IBM Blue Gene: Architecture and Software Overview

Alexander Pozdneev

Research Software Engineer, IBM

June 23, 2017 — International Summer Supercomputing Academy

© 2017 IBM Corporation

Outline

1. Blue Gene in the HPC world
2. Blue Gene/P architecture
3. Blue Gene/P Software
4. Advanced information
5. References
6. Conclusion

© 2017 IBM Corporation

Outline

1. Blue Gene in the HPC world
   - Brief history
   - TOP500
   - Graph500
   - Gordon Bell awards
   - Installation sites
2. Blue Gene/P architecture
3. Blue Gene/P Software
4. Advanced information
5. References
6. Conclusion

© 2017 IBM Corporation

Brief history of Blue Gene architecture

- **1999** — US$100M research initiative, novel massively parallel architecture, protein folding
- **2003,** Nov — Blue Gene/L first prototype — TOP500, #73
- **2004,** Nov — 16 Blue Gene/L racks at LLNL — TOP500, #1
- **2007** — The 2nd generation — Blue Gene/P
- **2009** — National Medal of Technology and Innovation
- **2012** — The 3rd generation — Blue Gene/Q
- Features:
  - Low frequency, low power consumption
  - Fast interconnect balanced with CPU
  - Light-weight OS

© 2017 IBM Corporation

Blue Gene in TOP500

Date	#	System	Model	Nodes	Cores
Jun'17	5	Sequoia	Q	96k	1.5M
Jun'16 – Nov'16	4	Sequoia	Q	96k	1.5M
Jun'13 – Nov'15	3	Sequoia	Q	96k	1.5M
Nov'12	2	Sequoia	Q	96k	1.5M
Jun'12	1	Sequoia	Q	96k	1.5M
Nov'11	13	Jugene	P	72k	288k
Jun'11	12	Jugene	P	72k	288k
Nov'10	9	Jugene	P	72k	288k
Jun'10	5	Jugene	P	72k	288k
Nov'09	4	Jugene	P	72k	288k
Jun'09	3	Jugene	P	72k	288k
Nov'08	4	DOE/NNSA/LLNL	L	104k	208k
Jun'08	2	DOE/NNSA/LLNL	L	104k	208k
Nov'07	1	DOE/NNSA/LLNL	L	104k	208k
Jun'07	1	DOE/NNSA/LLNL	L	64k	128k
Jun'06 / Nov'06	1	DOE/NNSA/LLNL	L	64k	128k
Nov'05	1	DOE/NNSA/LLNL	L	64k	128k
Jun'05	1	DOE/NNSA/LLNL	L	32k	64k
Nov'04	1	DD2 beta-system	L	16k	32k
Jun'04	4	DD1 prototype	L	4k	8k

© 2017 IBM Corporation

Blue Gene in Graph500

Date	#	System	Model	Nodes	Cores	Scale	GTEPS
Jun'17	3	Sequoia	Q	96k	1.5M	41	23751
Nov'16	3	Sequoia	Q	96k	1.5M	41	23751
Jun'16	2	Sequoia	Q	96k	1.5M	41	23751
Nov'15	2	Sequoia	Q	96k	1.5M	41	23751
Jul'15	2	Sequoia	Q	96k	1.5M	41	23751
Nov'14	1	Sequoia	Q	96k	1.5M	41	23751
Jun'14	2	Sequoia	Q	64k	1M	40	16599
Nov'13	1	Sequoia	Q	64k	1M	40	15363
Jun'13	1	Sequoia	Q	64k	1M	40	15363
Nov'12	1	Sequoia	Q	64k	1M	40	15363
Jun'12	1	Sequoia/Mira	Q	32k	512k	38	3541
Nov'11	1	BG/Q prototype	Q	4k	64k	32	253
Jun'11	1	Interpid/Jugene	P	32k	128k	38	18
Nov'10	1	Interpid	P	8k	32k	36	7

© 2017 IBM Corporation

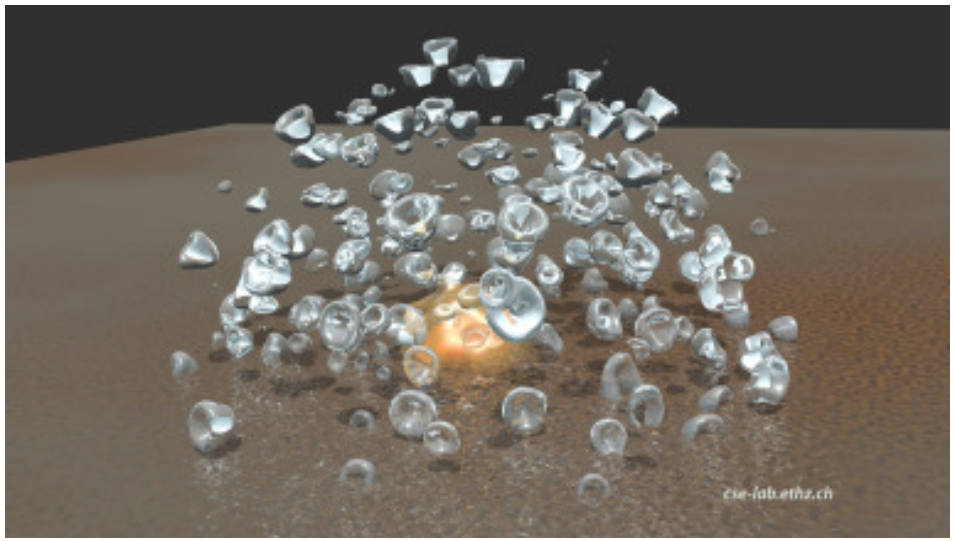
Gordon Bell prize — 8 awards

- **2015** — An Extreme-Scale Implicit Solver for Complex PDEs: Highly Heterogeneous Flow in Earth's Mantle
- **2013** — 11 PFLOP/s simulations of cloud cavitation collapse
- **2009** — The cat is out of the bag: Cortical simulations with 10^9 neurons, 10^{13} synapses
- **2008** — Linearly scaling 3D fragment method for large-scale electronic structure calculations
- **2007** — Extending stability beyond CPU millennium: A micron-scale atomistic simulation of Kelvin-Helmholtz instability
- **2006** — Large-scale Electronic Structure Calculations of High-Z Metals on the Blue Gene/L Platform
- **2006** — The Blue Gene/L supercomputer and quantum ChromoDynamics
- **2005** — 100+ TFlop Solidification Simulations on Blue Gene/L

© 2017 IBM Corporation

11 PFLOP/s simulations of cloud cavitation collapse

- ACM Gordon Bell Prize 2013
- IBM Blue Gene/Q Sequoia
  - 96 racks, 20.1 PFLOPS peak
  - 1.5M cores (6.0M threads)
  - 73% peak (14.4 PFLOPS)
  - 13T grid points
  - 15k cavitation bubbles
- Destructive capabilities of collapsing bubbles
  - high pressure fuel injectors and propellers
  - shattering kidney stones
  - cancer treatment
- doi: 10.1145/2503210.2504565
- https://youtu.be/zfG9solC6_Y

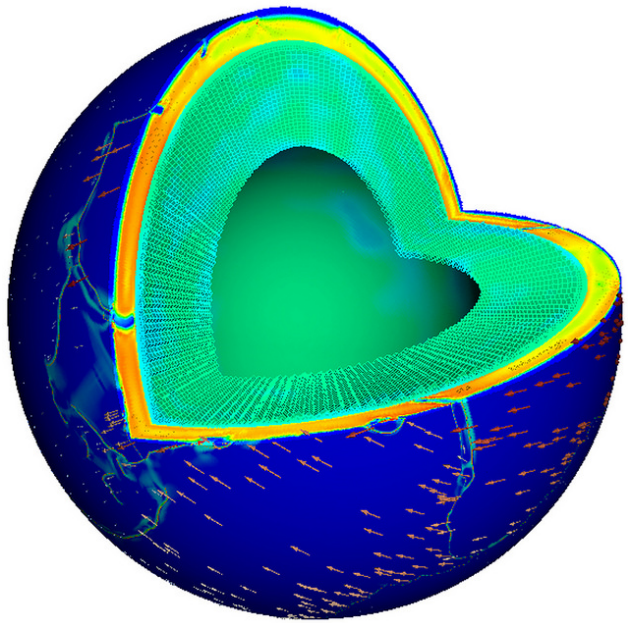


- IBM Research Zurich
- ETH Zurich
- Technical University of Munich
- Lawrence Livermore National Laboratory (LLNL)

© 2017 IBM Corporation

An Extreme-Scale Implicit Solver for Complex PDEs

- ACM Gordon Bell Prize 2015
- IBM Blue Gene/Q Sequoia
  - 96 racks, 20.1 PFLOPS peak
  - 1.5M cores
  - 97% parallel efficiency
  - 602B DOF
- Dynamics of the earth's mantle and crust
  - mantle convection
  - thermal and geological evolution of the planet
  - plate tectonics
- doi: 10.1145/2807591.2807675
- http://youtu.be/cHmk9-v6H1Y



- IBM Research Zurich
- The University of Texas at Austin
- New York University
- California Institute of Technology

© 2017 IBM Corporation

Some of Blue Gene installation sites

- USA:
  - DOE/NNSA/LLNL (2)
  - DOE/SC/ANL (3)
  - University of Rochester
  - Rensselaer Polytechnic Institute
  - IBM Rochester / T.J. Watson (4)
- Germany:
  - Forschungszentrum Juelich (FZJ)
- Italy:
  - CINECA
- Switzerland:
  - Ecole Polytechnique Federale de Lausanne
  - Swiss National Supercomputing Centre (CSCS)
- Poland:
  - Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw
- France:
  - CNRS/IDRIS-GENCI
  - EDF R&D
- UK:
  - University of Edinburgh
  - Science and Technology Facilities Council — Daresbury Laboratory
- Japan:
  - High Energy Accelerator Research Organization/KEK (2)
- Australia:
  - Victorian Life Sciences Computation Initiative
- Canada:
  - Southern Ontario Smart Computing Innovation Consortium/University of Toronto

