

## BG/Q Architecture

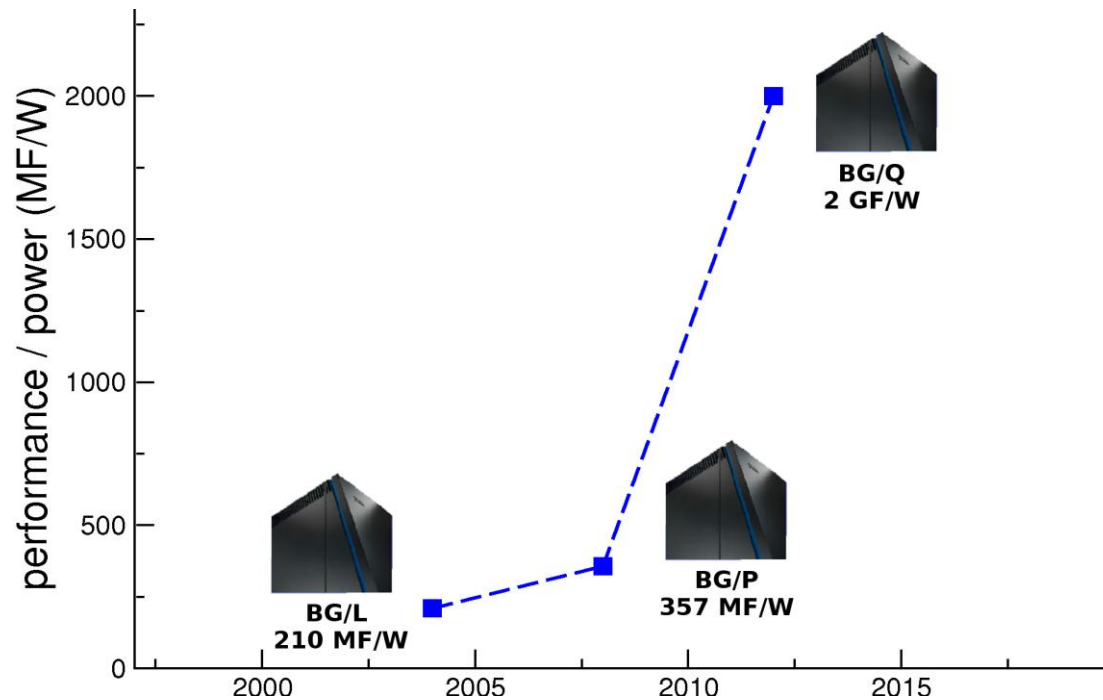
Carlo Cavazzoni  
Graziella Ferini  
HPC department, CINECA



# 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
  - ❑ The “shape”: meaning and consequences
  
- ❑ **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
<b>BG/L</b>	0.596	210	104	5.7	2.02	2048
<b>BG/P</b>	1	357	72	13.9	4.96	4096
<b>BG/Q</b>	20	2000	96	209	20.83	16384

Towards higher and higher:

- Performance
- Efficiency
- Density of cores per rack

## Blue Gene/Q

### Features:

- Is among the most powerful architectures
- Is among the most “green”
- Incorporates innovations that enhance 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



# PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE

## TOP10 November 2012

- 1 Titan - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x
- 2 Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom
- 3 K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect
- 4 Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom
- 5 JUQUEEN - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect
- 6 SuperMUC - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR
- 7 Stampede - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi
- 8 Tianhe-1A - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050
- 9 **Fermi - BlueGene/Q, Power BQC 16C 1.60GHz, Custom**
- 10 DARPA Trial Subset - Power 775, POWER7 8C 3.836GHz, Custom Interconnect

<http://www.top500.org>



## The Green500 List

Listed below are the November 2012 The Green500's energy-efficient supercomputers ranked from 1 to 10.

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	2,499.44	National Institute for Computational Sciences/University of Tennessee	Beacon - Appro GreenBlade GB824M, Xeon E5-2670 8C 2.600GHz, Infiniband FDR, Intel Xeon Phi 5110P	44.89
2	2,351.10	King Abdulaziz City for Science and Technology	SANAM - Adtech ESC4000/FDR G2, Xeon E5-2650 8C 2.000GHz, Infiniband FDR, AMD FirePro S10000	179.15
3	2,142.77	DOE/SC/Oak Ridge National Laboratory	Titan - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x	8,209.00
4	2,121.71	Swiss Scientific Computing Center (CSCS)	Todi - Cray XK7 , Opteron 6272 16C 2.100GHz, Cray Gemini interconnect, NVIDIA Tesla K20 Kepler	129.00
5	2,102.12	Forschungszentrum Juelich (FZJ)	JUQUEEN - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect	1,970.00
6	2,101.39	Southern Ontario Smart Computing Innovation Consortium/University of Toronto	BGQdev - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect	41.09
7	2,101.39	DOE/NNSA/LLNL	rzuseq - BlueGene/Q, Power BQC 16C 1.60GHz, Custom	41.09
8	2,101.39	IBM Thomas J. Watson Research Center	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	41.09
9	2,101.12	IBM Thomas J. Watson Research Center	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	82.19
10	2,101.12	Ecole Polytechnique Federale de Lausanne	CADMOS BG/Q - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect	82.19



