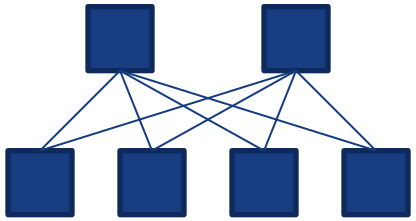
3D Torus for InfiniBand

HPC@mellanox.com

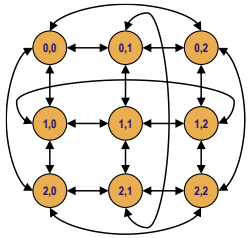
InfiniBand Supports Multiple Network Topologies

- Full support for Fat-tree (CLOS), Mesh, 3D-Torus topologies

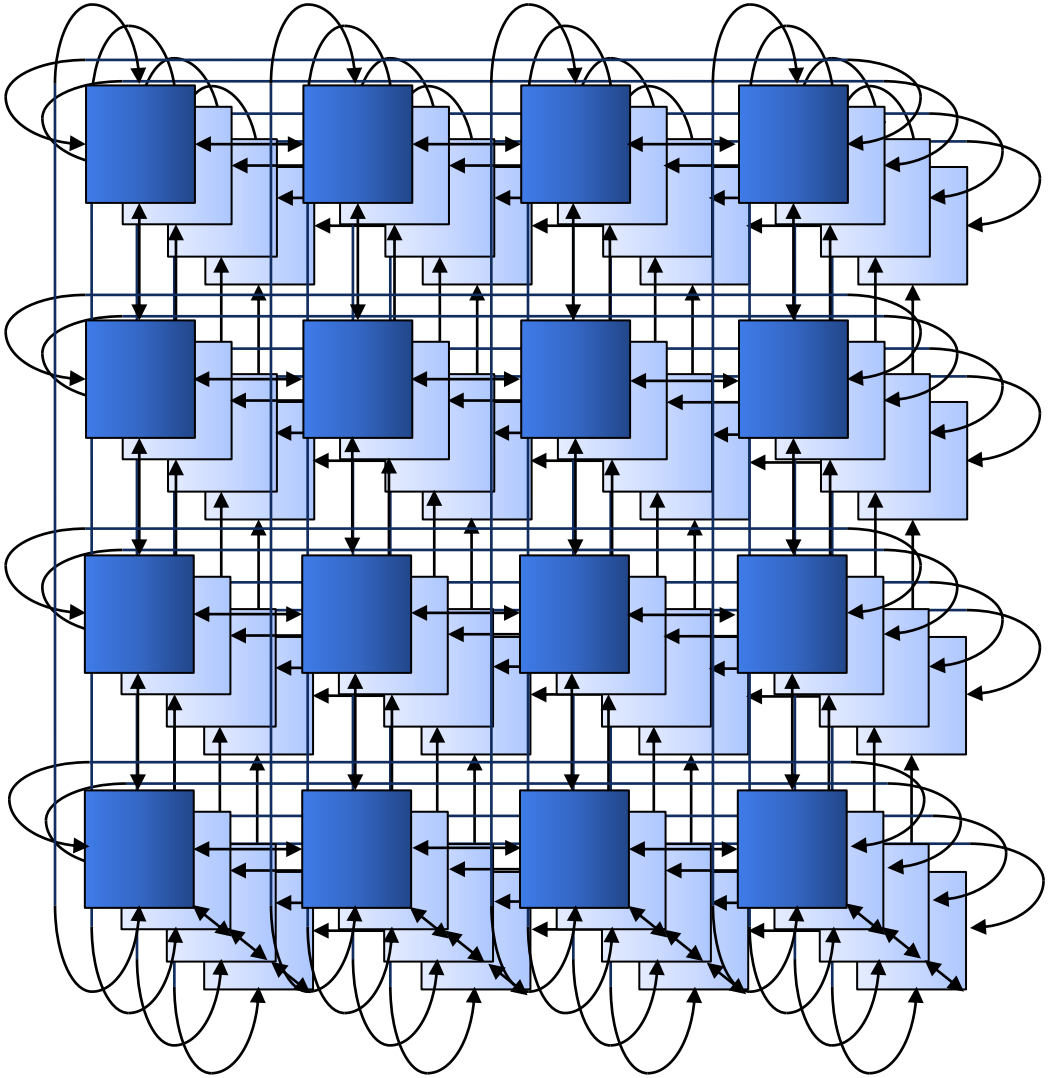
- CLOS (fat-tree)
  - Can be fully non-blocking (1:1) or blocking (x:1)
  - Typically enables best performance
    - Non blocking bandwidth, lowest network latency



