NVMe-based BeeGFS as a next-generation scratch filesystem for High Performance Computing and Artificial Intelligence / Machine Learning workloads

Greg Lehman, Igor Zupanovic, Jacob Anders, Rene Tyhouse, Garry Swan, Joseph Antony

13 March 2019

Who are we?

www.csiro.au

At CSIRO we do the extraordinary every day.

As Australia's national science agency, we innovate for tomorrow while delivering impact today – for our customers, all Australians and the world.

CSIRO innovations

WiFi
WLAN

POLYMER BANKNOTES

AEROGARD

BARLEYmax™

RELENZA FLU TREATMENT

TOTAL WELLBEING DIET

HENDRA VACCINE

EXTENDED WEAR CONTACTS

SOFTLY WASHING LIQUID

SELF TWISTING YARN

RAFT POLYMERISATION

NOVACQ™ PRAWN FEED

CSIRO IM&T Scientific Computing

CSIRO

IM&T Scientific Computing

~100
talented staff

80+
collaborative eResearch projects every 6 months

Working with over
2600+
customers

~5 Million
CPU hours per month

1500m²
data centre floor space across Australia

~2 Petaflops
aggregate performance

2000
published collections in data.csiro.au

~40PB
primary data holdings

IMT eResearch Supports End-to-End Science

Information Services

Advanced Collaboration

Workflow Services

Outreach & eResearch Planning

Research Data Services

Scientific Computing & Visualisation

eResearch services supporting the science workflow
(Diagram based on Bath University's 'Research360 InstitutionalResearch Lifecycle')

IMT eEnablement Services:
High Speed Networks, Application Development, Tele-presence, Office Productivity, Collaboration Tools

Scientific Computing

Platforms Group

- Systems
- Data
- Facilities

National Collaborations

Services Group

- Science Applications
- Visualisation
- User Services
- Data Processing

Systems

The Systems team manages:

- **Pearcey** – General purpose cluster. Upgraded to 230 Haswell nodes, 4480 cores, FDR Infiniband
- **Ruby** – SGI UV3000 NUMA System hosting 8TB and 640 cores from a single operating system
- **Bragg** – 384 Nvidia Kepler GPU's and Xeon Phi enabled system; 128 nodes. Top 500 System ~ 1M CUDA cores
- **HTCondor** – Cycle harvesting service across ~ 4400 desktops (360 CPU years of compute in the last year)

Systems services are:

- **Used by > 2600 CSIRO scientists & affiliates**
  - ~4 million CPU hours of HPC jobs per month
  - ~1 million CPU hours of HTCondor jobs per month
- **An essential contribution to CSIRO's science and research portfolio**

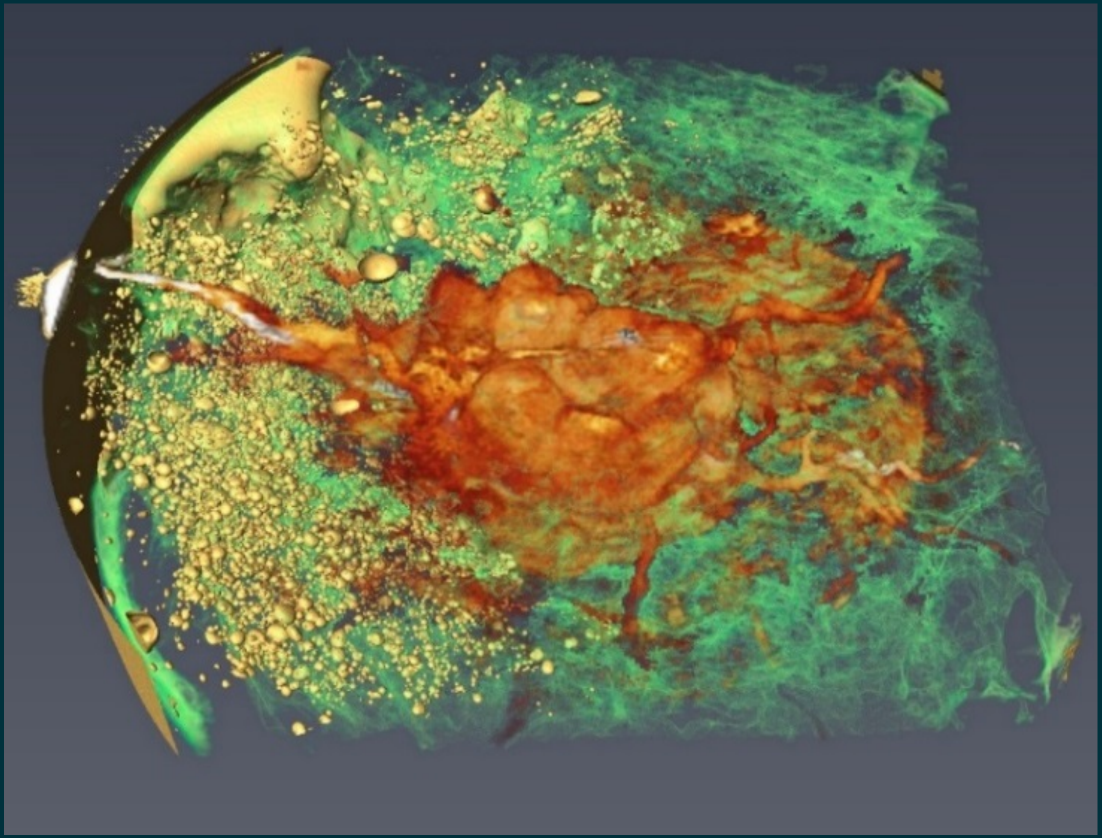
The CSIRO 'Bragg' and 'Pearcey' supercomputers

