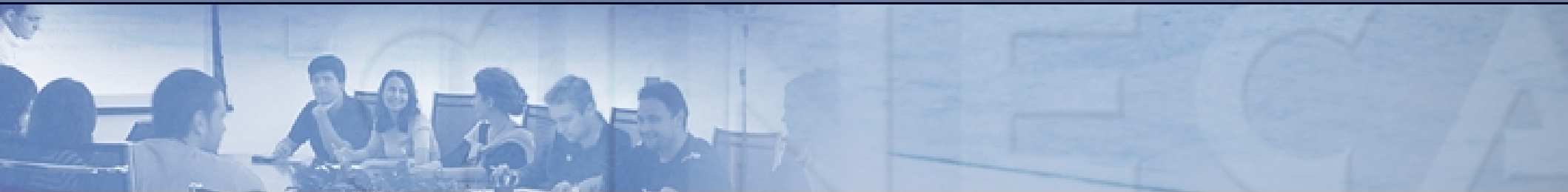


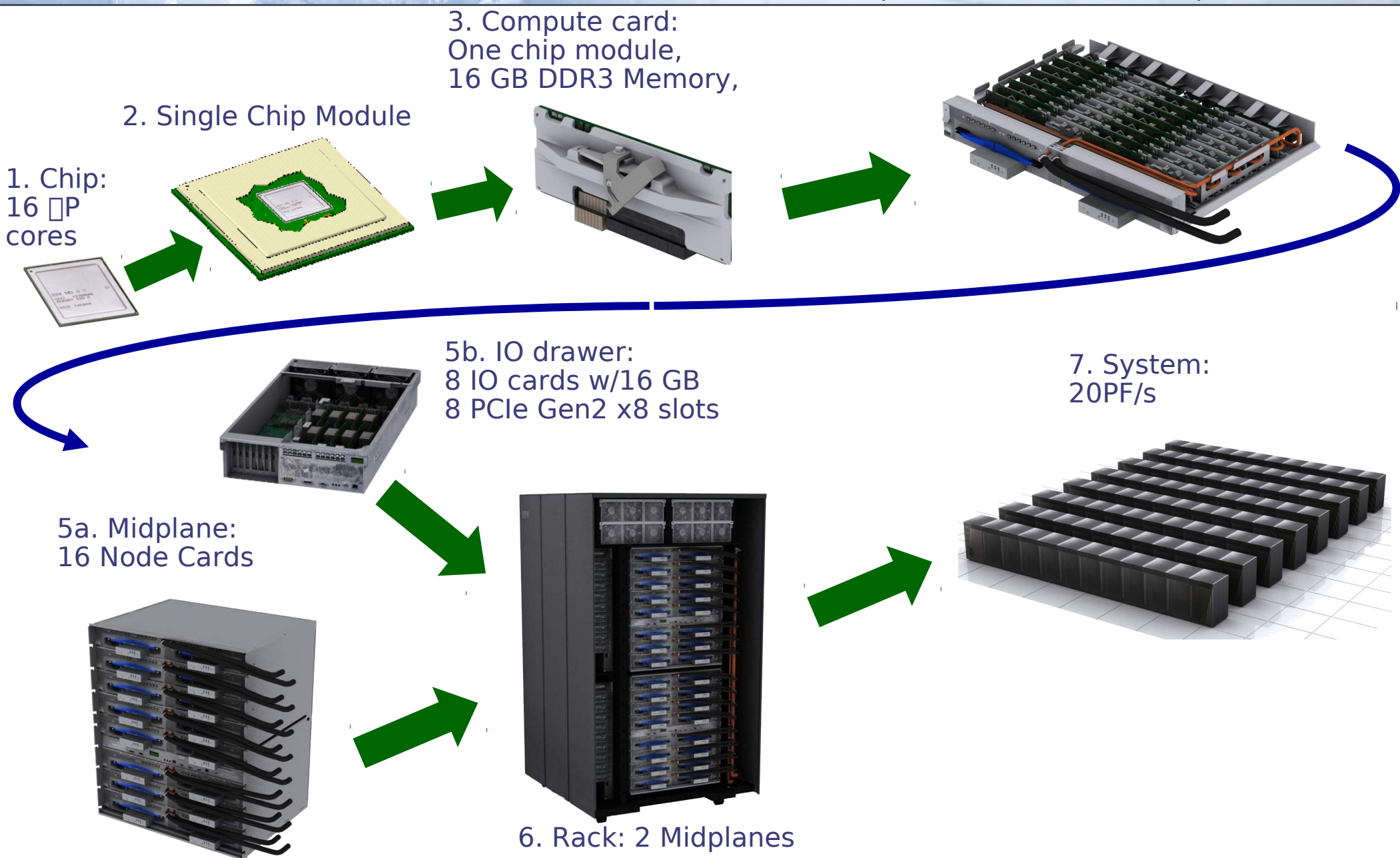
Network topology on FERMI



g.muscianisi@cineca.it
silvia.giuliani@cineca.it

Why the network topology is important

- Feature in a BG/Q system
- How ask the computational resources for a “large jobs”
- How to map your own phisical domain into the BG/Q hardware



Blue Gene/Q System Configurations

Torus size, in nodes

Torus size, in midplanes

BG/Q Nodes form a 5D torus

Nodecards: 2x2x2x2x2

Midplanes: 4x4x4x4x2

- 4D are cabled to other midplanes

5th dimension: extent 2 (stays within nodecard)

6th dimension is cpu # within the node

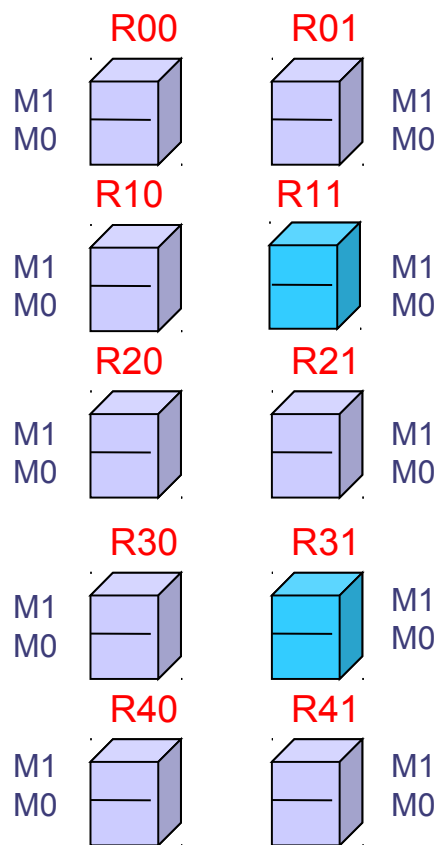
Dim. labels: ABCDE T

Different floor shapes

(Rows x Cols) for a given number of racks may correspond to the same, or to different torus shapes

