MVAPICH2 with Dual Rail 3D Torus Support on SDSC Data Intensive Gordon Cluster

Mahidhar Tatineni (mahidhar@sdsc.edu)

Amit Majumdar (majumdar@sdsc.edu)

MVAPICH2 User Group Meeting

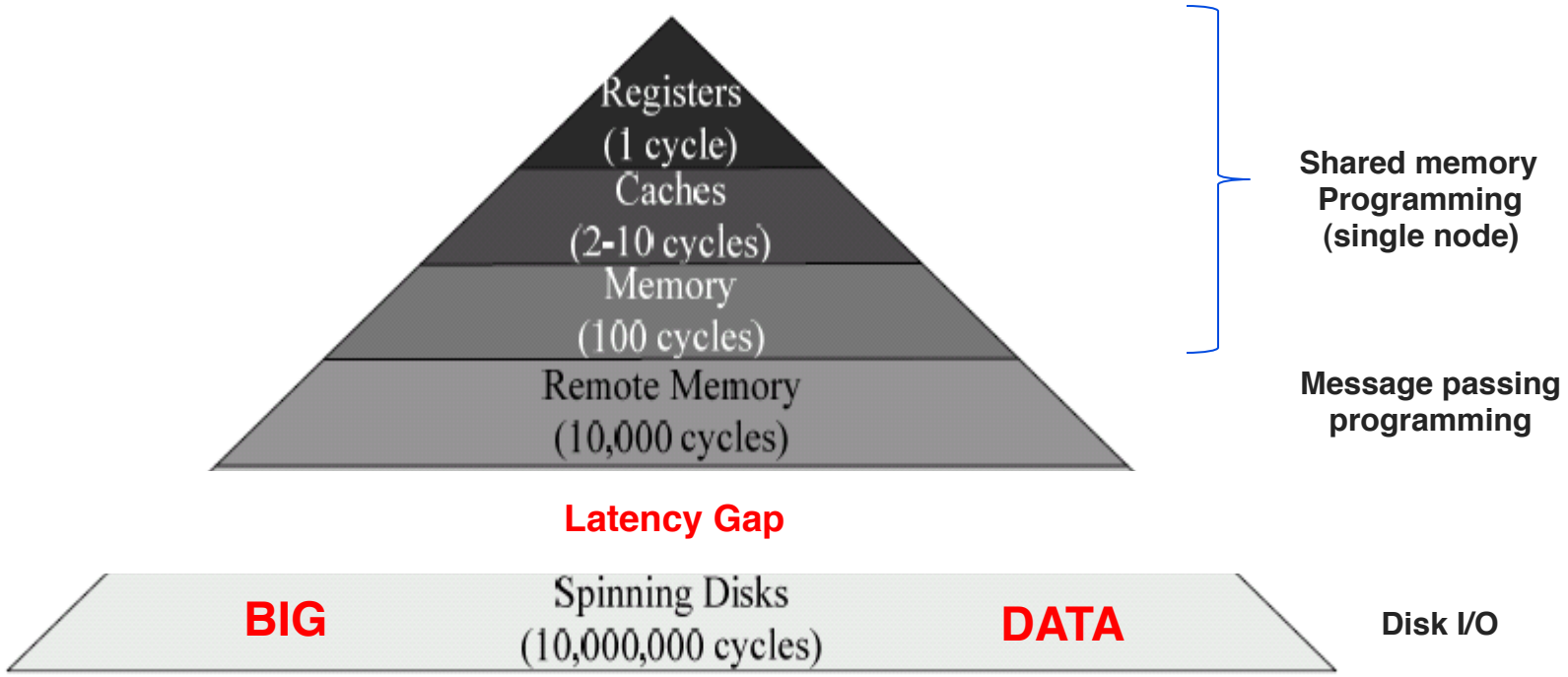
August 26 , 2013

Ref: Work by Pietro Cicotti, Robert Sinkovits, and Mahidhar Tatineni for Gordon acceptance benchmarking.

