Hadoop on the Gordon Data Intensive Cluster

Amit Majumdar, Scientific Computing Applications
Mahidhar Tatineni, HPC User Services
San Diego Supercomputer Center
University of California San Diego
Dec 18, 2012
WBDB2012.in

Gordon – An Innovative Data-Intensive Supercomputer

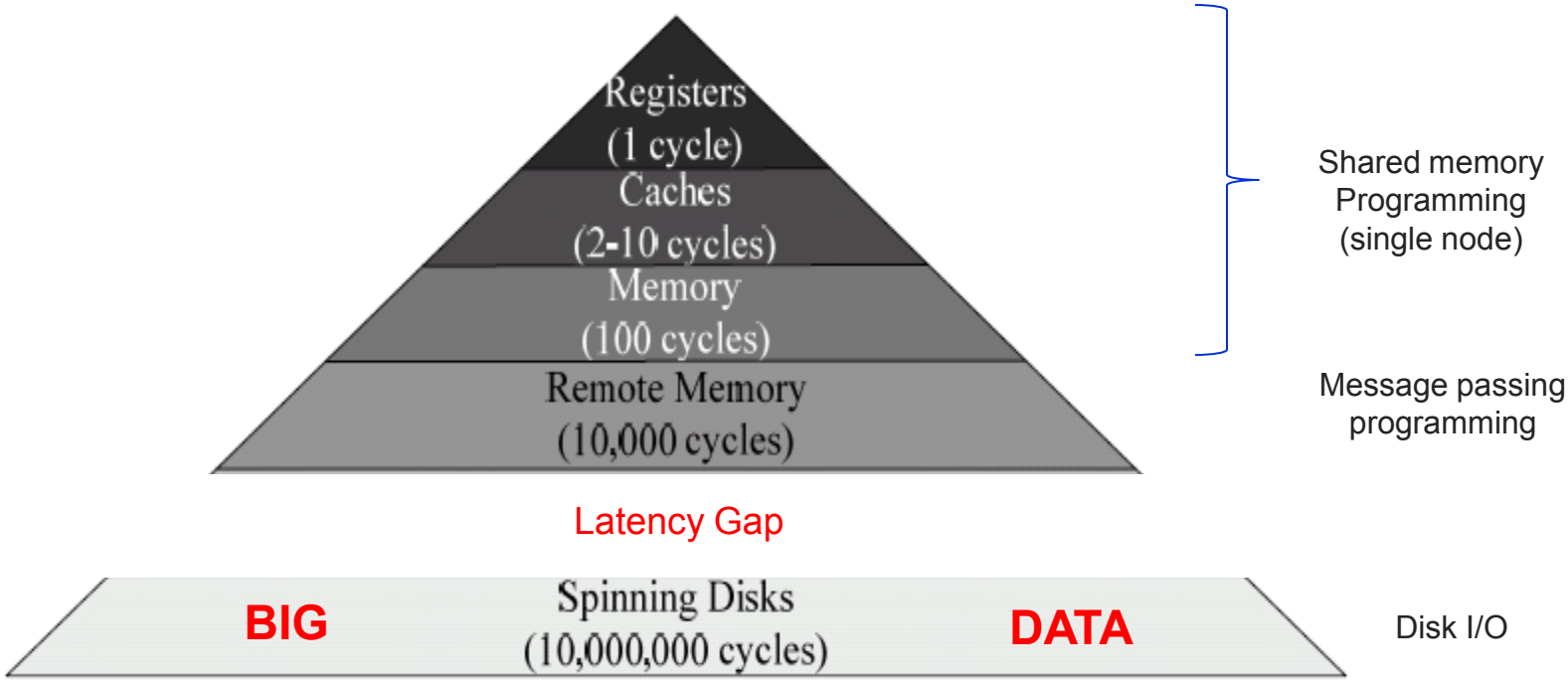
- Designed to accelerate access to massive amounts of data in areas of genomics, earth science, engineering, medicine, and others
- Emphasizes memory and IO over FLOPS.
- Appro integrated 1,024 node Sandy Bridge cluster
- 64 GB per node with 16 core Sandy Bridge
- 300 TB of high performance Intel flash
- 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)