Scientific Use-Cases Driving Storage

GPU-based Tomographic Reconstruction

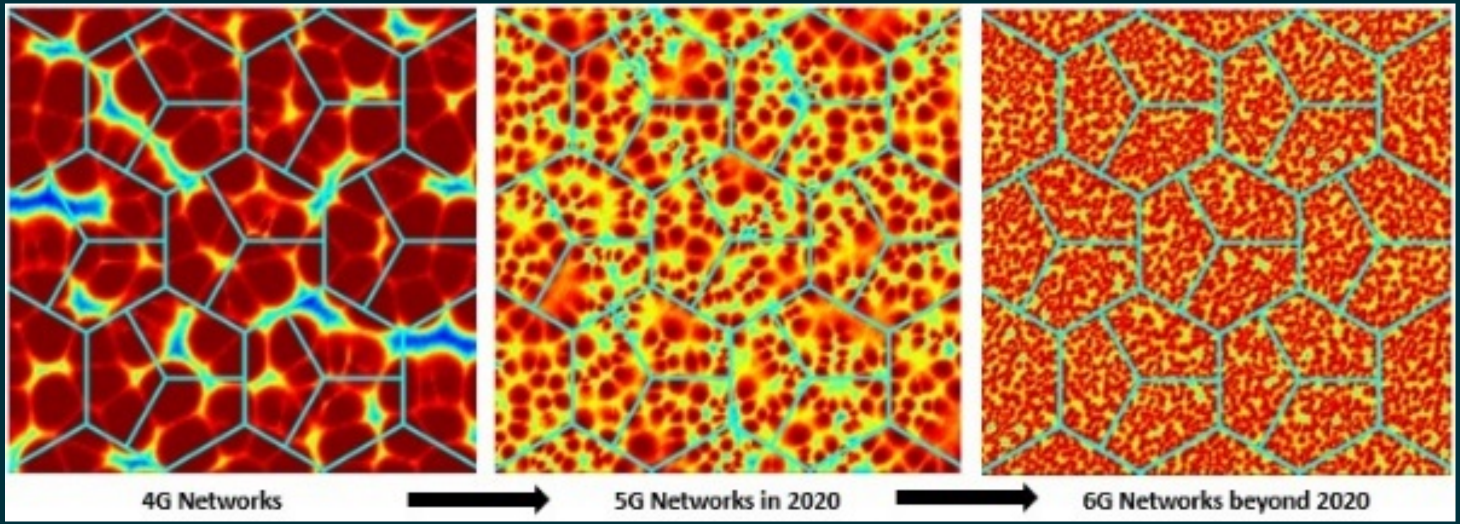
3D CT Reconstruction of an excised human breast containing a tumour (in red).

Imaged at the Imaging and Medical Beamline (IMBL) at the Australian Synchrotron



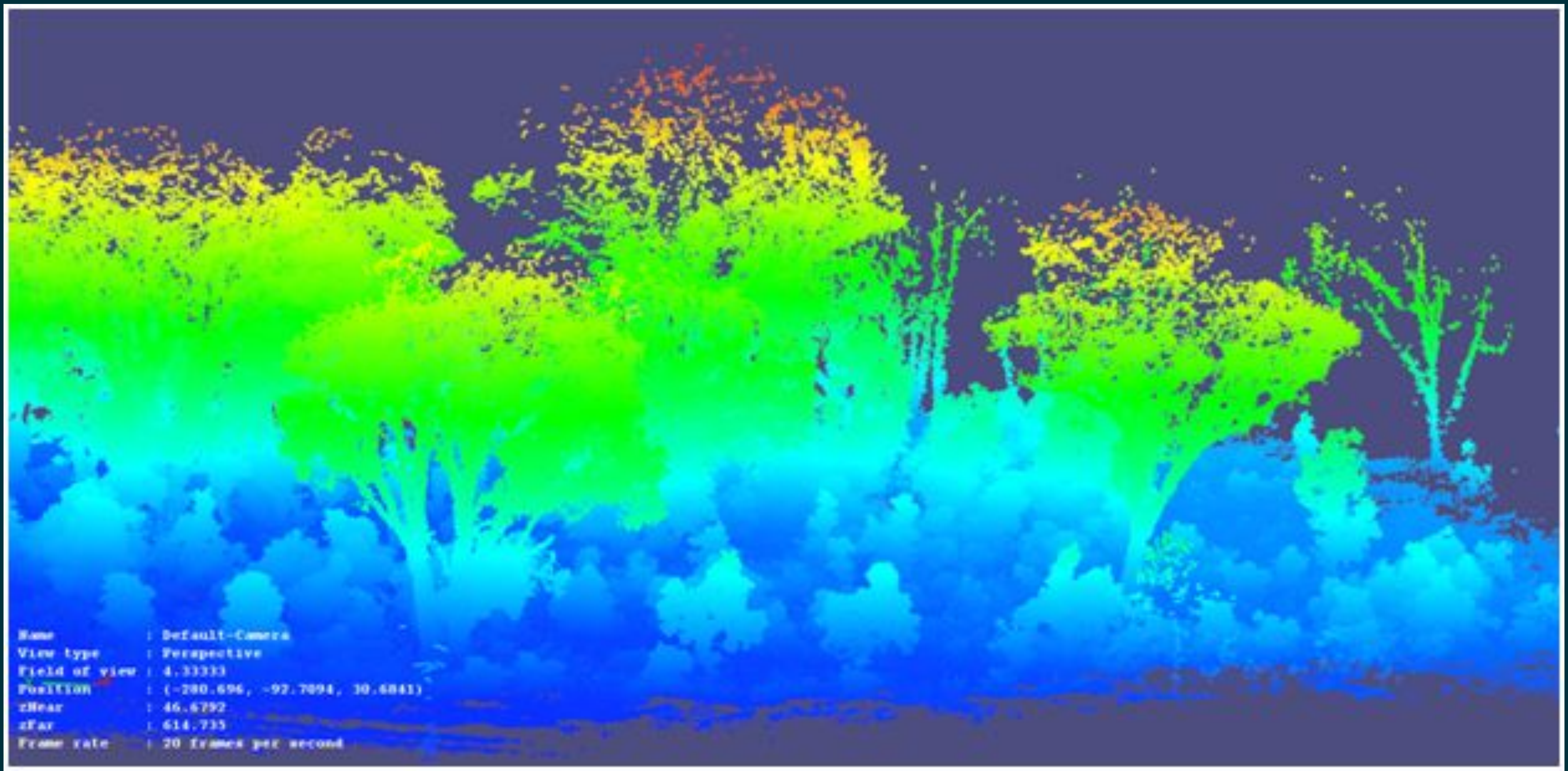
3D CT Reconstruction of breast tumour
Imaging and Medical Beamline, Australian Synchrotron