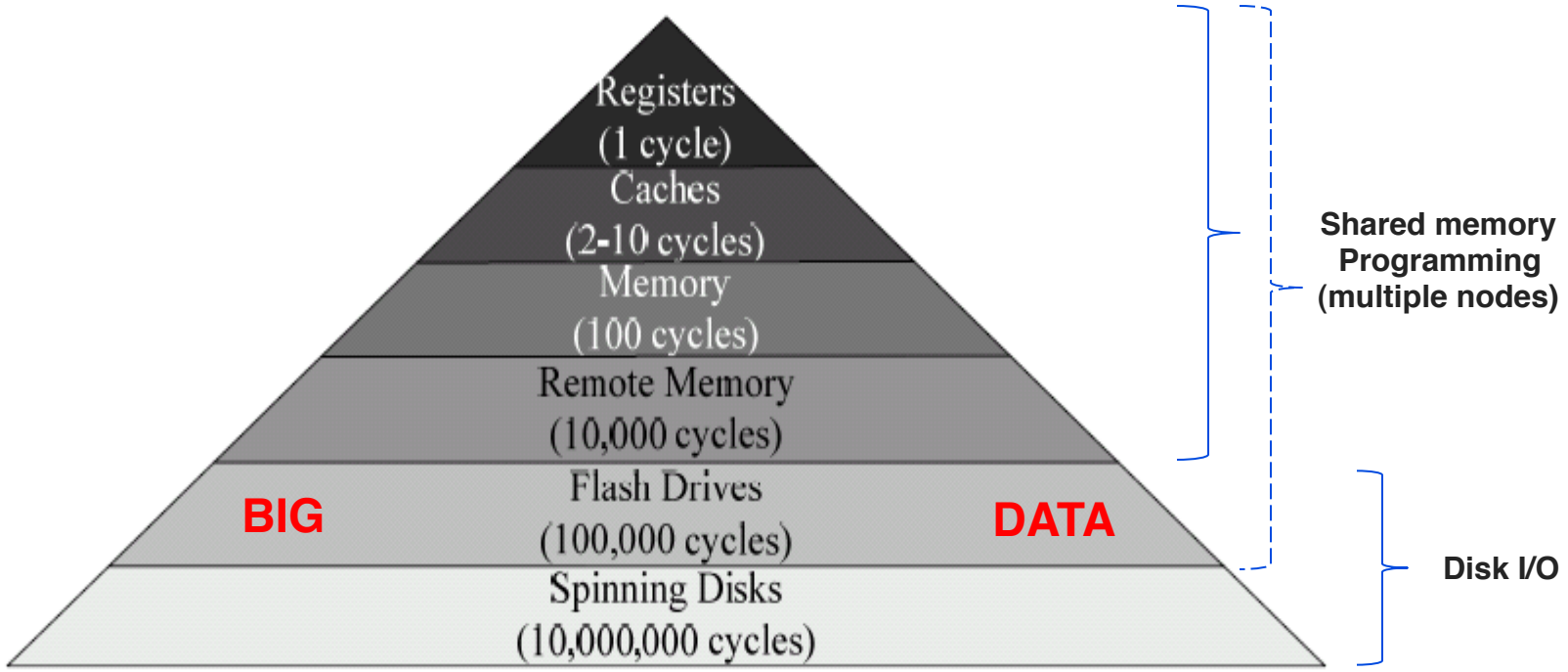
Gordon – Data Intensive Supercomputer

- Designed to accelerate access to massive amounts of data in areas of genomics, earth science, engineering, medicine, and others
- Emphasizes memory and IO over FLOPS.
- Appro integrated 1,024 node Sandy Bridge cluster
- 300 TB of high performance Intel flash
- Large memory supernodes via vSMP Foundation from ScaleMP
- 3D torus interconnect from Mellanox
- In production operation since February 2012
- Funded by the NSF and available through the NSF Extreme Science and Engineering Discovery Environment program (XSEDE)

The Memory Hierarchy of a Typical Supercomputer



The Memory Hierarchy of Gordon

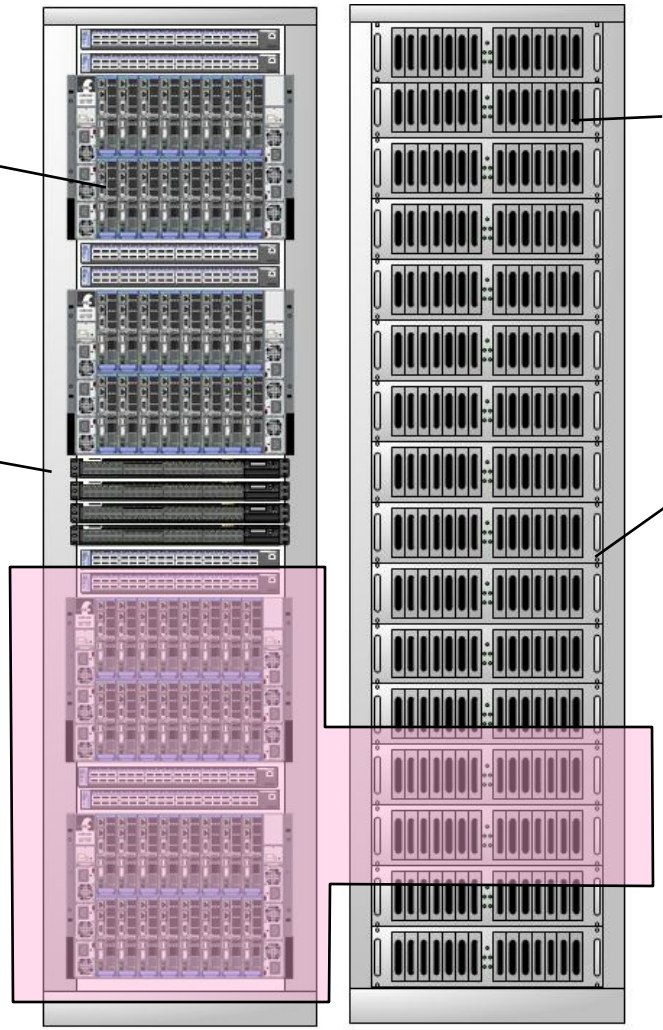


Gordon Design Highlights

- **1,024 2S Xeon E5 (Sandy Bridge) nodes**
- **16 cores, 64 GB/node**
- Intel Jefferson Pass **mobo**
- **PCI Gen3**

- **3D Torus**
- Dual rail QDR

- **Large Memory vSMP Supernodes**
- 2TB DRAM
- **10 TB Flash**

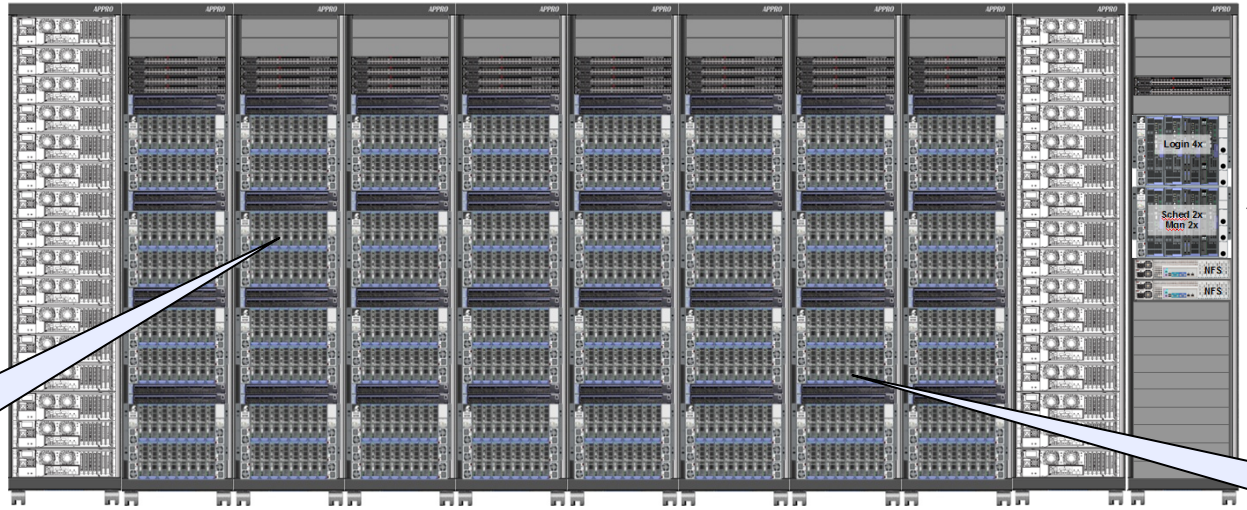


- 300 GB Intel 710 eMLC SSDs
- **300 TB aggregate**

- **64, 2S Westmere I/O nodes**
- 12 core, **48 GB/node**
- **4 LSI controllers**
- **16 SSDs**
- **Dual 10GbE**
- SuperMicro mobo
- **PCI Gen2**