Gordon Design: Two Driving Ideas

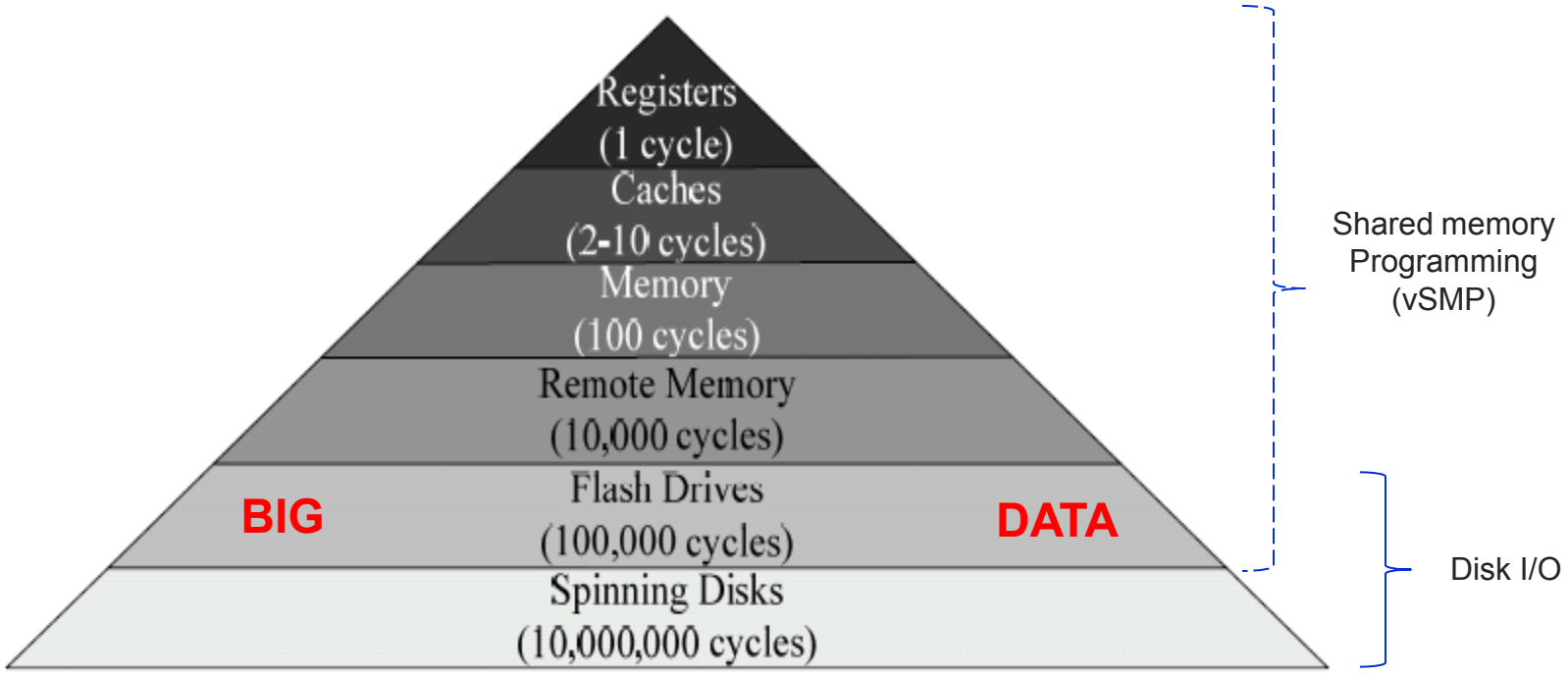
- **Observation #1: Data keeps getting further away from processor cores ("red shift")**
  - Do we need a new level in the memory hierarchy?

- **Observation #2: Many data-intensive applications are serial and difficult to parallelize**
  - Would a large, shared memory machine be better from the standpoint of researcher productivity for some of these?
  - ➔Rapid prototyping of new approaches to data analysis

