BoF: Architecture, Innovative Implementations and Development Plans

ISC, 2019

Introduction of Speakers

Frank Baetke
President, EOFS

Frank Herold
CEO, ThinkParQ

Dr. Peter Rösch
Chief Archiect &
Head of
Development,
ThinkParQ

Rene Tyhouse
Chief Technical
Architect, CSIRO

Agenda

- Overview of ThinkParQ: Frank Herold
- Overview of the Latest Version of BeeGFS: Dr. Peter Rösch
- Q&A
- Innovative Customer Implementation: CSIRO, Australia: Rene Tyhouse
- Q&A
- Overview of the BeeGFS Development Plans: Dr. Peter Rösch
- Q&A
- Survey
- Close and wrap-up: Frank Herold

The ThinkParQ behind BeeGFS

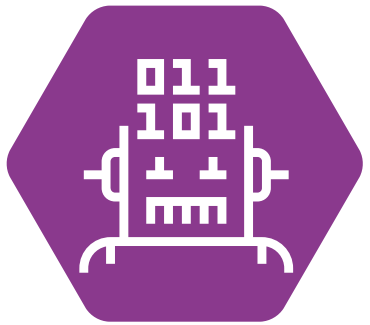
Frank Herold

About

- Established in 2014
  - Continuous growth since the beginning,
  - seeing lot of momentum in North America/APAC region on top of EMEA
- Focus on R&D
  - 70+% of the team,
  - added already another 30% in cy18/19)
- Independent
- X rankings in the top 20 on the IO-500 list.
- Awarded the HPCwire 2018 Best Storage Product or Technology Award

Delivering solutions for

HPC



AI / Deep Learning

Life Sciences

Oil and Gas

Standard and Enterprise Features

Standard Features:

- Distributed File system
- Per directory striping information
- Commandline or GUI based setup
- Statistics and Monitoring
- BeeOND

BeeGFS Enterprise Features (support contract required):

- High Availability
- Quota Enforcement
- Access Control Lists (ACLs)
- Storage Pools

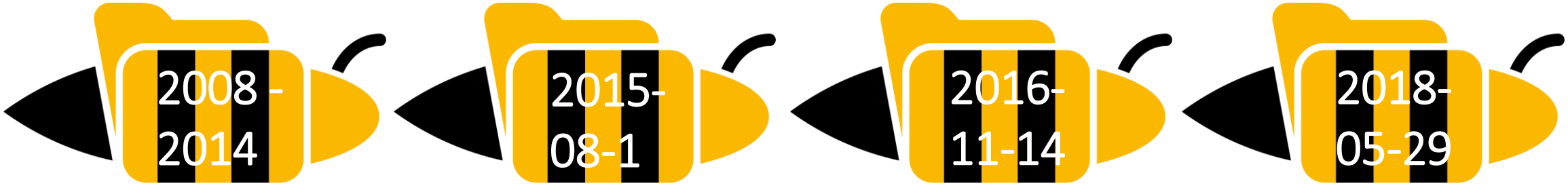
Overview of the Latest Version of BeeGFS

Dr. Peter Rösch

BeeGFS - Design Philosophy

- Designed for Performance, Scalability, Robustness and Ease of Use
- Distributed Metadata
- No Linux patches, on top of EXT, XFS, ZFS, BTRFS, ..
- Scalable multithreaded architecture
- Supports RDMA / RoCE & TCP (InfiniBand, Omni-Path, 100/40/10/1GbE, …)
- Easy to install and maintain (user space servers)
- Robust and flexible (all services can be placed independently)
- Hardware agnostic

Release History



FhGFS
2008 – 2014

BeeGFS 2015.03-r1 (latest: 2015.03-r27 of 2017-10-25)

BeeGFS 6.0 (latest: 6.19 of 2018-08-28)