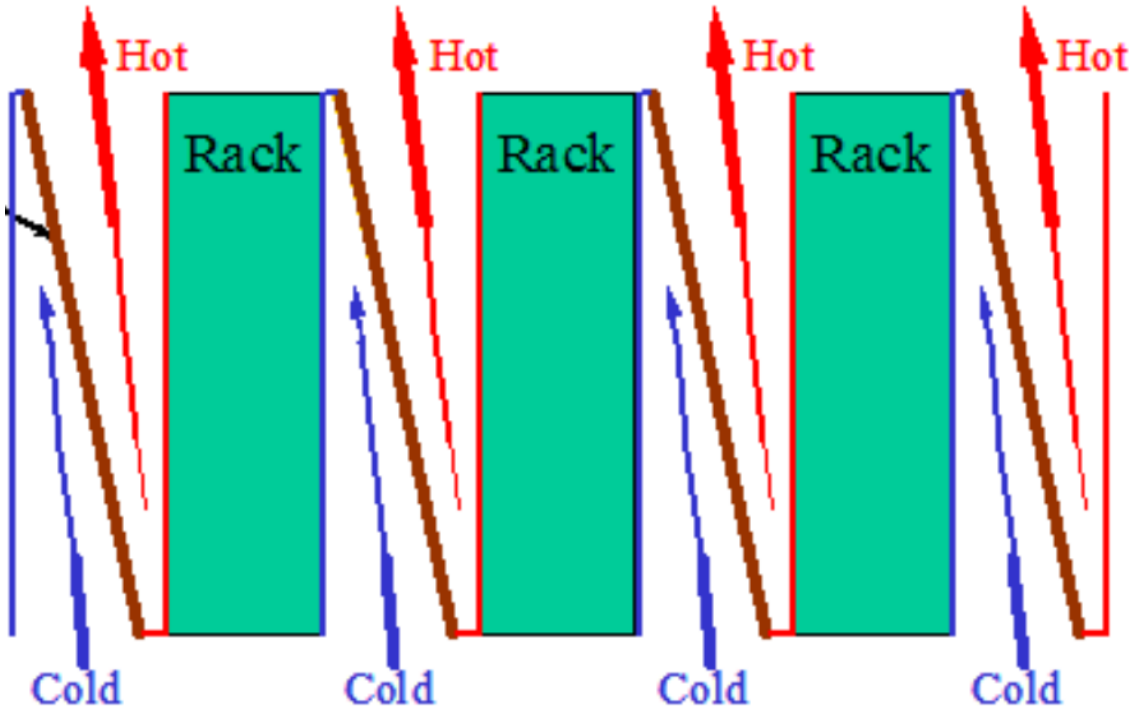
© 2017 IBM Corporation

Outline

1. Blue Gene in the HPC world
2. Blue Gene/P architecture
   - Air-cooling
   - Hardware overview
   - BG/P at Lomonosov MSU
   - Execution process modes
   - Computing node
   - Memory subsystem
   - Networks
3. Blue Gene/P Software
4. Advanced information
5. References
6. Conclusion

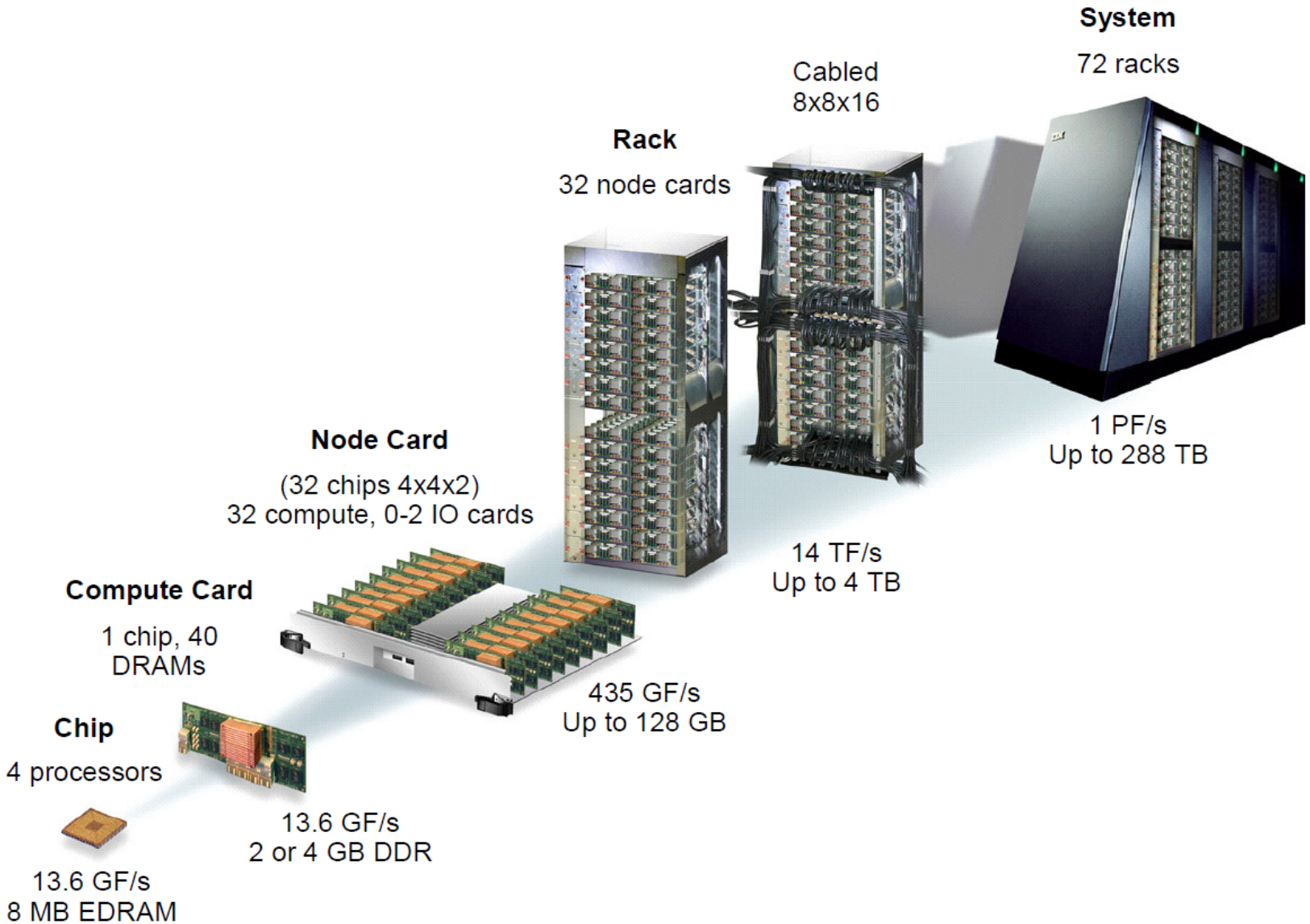
© 2017 IBM Corporation

Blue Gene/P air-cooling



© 2017 IBM Corporation

Blue Gene/P hardware overview



© 2017 IBM Corporation

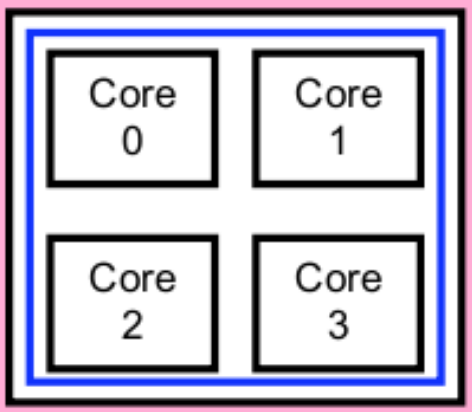
Blue Gene/P at Lomonosov MSU

- Two racks
  - 2048 nodes
  - 8192 cores
- 4 TB RAM
  - 2 GB per node
  - 0.5 GB per core
- Performance:
  - 27.2 TFLOPS peak
  - 23.2 TFLOPS on LINPACK (85% peak)