# FERMI @ CINECA

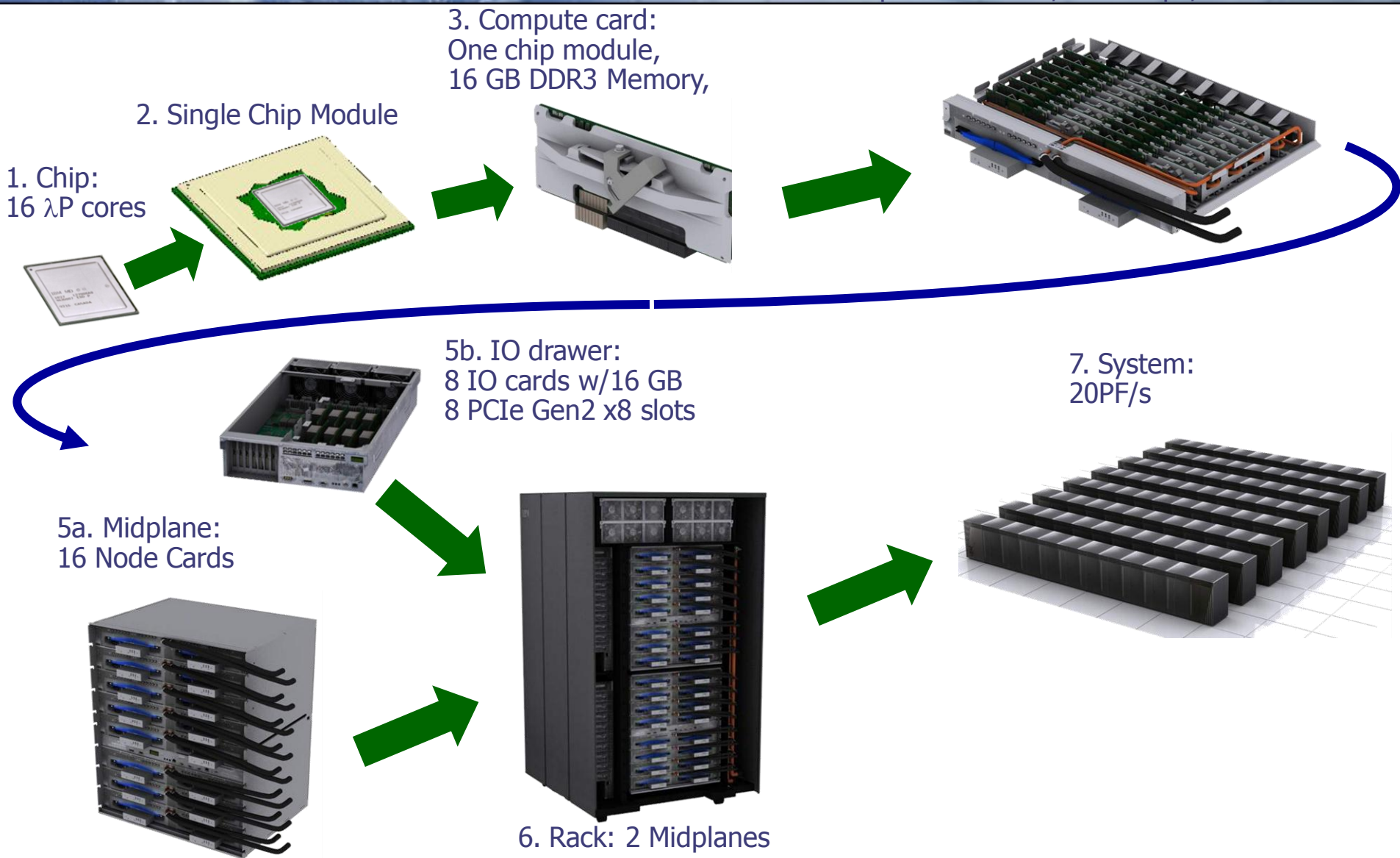
## PRACE Tier-0 System

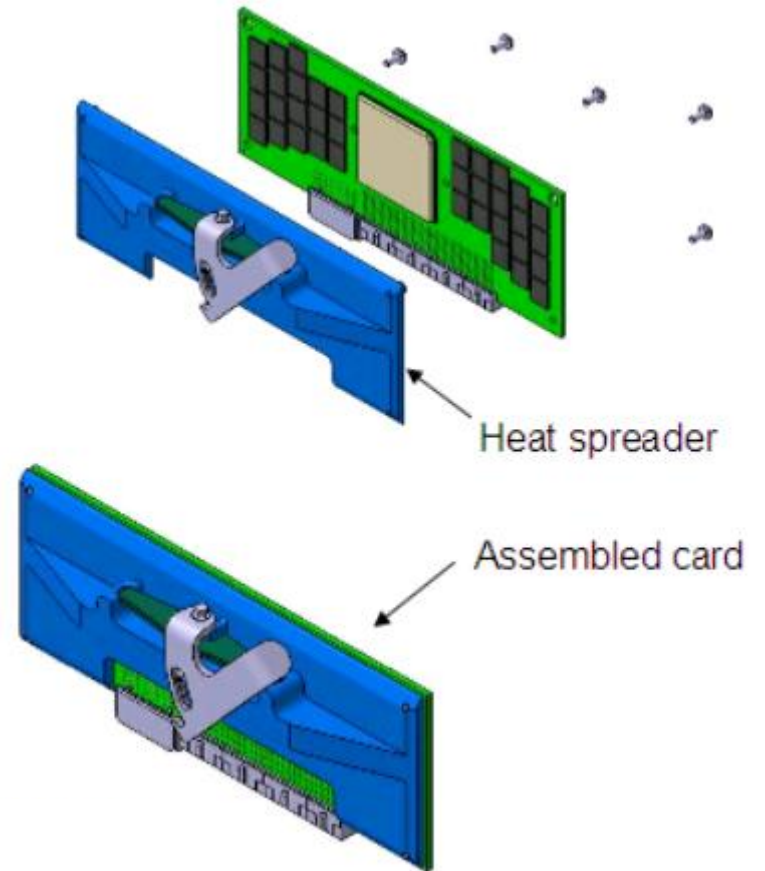
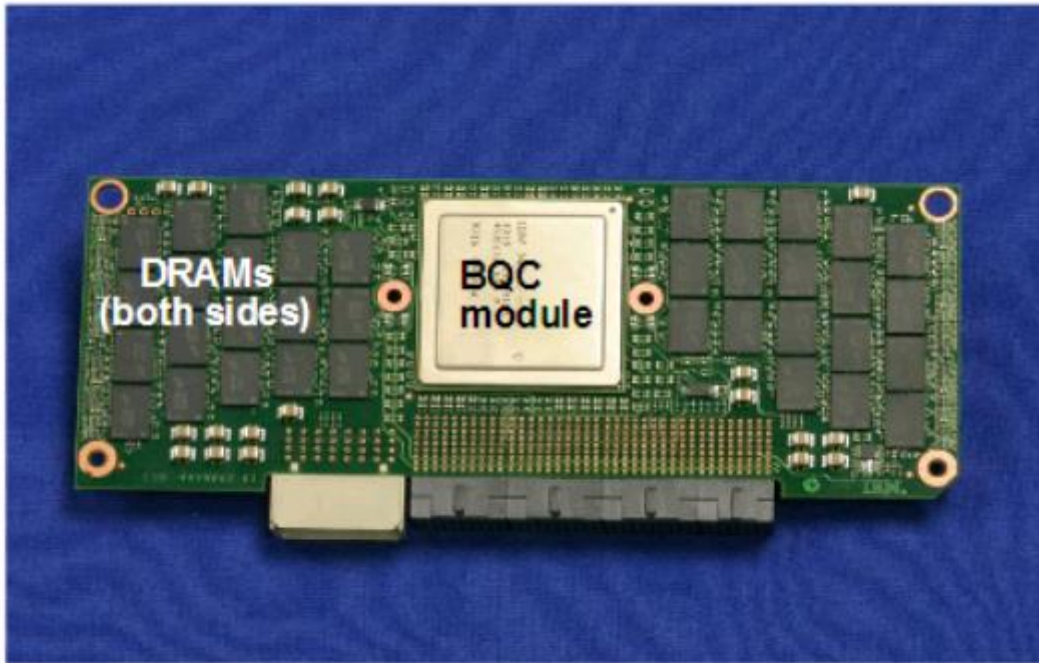
Architecture: 10 BGQ Frame  
Model: IBM-BG/Q  
Processor Type: IBM PowerA2, 1.6 GHz  
Computing Cores: 163840  
Computing Nodes: 10240  
RAM: 1 GByte / core  
Internal Network: 5D Torus  
Disk Space: 2 PByte of scratch space  
Peak Performance: 2 PFlop/s  
Power Consumption: 1 MWatt