This list is not complete;
other configs are possible...
up to 16x16 = 256 racks

| Racks | Rows | Col. | A | B | C | D | E | A | B | C | D | Nodes | Cores | Threads |
|-------|------|------|----|----|----|----|---|---|---|---|---|--------|-----------|-----------|
| mid | 1 | 1 | 4 | 4 | 4 | 4 | 2 | 1 | 1 | 1 | 1 | 512 | 8.192 | 32.768 |
| 1 | 1 | 1 | 4 | 4 | 4 | 8 | 2 | 1 | 1 | 1 | 2 | 1.024 | 16.384 | 65.536 |
| 2 | 1 | 2 | 4 | 4 | 8 | 8 | 2 | 1 | 1 | 2 | 2 | 2.048 | 32.768 | 131.072 |
| 3 | 1 | 3 | 4 | 4 | 12 | 8 | 2 | 1 | 1 | 3 | 2 | 3.072 | 49.152 | 196.608 |
| 4 | 1 | 4 | 8 | 4 | 8 | 8 | 2 | 2 | 1 | 2 | 2 | 4.096 | 65.536 | 262.144 |
| 4 | 2 | 2 | 4 | 8 | 8 | 8 | 2 | 1 | 2 | 2 | 2 | 4.096 | 65.536 | 262.144 |
| 6 | 1 | 6 | 12 | 4 | 8 | 8 | 2 | 3 | 1 | 2 | 2 | 6.144 | 98.304 | 393.216 |
| 6 | 2 | 3 | 4 | 8 | 12 | 8 | 2 | 1 | 2 | 3 | 2 | 6.144 | 98.304 | 393.216 |
| 6 | 3 | 2 | 4 | 12 | 8 | 8 | 2 | 1 | 3 | 2 | 2 | 6.144 | 98.304 | 393.216 |
| 8 | 1 | 8 | 8 | 8 | 8 | 8 | 2 | 2 | 2 | 2 | 2 | 8.192 | 131.072 | 524.288 |
| 8 | 2 | 4 | 8 | 8 | 8 | 8 | 2 | 2 | 2 | 2 | 2 | 8.192 | 131.072 | 524.288 |
| 8 | 4 | 2 | 8 | 8 | 8 | 8 | 2 | 2 | 2 | 2 | 2 | 8.192 | 131.072 | 524.288 |
| 10 | 5 | 2 | 4 | 20 | 8 | 8 | 2 | 1 | 5 | 2 | 2 | 10.240 | 163.840 | 655.360 |
| 12 | 3 | 4 | 8 | 12 | 8 | 8 | 2 | 2 | 3 | 2 | 2 | 12.288 | 196.608 | 786.432 |
| 12 | 2 | 6 | 12 | 8 | 8 | 8 | 2 | 3 | 2 | 2 | 2 | 12.288 | 196.608 | 786.432 |
| 16 | 2 | 8 | 16 | 8 | 8 | 8 | 2 | 4 | 2 | 2 | 2 | 16.384 | 262.144 | 1.048.576 |
| 16 | 4 | 4 | 8 | 16 | 8 | 8 | 2 | 2 | 4 | 2 | 2 | 16.384 | 262.144 | 1.048.576 |
| 20 | 5 | 4 | 8 | 20 | 8 | 8 | 2 | 2 | 5 | 2 | 2 | 20.480 | 327.680 | 1.310.720 |
| 24 | 6 | 4 | 8 | 12 | 16 | 8 | 2 | 2 | 3 | 4 | 2 | 24.576 | 393.216 | 1.572.864 |
| 24 | 2 | 12 | 8 | 8 | 12 | 16 | 2 | 2 | 2 | 3 | 4 | 24.576 | 393.216 | 1.572.864 |
| 28 | 7 | 4 | 8 | 28 | 8 | 8 | 2 | 2 | 7 | 2 | 2 | 28.672 | 458.752 | 1.835.008 |
| 32 | 8 | 4 | 8 | 16 | 16 | 8 | 2 | 2 | 4 | 4 | 2 | 32.768 | 524.288 | 2.097.152 |
| 32 | 4 | 8 | 8 | 16 | 16 | 8 | 2 | 2 | 4 | 4 | 2 | 32.768 | 524.288 | 2.097.152 |
| 40 | 5 | 8 | 8 | 20 | 16 | 8 | 2 | 2 | 5 | 4 | 2 | 40.960 | 655.360 | 2.621.440 |
| 48 | 6 | 8 | 16 | 12 | 16 | 8 | 2 | 4 | 3 | 4 | 2 | 49.152 | 786.432 | 3.145.728 |
| 48 | 4 | 12 | 8 | 16 | 12 | 16 | 2 | 2 | 4 | 3 | 4 | 49.152 | 786.432 | 3.145.728 |
| 48 | 3 | 16 | 8 | 12 | 16 | 16 | 2 | 2 | 3 | 4 | 4 | 49.152 | 786.432 | 3.145.728 |
| 56 | 7 | 8 | 8 | 28 | 16 | 8 | 2 | 2 | 7 | 4 | 2 | 57.344 | 917.504 | 3.670.016 |
| 64 | 8 | 8 | 8 | 16 | 16 | 16 | 2 | 2 | 4 | 4 | 4 | 65.536 | 1.048.576 | 4.194.304 |
| 64 | 4 | 16 | 8 | 16 | 16 | 16 | 2 | 2 | 4 | 4 | 4 | 65.536 | 1.048.576 | 4.194.304 |
| 72 | 9 | 8 | 12 | 12 | 16 | 16 | 2 | 3 | 3 | 4 | 4 | 73.728 | 1.179.648 | 4.718.592 |
| 80 | 10 | 8 | 8 | 20 | 16 | 16 | 2 | 2 | 5 | 4 | 4 | 81.920 | 1.310.720 | 5.242.880 |
| 96 | 12 | 8 | 16 | 12 | 16 | 16 | 2 | 4 | 3 | 4 | 4 | 98.304 | 1.572.864 | 6.291.456 |
| 96 | 8 | 12 | 16 | 16 | 12 | 16 | 2 | 4 | 4 | 3 | 4 | 98.304 | 1.572.864 | 6.291.456 |
| 96 | 6 | 16 | 16 | 12 | 16 | 16 | 2 | 4 | 3 | 4 | 4 | 98.304 | 1.572.864 | 6.291.456 |



10 racks

- 5 rows
- 2 columns

20 midplanes

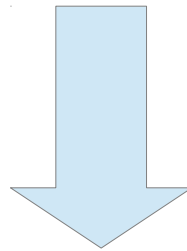
- 2 midplanes for each rack

| 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

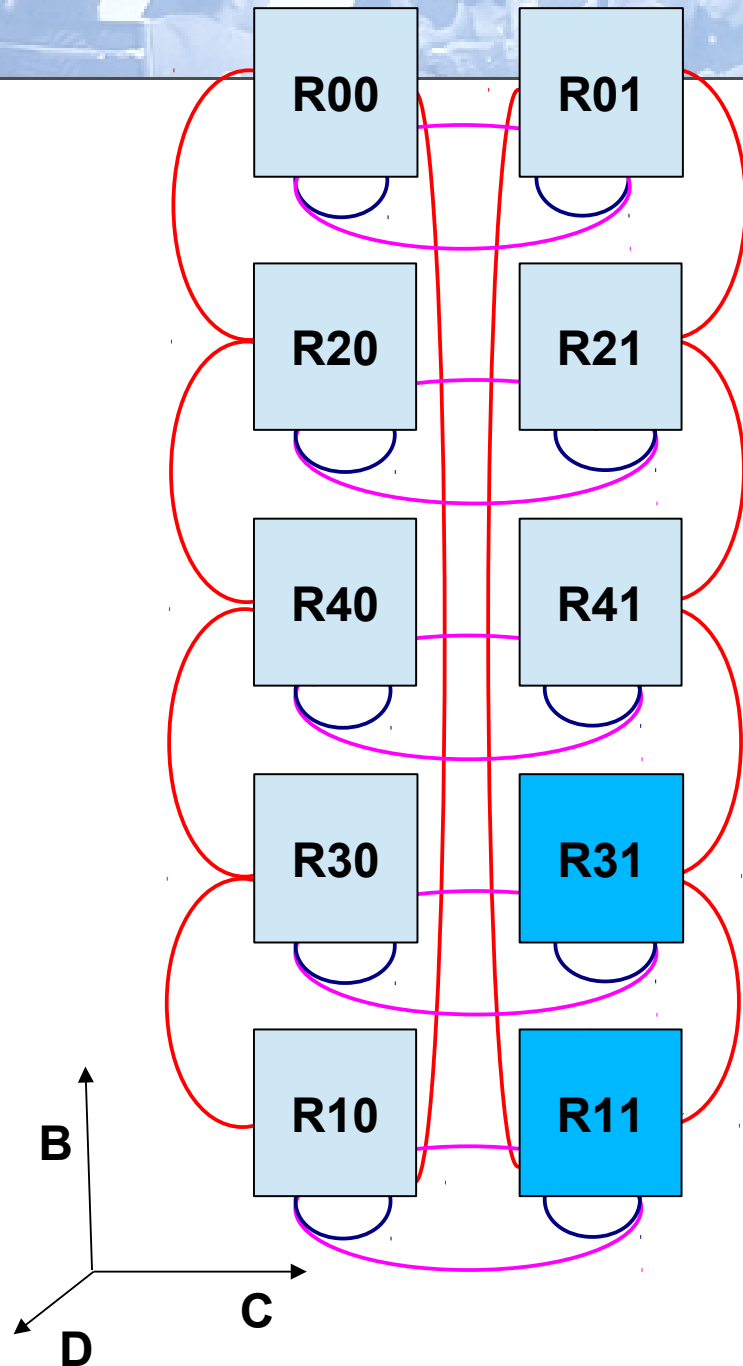
Fixed the number of the Racks in a BG/Q configuration



