shaping tomorrow with you

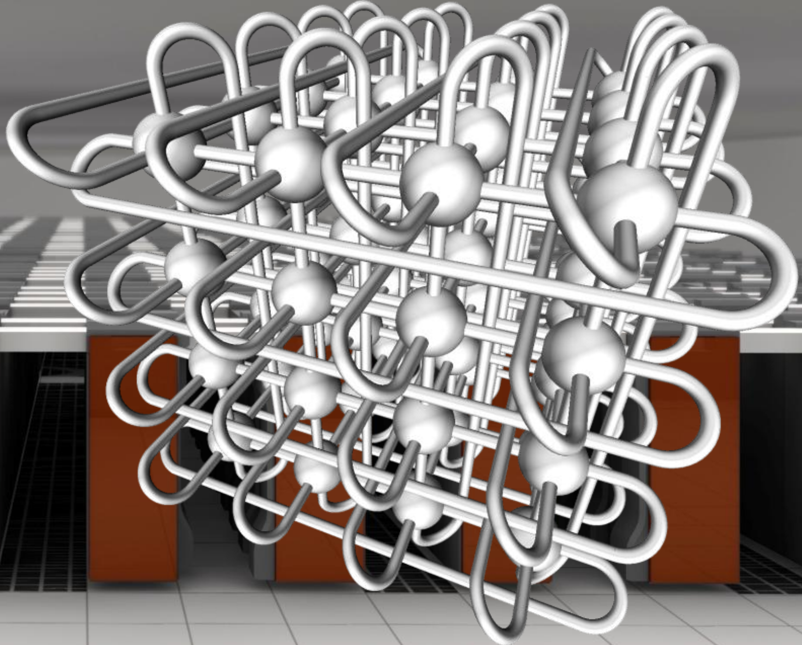
Tofu: A 6D Mesh/Torus Interconnect

Next Generation Technical Computing Unit

Fujitsu Limited

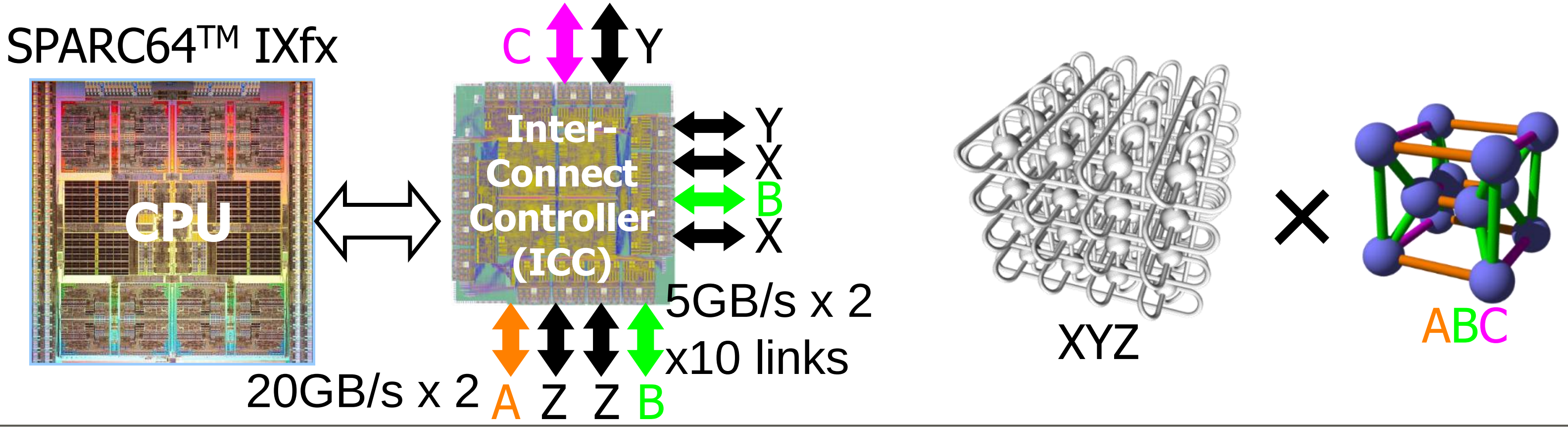
Tofu: A 6D mesh/torus interconnect

- High communication performance
- High system scalability
- High fault-tolerance



