Building the Open Storage Network

Alex Szalay
The Johns Hopkins University

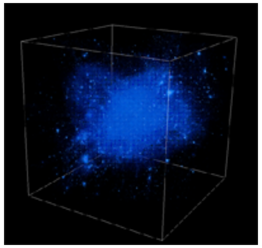
Institute for Data Intensive Engineering and Science
idies

OSN Mission Statement

The mission of OSN is to provide a low-cost, high-quality, sustainable, distributed storage cloud for the NSF research community.

Emerging Trends in Science

- Broad sociological changes
  - *Convergence of Physical and Life Sciences*
  - *Data collection in ever larger collaborations*
  - *Virtual Observatories: CERN, IVOA, NCBI, NEON, OOI,…*
  - *Analysis decoupled, off archived data by smaller groups*
- Scientific data sets moving from 100TBs to PBs
  - *While the data are here, analysis solutions are not*
  - *Data preservation and curation needs to be reinvented*
- National infrastructure doesn’t map onto new needs



Computational Infrastructure

- The NSF has invested significant funds into high performance computing, both capacity and capability
  - *These systems form XSEDE, a national scale organization with excellent support infrastructure*
  - *The usage of these machines is quite broad, and gradually transitioning from HPC simulations to include more and more large data analysis tasks*
- Most large MREFC projects still build their own computational infrastructure in a vertical fashion

Networking Infrastructure

- The NSF has invested about $150M to bring hugh-speed connectivity to over 200 universities in the CC-NIE and CC* programs
- The Internet2 provides a stable high-speed backbone at multiple 100G lines
- There are several peering points to ESNet, NASA and commercial cloud providers

Storage Infrastructure

- Storage largely balkanized
  - *Every campus/project does its own specific vertical system*
  - *As a result, lots of incompatibilities and inefficiencies*
  - *People are only interested in building minimally adequate*
  - *As a result, we build storage tiers 'over and over'*
  - *Big projects need petabytes, also lots of 'long tail' data*
- Cloud storage not a good match at this point for PBs
  - *Amazon, Google, Azure too expensive:*
    they force you to buy the storage every month
  - *Wrong tradeoffs: cloud redundancies too strong for science*
  - *Getting data in (and out) is very expensive*

Everybody needs a reliable, industrial strength storage tier!

Opportunity

- The NSF has funded 150+ universities to connect to Internet2 at high speeds (40-100G) for ~$150M
- Ideal for a large national distributed storage system:
  - *Place a 1-2PB storage rack at each of these sites (~200PB)*
  - *Create a redundant interconnected storage substrate using an industrial strength erasure code storage*
  - *Incredible aggregate bandwidth, easy flow between the sites*
  - *Can also act as gateways to cloud providers*
  - *Automatic compatibility, simple standard API (S3)*
  - *Implement a set of simple policies*
  - *Enable sites to add additional storage at their own cost*
  - *Variety of services built on top by the community*
- Estimated Cost: $20-40M

