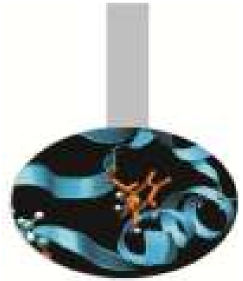
BG/Q Architecture

Carlo Cavazzoni

Mirko Cestari

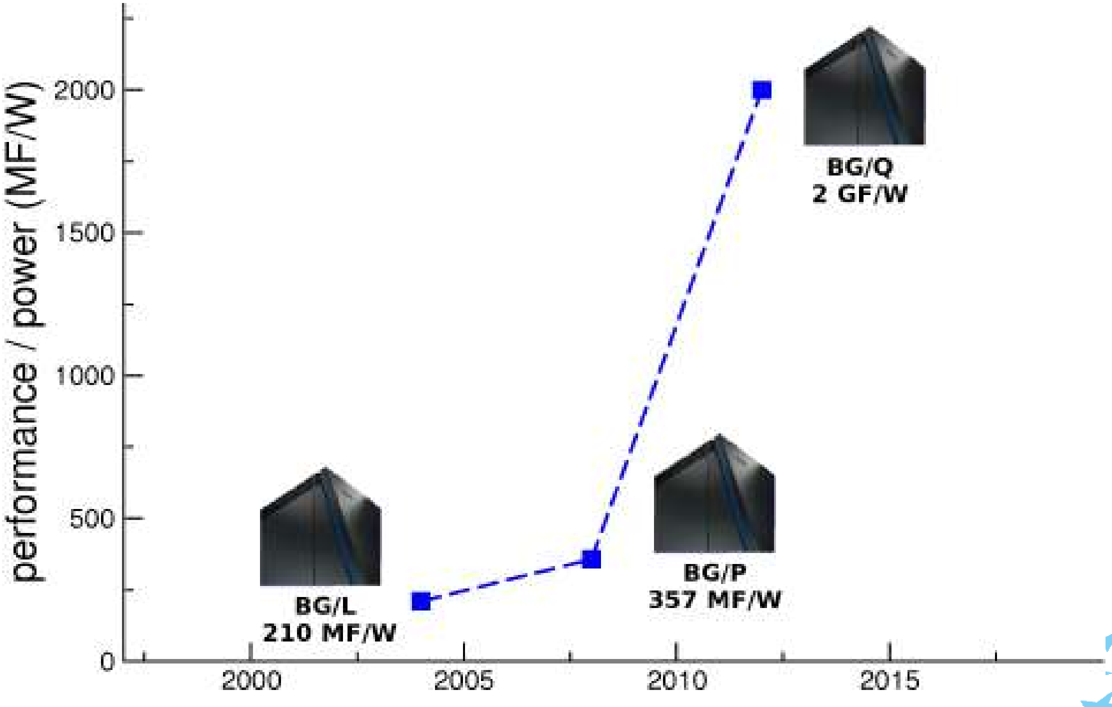
HPC department, CINECA

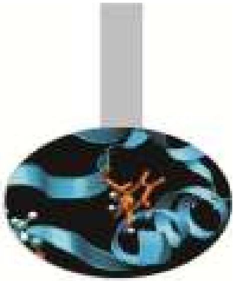
Outline



- ❑ **What is BG**
  - ❑ The Blue Gene family of supercomputers: evolution and challenges
- ❑ **Overview of Blue Gene/Q architecture**
  - ❑ FERMI configuration
  - ❑ Basic concepts: packaging hierarchy, partitions and compute blocks
- ❑ **A closer look**
  - ❑ Inside the processor and the chip
  - ❑ The QUAD FPU

- ❑ BG is a massively parallel supercomputer
- ❑ It holds different types of nodes (and networks)
- ❑ It is designed to have high energy-efficiency (performance/power)





BLUE GENE EVOLUTION

	Total		Biggest Config	Per rack		
	Performance [PF]	Efficiency [MF/W]	Max # of racks	Performance [TF]	Efficiency	# of cores
BG/L	0.596	210	104	5.7	2.02	2048
BG/P	1	357	72	13.9	4.96	4096
BG/Q	20	2000	96	209	20.83	16384

Towards higher and higher:

- Performance
- Efficiency
- Density of cores per rack

Blue Gene/Q

Features:

- among the most powerful architectures
- among the most “green”
- multi-core/multi-threaded computing
- Has an innovative design (system-on-a-chip)

… and objectives:

- Laying groundwork for Exascale computing
- Reduce total cost of ownership

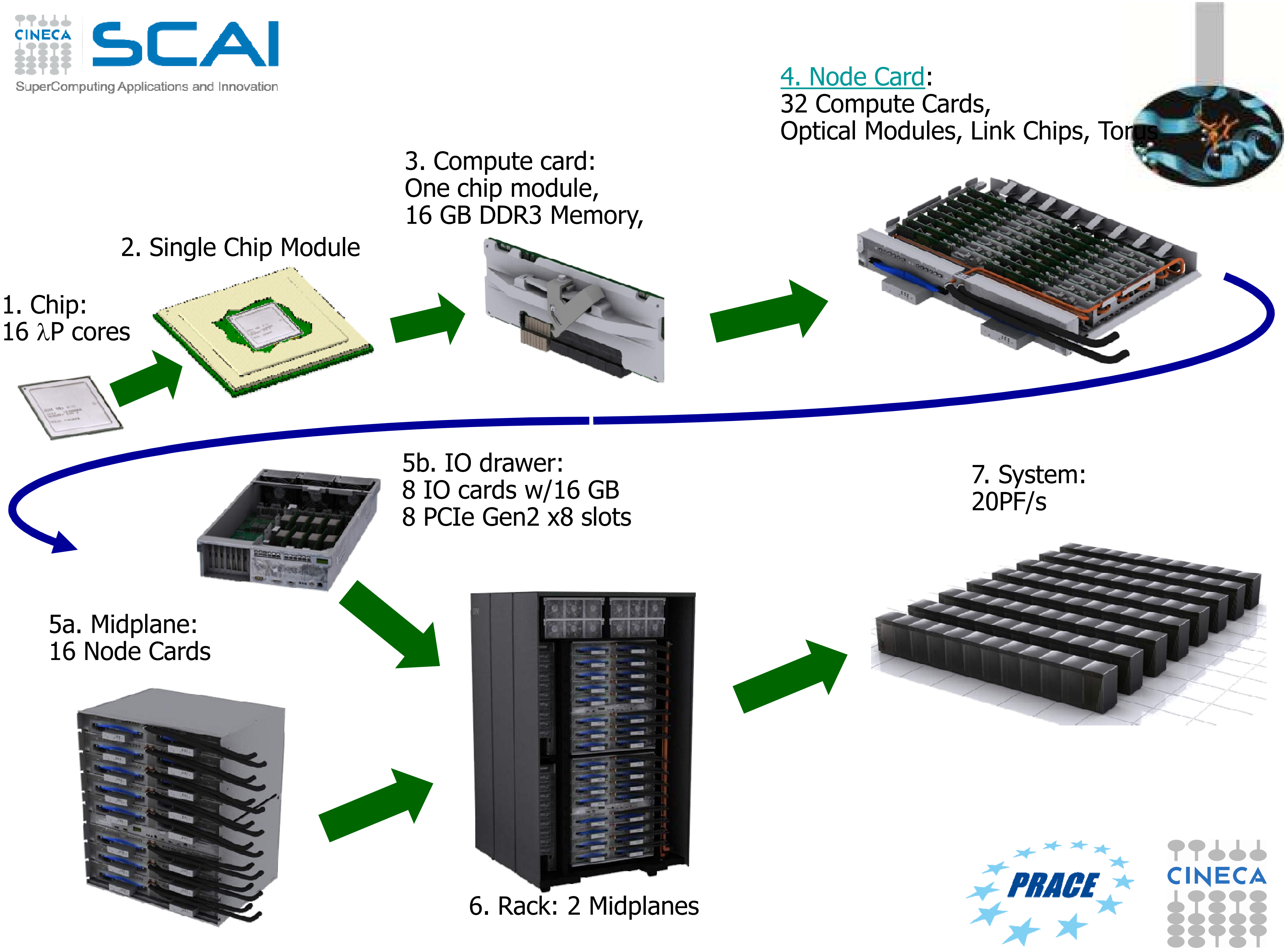
TOP500 November 2013

1	Tianhe-2- TH-IVB-FEP, Xeon E5-2692 2.20 GHz, TH Express-2, Intel Xeon Phi
2	Titan - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x
3	Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom
4	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect
5	Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom
6	Piz Daint – Cray XC30, Xeon E5-2670 8C 2.60GHz, Aries Interconnect, NVIDIA K20x
7	Stampede - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi
8	JUQUEEN - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect
9	Vulcan - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect
10	SuperMUC - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR
16	**Fermi - BlueGene/Q, Power BQC 16C 1.60GHz, Custom Interconnect**

http://www.top500.org

FERMI @ CINECA
PRACE Tier-0 System

Architecture: 10 BGQ Frame
Model: IBM-BG/Q
Processor Type: IBM PowerA2, 1.6 GHz
Computing Cores: 163840
Computing Nodes: 10240
RAM: 1 GByte / core
Internal Network: 5D Torus
Disk Space: 2 PByte of scratch space
Peak Performance: 2 PFlop/s
Power Consumption: 1 MWatt