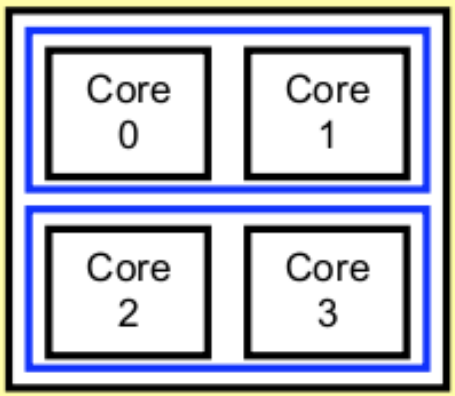
© 2017 IBM Corporation

Execution process modes

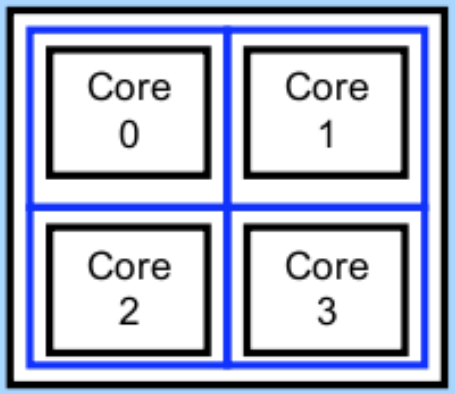
SMP Mode
1 Process
1-4 Threads/Process



Dual Mode
2 Processes
1-2 Threads/Process

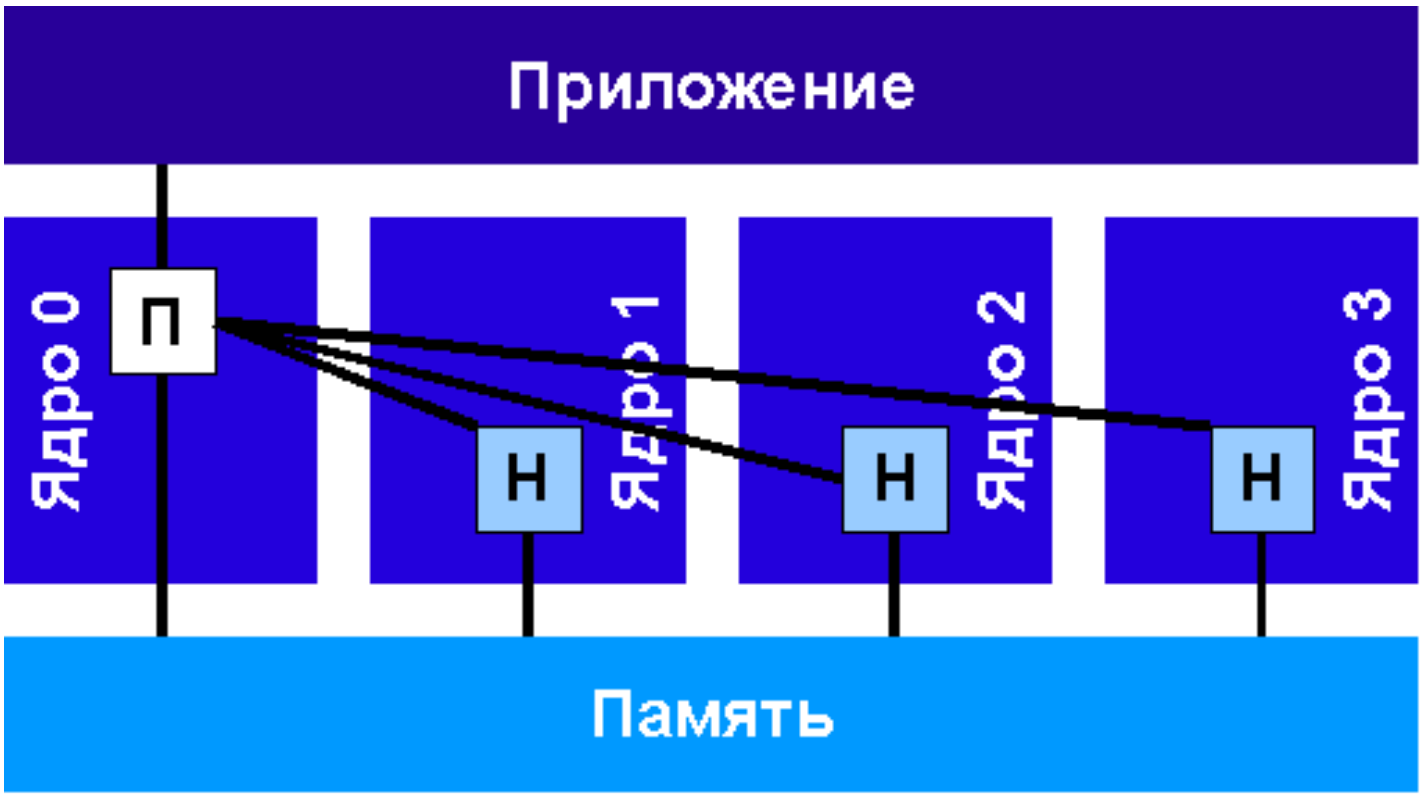


Quad Mode (VNM)
4 Processes
1 Thread/Process



© 2017 IBM Corporation

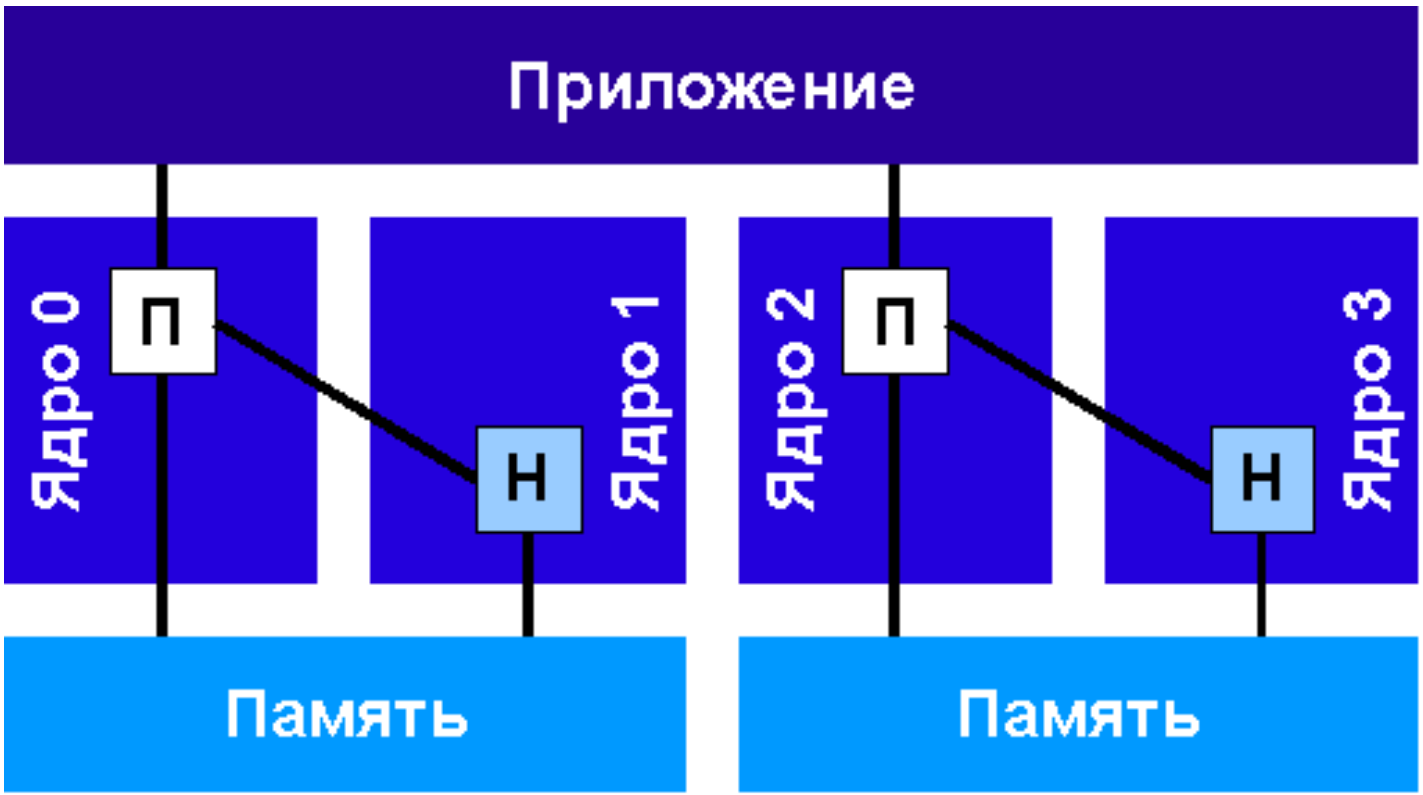
Symmetrical multiprocessing mode (SMP)




```
mpisubmit.bg -n 32 -m SMP -e "OMP_NUM_THREADS=4" a.out
```


© 2017 IBM Corporation

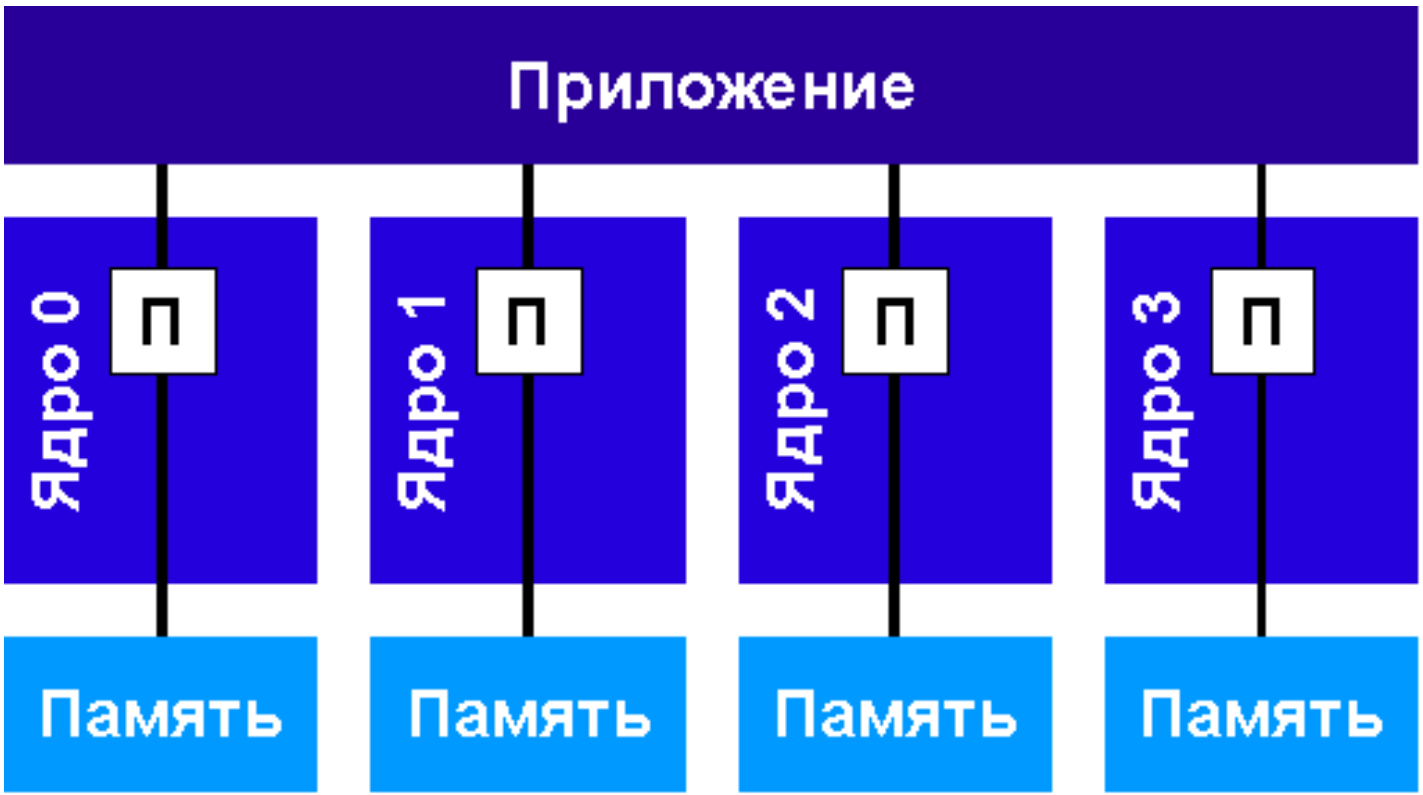
Dual node mode (DUAL)




```
mpisubmit.bg -n 32 -m DUAL -e "OMP_NUM_THREADS=2" a.out
```


