
An Introduction to the Gordon Architecture

Gordon Summer Institute & Cyberinfrastructure Summer Institute for Geoscientists

August 8-11, 2011

*Shawn Strande
Gordon Project Manager
San Diego Supercomputer Center*

Gordon Design Partnership



Funding and Oversight from the Office of
Cyberinfrastructure



Design, Deployment, Support, Management



System Integrator



Sandy Bridge and Westmere processors, flash
memory, compute node motherboards



vSMP Foundation Memory Aggregation Software



3D Torus Subnet Manager, IB Switches, HCAs

Gordon Design Innovations

- **Intel Sandy Bridge Processor:** First NSF system to be deployed with Intel's next generation chip
- **Flash memory:** 300 TB of high performance Intel flash via 64 I/O nodes made available as a local resource to applications
- **vSMP Memory Aggregation Software:** Create large SMP machines that can be reconfigured based on demand
- **3D torus interconnect:** Coupled with the dual rail QDR network provides a cost effective, power efficient, and fault tolerant interconnect
- **Data Oasis:** 100GB/s, 4 PB Lustre file system

Gordon is an integrated, data intensive supercomputer built from commodity hardware and software.

We did a few things to reduce risk without taking the fun out of it

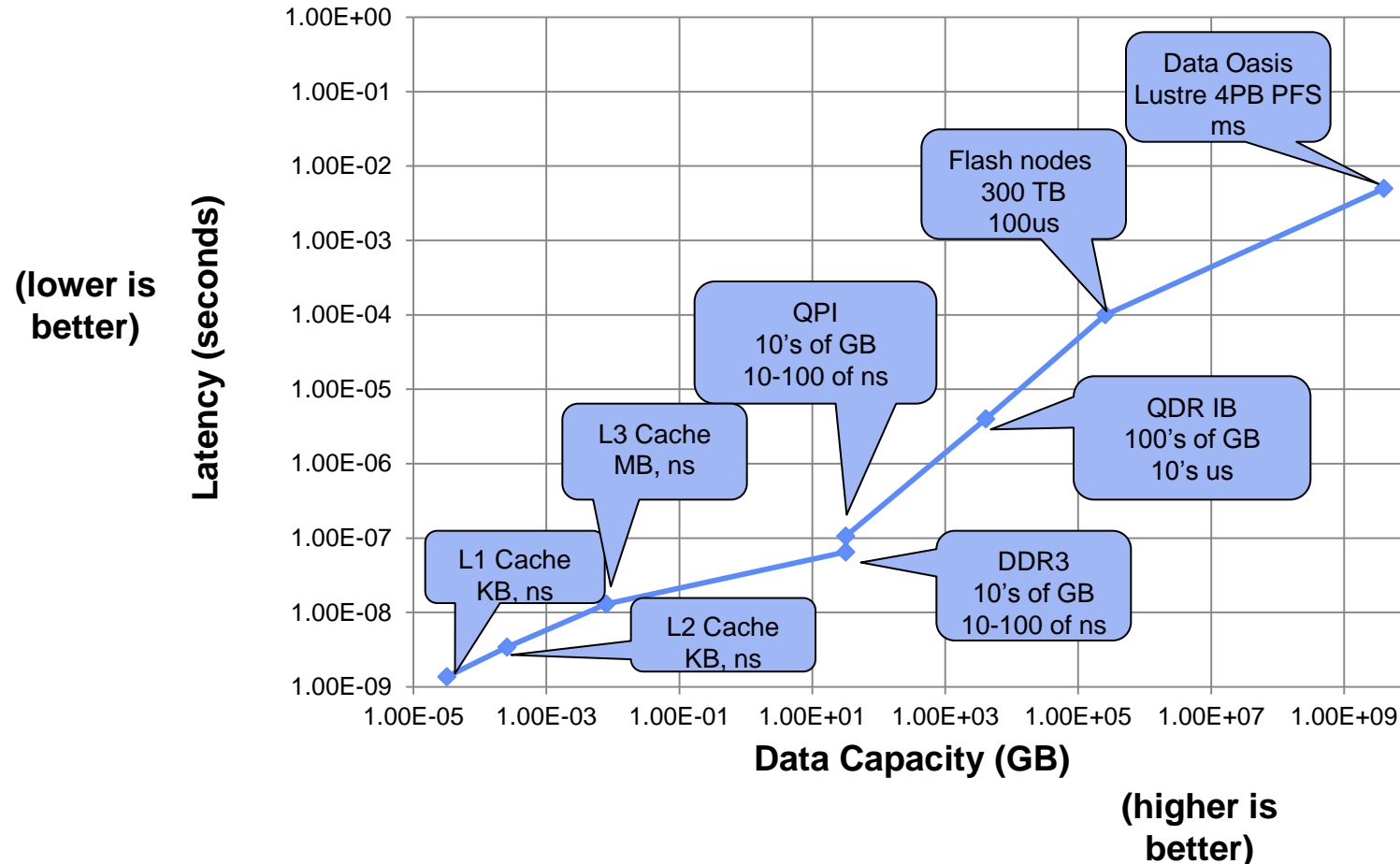
- **Deploy a prototype system (Dash)**
 - Test and verify flash performance of multiple vendors
 - Investigate file systems and protocols
 - Perform application benchmarks
 - Work with early users to explore data intensive applications
- **Use interim milestones to chart progress, and make course corrections as needed**
- **Get hardware as early as possible**
- **Test, analyze, reconfigure, repeat**
- **Work to maintain good partner relationships**
- **Stay focused on the big picture**
- **Cultivate and work with people who are: experts in their field – energetic - team players**

Gordon is not about FLOPS, but ...

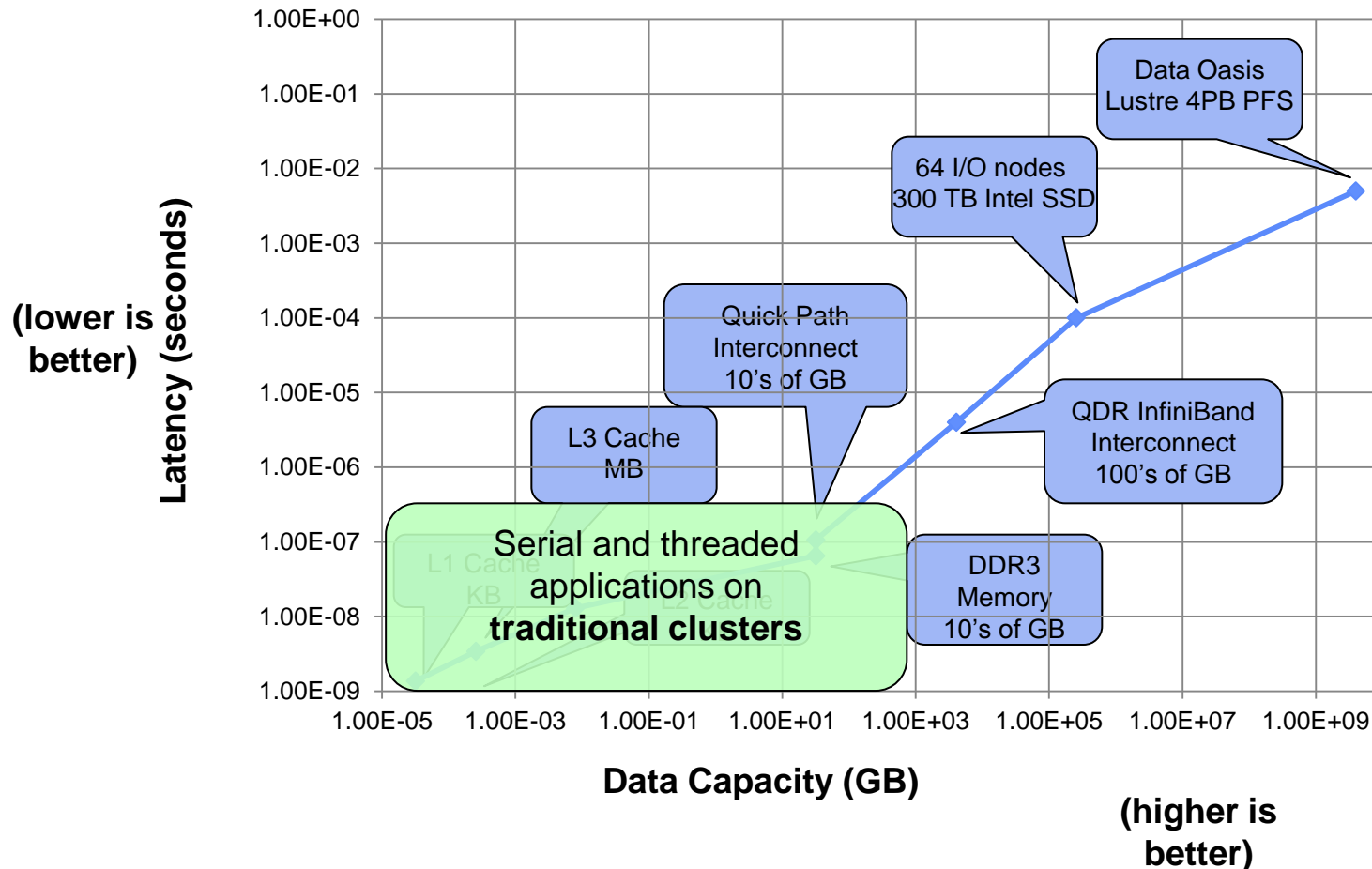
Rank	Site	Computer/Year Vendor	Cores	R _{max}	R _{peak}
1	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011 Fujitsu	548352	8162.00	8773.63
2	National Supercomputing Center in Tianjin China	Tianhe-1A - NUDT TH MPP, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C / 2010 NUDT	186368	2566.00	4701.00
38	Japan Atomic Energy Agency (JAEA) Japan	BX900 Xeon X5570 2.93GHz , Infiniband QDR / 2009 Fujitsu	17072	191.40	200.08
39	King Abdullah University of Science and Technology Saudi Arabia	Shaheen - Blue Gene/P Solution / 2009 IBM	65536	190.90	222.82
40	Shanghai Supercomputer Center China	Magic Cube - Dawning 5000A, QC Opteron 1.9 Ghz, Infiniband, Windows HPC 2008 / 2008 Dawning	30720	180.60	233.47
41	Government France	Cluster Platform 3000 BL2x220, L54xx 2.5 Ghz, Infiniband / 2009 Hewlett-Packard	24704	179.63	247.04
42	Taiwan National Center for High-performance Computing Taiwan	ALPS - Acer AR585 F1 Cluster, Opteron 12C 2.2GHz, QDR infiniband / 2011 Acer Group	26244	177.10	231.86
43	EDF R&D France	Ivanhoe - iDataPlex, Xeon X56xx 6C 2.93 GHz, Infiniband / 2010 IBM	16320	168.80	191.27
44	Swiss Scientific Computing Center (CSCS) Switzerland	Monte Rosa - Cray XT5 SixCore 2.4 GHz / 2009 Cray Inc.	22032	168.70	211.51