The Memory Hierarchy of a Typical Supercomputer



The Memory Hierarchy of Gordon

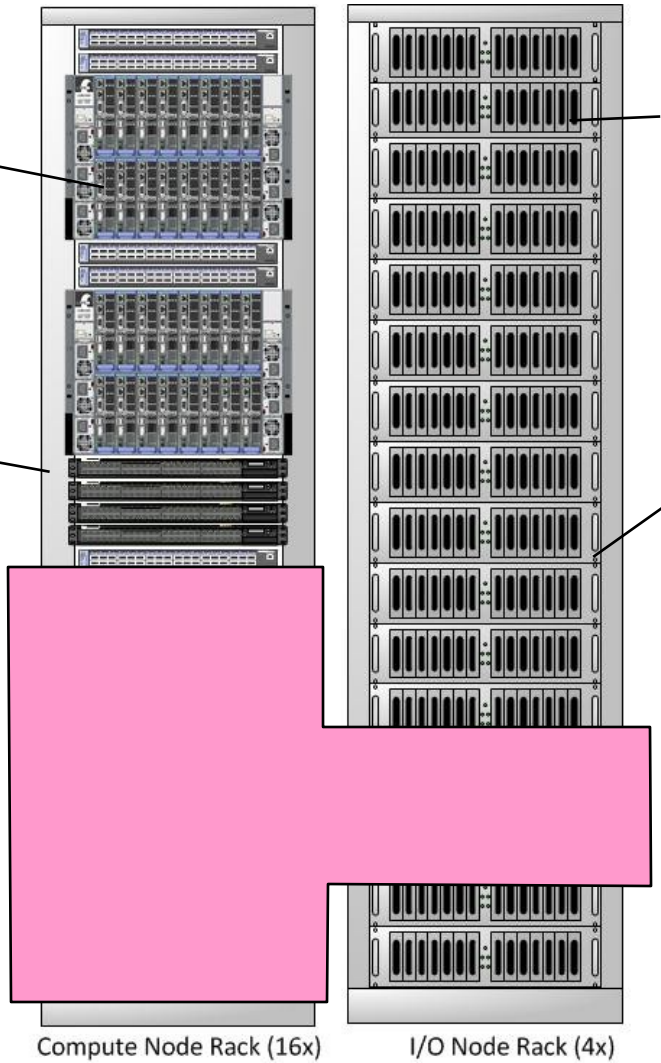


Gordon Design Highlights

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

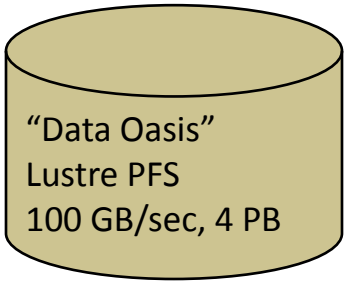
- 3D Torus
- Dual rail QDR

- 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



(Some) SSDs 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)	270 / 210 MB/s	100-150 MB/s	✔
IOPS (r/w)	38,500 / 2000	100	✔
Power consumption (when doing r/w)	2-5 W	6-10 W	✔
Price/GB	$3/GB	$.50/GB	-
Endurance	2-10PB	N/A	✔
Total Cost of Ownership	Jury is still out.		

Hadoop Overview

- **Hadoop framework extensively used for scalable distributed processing of large datasets. High interest from user base.**

- **Gordon architecture presents potential performance benefits with high performance storage (SSDs) and high speed network (QDR IB).**

- **Storage options viable for Hadoop on Gordon**
  - SSD via **iSCSI Extensions for RDMA (**iSER)
  - Lustre filesystem (Data Oasis), persistent storage

- **Approach to running Hadoop within the scheduler infrastructure – myHadoop (developed at SDSC).**

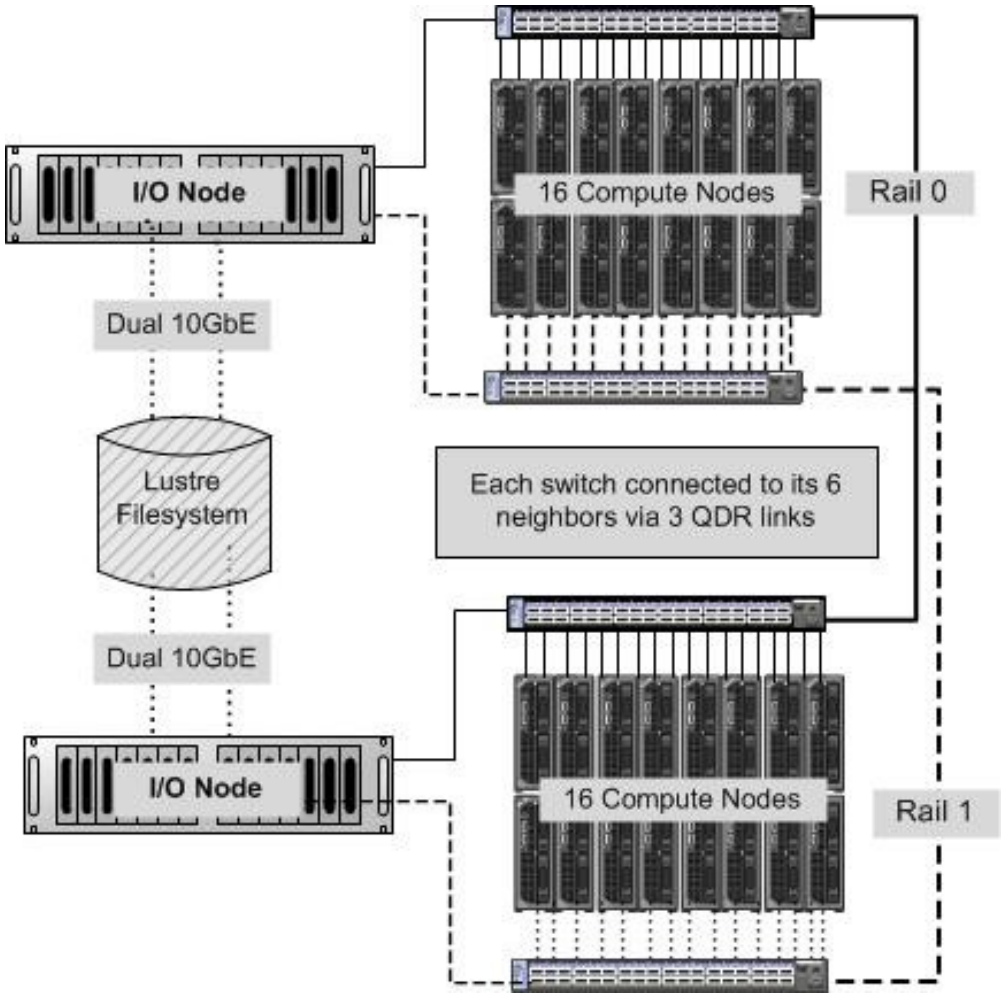
Overview (Contd.)