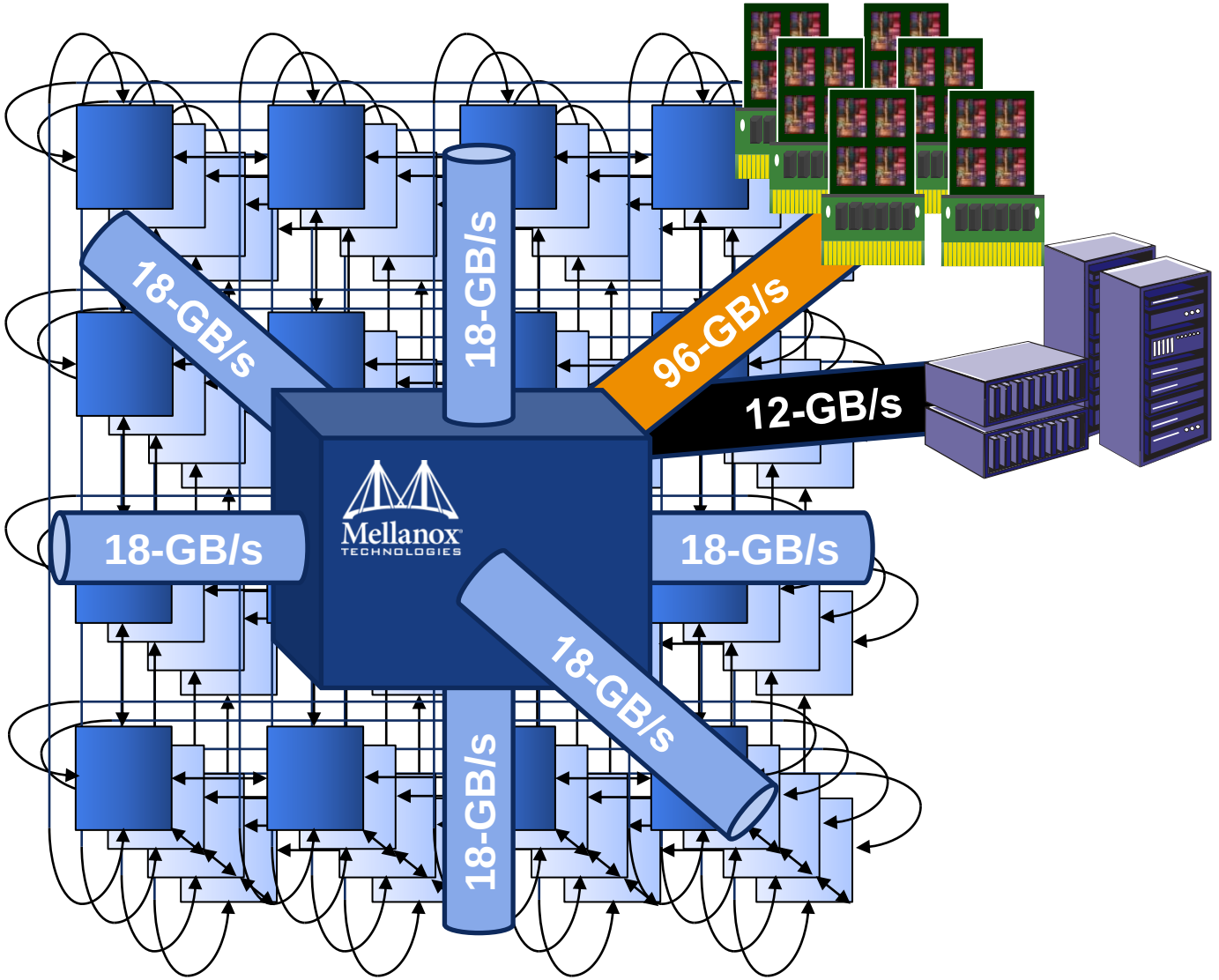
- Mesh or 3D Torus
  - Blocking network, cost-effective for systems at scale
  - Great performance solutions for applications with locality
  - Support for dedicate sub-networks
  - Simple expansion for future growth



What is a 3D-Torus Topology?

