Understanding the Impact of Interconnect Failures on System Operation

20 Years of Excellence in Computational Science

OLCF

Oak Ridge Leadership Computing Facility

1992–2012

Presented by:

Matt Ezell

HPC Systems Administrator

Oak Ridge Leadership Computing Facility (OLCF)

National Center for Computational Sciences (NCCS)

CUG 2013

May 8, 2013

Napa Valley, CA

INTRODUCING TITAN

Advancing the Era of Accelerated Computing

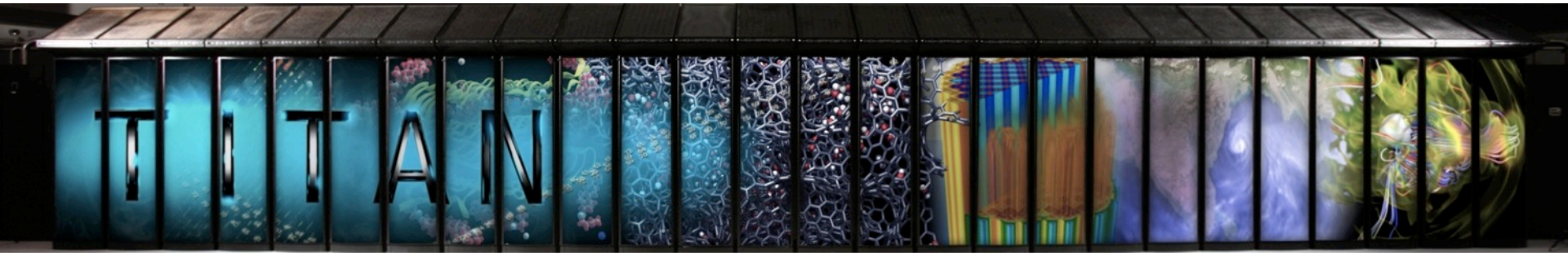
U.S. DEPARTMENT OF ENERGY | Office of Science

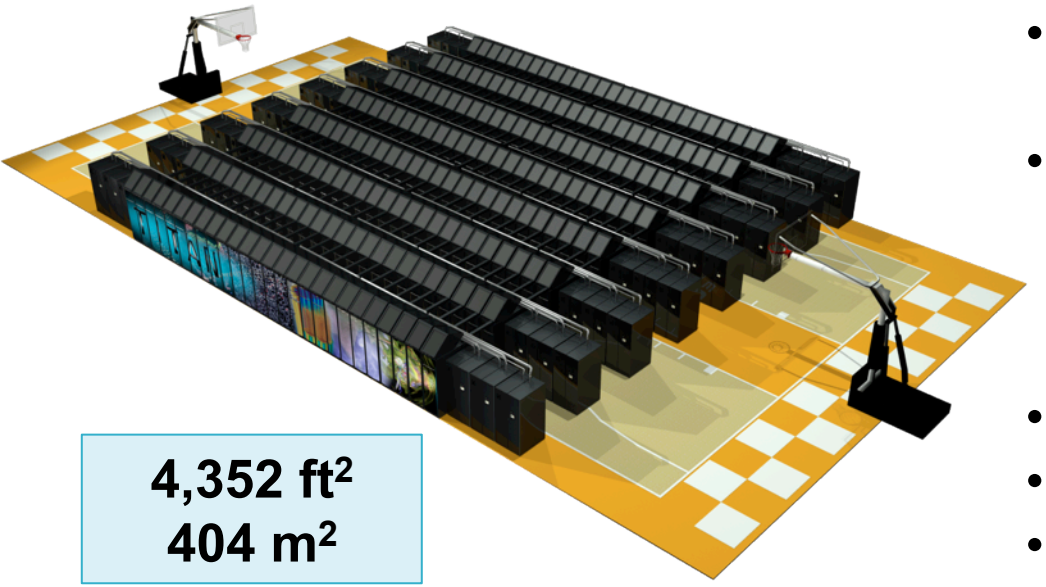
OAK RIDGE National Laboratory

OLCF
OAK RIDGE LEADERSHIP COMPUTING FACILITY

ORNL's "Titan" Hybrid System: World's Most Powerful Computer

#1 TOP500 SUPERCOMPUTER SITES #3 THE GREEN 500





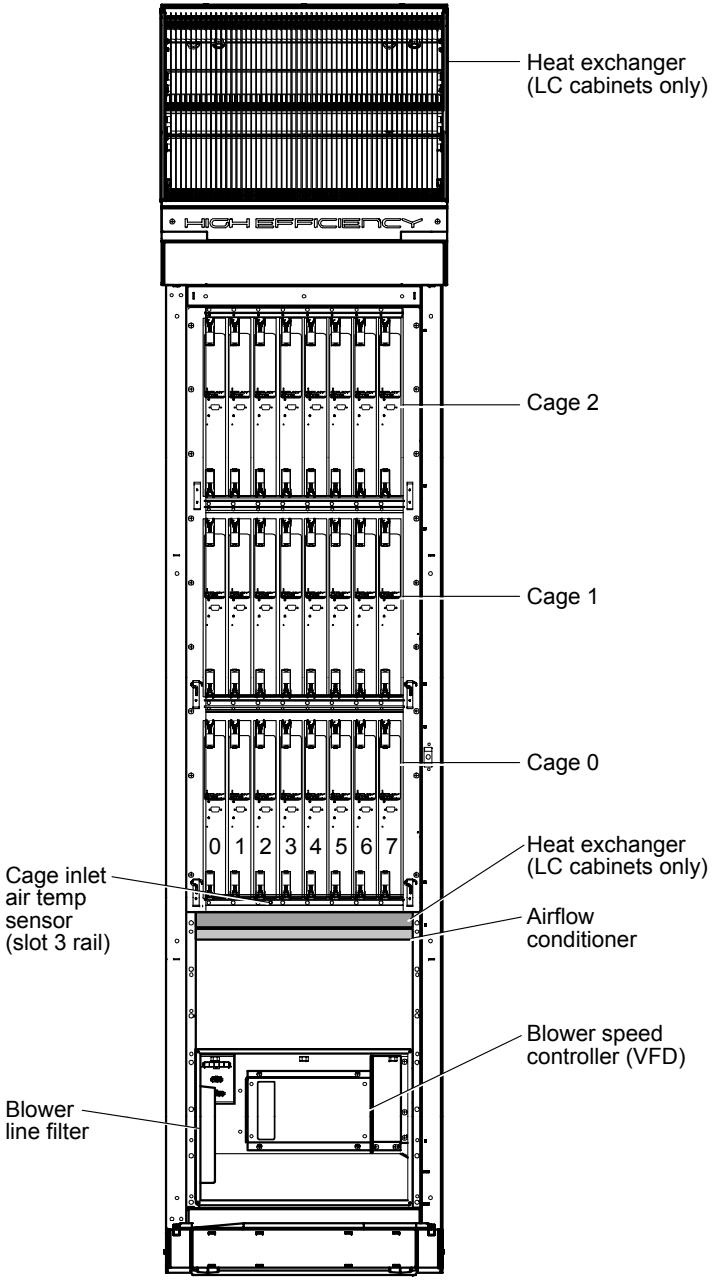
4,352 ft²
404 m²

SYSTEM SPECIFICATIONS:

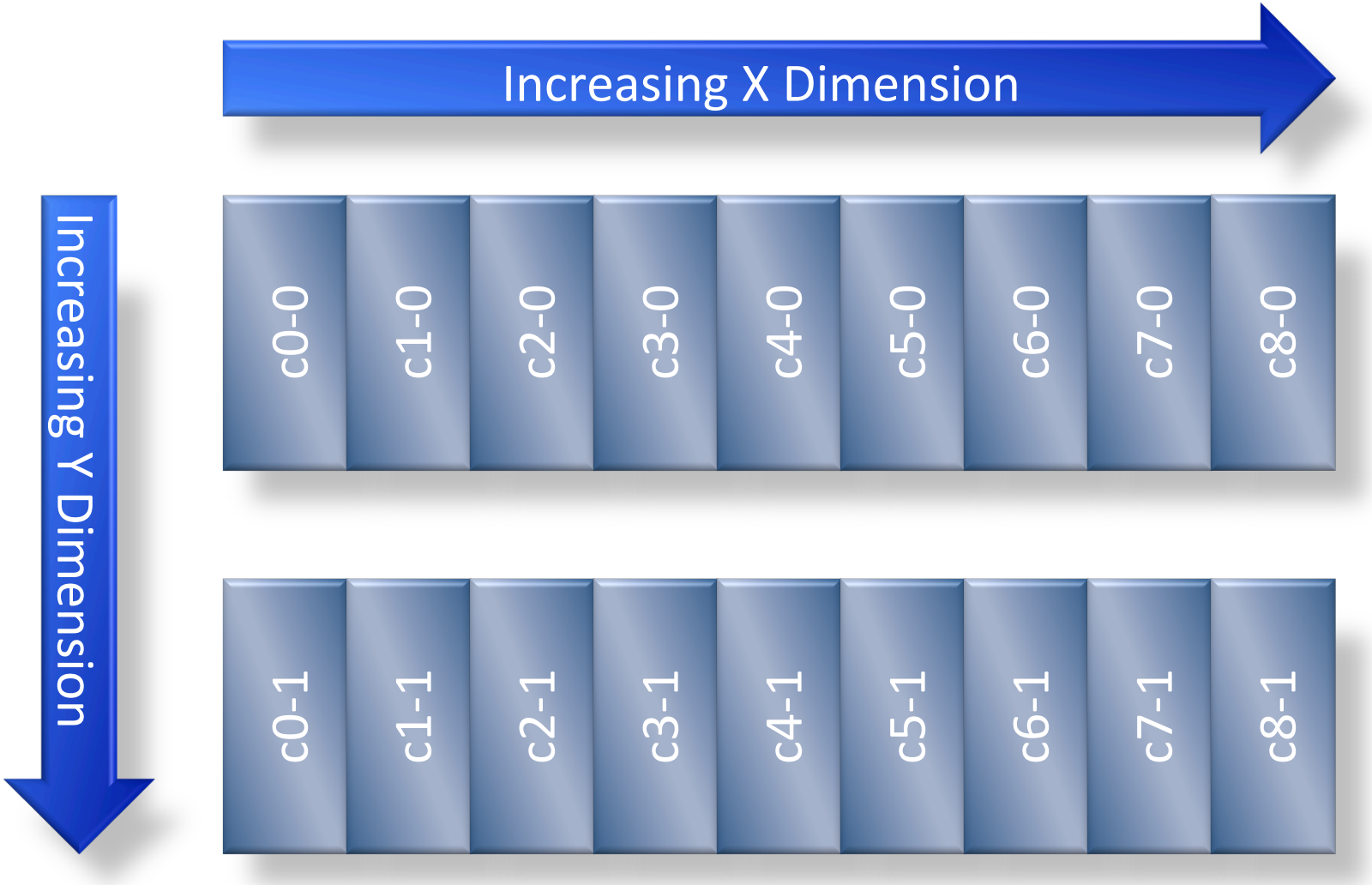
- Peak performance of 27.1 PF
  - 24.5 GPU + 2.6 CPU
- 18,688 Compute Nodes each with:
  - 16-Core AMD Opteron CPU
  - NVIDIA Tesla "K20x" GPU
  - 32 + 6 GB memory
- 512 Service and I/O nodes
- 200 Cabinets
- 710 TB total system memory
- Cray Gemini 3D Torus Interconnect
- 8.9 MW peak power

Cabinet Design

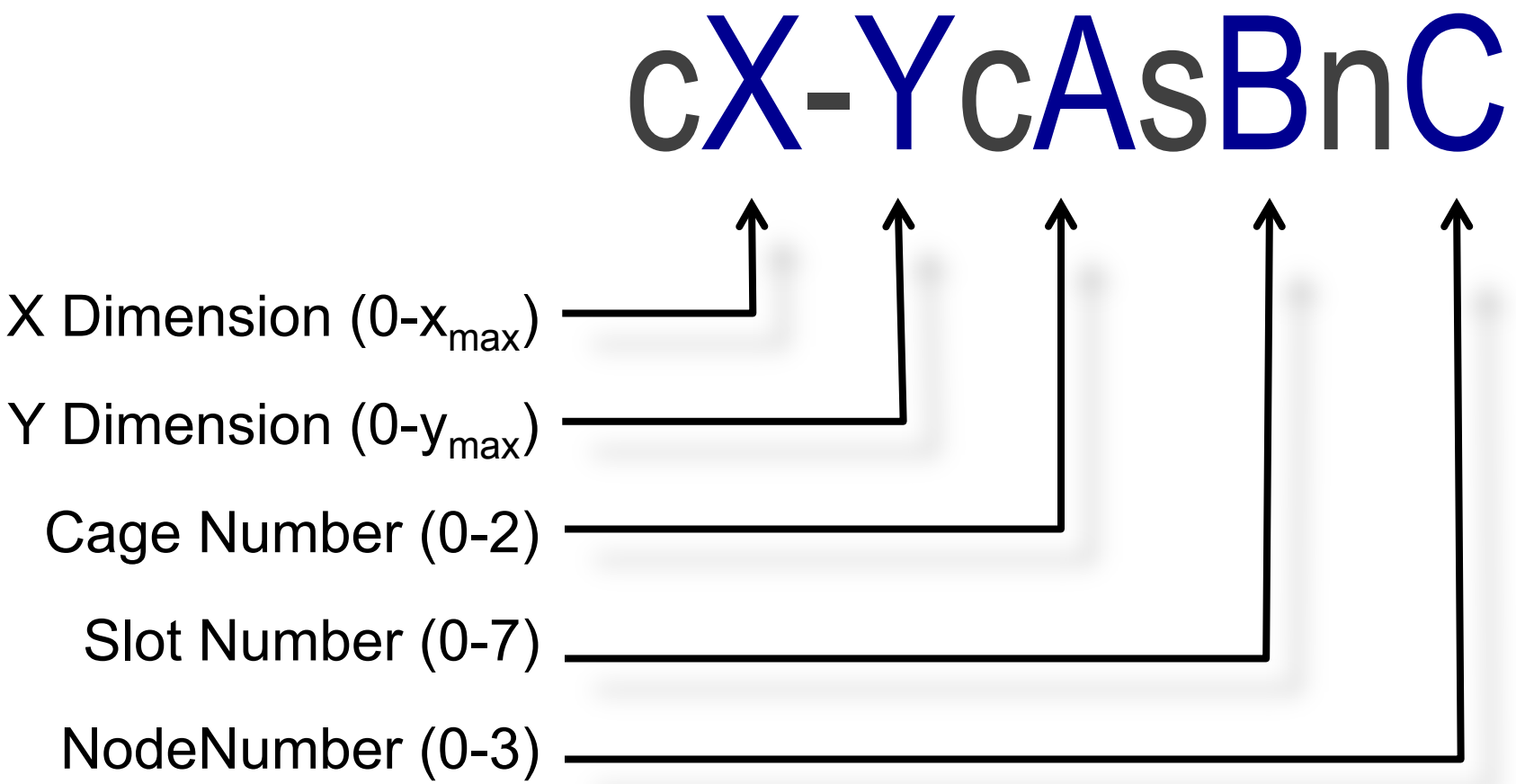
- One “L1” controller per cabinet
- Three cages (“chassis”) per cabinet
- Eight “modules” (“blades” or “slots”) per chassis
- One L0 controller, four nodes, and two Gemini NICs per module