Simulations of 5G Wireless and Beyond



Evaluation of large scale network end-points from 4G, 5G wireless networks and beyond

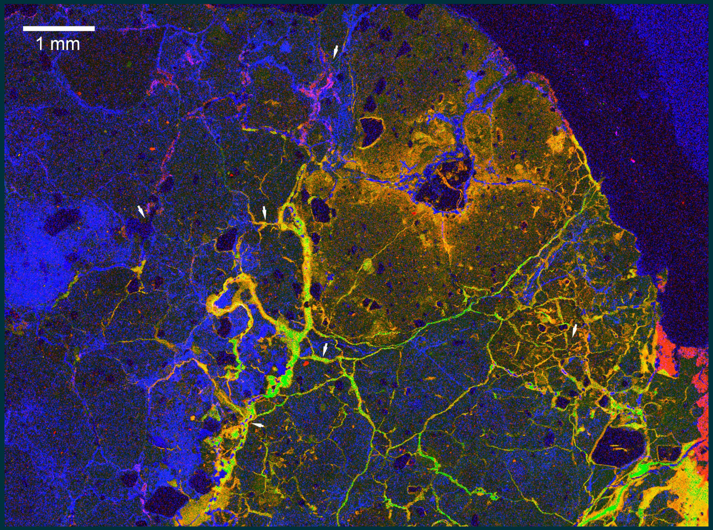
3D Vegetation Mapping and Analysis

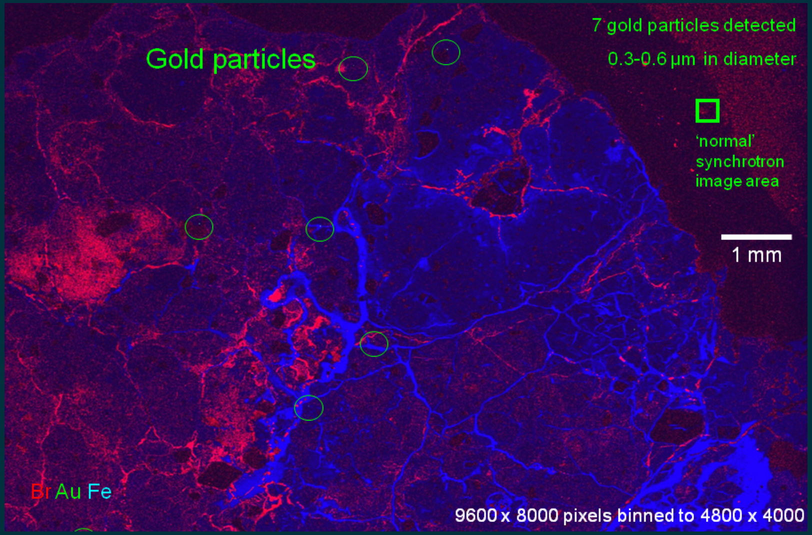


Generating vegetation cover maps in 3D from data acquired via a Zebedee handheld laser scanner

Maia X-Ray Imaging

- Synchrotron x-ray fluorescence (SXRF) imaging is a powerful technique used in the biological, geological, materials and environmental sciences, medicine and cultural heritage
- Digital images of microscopic or nanoscopic detail are built, pixel by pixel, by scanning the sample through the beam
- The resulting x-ray fluorescence radiation is characteristic of the chemical elements in that pixel. This is used to quantify the chemical composition of the sample, including important trace elements, and to build up element images of the sample
- CSIRO worked with the Brookhaven National Laboratory (BNL) to develop the Maia x-ray microprobe detector system.
- The system combines BNL's custom detector arrays and application-specific integrated circuits, with our high-speed data capture hardware and real-time spectral analysis algorithms
- Reconstruction algorithms run on HPC resources and need fast storage





Maia RGB image collected at the Australian Synchrotron of a clay sample from the Mt Gibson gold deposit in Western Australia (green = iron, blue = bromine, red = arsenic).

Capable Storage Underpins Next Generation Applied Industrial Science Applications

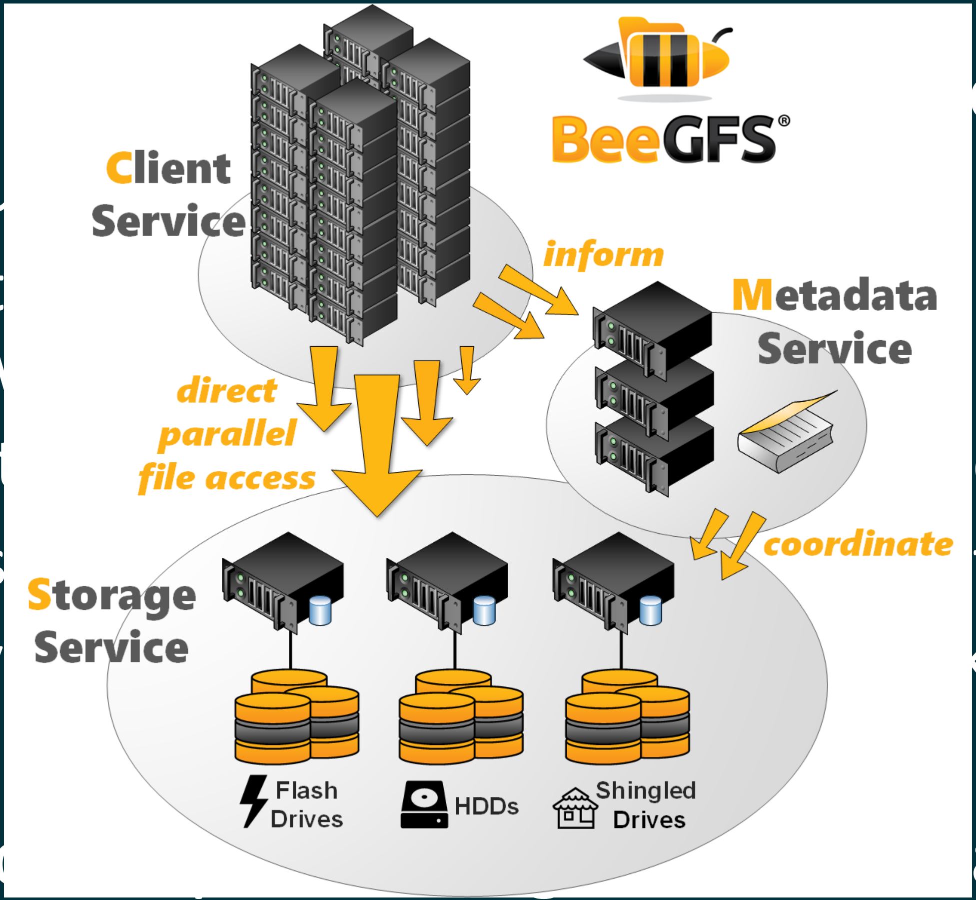
Storage Drivers

- The challenge faced by the IM&T Scientific Computing Team was to
  - Simultaneously optimize for high IOPS and high bandwidth workloads
  - Needs to be extremely power and rack efficient
  - Needs to be parallel, POSIX compliant filesystem
  - Ability to support HPC and AI/ML workloads