Tofu interconnect

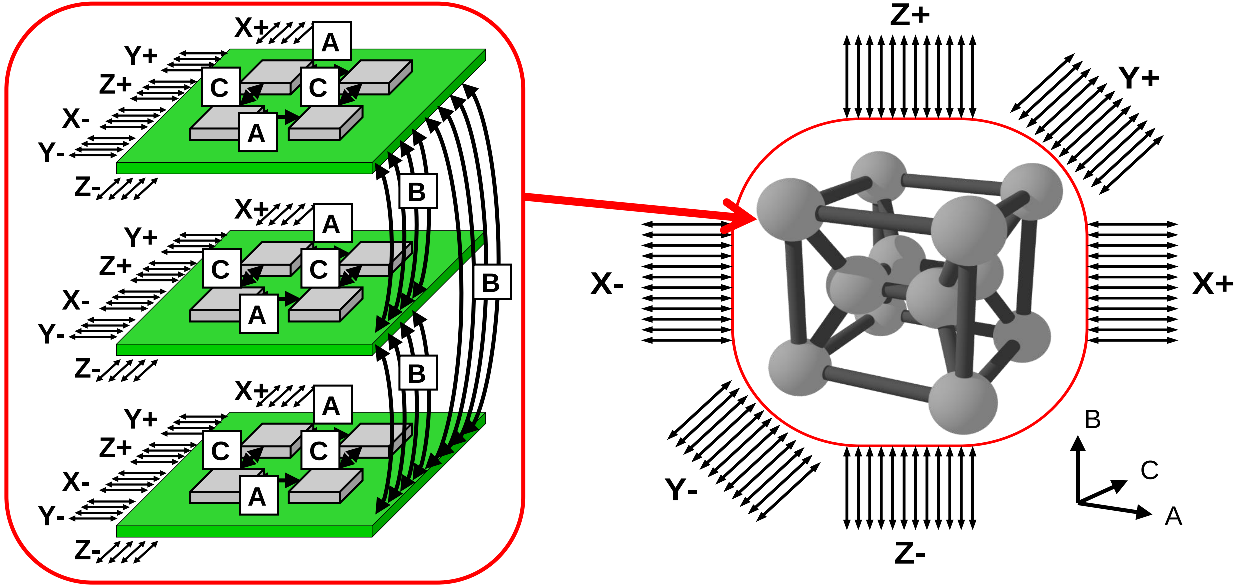
- Highly scalable and usable direct network (6D mesh/torus)
  - 10 redundant high BW links, 4 RDMA engines (4x2 simultaneous transfer)
  - Good collective communication performance with Tofu original algorithms
- Tofu barrier for barrier & reduction in H/W
- Direct attached interconnect controller



Tofu realizes scalable systems beyond 100,000 nodes
With low power consumption, low latency, and high BW