1. Chip:
16 λP cores

2. Single Chip Module

3. Compute card:
One chip module,
16 GB DDR3 Memory,

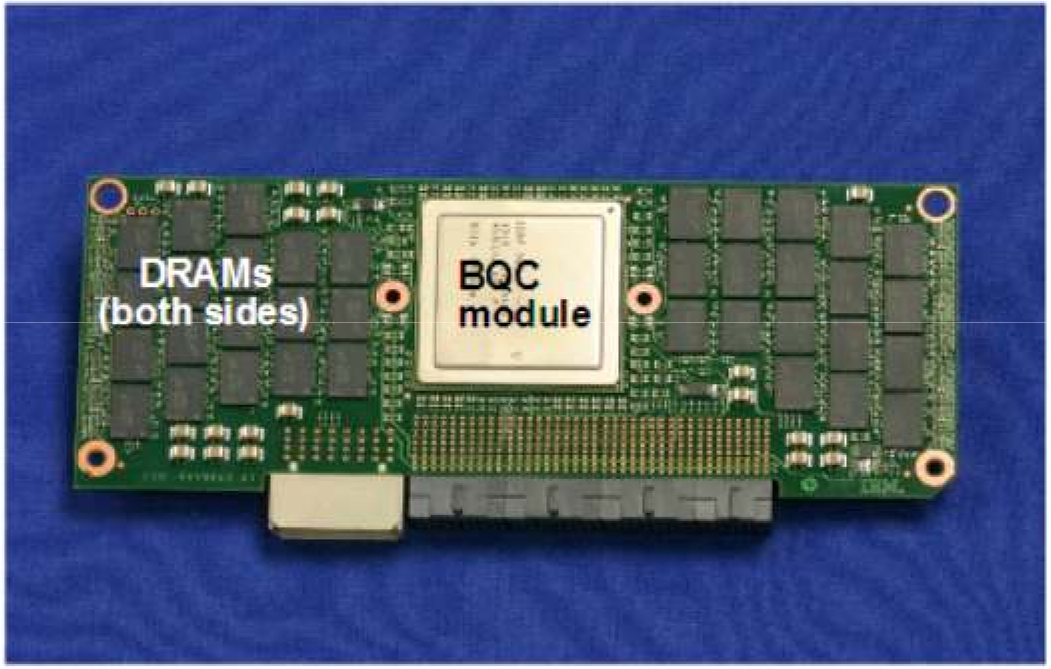
4. Node Card:
32 Compute Cards,
Optical Modules, Link Chips, Torus

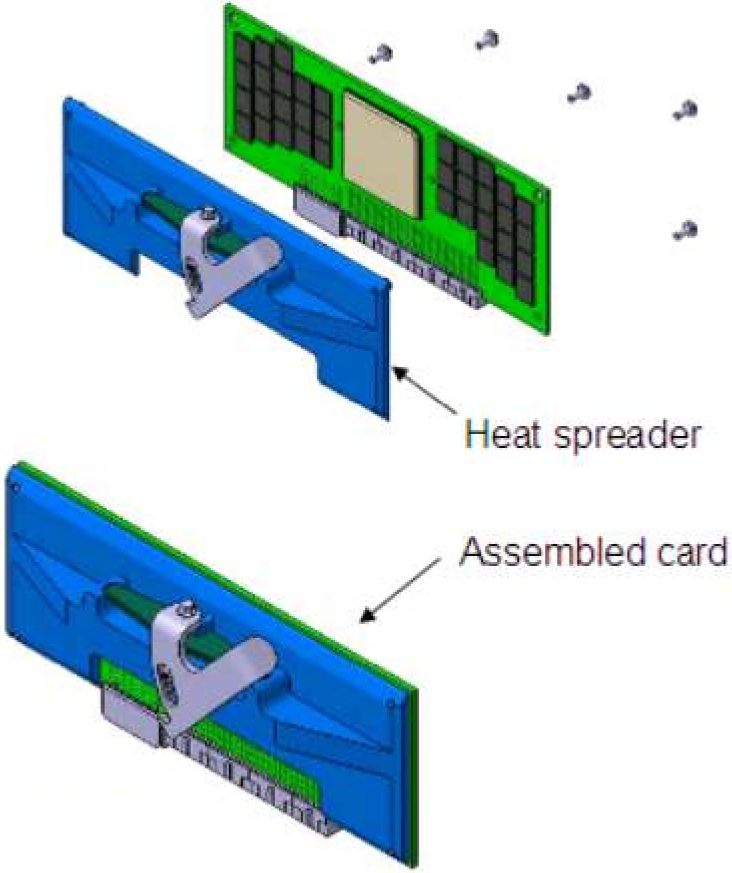
5b. IO drawer:
8 IO cards w/16 GB
8 PCIe Gen2 x8 slots

5a. Midplane:
16 Node Cards

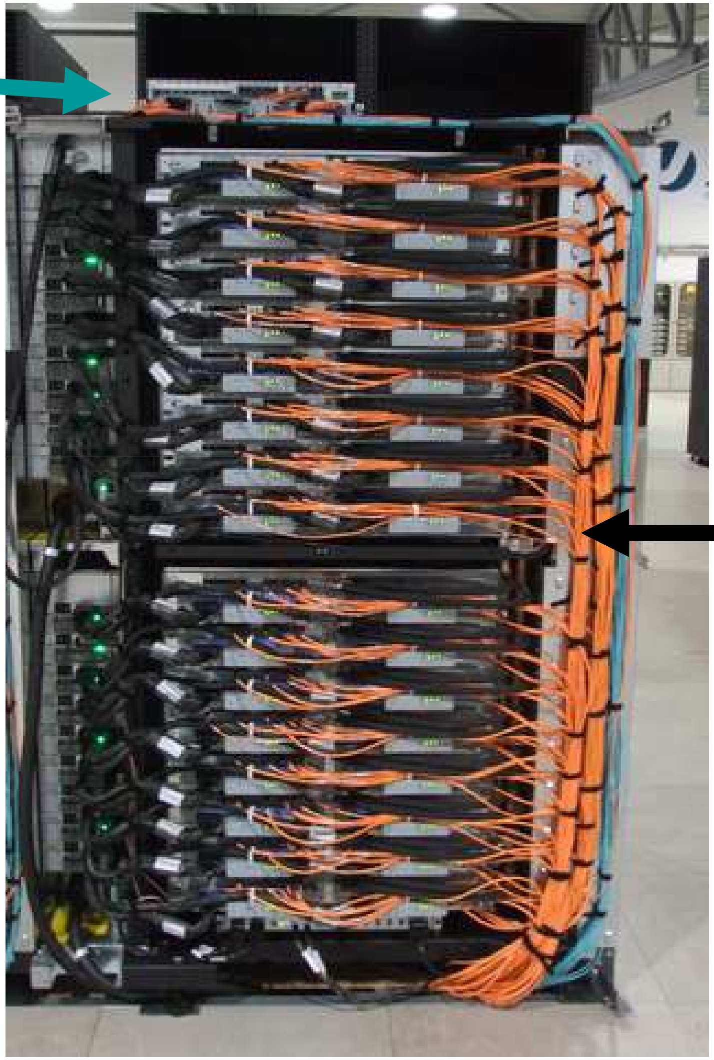
6. Rack: 2 Midplanes

7. System:
20PF/s


DRAMs
(both sides)
BQC
module

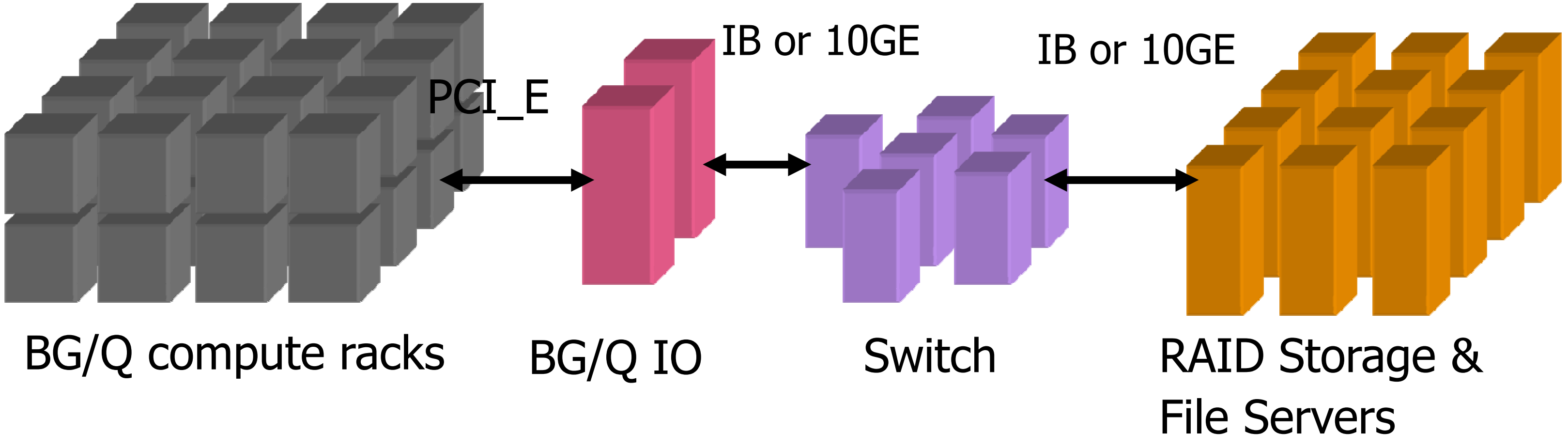

Heat spreader
Assembled card

Point-to-point fiber cables,
attaching the 8 I/O nodes
(on top of rack)
to compute nodes
(on 8 node cards)