A conservative estimate puts Gordon in the top 50 on the Top 500 list

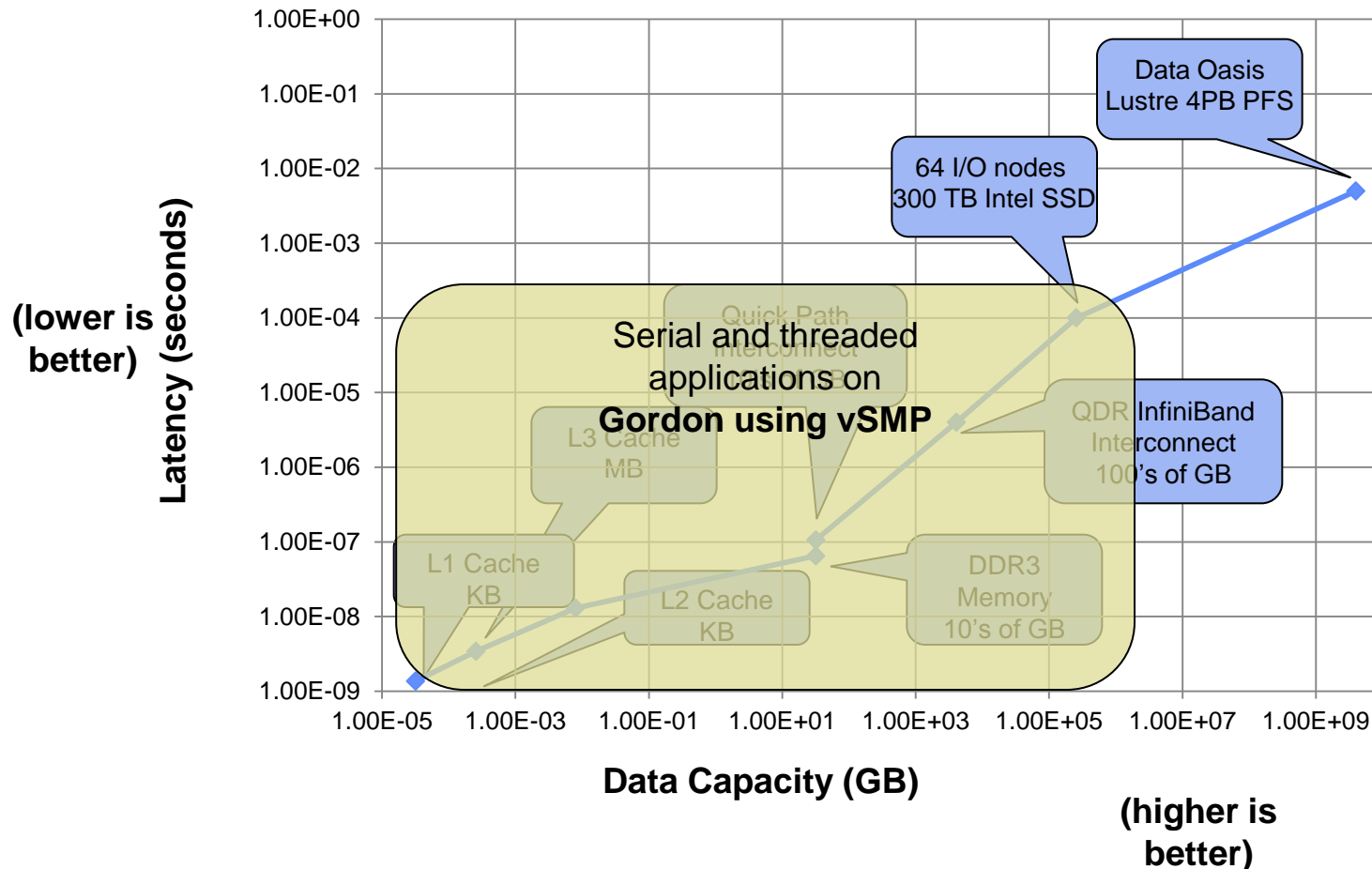
Access to Big Data Comes with a Latency Penalty



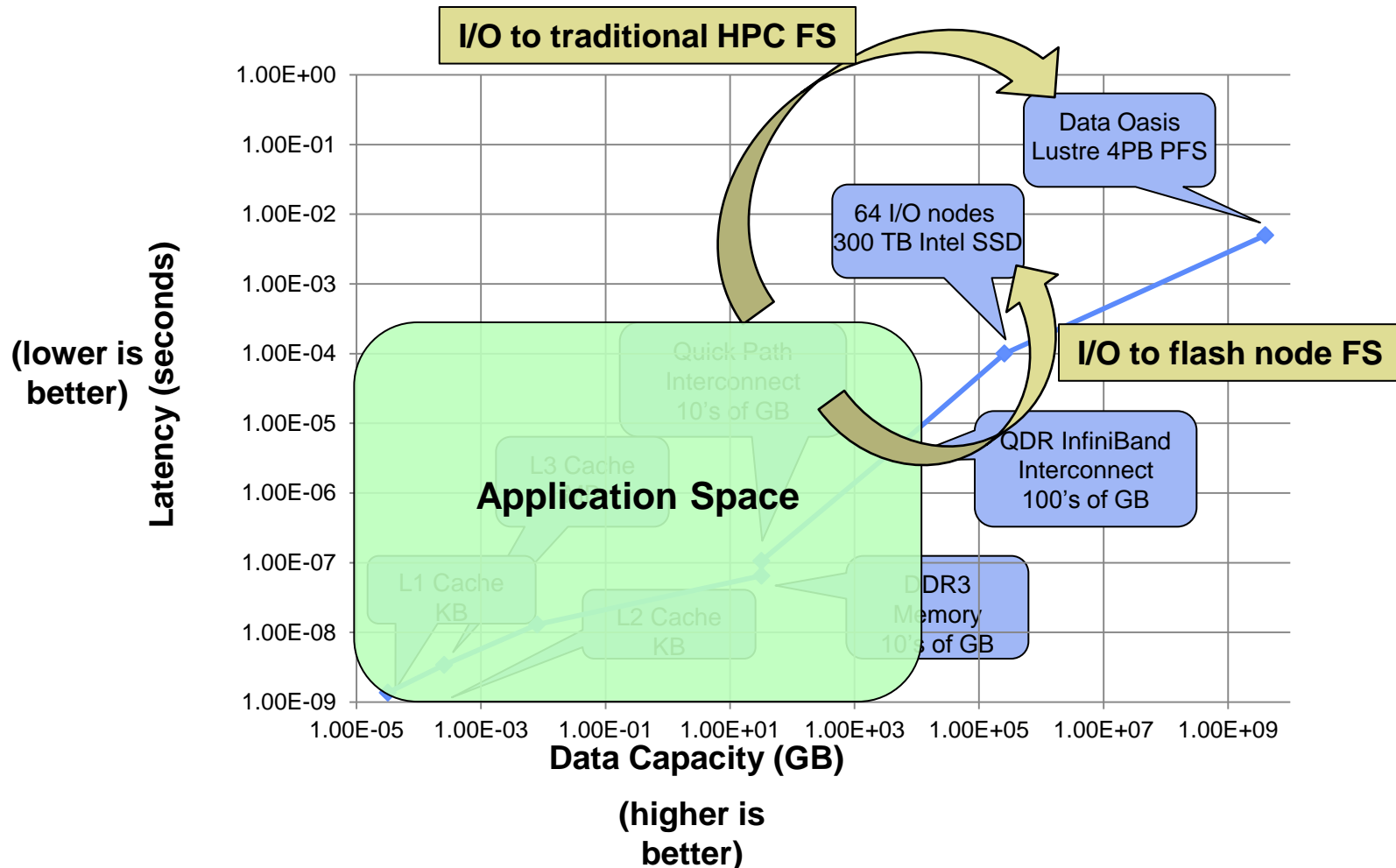
Gordon Architecture Bridges the Memory Capacity Gap



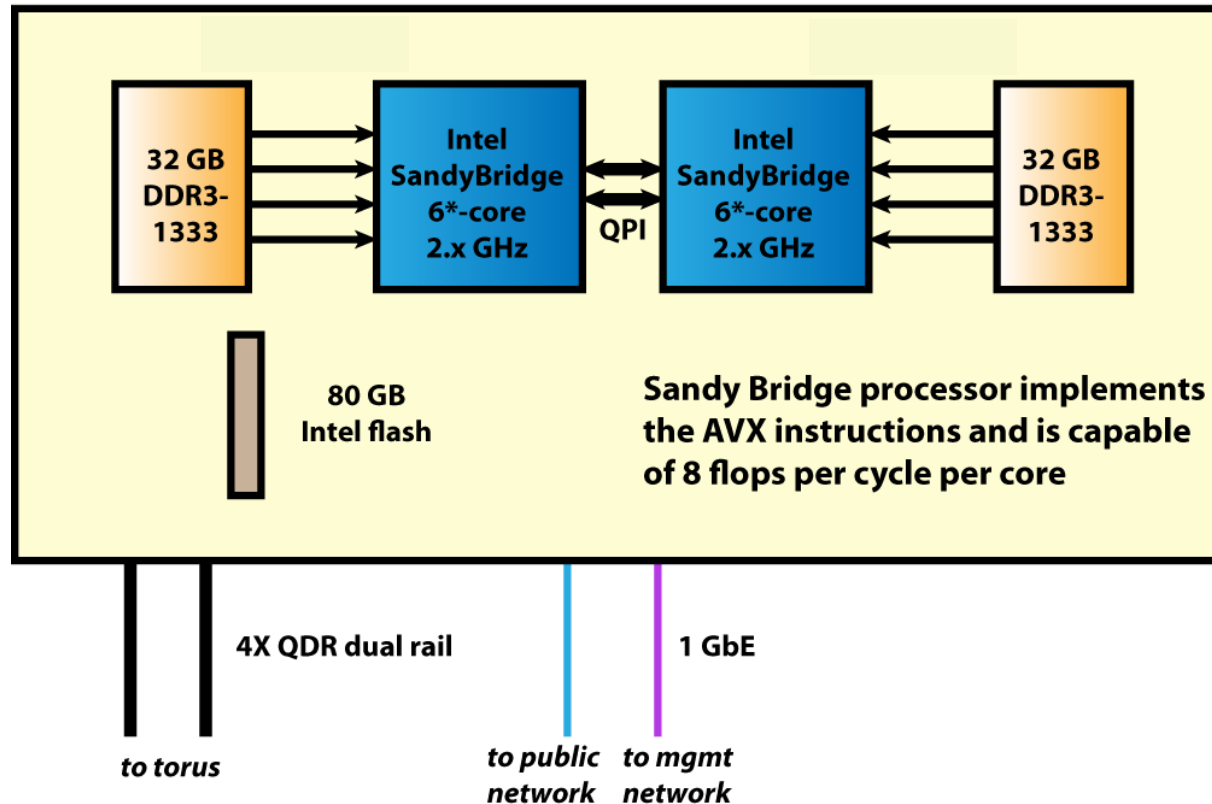
Gordon Architecture Bridges the Memory Capacity Gap