Cabinet Numbering – Overhead View



A Node's "C-Name" or Physical Name

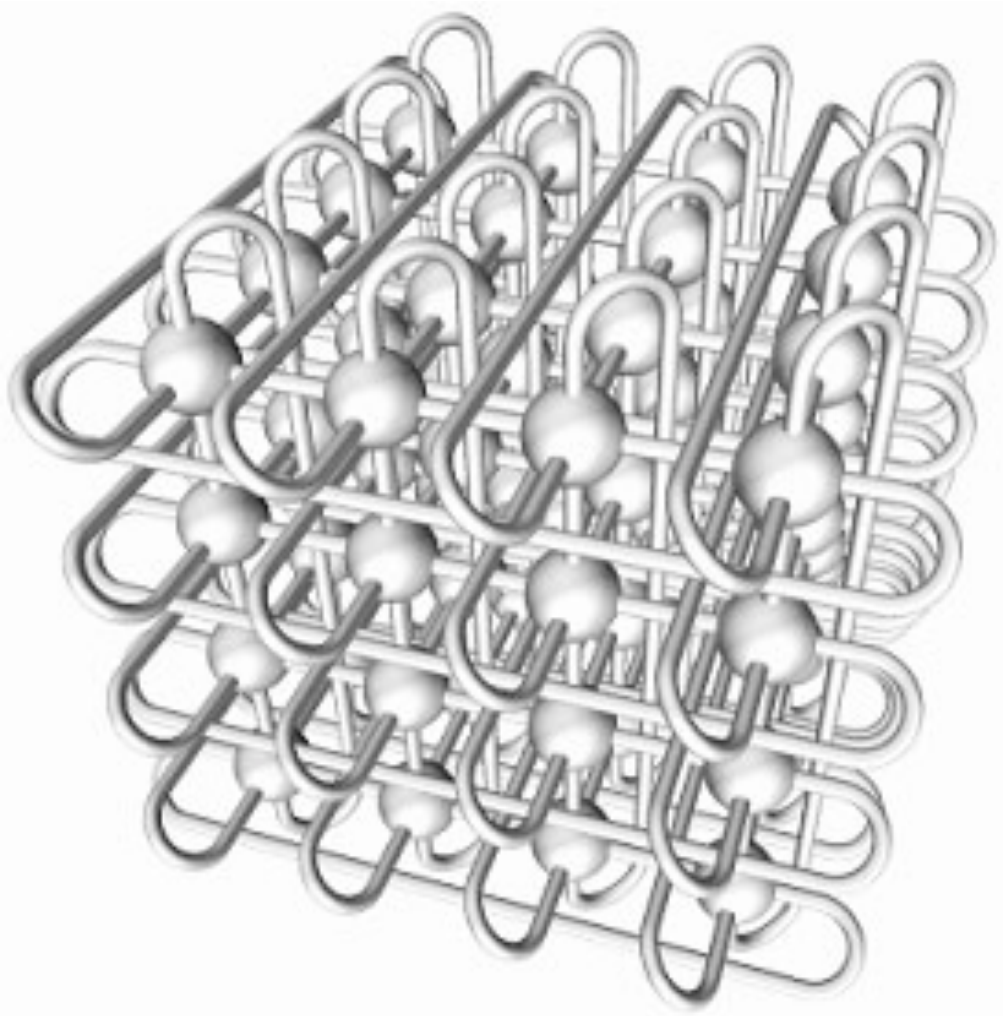


A Node's ID, or NID

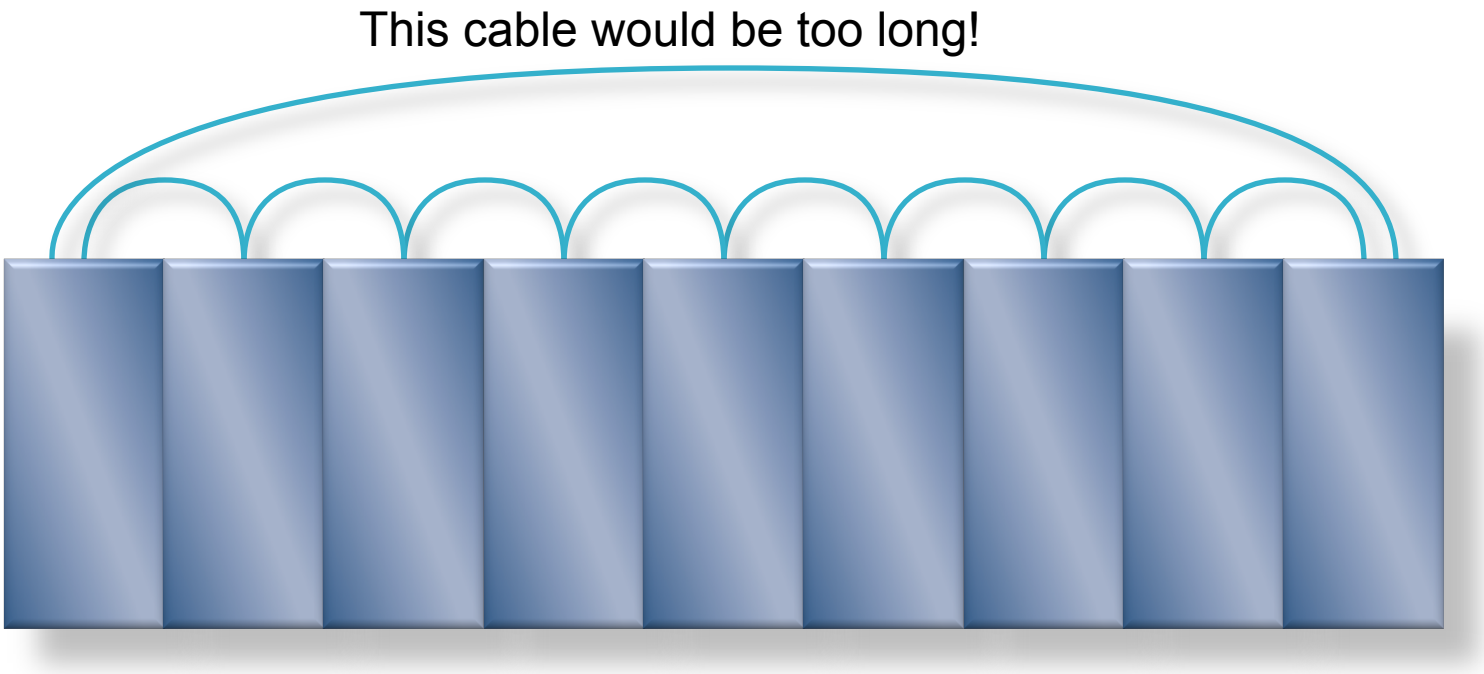
- Looks something like nid######
- Not very human friendly, used for routing purposes
- Can determine NID, cname, and topology coordinates from Cray-provided RCA module

```
// Figure out my rank and nid
int mynid;
int myrank;
rca_mesh_coord_t topo;
MPI_Comm_rank(MPI_COMM_WORLD,&myrank);
PMI_Get_nid(myrank, &mynid);
// Figure out my coordinates
rca_get_meshcoord(mynid, &topo);
// Get a list of all nodes
rs_node_array_t nodelist;
rca_get_sysnodes(&nodelist)
```