- **Network options**
  - 1 GigE – low performance
  - IPoIB
- **Results**
  - DFS – Benchmark runs done on both SSD and Lustre.
  - TeraSort – Scaling study with sizes 100GB-1.6TB, 4-128 nodes.
- **Future/Ongoing Work**
  - UDA - Unstructured Data Accelerator from Mellanox
    http://www.mellanox.com/related-docs/applications/SB_Hadoop.pdf
  - Lustre for Hadoop – Can provide another persistent data option.

Gordon System Specification

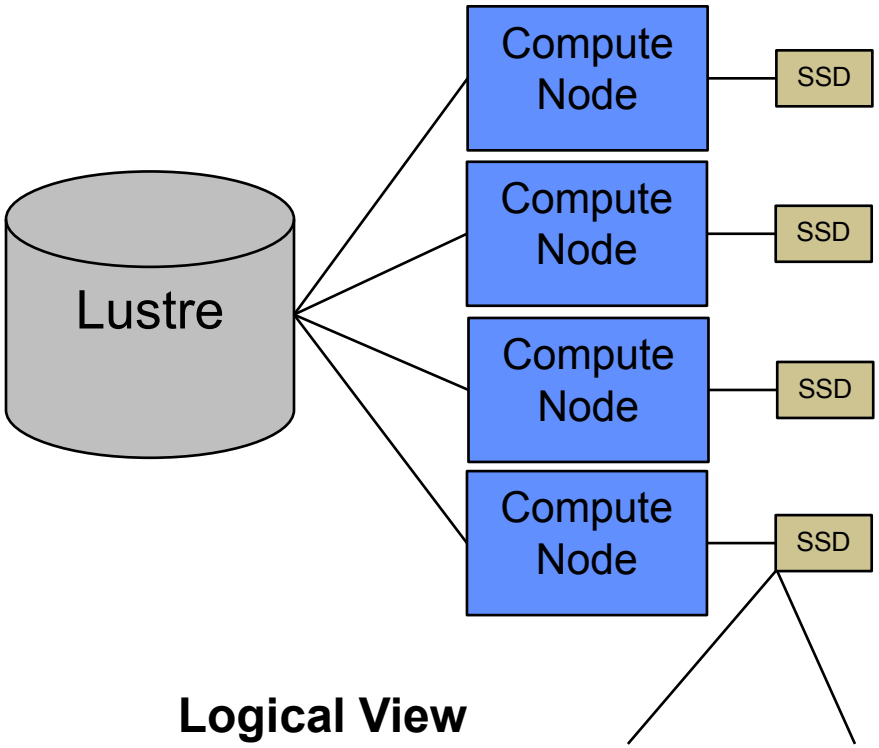
Intel Sandy Bridge Compute node	
Sockets	2
Cores	16
Clock speed	2.6
DIMM slots per socket	4
DRAM capacity	64 GB
Intel Flash I/O Node	
NAND flash SSD drives	16
SSD capacity per drive/Capacity per node/total	300 GB / 4.8 TB / 300 TB
Flash bandwidth per drive (read/write) Flash bandwidth per node (write/read)	270 MB/s / 210 MB/s 4.3 /3.3 GB/s
SMP Super-Node	
Compute nodes	32
I/O nodes	2
Addressable DRAM	2 TB
Addressable memory including flash	12TB
Gordon	
Compute Nodes	1,024
Total compute cores	16,384
Peak performance	341TF
Aggregate memory	64 TB
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 µs – 2.5 µs
Disk I/O Subsystem	
Total storage	4.5 PB (raw)
I/O bandwidth	100 GB/s
File system	Lustre

