SDSC's Data Oasis

Balanced performance and cost-effective Lustre file systems.

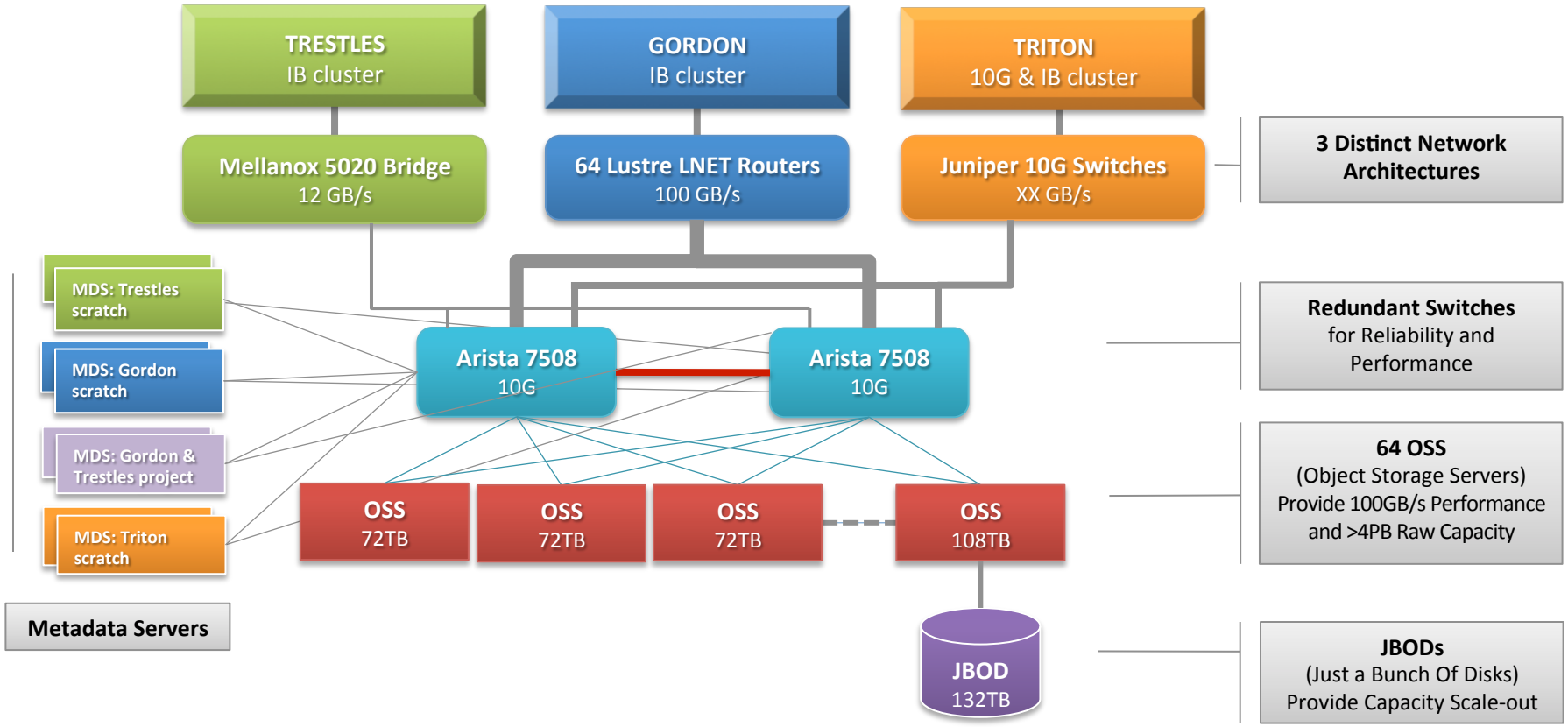
Lustre User Group 2013
(LUG13)

Rick Wagner
San Diego Supercomputer Center
Jeff Johnson
Aeon Computing
April 18, 2013

Data Oasis

- **High performance, high capacity Lustre-based parallel file system**
- **10GbE I/O backbone for all of SDSC's HPC systems, supporting multiple architectures**
- **Integrated by Aeon Computing using their EclipseSL**
- **Scalable, open platform design**
- **Driven by 100GB/s bandwidth target for *Gordon***
- **Motivated by $/TB and $/GB/s**
  - $1.5M = 4@MDS + 64@OSS = 4PB = 100GB/s
- **6.4PB capacity and growing**
- **Currently Lustre 1.8.7**