3D Torus

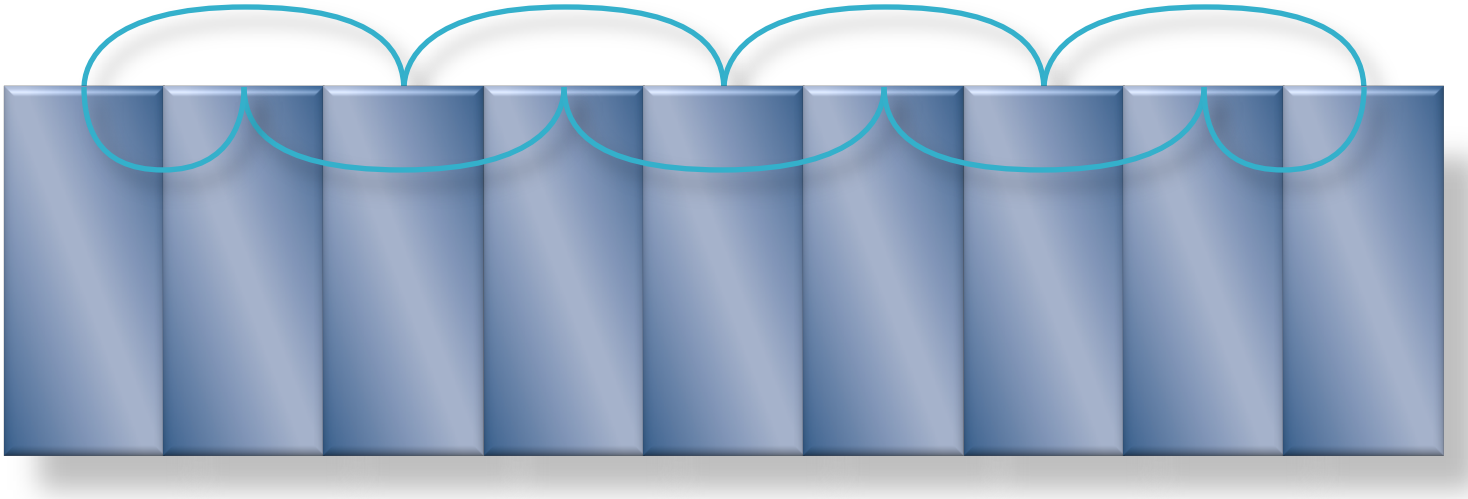


Topology – Cabling

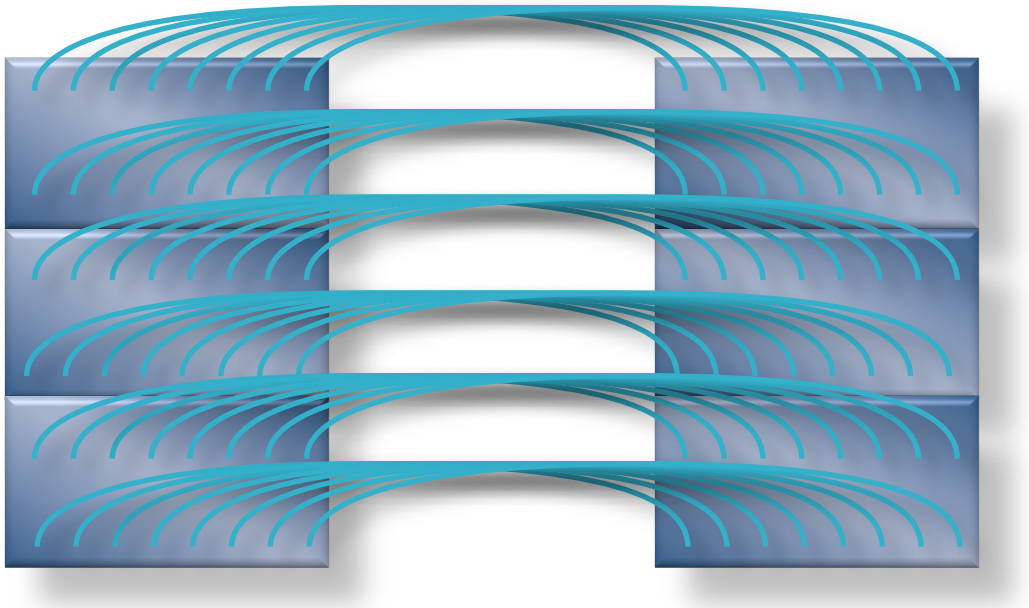


Topology – Cabling – Folded Torus

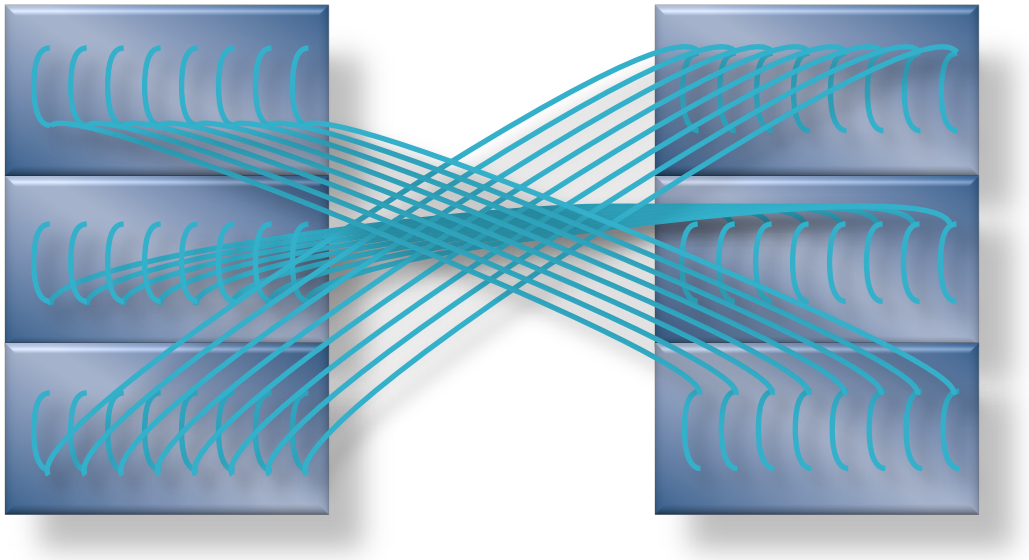
Reduces maximum cable length



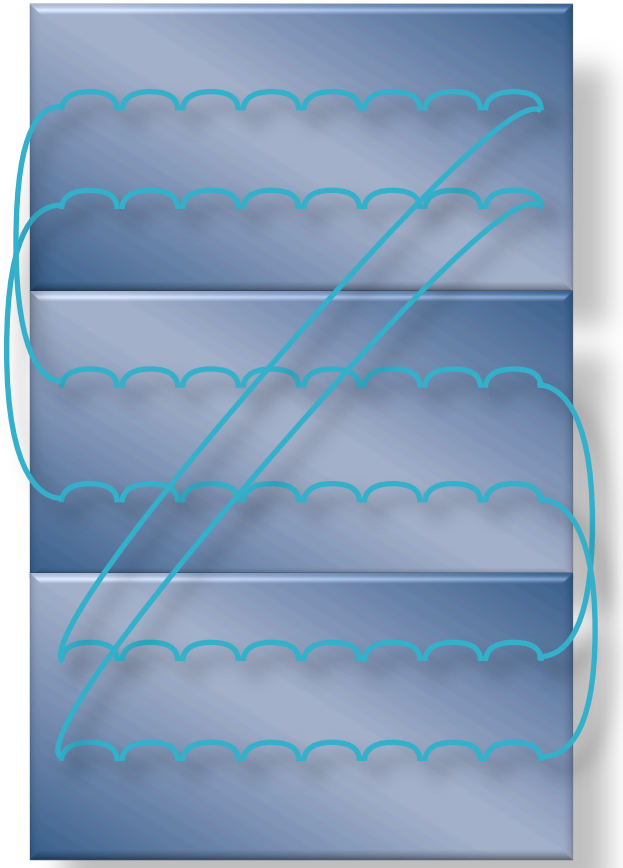
X Connectivity



Y Connectivity



Z Connectivity



Gemini Design

- Custom ASIC
- 2 NICs per chip
- 48 port router, tile-based design
- 8 tiles dedicated to NICs
- 40 tiles dedicated to external connectivity:

8 X+	8 Z+	4 Y+
8 X-	8 Z-	4 Y-

Gemini Tile Design

	0	1	2	3	4	5	6	7
0	Z+	Z+	X-	X-	X+	X+	Z-	Z-
1	Z+	Z+	X-	X-	X+	X+	Z-	Z-
2	Z-	Z-	Z-	P	P	Z+	Z+	Z+
3	X-	X-	Z-	P	P	Z+	X+	X+
4	X-	X-	Y-	P	P	Y+	X+	X+
5	Y-	Y-	Y-	P	P	Y+	Y+	Y+

How fast are Gemini Links?

- It depends!
- Link speed depends on link type
- There are 3 lanes associated with a network tile
- Protocol overhead is around 35% for large messages

Link Type	Link Data Rate	Number Links	Raw Bitrate	Data Rate
Y-Mezzanine	6.25 gbps	12	9.375 GB/s	~ 6 GB/s
Z-Backplane	5.0 gbps	24	15 GB/s	~ 9.75 GB/s
X, Z Cable	3.125 gbps	24	9.375 GB/s	~ 6 GB/s
Y Cable	3.125 gbps	12	4.6875 GB/s	~ 3 GB/s

Cray XK7 Compute Node

XK7 Compute Node Characteristics
AMD Opteron 6200 Interlagos 16 core processor @ 2.2GHz 32GB 1600 MHz DDR3
NVIDIA K20x (Kepler) with 6GB GDDR5 memory
