MPI-HW-Topo WG

FUJITSU

shaping tomorrow with you

HW-Topology Aware MPI Extension on Fujitsu PRIMEHPC FX10, FX100

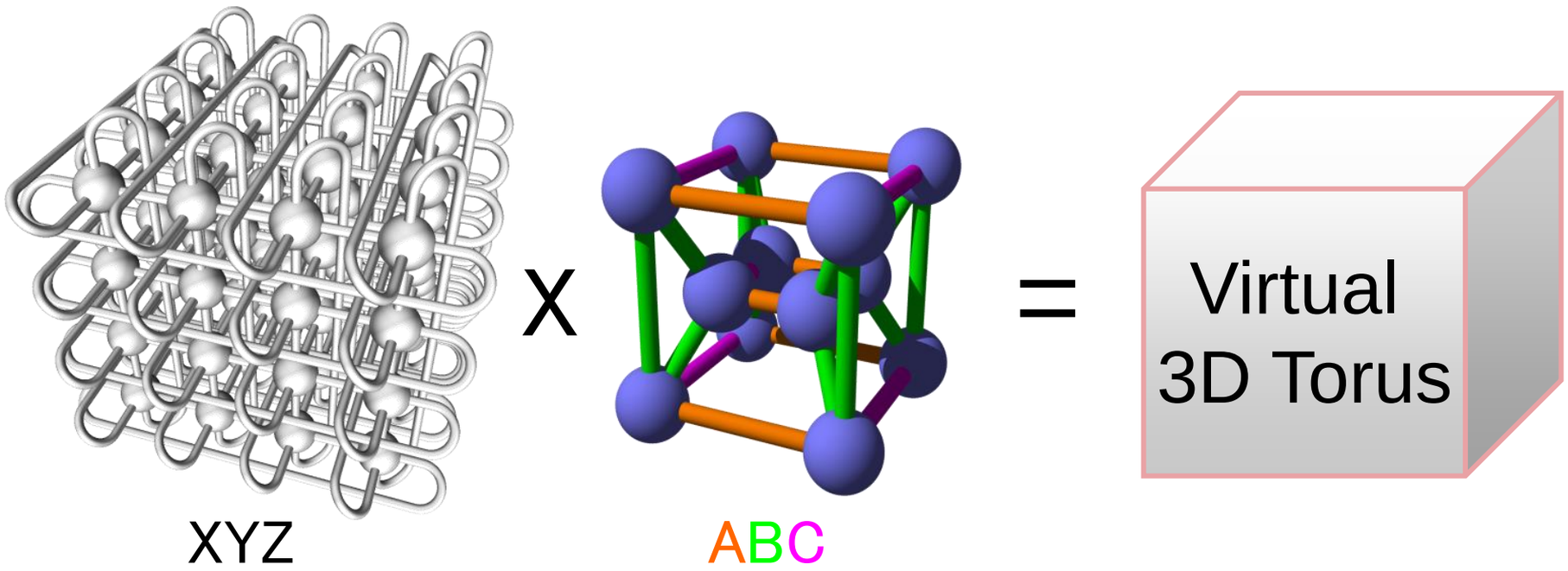
2018/04/12

Shinji Sumimoto,
Fujitsu

Network Architecture of the K Computer

FUJITSU

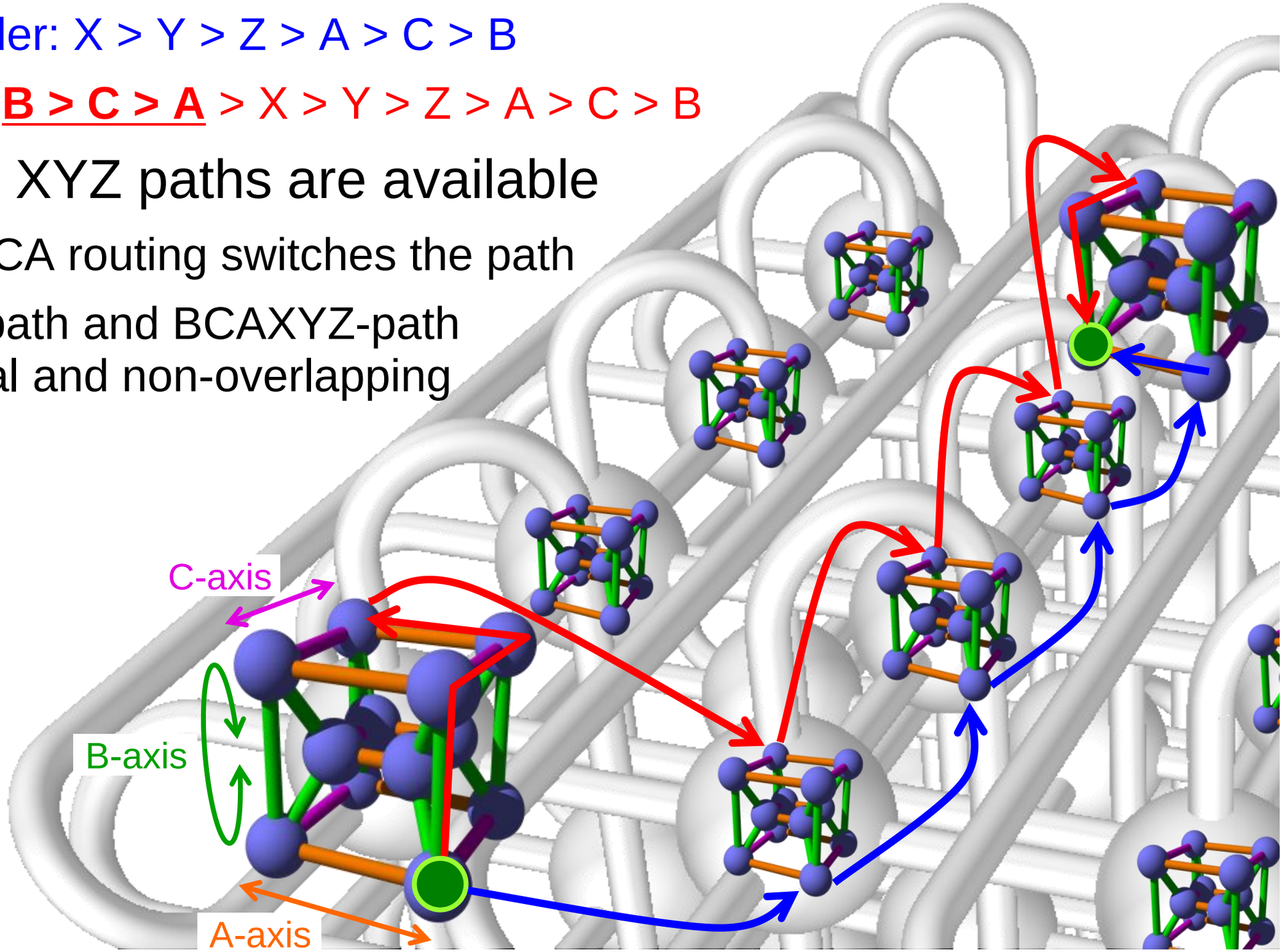
- 6D mesh/torus (24x18x16x2x3x2)
  - direct network (i.e. no switches)
- Providing 3D torus view to the users
  - Three extra axes are used to construct torus
    - even when a part of the K computer is used
    - even when some nodes are under maintenance