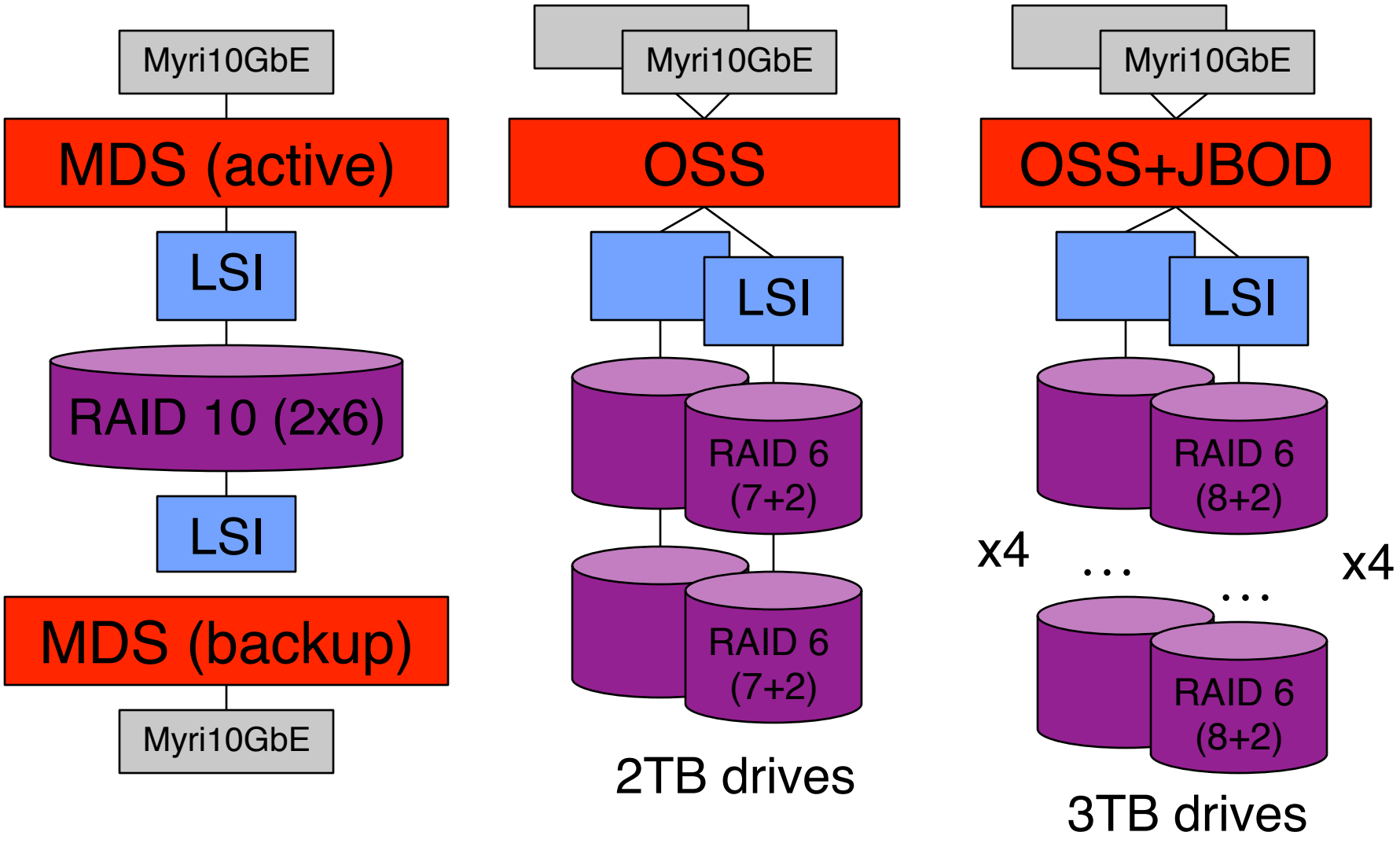
Data Oasis Heterogeneous Architecture



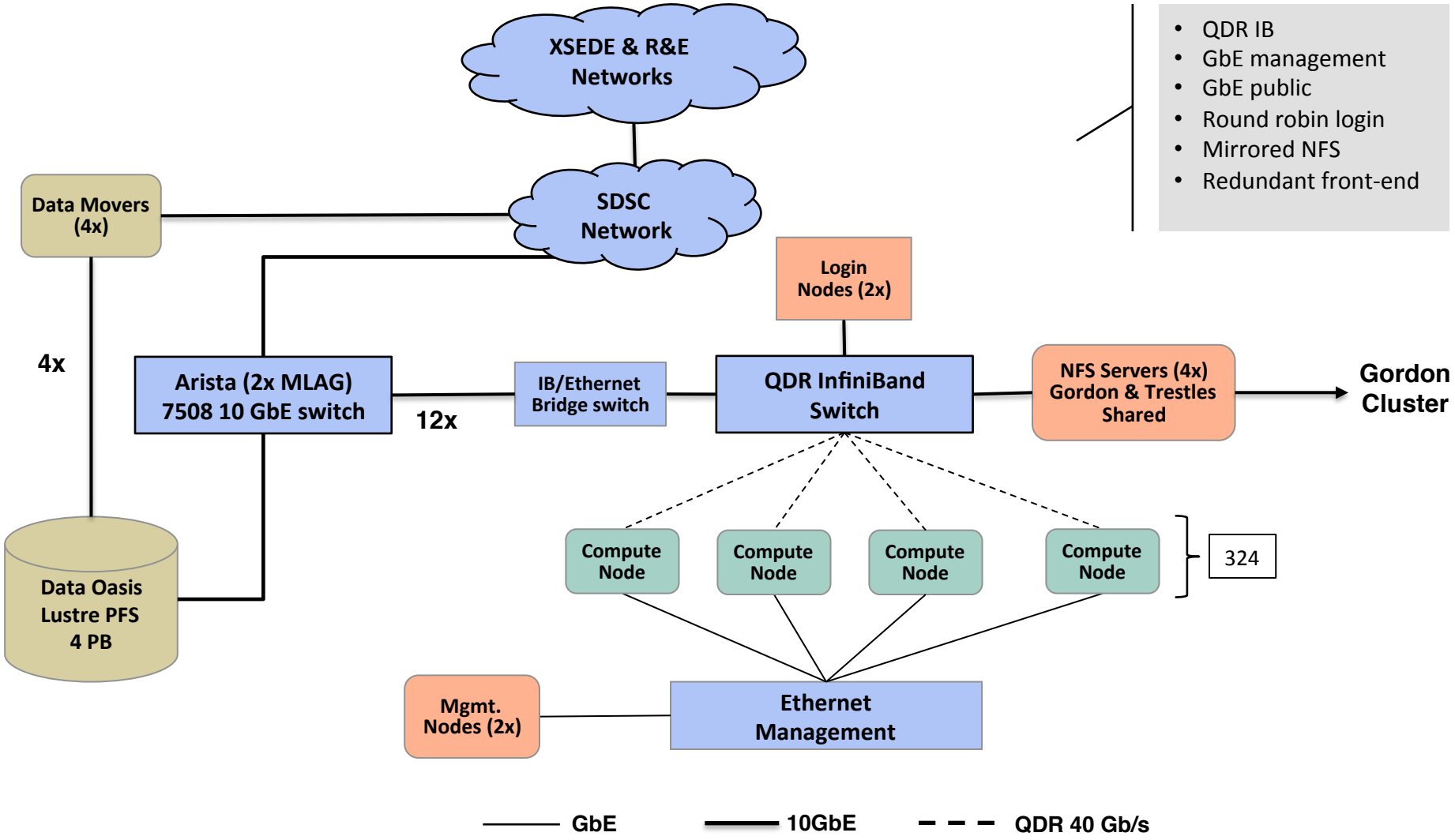
File Systems

File System	Clusters	OSSes	JBODs	Capacity (RAW)
Monkey	*Gordon*	32	0	2.3PB
Meerkat	*Gordon & Trestles*	8	8	1.9PB
Puma	*Trestles*	8	0	576TB
Dolphin	*Triton*	16	0	1.2PB
Rhino	Development	4	4	480TB