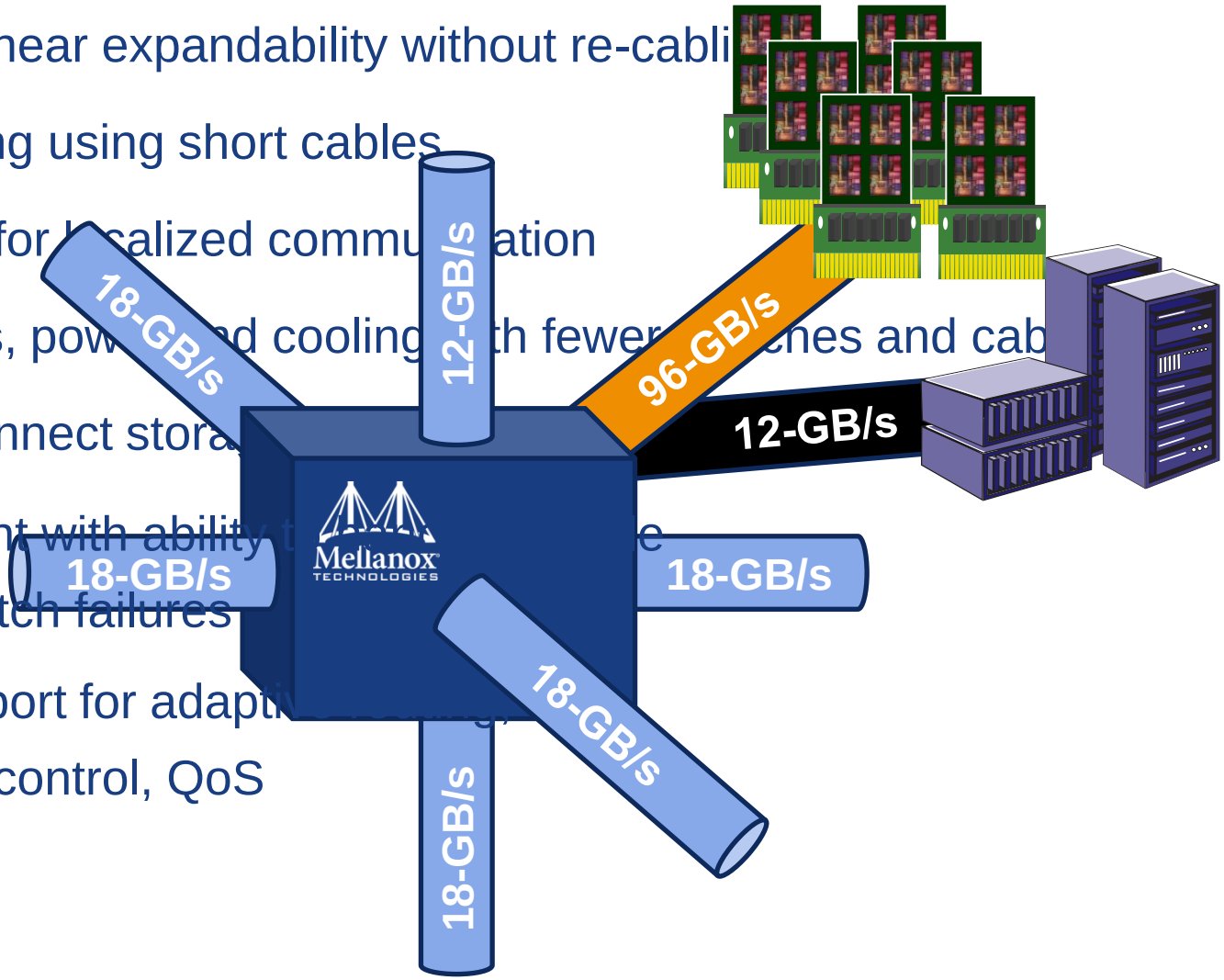


What is a 3D-Torus Topology?