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



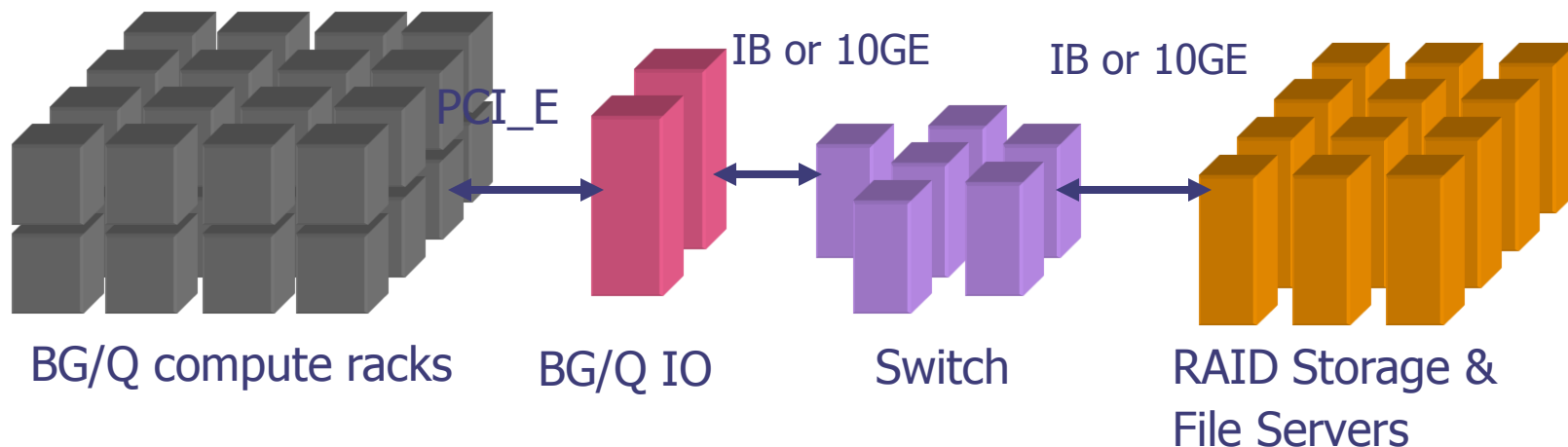


Point-to-point fiber cables,  
attaching the 8 I/O nodes  
(on top of rack)  
to compute nodes  
(on 8 node cards)



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
- Two I/O configurations (one traditional, one conceptual)

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)

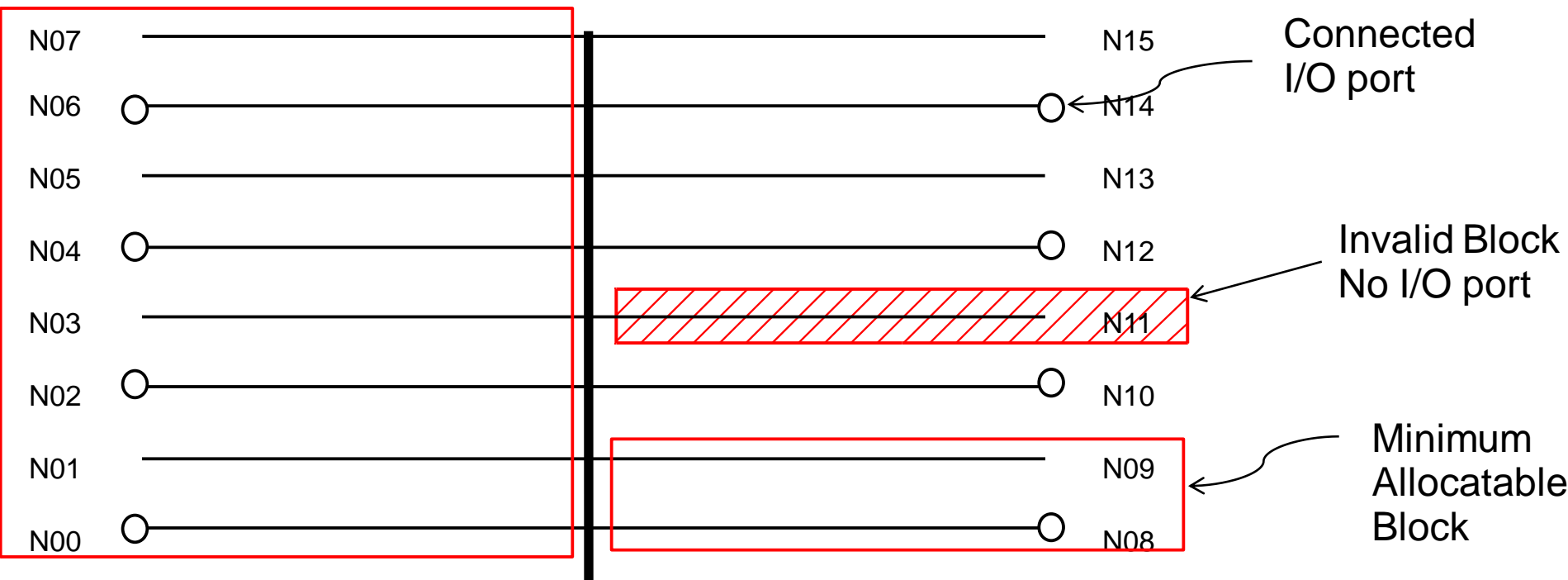
8 racks with 8 I/O nodes (2048 cores per I/O node)

# 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)

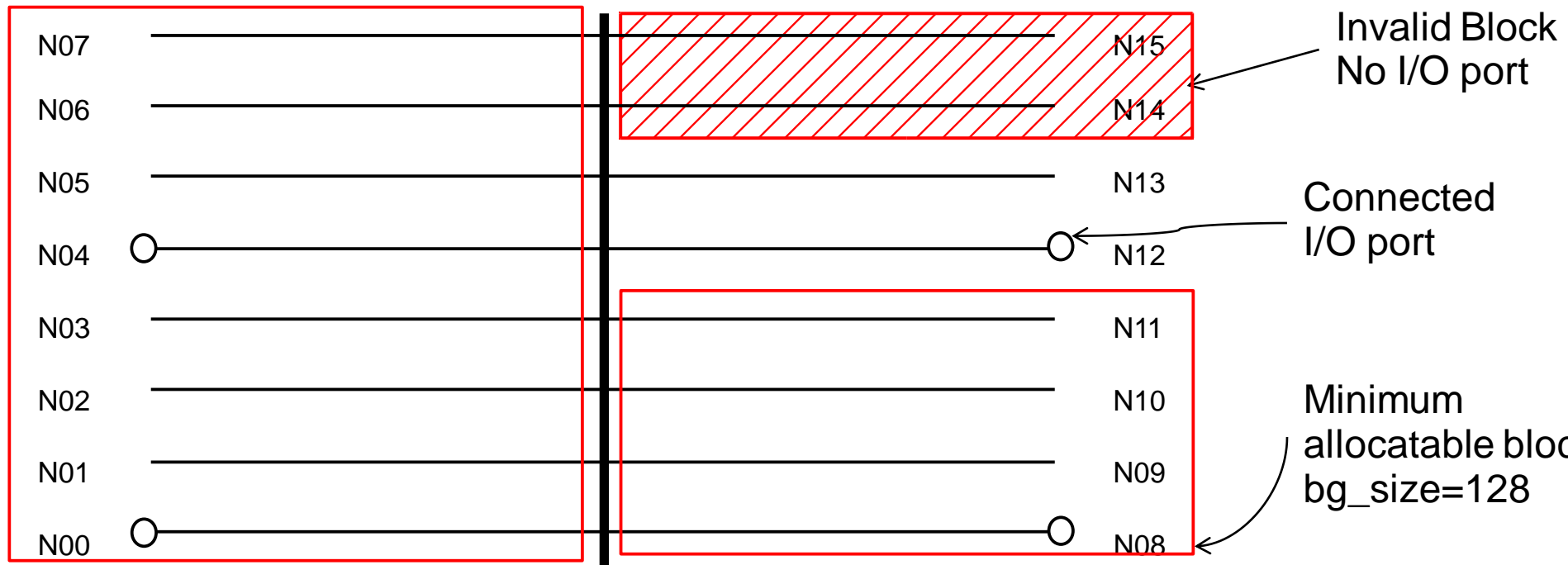
# MidPlane in FERMI RACK: R11, R31



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:

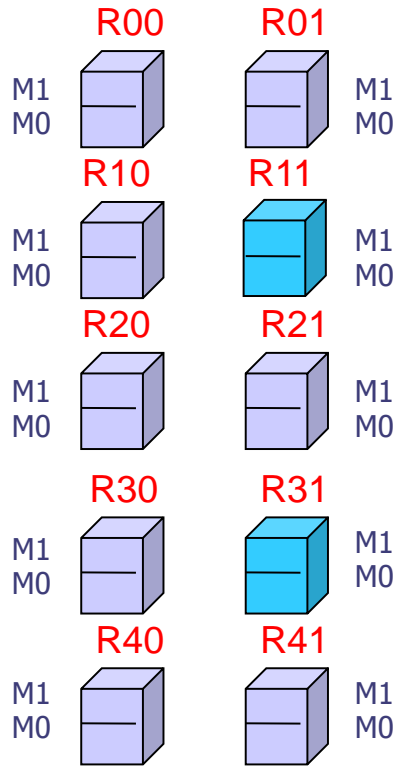
- made up of one or more node boards within a single midplane
- always multiple of 32 nodes
- Not a torus in all five dimensions (see later)

## ▪ Large blocks:

- made up of one or more complete midplanes
- always multiple of 512 nodes
- Can be a torus in all five dimensions

# 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
- 11<sup>th</sup> port for auto-routing to IO fabric



**10 racks**

- 5 rows
- 2 columns

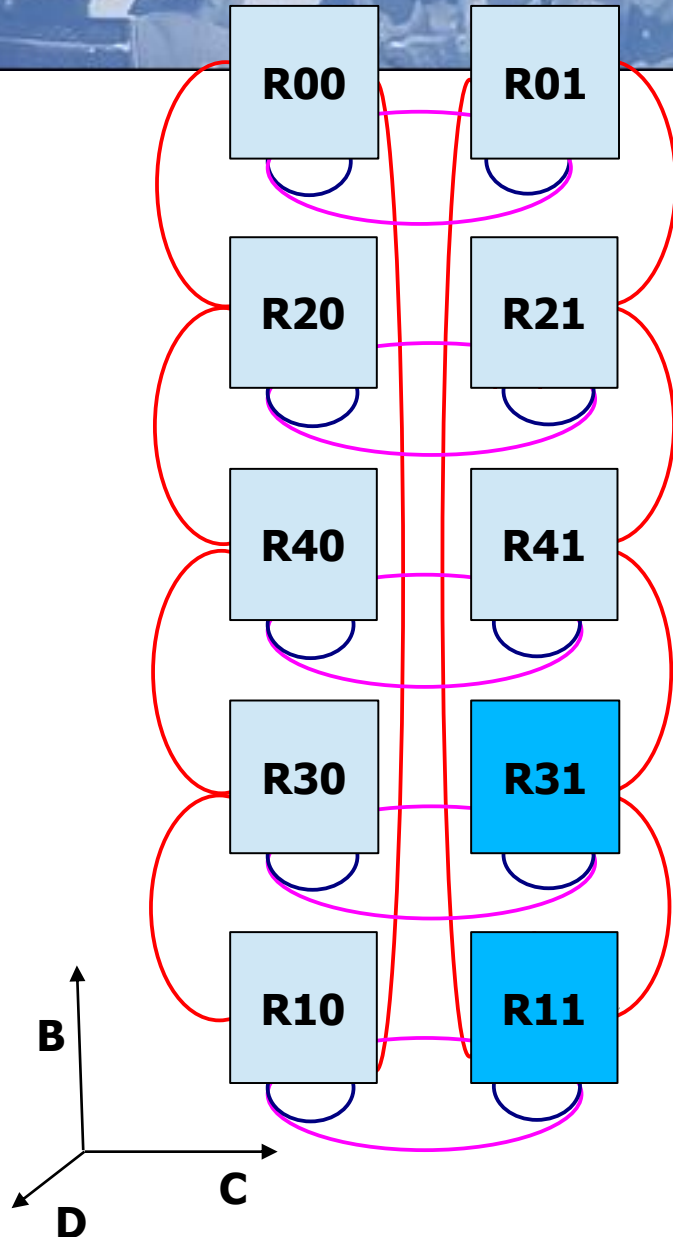
**20 midplanes**

- 2 midplanes for each rack

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

— Rack with 8 IO Nodes

— Rack with 16 IO Nodes



## 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 represented

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

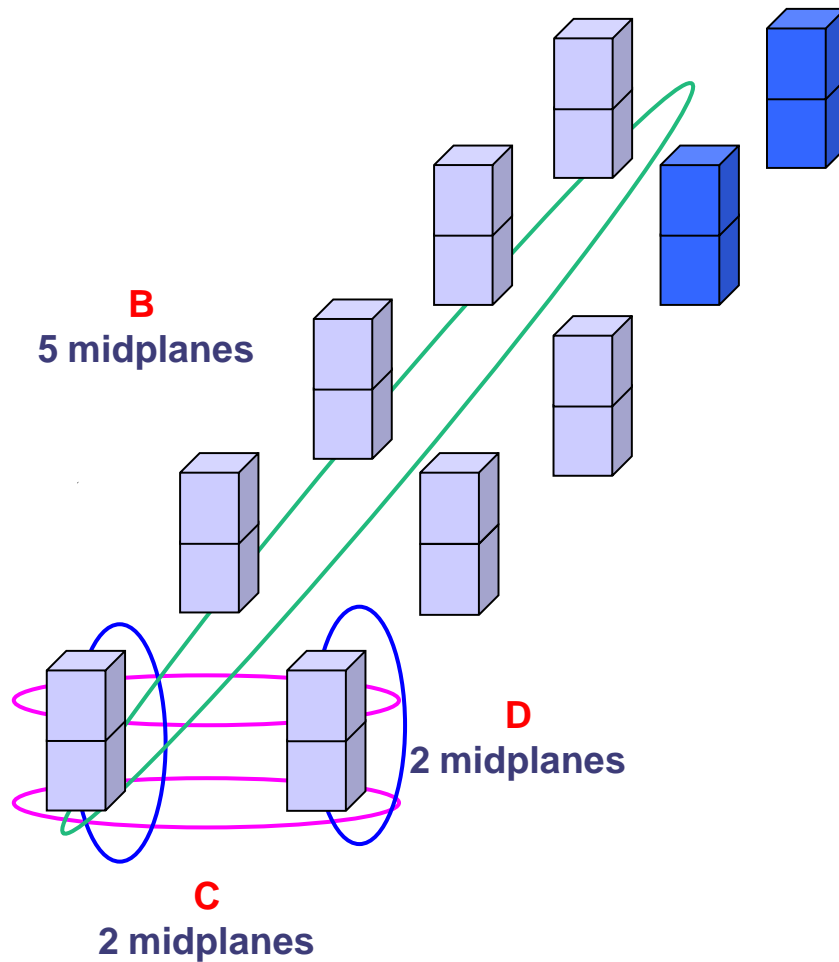
**SHAPE of FERMI =**

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

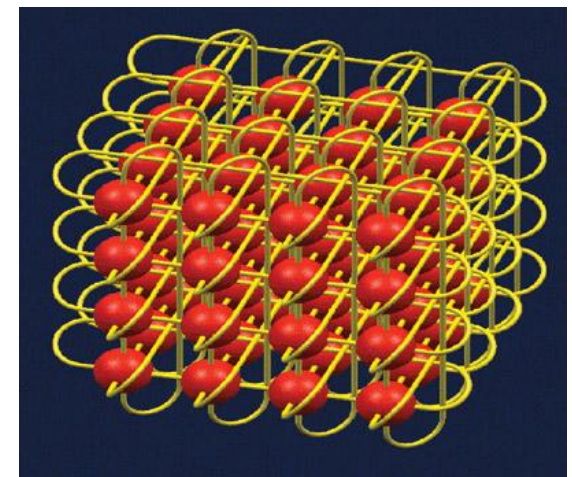
$$1 \times 5 \times 2 \times 2 = 20 \text{ MidPlanes}$$