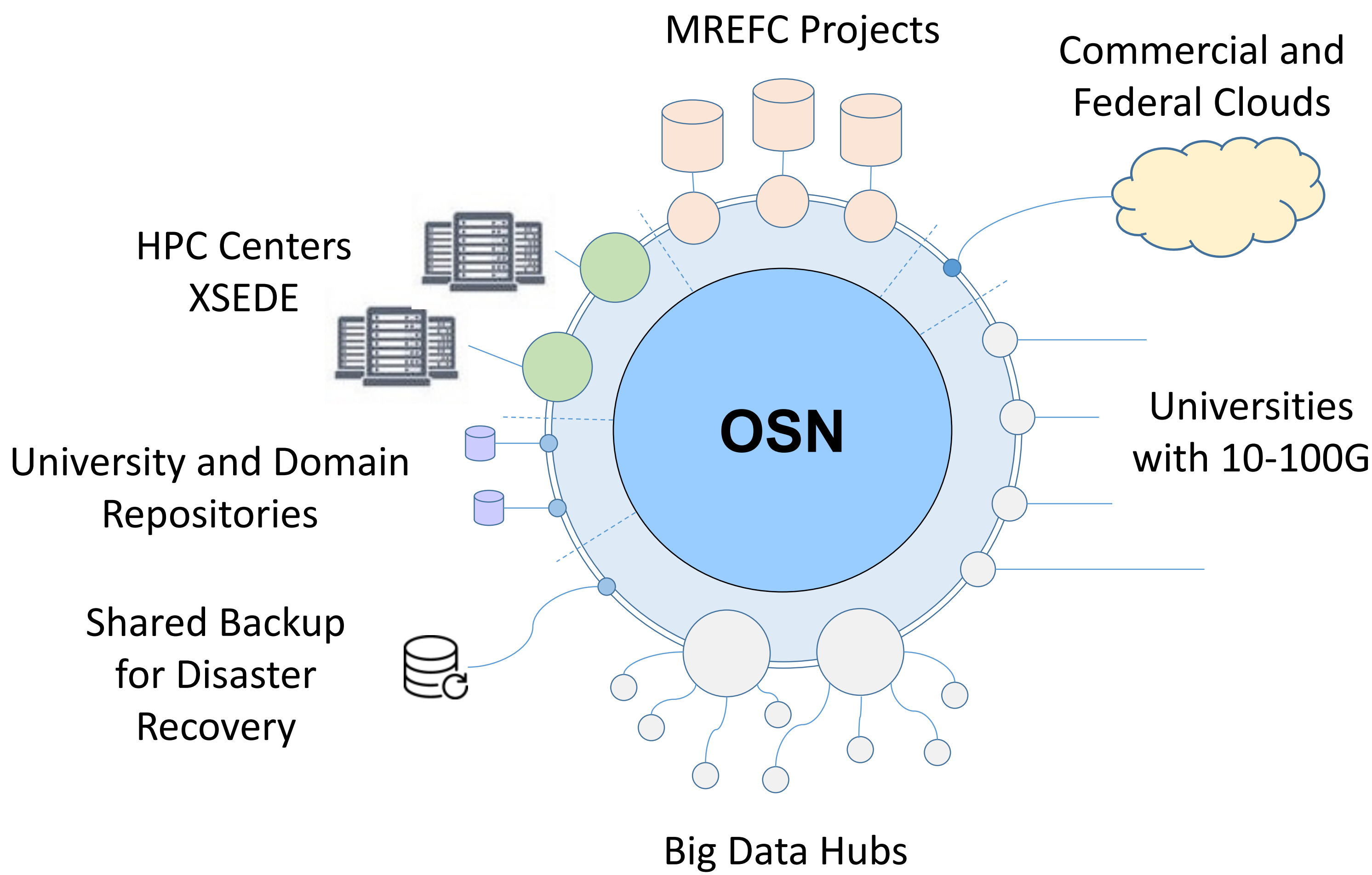
System could be the world's largest academic storage facility

Transformative Impact

- Totally change the landscape for academic Big Data
  - *Create a homogeneous, uniform storage tier for science*
  - *Liberate communities to focus on analytics and preservation*
  - *Amplify the NSF investment in networking*
  - *Very rapidly spread best practices nationwide*
  - *Universities can start thinking about PB-scale projects*
- Impact unimaginable
  - *Links to XSEDE, NDS, RDA, Globus*
  - *Big Data projects can use it for data distribution*
    - LHC, LSST, OOI, genomics
  - *Small projects can build on existing infrastructure*
  - *Enable a whole ecosystem of services to flourish on top*
  - *Would provide "meat" for the Big Data Hub communities*
    - Enable nation-wide smart cities movement

New opportunity for federal, local, industrial, private partnership

Connections



Questions, Tradeoffs

Cannot do "everything for everybody"!

- Where to draw the line? Use the 80-20 rule…
  - *Build the 20% of possible, that serves 80% of needs*
- Hierarchical or flat?
  - *A single central 'science cloud' vs a totally flat ring?*
  - *Or 4-6 big sites with 10-20PB, the rest flat with 1-2PB?*
- Object-store or POSIX
  - *Keep it simple, focus on large objects, S3 interface*
- This is really a social engineering challenge
  - *Teach the universities how to be comfortable with PB data*
  - *Centralized may be more efficient, but will have trust issues*
  - *Giving each university its own device speeds up adaptation*

High-Level Architecture

- Should there be any computing on top?
  - *A lightweight analytics tier makes system much more usable*
  - *A set of virtual machines for front ends*
  - *But these also add complexity?*
  - *Everybody needs similar storage, analytics tier more diverse*
  - *Some need HPC, others Beowulf/ Hadoop/ TensorFlow/ ??*
- Focus on simplicity
  - *Everybody needs storage, keep it storage only*
  - *Create a simple appliance with 1-2PB of storage*
  - *100G interfaces, straddling the campus firewall and DMZ*
- Software stack
  - *Ultra simple object-store interface, converging on S3*
  - *Management and monitoring*