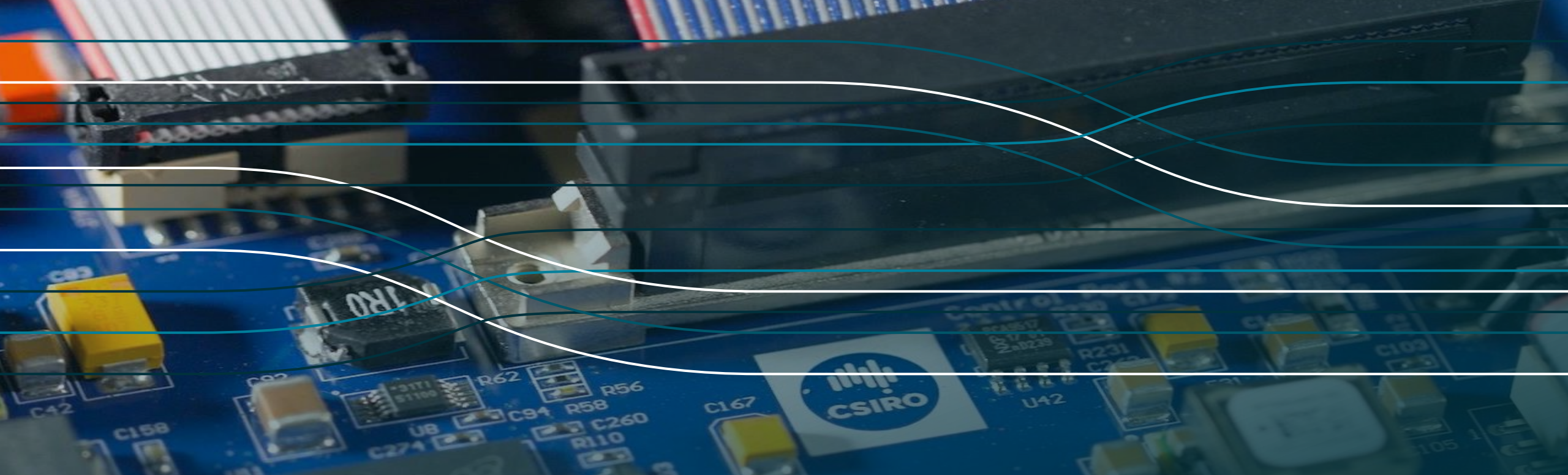
BeeGFS 7.0 (latest: 7.1.3 of 2019-05-10)

Actual release 7.1.3

- Added support for Kernel 4.19.x
  - Due to an issue in Kernel 4.19, this will only work from Kernel version 4.19.1 and newer.
- Fixes
  - Fixed possible deadlock situation in internal lock management layer that could have led to management daemon stalling.
  - Fixed a resource issue with IB connections by enforcing release of IB queue pair.
  - Fixed an issue which prevented the order of interfaces in the connInterfacesFile to be applied correctly.
  - Fixed include problem with Mellanox OFED 4.5