For **large block jobs** ( $\geq 1\text{MP}$ ) two connectivity between midplanes are provided:

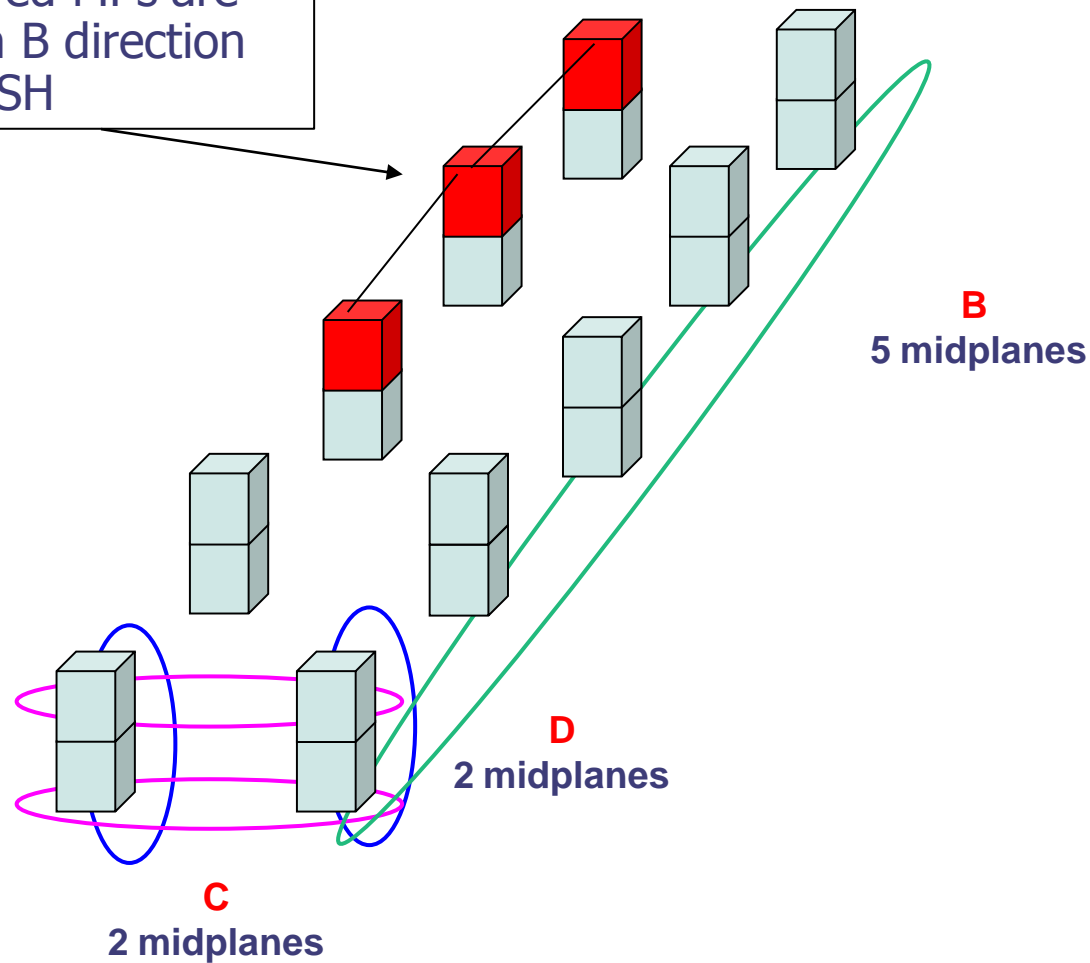
- **Torus** : periodic boundary conditions (e.g. “close line”) in all the dimensions A, B, C and D.
- **Mesh** : almost one dimension is not like a “close line”



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



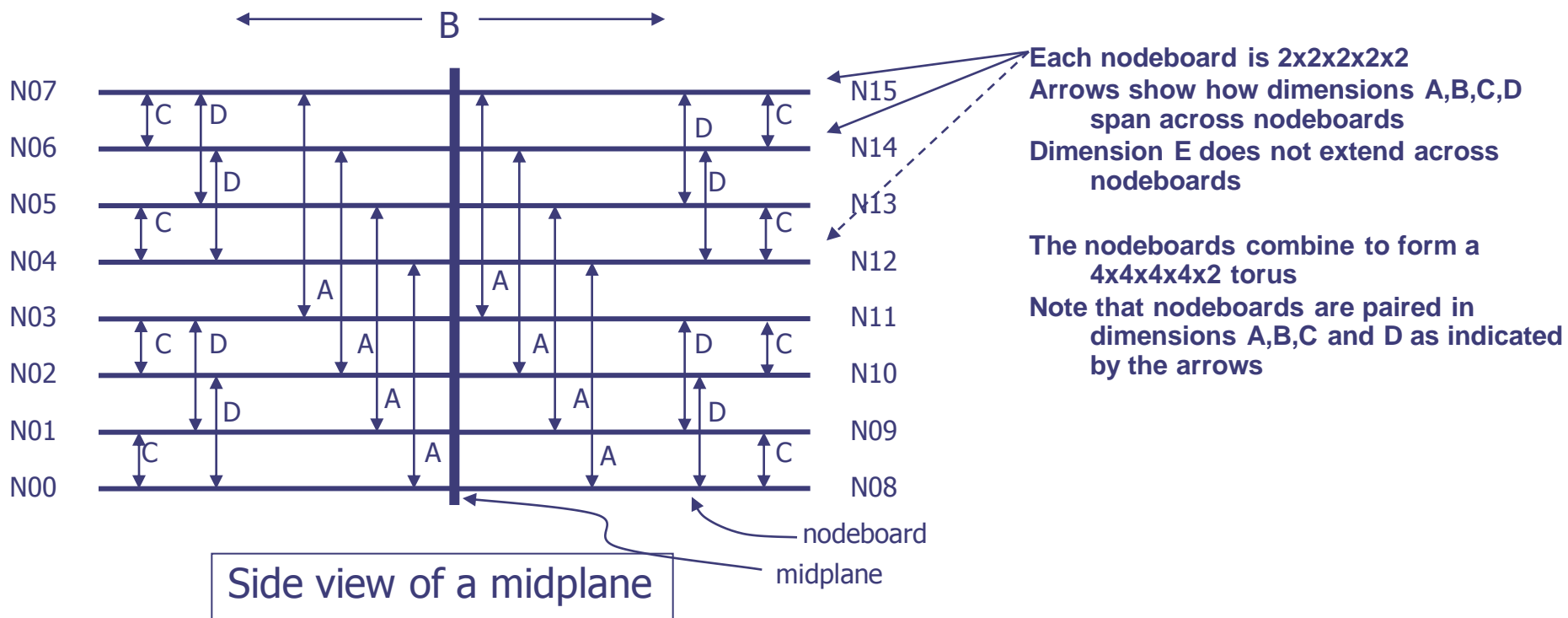
▪ The 3 red MPs are linked in B direction as a MESH