- We ended up choosing an NVMe based system driven by the BeeGFS filesystem

Storage Drivers

- The ch cientific Compu
  - Simult and high bandw
  - Need t efficient
  - Needs filesystem
  - Ability kloads

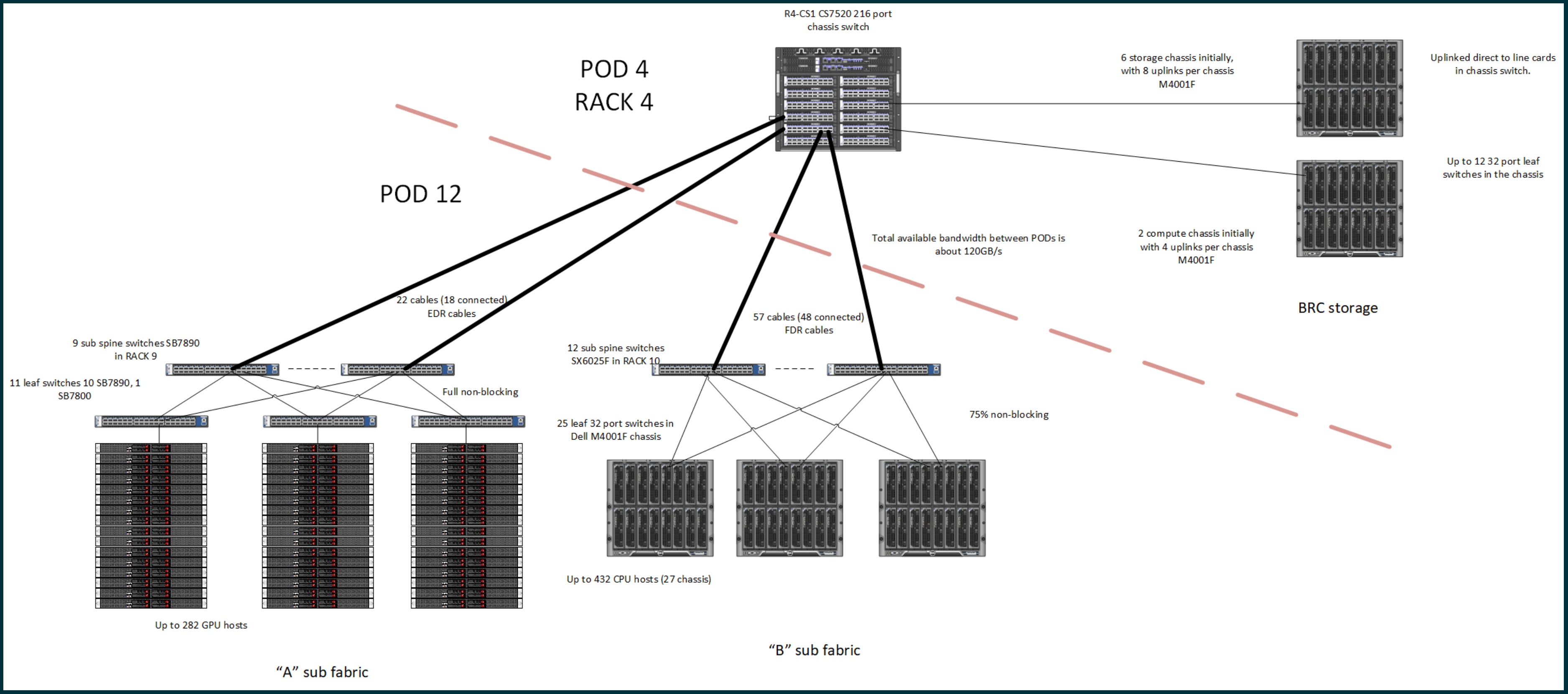


- We en ased system driven by the BeeGFS filesystem

