GHz, Infiniband FDR	79.80
37	919.44	Institute of Process Engineering, Chinese Academy of Sciences	Mole-8.5 Cluster, Xeon X5520 4C 2.27 GHz, Infiniband QDR, NVIDIA 2050	540.00
38	910.90	Max-Planck-Gesellschaft MPI/IPP	iDataPlex DX360M4, Xeon E5-2670 8C 2.600GHz, Infiniband FDR	205.70
39	908.83	Leibniz Rechenzentrum	iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR	2841.00
40	901.61	Georgia Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5660 2.8Ghz, nVidia Fermi, Infiniband QDR	117.90

119	410.51	Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw	Power 775, POWER7 8C 3.836GHz, Custom	156.70
120	410.44	IBM Poughkeepsie Benchmarking Center	Power 775, Power7 3.836 GHz	172.40
121	408.97	DOE/Bettis Atomic Power Laboratory	Atipa Cluster, Xeon X56xx (Westmere-EP) 2.66 GHz, Infiniband QDR	163.02
122	408.97	Knolls Atomic Power Laboratory	Atipa Cluster, Xeon X56xx (Westmere-EP) 2.66 GHz, Infiniband QDR	163.02
123	406.87	Slovak Academy of Sciences (SAV)	Power 775, POWER7 8C 3.836GHz, Custom	188.10
124	405.38	Swiss Scientific Computing Center (CSCS)	Cray XE6, Opteron 6272 16C 2.10 GHz, Cray Gemini interconne	780.00
125	404.45	CLUMEQ - McGill University	iDataPlex DX3 M3, Xeon 2.66, Infiniband	337.00
126	398.07	United Kingdom Meteorological Office	Power 775, POWER7 8C 3.836GHz, Custom	313.40
127	378.77	King Abdullah University of Science and Technology	Blue Gene/P Solution	504.00
128	378.77	EDF R&D	Blue Gene/P Solution	252.00
129	378.76	IDRIS	Blue Gene/P Solution	315.00
130	377.99	Rice University	BlueGene/P, PowerPC 450 4C 850 MHz, Proprietary	189.00
131	377.48	DOE/SC/Oak Ridge National Laboratory	Cray XK6, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA 2090	5142.00
132	377.22	IBM Poughkeepsie Benchmarking Cen	Power 775, Power7 3.836 GHz	423.10
133	369.09	Environment Canada	Power 775, POWER7 8C 3.84 GHz, Custom	501.50
134	369.09	Environment Canada	Power 775, POWER7 8C 3.84 GHz, Custom	501.50
135	368.64	Automotive	BladeCenter HS22 Cluster, WM Xeon 6-core 2.93Ghz, Infiniband	240.40

166	318.69	NOAA/Oak Ridge National Laboratory	Cray XT6-HE, Opteron 6100 12C 2.1GHz	610.00
167	311.76	NASA/Ames Research Center/NAS	SGI Altix ICE X/8200EX/8400EX, Xeon 54xx 3.0/5570/5670/E5-2670 2.93/2.6/3.06/3.0 Ghz, Infiniband QDR/FDR	3987.00
168	305.33	Electronics	xSeries x3550M3 Cluster, Xeon X5650 6C 2.66 GHz, Infiniband QDR	251.10
169	305.33	Electronics	xSeries x3550M3 Cluster, Xeon X5650 6C 2.66 GHz, Infiniband QDR	251.10
170	305.33	Electronics	xSeries x3550M3 Cluster, Xeon X5650 6C 2.66 GHz, Infiniband QDR	241.80
171	297.44	National Institute for Computational Sciences/University of Tennessee	Cray XT5-HE Opteron Six Core 2.6 GHz	3090.00
172	297.26	National Computational Infrastructure National Facility (NCI-NF)	Sun Blade x6048, Xeon X5570 2.93 Ghz, Infiniband QDR	425.22
173	288.73	Lawrence Livermore National Laboratory	Xtreme-X , Xeon E5-2670 8C 2.600GHz, Infiniband QDR	1203.20
174	286.74	Defence	BladeCenter HX5, Xeon E7-4870 10C 2.40 GHz, Infiniband QDR	226.20
175	281.56	IT Service Provider	Cluster Platform 3000 280c G7, Xeon X5672 4C 3.20GHz, Infiniband QDR	216.03
176	278.89	DOE/NNSA/LANL/SNL	Cray XE6, Opteron 6136 8C 2.40GHz, Custom	3980.00
177	277.76	Lawrence Livermore National Laboratory	Dell DCS Xanadu 2.5, Xeon E55xx 2.4Ghz, Infiniband DDR	260.69
178	277.29	University of Alaska - Arctic Region Supercomputing Center	Cray XE6 8-core 2.3 GHz	320.68
179	271.71	Centre for High Performance Computing	Blade X6275/ PowerEdge C6100 Cluster, Xeon X5570/X5670/ 4C 2.93 GHz, Infiniband QDR	225.72
180	268.22	Cray Inc.	Cray XE6, Opteron 6272 16C 2.1/2.2/2.3 GHz, Cray Gemini interconnect	467.90