"Data Oasis"
Lustre PFS
100 GB/sec, 4 PB

Gordon Layout: 21x48U Racks



128x36 port Mellanox InifiniScale QDR

Service Nodes
NFS, Login,
Rocks Mgmt
Core Ethernet switches

Compute Nodes
16 racks
64 nodes/rack

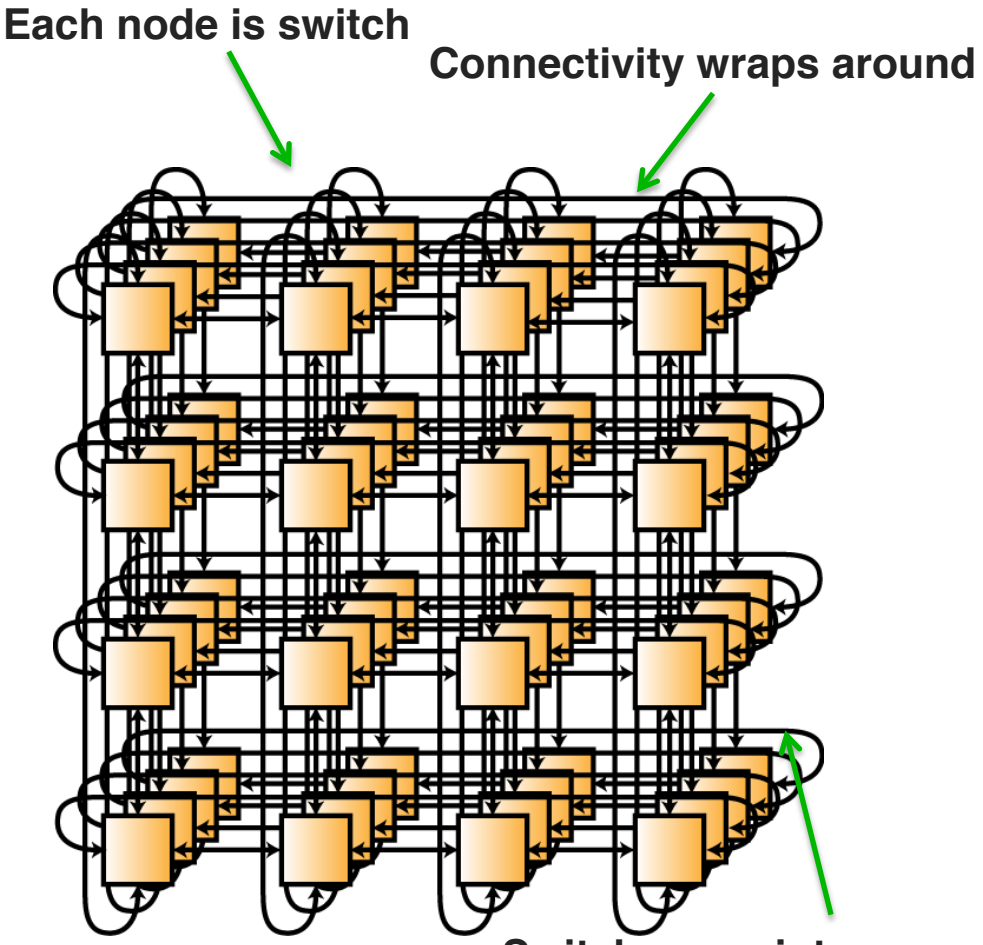
Isolation Bases

Hot Aisle Containment

64x36 port Juniper Ethernet Edge switches

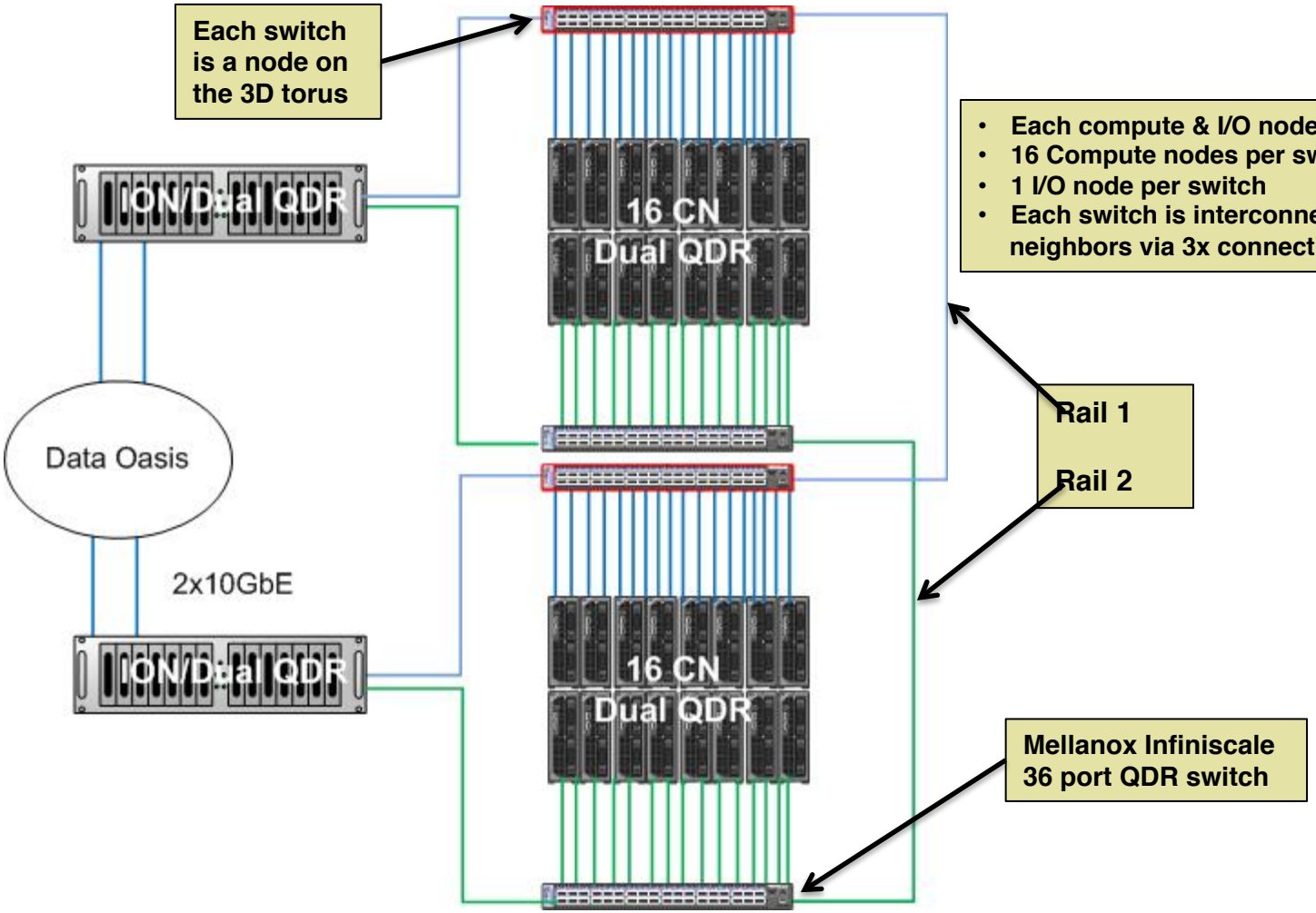
64 I/O nodes
4 Racks, 16 nodes/rack
64 TB/rack