Gordon Architecture Bridges the I/O Latency Gap



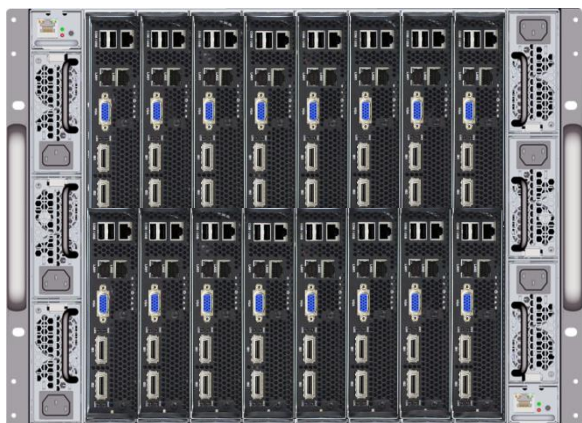
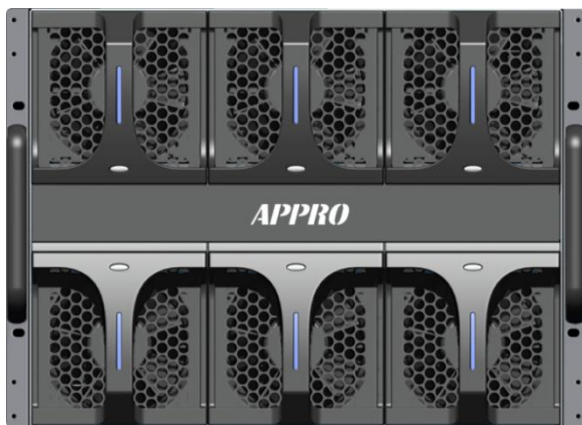
Dual – Socket Compute Node Based on the Intel Sandy Bridge Processor



summary
64 GB DRAM
12+ cores
2.0+ GHz
80 GB flash
system disk

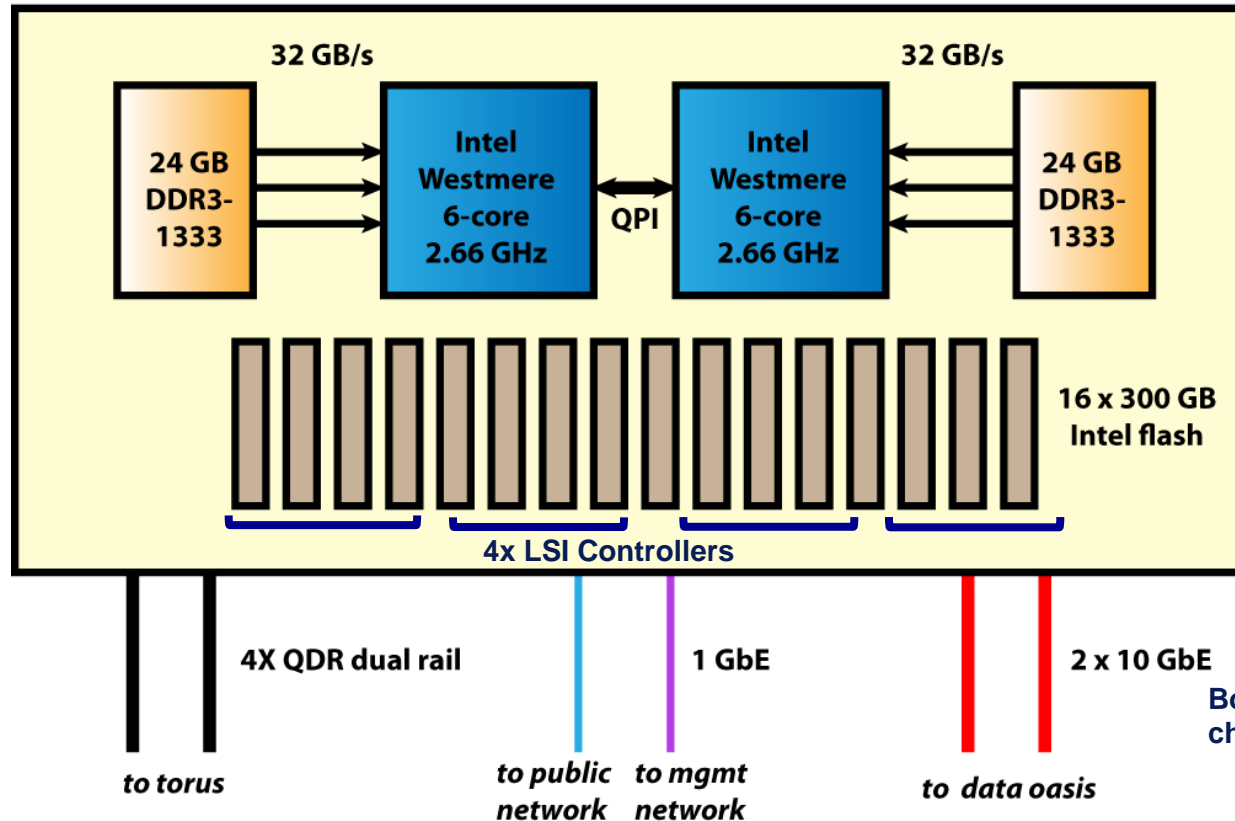
For more information on AVX, see <http://software.intel.com/en-us/avx/>

Gordon Compute Node Subrack



- Intel Motherboard
- 8RU Subrack
- 16x 2P Intel Sandy-Bridge Blades per subrack; 4 subrack/48U rack
- Six high-efficiency 1625W hot-swappable power supplies in N+1 configuration
- Support for dual-redundant platform management modules
- Six hot-swappable, redundant fan modules
- Shared reduces power consumption by up to 20W per blade over previous design

Gordon I/O Node Based on a Dual-socket Westmere Processor



summary

- 48 GB DRAM
- 12 cores
- 2.66 GHz
- 4.8 TB flash
- 2x 80GB SSD system

Gordon IO Node



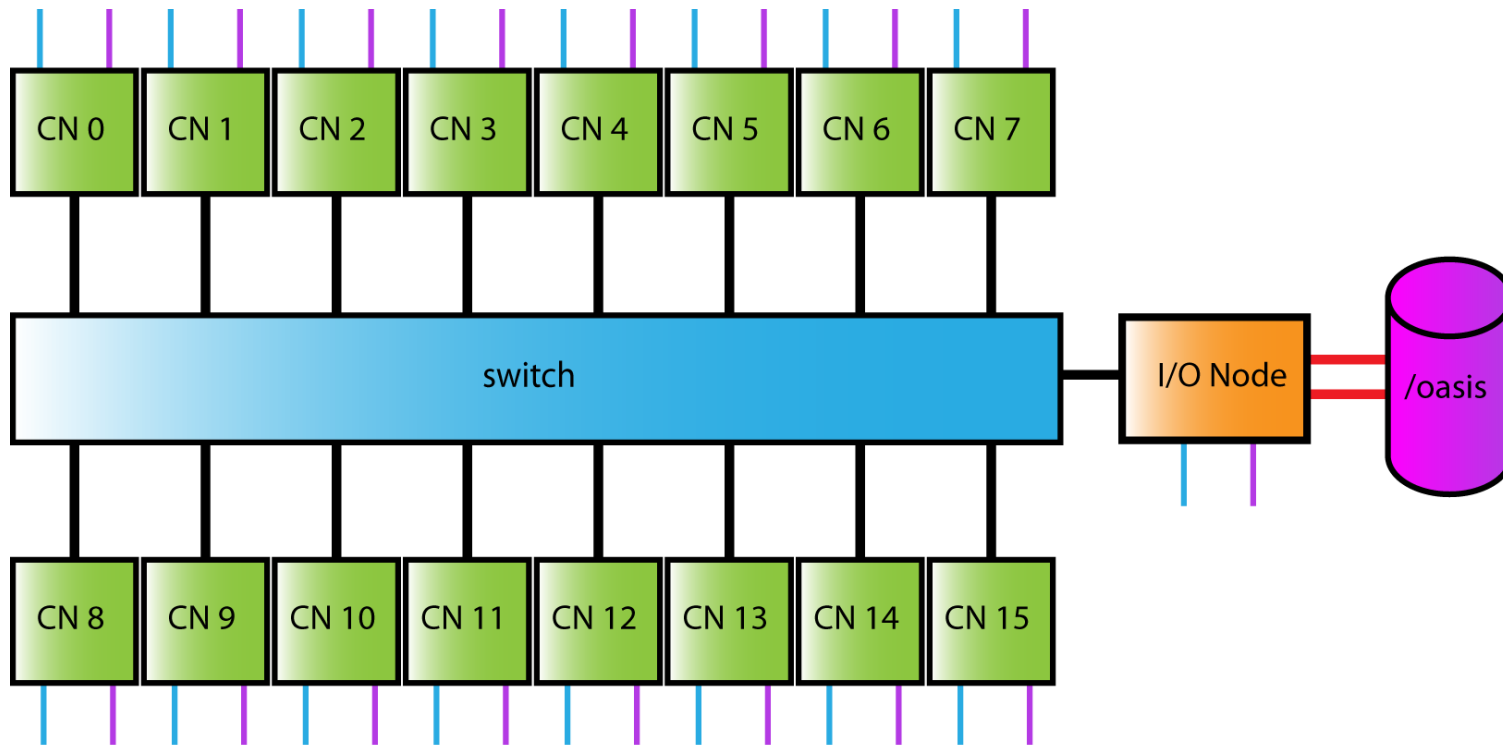
- Dual Intel® Xeon® Westmere processors
- 6 cores/processor
- 48 GB DDR3 1333 MHz RAM
- Seven (7) PCI-E 2.0 x8 slots
- 16 x 3.5" Hot-swappable SAS / SATA Bays
- 16 x Intel eMLC SSD's
- 2 x Intel Postville 80 GB SSD (system)
- 4 x LSI 9211-4i Controllers (1 for every 4 SSDs)
- 2 x Mellanox IB HCA Cards, QDR
- Broadcom BCM57711 Dual Port 10GbE
- Integrated IPMI 2.0 with Dedicated LAN



The I/O node was an SDSC/Appro design partnership. Extensive testing of all components occurred before Unit #1 was built and accepted.

