MODERN HARDWARE ARCHITECTURES

Tulin Kaman

Department of Applied Mathematics and Statistics

Stony Brook/BNL New York Center for Computational Science

tkaman@ams.sunysb.edu

Aug 23, 2012

- Listing the 500 most powerful computers in the world
- Updated twice a year
  - June : International Supercomputing Conference in Germany
  - November : Supercomputing Conference SC
- computers ranked by their performance on the LINPACK Benchmark
- http://www.top500.org

	NAME	SPECS	SITE	COUNTRY	CORES	R_{max} Pflop/s
1	**Sequoia**	IBM BlueGene/Q, Power BQC 16C 1.60 GHz, Custom interconnect	DOE / NNSA / LLNL	USA	1,572,864	16.33
2	**K computer**	Fujitsu SPARC64 VIIIfx 2.0GHz, Tofu interconnect	RIKEN AICS	Japan	705,024	10.51
3	**Mira**	IBM BlueGene/Q, Power BQC 16C 1.60 GHz, Custom interconnect	DOE / SC / ANL	USA	786,432	8.153
4	**SuperMUC**	IBM iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband QDR	Leibniz Rechenzentrum	Germany	147,456	2.897
5	**Tianhe-1A**	NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050	NUDT/NSCC/Tianjin	China	186,368	2.566

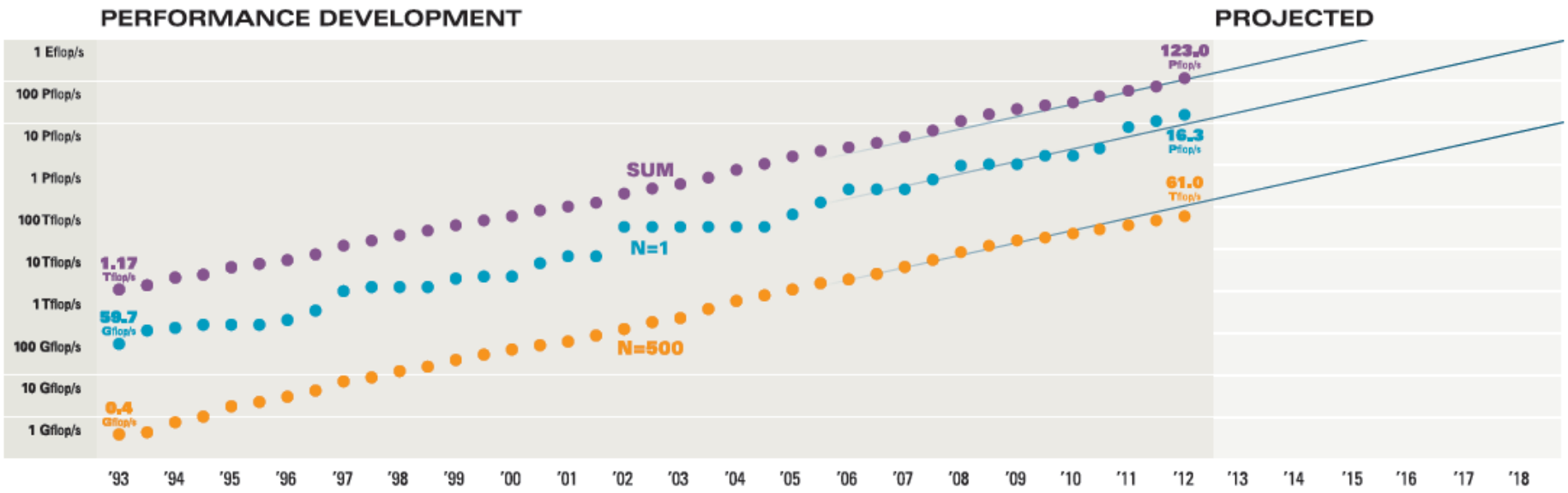
Sequoia
the world's top supercomputer

- For the first time since November 2009, a United States supercomputer sits atop the TOP500 list of the world's top supercomputers.