Cray Gemini High Speed Interconnect
Four compute nodes per blade 24 blades per cabinet 200 cabinets

Gemini Routing

- Dimension ordered routing
  - In general, X ➔ Y ➔ Z
  - Must “break” the rules when faults exist
  - This can lead to link sharing and/or additional hops (read: performance problems)
  - Some configurations become unrouteable
- Dimension order retry
  - Fairly new option that allows alternate orderings
  - Could have **huge** impacts on optimal placement
  - Might also change LNET routing characteristics

Fault Tolerance

- Lanes within a link can degrade transparently
  - Might not be transparent performance-wise
- Failures of links require a re-route
  - Traffic quiesces to allow re-calculated routes to be asserted
  - Some traffic must be delivered in-order
- This also supports “warm swap” capabilities to change out failed hardware

Gemini Lane Degrade

```
130325 15:51:00  c0-0c1s0g0l27     c0-0c1s1g0l17       1 Mode Exchanges
130325 15:51:00  c0-0c1s0g0l27     c0-0c1s1g0l17       1 RX lanemask=3
130325 15:51:00  c0-0c1s0g0l27       ***ERROR*** Gemini LCB lane(s) reinit failed
130325 15:52:00  c0-0c1s1g0l17     c0-0c1s0g0l27       1 TX lanemask=3
```

Lane Mask Value	Lane Status
7	All lanes working
3, 5, 6	One lane degraded
1, 2, 4	Two lanes degraded