Building Blocks

- Scalable element (SE)
  - *300TB of storage in single server*
  - *Support 40G interface for sequential read/write*
  - *Should saturate 40G for read, about half for write*
- Stack of multiple SEs
  - *Aggregated to 100G on a fast TOR switch, now becoming quite inexpensive (<$20K)*
- These can also exist inside the university firewall
  - *But purchased on local funds, storing local data*

OSN Software Requirements

- Functional
  - *does what is needed*
- Robust
  - *it is highly available*
- Secure
  - *ensures that only authorized entities can access its resources*
- Performant
  - *allows applications to make good use of petascale storage and high-speed networks*

Management

- Who owns it?
  - *OSN storage should remain in a common namespace*
  - *This would enable uniform policies and interfaces*
- Software management
  - *Central management of software stack (push)*
  - *Central monitoring of system state*
- Hardware management
  - *Local management of disk health*
  - *Universities should provide management personnel*
- Policy management
  - *This is **hard** and requires a lot more discussion*
- Monitoring
  - *Two tier, store all events and logs locally, send only alerts up*
  - *Try to predict disk failures, preventive maintanance*
- Establish metrics for success

Systems Management

- OSN servers will netboot into a minimal Linux distribution, running Kubernetes container management, intrusion detection (e.g., OSSEC), and other core remote monitoring and management software.
- All OSN other software will be deployed as Docker containers, including storage system (e.g., Ceph), storage management (e.g., allocation and accounting), and storage access (e.g., Globus Connect Server).

Storage Access

- Basic data access is based upon the S3 protocol
- Can be later augmented by a few well-justified APIs
- Authentication via Oauth, enabling InCommon
- Value added Globus services:
  - *Transfer: fire and forget*
  - *Replicate*
  - *Identifiers*
  - *Search*
  - *Automate*
  - *Manage*