Data Oasis Servers

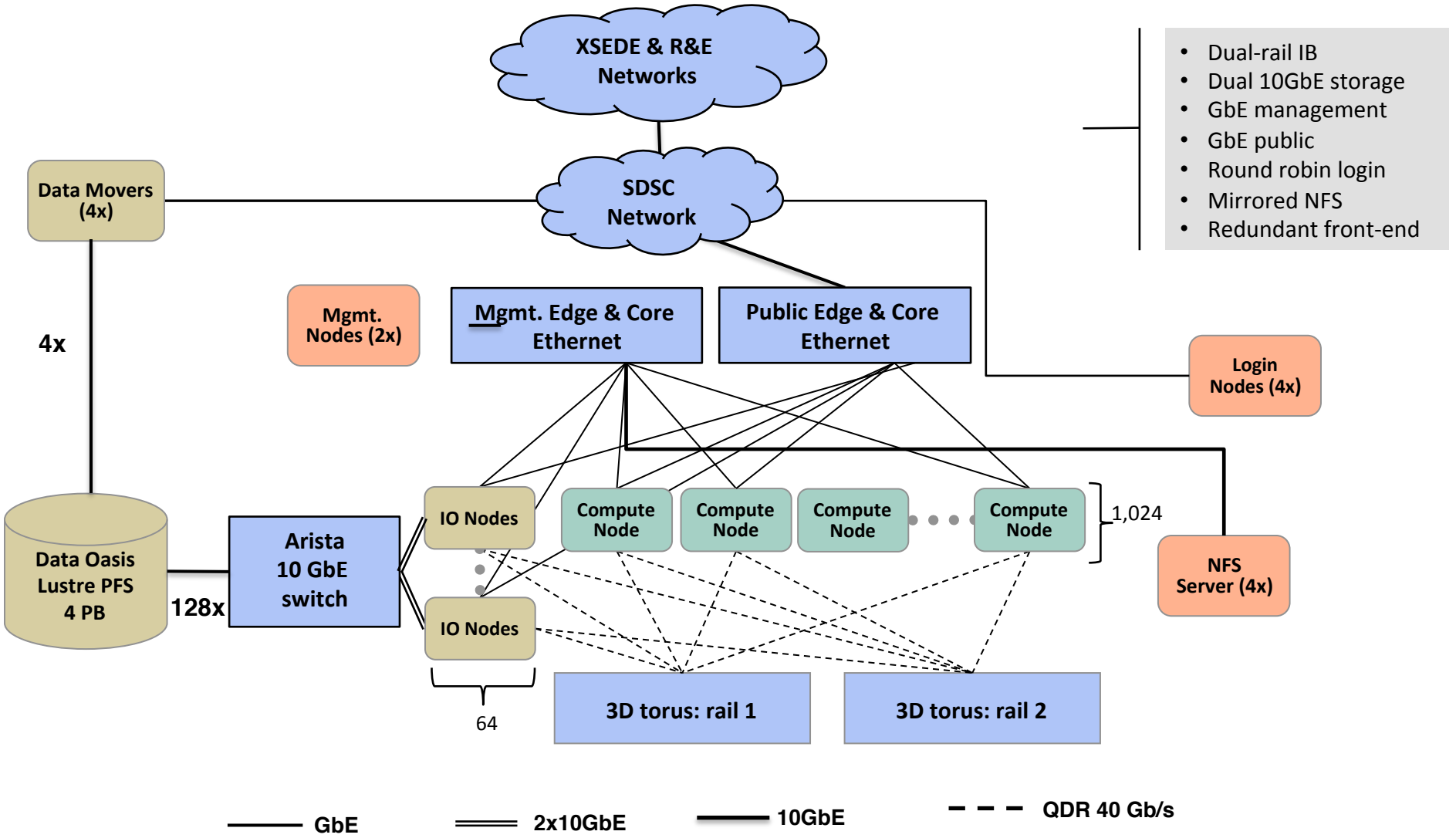


Trestles Architecture



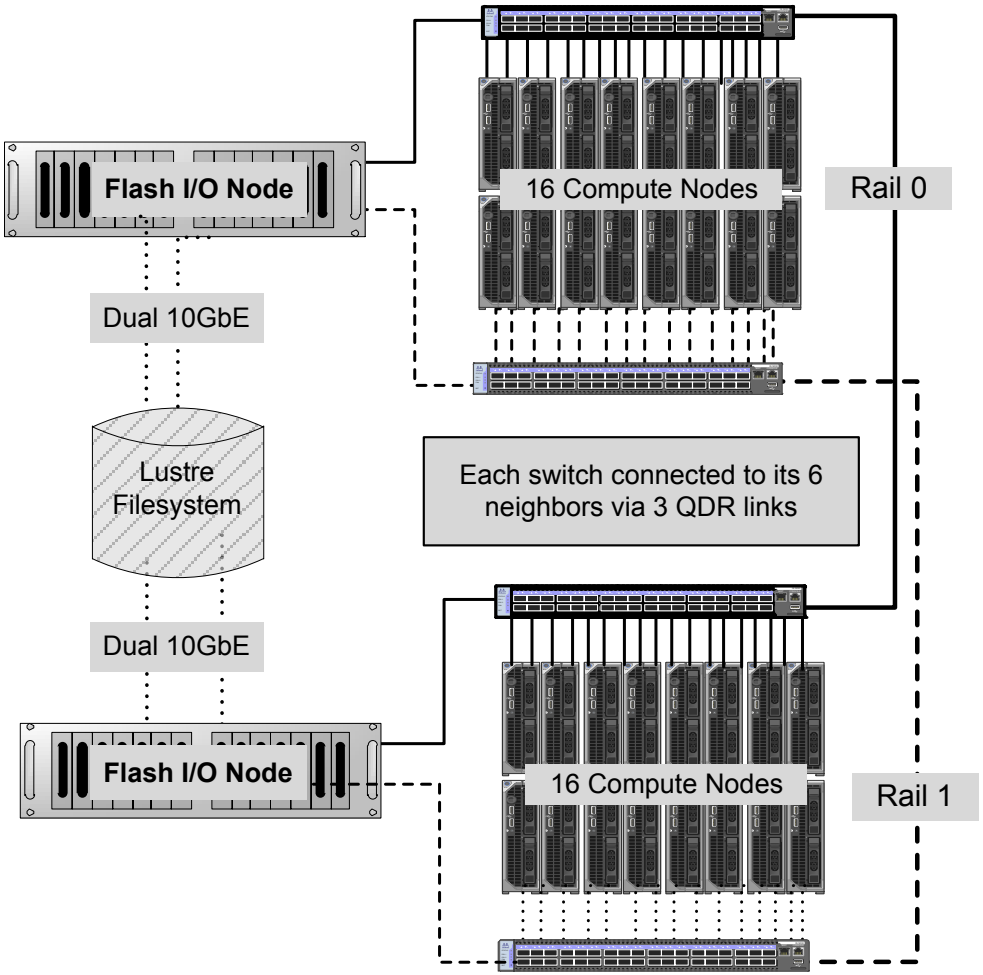
- QDR IB
- GbE management
- GbE public
- Round robin login
- Mirrored NFS
- Redundant front-end