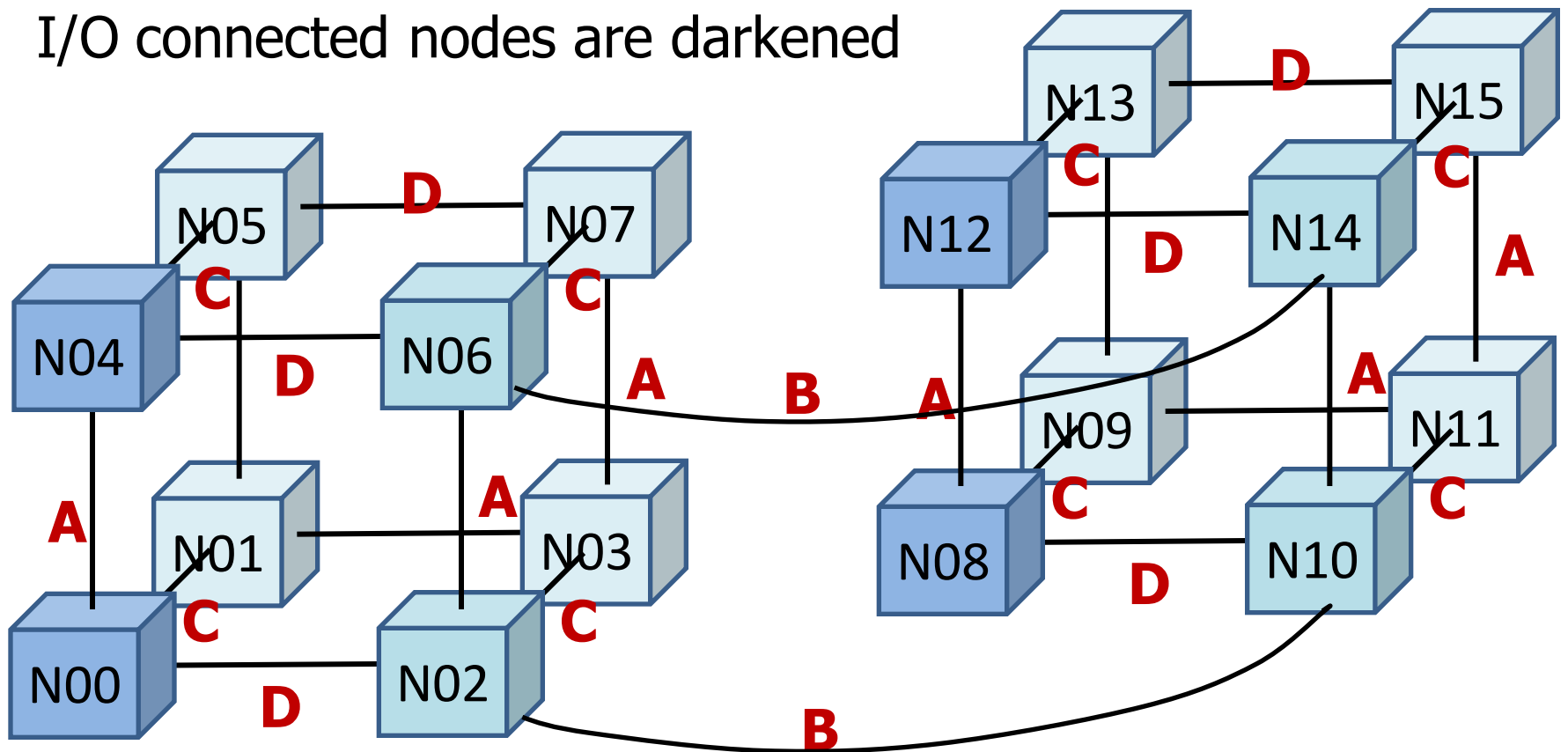
# 5-D torus wiring in a Midplane

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.