4D torus fiber cables,
connecting the
midplane to
other midplanes
(in same and other racks)

BG/Q I/O architecture



External, independent and dynamic I/O system

- I/O nodes in separate drawers/rack with private interconnections and full Linux support
- PCI-Express Gen 2 on every node with full sized PCI slot

•BlueGene Classic I/O with GPFS clients on the logical I/O nodes

•Similar to BG/L and BG/P

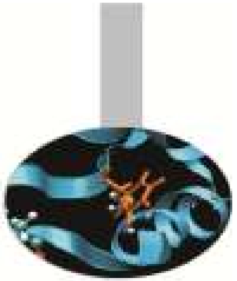
•Uses InfiniBand switch

•Uses DDN RAID controllers and File Servers

•BG/Q I/O Nodes are not shared between compute partitions

- **IO Nodes are bridge data from function-shipped I/O calls to parallel file system client**

•Components balanced to allow a specified minimum compute partition size to saturate entire storage array I/O bandwidth



I/O nodes – node cards ratio

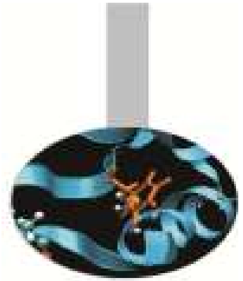
Blue Gene/Q has a Flexible I/O nodes – node cards ratio

8 I/O nodes per I/O drawer
4 I/O drawers per rack (maximum)

→

up to 32 I/O nodes per rack
= **1 I/O node per 512 compute cores**

FERMI configuration:
2 racks with 16 I/O nodes (1024 cores per I/O node)
8 racks with 8 I/O nodes (2048 cores per I/O node)



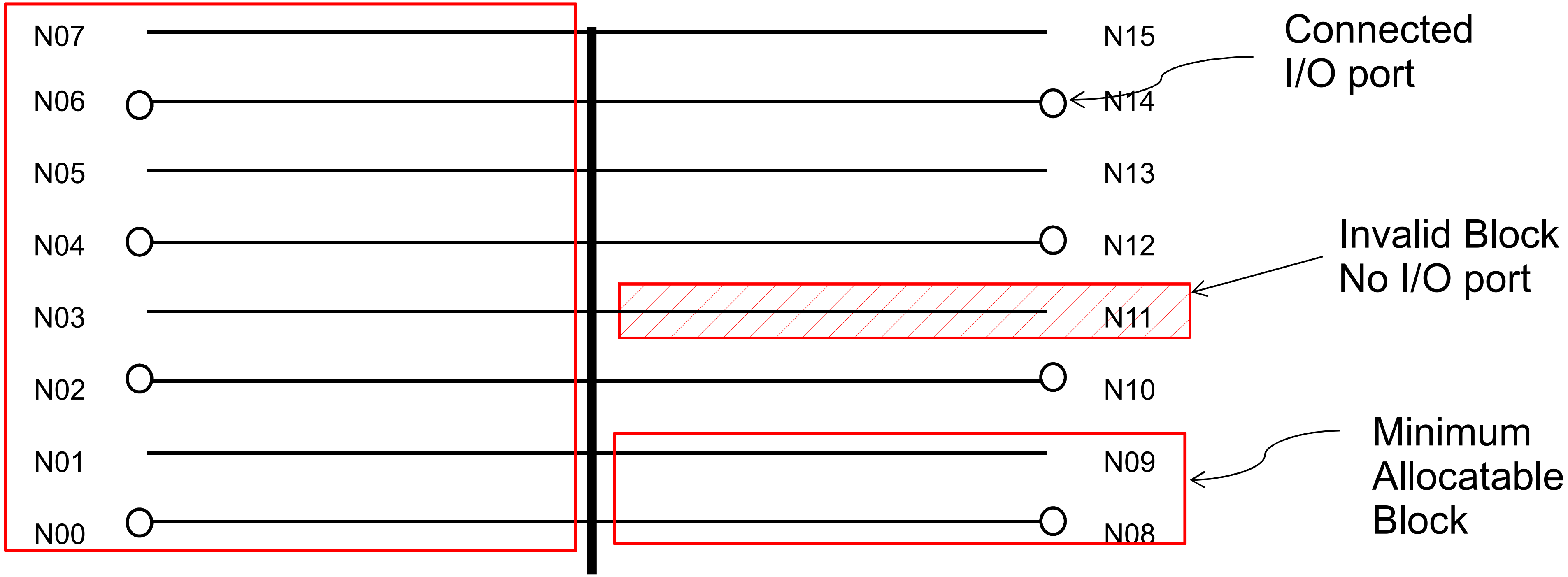
Ok, but… why should I care?

The number of I/O nodes per rack constraints:

-I/O bandwidth to/from compute racks
(each I/O node has 2 links (4GB/s in 4GB/s out))

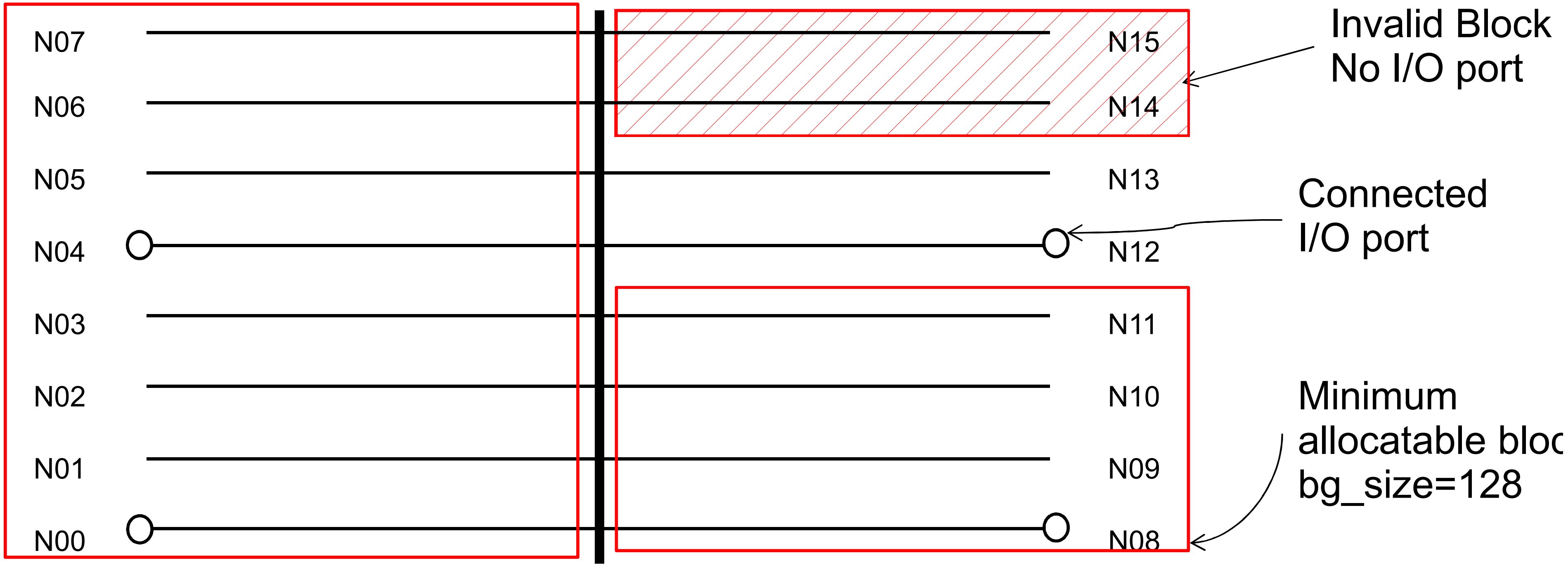
-The minimum partition allocatable on a BG/Q system ("small block" jobs)
For FERMI:
bg_size=64 (jobs running on R11 and R31)
bg_size=128 (jobs running on the other racks)

MidPlane in FERMI RACK: R11, R31

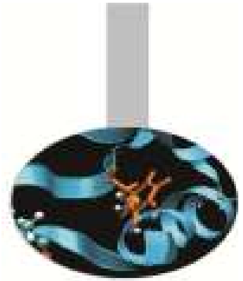


Example:
N08 – N09 = 64 Compute Cards (2x2x4x2x2)

MidPlane in FERMI / {R11 R31}



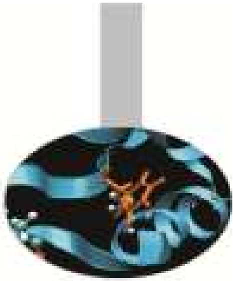
Example:
N08 – N09 – N10 – N11 = 128 Compute Cards (2x2x4x4x2)



Compute blocks on Fermi

- **Small blocks:**
- contains one or more node boards within a single midplane
- always multiple of 32 nodes