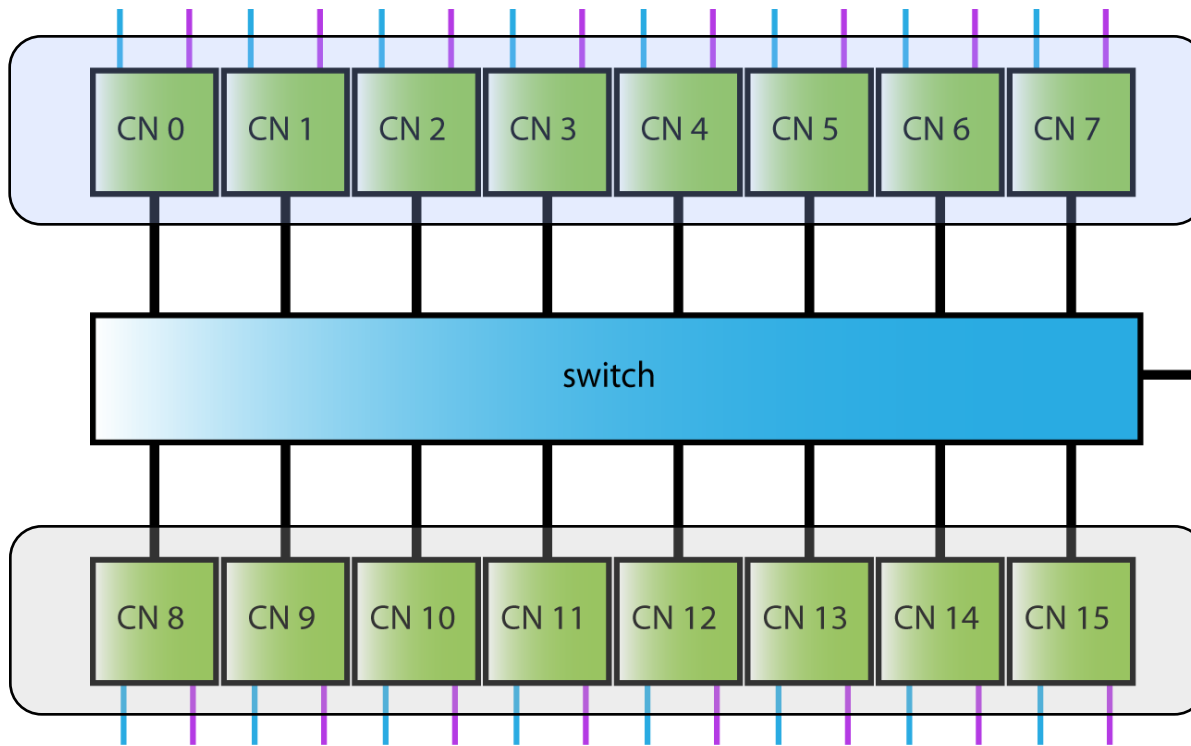
Compute and I/O node Configuration

Single Rail



- 4X QDR InfiniBand (32 Gb/s actual data rate)
- 10 GbE
- 1 GbE (to public network)
- 1 GbE (to management network)

Options for specifying flash resource

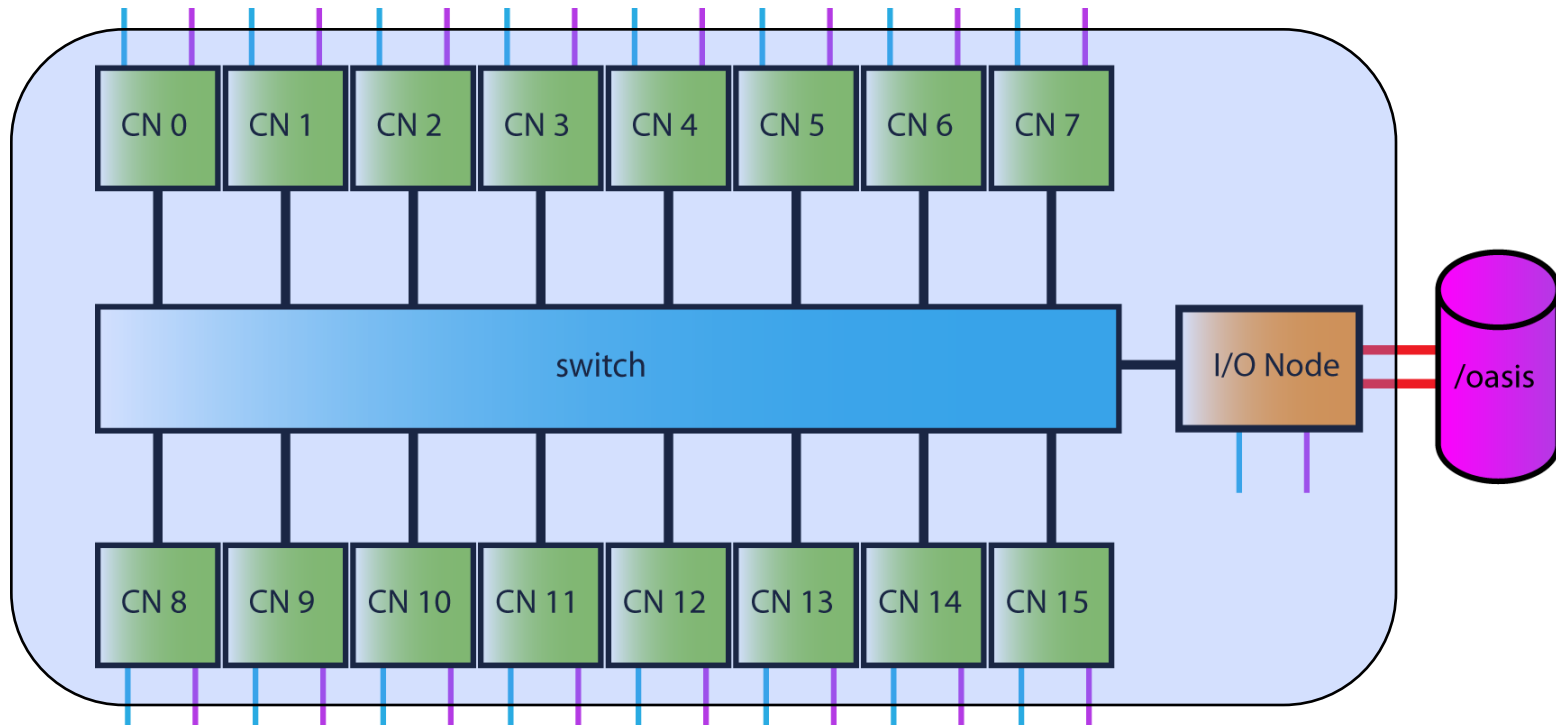


Asking for more
1st user requests 8 compute nodes and 4.8 TB flash

Asking for less
2nd user requesting 8 compute nodes and no flash can use other 8 nodes on this switch

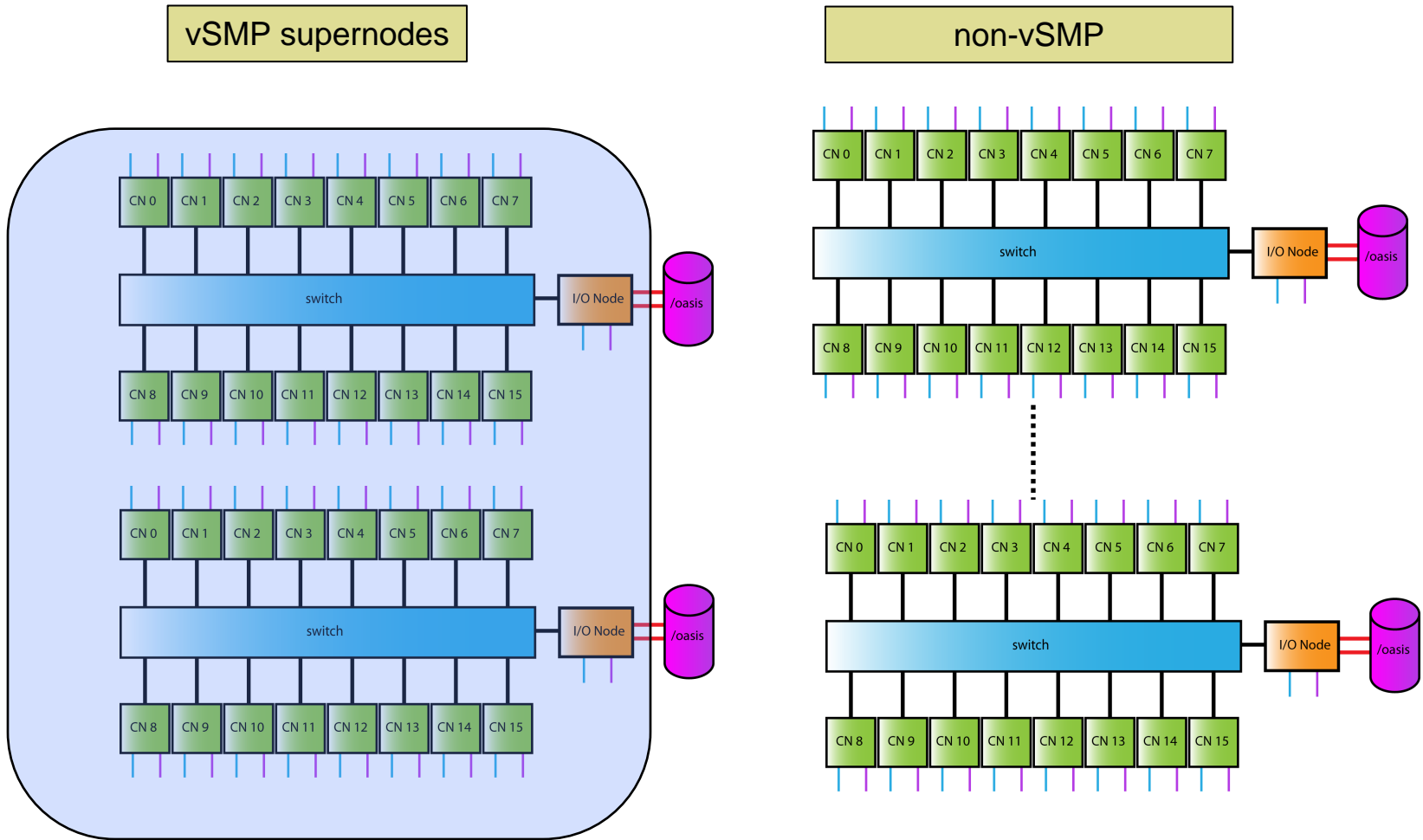
- 4X QDR InfiniBand (32 Gb/s actual data rate)
- 10 GbE
- 1 GbE (to public network)
- 1 GbE (to management network)