SHAPE of a BG/Q system =
number of midplanes on A, B, C, D directions

$$A \times B \times C \times D$$

Midplanes CABLING



B dimension

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

C dimension

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

D dimension

- connection among 2 midplains 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 |

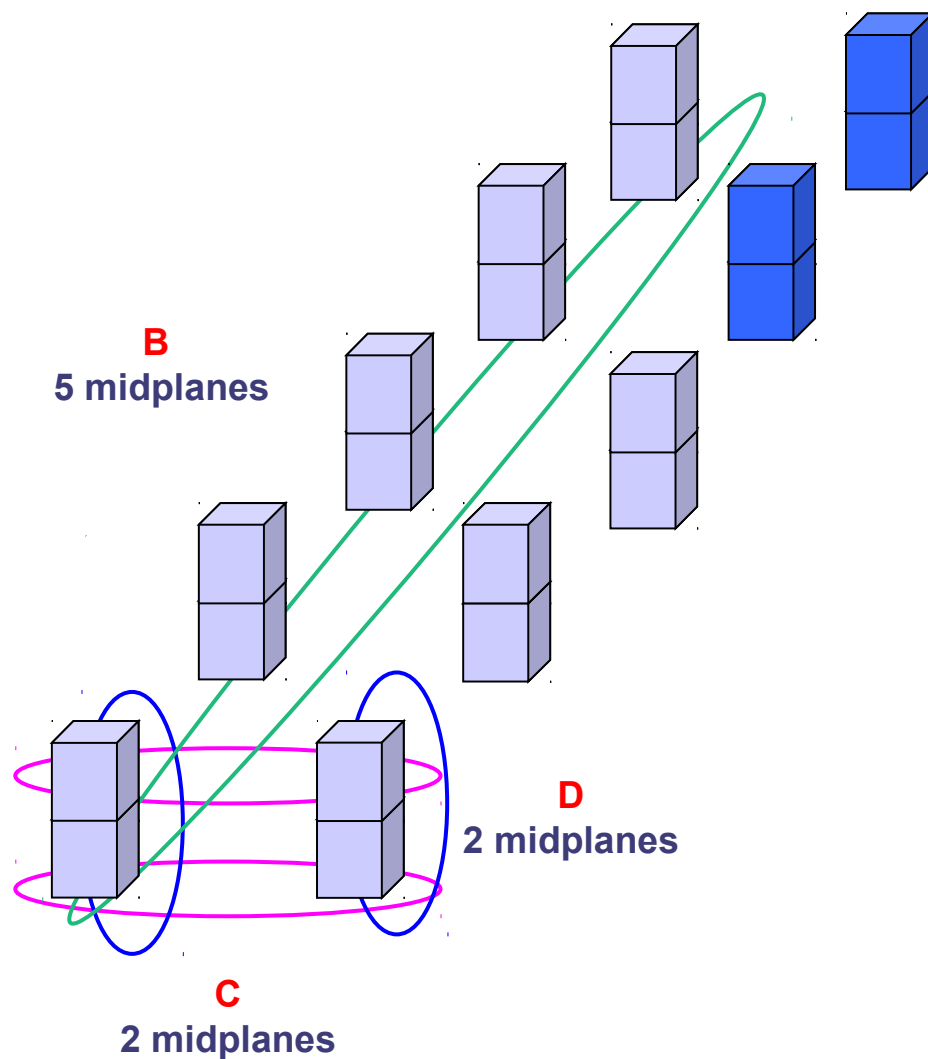
SHAPE of FERMI =

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

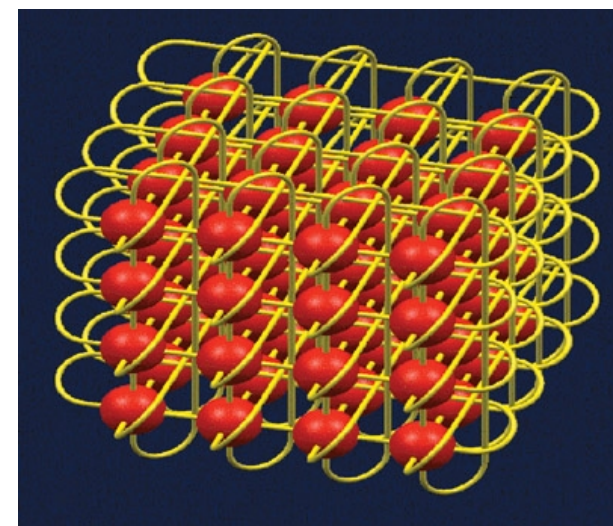
$$1 \times 5 \times 2 \times 2 = 20 \text{ MidPlanes}$$