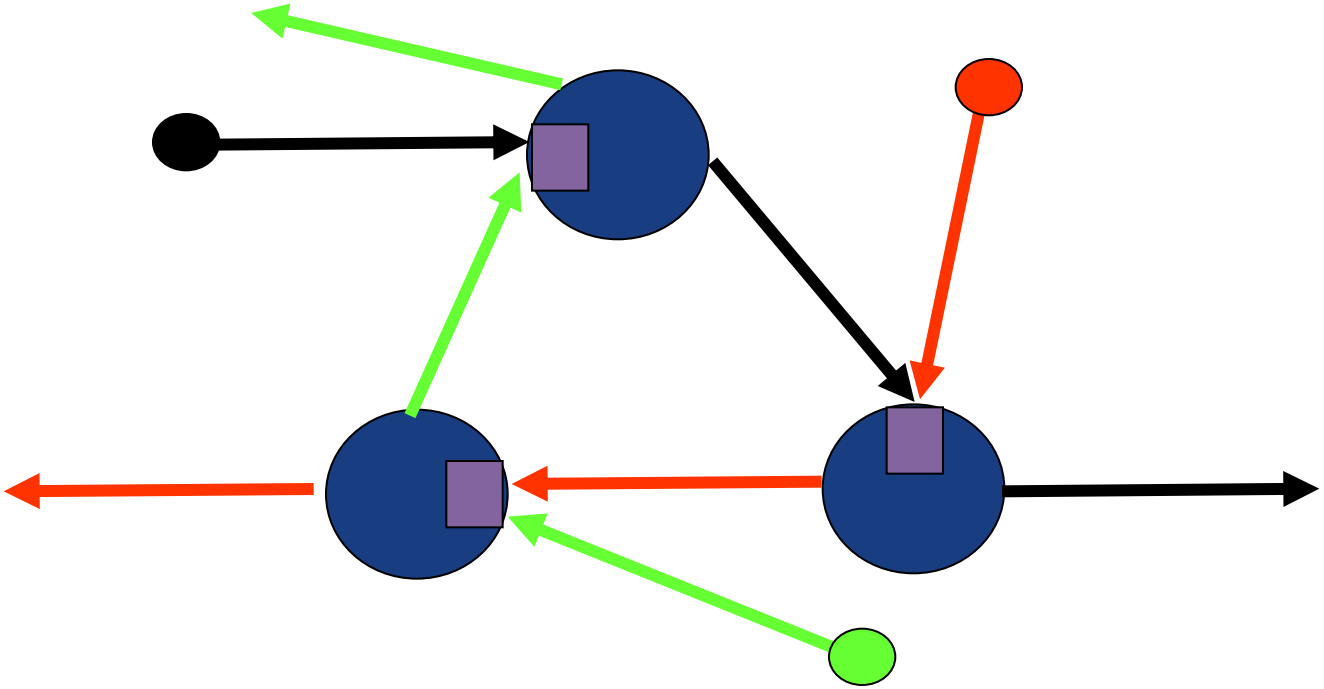
Benefits of 3D Torus

- Allows for linear expandability without re-cabli
- Simple wiring using short cables
- Works well for l calized commu ation
- Lower costs, pow d cooling th fewer ches and cab
- Ability to connect stora
- Fault tolerant with ability t e link and switch failures
- Built in support for adapti congestion control, QoS