Power data in KW for entire system

next

Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom / 2011 IBM	1572864	16324.75	20132.66	7890.0
2	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011 Fujitsu	705024	10510.00	11280.38	12659.9
3	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom / 2012 IBM	786432	8162.38	10066.33	3945.0
4	Leibniz Rechenzentrum Germany	**SuperMUC** - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR / 2012 IBM	147456	2897.00	3185.05	3422.7
5	National Supercomputing Center in Tianjin China	**Tianhe-1A** - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 / 2010 NUDT	186368	2566.00	4701.00	4040.0
6	DOE/SC/Oak Ridge National Laboratory United States	**Jaguar** - Cray XK6, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA 2090 / 2009 Cray Inc.	298592	1941.00	2627.61	5142.0
7	CINECA Italy	**Fermi** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom / 2012 IBM	163840	1725.49	2097.15	821.9
8	Forschungszentrum Juelich (FZJ) Germany	**JuQUEEN** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom / 2012 IBM	131072	1380.39	1677.72	657.5
9	CEA/TGCC-GENCI France	**Curie thin nodes** - Bullx B510, Xeon E5-2680 8C 2.700GHz, Infiniband QDR / 2012 Bull	77184	1359.00	1667.17	2251.0
10	National Supercomputing Centre in Shenzhen (NSCS) China	**Nebulae** - Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050 / 2010 Dawning	120640	1271.00	2984.30	2580.0
11	NASA/Ames Research Center/NAS United States	**Pleiades** - SGI Altix ICE X/8200EX/8400EX, Xeon 54xx 3.0/5570/5670/E5-2670 2.93/2.6/3.06/3.0 Ghz, Infiniband QDR/FDR / 2011 SGI	125980	1243.00	1731.84	3987.0
12	International Fusion Energy Research Centre (IFERC), EU(F4E) - Japan Broader Approach collaboration Japan	**Helios** - Bullx B510, Xeon E5-2680 8C 2.700GHz, Infiniband QDR / 2011 Bull	70560	1237.00	1524.10	2200.0

June 2012

Rank	Installation Site	Machine	Number of nodes	Number of cores	Problem scale	GTEPS
1	DOE/SC/Argonne National Laboratory	Mira/BlueGene/Q	32768	524288	38	3541.00
1	LLNL	Sequoia/Blue Gene/Q	32768	524288	38	3541.00
2	DARPA Trial Subset, IBM Development Engineering	Power 775, POWER7 8C 3.836 GHz	1024	32768	35	508.05
3	Information Technology Center, The University of Tokyo	Oakleaf-FX (Fujitsu PRIMEHPC FX 10)	4800	76800	38	358.10
4	GSIC Center, Tokyo Institute of Technology	HP Cluster Platform SL390s G7 (three Tesla cards per node)	1366	16392	35	317.09
5	Brookhaven National Laboratory	BLUE GENE/Q	1024	16384	34	294.29
6	DOE/SC/Argonne National Laboratory	Vesta/BlueGene/Q	1024	16384	34	292.36
7	NASA-Ames / Parallel Computing Lab, Intel Labs	Pleiades - SGI ICE-X, dual plane hypercube FDR infiniband, E5-2670 "sandybridge"	1024	16384	34	270.33
8	NERSC/LBNL	XE6	4817	115600	35	254.07
9	NNSA and IBM Research, T.J. Watson	NNSA/SC Blue Gene/Q Prototype II	4096	65536	32	236.00
10	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.0 (CPU only)	1366	16392	36	202.68
11	Intel Dupont / Parallel Computing Lab, Intel Labs	Endeavor; Dual-socket 2.6GHz SNB-EP	320	5120	32	115.94
12	LBL	Hopper	1800	43200	37	105.00

There is no such thing as an x86 supercomputer

Power is the limiting factor and traditional machines like Jaguar are 10 times less power efficient than Blue Gene/Q.

Intel is moving towards x86+MIC. MIC has the x86 toolchain but it is still an attached coprocessor and is not binary compatible with x86 CPUs.

AMD is moving towards CPU+APU. No one uses this hardware in HPC…

NVIDIA provides a disruptive software development environment and has all the issues associated with a heterogeneous system.

Even when platforms resemble commodity hardware, they still require extensive custom engineering to scale. Cray Gemini is substantially more scalable than Infiniband, for example.

IBM PERCS (formerly Blue Waters), is much less efficient than BG/Q…

What do you want for development?

The same:

- Compilers
- MPI
- Performance tools
- IDE
- Languages
- What else?

We have:

- GCC, LLVM
- MPICH2
- TAU, HPCToolkit, PAPI
- Eclipse
- Fortran, C/C++, Python

Do you take it for granted that you can use Python?

Do the portable compilers – i.e. not IBM XL or Cray – support the features?

OSS Compilers on BG/Q

ALCF has been frustrated by the compiler options on Blue Gene/P.

LLVM is an OSS compiler project driven by Apple and used by NVIDIA, Cray, Google, etc.

Hal Finkel (CSGF@ALCF) ported LLVM to Blue Gene/P and Blue Gene/Q and it supports features previously only available from the IBM compiler (auto and manual vectorization).

Collaborators are working on transactional memory in LLVM. We are exploring how to realize OpenMP and PGAS in LLVM.

We want to have TM and SE support in GCC (since these are OpenMP oriented), which will likely happen after the LLVM implementation.