Network Congestion and Throttling

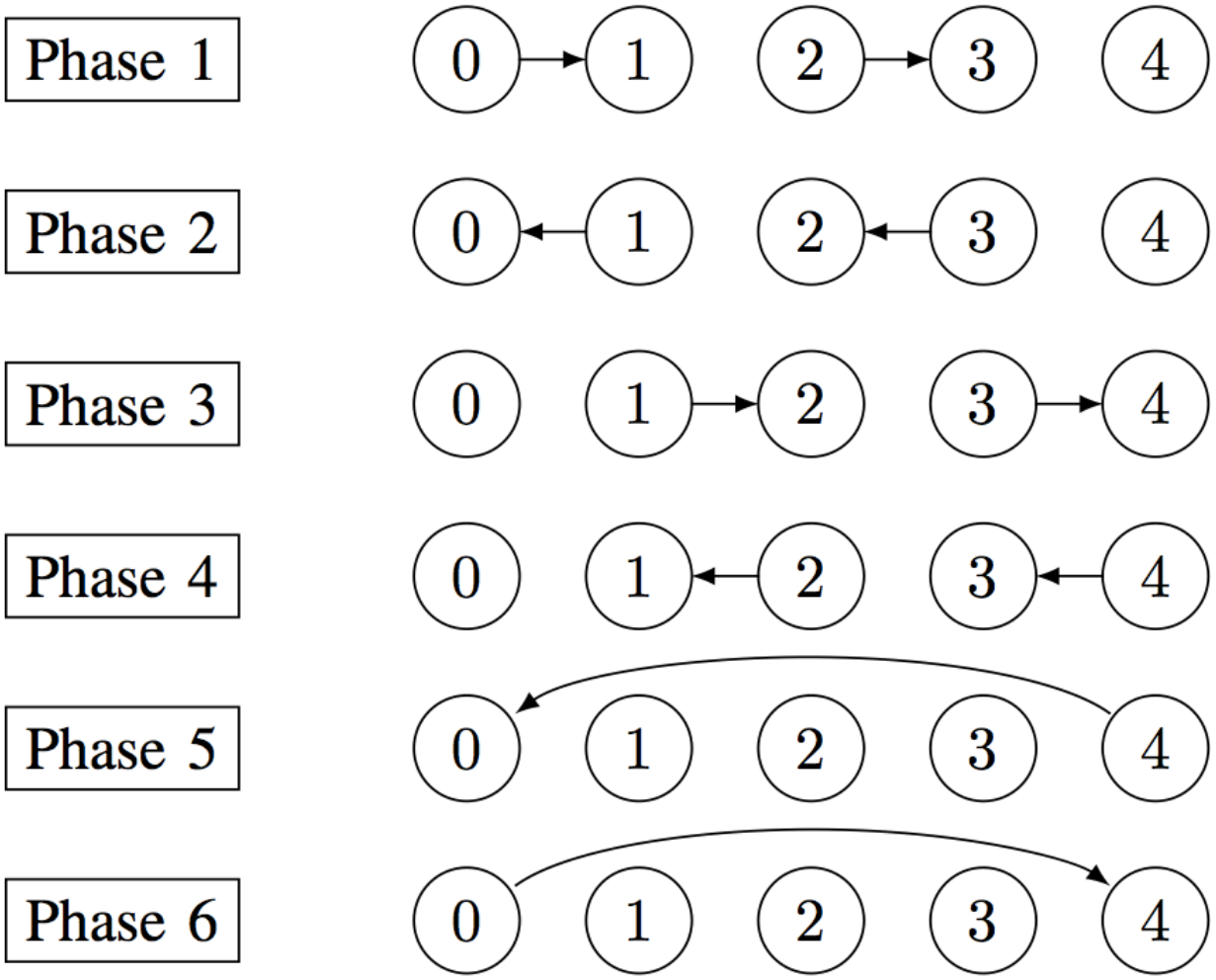
- "Bad" applications can overwhelm the network
  - Especially if they are placed non-optimally
- SMW watches for congestion and can "throttle" nodes that are injecting too many messages
- Modules may auto-throttle if they cannot talk to the SMW
  - And they can't tell you they are throttled if they cannot talk to the SMW
  - No cabinet heartbeat at the moment?
- Applications (or libraries) can use *balanced injection* to help mitigate congestion
- Impact on other applications depends on placement

The TopoBW Microbenchmark

- Created to understand network performance
- Use multiple iterations of concurrent asynchronous sends of large messages to fill the pipes
- Only transfer between nearest-neighbors
  - Network performance is the sum of the parts
  - Avoid link sharing