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

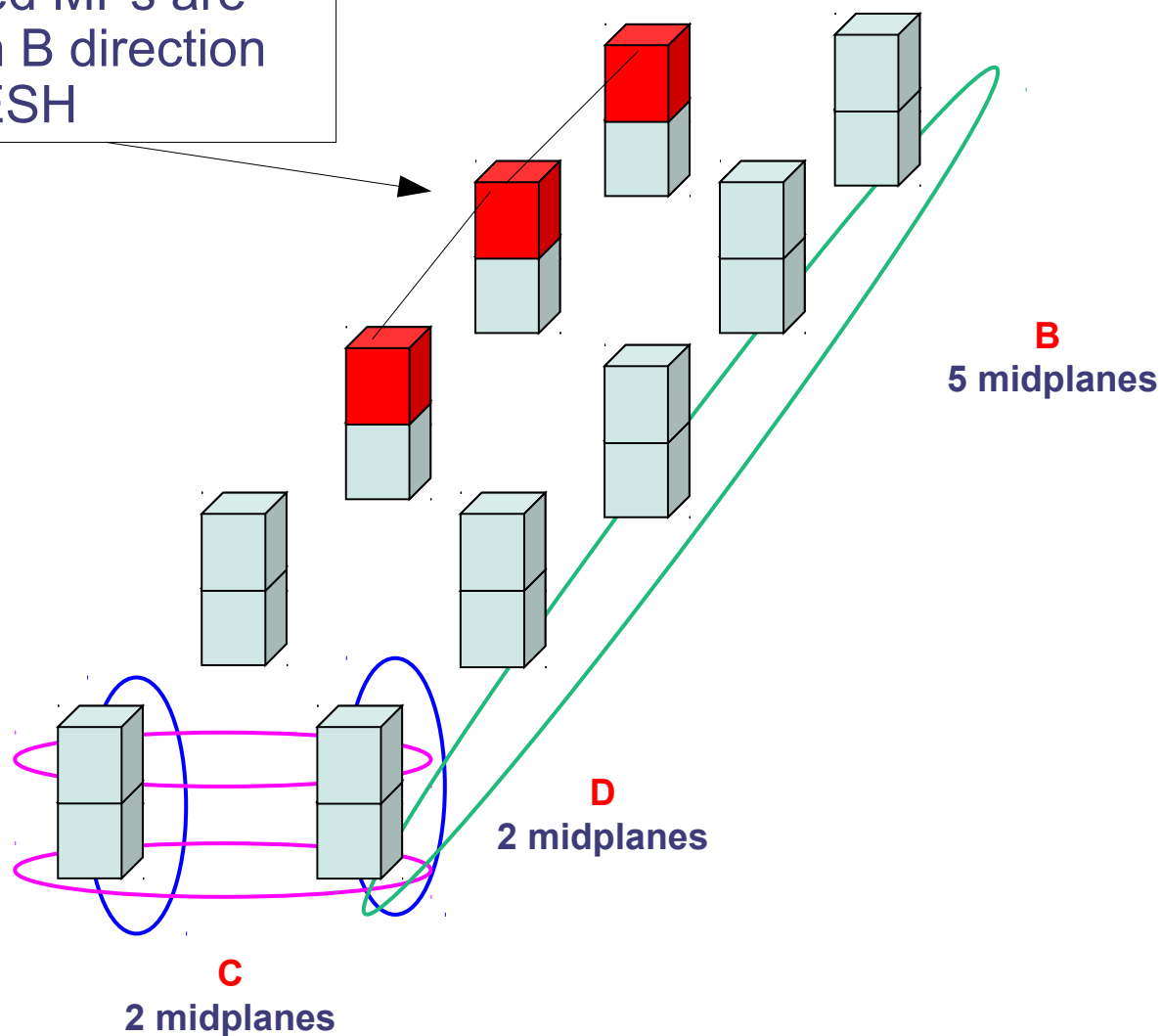
- **Torus** : periodic boundary conditions (e.g. “close line”) in all the dimensions A, B, C and D.
- **Mesh** : almost one dimension is not like a “close line”

