NeSI and your data: Scalable Storage

Fabrice Cantos
eResearch NZ 2019

New Zealand eScience Infrastructure

Overview

1. I/O System
2. NeSI Storage
3. Performance measure
4. Services & Future

Typical High Performance I/O System

Importance of Storage in HPC

Very few large scale applications of practical importance are NOT data intensive. *(Alok Choudhary)*

~5 % time spend in IO call vs compute

Storage RFP :

- Capacity
- Speed
- Resilience
- Budget

Moore's law in storage

Drive (NL-SAS) Capacity		Drive interface speed		CPU Single precision	
2010: 2 TB		113 MB/sec		200 Gflops	
2012: 4 TB	x7	138 MB/sec	x1.8	500 Gflops	x20
2014: 6TB		150 MB/sec		1000Gflops	
2016: 10TB		180 MB/s		1500 Gflops	
2019: 14TB		200 MB/s		2000 Gflops	

SSD Capacity		SSD speed	
2018: 4 TB		500-800 MB/sec	

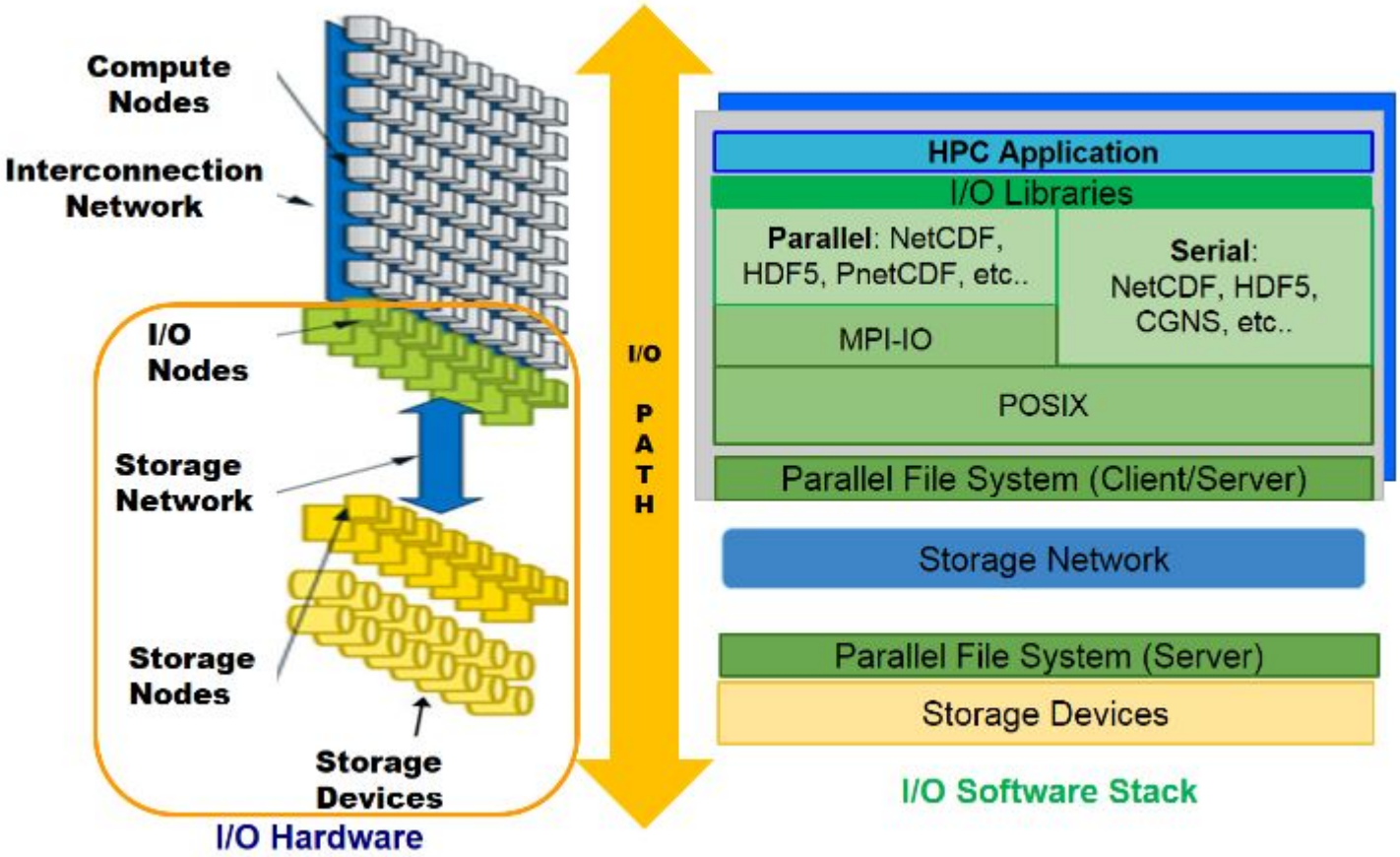
IOPS

- Block Size ?
- Serial / random ?