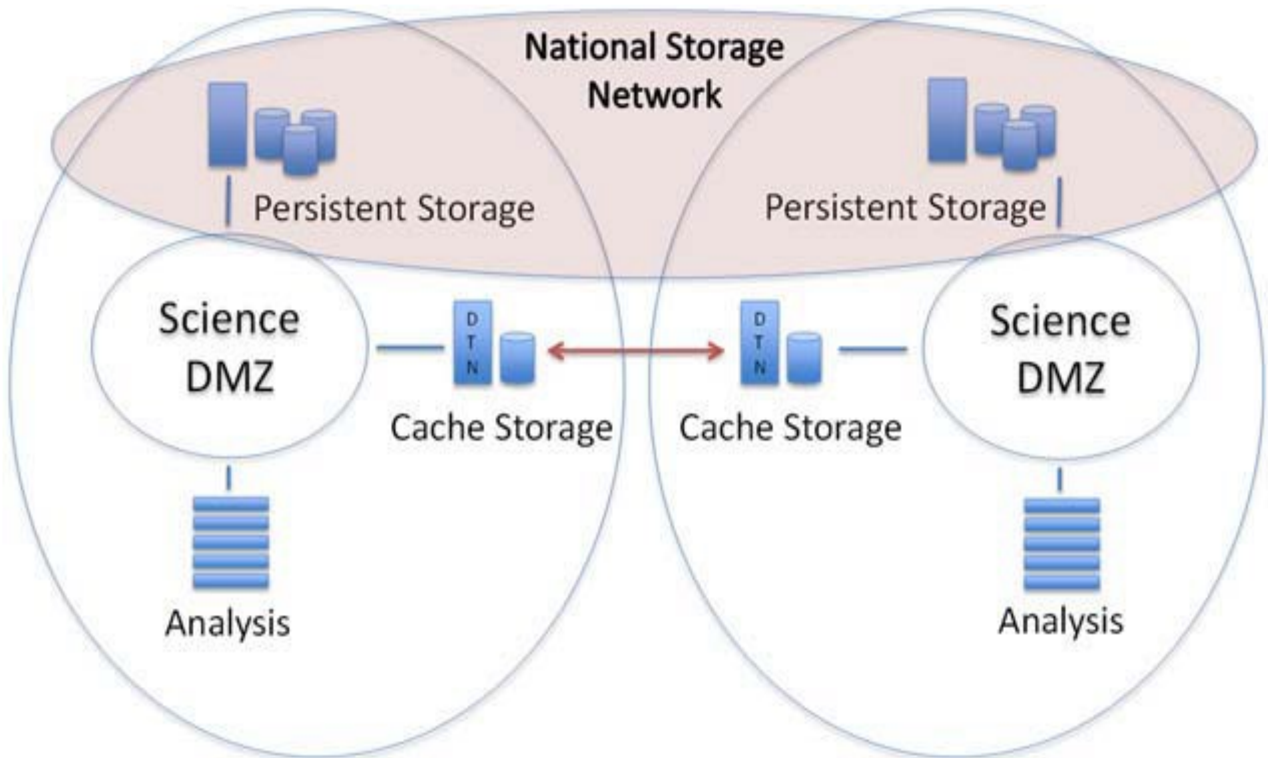
Security

- How do we make sure the system is secure?
  - *Appliances exist in DMZ*
  - *IPSEC across nodes?*
- How do we connect through the university firewalls?
  - *Second interface straddling firewall, access is subject to the university authentication*
  - *Only push/pull from the inside*
- Need lots more input from security experts

The Road Towards OSN

1. Establish public / private partnership
   - *Early seed founds from the Eric Schmidt Foundation (A. Szalay)*
   - *Pending NSF proposal with the Big Data Hubs (with C. Kirkpatrick and K. McHenry + 4 BDH)*
   - *Soon: NSF EAGER to support GLOBUS (I. Foster +S.Tuecke)*
2. Build community prototypes for different use cases, e.g.
   - *i. Move and process 1PB of satellite images to Blue Waters*
   - *ii. Move specific PB-scale MREFC data from Tier1 to Tier2 at a university for detailed sub-domain analytics (LSST)*
   - *iii. Create large simulation (cosmology or CFD) at XSEDE and move to a university to include in a NumLab*
   - *iv. Take a large set of LongTail data with small files and organize into larger containers, and explore usage models*