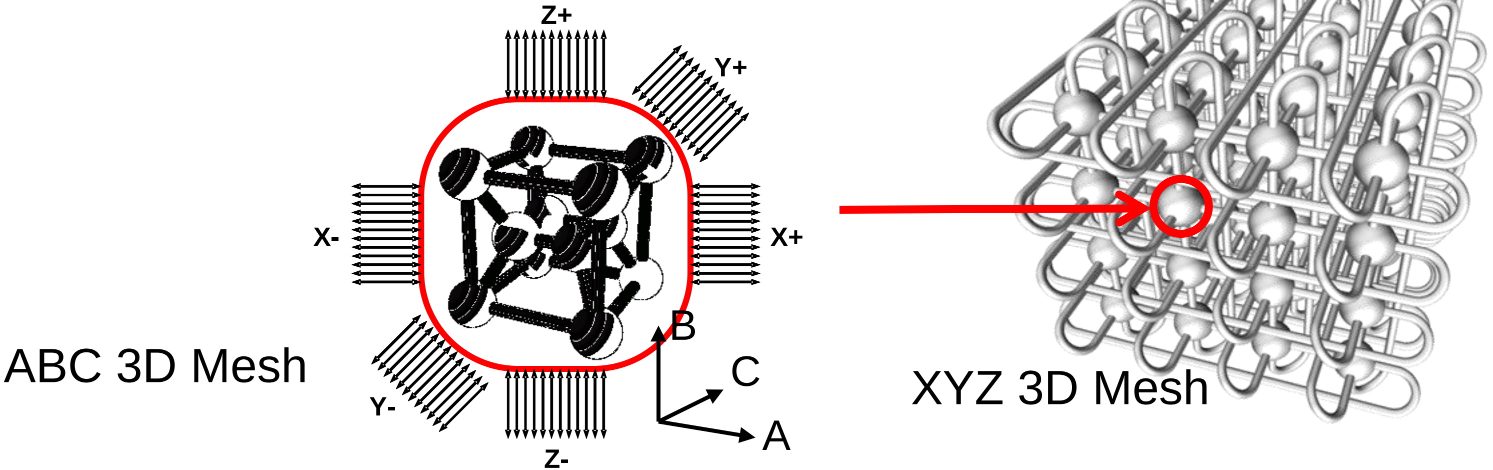
Node Group

- A node group is composed of 12 compute nodes.
- A- and C-Axis connect 4 compute node on a compute board.
- B-Axis connects 3 compute boards.



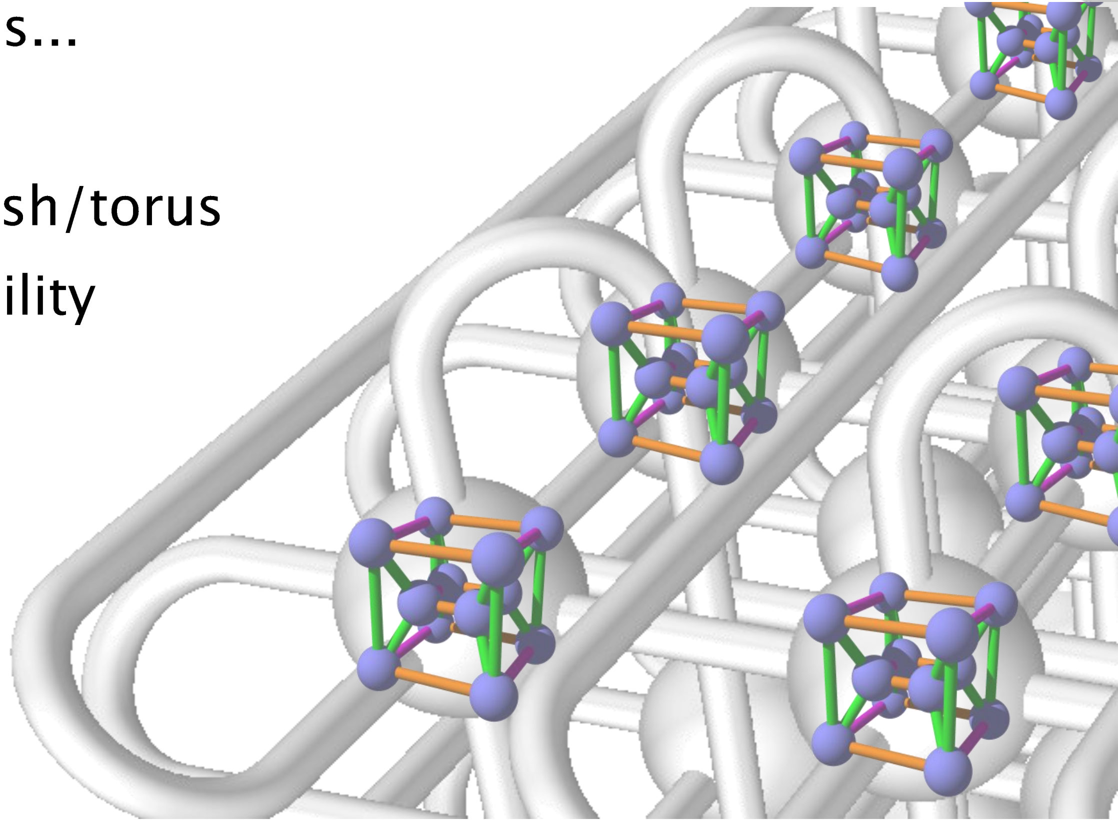
6D Mesh Topology

- All nodes have an address with six parameters (X,Y,Z,A,B,C).
- Total 6D Mesh is composed of ABC 3D Meshes and the XYZ 3D Mesh.
- ABC 3D Mesh
  - An ABC 3D Mesh connect 12 compute nodes.
- XYZ 3D Mesh
  - The XYZ 3D Mesh connects ABC 3D Mesh groups.