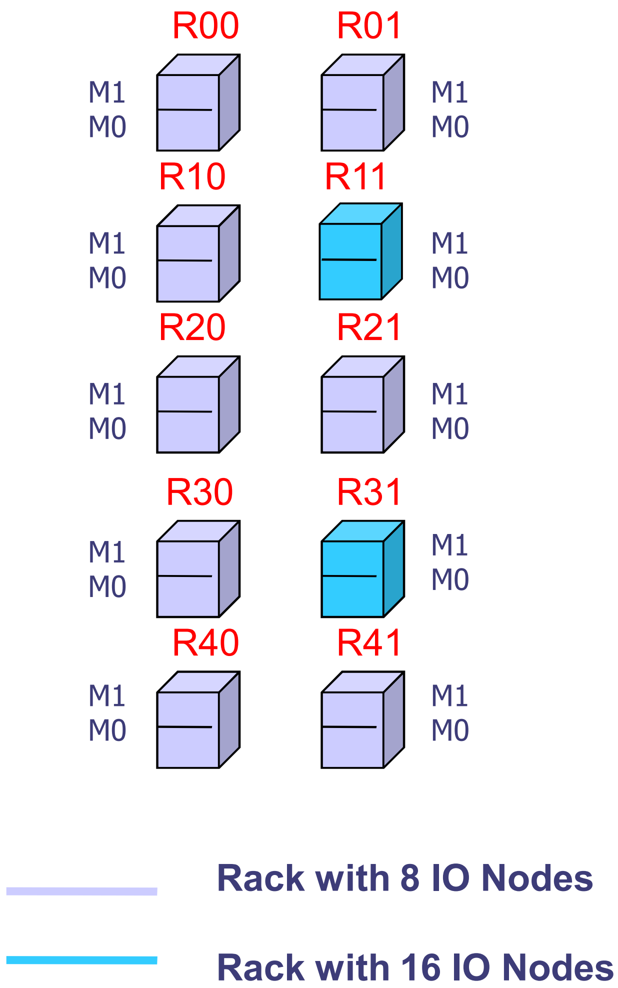
- **Large blocks:**
- contains one or more complete midplanes
- always multiple of 512 nodes



•New Network architecture:

- 5 D torus architecture sharing several embedded Virtual Network/topologies
  - 5D topology for point-to-point communication
    - 2 GB/s bidirectional bandwidth on all (10+1) links
    - Bisection bandwidth of 65TB/s (26PF/s) / 49 TB/s (20 PF/s) BGL at LLNL is 0.7 TB/s
  - Collective and barrier networks embedded in 5-D torus network.
- Floating point addition support in collective network
- 11th port for auto-routing to IO fabric

FERMI Configuration



Rack with 8 IO Nodes

Rack with 16 IO Nodes

10 racks

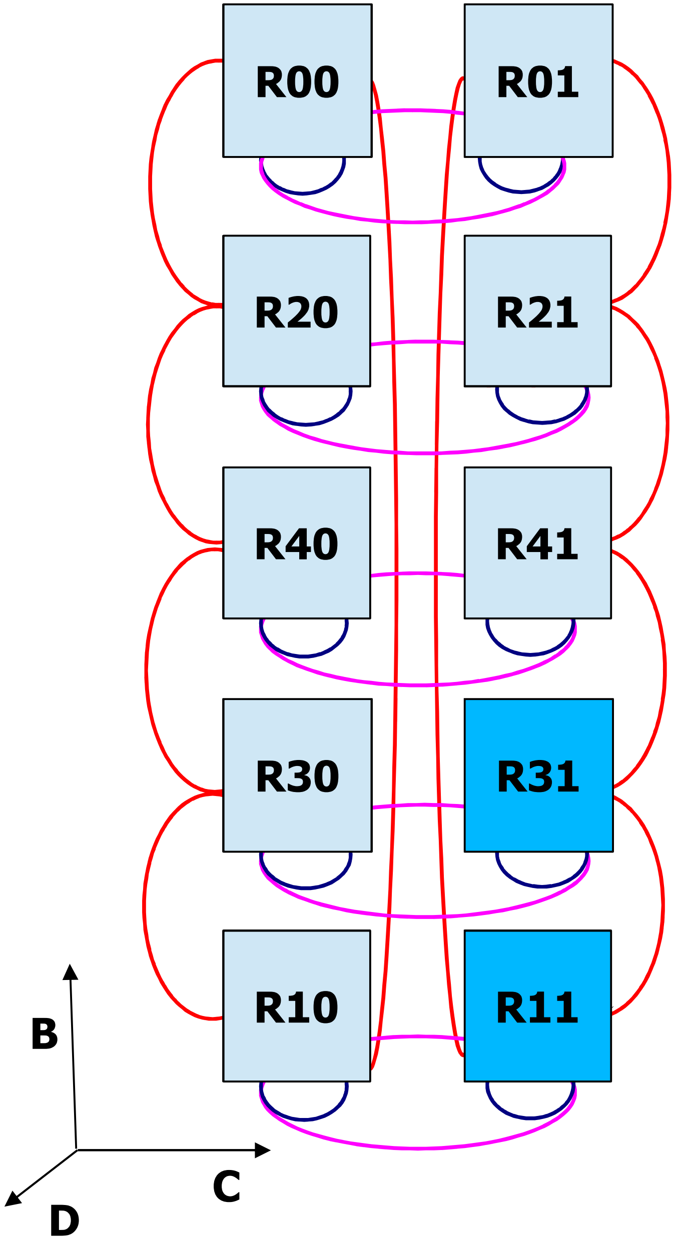
- 5 rows
- 2 columns

20 midplanes

- 2 midplanes for each rack

Racks	MP	Row	Col	A	B	C	D
10	20	5	2	1	5	2	2

Midplanes CABLING



B dimension

- connection among 2 midplanes goes down a column of racks
- on Fermi the number of the cables on the B dim is **5**

C dimension

- connection among 2 midplanes goes down a row of racks
- on Fermi the number of the cables on the C dim is **2**

D dimension

- connection among 2 midplanes in the same rack
- on Fermi the number of the cables on the D dim is **2**

A dimension

- the remaining direction, which can go down a row or column (or both). When two sets of cables go down a row or column, the longest cables define the A dimension
- on Fermi the number of the cables along the A dim is **1** and it is not rapresented

Racks	MP	Row	Col	A	B	C	D
10	20	5	2	1	5	2	2

FERMI Size in MidPlanes

SHAPE of FERMI =

number of midplanes in A, B, C, D directions

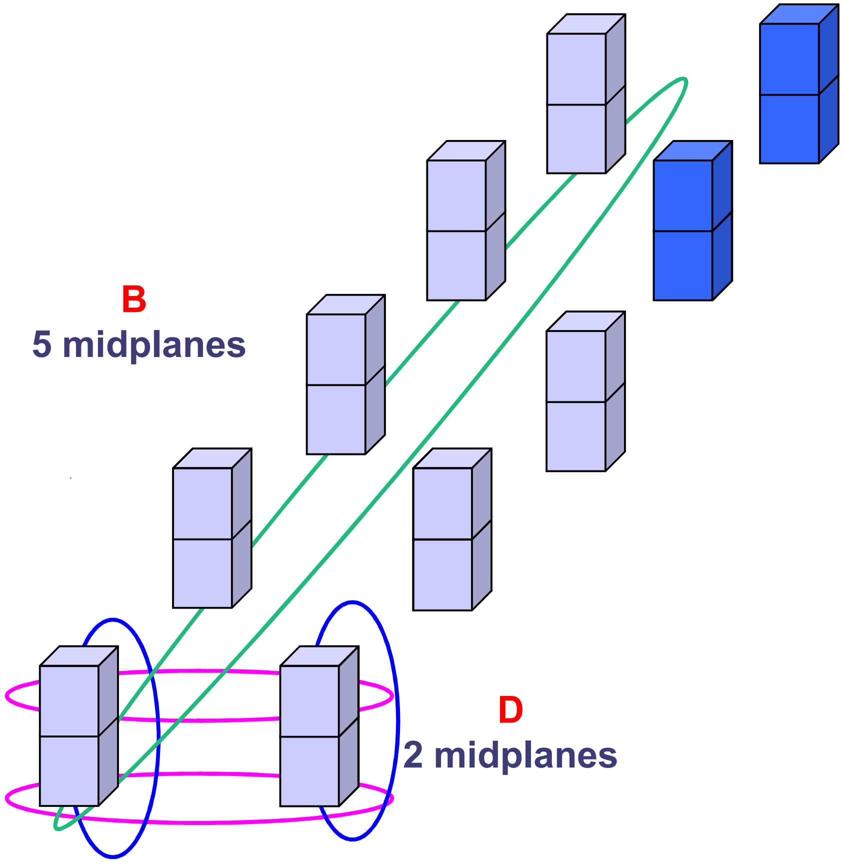
1 x 5 x 2 x 2 = 20 MidPlanes

Midplanes CONNECTIVITY

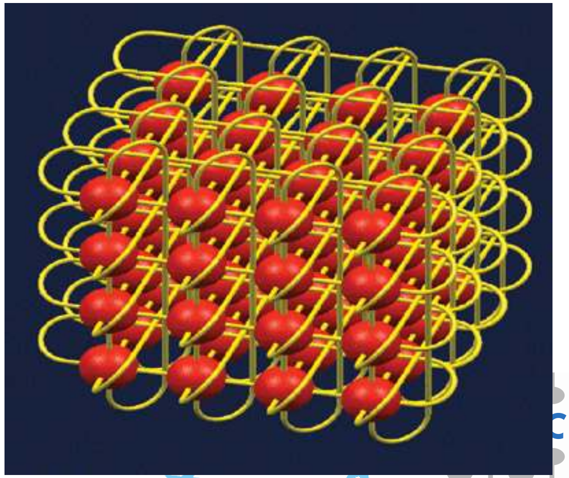
For **large block jobs** (>= 1MP) two connectivity between midplanes are provided:

- **Torus : periodic boundary conditions (e.g. "close line") in all the dimensions A, B, C and D.**
- **Mesh : almost one dimension is not like a "close line"**