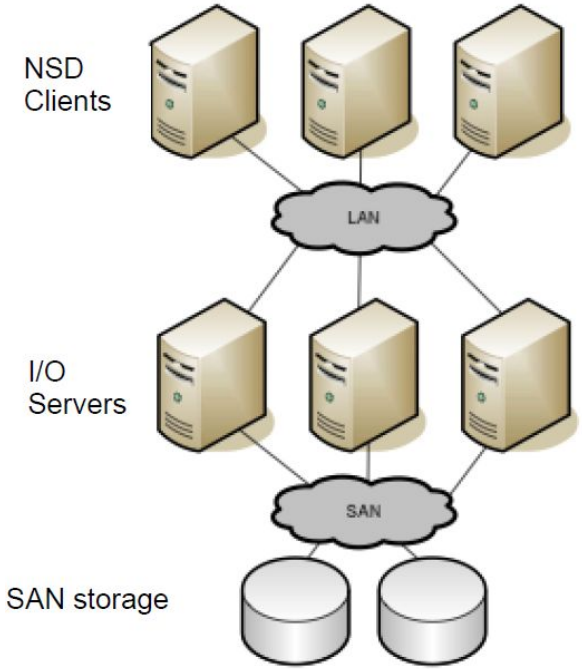
Technology	Latency
L1 CPU Cache	4 cycles (~1 nsec)
L2 CPU Cache	10 cycles
LLC CPU Cache	40 cycles
DRAM	240 cycles
NVRAM	2400 cycles
RDMA Read	6K cycles (2 usec)
FLASH Read	150K cycles (50 usec)
FLASH Write	1500K cycles (500 usec)
HDD Write min	1500K cycles (500 usec)*
HDD Read min	15000K cycles (5 msec)
HDD Read max	75000K cycles (25 msec)
Tape File Access	150000000K cycles (50 sec)

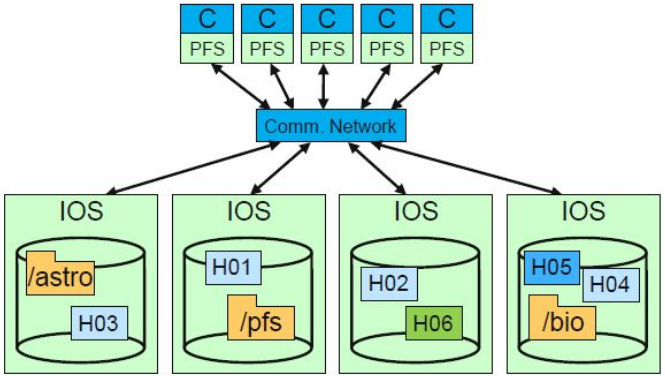
Device	IOPS
HDD	100 IOPS
SSD	~100,000 IOPS

IO System


Compute Nodes
Interconnection Network
I/O Nodes
Storage Network
Storage Nodes
Storage Devices
I/O Hardware
I/O PATH
HPC Application
I/O Libraries
Parallel: NetCDF, HDF5, PnetCDF, etc..
Serial: NetCDF, HDF5, CGNS, etc..
MPI-IO
POSIX
Parallel File System (Client/Server)
Storage Network
Parallel File System (Server)
Storage Devices
I/O Software Stack