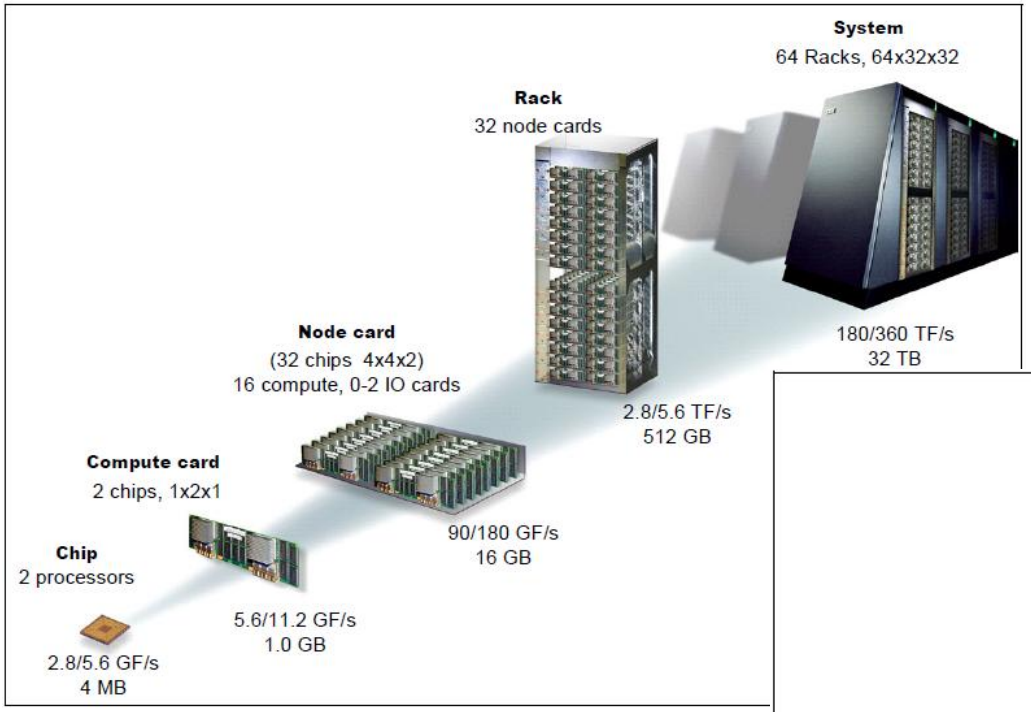
- Sequoia, the IBM BlueGene/Q system installed at the Department of Energy's Lawrence Livermore National Laboratory achieved 16.32 petaflop/s on the **LINPACK** benchmark using 1,572,864 cores.

- LINPACK: They measure how fast a computer solves a system of linear equations $Ax = b$