3D TORUS

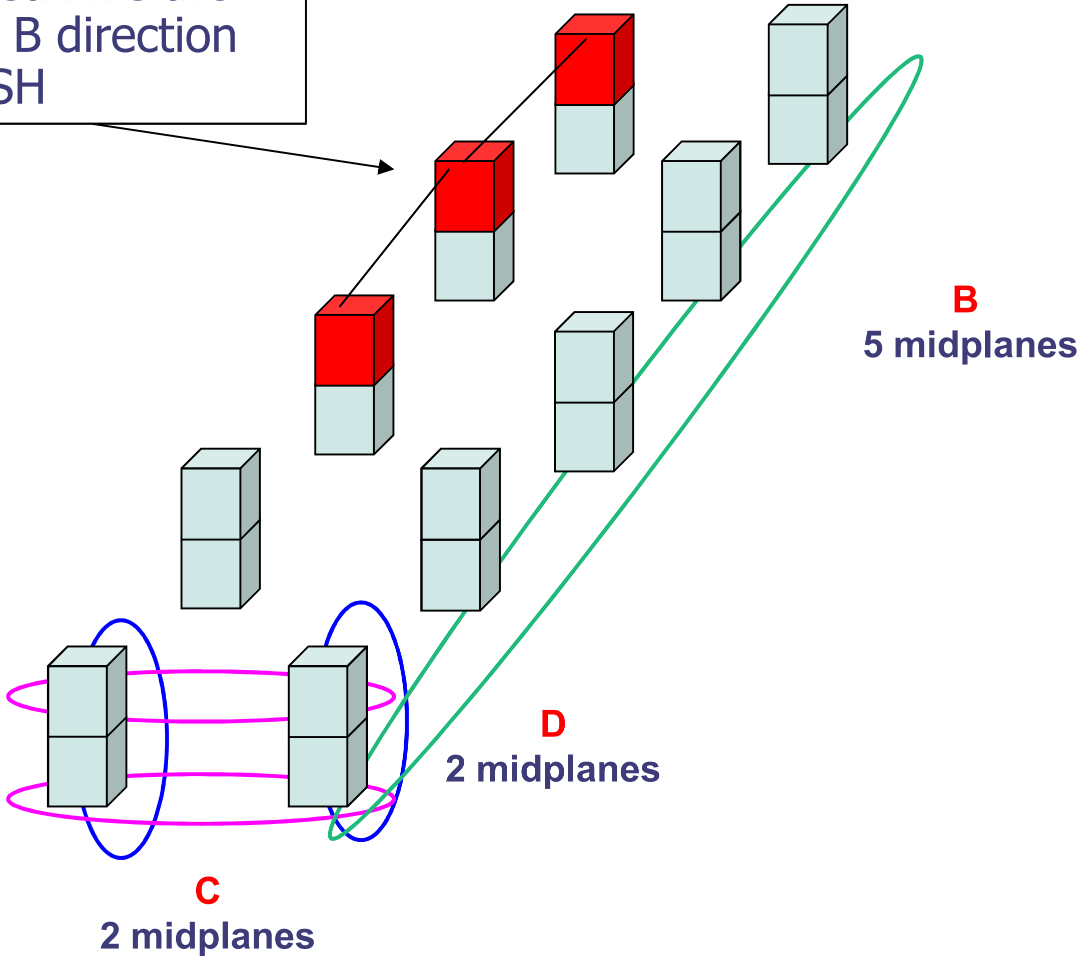


- 1 Midplane is the minimum TORUS available on a BlueGene/Q system



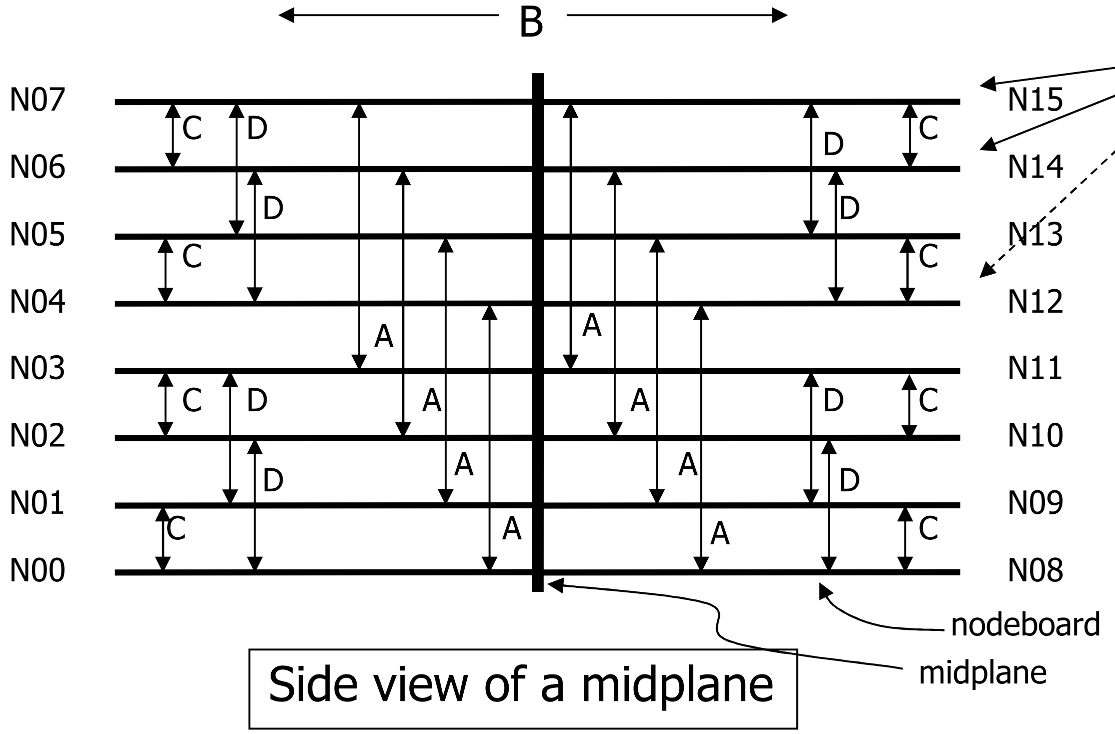
TORUS vs. MESH

- The 3 red MPs are linked in B direction as a MESH



5-D torus wiring in a Midplane

The 5 dimensions are denoted by the letters A, B, C, D, and E. The latest dimension E is always 2, and is contained entirely within a midplane.



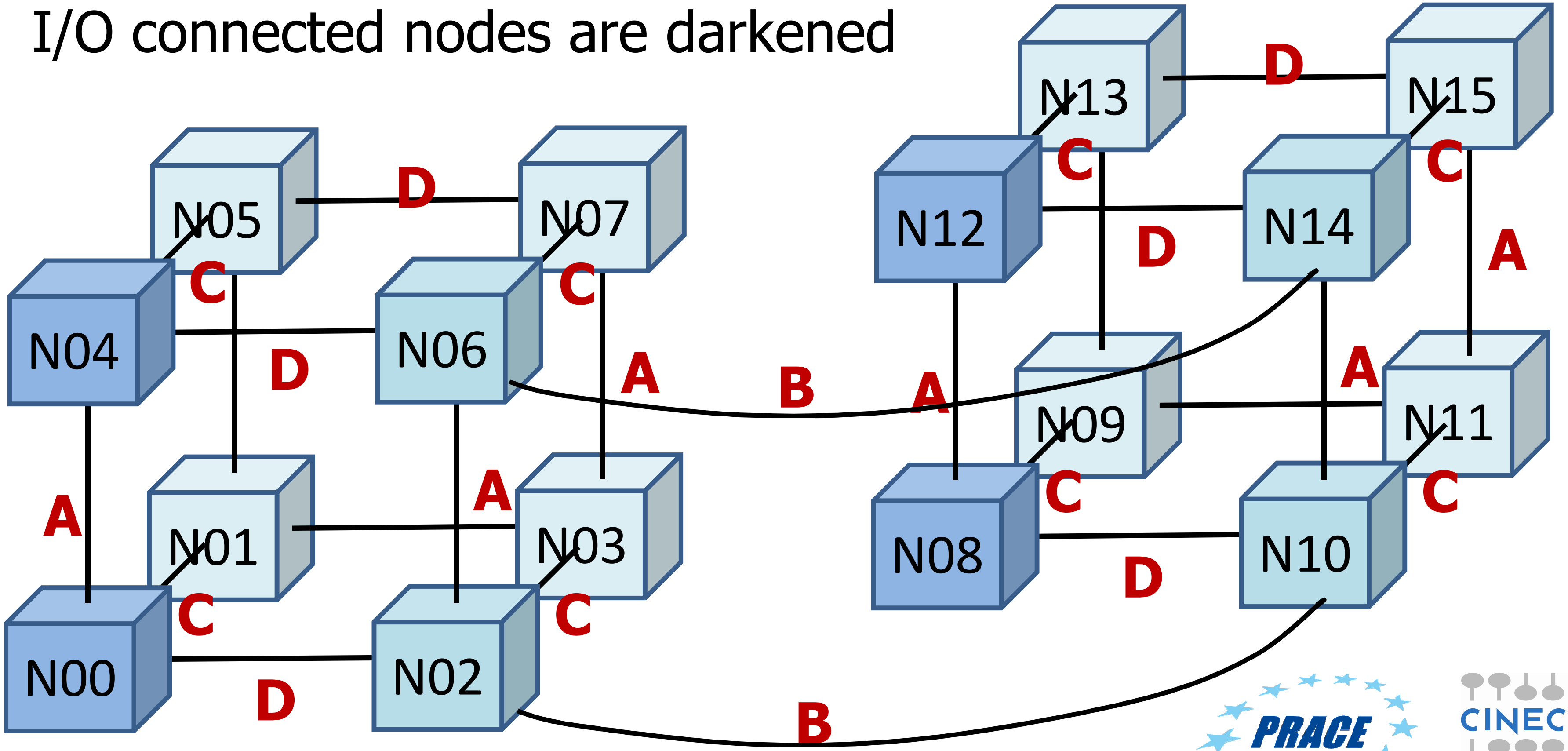
Side view of a midplane

- **Each nodeboard is 2x2x2x2x2**
- **Arrows show how dimensions A,B,C,D span across nodeboards**
- **Dimension E does not extend across nodeboards**