vSMP node configured from 16 compute nodes and one I/O node



To user, logically appears as a single, large SMP node with 1TB of memory

vSMP and non-vSMP Environments will Co-Exist



Flash Memory



OCZ Patriot G Skill Super Talent



Intel



Flash Drives are a Good Fit for Data Intensive Computing

	Flash Drive	Typical HDD	Good for Data Intensive Apps
Latency	< .1 ms	10 ms	✓
Bandwidth (r/w)	250 /170 MB/s	100 MB/s	✓
IOPS (r/w)	35,000/ 2000	100	✓
Power consumption (when doing r/w)	2-5 W	6-10 W	✓
MTBF	1M hours	1M hours	-
Price/GB	\$2/GB	\$.50/GB	-
Endurance	2-10PB	N/A	✓
Total Cost of Ownership	*** The jury is still out ***		



Apart from the differences between HDD and SSD it is not common to find local storage “close” to the compute. We have found this to be attractive in our Trestles cluster, which has local flash on the compute, but is used for traditional HPC applications (not high IOPS).

Flash Glossary of Terms

Term	Definition
NAND	Physical silicon MOSFET storage array. The flash storage media.
SLC: Single-level cell	Type of NAND storage device. Lower storage density for the same price. Higher endurance and performance than MLC. 1 bit of data per cell.
MLC: Multi-level cell	Type of NAND storage device. Higher storage density for the same price. Lower endurance and performance than SLC. 2 bits of data per cell.
eMLC: Enterprise MLC	Form of eMLC that uses high quality NAND, additional controller software, and overprovisioning to achieve higher performance and endurance.
IOPS: I/O operations per second	Key storage benchmark that is relevant for data intensive applications. The ability to sustain high IOPS provides performance for applications that exhibit high random access data patterns.
Wear Leveling	The process of distributing write operations uniformly over all of the blocks in the flash memory chips to preserve the life of the NAND. Controlled by the SSD controller software.
Write Amplification Factor	Ratio of actual erase operations to the minimum required to erase the data. Greater than 1 is bad. Less than 1 is possible with compression (though most HPC data is not compressible)
Overprovisioning	Space on the flash drive not available to the user that is used for garbage collection, wear leveling, and containing bad blocks. Overprovisioning can increase endurance.
Program/Erase (P/E) cycle	A flash memory cell must be erased before it can be written to. This causes cell wear out and is what sets the endurance of an SSD. Minimum erase size for SSD's is a block.
Endurance	How much data can be written to an SSD before it ceases to function properly. Set by the P/E cycle count. Lower for MLC flash.

Gordon Layout: 21x48U Racks



Hot Aisle Containment

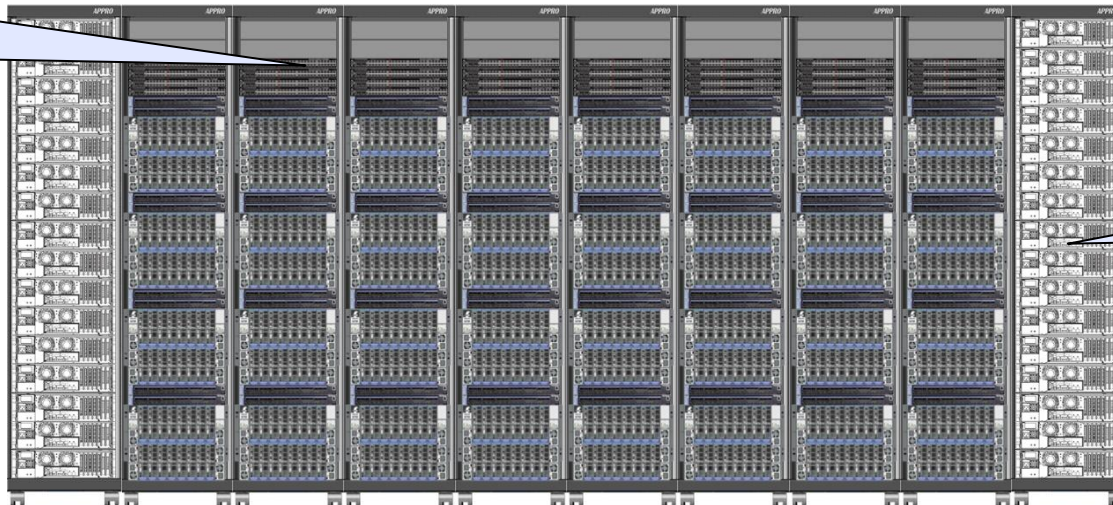
Isolation Bases

128x36 port
Mellanox
InfiniScale
QDR

Service Nodes
NFS, Login,
Rocks Mgmt
Core Ethernet
switches

Compute Nodes
16 racks
64 nodes/rack

128x36 port
Juniper
Ethernet Edge
switches

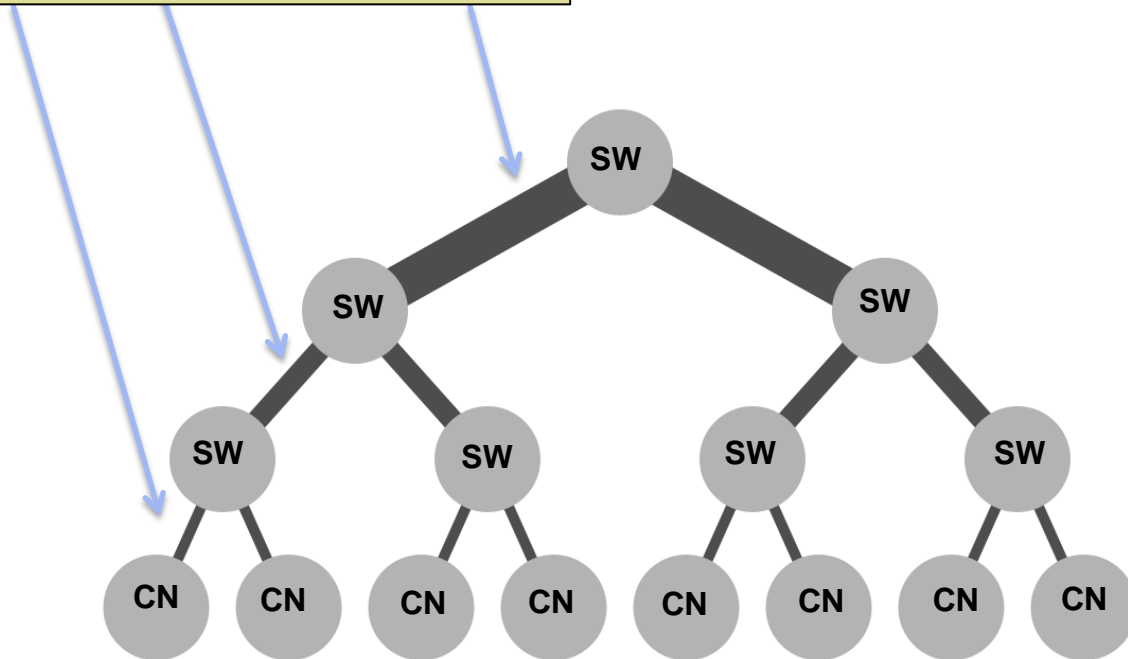


64 I/O nodes
4 Racks, 16 nodes/rack
64 TB/rack

Networking

Fat Tree Interconnects are excellent for some applications, though not a Gordon design point

Links get “fatter” as you go up the tree



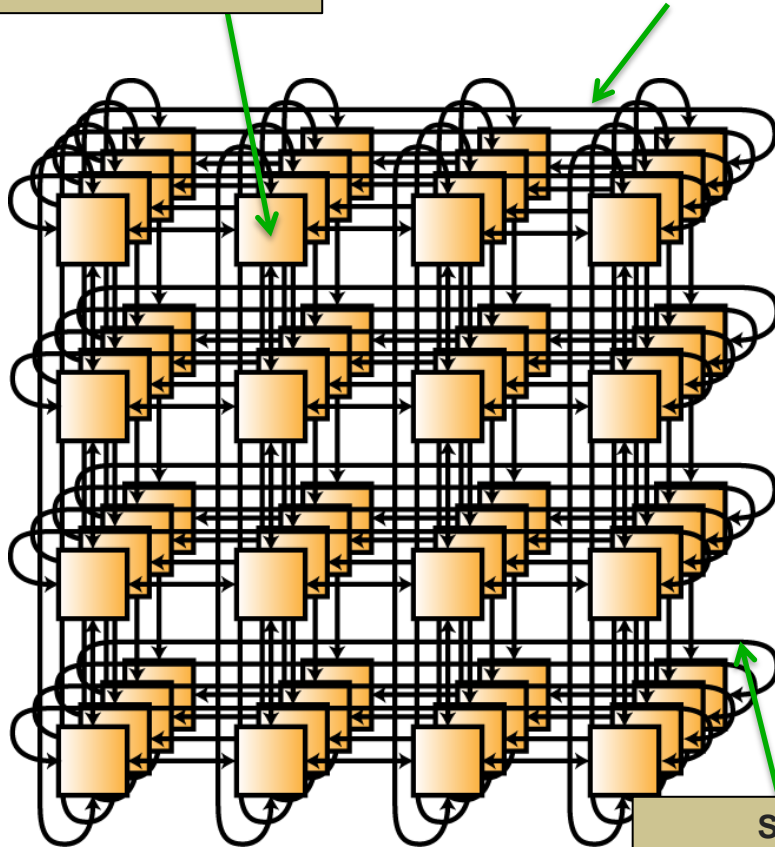
- Every node has equal access bandwidth to every other node in the cluster.
- This is great for large scale applications where nodes do a lot of communication with each other, e.g., MPI.