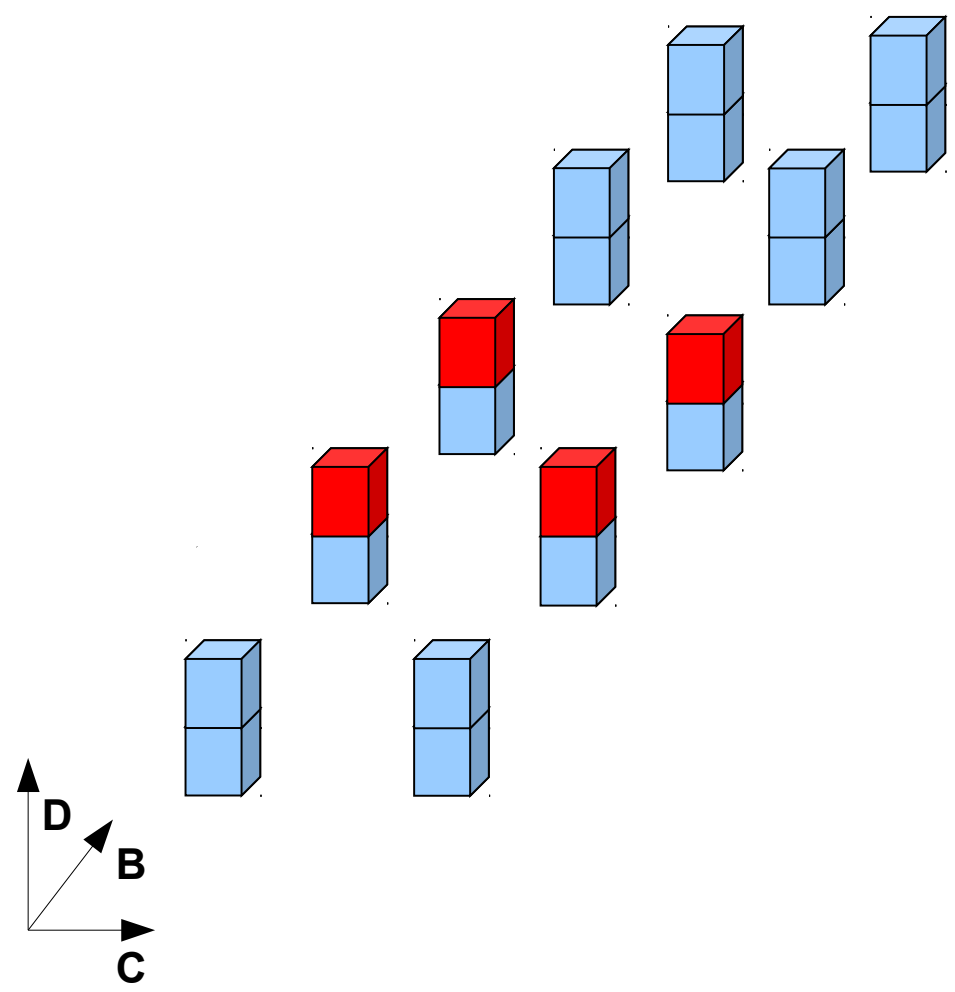


1 Midplane is the minimum TORUS available on a BlueGene/Q system

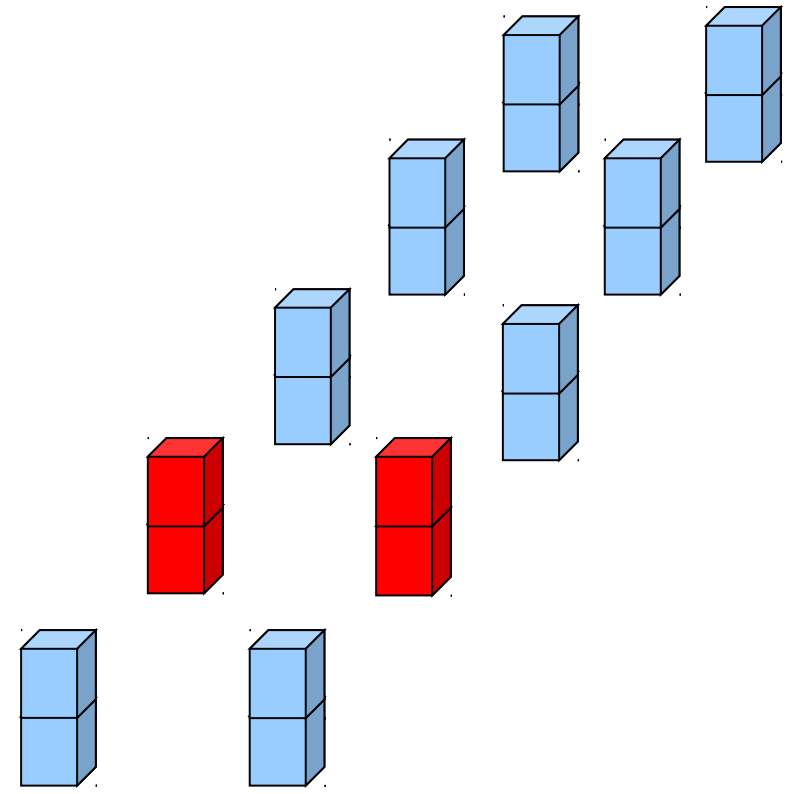


The 3 red MPs are linked in B direction as a MESH

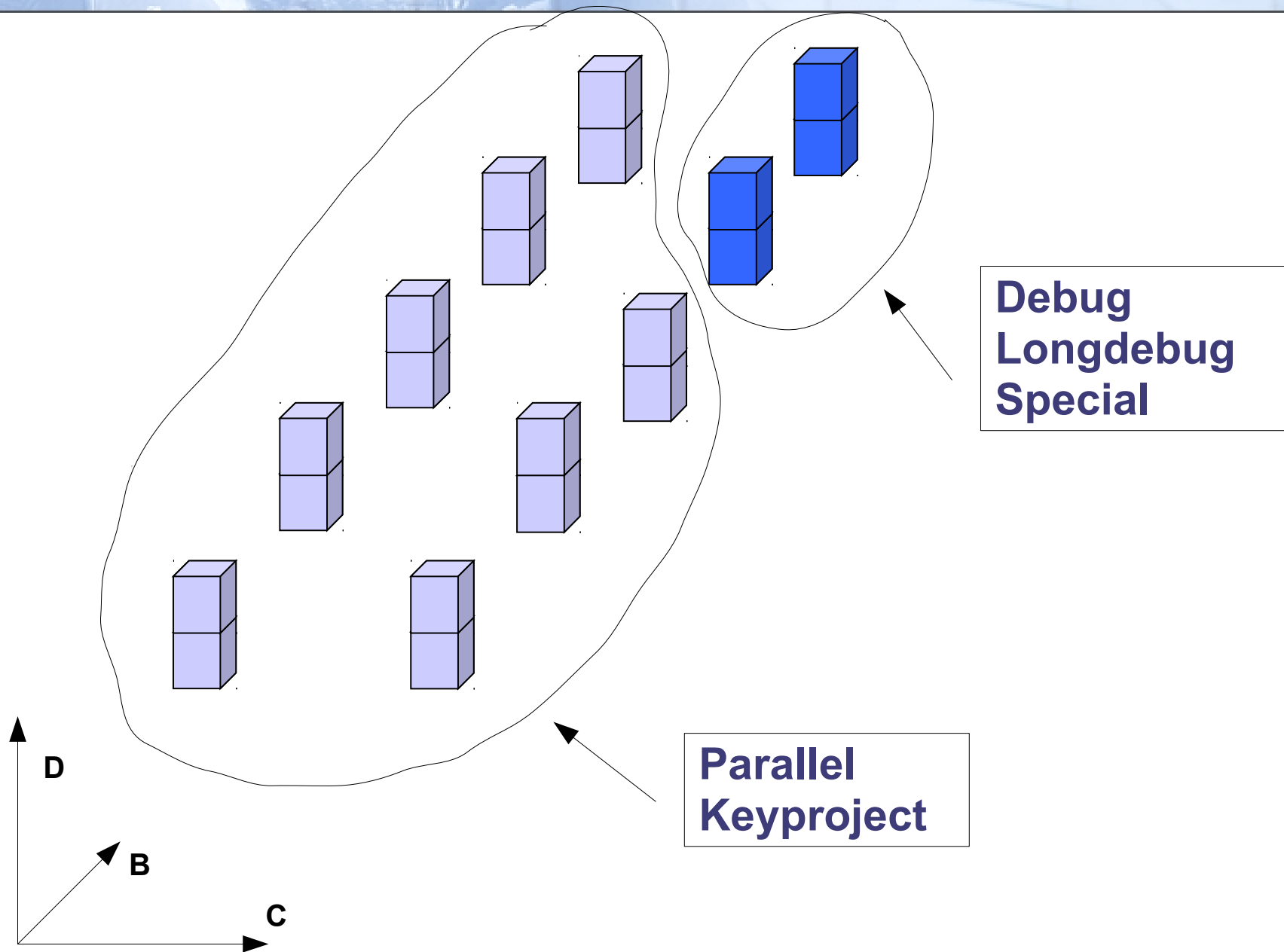




A Torus B Mesh C Torus D Torus



A Torus B Torus C Torus D Torus

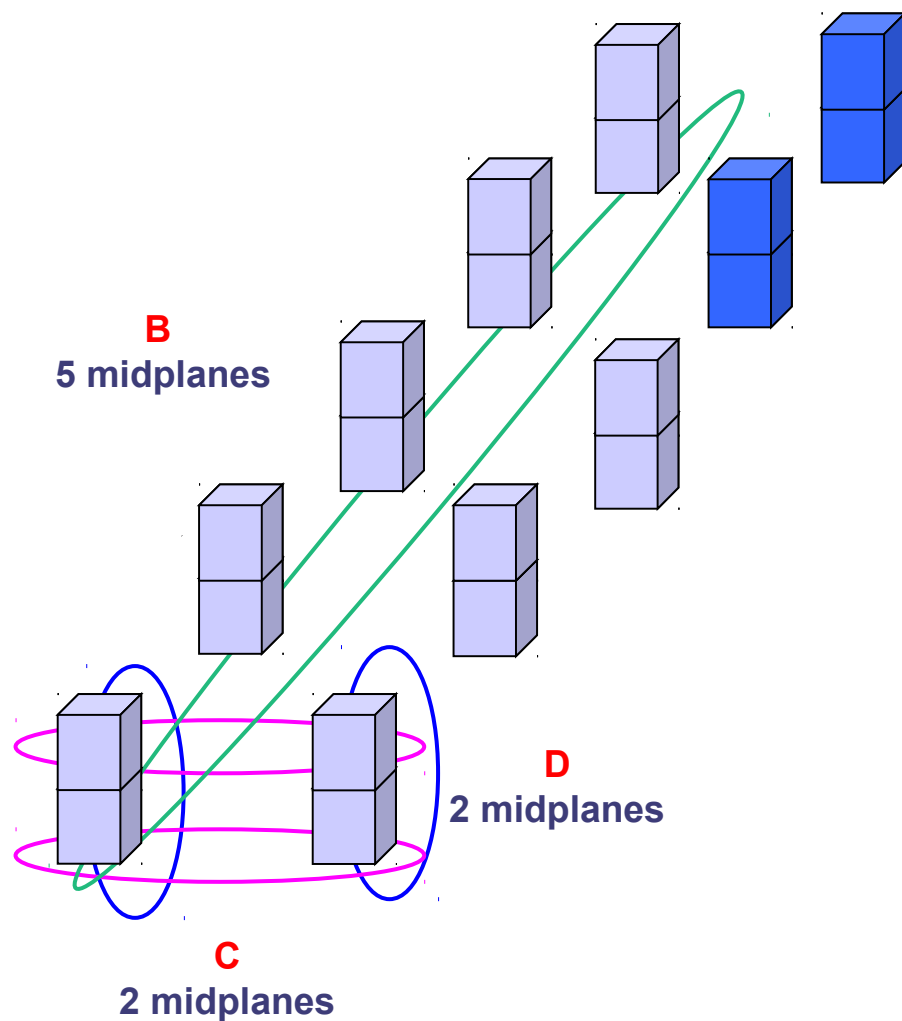


#@ bg_size = compute nodes number

OR

#@ bg_shape = $A \times B \times C \times D$

- **bg_shape** is usefull for mapping your phisical problem/domain into hardware
- **bg_shape** and **bg_size** keywords are mutually exclusive



FERMI shape = 1 x 5 x 2 x 2

| MP | bg_size | bg_shape | |
|----|---------|----------|-----|
| 3 | 1 536 | 1x1x3x1 | no |
| 3 | 1 536 | 1x3x1x1 | yes |
| 4 | 2 048 | 1x4x1x1 | yes |
| 4 | 2 048 | 1x1x2x2 | yes |
| 4 | 2 048 | 1x1x4x1 | no |
| 6 | 3 072 | 6x1x1x1 | no |
| 6 | 3 072 | 1x6x1x1 | no |
| 6 | 3 072 | 1x3x2x1 | yes |
| 6 | 3 072 | 1x2x3x1 | no |