   - *v. Interface to cloud providers (ingress/ egress/ compute)*
3. Build community initiative for large scale funding

OSN Concept



National Storage Network
Persistent Storage
Persistent Storage
Science DMZ
D T N
Cache Storage
D T N
Cache Storage
Science DMZ
Analysis
Analysis

Cost Components

1. OSN Node Initial Purchase ($140K) per institution
2. Operations of the Command Center: monitoring, software upgrades (2.5 FTE/$350K/yr for 20 nodes, 3 FTE for 100 nodes)
3. Licensing ($5-10K/institution/yr) for things like Globus, OS support. We expect this not to exceed 10% of the hardware purchase price/year, going into saturation after a certain number of units.
4. Labor to maintain OSN Node (.25 FTE/<$35K)
5. Resource allocation (.5FTE/$70K) added to XSEDE

Institutional Responsibilities

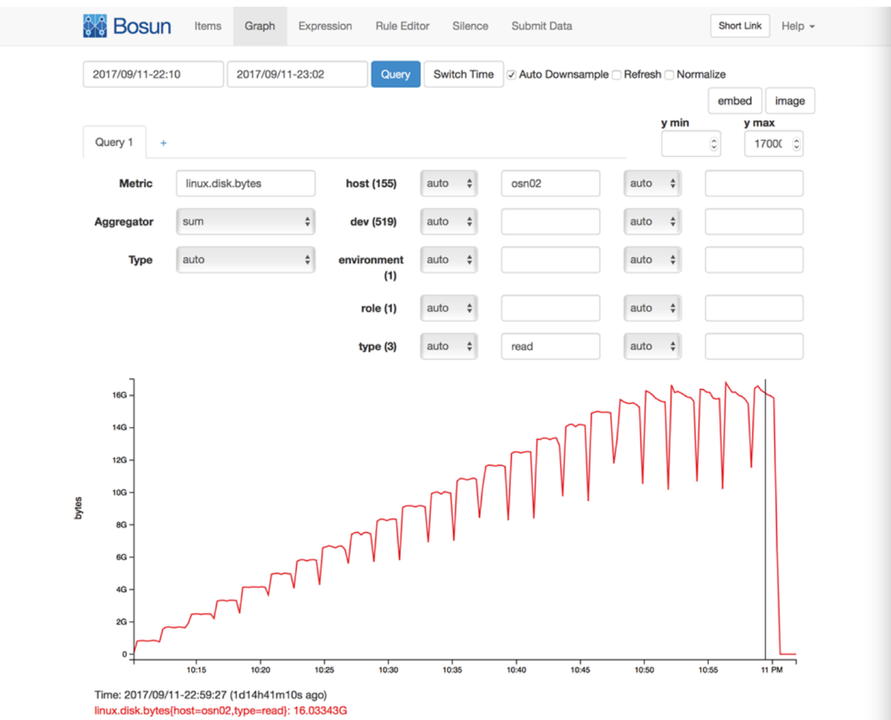
NSF	Host Institution
OSN Node Initial Purchase ($140K) per institution	Replacement equipment following an initial five year warranty period.
OSN Command Center - Technical Coordination (2.5 FTE/$350K/yr)	Labor to maintain OSN Node (25% FTE/$35K)
Licensing ($10K/institution/yr) for initial grant period	Licensing following grant period
Allocation (.5FTE/$70K) added to XSEDE	-