CN – Compute Node
SW - Switch

Gordon's 3D torus interconnect is ideal for data intensive computing

Each node is switch

Connectivity wraps around



Switches are interconnected by 3 links in each +/- x, y, z direction

Many data intensive applications do not require large number of compute nodes.

Gordon switches are connected in 4x4x4 3D torus.

The links are Quad Data Rate (QDR) InfiniBand. Each link is 40Gbps peak.

There are two rails – i.e., two complete tori. 64 nodes in each torus.

MPI can take advantage of dual rail to provide 2x the bandwidth

Maximum of six hops to get from one node to furthest node in cluster

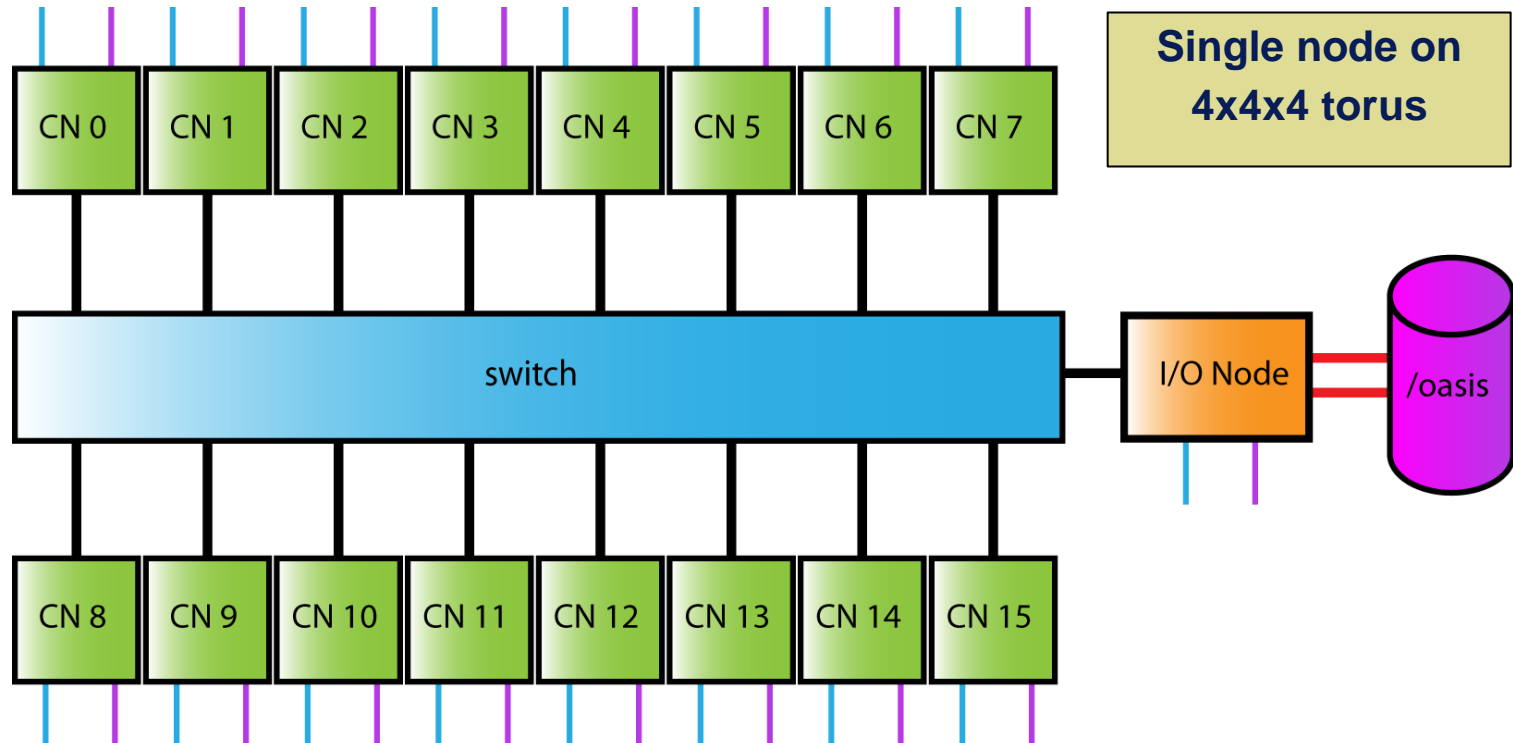
Fault tolerant, requires up to 40% fewer switches and 25-50% fewer cables than other topologies.

Scheduler will be aware of torus geometry and assign nodes to jobs accordingly.

Benefits of the dual rail 3D torus

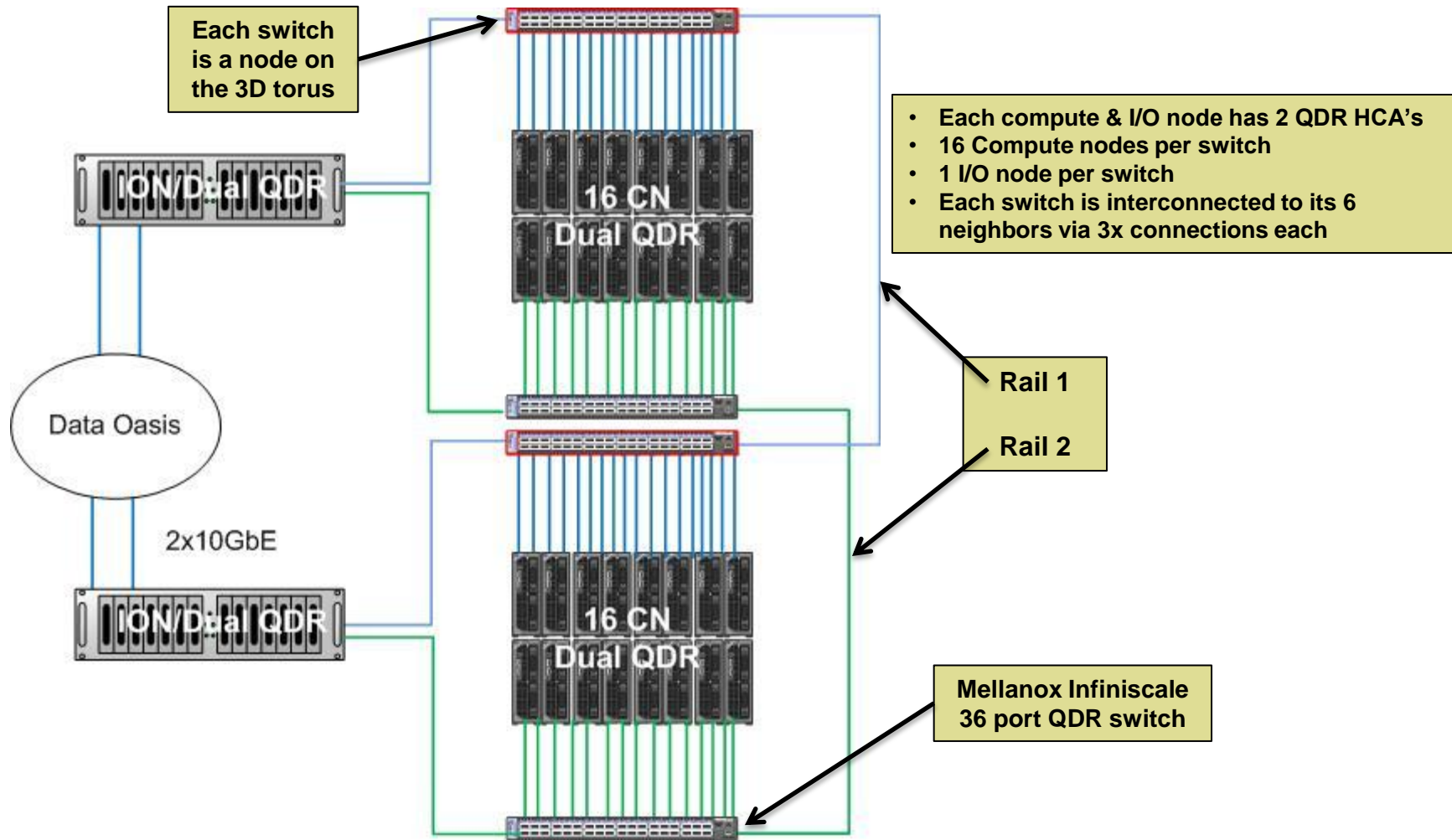
- **Linearly Expandable**
- **Simple wiring pattern**
- **Short Cables; Fiber Optic cables generally not required**
- **Lower Cost**
 - **40% fewer switches and cables – O(\$500-750K savings)**
- **Works well for localized communication, particularly in capacity environments**
- **Lower power and cooling costs**
- **Fault Tolerant within the mesh with alternate routing**
- **Fault Tolerant with dual-rails for routing algorithms**
- **QDR InfiniBand Network Hardware**
 - **4GB/sec QDR links**
 - **Switch to Switch 3 x 4Gb/sec**
 - **Total of 12 GB/sec in Each Direction (N, S, E, W, U, and D)**
 - **Compute Node to Switch 2 x 4GB/sec (1 per rail)**
 - **3.2 GB/sec usable Bandwidth per connection in each direction**
 - **Dual HCAs with 8X Gen 3 PCI-Express on Compute Nodes**