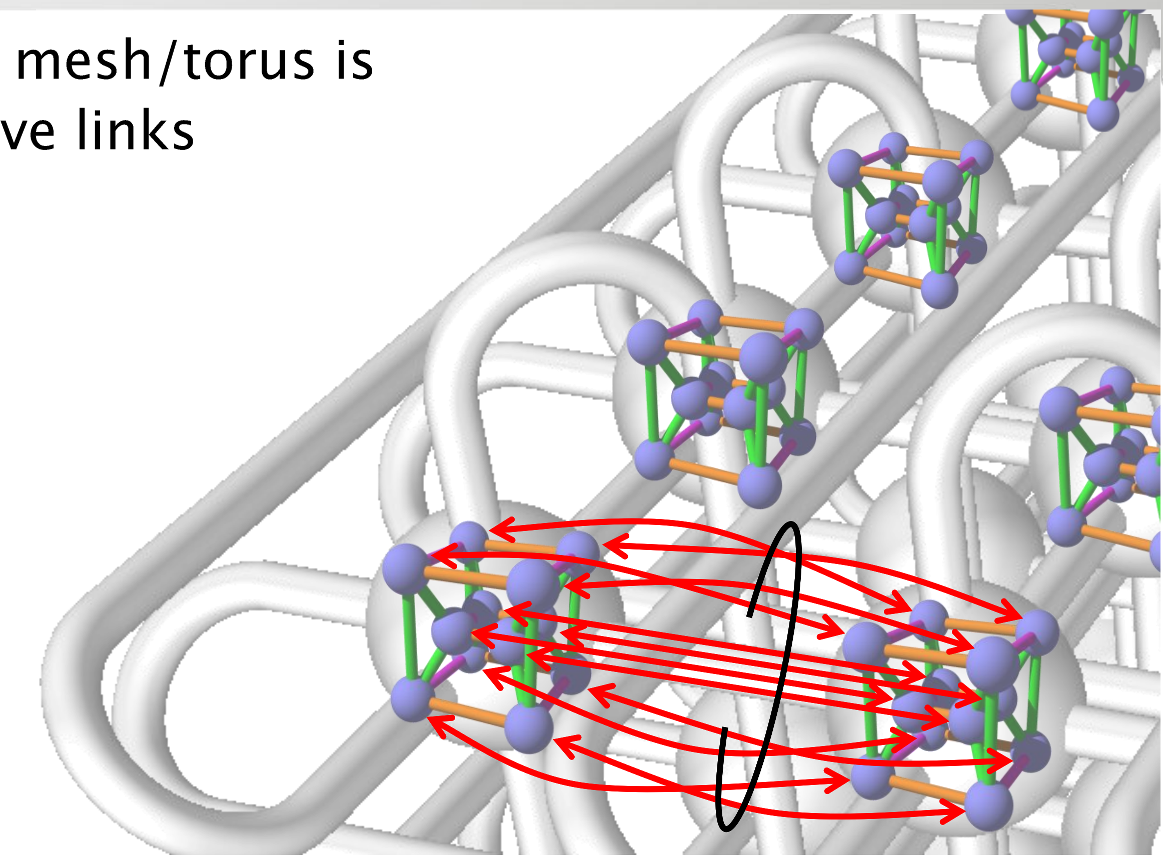
Network construction

- From the other perspectives…
  - Overlaid twelve *xyz* torus
  - X x Y x Z array of *abc* mesh/torus
- Twelve times higher scalability than the 3D torus network