$$Bandwidth = \frac{AmountTransferred}{TimeElapsed}$$

TopoBW Phases



TopoBW Output

```
c0-0c1s5n1 (nid00053), Rank  0/30, Topology [ 1, 0, 5] of [ 2, 1, 7], NeighborRanks [ 1,-1,-1,-1,-1,-1] PartnerRank -1
c0-0c1s7n0 (nid00048), Rank 24/30, Topology [ 1, 0, 7] of [ 2, 1, 7], NeighborRanks [-1,-1,27,25,-1,19] PartnerRank 25
c0-0c1s3n2 (nid00038), Rank 10/30, Topology [ 1, 1, 3] of [ 2, 1, 7], NeighborRanks [ 8,-1,11, 9,14,-1] PartnerRank 9
c0-0c1s7n3 (nid00047), Rank 27/30, Topology [ 1, 1, 7] of [ 2, 1, 7], NeighborRanks [29,-1,26,24,23,20] PartnerRank 26
c0-0c0s3n1 (nid00007), Rank  5/30, Topology [ 0, 0, 3] of [ 2, 1, 7], NeighborRanks [-1,11, 6, 8,-1,-1] PartnerRank 6
c0-0c1s2n1 (nid00059), Rank 15/30, Topology [ 1, 0, 2] of [ 2, 1, 7], NeighborRanks [-1,-1,16,14,-1,11] PartnerRank 16
c0-0c0s3n0 (nid00006), Rank  6/30, Topology [ 0, 0, 3] of [ 2, 1, 7], NeighborRanks [-1,12, 7, 5,-1,-1] PartnerRank 5
c0-0c0s5n0 (nid00010), Rank  2/30, Topology [ 0, 0, 5] of [ 2, 1, 7], NeighborRanks [-1,-1, 3, 1,-1,-1] PartnerRank 1
c0-0c1s3n0 (nid00056), Rank 12/30, Topology [ 1, 0, 3] of [ 2, 1, 7], NeighborRanks [ 6,-1, 9,11,16,-1] PartnerRank 11
c0-0c0s5n1 (nid00011), Rank  1/30, Topology [ 0, 0, 5] of [ 2, 1, 7], NeighborRanks [-1, 0, 2, 4,-1,-1] PartnerRank 2
c0-0c0s5n2 (nid00020), Rank  4/30, Topology [ 0, 1, 5] of [ 2, 1, 7], NeighborRanks [-1,-1, 1, 3,-1,-1] PartnerRank 3
c0-0c1s6n3 (nid00045), Rank 23/30, Topology [ 1, 1, 6] of [ 2, 1, 7], NeighborRanks [-1,-1,-1,-1,-1,27] PartnerRank -1
c0-0c1s1n3 (nid00035), Rank 22/30, Topology [ 1, 1, 1] of [ 2, 1, 7], NeighborRanks [-1,-1,-1,-1,20,13] PartnerRank -1
c0-0c1s0n2 (nid00032), Rank 21/30, Topology [ 1, 1, 0] of [ 2, 1, 7], NeighborRanks [-1,-1,18,20,26,-1] PartnerRank 20
c0-0c1s7n1 (nid00049), Rank 25/30, Topology [ 1, 0, 7] of [ 2, 1, 7], NeighborRanks [17,-1,24,26,-1,18] PartnerRank 24
c0-0c1s0n3 (nid00033), Rank 20/30, Topology [ 1, 1, 0] of [ 2, 1, 7], NeighborRanks [-1,-1,21,19,27,22] PartnerRank 21
c0-0c1s0n0 (nid00062), Rank 19/30, Topology [ 1, 0, 0] of [ 2, 1, 7], NeighborRanks [-1,-1,20,18,24,-1] PartnerRank 18
c0-0c1s3n1 (nid00057), Rank 11/30, Topology [ 1, 0, 3] of [ 2, 1, 7], NeighborRanks [ 5,-1,12,10,15,-1] PartnerRank 12
c0-0c0s7n1 (nid00015), Rank 17/30, Topology [ 0, 0, 7] of [ 2, 1, 7], NeighborRanks [-1,25,-1,28,-1,-1] PartnerRank -1
c0-0c1s7n2 (nid00046), Rank 26/30, Topology [ 1, 1, 7] of [ 2, 1, 7], NeighborRanks [28,-1,25,27,-1,21] PartnerRank 27
c0-0c0s3n2 (nid00024), Rank  8/30, Topology [ 0, 1, 3] of [ 2, 1, 7], NeighborRanks [-1,10, 5, 7,-1,-1] PartnerRank 7
c0-0c1s2n3 (nid00037), Rank 13/30, Topology [ 1, 1, 2] of [ 2, 1, 7], NeighborRanks [-1,-1,14,16,22, 9] PartnerRank 14
c0-0c1s0n1 (nid00063), Rank 18/30, Topology [ 1, 0, 0] of [ 2, 1, 7], NeighborRanks [-1,-1,19,21,25,-1] PartnerRank 19
c0-0c0s7n3 (nid00017), Rank 29/30, Topology [ 0, 1, 7] of [ 2, 1, 7], NeighborRanks [-1,27,28,-1,-1,-1] PartnerRank 28
c0-0c1s3n3 (nid00039), Rank  9/30, Topology [ 1, 1, 3] of [ 2, 1, 7], NeighborRanks [ 7,-1,10,12,13,-1] PartnerRank 10
c0-0c0s3n3 (nid00025), Rank  7/30, Topology [ 0, 1, 3] of [ 2, 1, 7], NeighborRanks [-1, 9, 8, 6,-1,-1] PartnerRank 8
c0-0c1s2n0 (nid00058), Rank 16/30, Topology [ 1, 0, 2] of [ 2, 1, 7], NeighborRanks [-1,-1,13,15,-1,12] PartnerRank 15
c0-0c0s7n2 (nid00016), Rank 28/30, Topology [ 0, 1, 7] of [ 2, 1, 7], NeighborRanks [-1,26,17,29,-1,-1] PartnerRank 29
c0-0c0s5n3 (nid00021), Rank  3/30, Topology [ 0, 1, 5] of [ 2, 1, 7], NeighborRanks [-1,-1, 4, 2,-1,-1] PartnerRank 4
c0-0c1s2n2 (nid00036), Rank 14/30, Topology [ 1, 1, 2] of [ 2, 1, 7], NeighborRanks [-1,-1,15,13,-1,10] PartnerRank 13
```

TopoBW Output

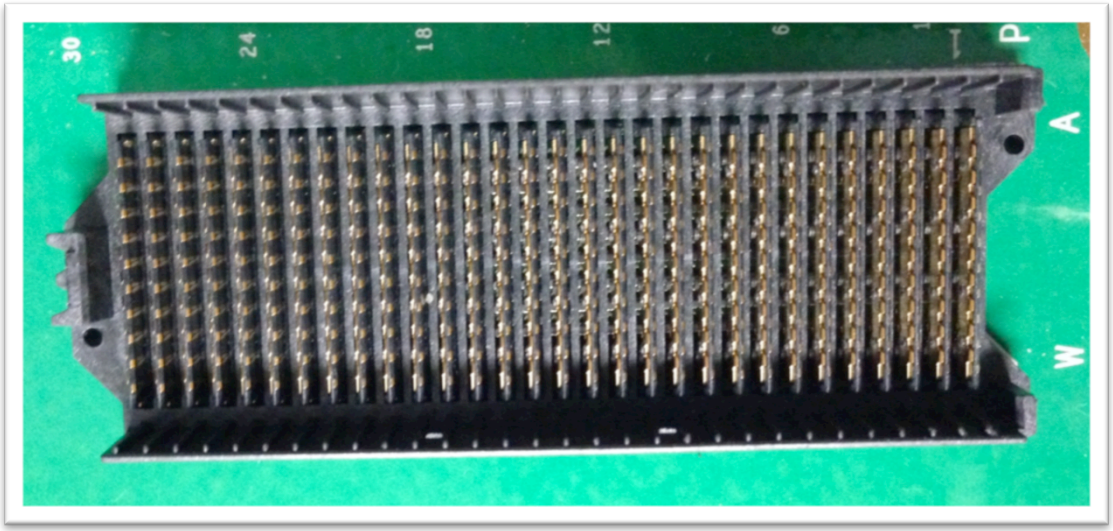
```
c0-0c0s5n1 X+CE c0-0c1s5n1 (nid00011 to nid00053):  5.0454 seconds = 5945.98 MB/sec
c0-0c0s3n2 X+CS c0-0c1s3n2 (nid00024 to nid00038): 10.0846 seconds = 2974.83 MB/sec
c0-0c0s3n3 X+CS c0-0c1s3n3 (nid00025 to nid00039): 10.0851 seconds = 2974.68 MB/sec
c0-0c0s3n0 Y+GS c0-0c0s3n1 (nid00006 to nid00007):  4.4262 seconds = 6777.84 MB/sec
c0-0c0s7n1 Y+ME c0-0c0s7n2 (nid00015 to nid00016):  5.0445 seconds = 5947.03 MB/sec
c0-0c1s0n3 Y+MS c0-0c1s0n0 (nid00033 to nid00062):  5.4767 seconds = 5477.70 MB/sec
c0-0c1s6n3 Z+BE c0-0c1s7n3 (nid00045 to nid00047):  4.4365 seconds = 6762.15 MB/sec
c0-0c1s2n0 Z+BS c0-0c1s3n0 (nid00058 to nid00056):  7.2851 seconds = 4118.02 MB/sec
c0-0c1s2n1 Z+BS c0-0c1s3n1 (nid00059 to nid00057):  7.2847 seconds = 4118.23 MB/sec
c0-0c1s7n0 Z+CS c0-0c1s0n0 (nid00048 to nid00062): 10.5132 seconds = 2853.55 MB/sec
```