IBM Spectrum Scale (formerly GPFS)

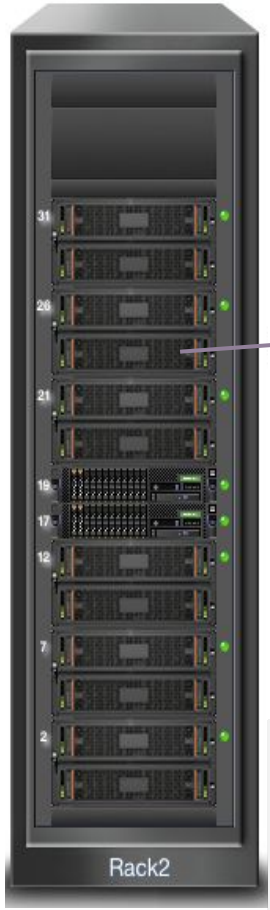


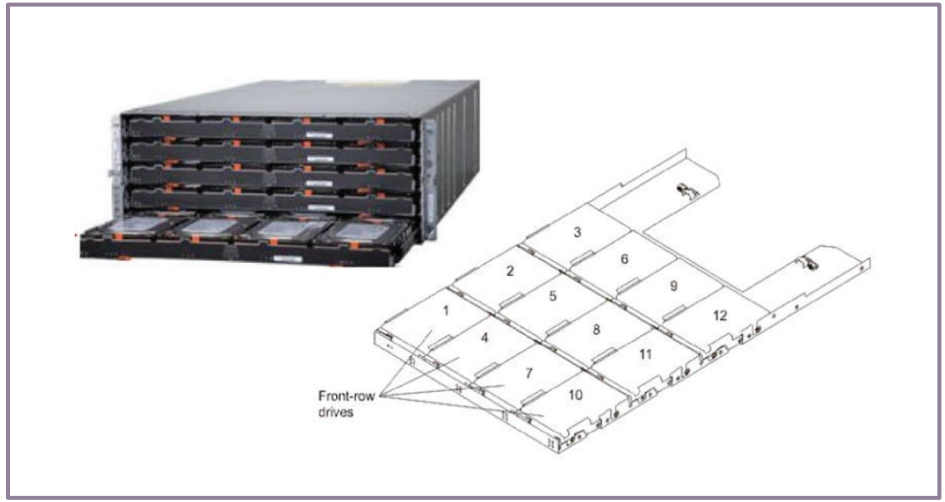


NeSI Storage

HPCF Storage

ESS Building Blocks



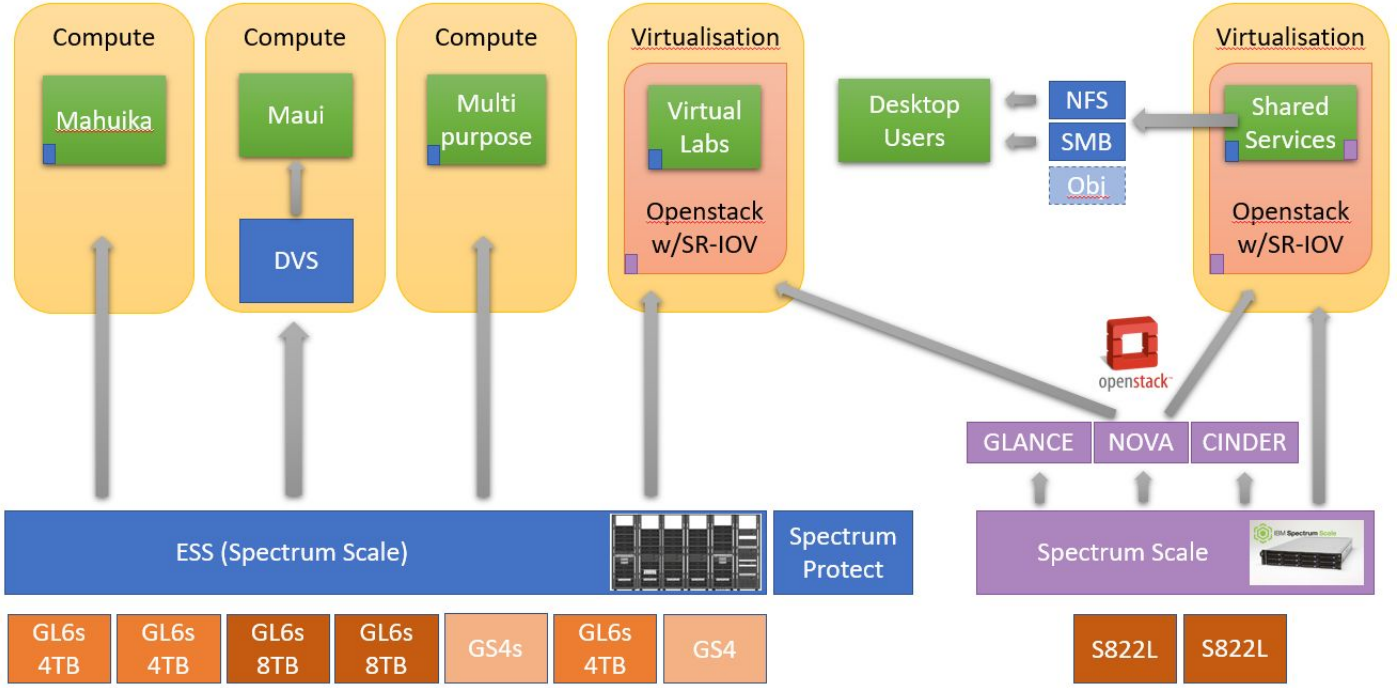


ESS Model	Qty	Per ESS Building block: Drive Size	RAW Capacity (TiB)	Recovery Groups	Pdisks per DA	Total Vdisks	Vdisk Size	Vdisk After Raid	Max Throughput per Vdisk (GB/s)	Min Throughput per Vdisk (GB/s)