Simplified single rail view of Gordon connectivity

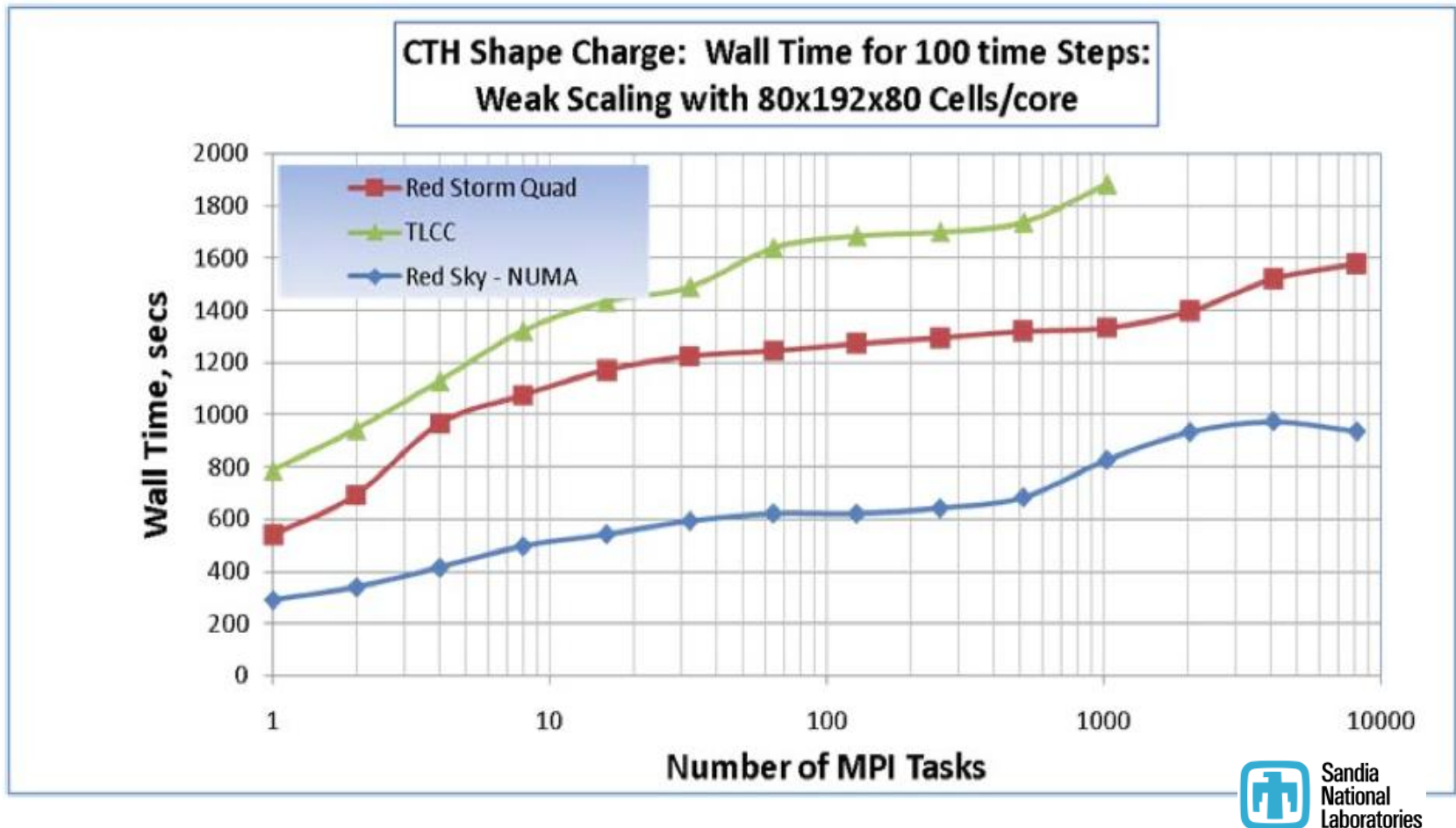


- 4X QDR InfiniBand (32 Gb/s actual data rate)
- 10 GbE
- 1 GbE (to public network)
- 1 GbE (to management network)

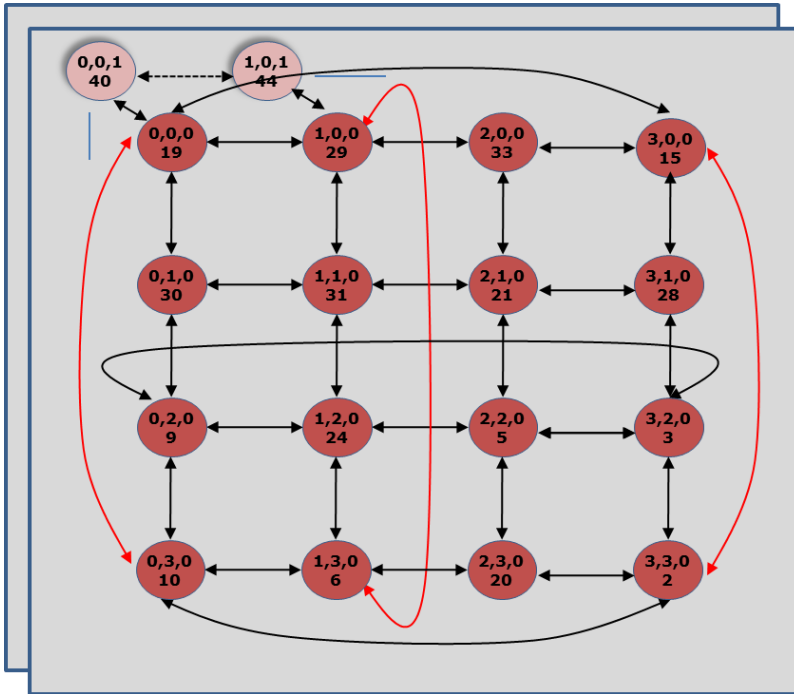
Switch Node IB Networking Configuration



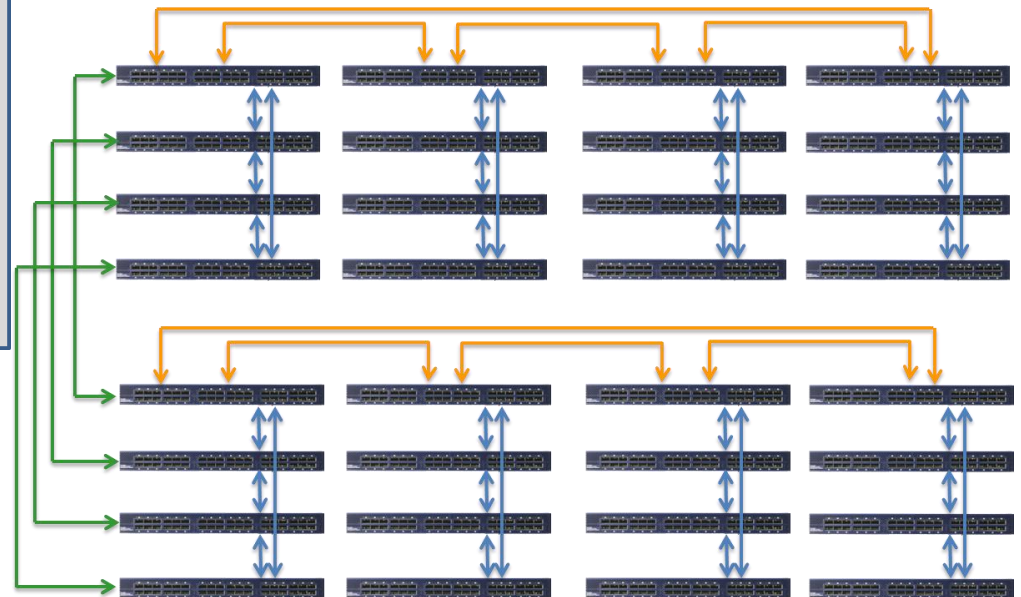
Sandia 3D Torus Performance (Red Sky)



Prototype 4x4x2 torus



- 32 Mellanox InfiniScale IB Switches (4x4x2)
- 32 Appro Greenblade Servers
- Open Fabrics Alliance (OFED) network software based on work at Sandia National Labs and supported by Mellanox



Live demo at SC'10

3D Torus Testing



- Lab Configuration 4X4X2 Single-Rail 3D Torus
- 32 Appro-provided Compute Nodes with 1 Compute Node per Switch Node
- Intel MPI Benchmark Tests Conducted
 - PingPong
 - PingPing
 - Sendrecv
 - Exchange
 - Allreduce
 - Reduce
 - Reduce_scatter
 - Allgather
 - Allgatherv
 - Gather
 - Gather
 - Scatter
 - Scatternv
 - Alltoall
 - Alltoallv
 - Bcast
 - Barrier