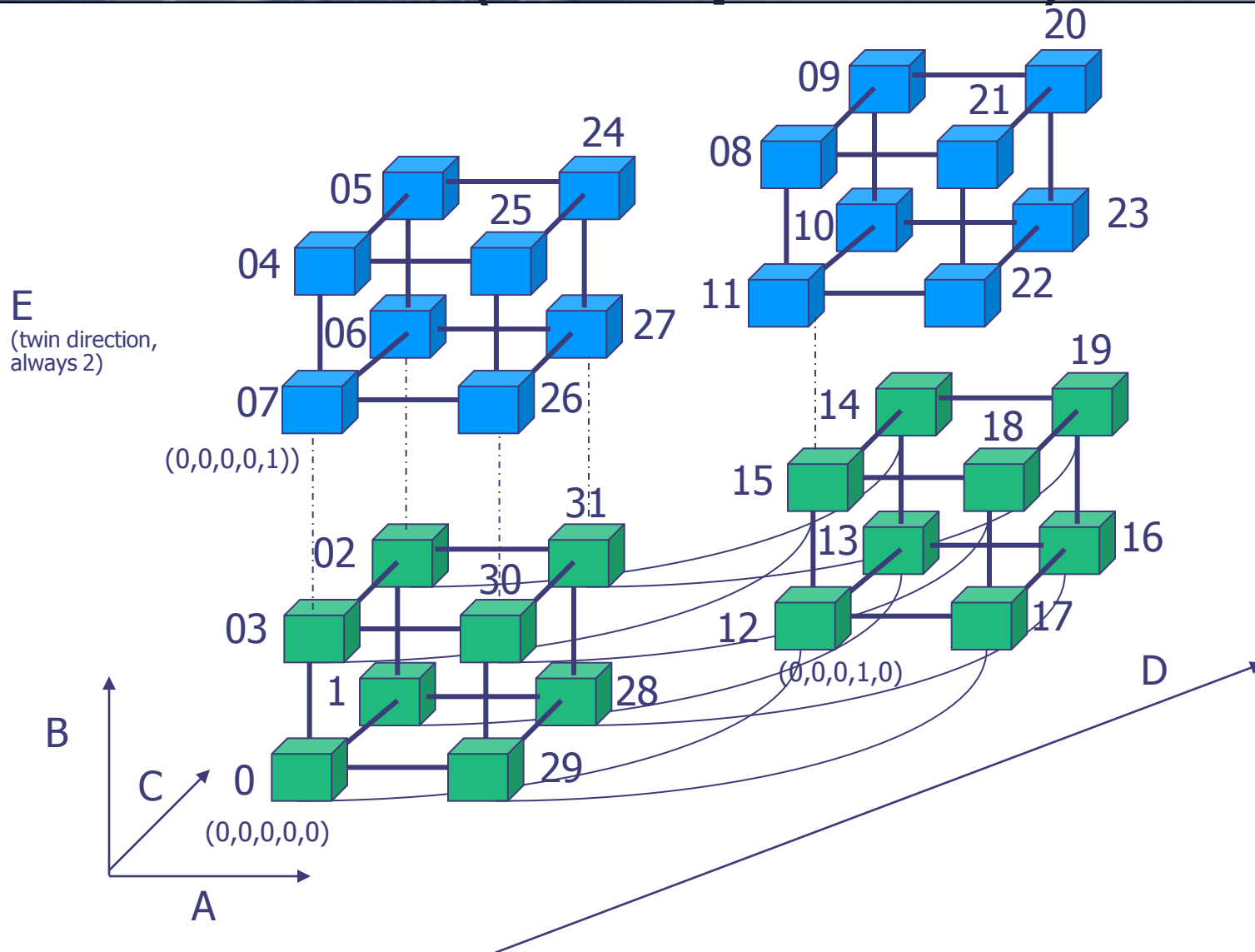


# 5-D torus in a Midplane

I/O connected nodes are darkened



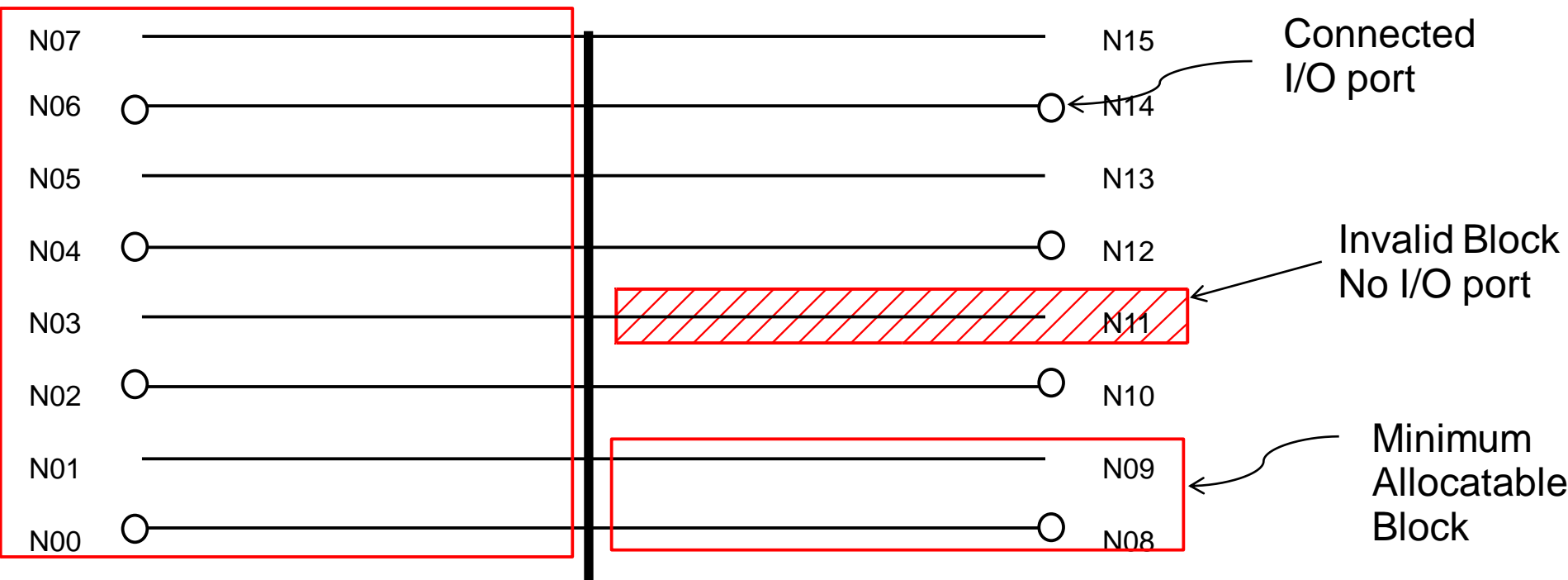
# Node Board (32 Compute Nodes): 2x2x2x2x2



# 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

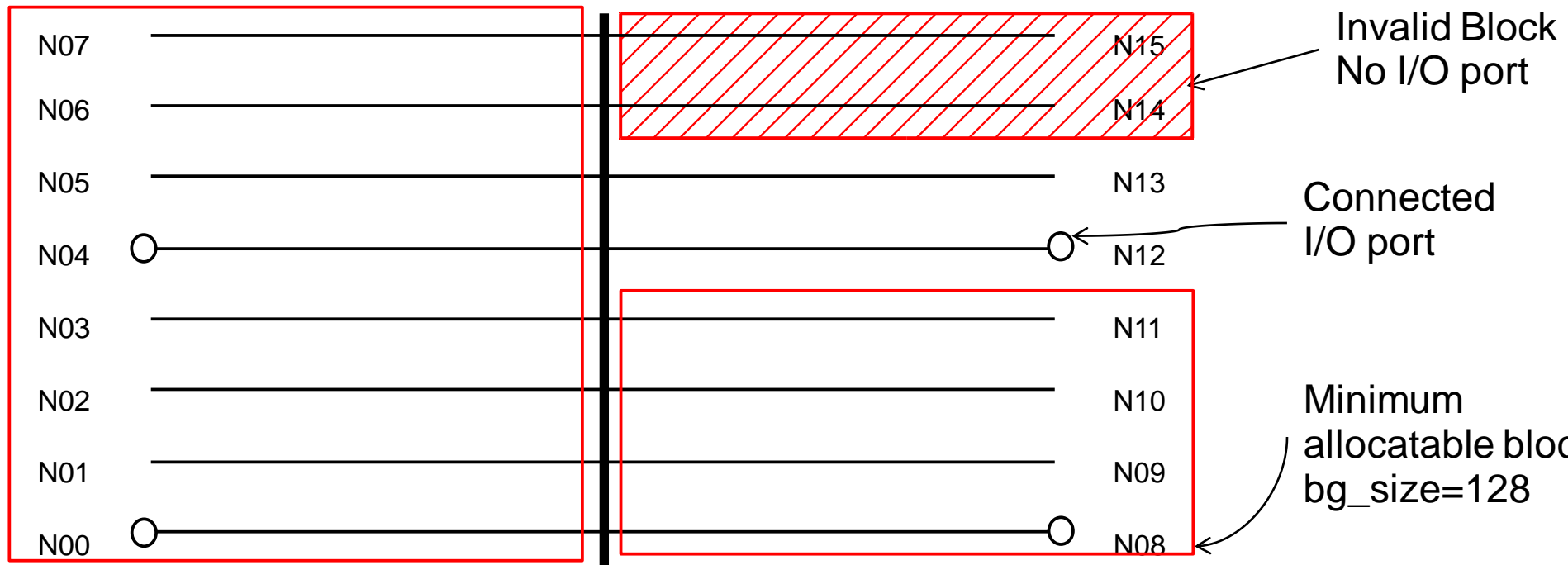
# MidPlane in FERMI RACK: R11 R31



Example:

N08 – N09 = 64 Compute Cards (2x2x4x2x2)

# MidPlane in FERMI / {R11 R31}



Example:

$N08 - N09 - N10 - N11 = 128$  Compute Cards (2x2x4x4x2)