© 2017 IBM Corporation

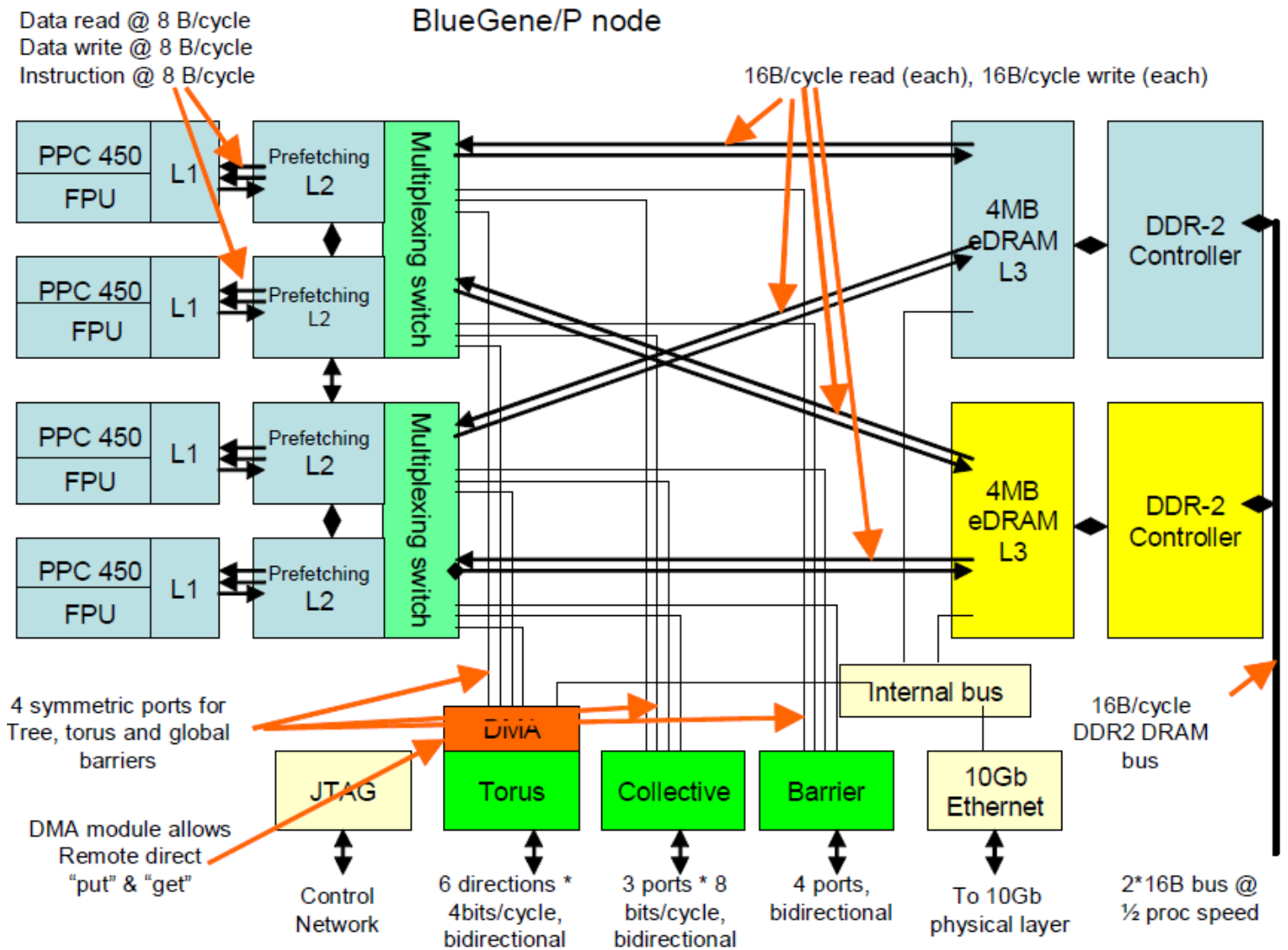
Virtual node mode (VN)




```
mpisubmit.bg -n 32 -m VN -e "OMP_NUM_THREADS=1" a.out
```


© 2017 IBM Corporation

Blue Gene/P computing node



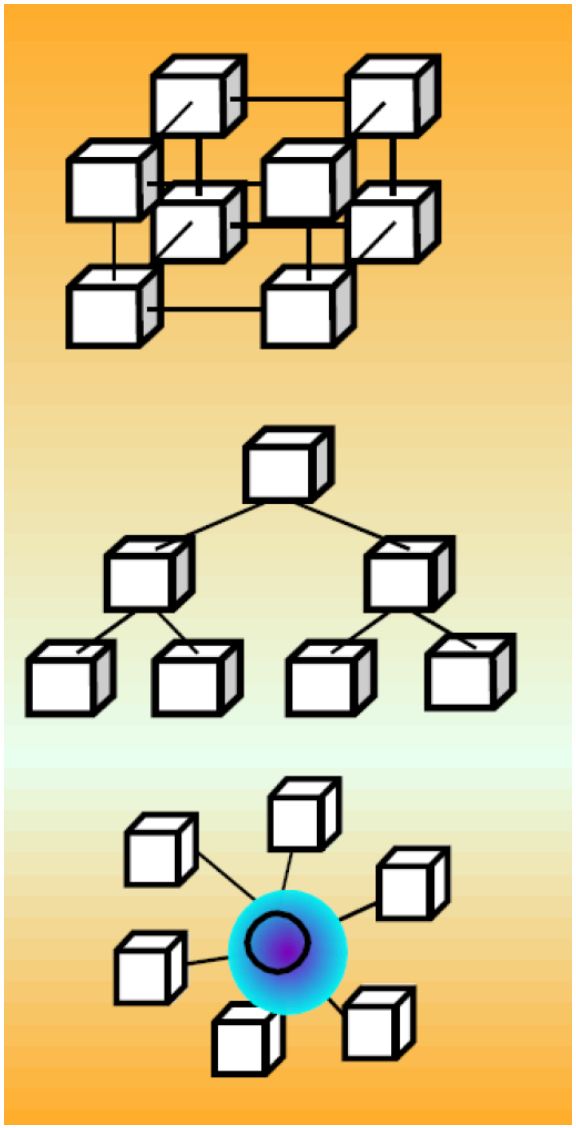
© 2017 IBM Corporation

Blue Gene/P memory subsystem

Cache	Total per node	Size	Replacement policy	Associativity
L1 instruction	4	32 KB	Round-Robin	▸ 64-way set-associative ▸ 16 sets ▸ 32-byte line size
L1 data	4	32 KB	Round-Robin	▸ 64-way set-associative ▸ 16 sets ▸ 32-byte line size
L2 prefetch	4	14 x 256 bytes	Round-Robin	▸ Fully associative (15-way) ▸ 128-byte line size
L3	2	2 x 4 MB	Least recently used	▸ 8-way associative ▸ 2 bank interleaved ▸ 128-byte line size
Double data RAM (DDR)	2	▸ Minimum 2 x 512 MB ▸ Maximum 4 GB	N/A	▸ 128-byte line size

© 2017 IBM Corporation

Blue Gene/P networks



- 3D torus
  - 3.4 Gbit/s per each of 12 ports (5.1 GB/s per node)
  - Hardware latency: 0.5 ms (nearest neighbours), 5 ms (round-trip)
  - MPI latency: 3 ms, 10 ms
  - Main communication network
  - Many collectives take benefit of it
- Collectives (tree)
  - Global one-to-all communications (broadcast, reduction)
  - 6.8 Gbit/s per port
  - Round-trip latency: 1.3 ms (hardware), 5 ms (MPI)
  - Connects computing nodes and I/O nodes
  - Collectives on `MPI_COMM_WORLD`
- Global interrupts
  - Worst-case latency: 0.65 ms (hardware), 5 ms (MPI)
  - `MPI_Barrier()`

© 2017 IBM Corporation

Outline

1. Blue Gene in the HPC world
2. Blue Gene/P architecture
3. Blue Gene/P Software
   - Software
   - Compilers
   - Managing jobs
   - Quick reference
4. Advanced information
5. References
6. Conclusion

© 2017 IBM Corporation

Blue Gene/P software

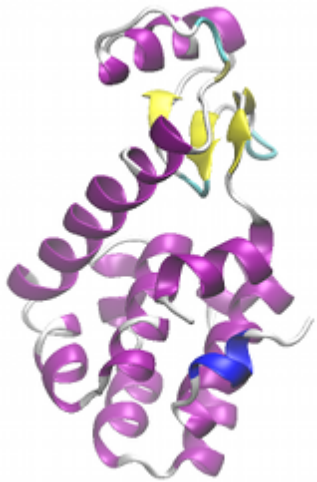
System software

- IBM XL and GCC compilers
- ESSL mathematical library (BLAS, LAPACK, etc.)
- MASS mathematical library
- LoadLeveler job scheduler

Application software and libraries

- FFTW
- GROMACS
- METIS
- ParMETIS
- szip
- zlib

`http://hpc.cmc.msu.ru/bgp/soft`



© 2017 IBM Corporation

Blue Gene/P compilers

Compiler		C/C++		Fortran			
		C	C++	Fortran 77	Fortran 90	Fortran 95	Fortran 2003
GCC		mpicc	mpicxx	mpif77	—	—	—
IBM XL	non-threaded	mpixlc	mpixlcxx	mpixlf77	mpixlf90	mpixlf95	mpixlf2003
	thread-safe	mpixlc_r	mpixlcxx_r	mpixlf77_r	mpixlf90_r	mpixlf95_r	mpixlf2003_r

Cross-compiling!