IBM ESS GS4s		1 3.84TB SSD	320	2	48	100	3.2	2.56		0.3
IBM ESS GL6s		2 4TB	1780	2	251	10	178	142.4	17.5	3.5
IBM ESS GL6s		2 8TB	3560	2	251	20	178	142.4	17.5	1.75

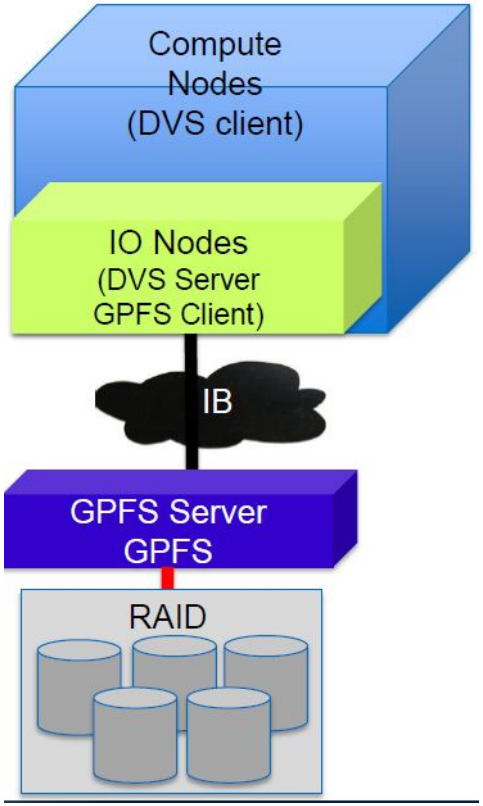
Filesystem Capacity

File System	File Set	File Set Size	% of Capacity	% of Bandwidth
/scale_wlg_persistent	/home	200 TB	26%	20%
	/project	2000 TB		
/scale_wlg_nobackup	/nobackup	5400 TB	63%	70%
/scale_wlg_nearline	/nearline	550 TB	7%	10%