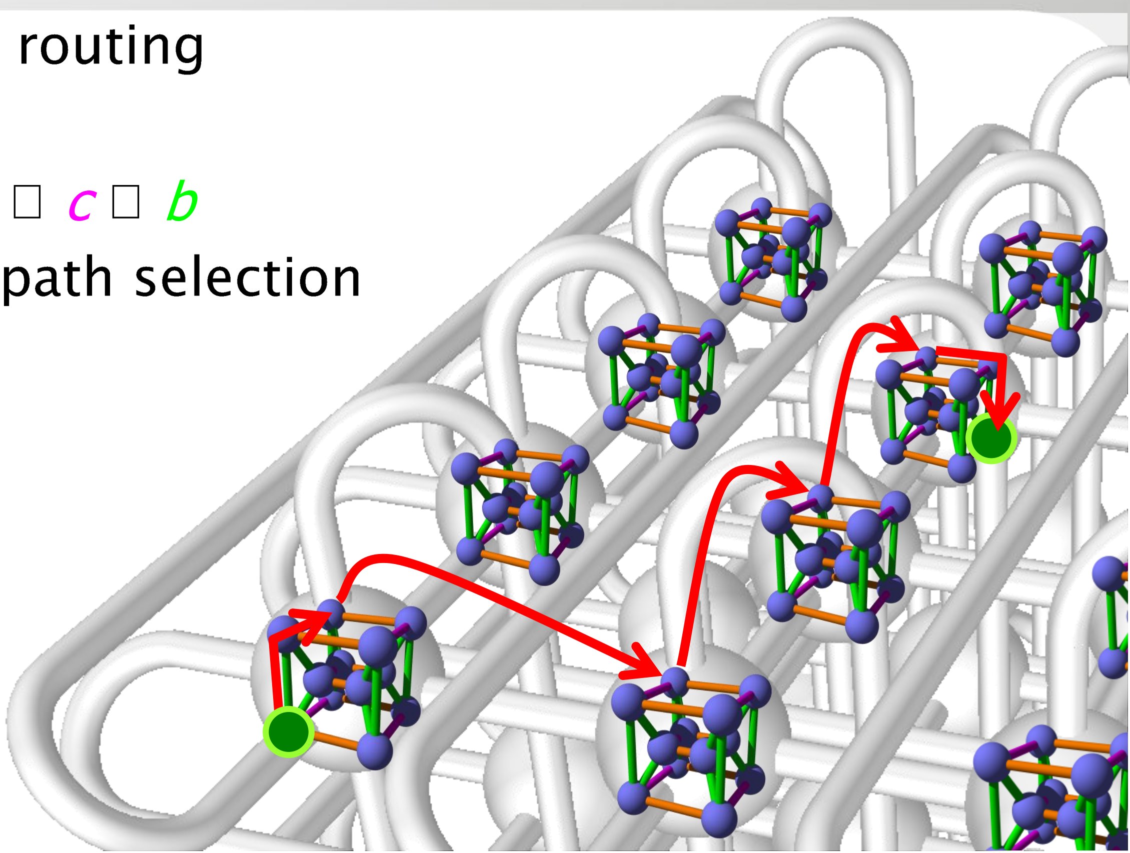
Network construction cont.