© 2017 IBM Corporation

IBM XL compilers flags

- `-qarch=450 -qtune=450`
  - exploit only one of two FPUs
  - use when data is not 16-byte aligned
- `-qarch=450d -qtune=450`
  - alignment!
- `-O3` (`-qstrict`)
  - minimal optimiazation level for Double Hammer FPU
- `-O3 -qhot`
- `-O4` (`-qnoipa`)
- `-O5`
- `-qreport -qlist -qsource`
  - a lot of useful information in `.lst` file
- `-qsmp=omp`, `-qsmp=auto`

© 2017 IBM Corporation

Job submission

`mpisubmit.bg [SUBMIT_OPTIONS] EXECUTABLE [-- EXE_OPTIONS]`

- `-n NUMBER` — number of compute nodes
- `-m MODE` — mode (`SMP`, `DUAL`, `VN`)
- `-w HH:MM:SS` — maximum wallclock time
- `-e ENV_VARS` — environment variables

Example:

- `$> mpisubmit.bg -n 256 -m DUAL -w 00:05:00 -e "OMP_NUM_THREADS=2 SOME_ENV_VAR=8x" a.out -- 3.14`

More info:

- `$> mpisubmit.bg --help`
- `http://hpc.cmc.msu.ru/bgp/jobs/mpisubmit`

© 2017 IBM Corporation

Job input and output

- Default stdout: `<exec>.$(jobid).out`
- Default stderr: `<exec>.$(jobid).err`
- Customization (`mpisubmit.bg`):
  - `--stdout STDOUT_FILE_NAME`
  - `--stderr STDERR_FILE_NAME`
  - `--stdin STDIN_FILE_NAME`

© 2017 IBM Corporation

Working with a job scheduler queue

- `llq` — retrieve job queue info
- `llq -l JOB_ID` — retrieve info about specific job
- `llmap` — retrieve job queue info (“Blue Gene/P style”)
- `llcancel JOB_ID` — cancel the job

© 2017 IBM Corporation

Quick reference: Compilation

MPI + OpenMP:

- **C**: `mpixlc_r -qsmp=omp -O3 hello.c -o hello`
- **C++**: `mpixlcxx_r -qsmp=omp -O3 hello.cxx -o hello`
- **Fortran**: `mpixlf90_r -qsmp=omp -O3 hello.F90 -o hello`

© 2017 IBM Corporation

Quick reference: Job submission

MPI + OpenMP, 32 nodes, 15 mins

- **SMP** (4 cores per MPI process):
  ```
  mpisubmit.bg -n 32 -m SMP -w 00:15:00 -e
  "OMP_NUM_THREADS=4" hello
  ```
- **DUAL** (2 cores per MPI process):
  ```
  mpisubmit.bg -n 32 -m DUAL -w 00:15:00 -e
  "OMP_NUM_THREADS=2" hello
  ```
- **VN** (1 core per MPI process):
  ```
  mpisubmit.bg -n 32 -m VN -w 00:15:00 -e
  "OMP_NUM_THREADS=1" hello
  ```


© 2017 IBM Corporation

Quick reference: Job queue

- Job queue status
  - `llq`
  - `llmap`
- Cancel the job
  - `llcancel JOB_ID`

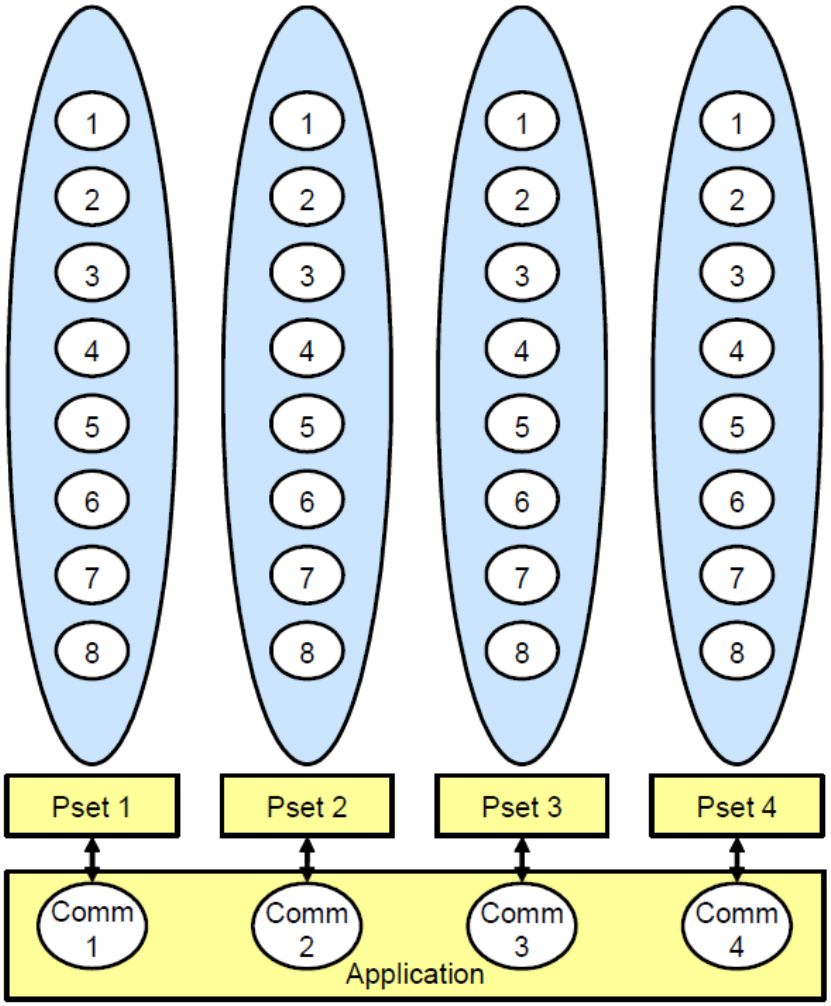
© 2017 IBM Corporation

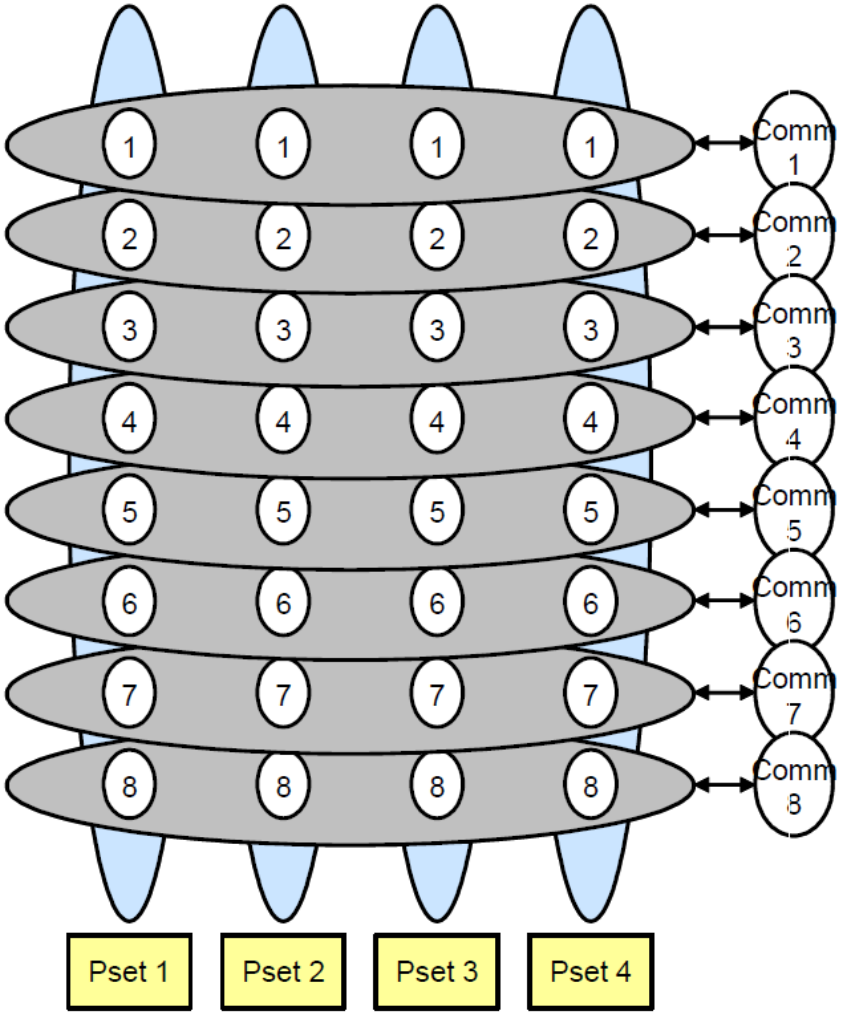
Outline

1. Blue Gene in the HPC world
2. Blue Gene/P architecture
3. Blue Gene/P Software
4. Advanced information
   - I/O nodes psets
   - Double hammer FPU
5. References
6. Conclusion