Subrack Level Architecture



16 compute nodes per subrack

Primary Storage Option for Hadoop: One SSD per Compute Node (only 4 of 16 compute nodes shown)

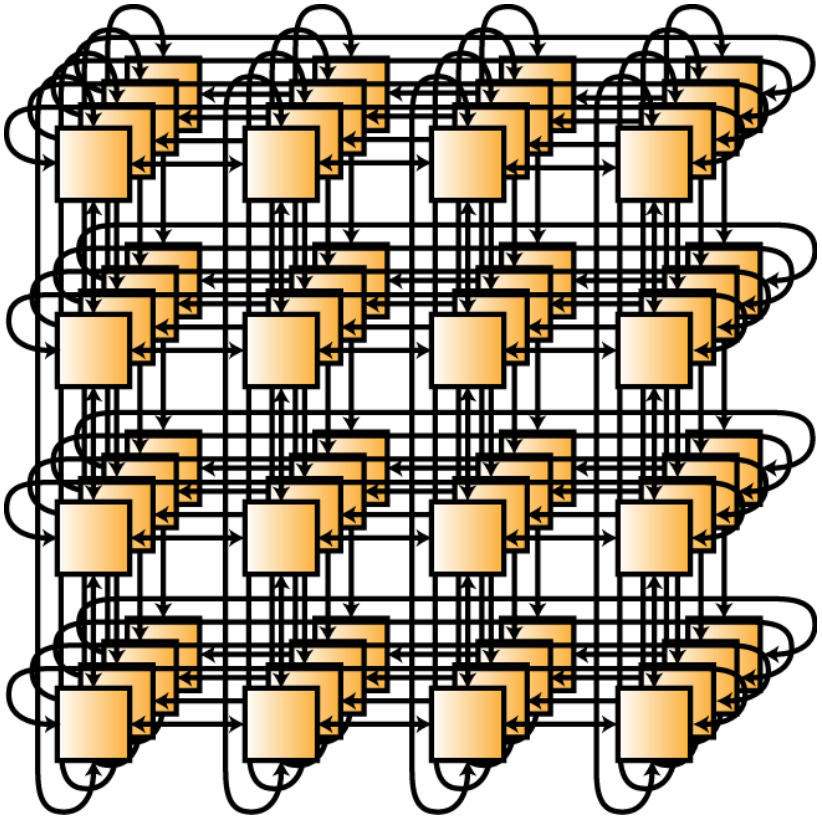


Logical View

File system appears as: /scratch/$USER/$PBS_JOBID

- **The SSD drives on each I/O node are exported to the compute nodes via the iSCSI Extensions for RDMA (iSER).**
- **One 300 GB flash drive exported to each compute node appears as a local file system**
- **For the Hadoop cluster, 300 GB of disk available as "datanode"**
- **Each compute node has 2 IB connections**
- **iSER and Lustre use rail 1**
- **Hadoop network (IPoIB or UDA) uses rail 0**
- **Lustre parallel file system is mounted identically on all nodes with aggregate BW of 100 GB/sec**
- **This exported SSD storage can serve as the local storage used by HDFS**
- **Network BW over rail 1 and the aggregate I/O BW from the 16 SSD drives are well matched**
- **No network bottleneck in physically locating the flash drives in the I/O node**