Projected Costs

		Yr 1	Yr 2	Yr 3	Yr 4	Yr 5	total
Hardware [units]		10	10	30	50	0	
Software [units]		10	20	50	100	100	
Command C [FTE]		2	2	2.5	3	3	
Resource A. [FTE]		0.5	0.5	0.5	1	1	
	cost/unit	cost [$K]	cost [$K]	cost [$K]	cost [$K]	cost [$K]	cost [$K]
Hardware	140	1400	1400	4200	7000	0	14000
Software license	10	100	200	500	1000	1000	2800
Command Ctr.	140	350	350	350	420	420	1890
Resource Alloc.	140	70	70	70	140	140	490
Contingency		130	130	180	240	140	820
total cost in year		2050	2150	5300	8800	1700	20000
cumulative cost		2050	4200	9500	18300	20000	

Testing Phase1 HW

Single SuperMicro server with 2 disk boxes and 2x44 8TB HGST drives, running ZFS



700TB/box, saturates around 14GB/sec

Brian Mohr and Alainna White (JHU)

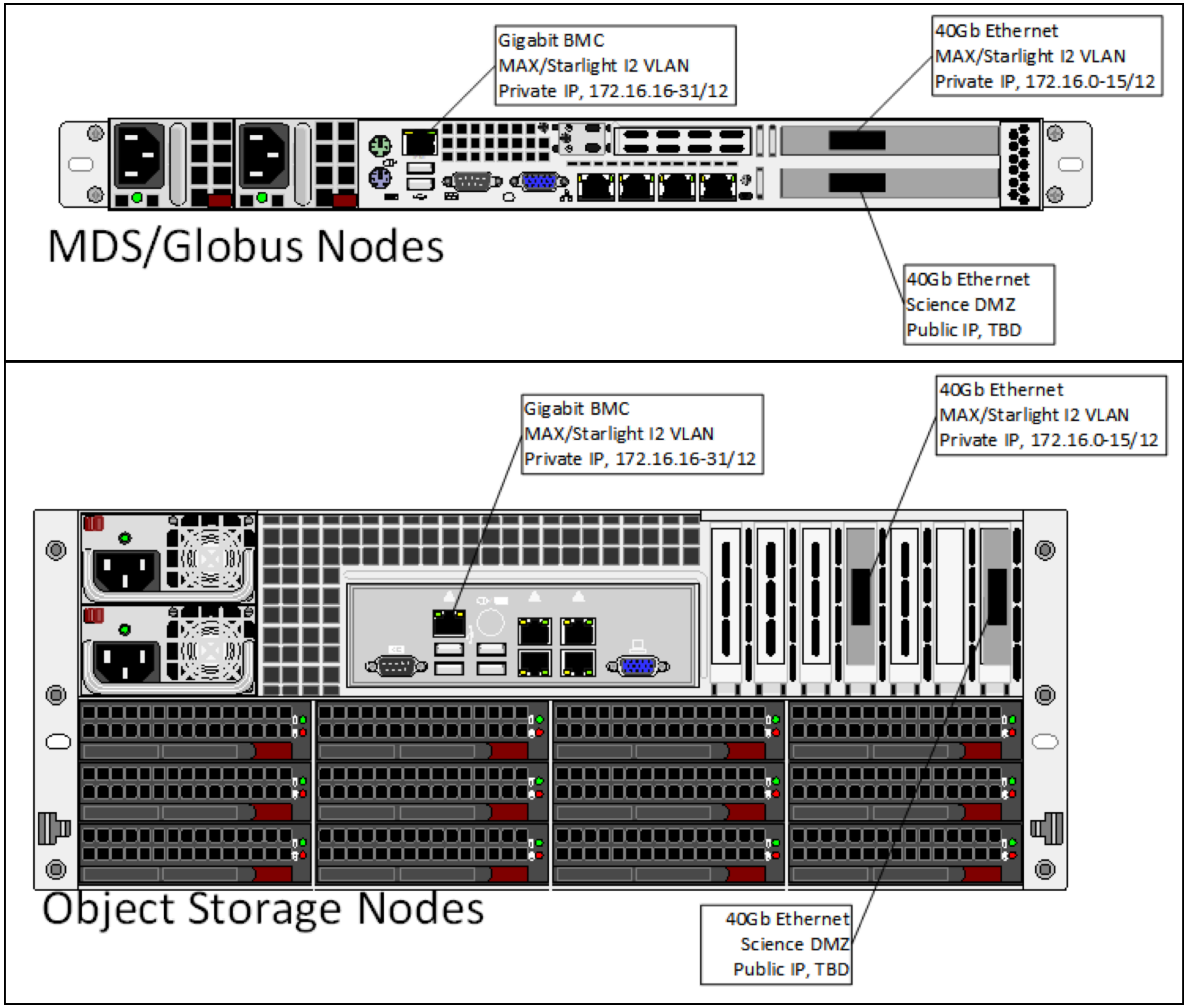