Gordon Architecture: 3D Torus of Switches

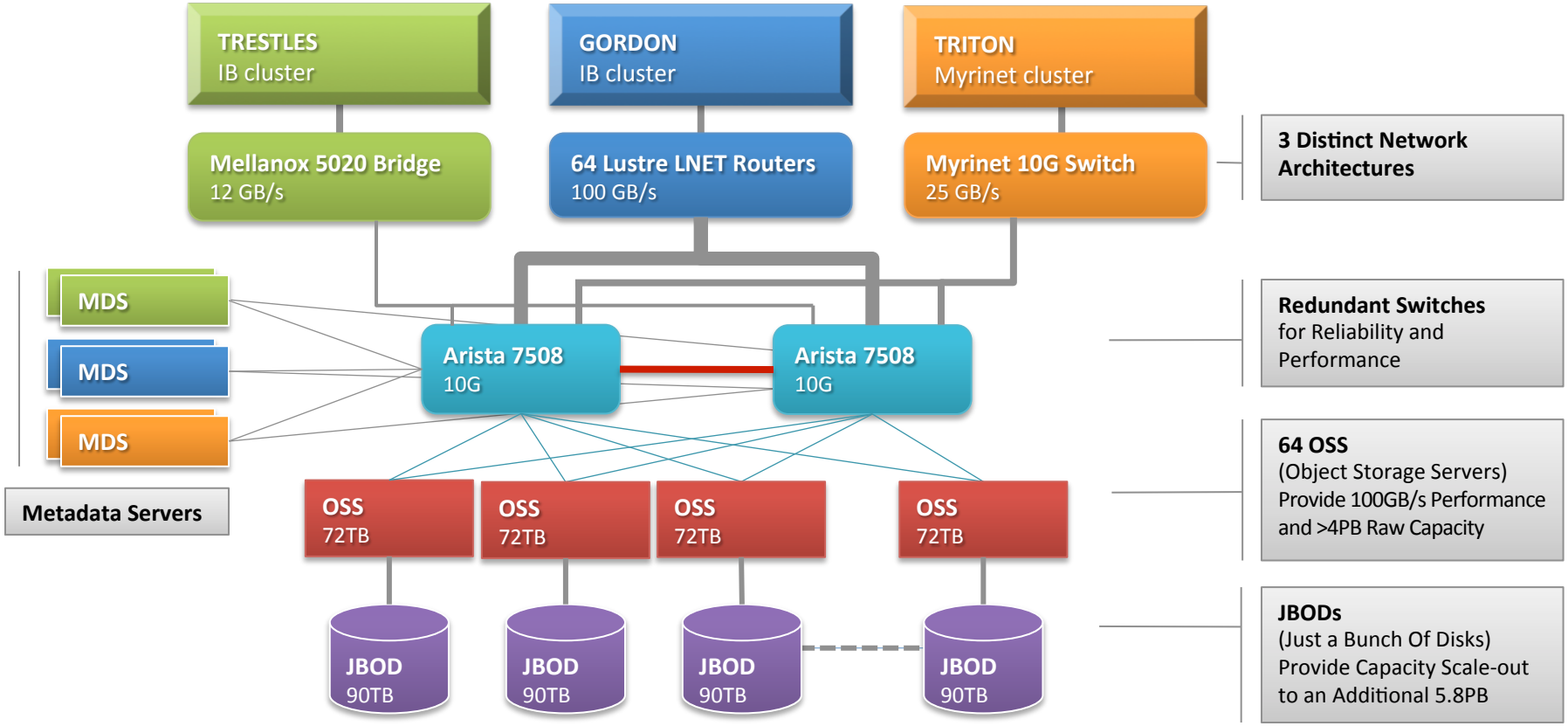


- Switches are connected in 4x4x4 3D torus
- Linearly expandable
- Short Cables- Fiber Optic cables generally not required
- Lower Cost :40% as many switches, 25% to 50% fewer cables
- Works well for localized communication
- Fault Tolerant within the mesh with 2QoS Alternate Routing
- Fault Tolerant with Dual-Rails for all routing algorithms
- Two rails – i.e., two complete tori with 64switch nodes in each torus
- Maximum of 6 hops