- Each pair of adjacent *abc* mesh/torus is interconnected with twelve links



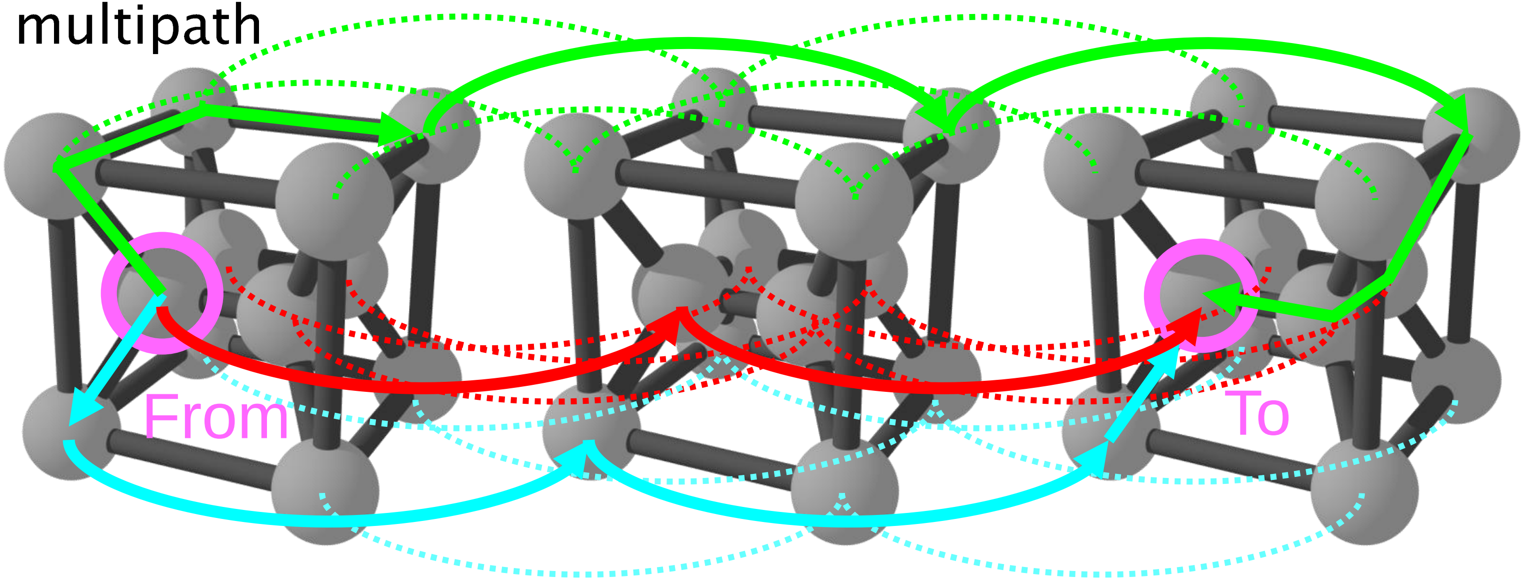
Routing algorithm

- Extended dimension order routing
  - Additional *abc* traversal
  - *b*→*c* □*a* □ *x*□ *y*□ *z*⇒ *a* □ *c*□ *b*
  - The first *abc* traversal is path selection



Multiple Paths

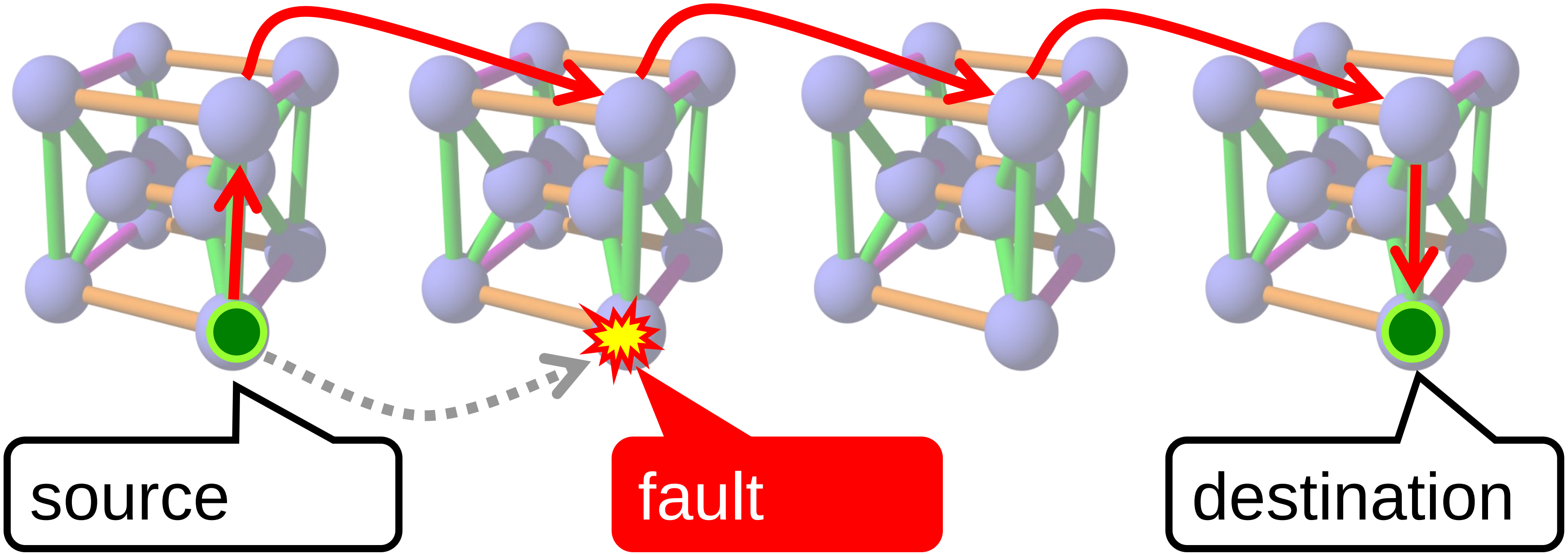
- The proactive routing algorithm allows 12 routing paths.
- Detouring faulty nodes
- Trunking multipath



3 example paths out of 12 possible paths

Detouring faulty nodes

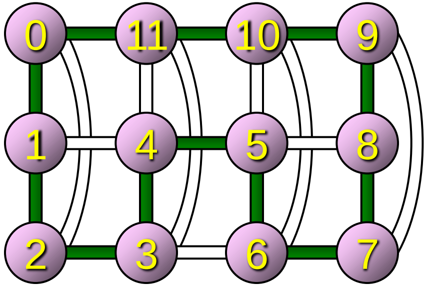
- Multipath routing allows to detour faulty nodes