Phase2 Hardware

- Based on Supermicro and Mellanox

Item	Price ea	Quantity	Total
APC AR3340 Rack	$ 2,000	1	$ 2,000
Switched/Metered PDU	$ 1,975	2	$ 3,950
Storage Nodes (SuperMicro)	$ 18,500	5	$ 92,500
4U chassis			
1 Intel SKL6140 CPU (18/36, 2.3GHz)			
256GB 2666MHz ECC DDR4 RAM			
2 240GB SATA SSD (Boot)			
2 Mellanox ConnectX-4 EN single port			
MetaData/Globus Nodes	$ 5,000	3	$ 15,000
1U chassis			
1 Intel SKL5115 CPU (10/20, 2.4GHz)			
96GB 2666MHz ECC DDR4 RAM			
2 240GB SATA SSD (Boot)			
2 Samsung 512GB m.2 NVMe SSDs			
Mellanox 100Gb Ethernet Switch	$ 15,000	1	$ 15,000
32 ports QSFP28			
Five year warranty			
Total Hardware Cost (2017 pricing)			$ 128,450

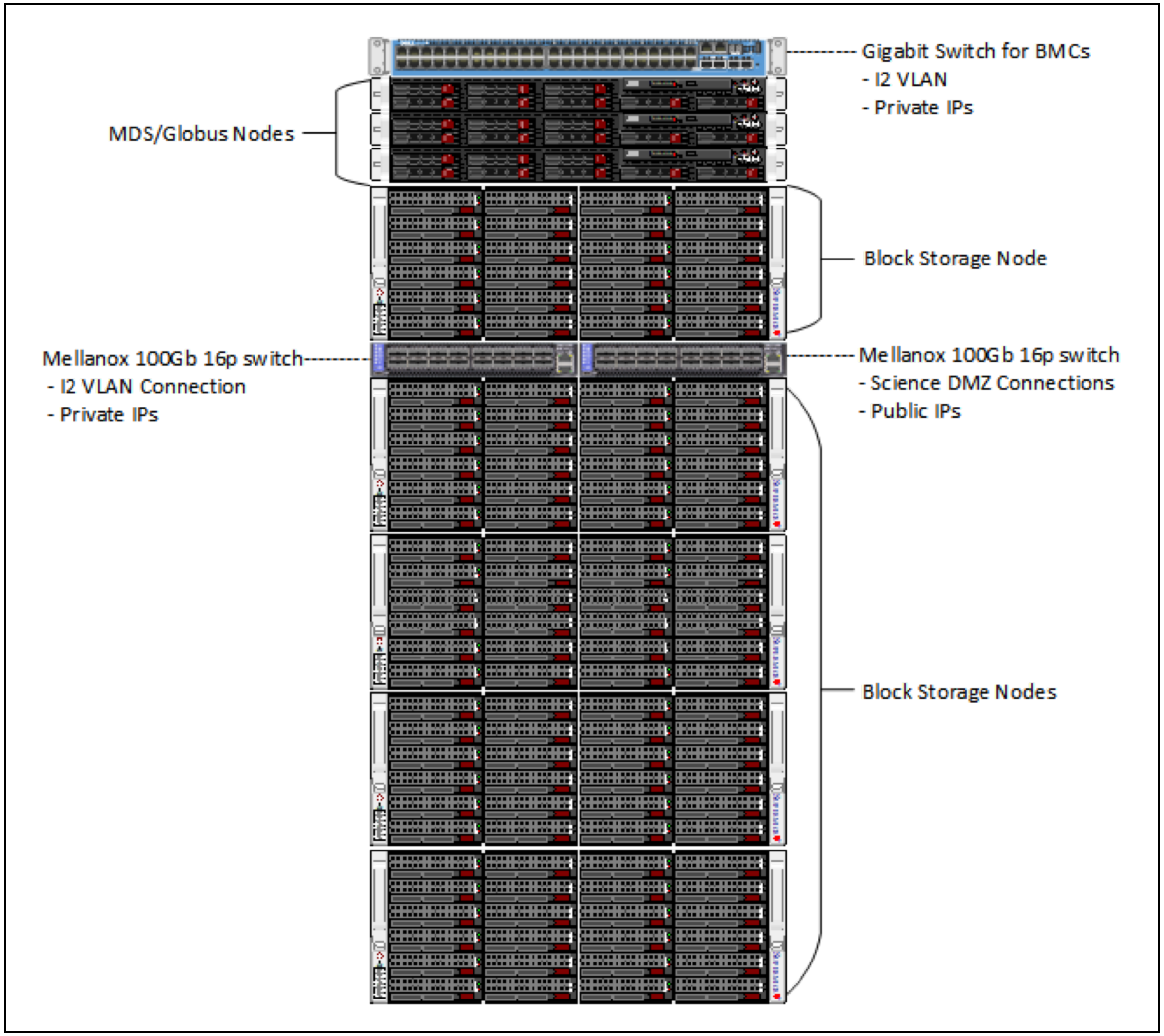
Brian Mohr and Alainna White (JHU)