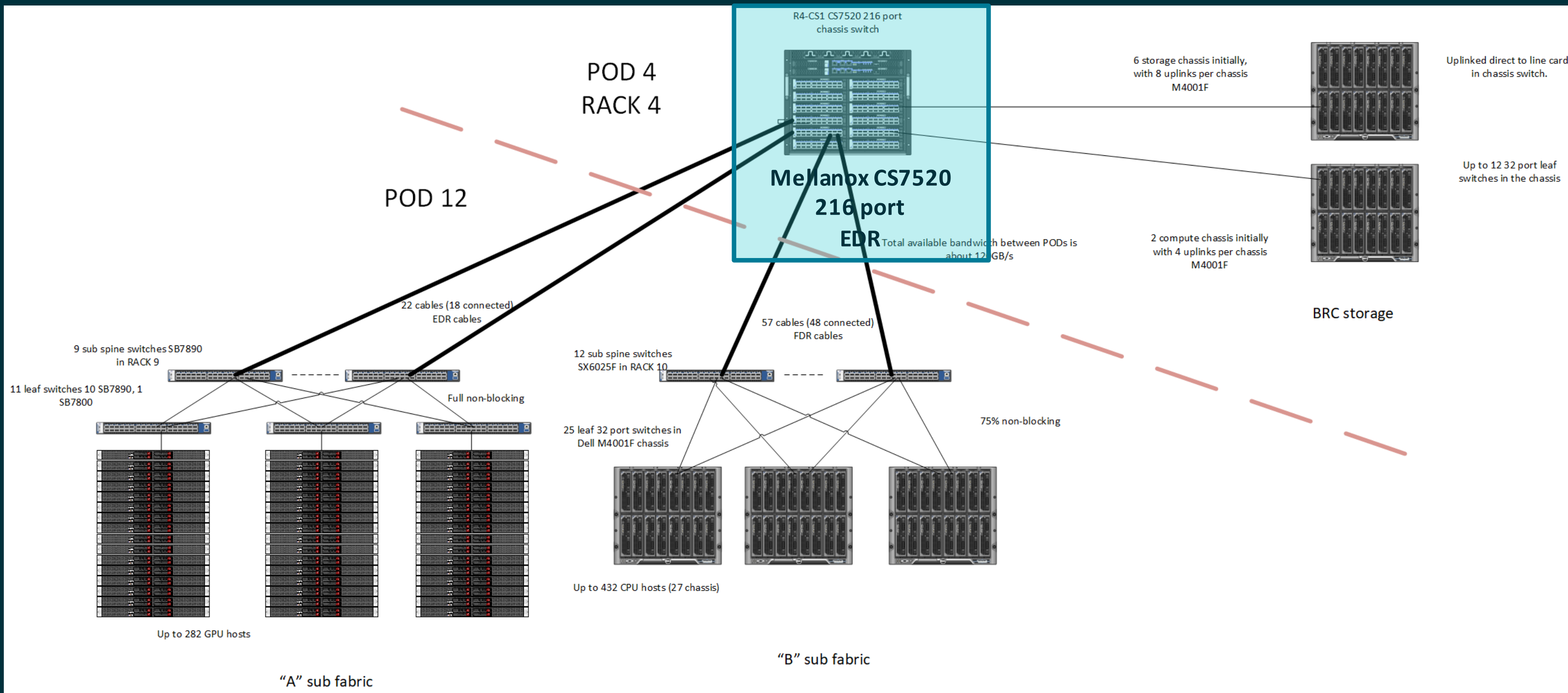
Hardware Building Blocks

- Current Networking Topology
- Metadata Service Building Blocks
- Storage Service Building Blocks

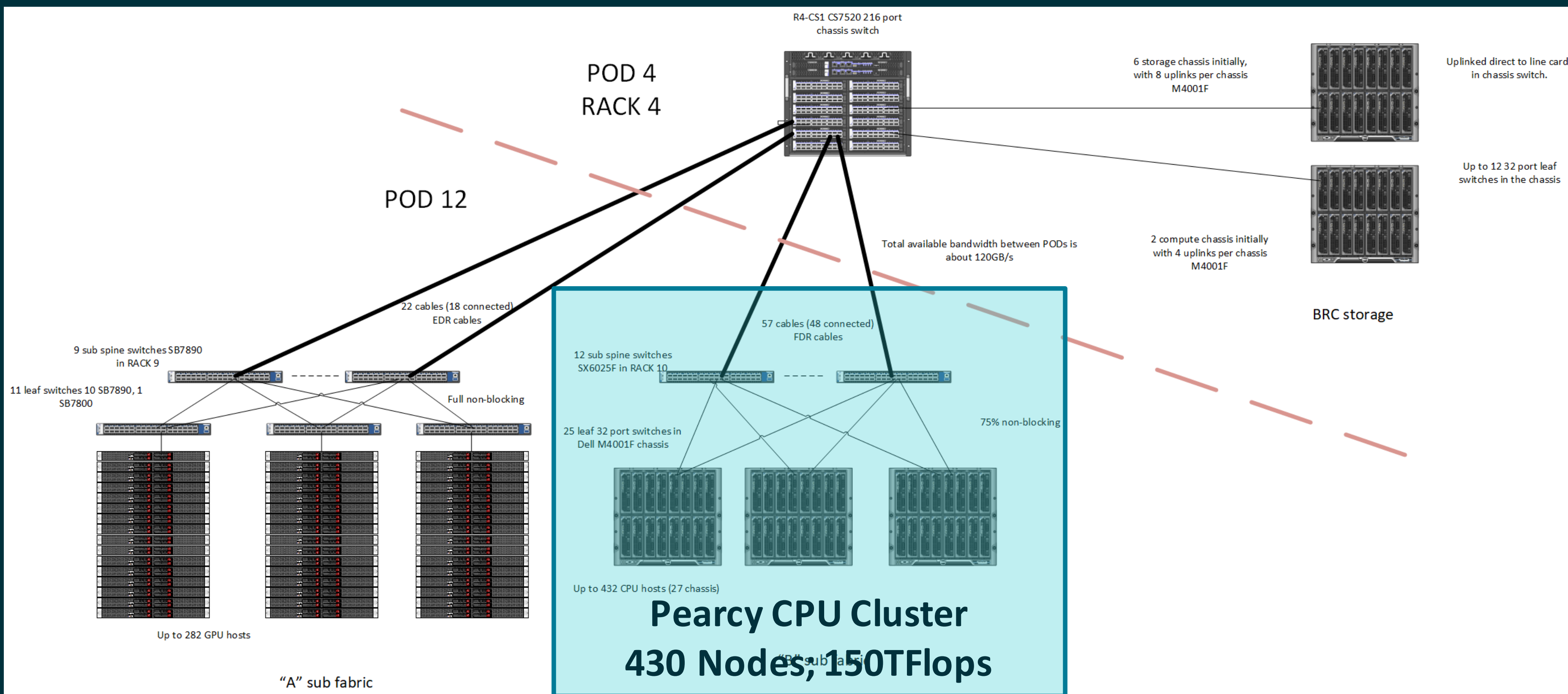
Switch Centric View of Compute and Storage Clusters at the CDC Facility