Application Torus

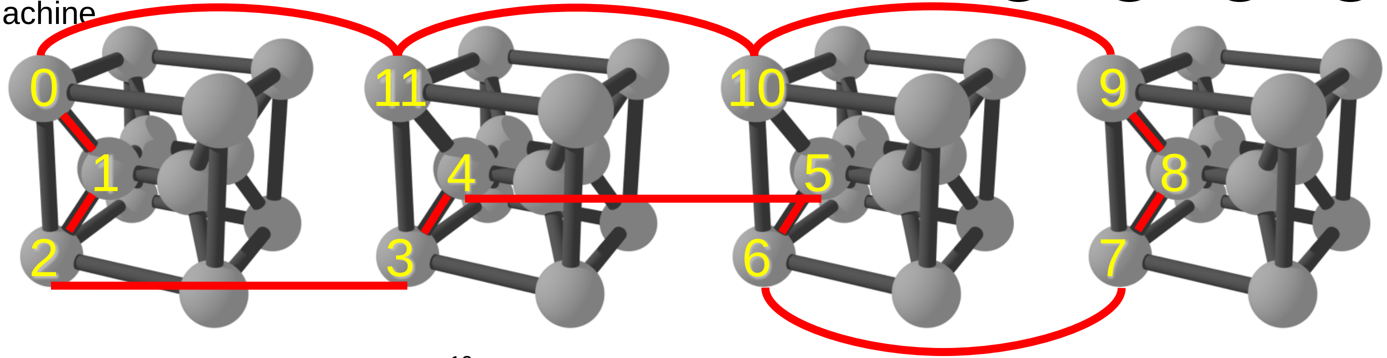
- An application torus is allocated to each job.
- An application torus is physically a 6D submesh of a machine.
- 1D to 3D application tori are supported.
- One dimension of an application torus is rendered by folding together several machine dimensions

2 dimensional slice view



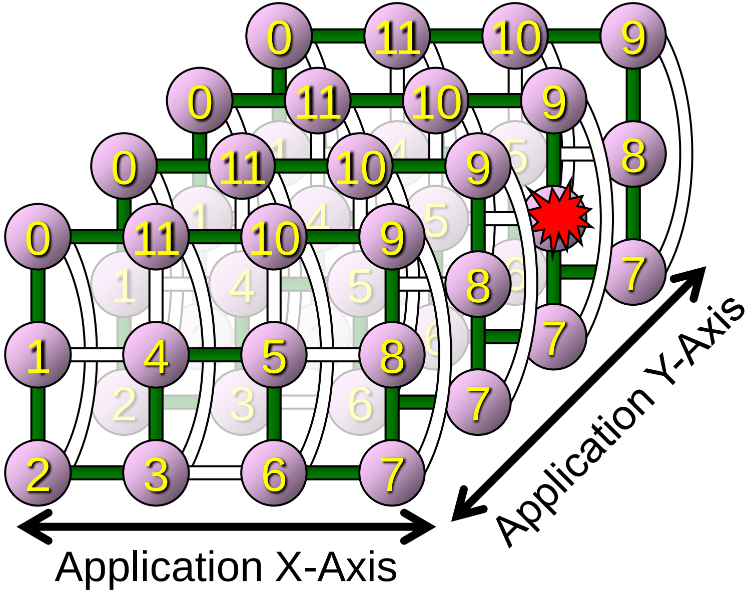
Example)
One application dimension rendered on two dimensional slice of a machine

4 dimensional slice view

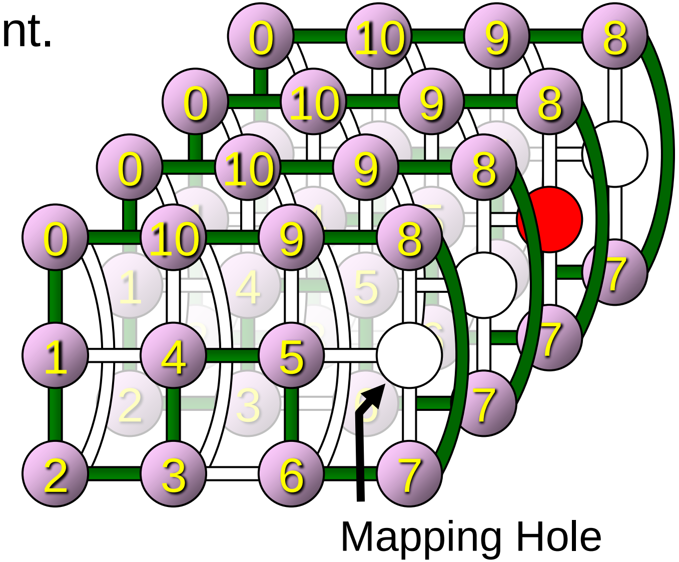


Graceful Degradation

- The job management system may run a job on a 6D machine submesh with a faulty node.