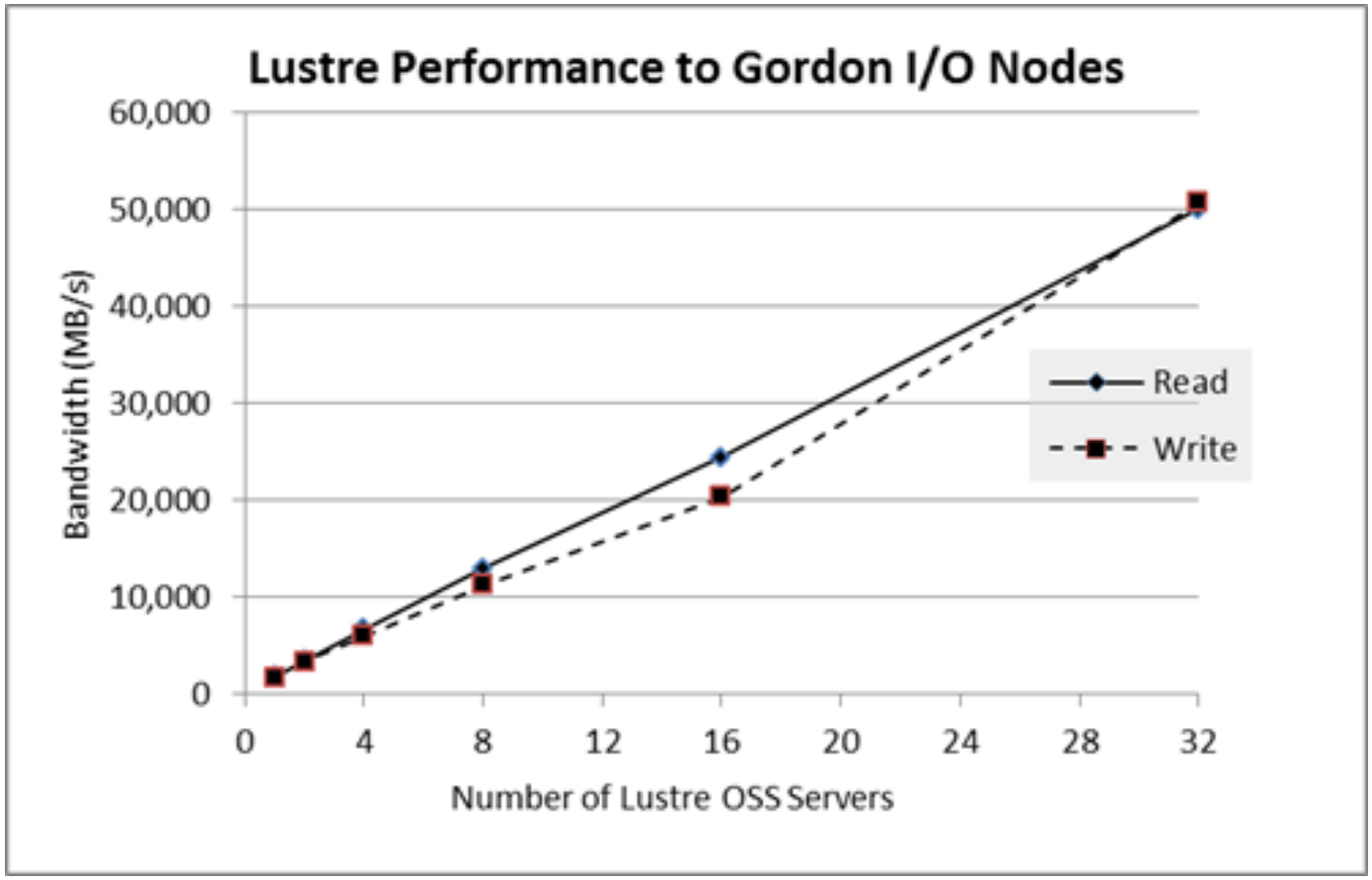
Torus Node IB Networking



Data Oasis Heterogeneous Architecture
Lustre-based Parallel File System



Data Oasis Performance



MVAPICH2 on Gordon

- MVAPICH2 [Current version is 1.9] is the default MPI implementation on Gordon.
- Compiled with --enable-3dtorus-support flag. Multi-rail support is also in place.
- LIMIC2 [Current version on system is 0.5.6]
- SSDs on Gordon are in I/O nodes. Exported to the compute nodes via iSER. Rail 1 (mlx4_1) is used for this part.
- I/O nodes also serve as lustre routers. Again I/O traffic is going on rail 1 (mlx4_1).
- Given I/O traffic, both to lustre and SSDs (local scratch) can saturate rail 1, default recommendation is to run MVAPICH2 with one rail [MV2_IBA_HCA=mlx4_0, MV2_NUM_HCAS=1]

InfiniBand Bandwidth Performance

Half-duplex IO bandwidth for each of a Compute nodes QDR InfiniBand channels.

	IB half-duplex speed (MB/s)	
	Rail 0 (on-board)	Rail 1 (add-on)
Min	3,830	3,250
Max	3,883	3,380
Avg.	3,867	3,376
Std Dev.	7.971	9.508

Full duplex bandwidth between Compute nodes using a single QDR InfiniBand channel.

	IB full-duplex speed (MB/s)	
	Rail 0 (on-board)	Rail 1 (add-on)
Min	6,613	5,746
Max	7,515	6,457
Avg.	7,505	6,388
Std. Dev.	35.19	71.47

The add-on IB bandwidth performance is limited by the PCIe riser card design which is based on the Gen2 spec.

Latency Performance

- **The average (ping-pong/2) latency between pairs of compute nodes (total of 1024) was measured with the Intel ping-pong benchmark.**
- **This includes the software latency, driver latency, HCA firmware latency, and up to a maximum of five switches.**
- **The test was run from nodes on all switches to four random nodes throughout the torus ensuring that the maximum number of switch hops was included.**

	Latency (μs)	
	Rail 0	**Rail 1**
Min	1.03	1.67
Max	1.85	2.57
Avg.	1.44	2.16
Std. Dev.	.168	.177

Full-duplex Any-Any Compute Node IB Bandwidth

- **Full-duplex bandwidth between Compute nodes using two QDR InfiniBand channels operating in parallel.**
- **No switch congestion as only two nodes running the test at any given time.**

testNode#1:testNode#2	IB rate, MB/s
gcn-19-22:gcn3-47	10,251
gcn-19-64:gcn-5-54	10,400
gcn-2-71:gcn-4-21	10,424

Inter-switch Link Performance

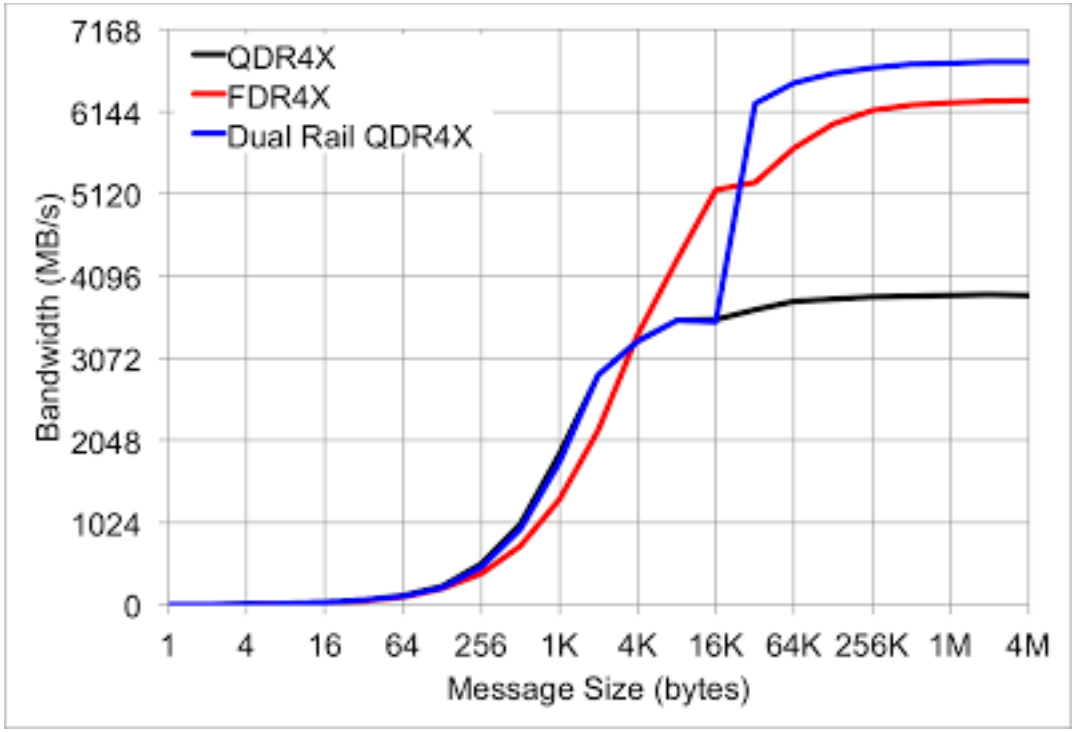
- Performance of inter-switch links was measured.
- With 3 links between each pair of switches and 2 rails, the expected half duplex performance is expected to be approximately three times the sum of rail 0 and rail 1 half-duplex rates = (3x(3.4+3.8)=21).