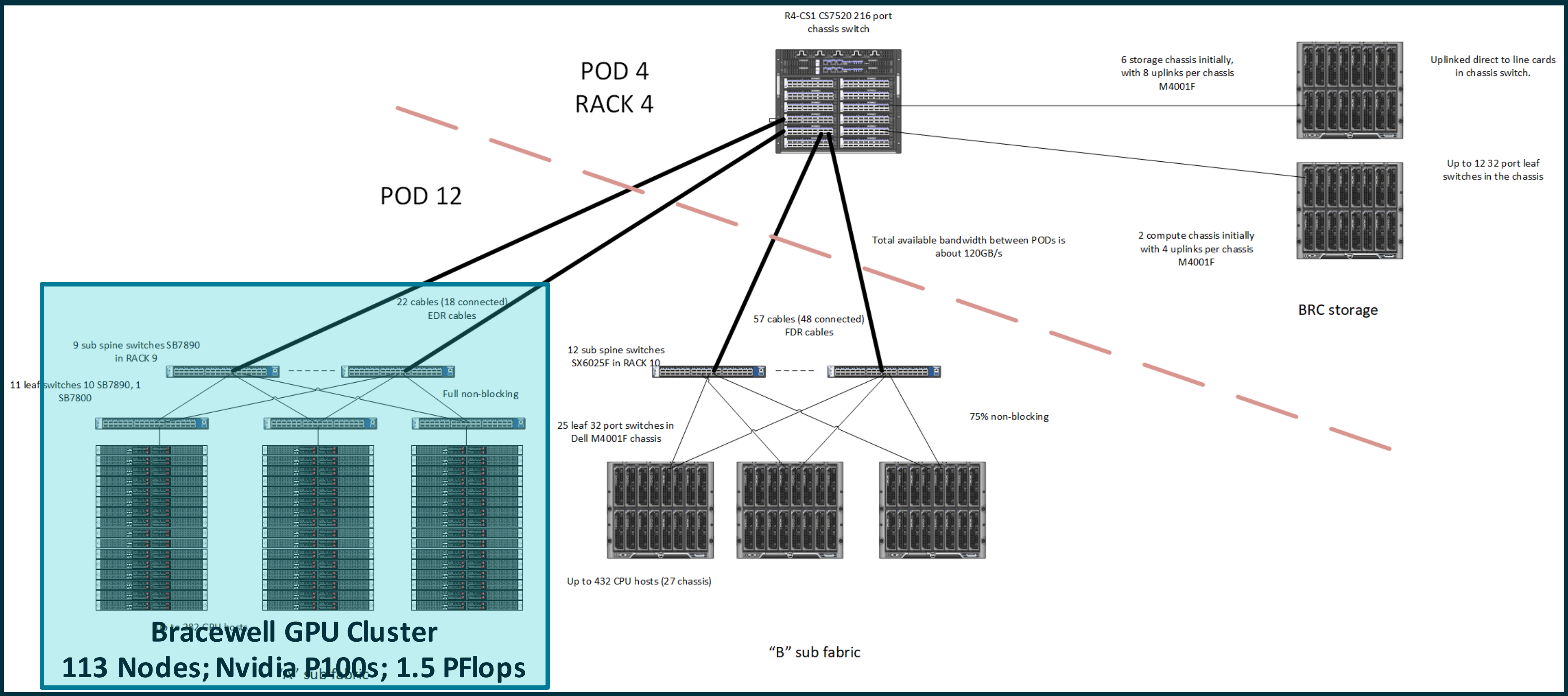
Switch Centric View of Compute and Storage Clusters at the CDC Facility



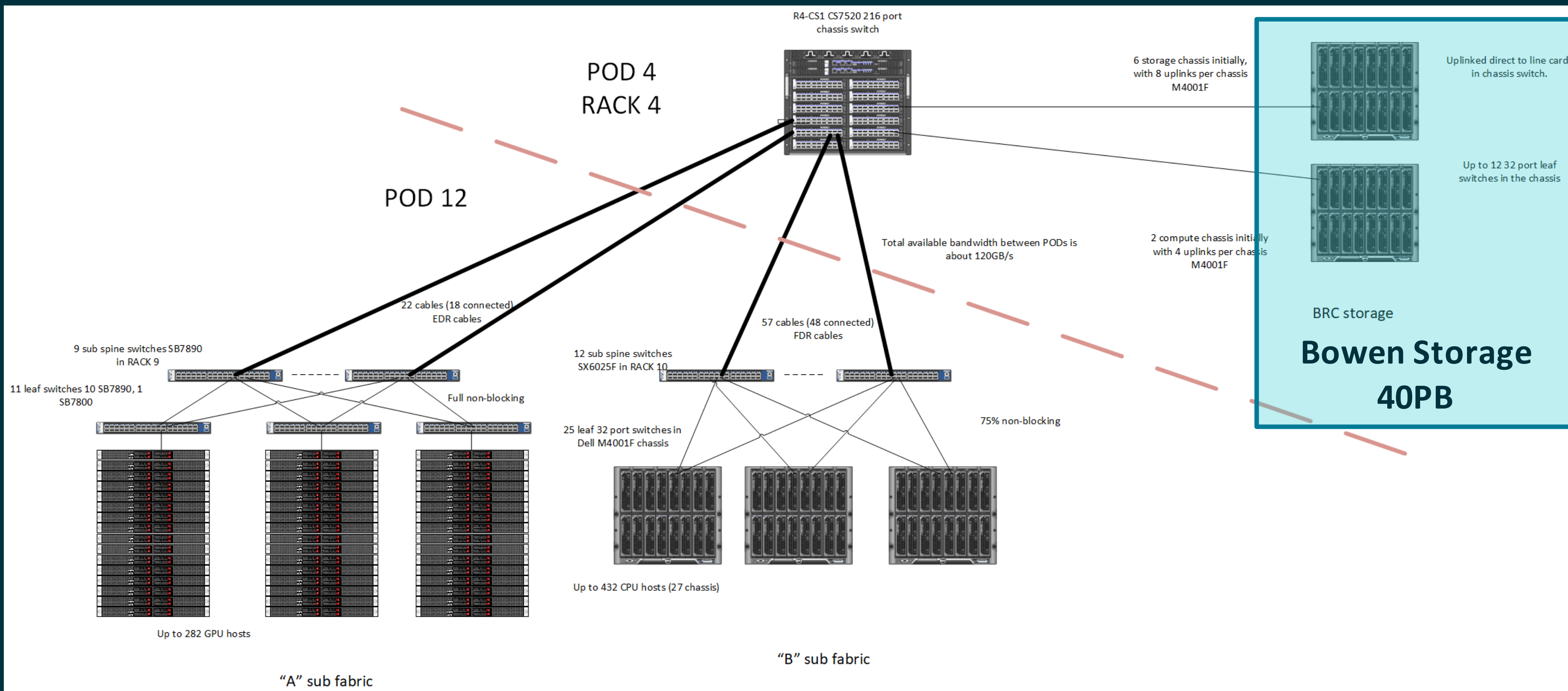
Switch Centric View of Compute and Storage Clusters at the CDC Facility