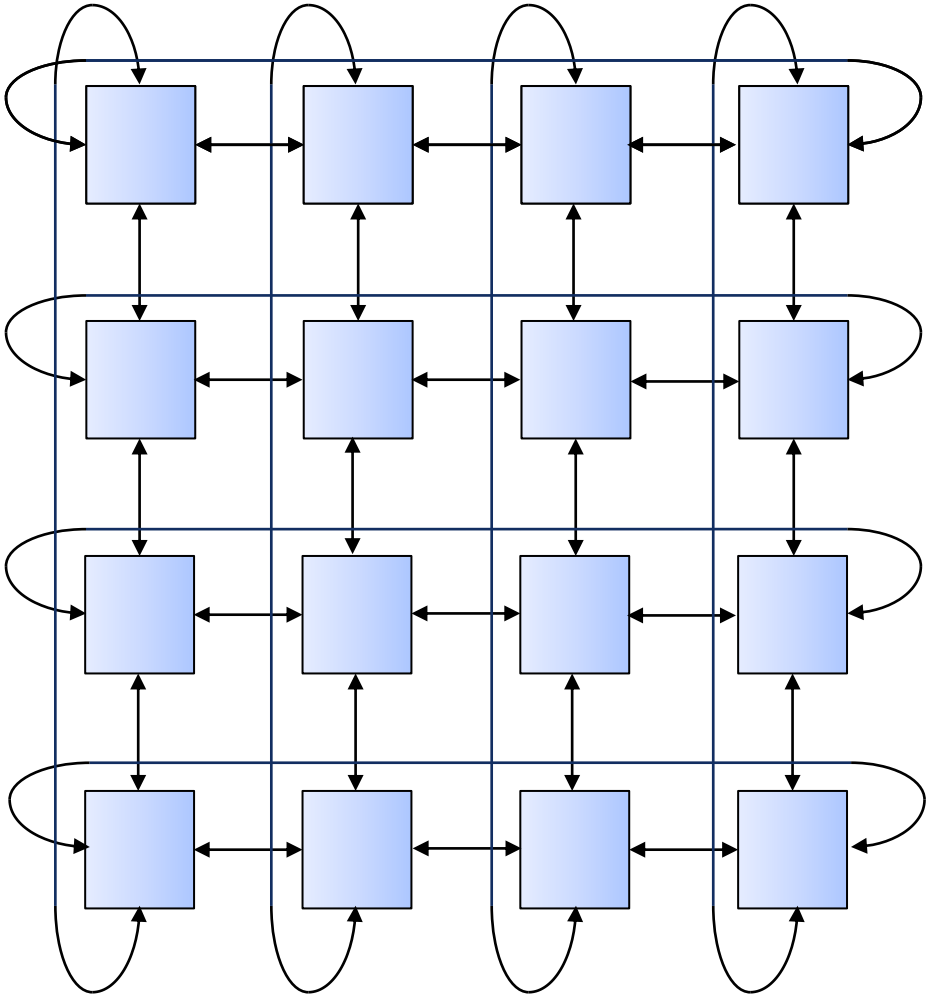


Loop Topologies

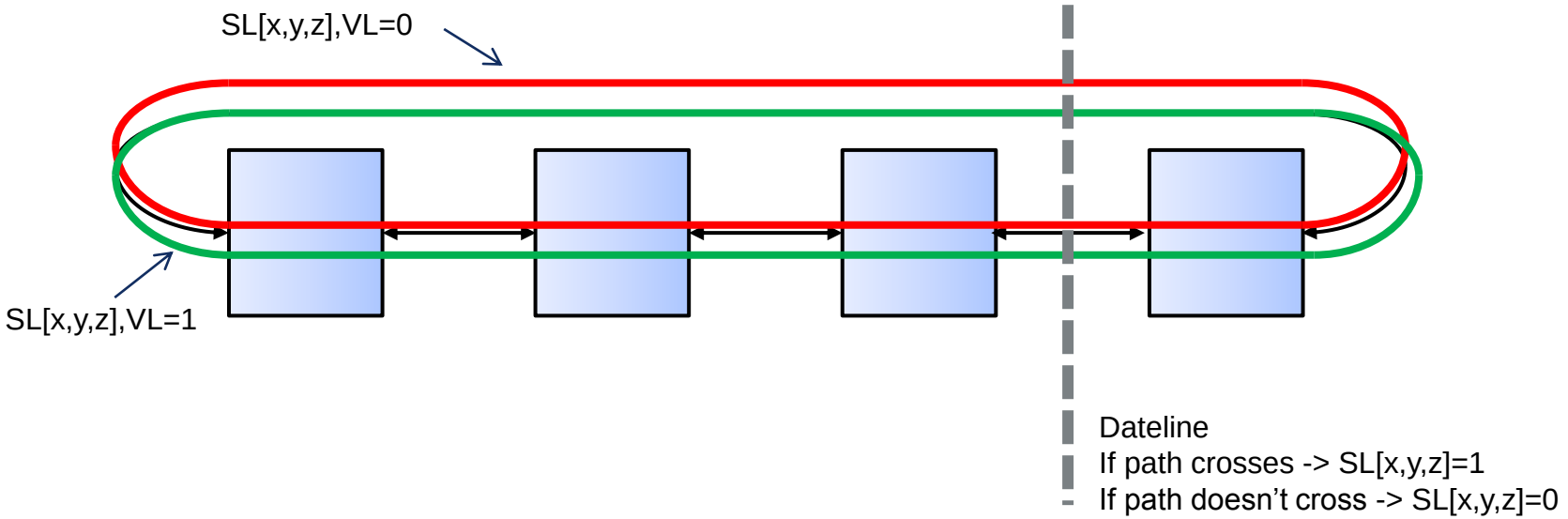
- With any loop topology, credit deadlocks need to be accounted for