Aggregated measured bandwidth of inter-switch links

	IB rate, MB/s
Min	19,224
Max	20,737
Avg.	20,456
Std. Dev.	160.7

Dual Rail QDR vs FDR
OSU Bandwidth Test

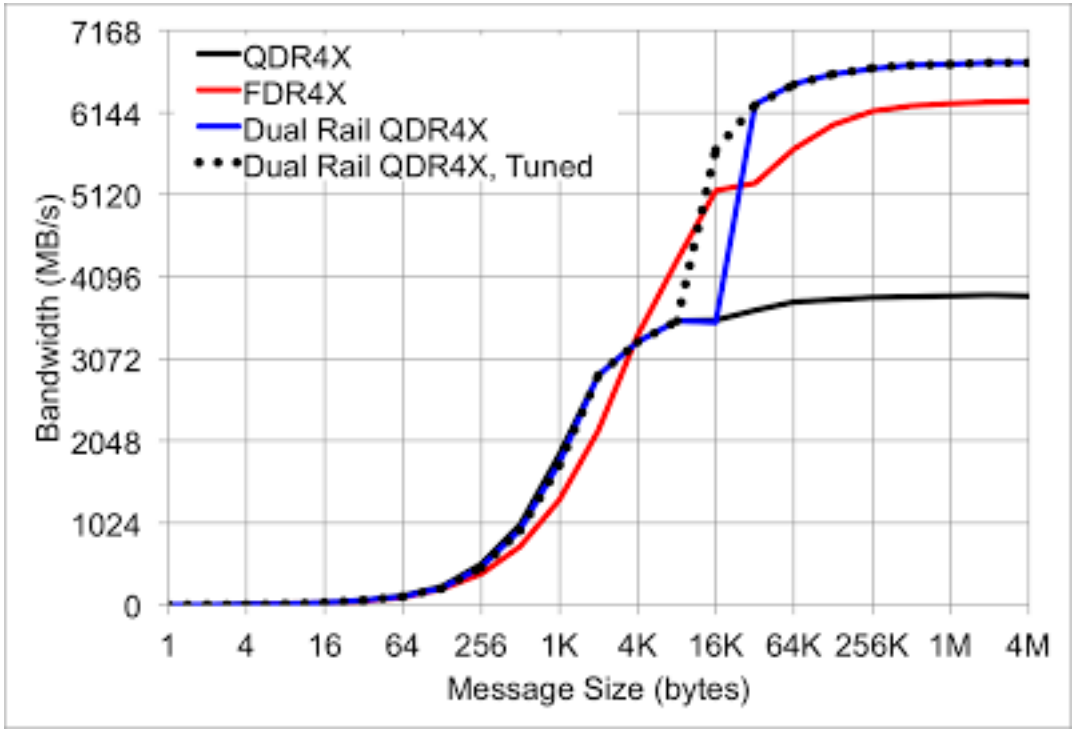
- **MVAPICH2 out of the box without any tuning**



*Tests done by Glenn Lockwood (SDSC)

Dual Rail QDR vs FDR
OSU Bandwidth Test

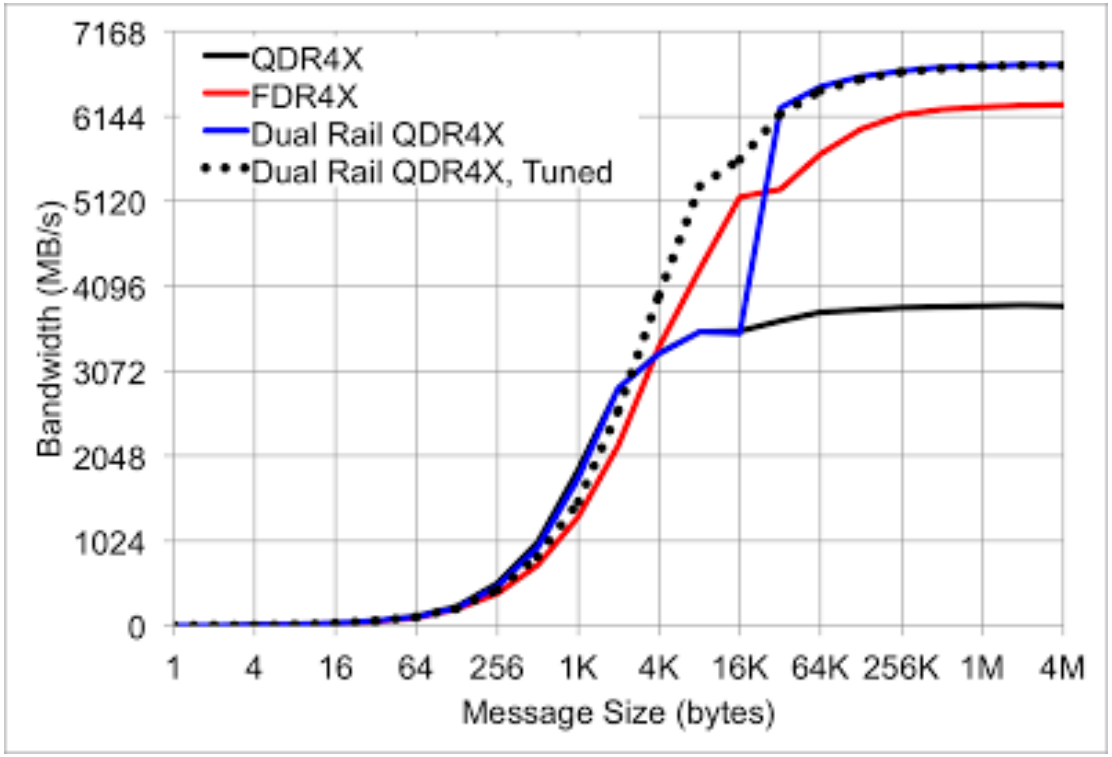
- **MV2_RAIL_SHARING_LARGE_MSG_THRESHOLD=8k**



*Tests done by Glenn Lockwood (SDSC)

Dual Rail QDR vs FDR
OSU Bandwidth Test

- **MV2_SM_SCHEDULING=ROUND_ROBIN**
- **In new version this is MV2_RAIL_SHARING_POLICY, default**



*Tests done by Glenn Lockwood (SDSC)

HPCC performance

A comprehensive set of HPCC benchmarks were run spanning four different conditions to determine the impact of core count and interconnect properties on performance.