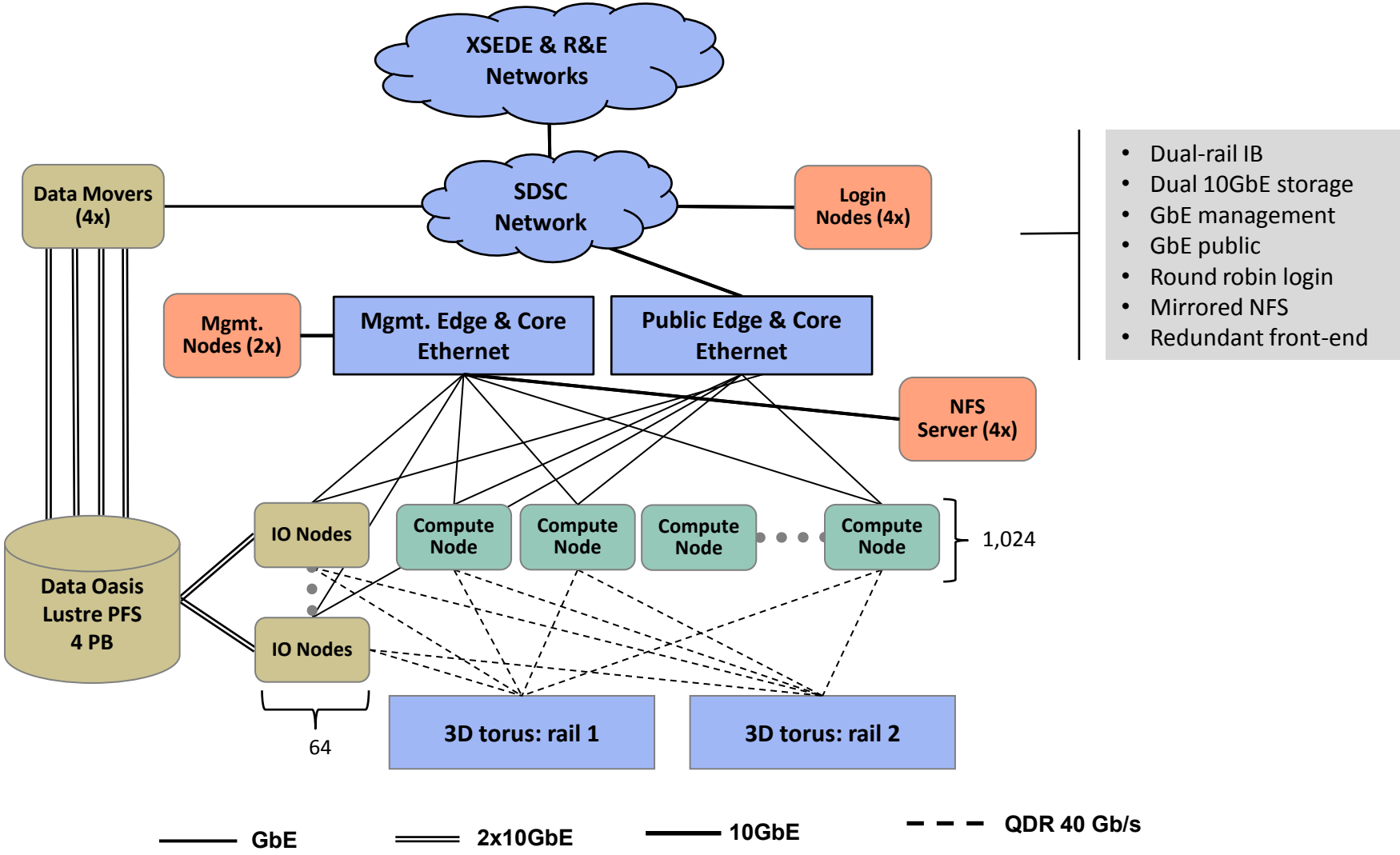
3D Torus of Switches



- Linearly expandable
- Simple wiring pattern
- 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

3rd dimension wrap-around not shown for clarity

Gordon Network Architecture



- Dual-rail IB
- Dual 10GbE storage
- GbE management
- GbE public
- Round robin login
- Mirrored NFS
- Redundant front-end

myHadoop – Integration with Gordon Scheduler

- **myHadoop* was developed at SDSC to help run Hadoop within the scope of the normal scheduler.**
- **Scripts available to use the node list from the scheduler and dynamically change Hadoop configuration files.**
  - mapred-site.xml
  - masters
  - slaves
  - core-site.xml
- **Users can make additional hadoop cluster changes if needed**
- **Nodes are exclusive to the job [does not affect the rest of the cluster].**

***myHadoop link: http://sourceforge.net/projects/myhadoop/**

Hadoop on Gordon – Storage Options

- **Primary storage option is to use SSDs which are available via iSCSI Extensions for RDMA (iSER) on each compute node.**

- **Prolog creates job specific SSD filesystem location which can be used for HDFS storage. myHadoop framework leveraged to do this dynamically.**

- **Lustre storage is available for persistent option. DFS benchmark results show very good performance comparable to SSD storage.**