TopoBW Output

```
Bandwidth X
    min: 2974.68 on c0-0c0s3n3
    max: 5947.06 on c0-0c1s5n1
    avg: 3717.81
Bandwidth Y
    min: 5437.83 on c0-0c1s3n1
    max: 6781.81 on c0-0c1s3n3
    avg: 6124.41
Bandwidth Z
    min: 2667.21 on c0-0c1s0n0
    max: 6762.15 on c0-0c1s6n3
    avg: 4364.39
```

TopoBW Results

- Low results can indicate lane degrades
  - Look in the *netwatch* file to verify
- Found two slow Geminis on Titan
  - Modules returned to Cray for PFA
  - Faulty connectors discovered
    - Mezzanine NexLev
    - *XU7000* Opteron socket



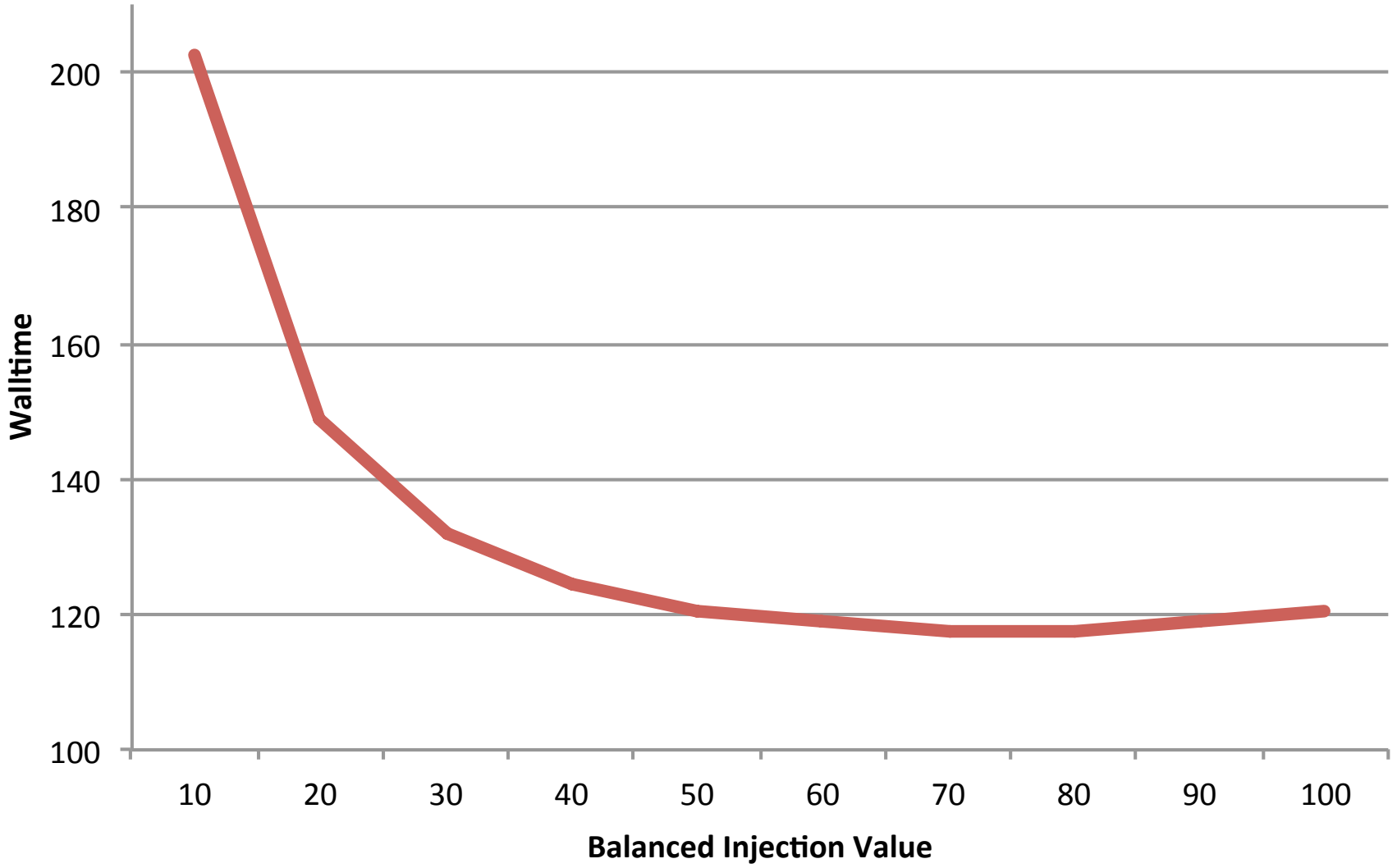
TopoBW Limitations

- Single node injection bandwidth is less than Z-backplane bandwidth
  - Won't always see minor degradations IF you don't have both nodes participating
- Needs to run on full system if you want to trust results
  - Because it can't tell what other nodes are doing

TopoBW and MMRs

- Previously, network performance data was only available through CrayPat
  - TopoBW needs much finer-grained access to the counters
- Early work to read Gemini MMR data
  - Uses not-well-documented access to *gpcd* device
  - Should allow per-link statistics
  - Analysis still ongoing
- March PE release supports Gemini performance data through PAPI
  - Probably makes sense to move to this “supported” mechanism

S3D Balanced Injection Sweep



So How Do Interconnect Failures Impact System Operation?

- It depends!
- Is the network the bottleneck for your application?
  - Is it bandwidth-limited or latency-sensitive?
- How well did the scheduler place your application?
  - Today, probably not very well
  - Hopefully this will improve in the future
- Extent of impact is difficult to quantify
  - And expensive to test
  - How do you “break” parts or simulate their failure?

What we do know

- Lane degrades reduce available bandwidth
- Missing Geminis cause routing turns that will oversubscribe links
- Throttling greatly reduces injection bandwidth
- Re-route must quiesce traffic, but this is temporary

Conclusions

- Gemini was a large step forward toward resilience in Cray systems
- Faults or congestion can lead to suboptimal network performance
  - This can be difficult to quantify
  - Extent of degradation depends on communication pattern
- Cray provided diagnostics do not find all errors
- ORNL-developed TopoBW can help spot performance anomalies

Questions?

ezellma@ornl.gov