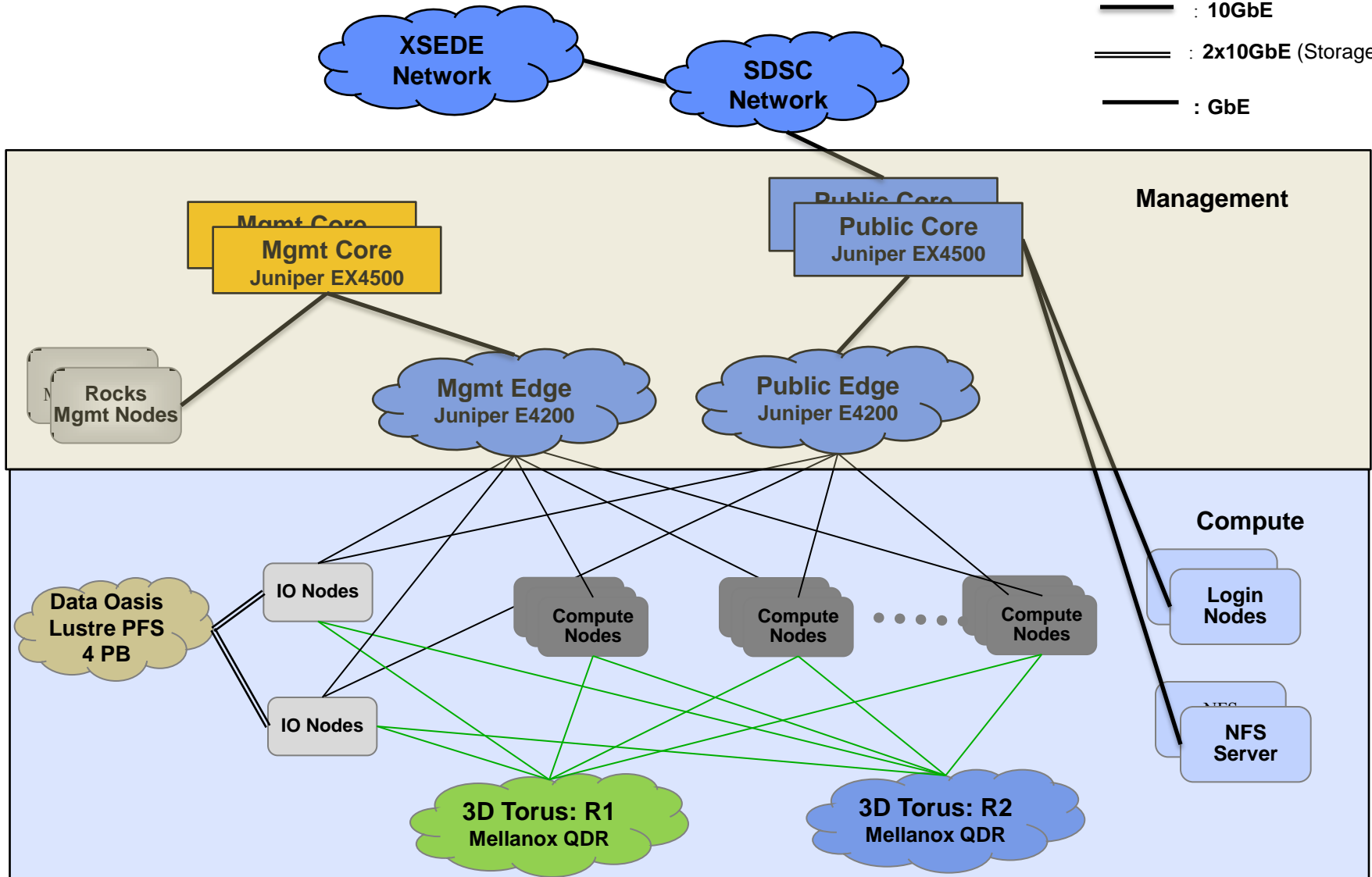
Tests

- Measurements with contention
 - Latch up - No Credit Loops or Latch up conditions were discovered
 - Measurements without contention
 - Latency - Max latency = 1.6 uSec + (5 * 133 nSec)= 2.265 uSec**
 - Bandwidth - 3.411GB/s**
- Failover
 - Links Removed with Intel MPI Benchmark (IMB) running**
 - Test completed without errors.**
- Optimized Routing
 - Routes generated and traced to determine that correct paths were used.

Results

Gordon Network Architecture

-  : InfiniBand QDR 40 Gb/s
-  : 10GbE
-  : 2x10GbE (Storage)
-  : GbE

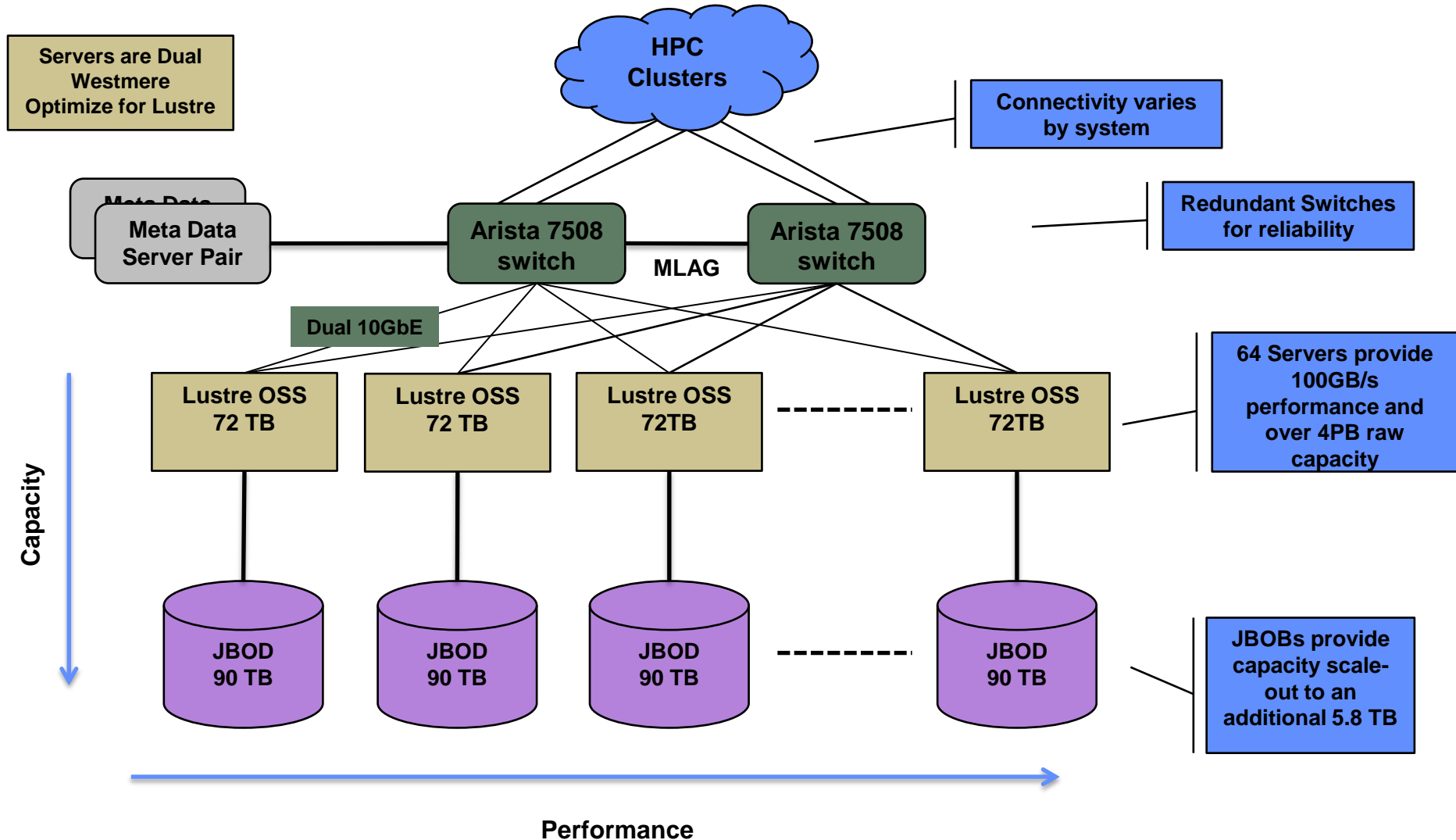


Gordon System Specification

INTEL SANDY BRIDGE COMPUTE NODE	
Sockets & Cores	2 & 12*
Clock speed	2.0 *
DIMM slots per socket	4
DRAM capacity	64 GB
INTEL FLASH I/O NODE	
NAND flash SSD drives	16 or more
SSD capacity per drive/Capacity per node	256 GB / 4 TB *
Bandwidth per drive (r/w)	250 MB/s / 170 MB/s
IOPS per drive (r/w)	35,000 / 2000
SMP SUPER-NODE (VIA VSMP)	
Compute nodes / I/O Nodes	32 / 2
Addressable DRAM	2 TB
Addressable memory including flash	10 TB
GORDON	
Compute Nodes	1,024
Total compute cores	12,288
Peak performance	200 TF *
Aggregate memory	64 TB DRAM; 256 TB flash
INFINIBAND INTERCONNECT	
Aggregate torus BW	9.2 TB/s
Type	Dual-Rail QDR InfiniBand
Link Bandwidth	8 GB/s (bidirectional)
Latency (min-max)	1.25 μ s – 2.5 μ s
LUSTRE-BASED DISK I/O SUBSYSTEM	
Total storage	4 PB (raw)
I/O bandwidth	100 GB/s

* May be revised

Data Oasis Performance and Capacity Architecture



Some final words

Dash Prototype

Software

Resource Scheduling

Allocations

Dash Prototype vs. Gordon

	Dash	Gordon
Number of Compute Nodes	64	1,024
Number of I/O Nodes	4	64
Compute node processors	Intel Nehalem	Intel Sandy Bridge
Compute node memory	48 GB	64 GB
I/O node flash	Intel X25E SLC	Intel eMLC
Flash Capacity per Node	1 TB	4.8 TB
vSMP Supernode Size	16 nodes/768GB	32 nodes/2TB
InfiniBand Network	Single Rail, Fat Tree, DDR	Dual Rail, 3D Torus, QDR
Resource Management	Torque	SLURM

When considering benchmark results and scalability, keep in mind that nearly every major feature of Gordon will be an improvement over Dash.

Gordon Systems Software Stack

Application	Function
Rocks	Cluster Management
CentOS	Operating System
vSMP Foundation	SMP Aggregation
tgtd	iSER storage daemon
XFS, OCFS	I/O node/flash file system
Lustre	Data Oasis Parallel File System
OpenSM/ OFED	3D torus subnet manager IB networking stack
SLURM	Resource Manager

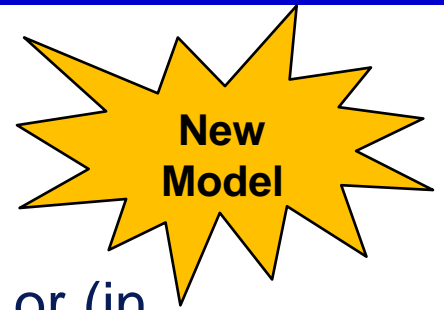
Cluster management and Job scheduling



Cluster management and job scheduling will be handled using the Simple Linux Utility for Resource Management (SLURM)

- Open source, highly scalable
- Well supported by LLNL
- Deployed on many of the world's largest systems
- Advanced reservations
- Backfill scheduling
- Topology aware – critical for Gordon's 3D torus

Gordon dedicated I/O nodes allocations and usage



- Can request long-term dedicated use of one or (in exceptional cases) two I/O nodes
Four dedicated compute nodes will be awarded for each compute node unless justification is made for more
- Usage scenarios
 - Hosting/analysis of community data sets
 - Very large data sets with “hot” results
 - Science Gateways:
www.teragrid.org/web/science-gateways
 - Other special cases that we haven’t even thought of, but maybe you have

Deploying Gordon Open Questions

- Fraction of machine deployed as vSMP nodes
- Size of vSMP nodes
- Number of I/O nodes allocated as dedicated
- Fraction of machine available for interactive jobs
- Fraction of I/O nodes used for visualization
- Size and length of queues

Answers to all of these questions will depend on the mix of allocations requests approved by committee, demand by user, and scheduling decisions to balance needs of users.

Thank you!

For more information

***<http://gordon.sdsc.edu>
gordoninfo@sdsc.edu***

***Shawn Strande
strande@sdsc.edu
(858)822-0277***