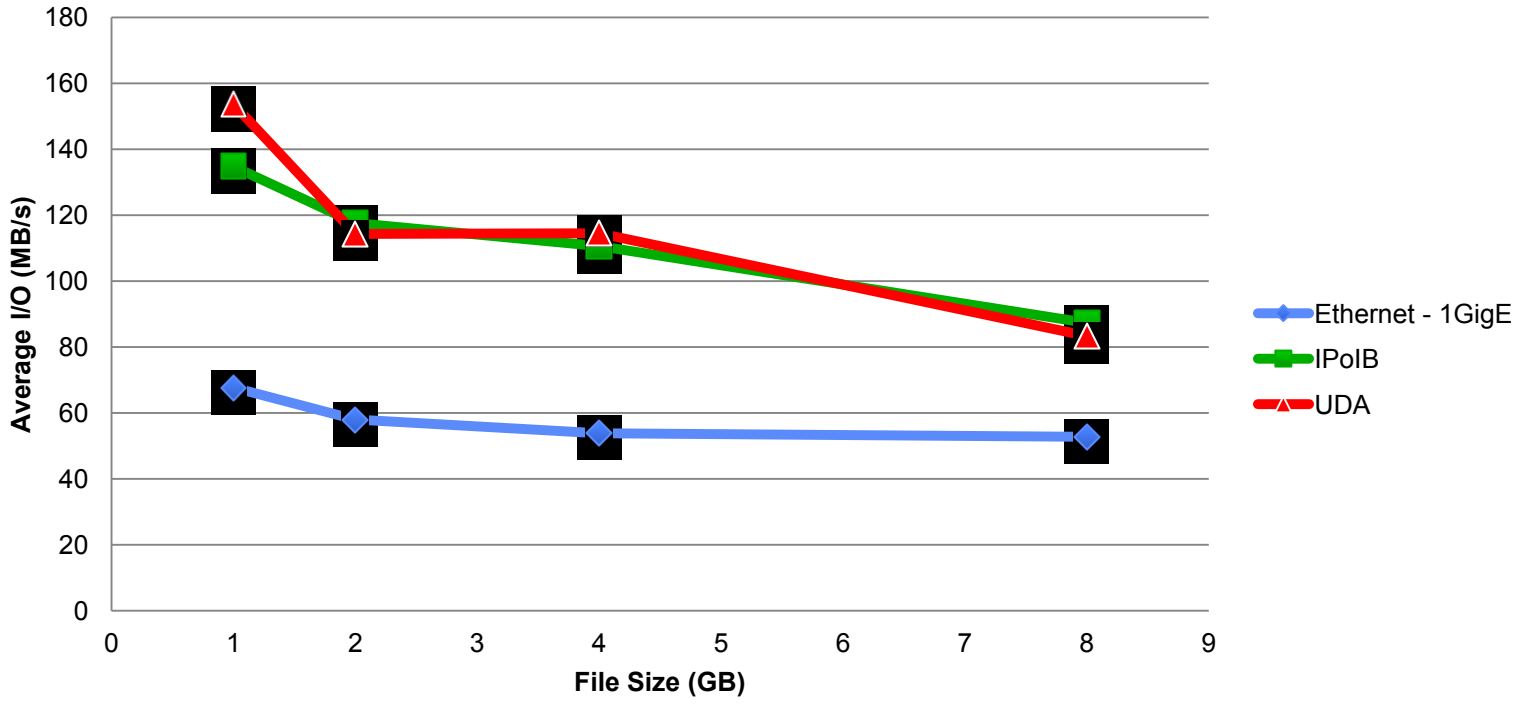
Hadoop on Gordon – Network Options

- **All Gordon compute nodes are dual QDR Infiniband connected. Additionally, a 1GigE connection is available.**

- **For a standard Apache Hadoop install IPoIB using one of the default QDR links is the best network option. myHadoop setup uses the ib0 addresses in dynamic configuration set up.**

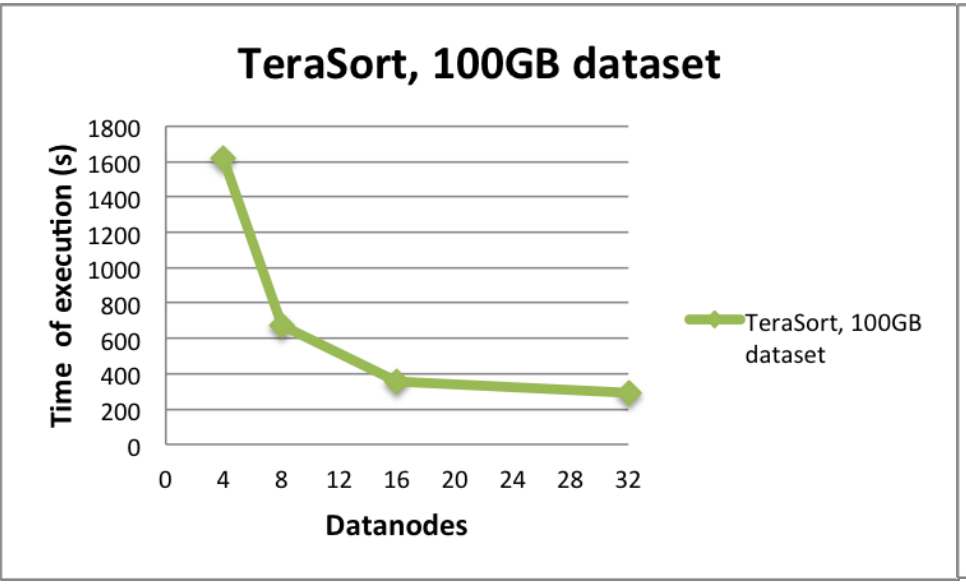
- **Mellanox's Unstructured Data Accelerator (UDA) is another option providing the middleware interface directly to high speed IB link. We are currently testing this on Gordon.**