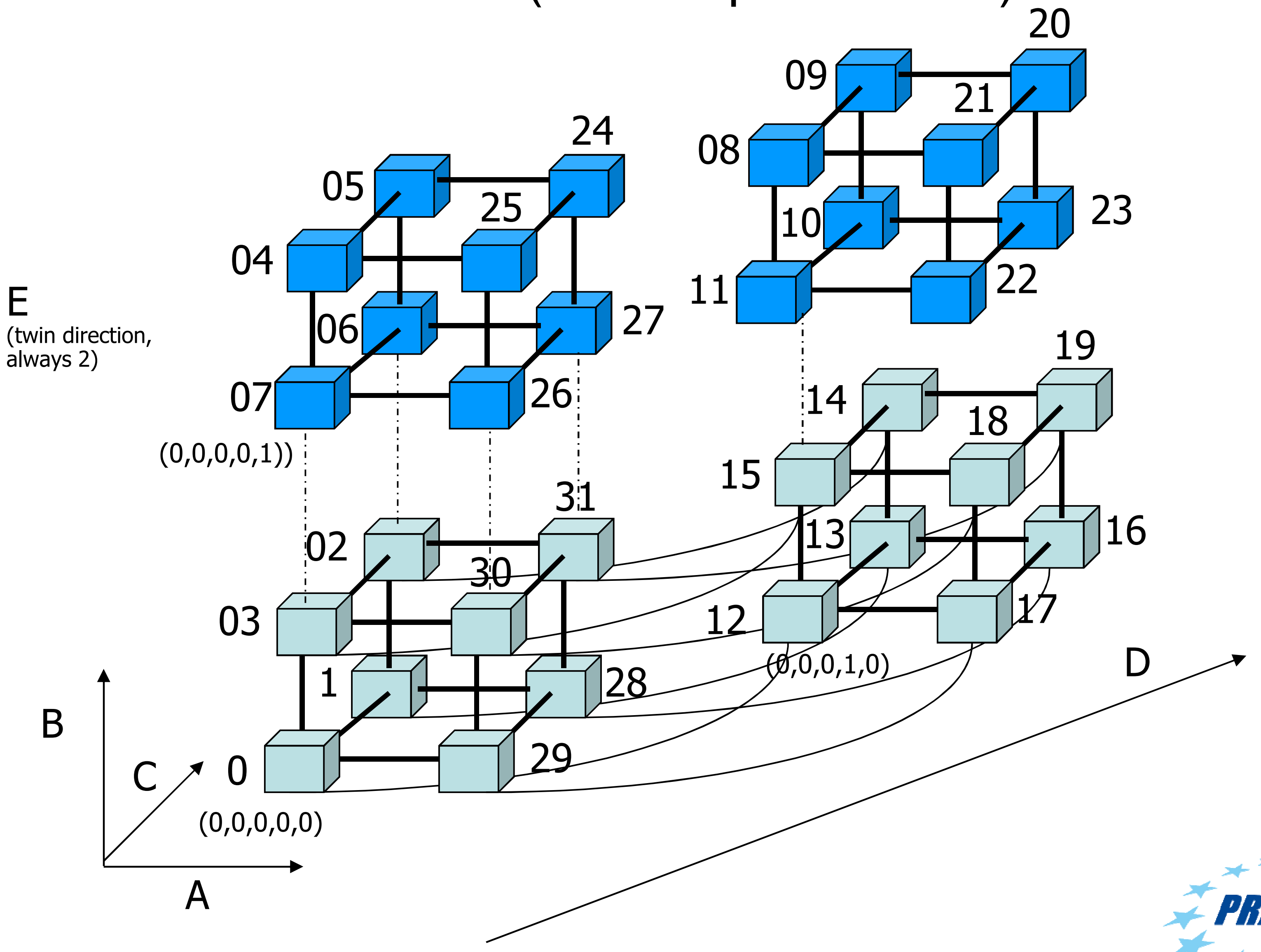
- **The nodeboards combine to form a 4x4x4x4x2 torus**
- **Note that nodeboards are paired in dimensions A,B,C and D as indicated by the arrows**