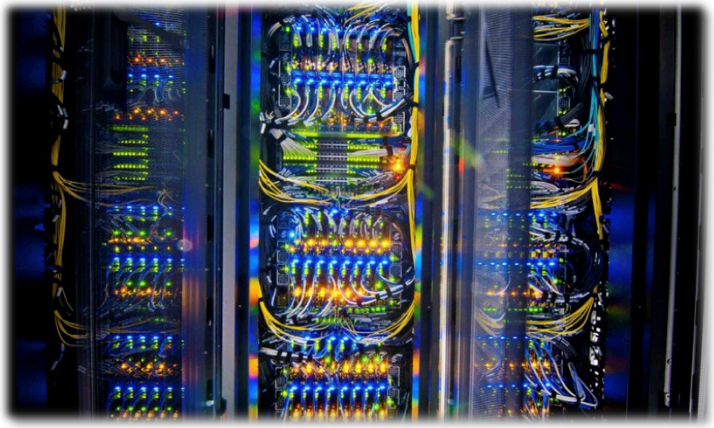
Gordon Network Architecture

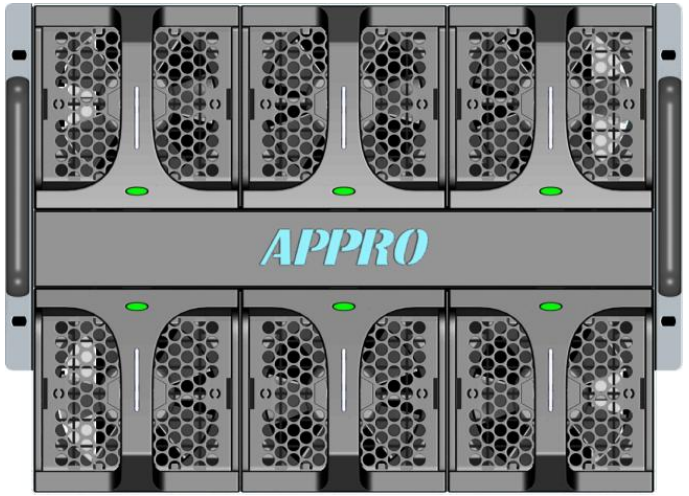


- Dual-rail IB
- Dual 10GbE storage
- GbE management
- GbE public
- Round robin login
- Mirrored NFS
- Redundant front-end