Hadoop on Gordon – TestDFSIO Results

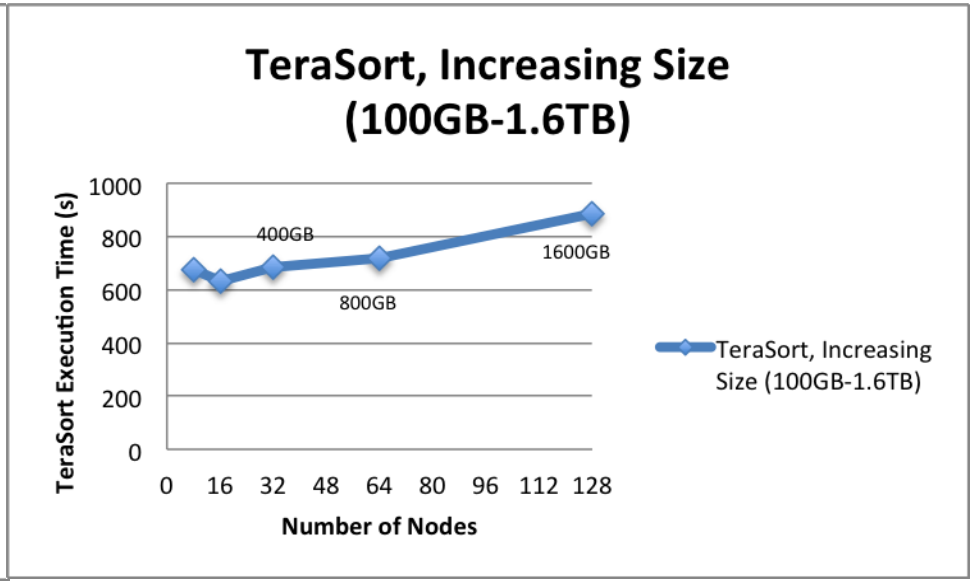


Results from the TestDFSIO benchmark. Runs were conducted using 4 datanodes and 2 map tasks. Average write I/O rates are presented.

Hadoop on Gordon – TeraSort Results



(a)



(b)

Results from the TeraSort benchmark. Data generated using teragen and is then sorted. Two sets or runs were performed using IPoIB, SSD storage – (a) Fixed dataset size of 100GB, number of nodes varied from 4 to 32, (b) Increasing dataset size with number of nodes.

Preliminary Hadoop on Gordon Tuning

- **Based on the"Optimizing Hadoop Deployments" whitepaper by Intel, October 2010, version 2.0**

- **Managed using the Rocks cluster management software, with a CentOS distribution (kernel version 2.6.34.7-1), and a custom mellanox ofed stack (v. 1.5.3) designed to enable the iSER export of SSD resources.**
- **The Hadoop cluster tests were done using Apache Hadoop version 1.0.4.**
- **The following three network options were used for the Hadoop cluster – 1) 1 GigE ethernet, 2) IPOIB using one Infiniband rail, and 3) UDA version 3.0.**
- **In addition to the storage and network options, tuning of**
  - HDFS parameters (dfs.block.size),
  - map/reduce parameters (io.sort.factor, io.sort.mb, and io.sort.record etc),
  - general configuration parameters (dfs.namenode.handler.count, and mapred.job.tracker.handler.count etc) was also considered.

Summary

- **Hadoop can be dynamically configured and run on Gordon using myHadoop framework.**
- **High speed Infiniband network allows for significant gains in performance using SSDs.**
- **TeraSort benchmark shows good performance and scaling on Gordon.**
- **Future improvements/tests include**
  - using Mellanox UDA for terasort benchmark
  - use lustre filesystem for HDFS storage