PERFORMANCE DEVELOPMENT



Blue Gene L/P system
from the microprocessor to the full system

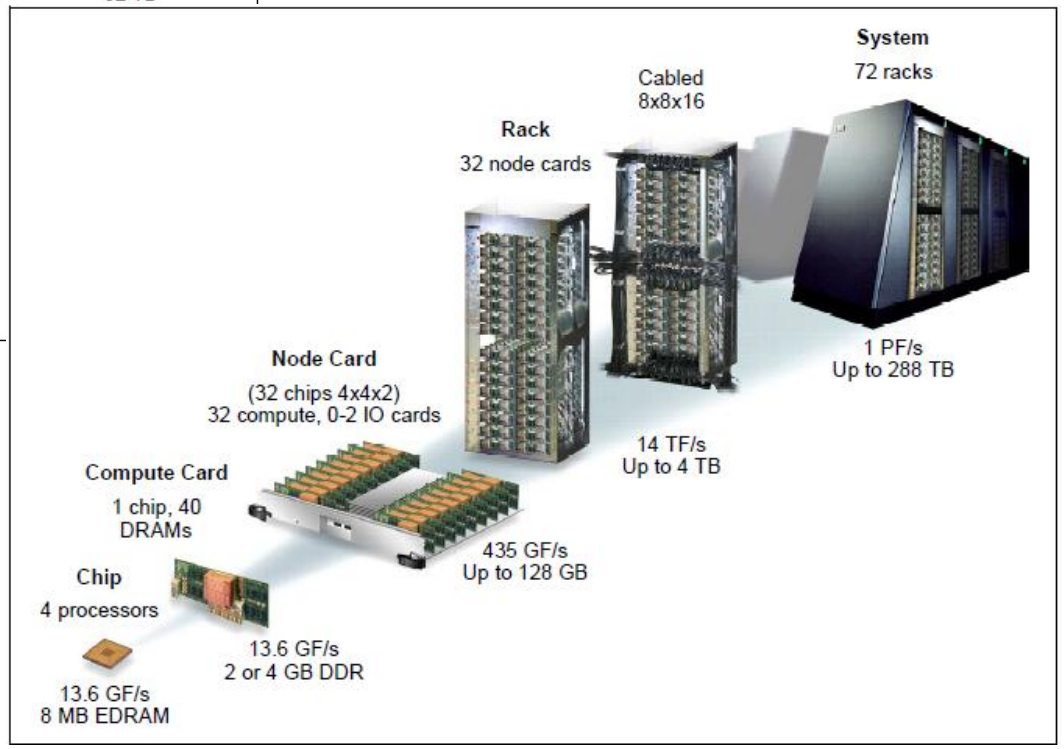


New York Blue/L : 18 racks

18432 compute nodes (36864 CPUs)

Blue Gene L

Blue Gene P



New York Blue/P : 2 racks

2048 compute nodes (8192 CPUs)