Routing Algorithm

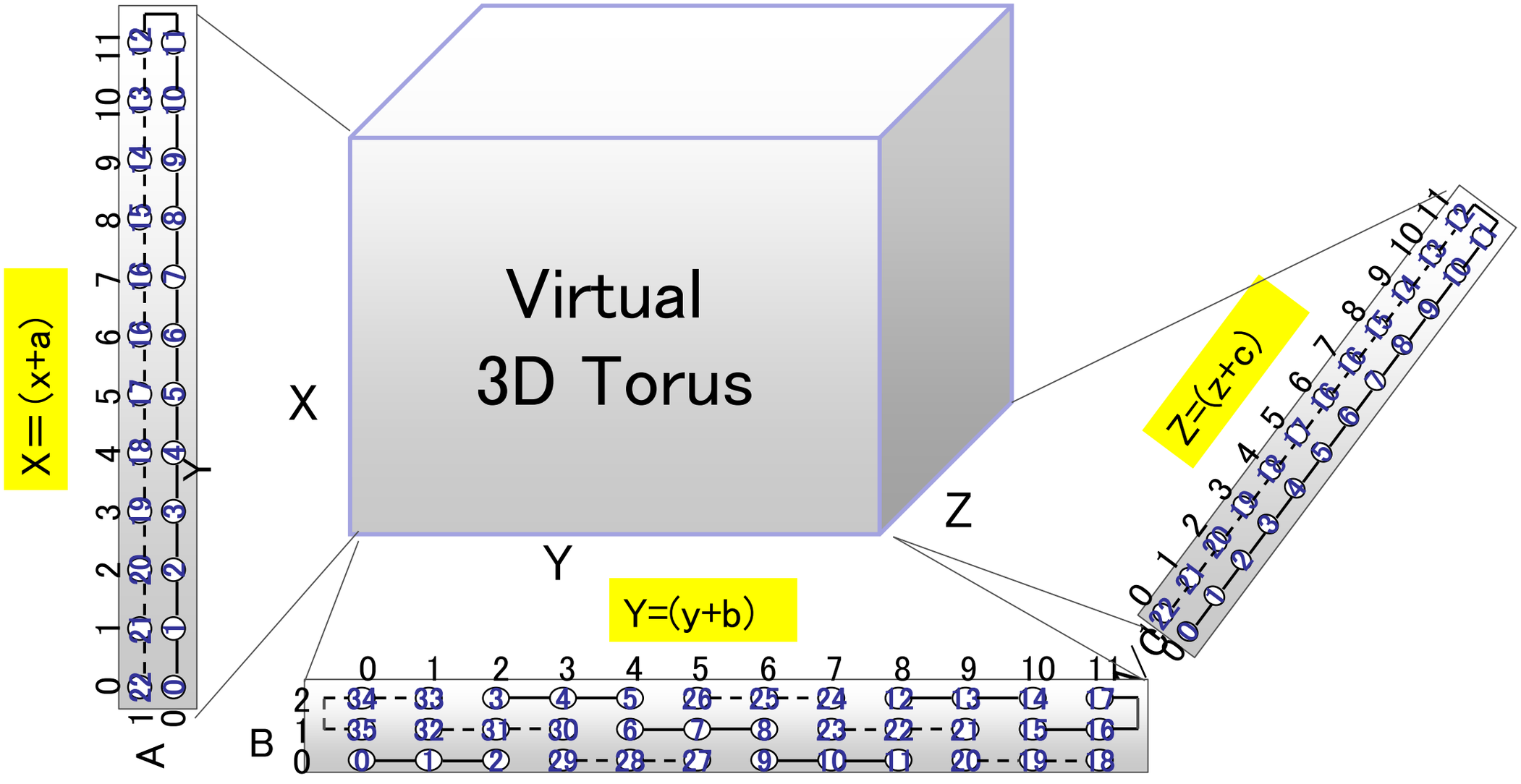
FUJITSU

- Extended dimension-order algorithm
  - Default order: X > Y > Z > A > C > B
  - Extended: **B > C > A** > X > Y > Z > A > C > B
- 3x2x2 = 12 XYZ paths are available
  - The first BCA routing switches the path
  - XYZACB-path and BCAXYZ-path are minimal and non-overlapping



Job Allocation and Rank Mapping on Tofu

- User can specify 1,2,3D network on job submission.
- Job scheduler makes torus network by combination of XYZ+ABC
  - User can specify the combination
- Basic Combination of 3D Torus Mapping: X=x+a, Y=y+b, Z=z+c



Fujitsu's Extension for Topology Aware MPI Applications

FUJITSU

- Providing Topology Aware MPI Application Environment with Runtime System and MPI Libraries in Fujitsu's Technical Computing Suite (TCS) system software

- The Runtime System:
  - User can specify a shape of 1D, 2D and 3D of torus shape using pjsub option
  - Ex: # pjsub -L node=4x4x4 job.sh
    # pjsub -L node=8x8 job.sh
- The Fujitsu MPI:
  - Programmer can get a shape and network topological information and program using Fuijtsu's MPI extension functions.

The Fujitsu MPI: Extended Functions

FUJITSU

Table 5.1 Rank query interface function list

Function name	Function overview
FJMPI_Topology_get_dimension	Gets the number of dimensions given to MPI_COMM_WORLD
FJMPI_Topology_get_shape	Gets the process shape given to MPI_COMM_WORLD
FJMPI_Topology_rank2x	Gets the X coordinate value from the rank number
FJMPI_Topology_rank2xy	Gets the XY coordinate value from the rank number
FJMPI_Topology_rank2xyz	Gets the XYZ coordinate value from the rank number
FJMPI_Topology_x2rank	Gets the rank number from the X coordinate value
FJMPI_Topology_xy2rank	Gets the rank number from the XY coordinate value
FJMPI_Topology_xyz2rank	Gets the rank number from the XYZ coordinate value
FJMPI_Topology_cart_reorder	Gets the value that determines the rank of a communicator with a Cartesian structure
FJMPI_Topology_sys_rank2xyzabc	Gets the Tofu coordinates from the rank number
FJMPI_Topology_sys_xyzabc2rank	Gets the rank number from the Tofu coordinates
FJMPI_Topology_rel_rank2xyzabc	Gets the relative Tofu coordinates from the rank number
FJMPI_Topology_rel_xyzabc2rank	Gets the rank number from the relative Tofu coordinates

From MPI User's Guide(PRIMEHPC FX100)

FUJITSU

shaping tomorrow with you