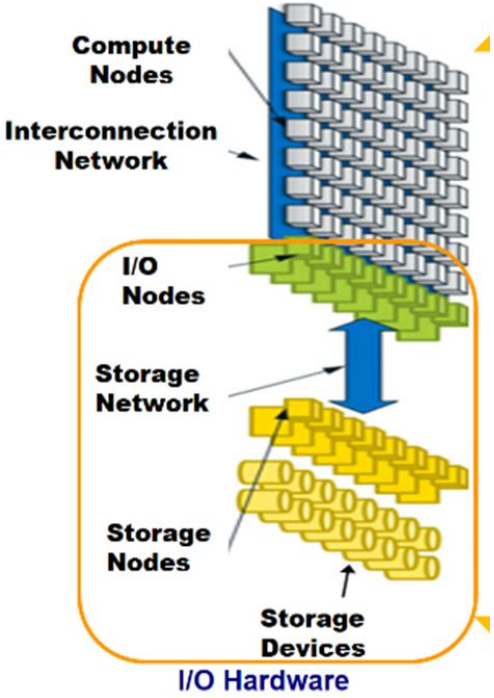
NeSI Storage



Maui IO Nodes - Cray DVS



Data Virtualization Service (DVS)



Performance Measures

IO Benchmarks

	Mahuika			Maui (via DVS)		
MDTEST (4KB creates, unique dirs.)/s	156,900			35,766 (23%!)		
	Read	**Write**	**Total**	**Read**	**Write**	**Total**
IOR **4KB** (GB/s) (Single Stream)	2.1	1.3		0.044 3.2 (w/IOBUF)	0.055 3.2 (w/IOBUF)	
IOR **8MB** (GB/s) (Single Stream)	5.1	3.3		2.5	2.3	
IOR (GB/s) **(total bandwidth)**	59.5	86.7	**146.3**	63.0	64.0	**126.9**

Comparaison with

FitzRoy

Pan

	Pan	FitzRoy	Mahuika	Maui
MDTEST (4KB creates, unique dirs.)/s		9926	156,900	35,766
(GB/s)				
IOR 4KB (Single Stream)		0.150	2.1/1.3	3.2/.3.2 (w/IOBUF!)
IOR 8MB (Single Stream)		1.1/ 1.3	5.1/3.3	2.5/2.3
IOR (total bandwidth)	4	8.1	146.3	126.9

Services

Services

- Snapshot or /home & /nesi/projects
- Quota Management
- Backups for disaster recovery /home & /nesi/projects
- Data Transfer Service (Globus)

Future Work

- Nearline storage with librarian tool
- Replication of filespace with NeSI collaborators
- Data Object service export
- SSD pool for hot pool data
- Read & Write local cache.
- Maui Native GPFS client

NeSI @ eResearch NZ - Talks & Workshops:

Monday 18 Feb

2:10 - 2:30 pm
Understanding research drivers for NZ's advanced research computing

2:30 - 2:50 pm
How NeSI helps Manaaki Whenua - Landcare Research monitor land cover changes

3:30 - 3:50 pm
NeSI Futures

4:30 - 5:30 pm
Training Community BoF

4:50 - 5:10 pm
Catering to domain (Genomics) specific eResearch needs

Tuesday 19 Feb

11:00 - 11:20 am
The NeSI HPC Computer and Data Analytics Service

11:00 am - 12:30 pm
Open Space Session - BYO topics!

1:30 - 1:50 pm
Visualization capabilities of NeSI's new high performance computers

1:30 - 1:50 pm
A day in the life of NeSI's Apps Support

1:50 - 2:10 pm
NeSI and your data: Scalable storage

1:50 - 2:10 pm
Research Software Engineering (RSE): What's in a name?

Tuesday 19 Feb (cont.)

2:10 - 2:30 pm
Scaling new data services at NeSI

2:30 - 2:50 pm
Insight into the new NeSI platforms

3:30 - 4:30 pm
(Inter)national collaborative research infrastructure strategies BoF

3:30 - 4:30 pm
Research Software Engineering BoF

4:30 - 5:30 pm
Research Cloud NZ BoF

Wednesday 20 Feb

11:10 am - 4:00 pm
Hacky Hour / Bring your own code