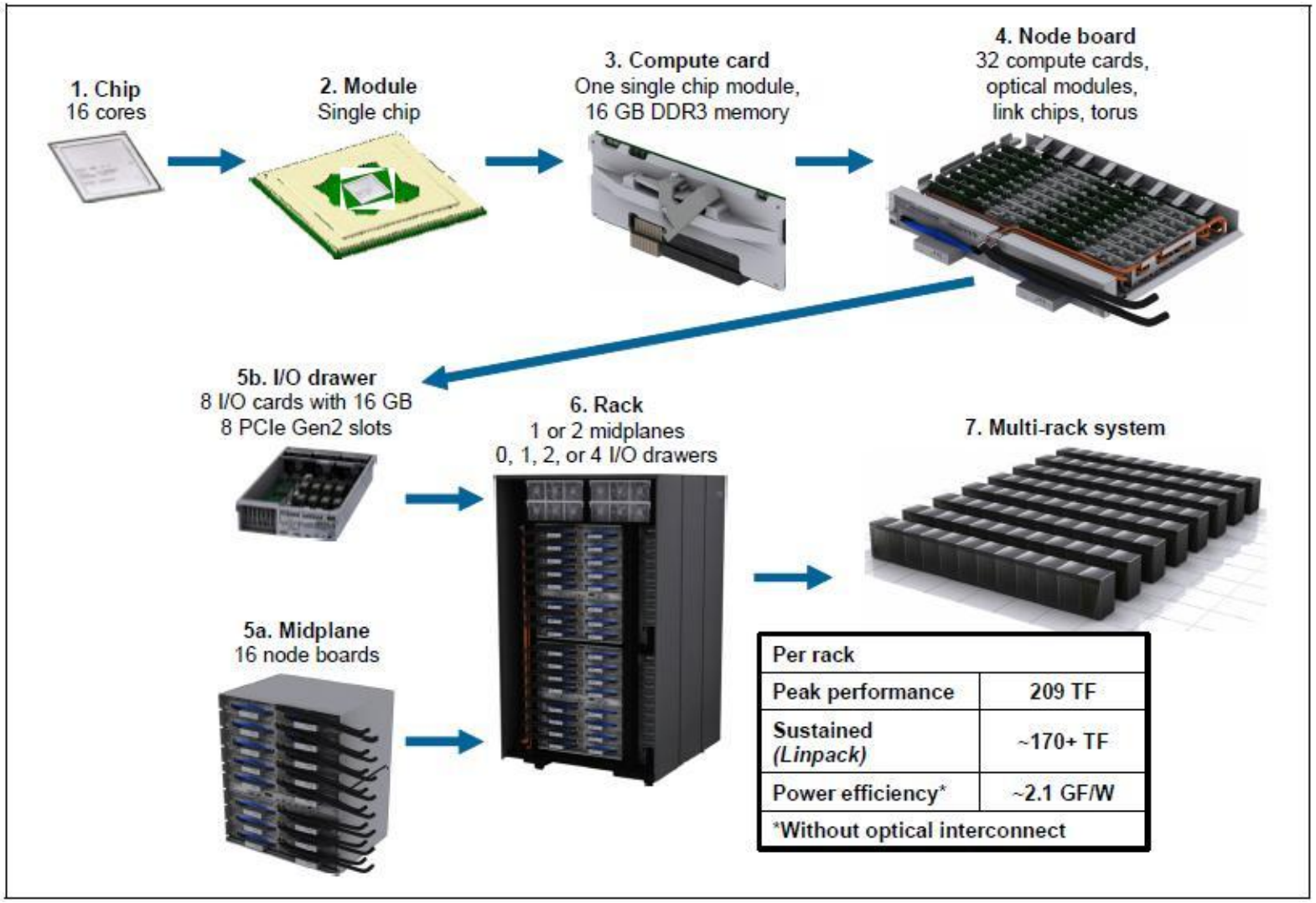
Comparison between BG L and BG P

Node		
Cores per node	2	4
Core clock speed	700 MHz	850 MHz
Cache coherency	Software managed	SMP
Private L1 cache	32 KB per core	32 KB per core
Private L2 cache	14 stream prefetching	14 stream prefetching
Shared L3 cache	4 MB	8 MB
Physical memory per node	512 MB-1 GB	2 GB or 4 GB
Main memory bandwidth	5.6 GBps	13.6 GBps
Peak performance	5.6 GFLOPS per node	13.6 GLOPS per node

The theoretical
peak peformance = (CPU speed) x (CPU Instruction per cycle) x (number of cores per node)
(per node)

Blue Gene Q system

from the microprocessor to the full system

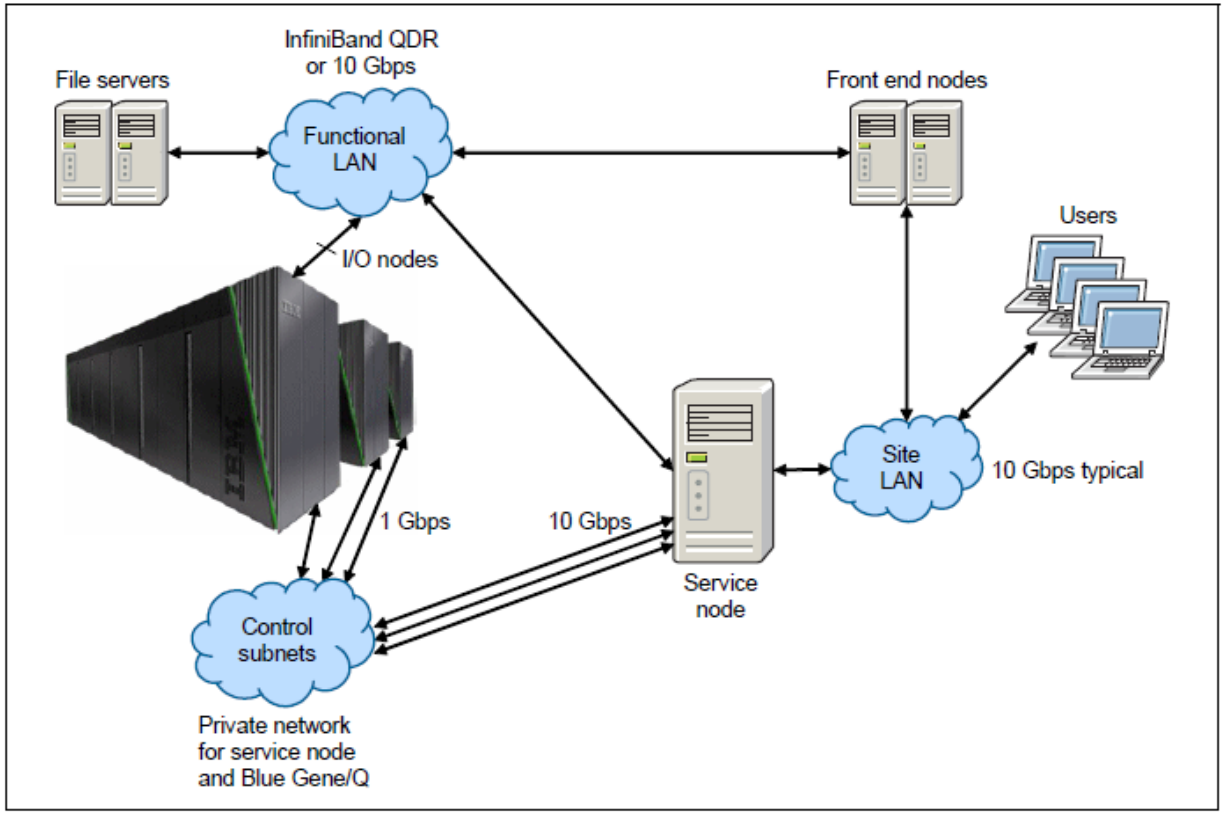


Per rack	
Peak performance	209 TF
Sustained *(Linpack)*	~170+ TF
Power efficiency*	~2.1 GF/W
*Without optical interconnect	