Torus-2QOS Routing Algorithm

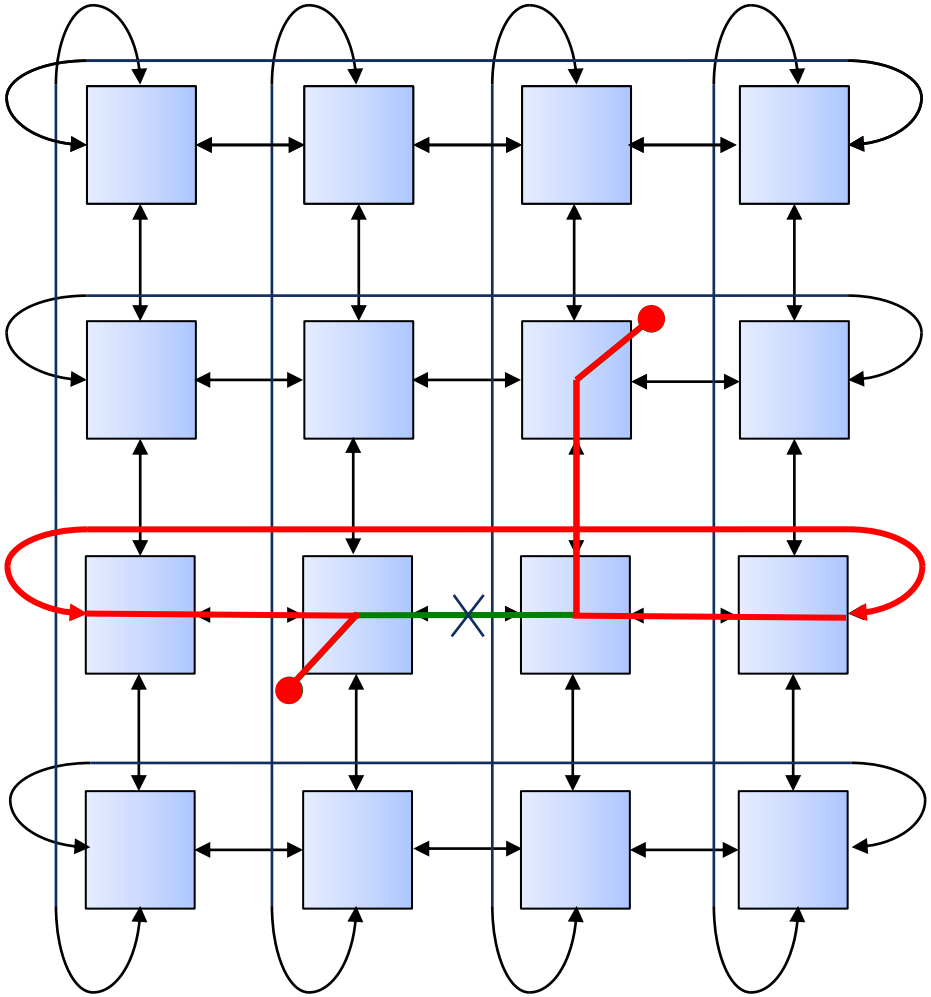


Torus-2QOS Routing Algorithm



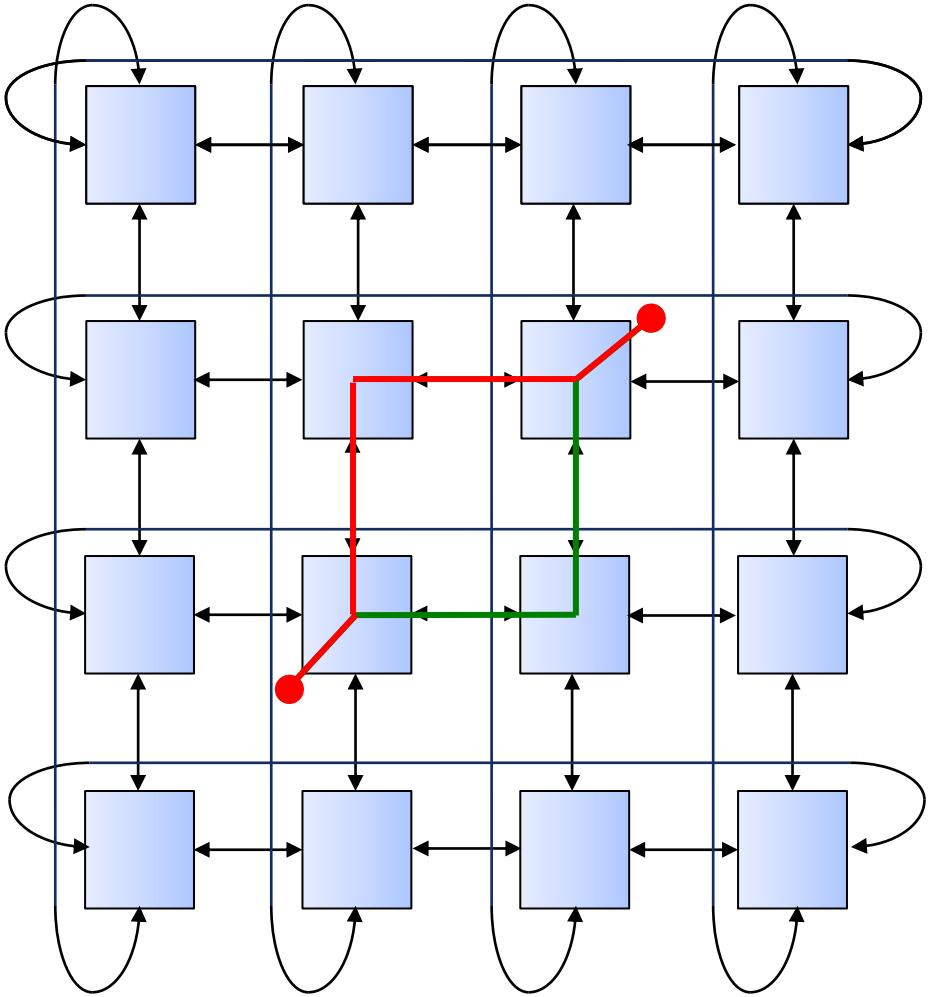
- Based on 'Dimension Ordered Routing' algorithm (e.g. route X, then Y, then Z)
- A single Service Level (SL) bit is used for each radix to dictate the Virtual Lane (or virtual switch buffers) used
- So, for 3-radix torus, 8 SLs and 2 separate virtual lanes are used
- InfiniBand supports 16 SLs so this allows for 2 levels of QOS in the fabric
- Algorithm allows for rerouting around multiple link faults and switch faults without changing SL

Link Failure Example



- Reroute on link failure without changing SL.
- Can handle multiple link failures in fabric