REMARK

@ bg_shape = A x B x C x D

We cannot require values of A, B, C, and D greater than the corresponding A, B, C, and D sizes of FERMI, otherwise, the job will never be able to start

FERMI shape = 1 x 5 x 2 x 2

#@ bg_shape = A x B x C x D

AND

#@ bg_rotate = true|false

- LL should consider all possible permutations of the given shape
- Set **bg_rotate** to "false" to have the A, B, C and D dimensions of the allocation block exactly as defined by the **bg_shape** job command file keyword

4 midplanes

@ bg_shape = 1 x 1 x 2 x 2

@ bg_rotate = true (default)

LL should consider all possible permutation of the given shape



1 x 1 x 2 x 2

1 x 2 x 1 x 2

1 x 2 x 2 x 1

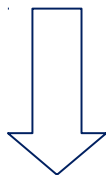
~~1 x 4 x 1 x 1~~

bg_shape & bg_rotate - example

4 midplanes

@ bg_shape = 1 x 1 x 2 x 2

@ bg_rotate = false



1 x 1 x 2 x 2

~~1 x 2 x 1 x 2~~

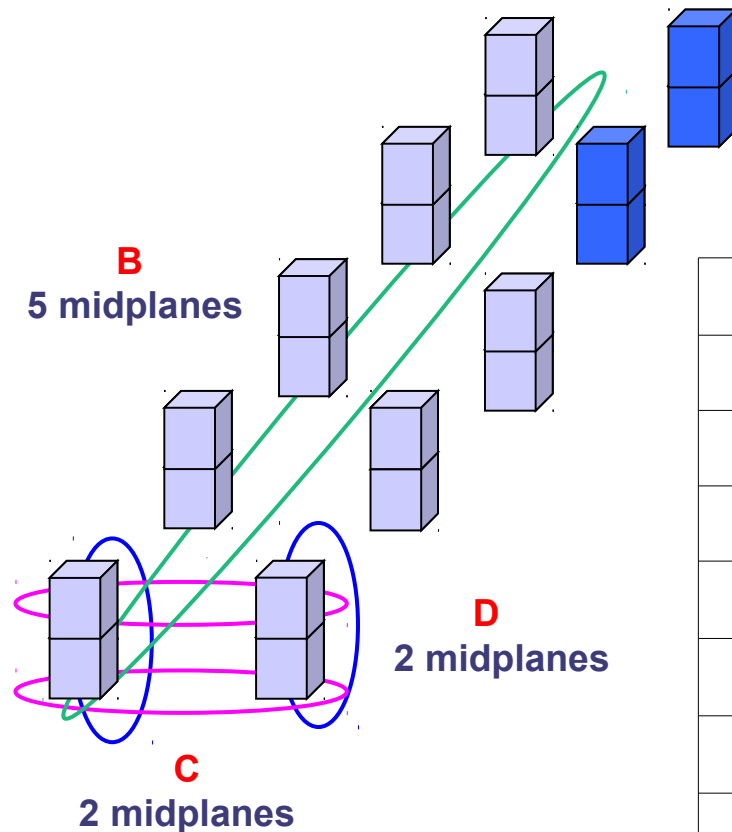
~~1 x 2 x 2 x 1~~

#@ bg_connectivity =
Torus|Mesh|Either|Xa Xb Xc Xd

- Can be used with **bg_size** and **bg_shape**
- Allow to specify the connectivity among the MPs in A, B, C and D dimensions.

Shape & Connectivity

– How to use –



| MP | bg_shape | bg_connectivity | |
|----|----------|-----------------|-----|
| 2 | 1x1x1x2 | Torus | yes |
| 2 | 1x1x2x1 | Torus | yes |
| 2 | 1x2x1x1 | Mesh | yes |
| 3 | 1x3x1x1 | Torus | no |
| 4 | 1x1x2x2 | Torus | yes |
| 4 | 1x2x1x2 | Torus | no |
| 4 | 1x2x2x1 | Mesh | yes |
| 4 | 1x4x1x1 | Torus | no |
| 5 | 1x5x1x1 | Torus | yes |
| 6 | 1x3x2x1 | Torus | no |
| 6 | 1x3x1x2 | Mesh | yes |

@ bg_shape = A x B x C x D

@ bg_connectivity = Torus

Midplanes must have Torus connectivity in all dimensions

bg_size

512

512 * 2

512 * 4

512 * 5 (special request)

512 * 10 (special request)

512 * 20 (special request)

bg_shape

1x1x1x1

1x1x1x2 and 1x1x2x1

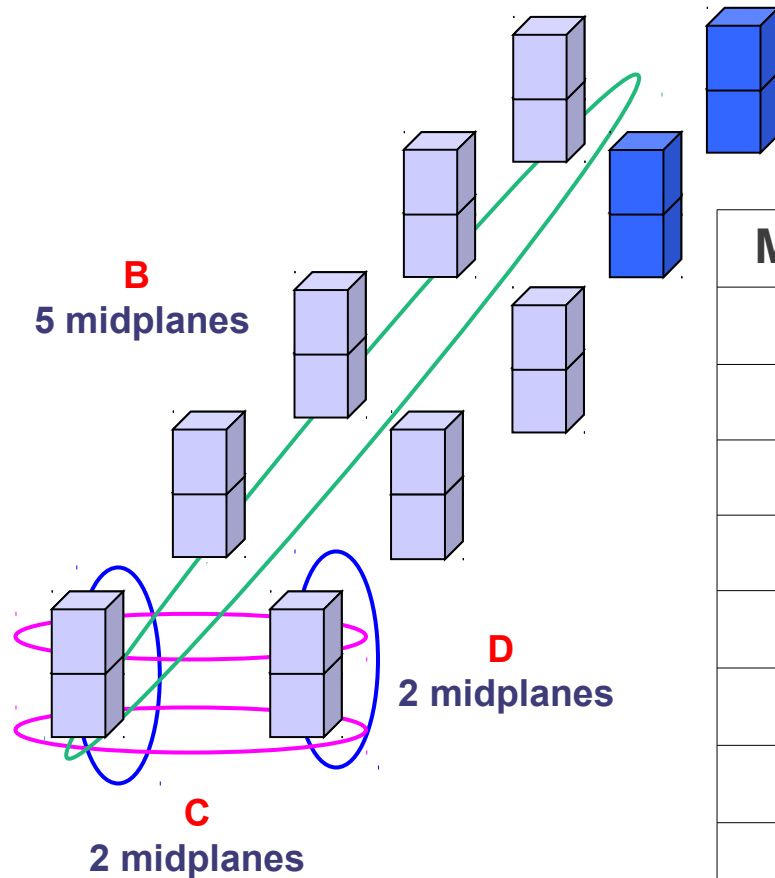
1x1x2x2

1x5x1x1

1x5x1x2 and 1x5x2x1

1x5x2x2

Shape & Connectivity - example



| MP | bg_shape | A | B | C | D |
|----|----------|-------|-------|-------|-------|
| 1 | 1x1x1x1 | Torus | Torus | Torus | Torus |
| 2 | 1x1x1x2 | Torus | Torus | Torus | Torus |
| 2 | 1x1x2x1 | Torus | Torus | Torus | Torus |
| 2 | 1x2x1x1 | Torus | Mesh | Torus | Torus |
| 3 | 1x3x1x1 | Torus | Mesh | Torus | Torus |
| 4 | 1x1x2x2 | Torus | Torus | Torus | Torus |
| 4 | 1x2x1x2 | Torus | Mesh | Torus | Torus |
| 4 | 1x2x2x1 | Torus | Mesh | Torus | Torus |
| 4 | 1x4x1x1 | Torus | Mesh | Torus | Torus |
| 5 | 1x5x1x1 | Torus | Torus | Torus | Torus |
| 6 | 1x3x2x1 | Torus | Mesh | Torus | Torus |
| 6 | 1x3x1x2 | Torus | Mesh | Torus | Torus |