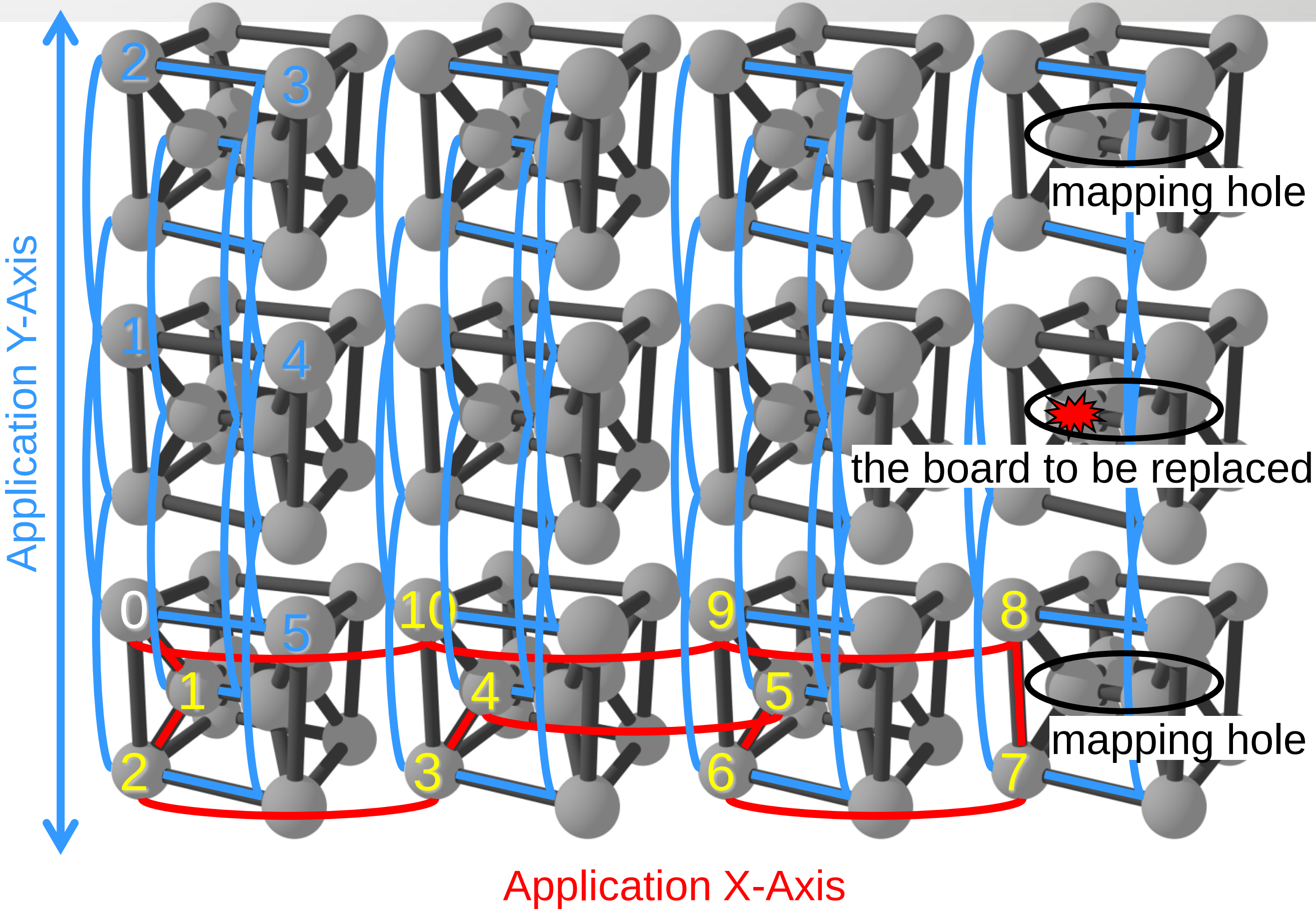


When a node failure occurs, the running job is force quitted and restarted from the user's checkpoint.



The 6D submesh can be reused. One of the app-dimensions is degraded by one hop.

Hot-swappable Compute Board



FUJITSU

shaping tomorrow with you