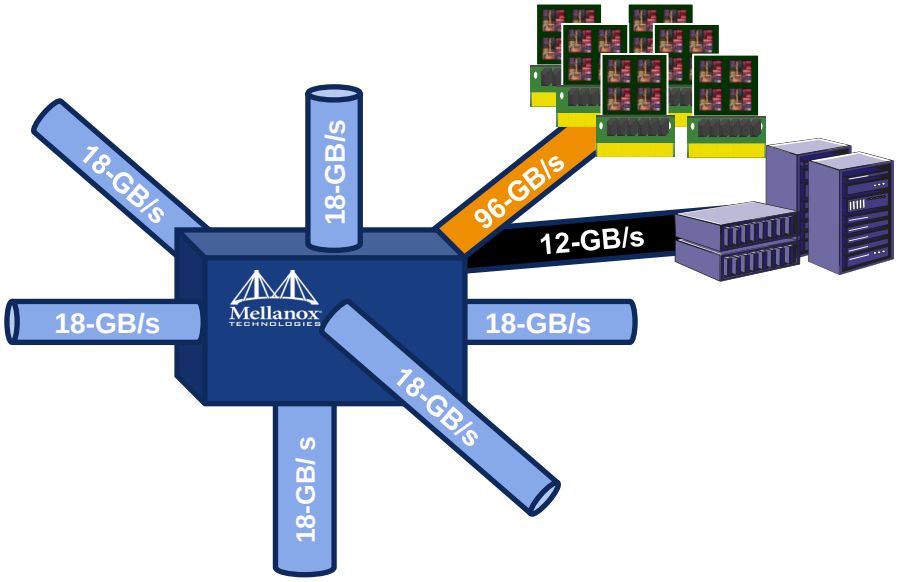
Switch Failure Example



- Reroute on switch failure without changing SL
- Uses unique InfiniBand capability of mapping SL->VL based on input/output port combination
- Illegal turn detected (Y->X route) and assigns it a different VL to avoid loop potential

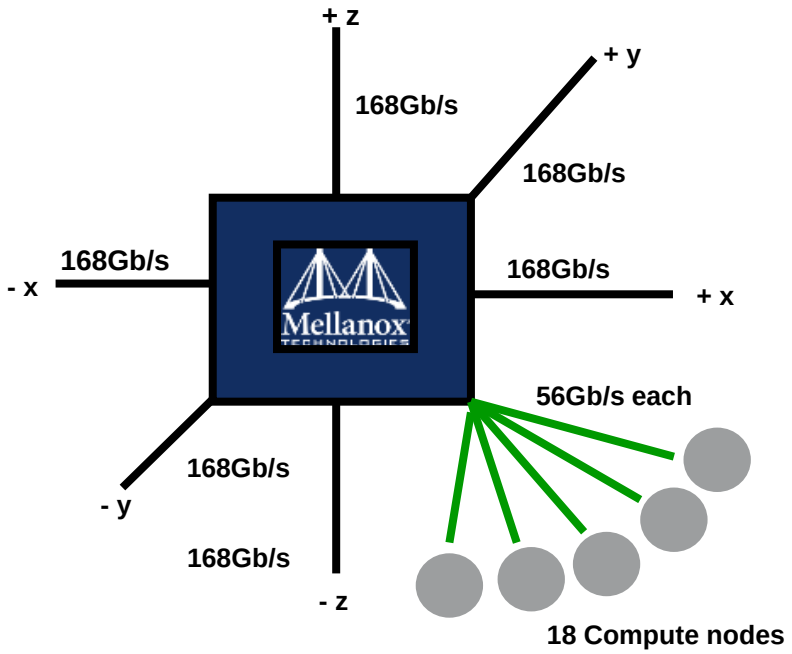
Torus2QOS Algorithm

- Expandable without re-cabling
- Simple wiring using short cables
- Fault tolerant with ability to handle multiple link and switch failures
- Two QOS levels
- GA in MLNX_OFED and being used in production on multiple clusters
  - Sandia Red Sky
  - San Diego Super Computing Gordon cluster



Network Properties

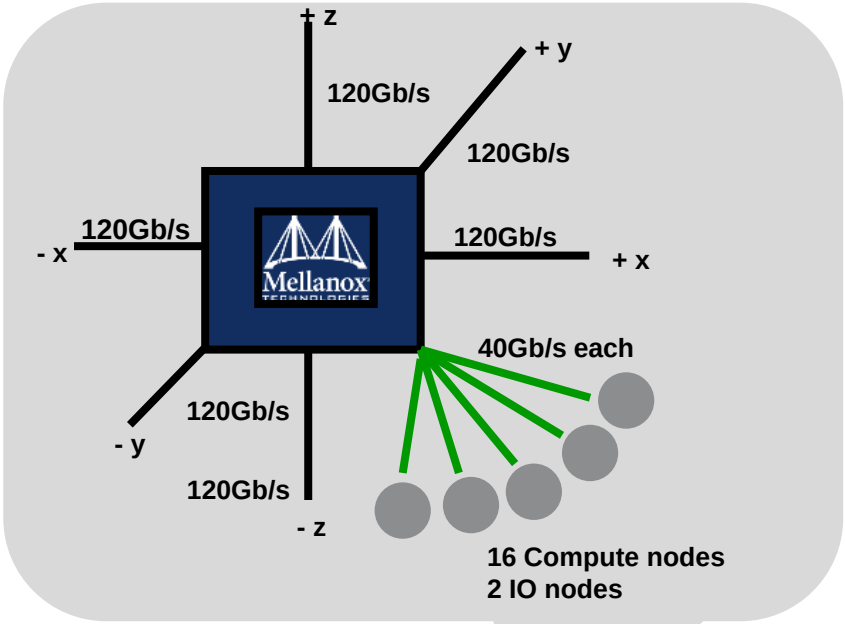
- Worse case latency is ‘1 + x/2 + y/2 + z/2’ switch hops (e.g. 4x4x4 torus is maximum 7 switch hops)
- Bisectional bandwidth is $2k^2$, where k is switch nodes in x,y or z directions (if uneven use 2xy, 2yz or 2xz)
- Flexible switch node configurations supported (e.g. 12 servers, and 224-Gb/s in each direction