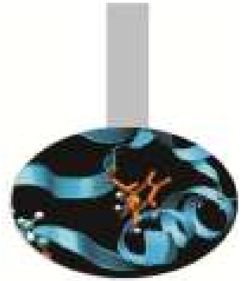
5-D torus in a Midplane

I/O connected nodes are darkened



Node Board (32 Compute Nodes): 2x2x2x2x2

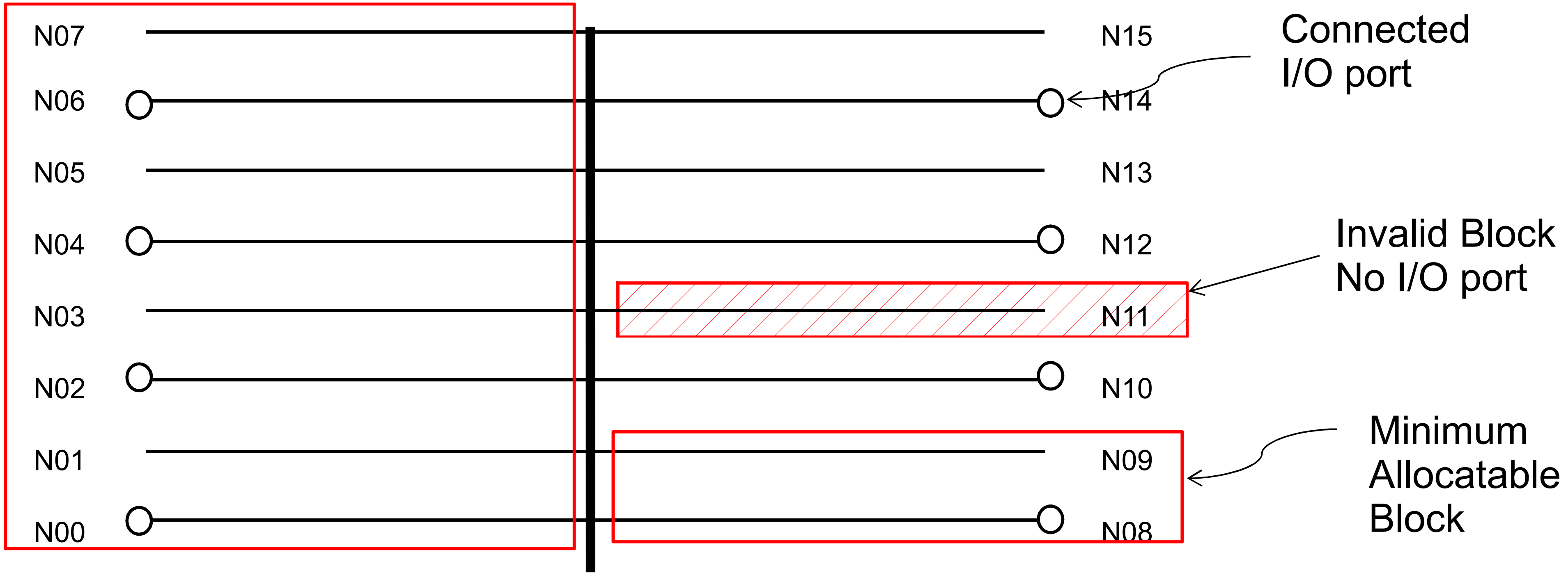




Network topology | Mesh versus torus

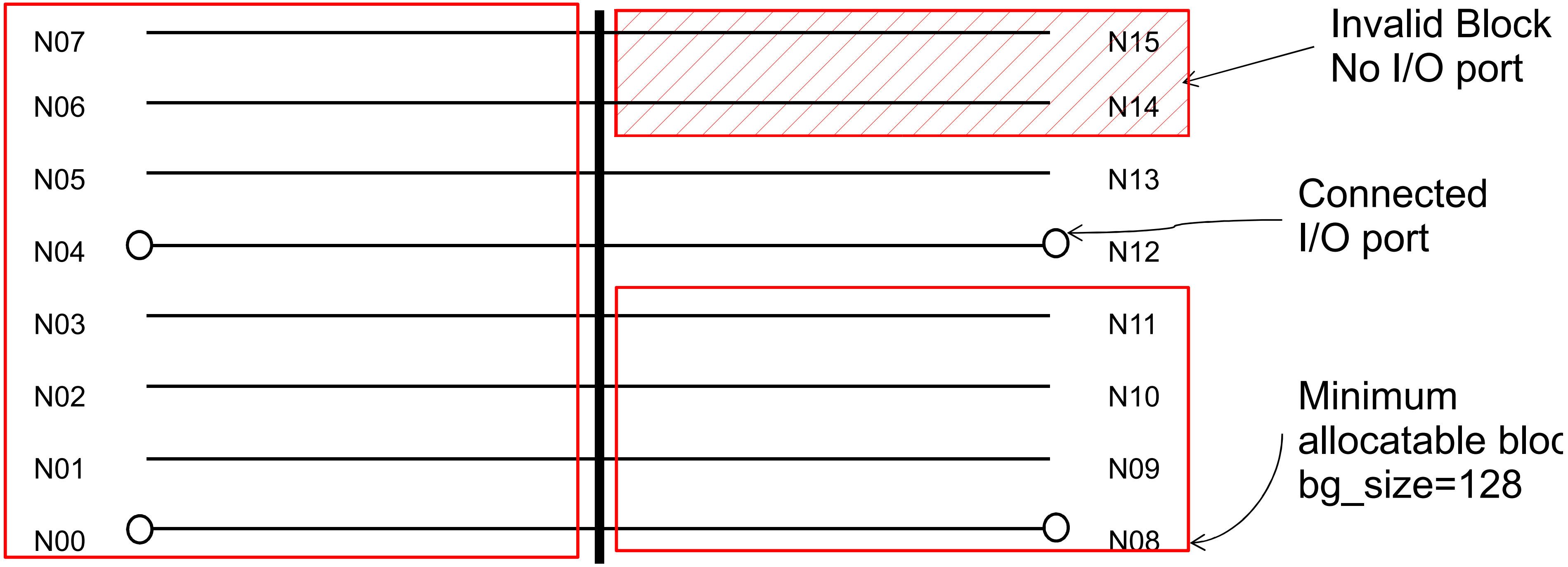
# Node Boards	# Nodes	Dimensions	Torus (ABCDE)
1	32	2x2x2x2x2	00001
2 (adjacent pairs)	64	2x2x4x2x2	00101
4 (quadrants)	128	2x2x4x4x2	00111
8 (halves)	256	4x2x4x4x2	10111

MidPlane in FERMI RACK: R11 R31



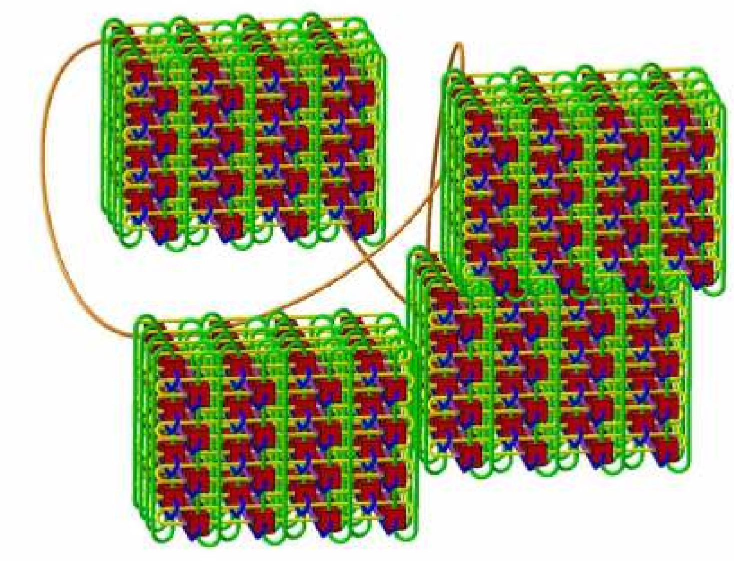
Example:
N08 – N09 = 64 Compute Cards (2x2x4x2x2)

MidPlane in FERMI / {R11 R31}



Example:
N08 – N09 – N10 – N11 = 128 Compute Cards (2x2x4x4x2)

Inter-Processor Communication



- **Integrated 5D torus**
  - –Virtual Cut-Through routing
  - –Hardware assists for collective & barrier functions
  - –FP addition support in network
  - –RDMA
    - Integrated on-chip Message Unit
- **2 GB/s raw bandwidth on all 10 links**
  - –each direction -- i.e. 4 GB/s bidi
  - –1.8 GB/s user bandwidth
    - protocol overhead
- **5D nearest neighbor exchange measured at 1.76 GB/s per link (98% efficiency)**
- **Hardware latency**
  - –Nearest: 80ns
  - –Farthest: 3us

    (96-rack 20PF system, 31 hops)
- **Additional 11th link for communication to IO nodes**
  - –BQC chips in separate enclosure
  - –IO nodes run Linux, mount file system
  - –IO nodes drive PCIe Gen2 x8 (4+4 GB/s)

    ↔ IB/10G Ethernet ↔ file system & world

Network Performance

- **All-to-all: 97% of peak**
- **Bisection: > 93% of peak**
- **Nearest-neighbor: 98% of peak**
- **Collective: FP reductions at 94.6% of peak**

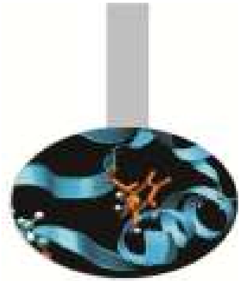
© 2011 IBM Corporation

BGQ PowerA2 processor

Carlo Cavazzoni, Mirko Cestari
HPC department, CINECA

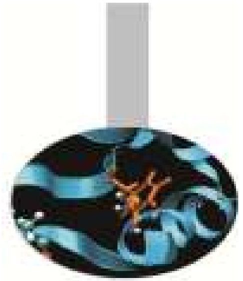
Power A2

- 64bit (was 32 bit for BG/L and BG/P)
- Power instruction set (Power1…Power7, PowerPC)
- RISC processors
- Superscalar
- Multiple Floating Point units
- SMT
- Multicore



PowerA2 chip, basic info

- 16 cores + 1 + 1 (17th Processor core for system functions)
- 1.6GHz
- system-on-a-chip design
- 16GByte of RAM at 1.33GHz
- 32MByte L2 cache, 64B L1 line cache
- Peak Perf 204.8 gigaflops
- power draw of 55 watts
- 45 nanometer copper/SOI process (same as Power7)
- Water Cooled

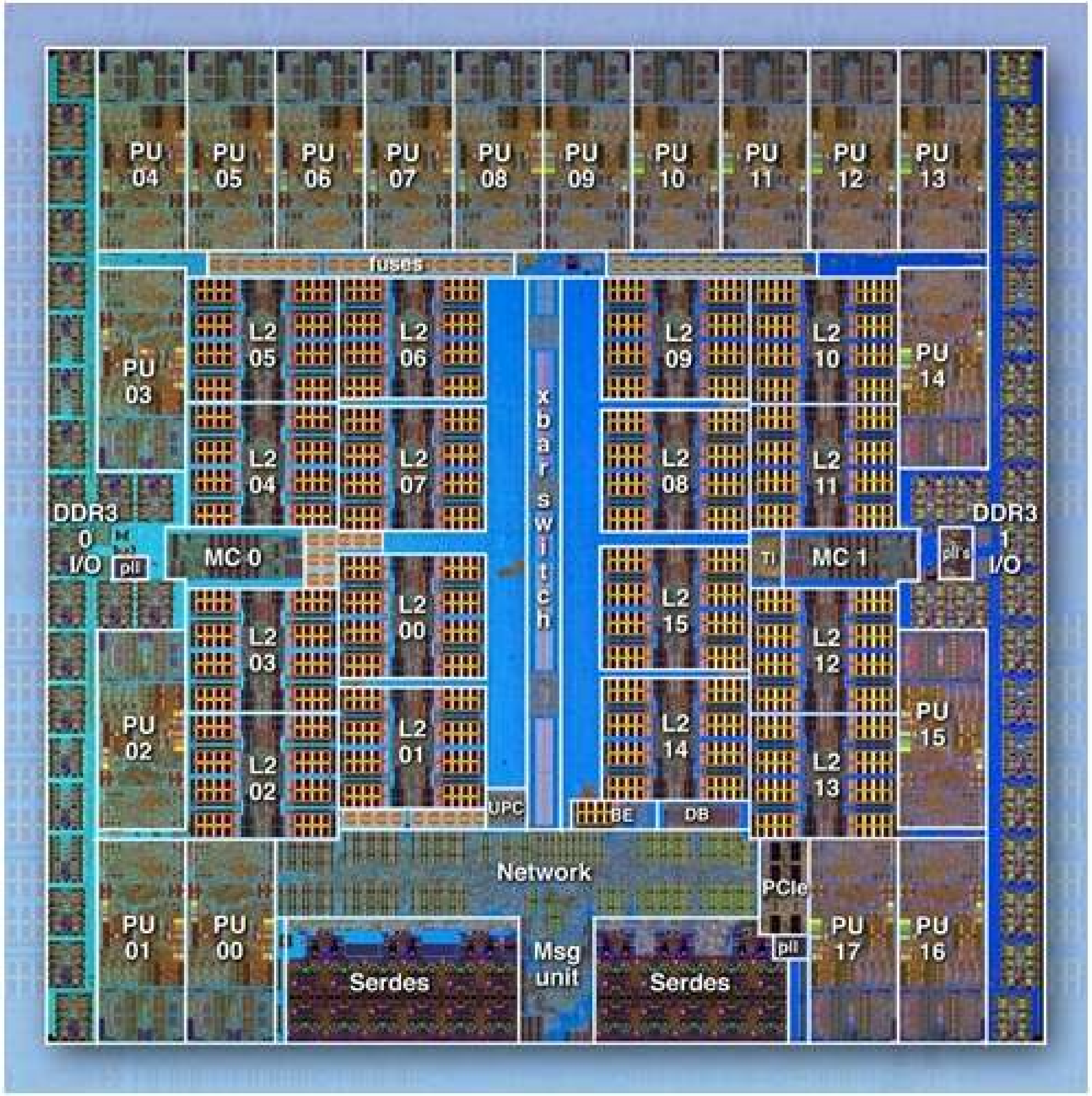


PowerA2 chip, more info

- Contains a 800MHz crossbar switch
  - links the cores and L2 cache memory together
  - peak bisection bandwidth of 563GB/sec
  - connects the processors, the L2, the networking
- 5D torus interconnect is also embedded on the chips
- Two of these can be used for PCI-Express 2.0 x8 peripheral slots.
- supports point-to-point, collective, and barrier messages and also implements direct memory access between nodes.