Summary for Large jobs (>1MP)

#@ bg_size = n° of compute nodes

#@ bg_connectivity = Mesh (default)

#@ bg_connectivity=
Torus|Either|Xa Xb Xc Xd

#@ bg_shape = A x B x C x D

#@ bg_rotate = true (default)
#@ bg_connectivity = Mesh (default)

#@ bg_rotate = false
#@ bg_connectivity = Mesh (default)

#@ bg_rotate = true (default)
#@ bg_connectivity = Torus|Either|Xa Xb Xc Xd

#@ bg_rotate = false
#@ bg_connectivity = Torus|Either|Xa Xb Xc Xd



Example

4 midplanes

#@bg_size = 2 048

#@bg_connectivity = Mesh

1 x 2 x 2 x 1

1 x 2 x 1 x 2

1 x 1 x 2 x 2

1 x 4 x 1 x 1

#@bg_size = 2 048

#@bg_connectivity = Torus

1 x 1 x 2 x 2



4 midplanes

```
#@bg_shape = 1x2x2x1  
#@bg_connectivity = Mesh  
#@bg_rotate = true
```

1 x 2 x 2 x 1

1 x 2 x 1 x 2

1 x 1 x 2 x 2

```
#@bg_shape = 1x2x2x1  
#@bg_connectivity = Mesh  
#@bg_rotate = false
```

1 x 2 x 2 x 1



Example

4 midplanes

```
#@bg_shape = 1x2x2x1  
#@bg_connectivity = Mesh  
#@bg_rotate = false
```

1 x 2 x 2 x 1

```
#@bg_shape = 1x2x2x1  
#@bg_connectivity = Torus  
#@bg_rotate = false
```

NEVER EXECUTED



Mapping



FERMI Configuration

N° of MPs 20

N° of Compute Nodes 10240

FERMI Size in Midplanes 1x5x2x2

Compute Nodes in a Midplane 4x4x4x4x2

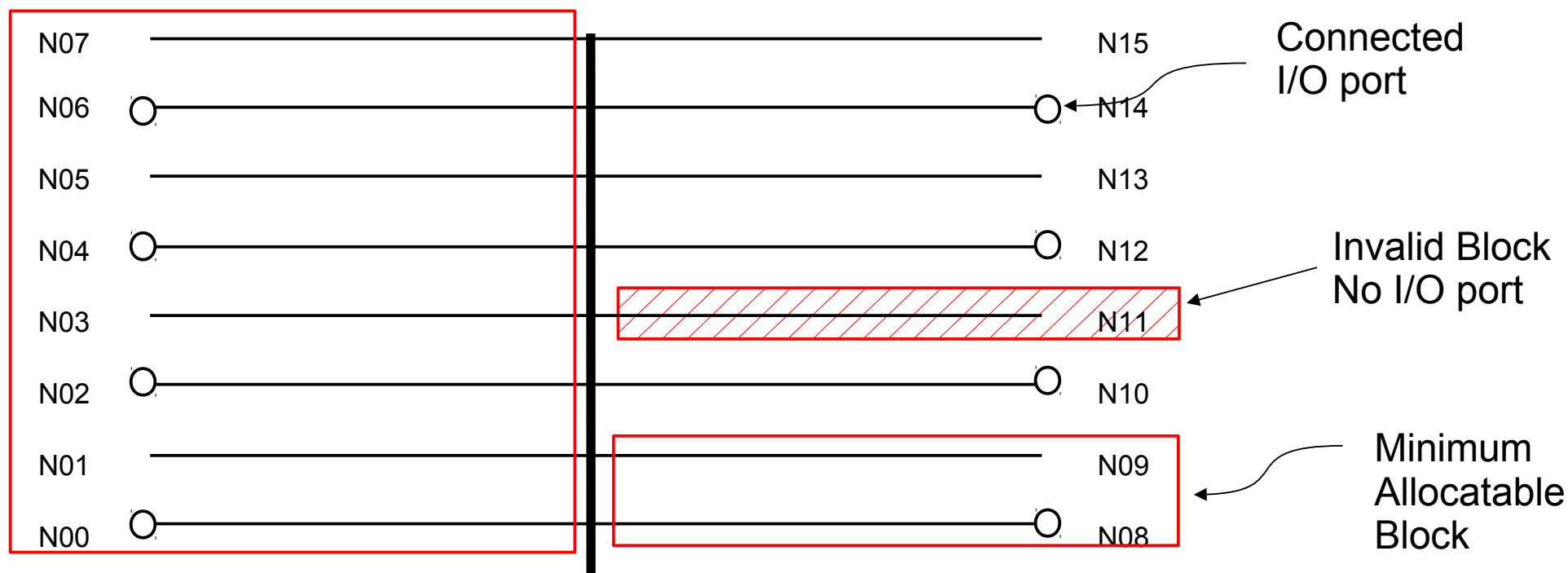
FERMI Size in Compute Nodes 4x20x8x8x2

E dimension.
It is always 2



Compute partizion – Small block

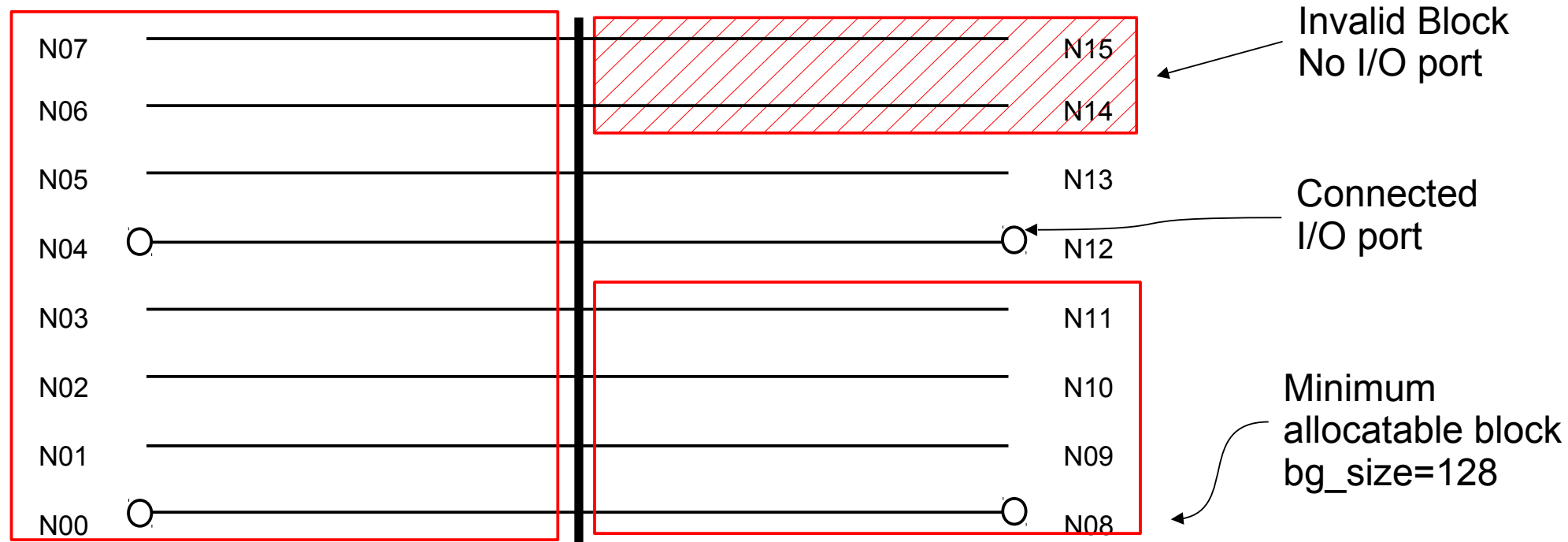
| # Node Cards | # Nodes | Torus Size, in Nodes | IsTorus (ABCDE) |
|--------------------|---------|----------------------|-----------------|
| 1 | 32 | 2x2x2x2x2 | 00001 |
| 2 (adjacent pairs) | 64 | 2x2x4x2x2 | 00101 |
| 4 (quadrants) | 128 | 2x2x4x4x2 | 00111 |
| 8 (halves) | 256 | 4x2x4x4x2 | 10111 |