Blue Gene Evolution

- **BG/L (5.7TF/rack)**
  - Dual-core system-on-chip,
  - 1 GB / Node
- **BG/P (13.9TF/rack)**
  - Quad core system-on-chip
  - 2 GB / Node
  - SMP support, OpenMP, MPI
- **BG/Q (209TF/rack)**
  - 16 core/64 thread
  - 16 GB / Node

Blue Gene system architecture



Front-end node: provides access users to edit and compile program, create job script file and submit jobs

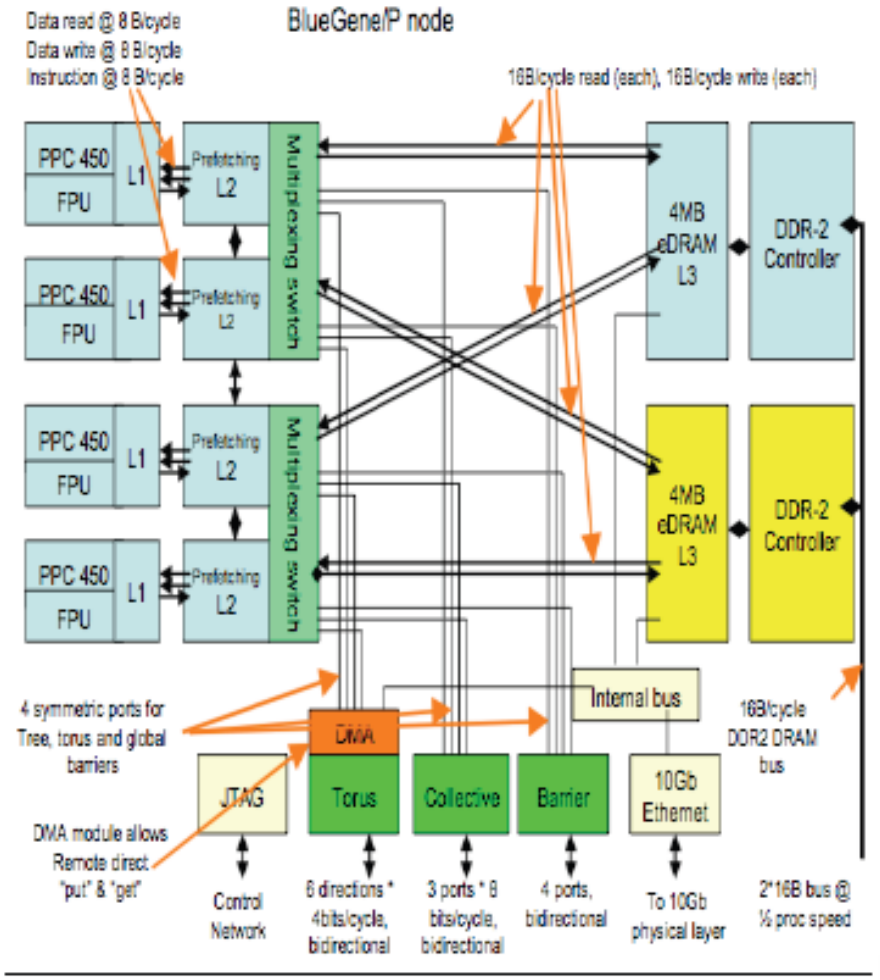
Service node: the manager of the system and used for controlling the system.

I/O nodes (IO): provide access to external devices through an Ethernet port to the 10 gigabit functional network and can perform file I/O operation