© 2017 IBM Corporation

Grouping processes in respect to I/O nodes



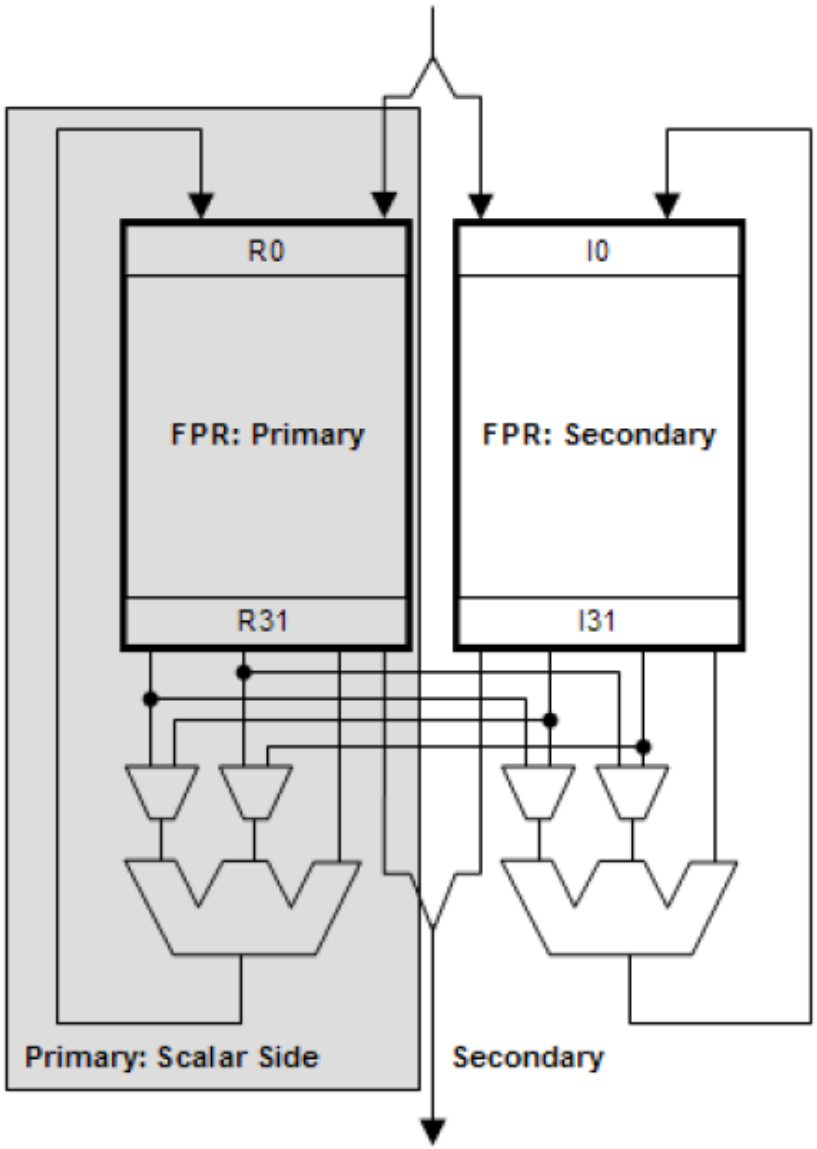


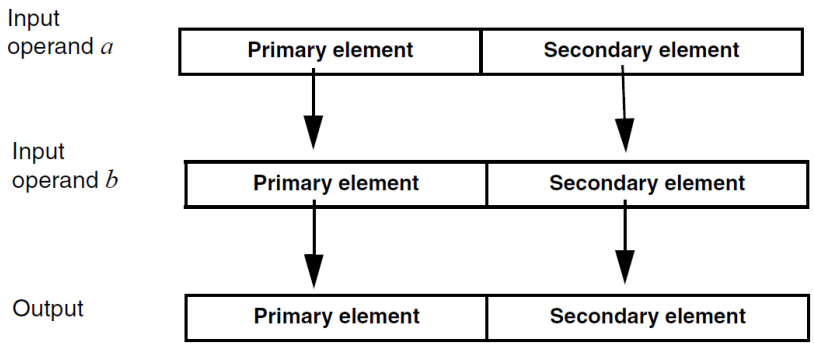

```
int MPIX_Pset_same_comm_create (MPI_Comm *pset_comm);
```

```
int MPIX_Pset_diff_comm_create (MPI_Comm *pset_comm);
```


© 2017 IBM Corporation

Double hammer FPU






```
#define MY_ALIGN __attribute__((aligned(16)))

int main() {
    double MY_ALIGN a[2] = { 2, 3 };
    double MY_ALIGN b[2] = { 4, 5 };
    double MY_ALIGN c[2] = { 6, 7 };
    double MY_ALIGN r[2]; // 2 * 4 + 6 = 14
                          // 3 * 5 + 7 = 22

    double _Complex ra = __lfpd(a);
    double _Complex rb = __lfpd(b);
    double _Complex rc = __lfpd(c);

    double _Complex rr = __fpmadd(rc, rb, ra);

    __stfpd(r, rr);
}
```

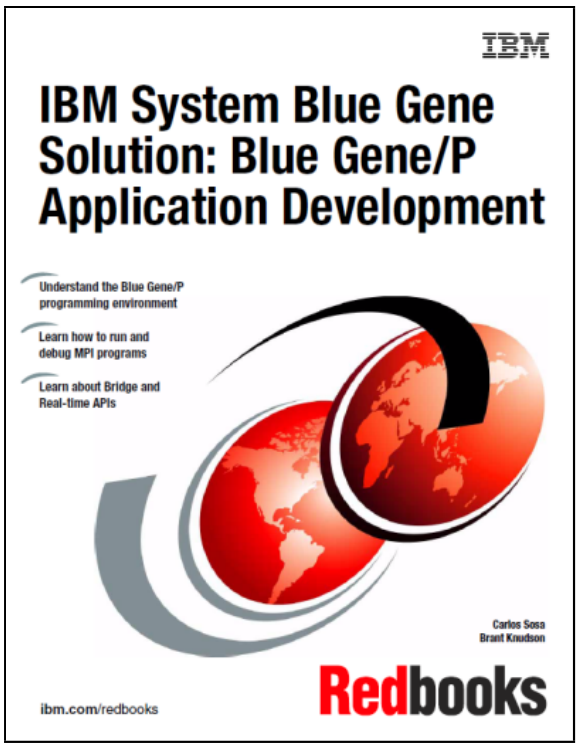

© 2017 IBM Corporation

Outline

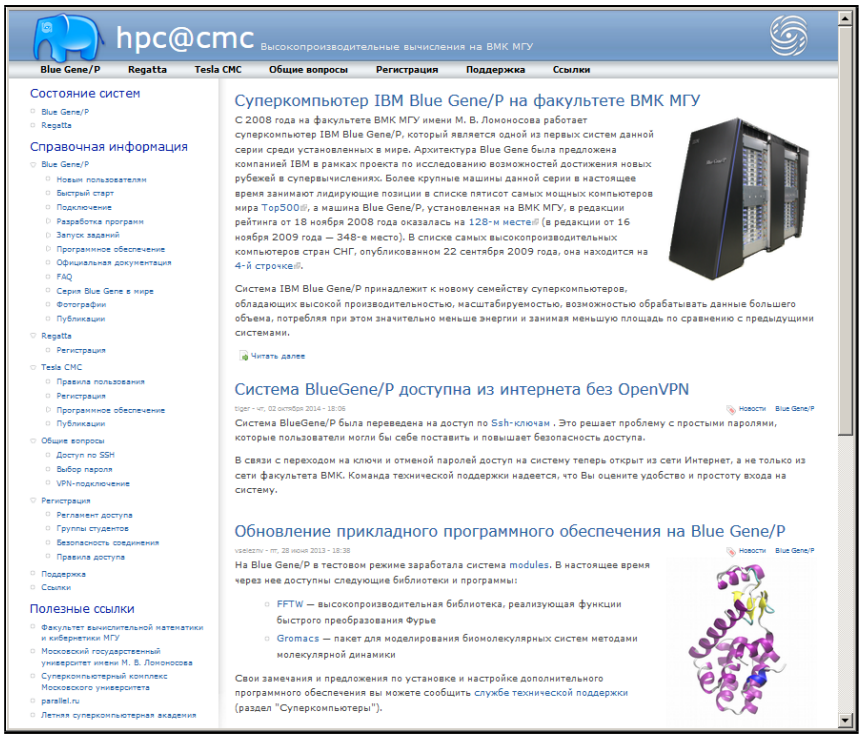
1. Blue Gene in the HPC world
2. Blue Gene/P architecture
3. Blue Gene/P Software
4. Advanced information
5. References
6. Conclusion

© 2017 IBM Corporation

References



http://www.redbooks.ibm.com/abstracts/sg247287.html



http://hpc.cmc.msu.ru

© 2017 IBM Corporation

Outline

1. Blue Gene in the HPC world
2. Blue Gene/P architecture
3. Blue Gene/P Software
4. Advanced information
5. References
6. Conclusion

© 2017 IBM Corporation

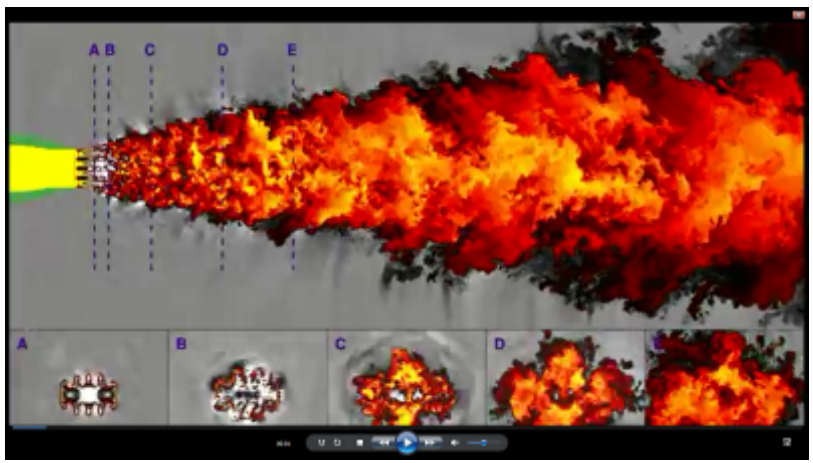
Conclusion

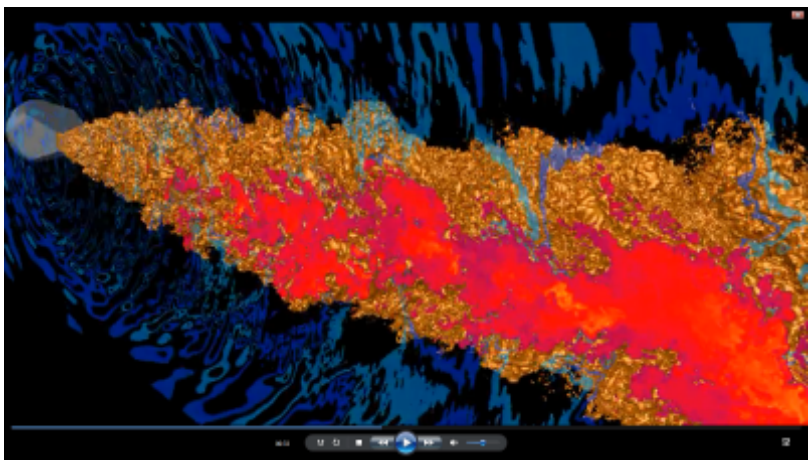
- Brief history of Blue Gene
  - BG/L, BG/P, BG/Q
  - National Medal of Technology and Innovation
  - TOP500, Graph500
  - Eight Gordon Bell awards
- Blue Gene/P architecture
  - Four cores per node
  - 2 GB RAM per node
  - 1024 nodes per rack
  - Two BG/P racks at Lomonosov MSU
  - Compilers and software