Example:

N08 – N09 = 64 Compute Cards (2x2x4x2x2)

MidPlane in FERMI / {R11 R31}

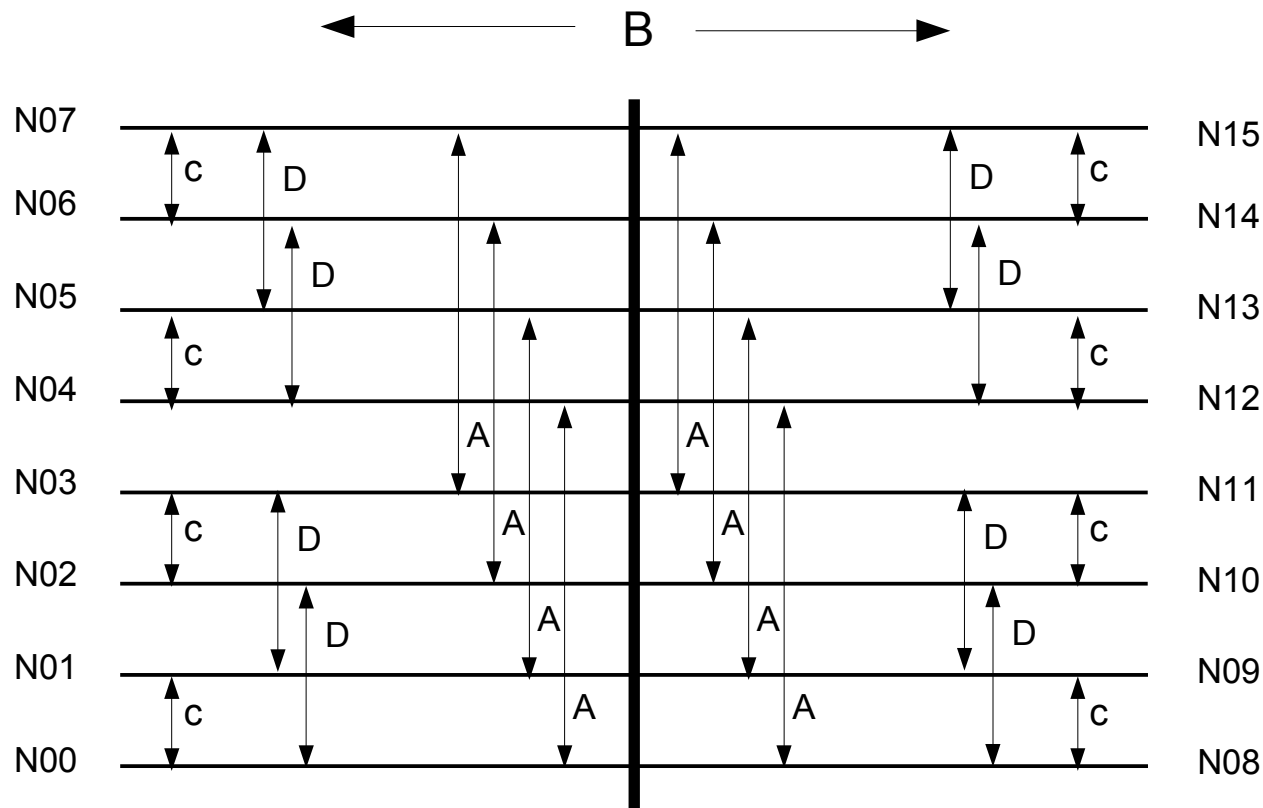


Example:

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

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

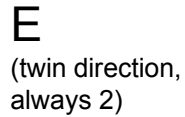
Arrows show how dimensions

A,B,C,D span across nodeboards

Dimension E does not extend across nodeboards

The nodeboards combine to form a 4x4x4x2 torus

Note that nodeboards are paired in dimensions A,B,C and D as indicated by the arrows



Example: 64 nodes, 64 task MPI, 1 rank/node, mapping ABCDET

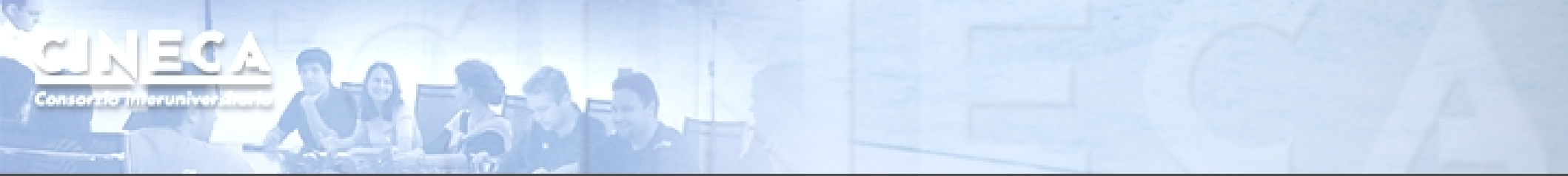
```
runjob --np 64 --ranks-per-node 1 --mapping ABCDET --exe  
mpi_procs_name.exe | sort
```

```
block shape: 2x2x4x2x2; torus: 00101
```

| | | | | | |
|------|---|----|----|---------------|----------------|
| Task | 0 | of | 64 | (0,0,0,0,0,0) | R11-M1-N14-J21 |
| Task | 1 | of | 64 | (0,0,0,0,1,0) | R11-M1-N14-J18 |
| Task | 2 | of | 64 | (0,0,0,1,0,0) | R11-M1-N14-J25 |
| Task | 3 | of | 64 | (0,0,0,1,1,0) | R11-M1-N14-J30 |
| Task | 4 | of | 64 | (0,0,1,0,0,0) | R11-M1-N14-J20 |
| Task | 5 | of | 64 | (0,0,1,0,1,0) | R11-M1-N14-J19 |
| Task | 6 | of | 64 | (0,0,1,1,0,0) | R11-M1-N14-J24 |
| Task | 7 | of | 64 | (0,0,1,1,1,0) | R11-M1-N14-J31 |

... ..

| | | | | | |
|------|----|----|----|---------------|----------------|
| Task | 58 | of | 64 | (1,1,2,1,0,0) | R11-M1-N15-J06 |
| Task | 59 | of | 64 | (1,1,2,1,1,0) | R11-M1-N15-J01 |
| Task | 60 | of | 64 | (1,1,3,0,0,0) | R11-M1-N15-J11 |
| Task | 61 | of | 64 | (1,1,3,0,1,0) | R11-M1-N15-J12 |
| Task | 62 | of | 64 | (1,1,3,1,0,0) | R11-M1-N15-J07 |
| Task | 63 | of | 64 | (1,1,3,1,1,0) | R11-M1-N15-J00 |



THANKS FOR ATTENTION !!!

QUESTIONS ???