1. **Single rail torus + 16 cores/node**
2. **Double rail torus + 16 cores/node**
3. **Single rail torus + 12 cores/node**
4. **Double rail torus + 12 cores/node**

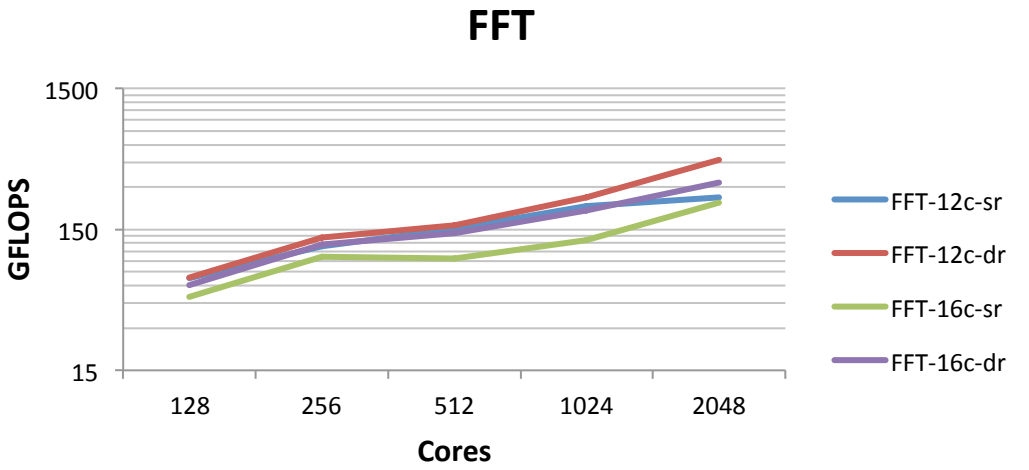
Also ran a 512-core benchmark using a non-contiguous set of nodes to investigate the impact of communications contention on performance

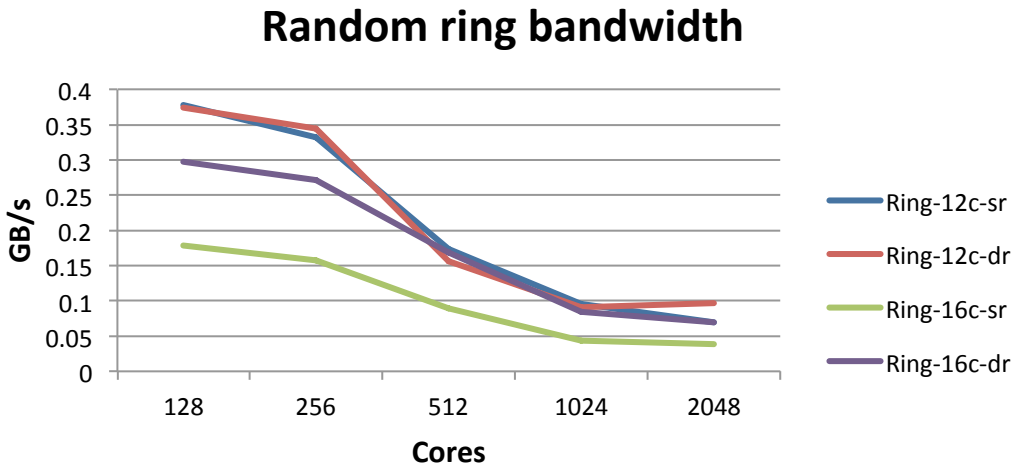
HPCC performance

Cores	G-HPL	G-PTRANS	G-FFTE	G-Rand Access	G- Triad	EP-Triad	EP-DGEMM	Rand Ring BW	Rand Ring Lat	HPL %peak
	Tflop/s	GB/s	Gflop/s	Gup/s	GB/s	GB/s	Gflop/s	GB/s	μs	%
128	2.25	27.00	68.3	0.904	598	4.67	19.34	0.374	4.3	84.5
256	4.50	57.00	131.4	1.595	1178	4.60	19.30	0.345	5.6	84.5
512	8.77	28.2	161.4	2.706	2350	4.59	19.41	0.156	7.1	82.4
1,024	17.02	83.9	254.6	4.316	4690	4.58	19.21	0.091	8.1	79.9
2,048	29.05	177.0	465.3	7.073	9359	4.57	19.38	0.097	8.6	68.2
16,160	284.5	219.7	679.0	15.751	55590	3.44	18.86	0.021	18.1	84.6

- **128-2048 core: dual-rail, 12 cores/node**
- **16160 core: single rail, 16 core/node**