Compute nodes (CN): run user application

CN and I/O node properties

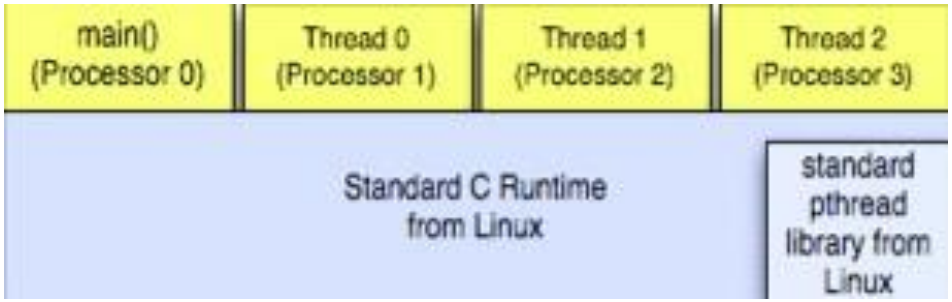
Node processors (compute and I/O)	Quad 450 PowerPC
Processor frequency	850 MHz
Coherency	Symmetrical multiprocessing
L1 Cache (private)	32 KB per core
L2 Cache (private)	14 stream prefetching
L3 Cache size (shared)	8 MB
Main store memory/node	2 GB or 4 GB
Main store memory bandwidth	16 GBps
Peak performance	13.6 GFLOPS (per node)



Execution Process Mode on BG P

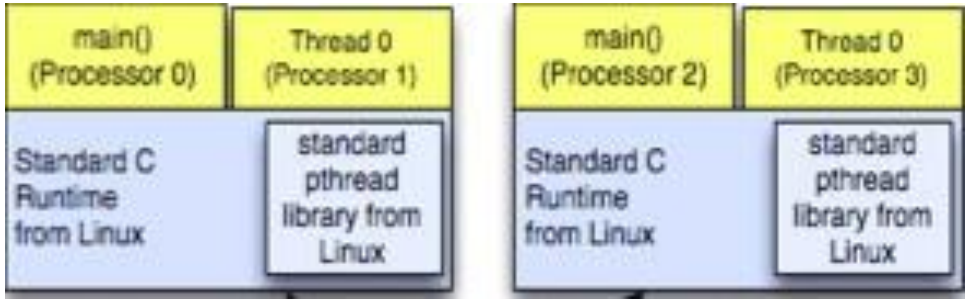
Symmetrical **M**ulti**P**rocessing (SMP) Node Mode

Each compute node executes a single MPI task per node with a maximum of 4 threads



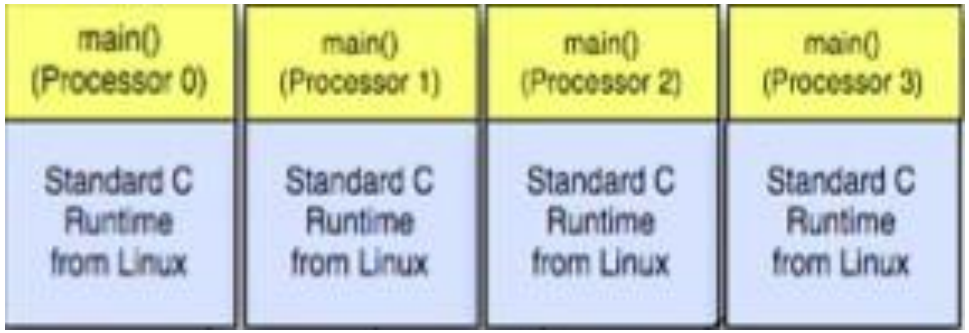
Dual **N**ode (DUAL) Mode

Each compute node executes two MPI tasks per node. Each task in this mode get the half memory. It runs two threads per task.

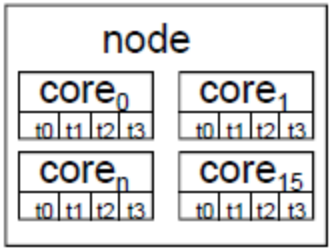


Virtual **N**ode (VN) Mode

In this mode, kernel runs four MPI task on each compute node.



Processes per node in BG Q



The number of threads that each process can have active at a given moment is dictated by the processes per node value.

Processes per node	Number of A2 cores per process	Maximum number of active hardware threads per process
1	16	64
2	8	32
4	4	16
8	2	8
16	1	4
32	2 processes per core	2
64	4 processes per core	1

The default mode for the **runjob** command is one process per node.

runjob ... -p 64 ...