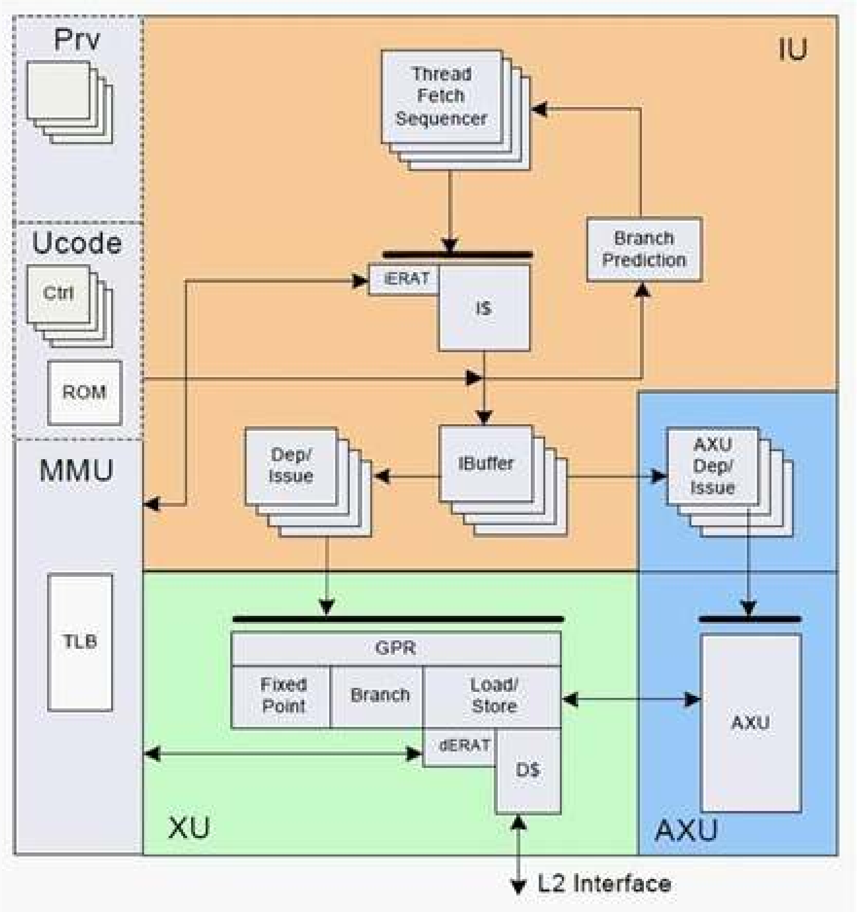
PowerA2 chip, layout

System-on-a-Chip design: integrates processors, memory and networking logic into a single chip



PowerA2 core

- •4 FPU
- •in-order dispatch, execu
- and completion
- •2-way concurrent issue.
- 1 branch/integer/load/
- 1 AXU (FP/vector).
- **•4 way SMT**

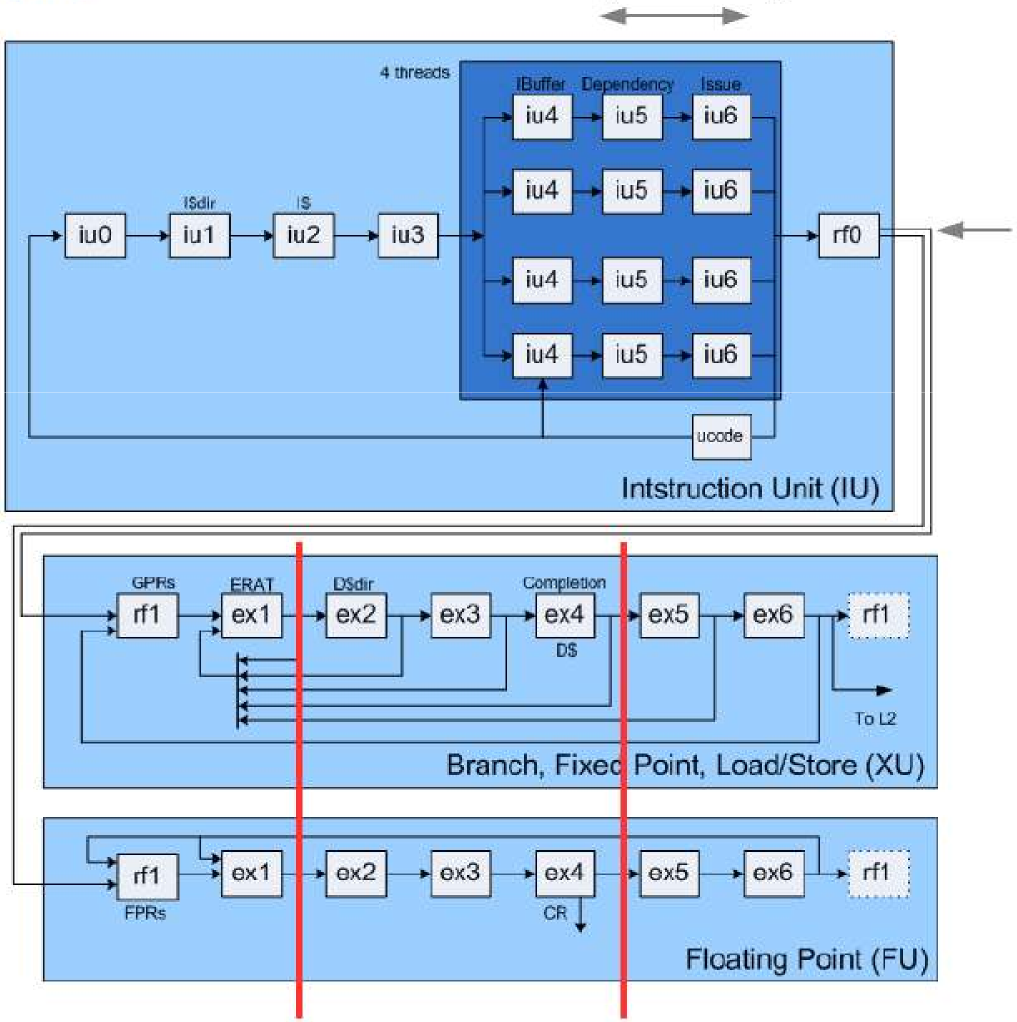


A2 Core | Pipeline

Primary Stall Stages

Front End

Back End



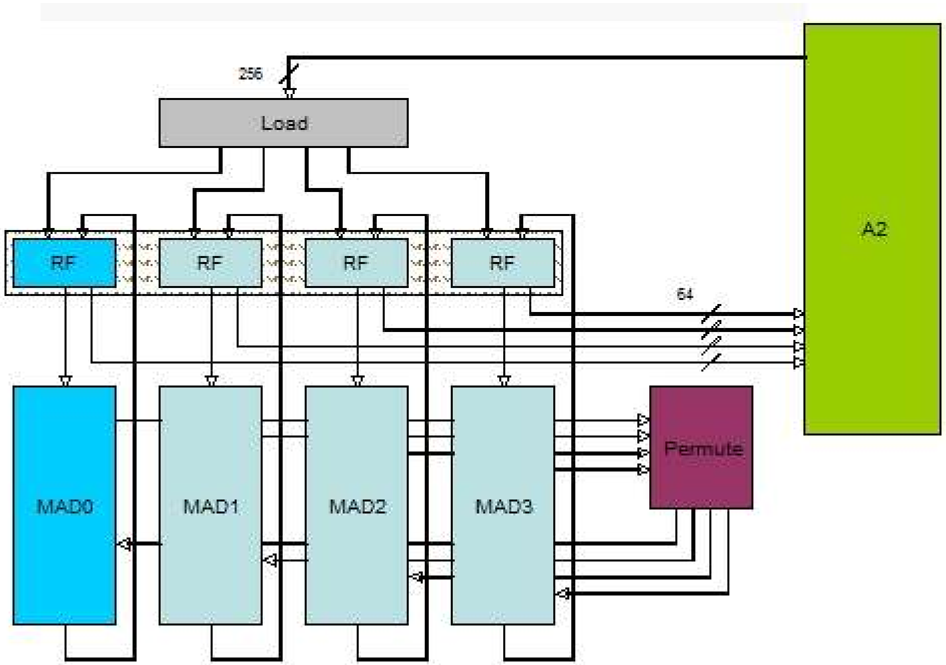
Instruction Issue

SMT, why?

- is a direct consequence of the in-order instruction core
- avoids instruction stall
- increases instruction throughput (not the peak performance)…
- … still can increase the overall performance of a (memory bound) application
- enables superscalar pipeline

PowerA2 FPU

- each processor has Quad FPU
- four-wide double precision SIMD instructions
- (or) two-wide complex arithmetic SIMD inst.
- six-stage pipeline
- permute instructions
- maximum of eight concurrent
- floating point operations
- per clock plus a load and a store.



Thanks for your attention!
Any question?