HPCC performance – bandwidth sensitive tests

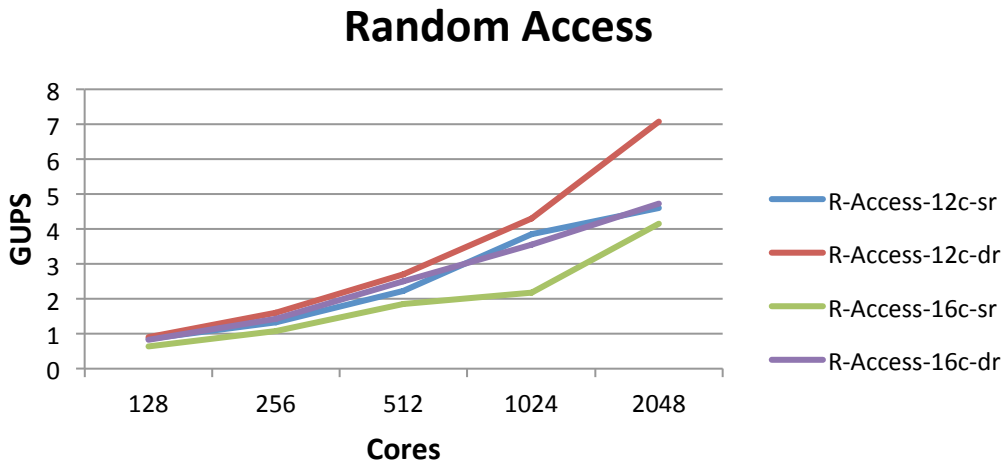


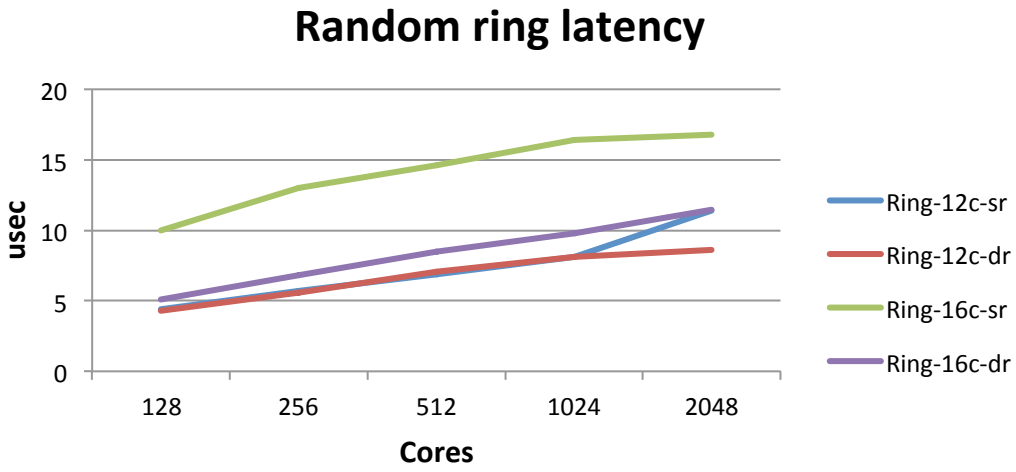


Bandwidth sensitive tests show better performance when using both rails.

Also see improvements from using 12 rather than 16 cores/node

HPCC performance – latency sensitive tests





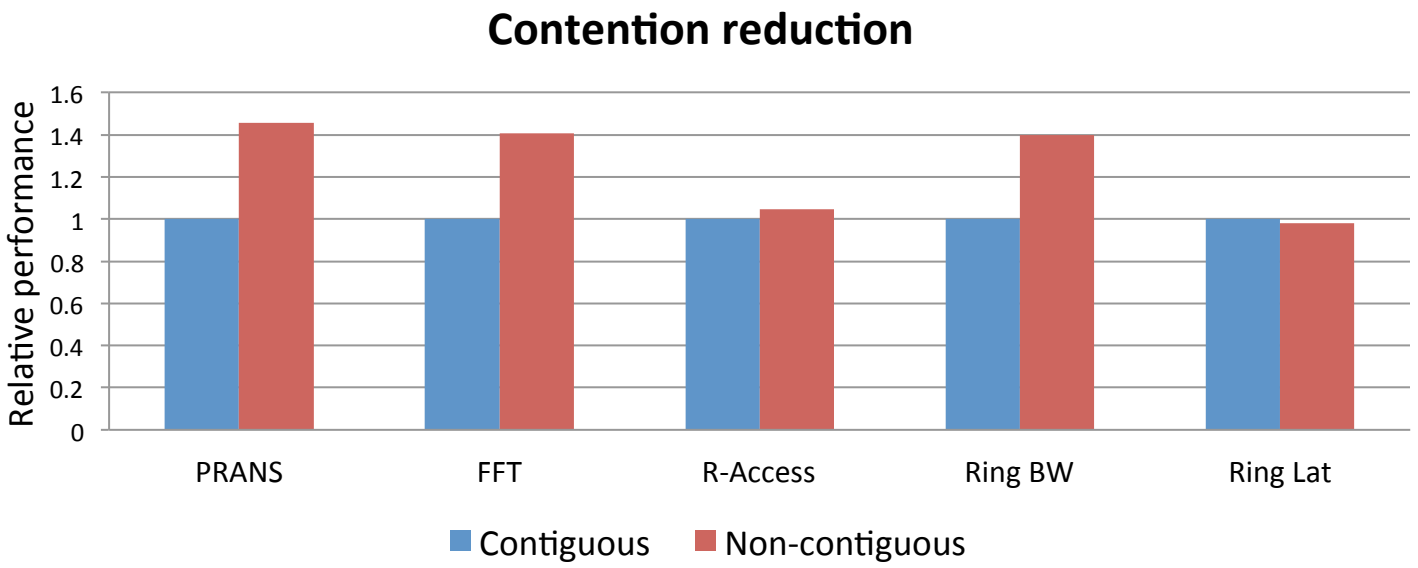
Latency sensitive tests show better performance when using both rails.

Also see improvements from using 12 rather than 16 cores/node

HPCC performance – contention

To assess the impact of communication contention, two 512-core HPCC runs were made in which the cores were:

1. **located on contiguous nodes within 2 neighboring subracks**
2. **spread across cluster using 3 compute nodes per switch (equal to number of inter-switch links)**



Note: The Gordon scheduler allows users to specify the number of switch hops to control job placement.

HPCC performance – summary

- Achieved 82-85% of peak for Linpack runs at smallest and largest core counts.
- Benchmarks dependent on memory bandwidth (STREAMS, DGEMM) do best when running with 12 cores/node.
- Benchmarks dependent on network performance (PTRANS, FFT, random ring/access) do best when using both rails.
- Minimizing contention results in 40% better performance on bandwidth sensitive HPCC benchmarks.
- Current SDSC scheduler allows users to specify maximum switch hops allowed for a job. Modifications for topology aware scheduling are being planned.

Summary

- Production Gordon stack features MVAPICH2 w/ --enable-3dtorus-support flag and dual rail support.
- Dual rail QDR performance competitive with FDR performance. MVAPICH2 environment variables such as MV2_RAIL_SHARING_LARGE_MSG_THRESHOLD and MV2_RAIL_SHARING_POLICY (earlier MV2_SM_SCHEDULING) can be used to tune performance.
- Gordon has oversubscription of switch to switch links. Spreading tasks to reduce contention can improve performance.
- Research ongoing to use topology aware scheduling to improve application performance on Gordon.
- Big Thank You to Dr. Panda's group! Gordon was the first production dual rail InfiniBand 3-D torus machine and the MVAPICH2 deployment was flawless out of the box.
- Acknowledgements : NSF Grant #0910847 (Gordon), #1147926 (SI2), and #0926692(STCI).