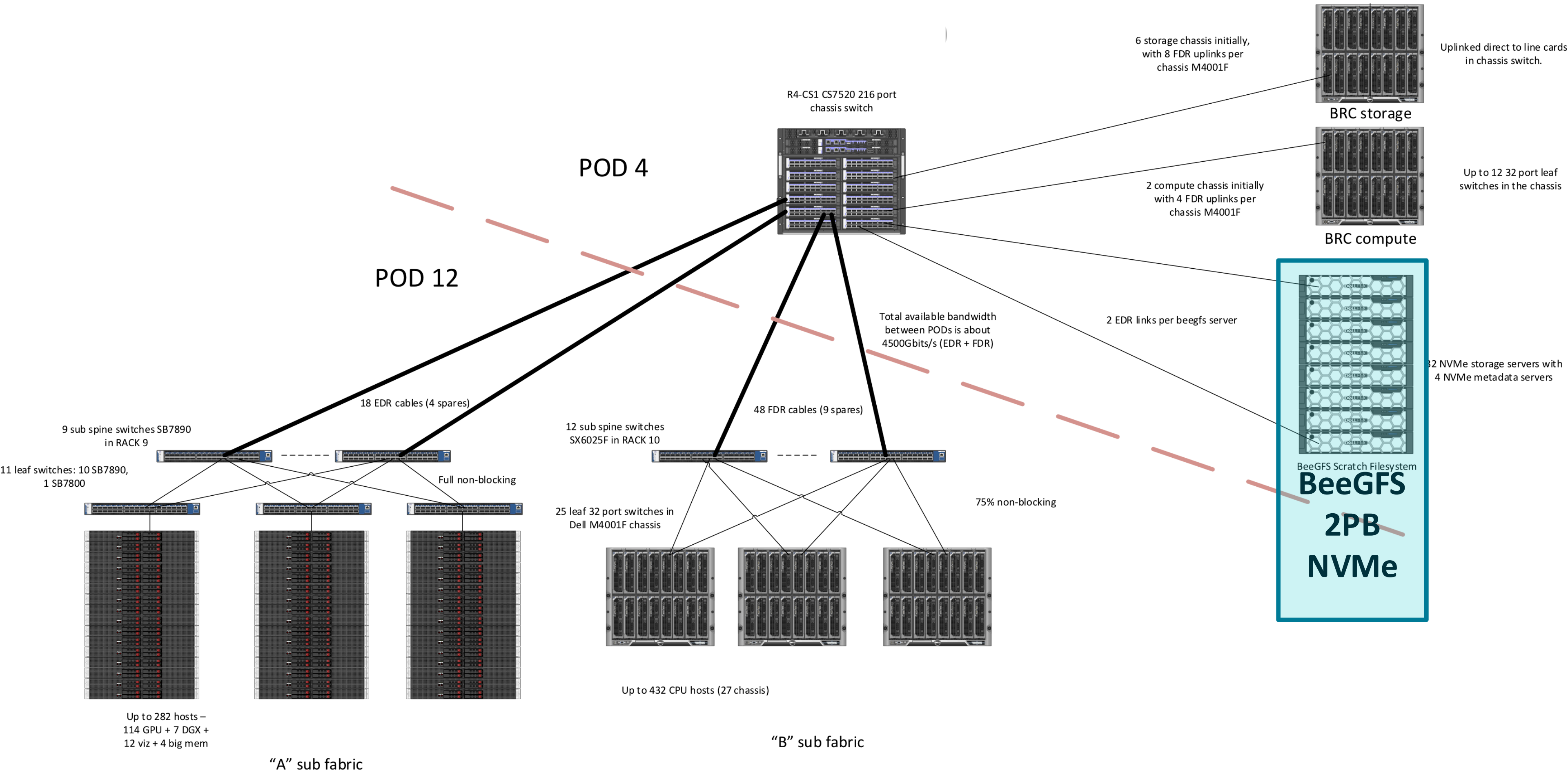
Switch Centric View of Compute and Storage Clusters at the CDC Facility



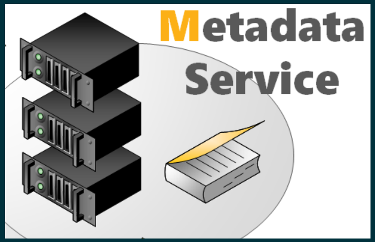
Switch Centric View of Compute and Storage Clusters at the CDC Facility



Switch Centric View of Compute and Storage Clusters at the CDC Facility

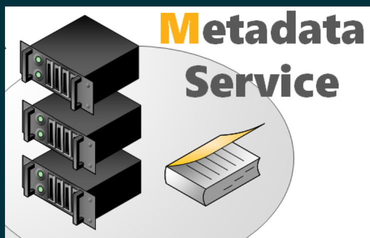


Metadata Service Building Blocks 1/2



- 4 **Metadata** servers
  - DellEMC R440
  - Dual Intel 6154
    - *3.0GHz 12 core, 384GB*
  - Dual *ConnectX-5 EDR*

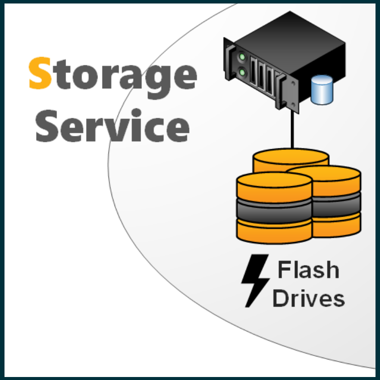
Metadata Service Building Blocks 2/2



- 4 **Metadata** servers
  - DellEMC R440
  - Dual Intel 6154
    - *3.0GHz 12 core, 384GB*
  - Dual *ConnectX-5 EDR*