# BGQ PowerA2 processor

Carlo Cavazzoni, HPC department, CINECA



# Power A2

64bit

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

32MByte cache

system-on-a-chip design

16GByte of RAM at 1.33GHz

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

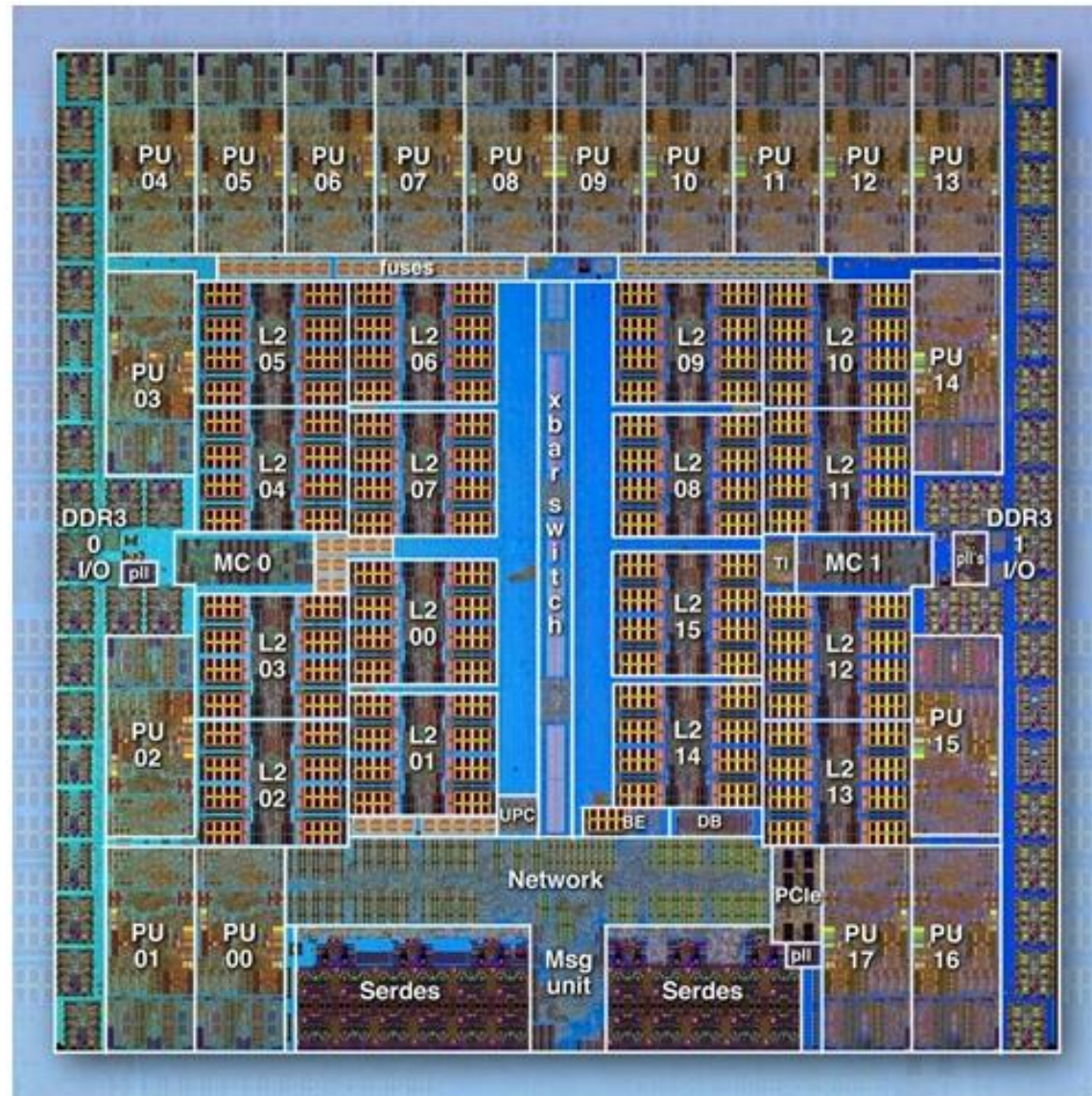
Two of these can be used for PCI-Express 2.0 x8 peripheral slots.

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

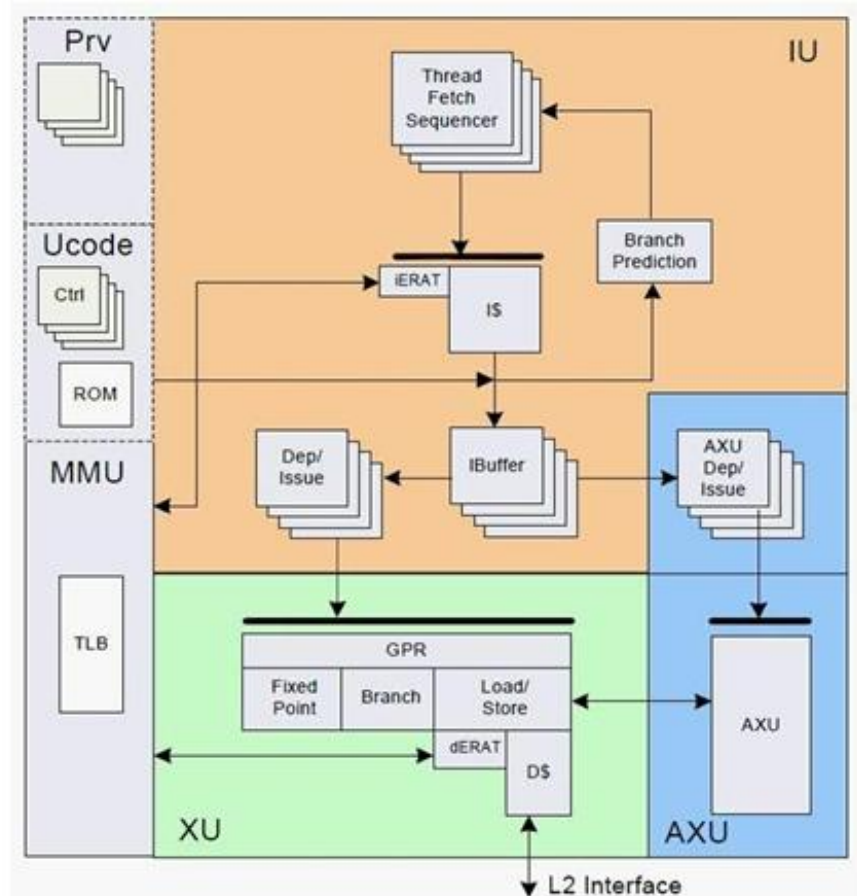
4 FPU

4 way SMT

64-bit instruction set - in-order dispatch,  
 execution, and completion

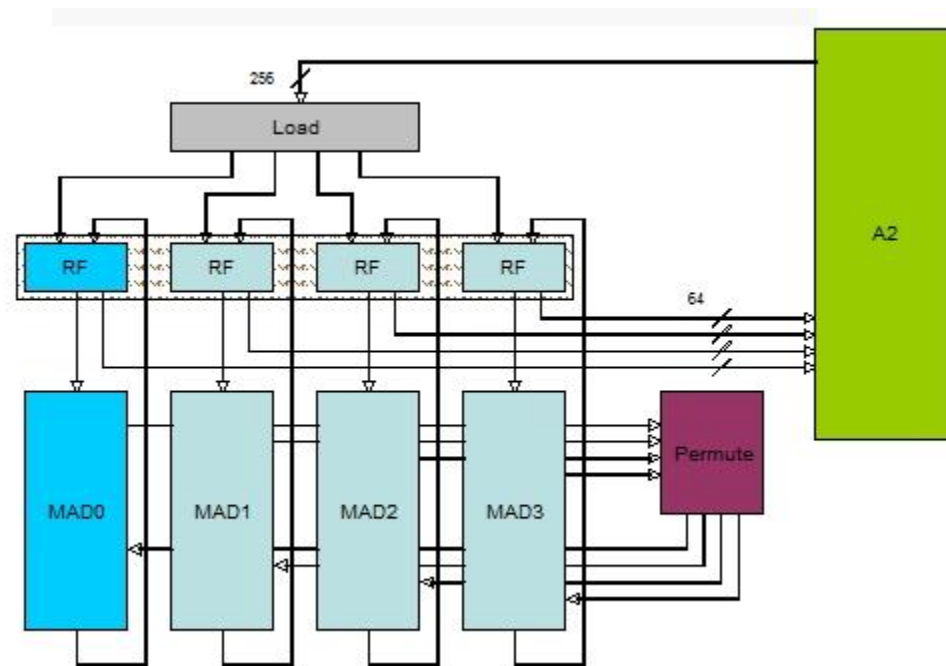
16KB of L1 data cache

16KB of L1 instructions cache



# PowerA2 FPU

Each FPU on each core has four pipelines  
execute scalar floating point instructions  
Quad pumped  
four-wide SIMD instructions  
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?