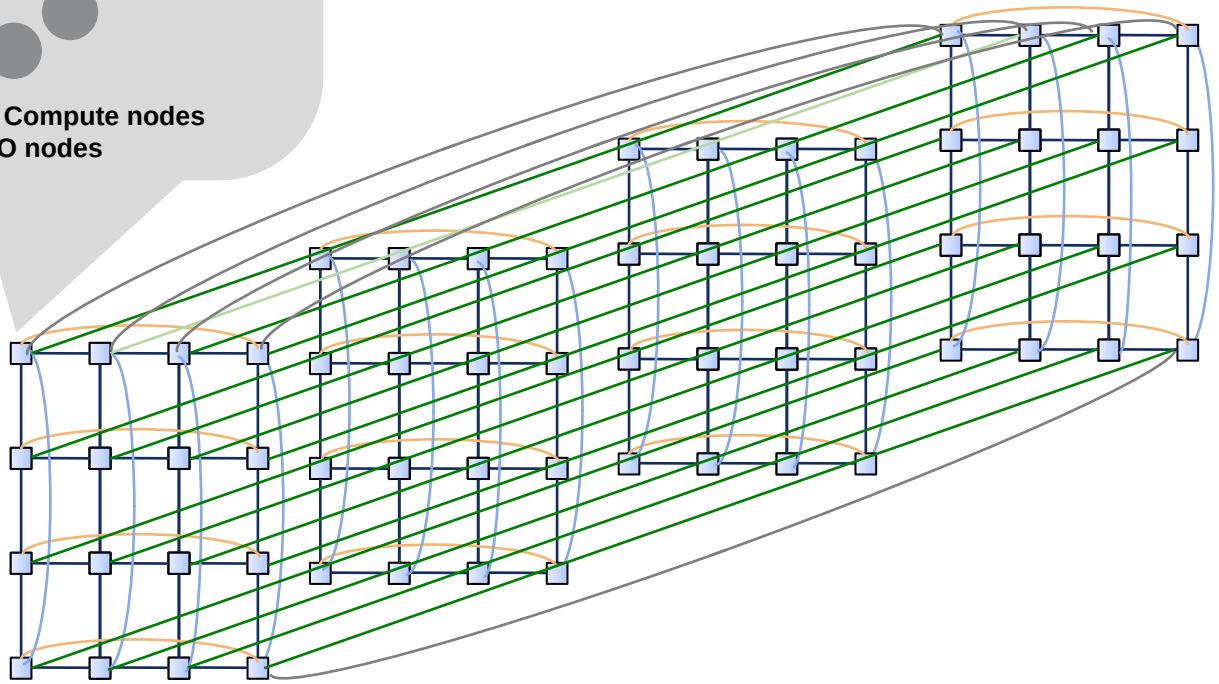


Cabling Techniques

- It is recommended to name switch nodes by their geographical location in the torus (e.g. switch_xyz, or switch_010…switch_003, etc)
- It is recommended that each direction on a switch node use the same ports (e.g. +x uses 1,2,3…-x uses 4,5,6… +y uses 7,8,9, etc)
  - switch_000 (1,2,3) => switch_100 (4,5,6)
  - switch_000 (4,5,6) => switch_300 (1,2,3)
  - switch_000 (7,8,9) => switch_010 (10,11,12)
  - switch_000 (10,11,12) => switch_030 (7,8,9)
  - switch_000 (13,14,15) => switch_001 (16,17,18)
  - switch_000 (16,17,18) => switch_003 (13,14,15)
- A cable technique to avoid the long cable in the final loopback connection is to cable switches using every other order
  - Example for 6-ary: 1 -> 3 -> 5 -> 6 -> 4 -> 2 -> 1

Example 3D Torus – SDSC Gordon





Thank You

HPC@mellanox.com