© 2017 IBM Corporation

Jet engine noise CFD simulation

- Supersonic jet noise simulation, effects of chevrons on jet noise
- 128k Blue Gene/P cores — ≈ 100 hours
- 1M Blue Gene/Q cores — ≈ 12 hours
- IBM Blue Gene Q Sequoia, Lawrence Livermore National Laboratory
- `https://youtu.be/cjoz5tncRUs`
- `https://youtu.be/uxT-VmY3OWc`
- Video: Joseph W. Nichols, Stanford Center for Turbulence Research

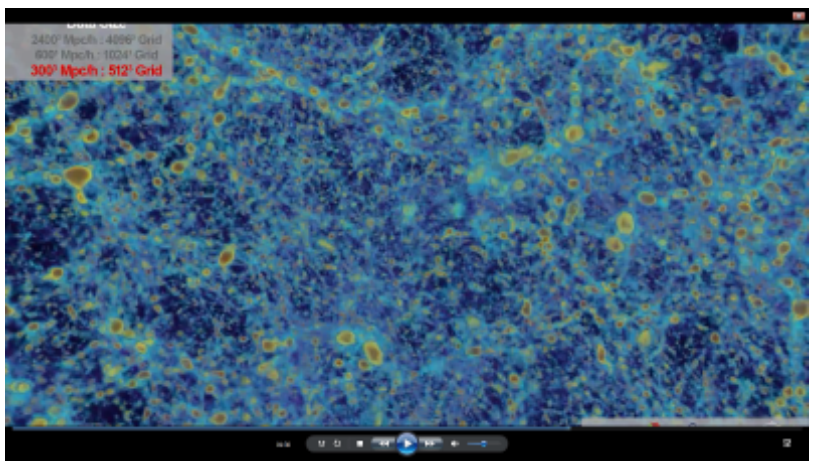


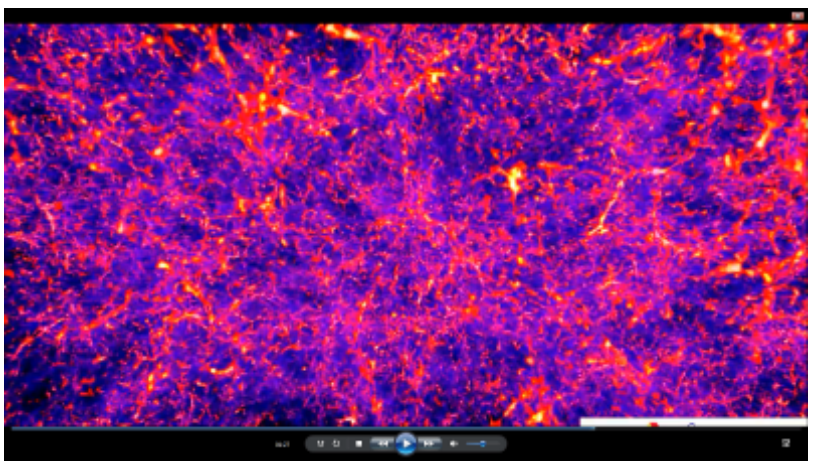


© 2017 IBM Corporation

Secrets of the Dark Universe

- The evolution of the Universe simulation, understanding the physics of the dark matter and energy
- 1 BG/Q rack — 68B particles
- 32 BG/Q racks — 1.1T particles
- `http://www.youtube.com/watch?v=tdv8yrJk4VE`
- `http://www.youtube.com/watch?v=-S-T_iTiAxQ`
- Video: ANL, LANL, LBNL

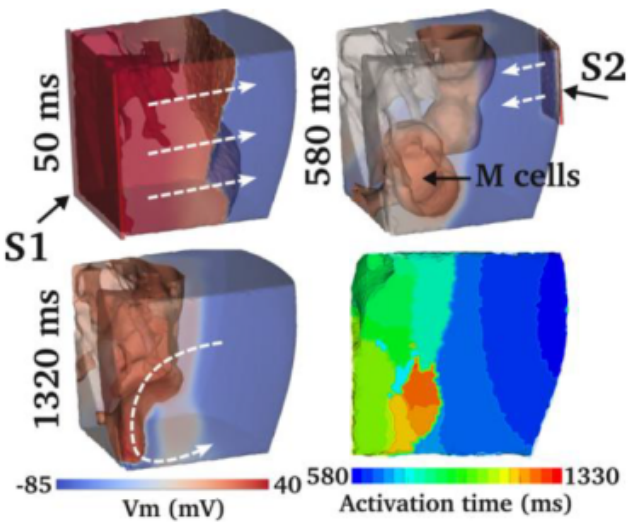


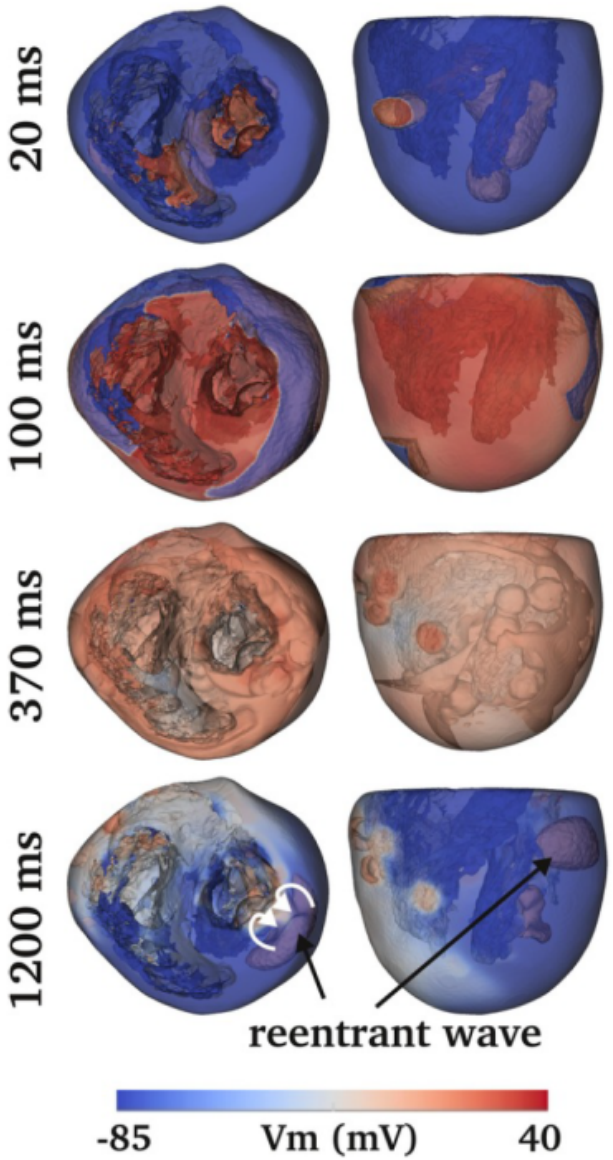


© 2017 IBM Corporation

Real-time modeling of human heart ventricles

- Simulation of drug-induced arrhythmias
- Resolution — 0.1 mm
- 768k Blue Gene/Q cores
- 43% peak
- http://dl.acm.org/citation.cfm?id=2388999
- LLNL, IBM Research, IBM Research Collaboratory for Life Sciences

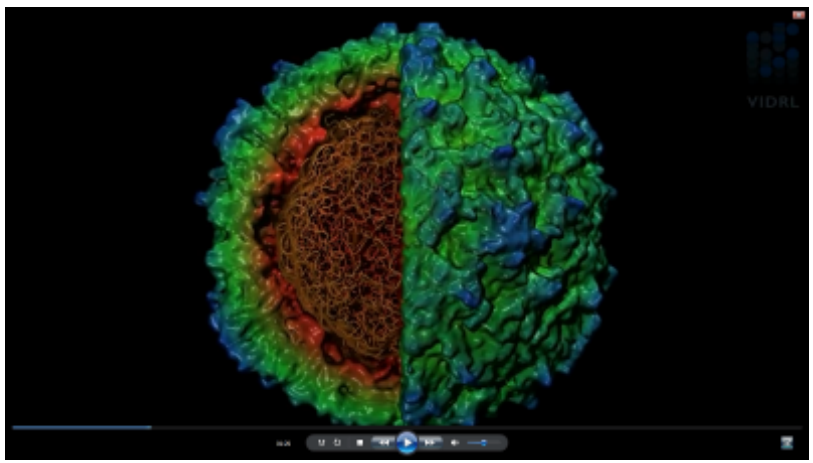


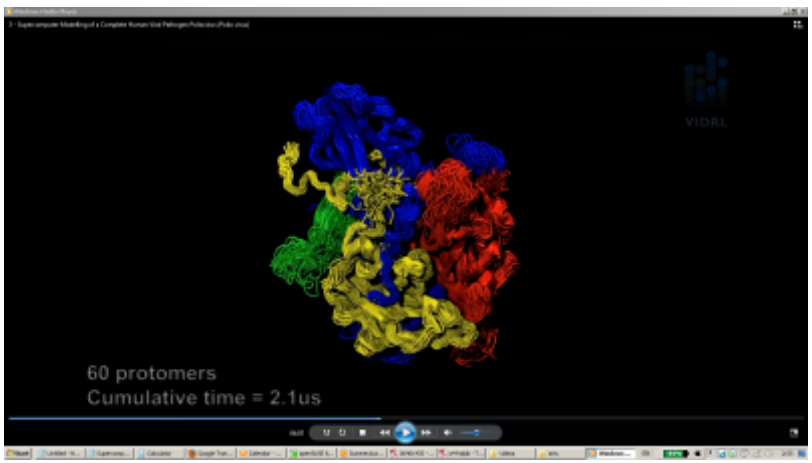


© 2017 IBM Corporation

Modelling of a complete human viral pathogen poliovirus

- Reconstruction and simulation of poliovirus
- Antiviral drugs, virus infection, modelling related viruses
- 3.3M–3.7M atoms
- Blue Gene/Q, Victorian Life Sciences Computing Initiative
- `http://www.youtube.com/watch?v=Nih0Qa673FY`