Compiling and linking applications on the Blue Gene/Q system

- GNU compilers: /bgsys/drivers/ppcfloor/gnu-linux/bin

C	powerpc64-bgq-linux-gcc
C++	powerpc64-bgq-linux-gfortran
Fortran	powerpc64-bgq-linux-g++

- XL Compilers: /opt/ibmcmp/vacpp/bg/12.1/bin

C	bgc89, bgc99, bgcc, bgxlc bgc89_r, bgc99_r, bgcc_r, bgxlc_r
C++	bgxlc++, bgxlc++_r, bgxlC, bgxlC_r
Fortran	bgf2003, bgf95, bgxlf2003, bgxlf90_r, bgxlf_r, bgf77, bgfort77, bgxlf2003_r, bgxlf95, bgf90, bgxlf, bgxlf90, bgxlf95_r

Compiling MPI programs with wrappers

- Library versions are installed in /bgsys/drivers/ppcfloor/comm

- Scripts compile and link MPI programs
  - **GNU Compilers:** /bgsys/drivers/ppcfloor/comm/bin/
    mpicc, mpicxx, mpif77, mpif90
  - **XL Compilers:** /bgsys/drivers/ppcfloor/comm/xl/bin
    mpixlc, mpixlcxx, mpixlf77, mpixlf90, mpixlf95, mpixlf2003
    mpixlc_r, mpixlcxx_r, mpixlf77_r, mpixlf90_r, mpixlf95_r, mpixlf2003_r

XL Compilers Options

- These options identify that the code is targeted

-qarch=440 -qtune=440 : for Blue Gene/L

-qarch=450 -qtune=450 : for Blue Gene/P

-qarch=qp -qtune=qp : for Blue Gene/Q

- Default options for XL compilers for BG/Q only

-q64 : The Blue Gene/Q compilers generate only 64-bit code

-qsimd=auto : indicate whether the compiler transforms code into a form that can use the floating-point instruction set

IBM XL Compilers Optimization Options

- IBM XL compilers have level of optimization support.
  - Basic command-line optimization

    -O0 = -qsimd=auto

    -O2 = -O0 -qmaxmem=8192 -qsimd=auto
  - Advanced command-line optimization

    -O3 = -O2 -qnostrict -qmaxmem=-1 -qhot=level=0 -qsimd=auto

    -O4 = -O3 -qhot –qipa -qarch=auto -qtune=auto -qcache=auto -qsimd=auto

    -O5 = "All of -O4" -qipa=level=2

    - -qsimd=auto : for BG/Q only
    - -qhot : enables and customizes high-order loop analysis and transformation.
    - -qipa : enables and customizes interprocedural analysis
    - -qsmp = omp : enables OpenMP

- Make sure that the application is first compiled and executed properly at low optimization levels.

How to program machines that can be built?

- Distributed - Memory Machines
  - Each node in the computer has a locally addressable memory space.
  - Parallel programs consists of cooperating processes, each with its own memory.
  - Processes send data to one another as messages. **Message Passing Interface (MPI)** is a widely used standard for writing message-passing programs.
- Shared - Memory Machines
  - Each core can access the entire data space.
  - In shared memory multi-core architectures, **OpenMP, Pthreads** can be used to implement parallelism.

New York Blue

- **18 racks Blue Gene L**
  - 18432 (18 x 1024) dual-processor Compute Nodes
  - 700 MHz PowerPC440 processors (a total of 36864)
  - The two cores on a chip share a 1 GB of DDR memory
- **2 racks Blue Gene P**
  - 2048 (2 x 1024) quad core-processor Compute Nodes
  - 850 MHz PowerPC450 processors (a total of 8192)
  - The four cores on a chip share a 2 GB of DDR memory
- **1 rack Blue Gene Q**
  - 1024 (1 x 1024) 16 core Compute Nodes
  - PowerPC® A2 core processors (a total of 16384)
  - 16 cores on a chip share a16 GB of memory

New Users

- **Getting a Computer Account on New York Blue**

Step 1: **Request permission to use "New York Blue/L and P" or the "BNL Blue Gene/Q“**

Step 2: **Obtain a BNL Guest Number**

Step 3: **Complete Cyber Security Training**

Step 4: **Apply for a "New York Blue/L and P" and/or a "BNL Blue Gene/Q" Computer Account**

Step 5: **Login to the Blue Gene ssh gateways and the front-end host**

Step 6: **Subscribe to the BNL Blue Gene users mailing lists**

http://www.bnl.gov/newyorkblue/

IBM Redbooks

- **IBM System Blue Gene Solution: Blue Gene/Q Application Development**, SG24-7948-00
  Redbooks, published 8 August 2012.

- **IBM System Blue Gene Solution: Blue Gene/Q System Administration**, SG24-7869-00
  Redbooks, published 13 June 2012.