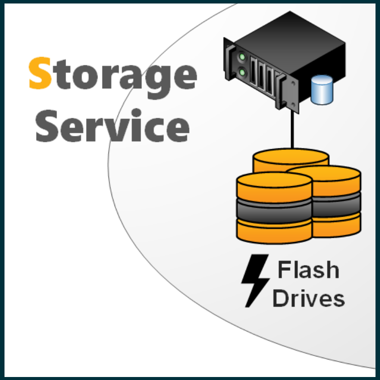
- **Intel P4600**
  - 24 x ***1.6TB Intel P4600 NVMe***
  - 3D NAND TLC
  - Random Reads ~ 5.6 million IOPS
  - Random Writes ~ 1.8 million IOPS
  - Active Power
    - 14.2 Watts (Write); 9 Watts (Read)
  - Idle Power
    - < 5 Watts

Storage Service Building Blocks 1/2



- 32 **Storage** servers
  - DellEMC R740xd
  - Dual Intel 6148
    - *2.4GHz 20 core, 192GB*
  - Dual *ConnectX-5 EDR*

Storage Service Building Blocks 2/2

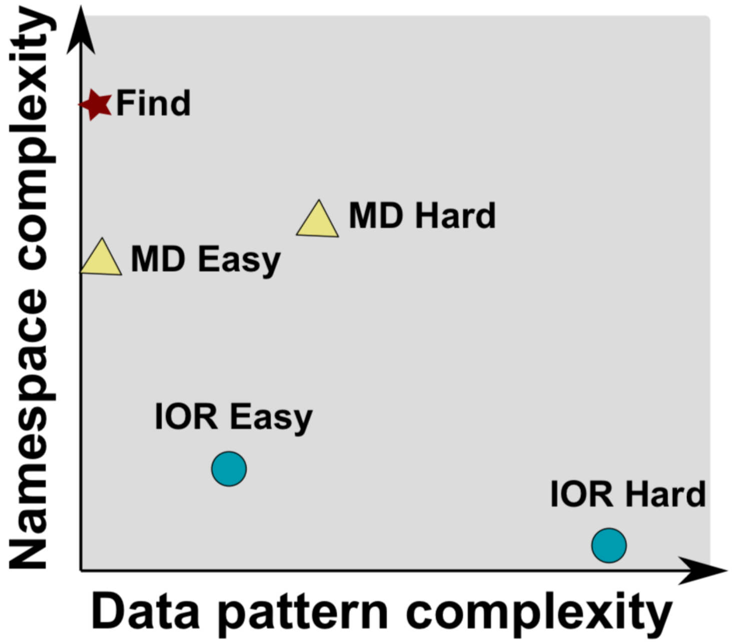


- 32 **Storage** servers
  - DellEMC R740xd
  - Dual Intel 6148
    - *2.4GHz 20 core, 192GB*
  - Dual *ConnectX-5 EDR*
- **Intel P4600**
  - 24 x ***3.2TB Intel P4600 NVMe***
  - 3D NAND TLC
  - Random Reads ~ 6.4 million IOPS
  - Random Writes ~ 2.3 million IOPS
  - Active Power
    - 21 Watts (Write); 10 Watts (Read)
  - Idle Power
    - < 5 Watts

IO500 Benchmark

www.csiro.au

Covered Access Patterns



- IOR-easy: optimal (large sequential) performance on POSIX files
- IOR-hard: small random performance on a shared POSIX file
- MD-easy: mdtest, per rank directory, with empty files
- MD-hard: more complex metadata operations on 3900 byte files
- find: query and filter files based on name and creation time
- Executing different patterns currently not covered (another dimension)

https://www.vi4io.org/_media/17-benchmarking-ws-io500.pdf

Benchmarking Phases

1. Create
   1. IOR-easy write
   2. IOR-hard write
   3. MD-easy create
   4. MD-hard create
2. Access
   5. IOR-easy read
   6. MD-easy stat
   7. IOR-hard read
   8. MD-hard stat
   9. find files
3. Cleanup
   10. MD-easy remove
   11. MD-hard remove

https://www.vi4io.org/_media/17-benchmarking-ws-io500.pdf

IO500 10 Node Challenge – ZFS backend

SC'18 Results

#	information							io500		
	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw	md
									GiB/s	kIOP/s
1	WekaIO		WekaIO		10	700	zip	58.25	27.05	125.43
2	Oak Ridge National Laboratory	Summit	IBM	Spectrum Scale	10	160	zip	44.30	9.84	199.48
3	DDN	Bancholab	DDN	Lustre	10	240	zip	31.50	6.33	156.69
4	IBM	Sonasad	IBM	Spectrum Scale	10	10	zip	24.24	4.57	128.61
5	KAUST	Shaheen II	Cray	DataWarp	10	80	zip	13.99	14.45	13.53

10 Clients; 16 Threads

```
[RESULT] BW   phase 1          ior_easy_write      40.874 GB/s : time 316.20 seconds
[RESULT] BW   phase 2          ior_hard_write       0.346 GB/s : time 506.62 seconds
[RESULT] BW   phase 3           ior_easy_read      54.308 GB/s : time 237.98 seconds
[RESULT] BW   phase 4           ior_hard_read       3.335 GB/s : time  52.58 seconds
[RESULT] IOPS phase 1       mdtest_easy_write     133.583 kiops : time 821.38 seconds
[RESULT] IOPS phase 2       mdtest_hard_write     138.079 kiops : time 312.33 seconds
[RESULT] IOPS phase 3                    find     388.800 kiops : time 133.70 seconds
[RESULT] IOPS phase 4        mdtest_easy_stat     367.286 kiops : time 133.35 seconds
[RESULT] IOPS phase 5        mdtest_hard_stat     137.877 kiops : time  28.31 seconds
[RESULT] IOPS phase 6      mdtest_easy_delete      54.482 kiops : time 920.06 seconds
[RESULT] IOPS phase 7        mdtest_hard_read      34.404 kiops : time 108.13 seconds
[RESULT] IOPS phase 8      mdtest_hard_delete      12.771 kiops : time 291.66 seconds
[SCORE] Bandwidth 7.11459 GB/s : IOPS 98.2653 kiops : TOTAL 26.4408
```

Summary

- Capable storage building blocks are needed for driving next generation applied industrial scientific applications

- CSIRO has invested in a 2PB NVMe solution which met performance and power criteria

- The POSIX compliant, BeeGFS parallel filesystem will be rolled out to users in Q1, 2019

CSIRO.

We imagine.
We collaborate.
We innovate.

Thank you