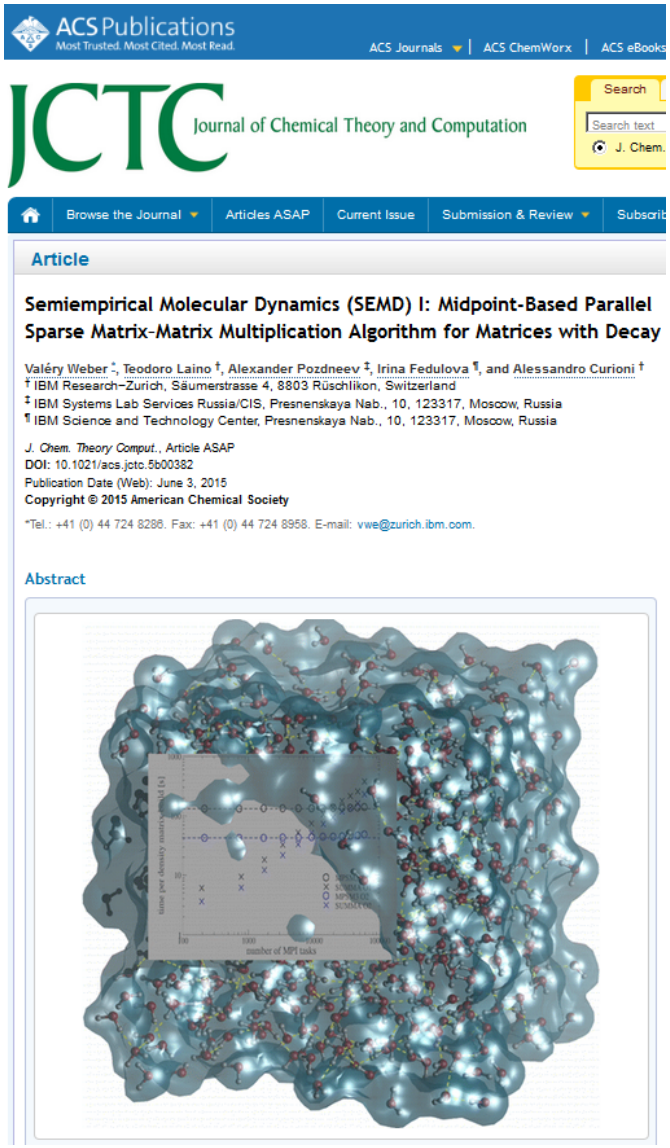




© 2017 IBM Corporation

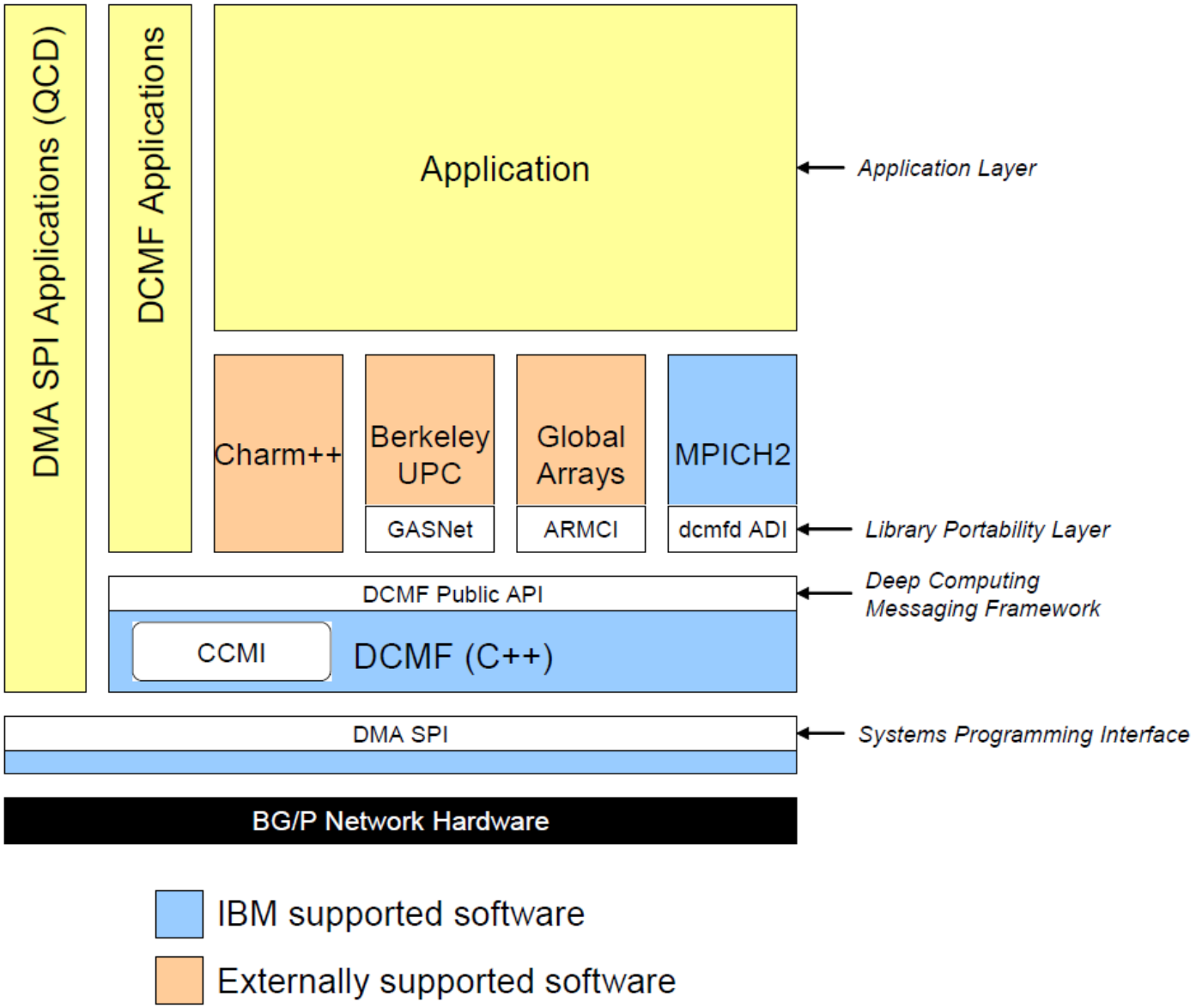
Parallel sparse matrix–matrix multiplication

- Semiempirical Molecular Dynamics (SEMD) I: Midpoint-Based Parallel Sparse Matrix–Matrix Multiplication Algorithm for Matrices with Decay
- Inspired by the midpoint method
- Approaching a perfect linear scaling
- Up to 185 193 processes on BG/Q
- doi: 10.1021/acs.jctc.5b00382



© 2017 IBM Corporation

Blue Gene/P APIs



© 2017 IBM Corporation

Disclaimer

All the information, representations, statements, opinions and proposals in this document are correct and accurate to the best of our present knowledge but are not intended (and should not be taken) to be contractually binding unless and until they become the subject of separate, specific agreement between us.

Any IBM Machines provided are subject to the Statements of Limited Warranty accompanying the applicable Machine.

Any IBM Program Products provided are subject to their applicable license terms. Nothing herein, in whole or in part, shall be deemed to constitute a warranty.

IBM products are subject to withdrawal from marketing and or service upon notice, and changes to product configurations, or follow-on products, may result in price changes.

Any references in this document to "partner" or "partnership" do not constitute or imply a partnership in the sense of the Partnership Act 1890.

IBM is not responsible for printing errors in this proposal that result in pricing or information inaccuracies.

© 2017 IBM Corporation

Правовая информация

IBM, логотип IBM, BladeCenter, System Storage и System x являются товарными знаками International Business Machines Corporation в США и/или других странах. Полный список товарных знаков компании IBM смотрите на узле Web: `www.ibm.com/legal/copytrade.shtml`.

Названия других компаний, продуктов и услуг могут являться товарными знаками или знаками обслуживания других компаний.

(c) 2016 International Business Machines Corporation. Все права защищены.

Упоминание в этой публикации продуктов или услуг корпорации IBM не означает, что IBM предполагает предоставлять их во всех странах, в которых осуществляет свою деятельность, информация о предоставлении продуктов или услуг может быть изменена без уведомления. За самой свежей информацией о продуктах и услугах компании IBM, предоставляемых в Вашем регионе, следует обращаться в ближайшее торговое представительство IBM или к авторизованным бизнес-партнерам.

Все заявления относительно намерений и перспективных планов IBM могут быть изменены без уведомления.

Информация о продуктах третьих фирм получена от производителей этих продуктов или из опубликованных анонсов указанных продуктов. IBM не тестировала эти продукты и не может подтвердить производительность, совместимость, или любые другие заявления относительно продуктов третьих фирм. Вопросы о возможностях продуктов третьих фирм следует адресовать поставщику этих продуктов.

Информация может содержать технические неточности или типографические ошибки. В представленную в публикации информацию могут вноситься изменения, эти изменения будут включаться в новые редакции данной публикации. IBM может вносить изменения в рассматриваемые в данной публикации продукты или услуги в любое время без уведомления.

Любые ссылки на узлы Web третьих фирм приведены только для удобства и никоим образом не служат поддержкой этим узлам Web. Материалы на указанных узлах Web не являются частью материалов для данного продукта IBM.

© 2017 IBM Corporation