Storage and Globus Nodes



Brian Mohr and Alainna White (JHU)

Single Appliance



Brian Mohr and Alainna White (JHU)

Phase2 HW Performance

	5-MINUTE TEST					
	RF	RR	RB	WF	WR	WB
osd-001	5401	2699	8042	4807	2149	7416
osd-002	5399	2671	7992	4747	2063	7296
osd-003	5408	2684	8018	4730	2069	7329
osd-004	5355	2695	7959	4707	2070	7250
osd-005	5353	2674	7939	4678	2058	7230
all	26916	13423	39950	23669	10409	36521
	30-SECOND TEST					
	RF	RR	RB	WF	WR	WB
osd-001	5596	2787	8363	5642	2959	8185
osd-002	5643	2778	8374	5679	2960	8312
osd-003	5603	2812	8393	5631	2973	8279
osd-004	5601	2799	8373	5647	2959	8253
osd-005	5590	2796	8350	5583	2948	8215
all	28033	13972	41853	28182	14799	41244

I/O numbers are in MB/sec, running Linux XFS for now

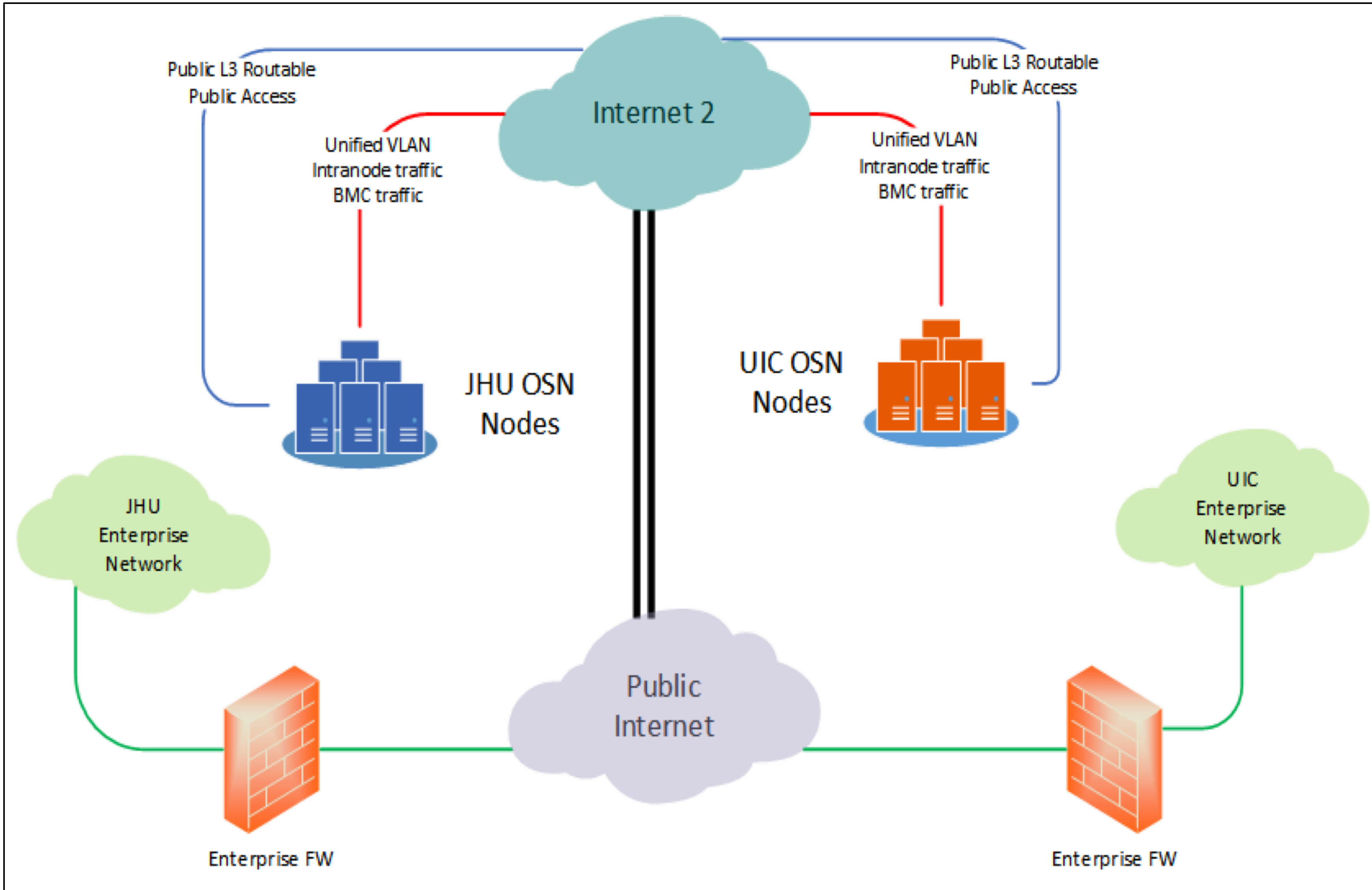
RF	Read Front backplane
RR	Read Rear backplane
RB	Read Both backplanes
WF	Write Front backplane
WR	Write Rear backplane
WB	Write Both backplanes

Brian Mohr and Alainna White (JHU)

Next Steps

- Validate low level HW performance across I2
  - *Connect appliances at JHU and StarLight*
- Deploy and optimize OSN V1
  - *initial set of authentication, authorization, data movement, and data sharing capabilities to support experimentation and validation*
- Deploy nodes at Big Data Hubs
  - *Start aggressive science use cases*
  - *Connect and test performance with cloud providers*
- Develop a design, based on community input
  - *backed up by experimental studies, for a more full-featured OSN Software Platform V2 to support full-scale production deployment*

Test Layout


Public L3 Routable
Public Access
Internet 2
Public L3 Routable
Public Access
Unified VLAN
Intranode traffic
BMC traffic
Unified VLAN
Intranode traffic
BMC traffic
JHU OSN
Nodes
UIC OSN
Nodes
JHU
Enterprise
Network
UIC
Enterprise
Network
Public
Internet
Enterprise FW
Enterprise FW

What is the Future?

- Over the next 5 years it will host and move much of the NSF generated academic data
- Will establish best practices and standards
- Open Data Services migrate one level up, built over **trusted** storage

- Some time in the next 10 years most academic data will migrate into the cloud due to economies of scale
- The OSN will not become obsolete, but becomes part of a hierarchical data caching system
- It will also provide impedance matching to the Tier0/1 to Tier2 center connectivity of MREFC instruments/projects

Summary

- High end computing has three underlying pillars
  - *Many-core computing/HPC / supercomputers*
  - *High Sped Networking*
  - *Reliable and fast data storage*
- The science community has heavily invested in first 2
  - *Supercomputer centers/XSEDE, Internet 2, CC-NIE, CC**
- Time for a coherent, national scale solution for data
  - *Needs to be distributed for wide buy-in and **TRUST***
- Only happens if the whole community gets behind it
- **Globus is at the heart of the system**