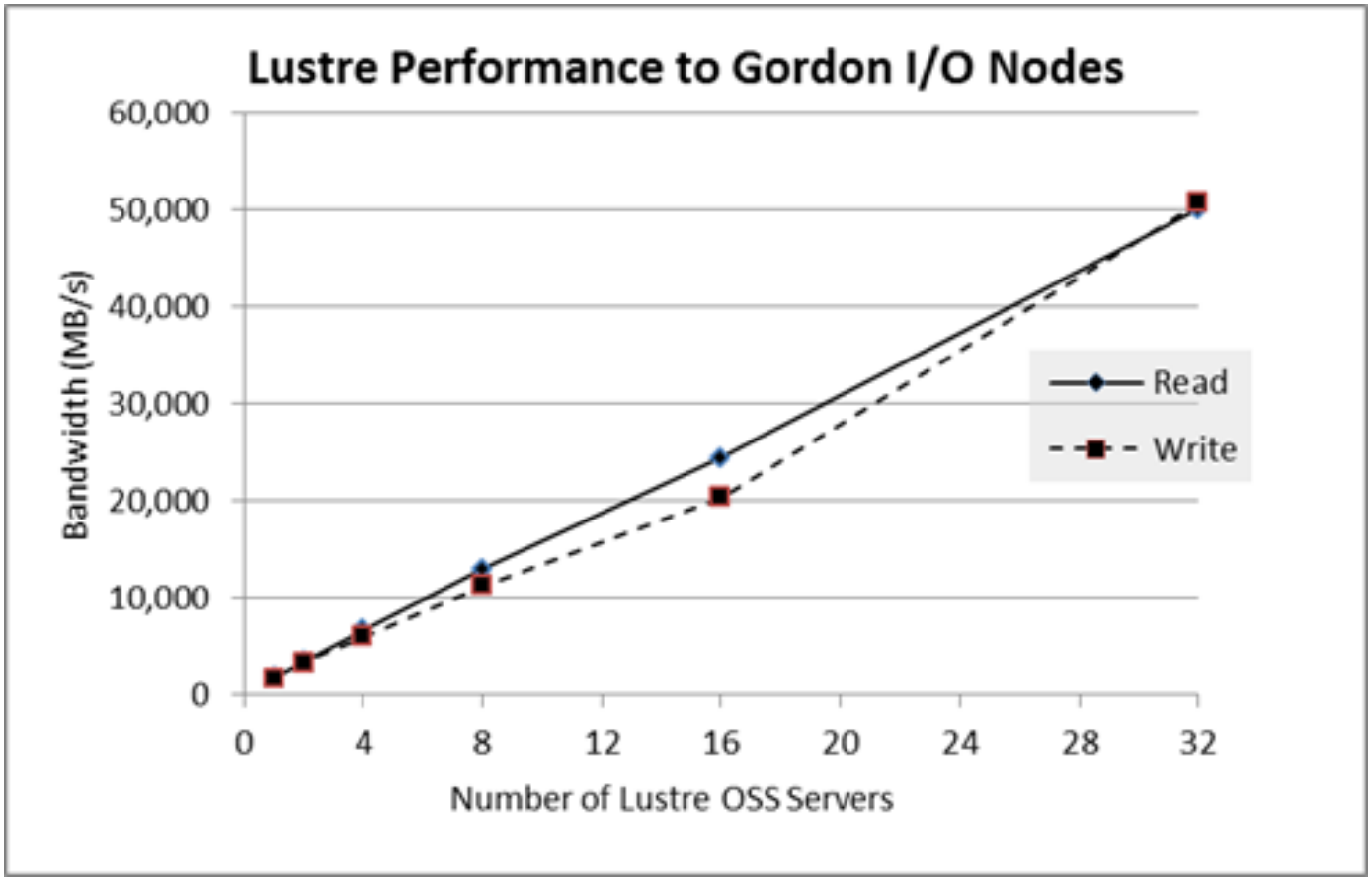
Gordon Network Design Detail







Data Oasis Performance – Measured from Gordon



Issues & The Future

- **LNET "death spiral"**
  - LNET tcp peers stop communicating, packets back up
- **We need to upgrade to Lustre 2.x soon**
  - Can't wait for MDS SMP improvements & DNE
- **Design drawback: juggling data is a pain**
- **Client virtualization testing**
  - SR-IOV very promising for o2ib clients
- **Watching the Fast Forward program**
  - *Gordon*'s architecture ideally suited to burst buffers
- **HSM**
  - Really want to tie Data Oasis to SDSC Cloud