Innovative Customer Implementation / Casestudy

Rene Tyhouse, CSIRO, Australia



Next Generation Scratch Filesystem

ISC 2019

Rene Tyhouse

June 2019

INFORMATION MANAGEMENT AND TECHNOLOGY (IMT)
www.csiro.au

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

Outline

About CSIRO

Scientific Computing overview

CSIRO's All Flash Scratch Filesystem

Filesystem Benchmarking

CSIRO – Australia's National Science Agency

5800
talented staff

$1billion
+ budget

Working with over **2800+** industry partners

55
sites across Australia

Top 1%
of global research agencies

Each year **6 CSIRO technologies** contribute $5 billion to the economy

We solve the greatest challenges through innovative science and technology

WiFi
WLAN

POLYMER BANKNOTES

TOTAL WELLBEING DIET

AEROGARD

HENDRA VACCINE

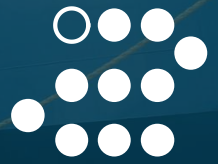
BARLEYmax™

EXTENDED WEAR CONTACTS

RELENZA FLU TREATMENT

SOFTLY WASHING LIQUID

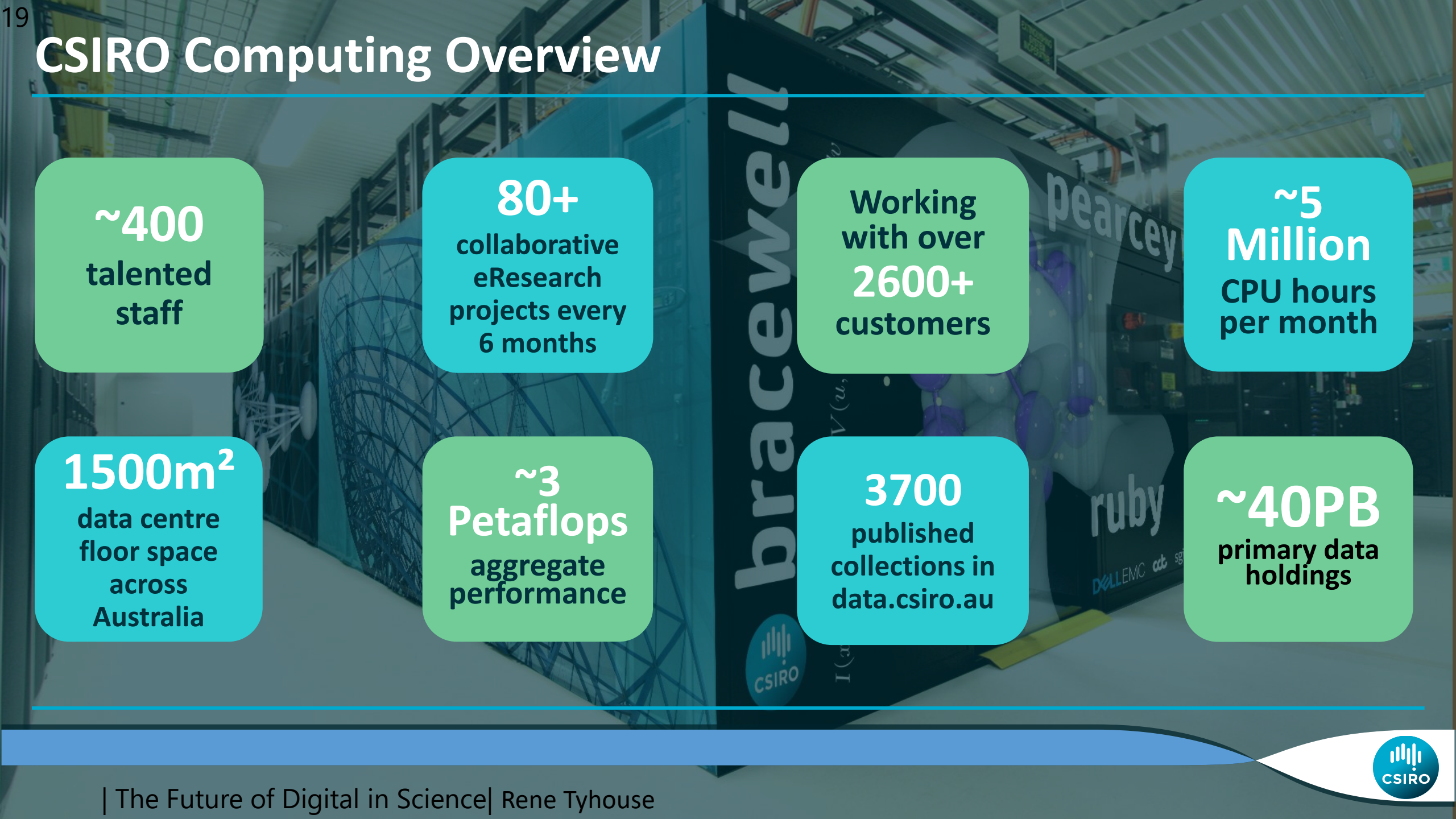
SELF TWISTING YARN



RAFT POLYMERISATION

NOVACQ™ PRAWN FEED

CSIRO Computing Overview



~400 talented staff

80+ collaborative eResearch projects every 6 months

Working with over **2600+** customers

~5 Million CPU hours per month

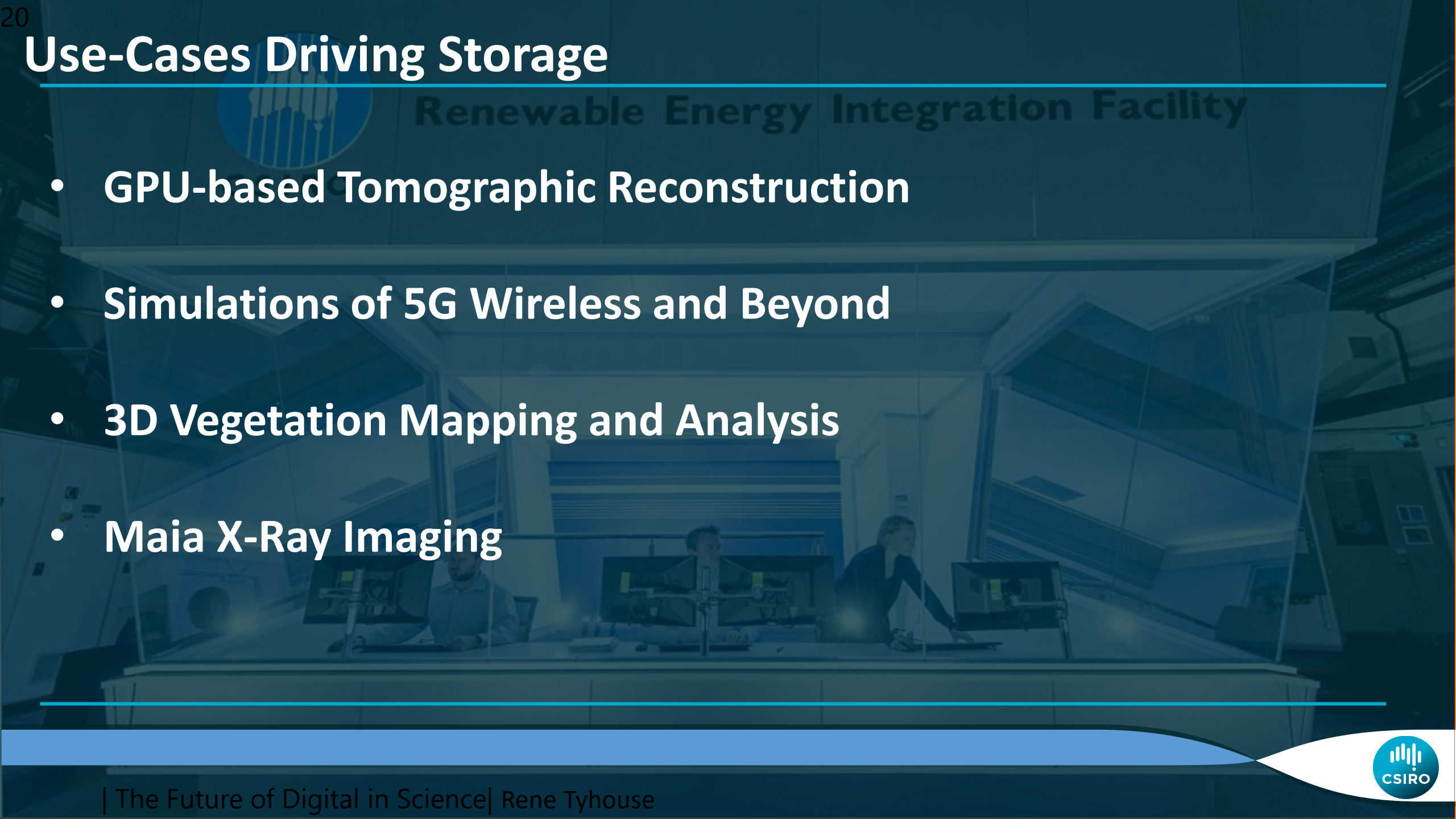
1500m² data centre floor space across Australia

~3 Petaflops aggregate performance

3700 published collections in data.csiro.au

~40PB primary data holdings

Use-Cases Driving Storage

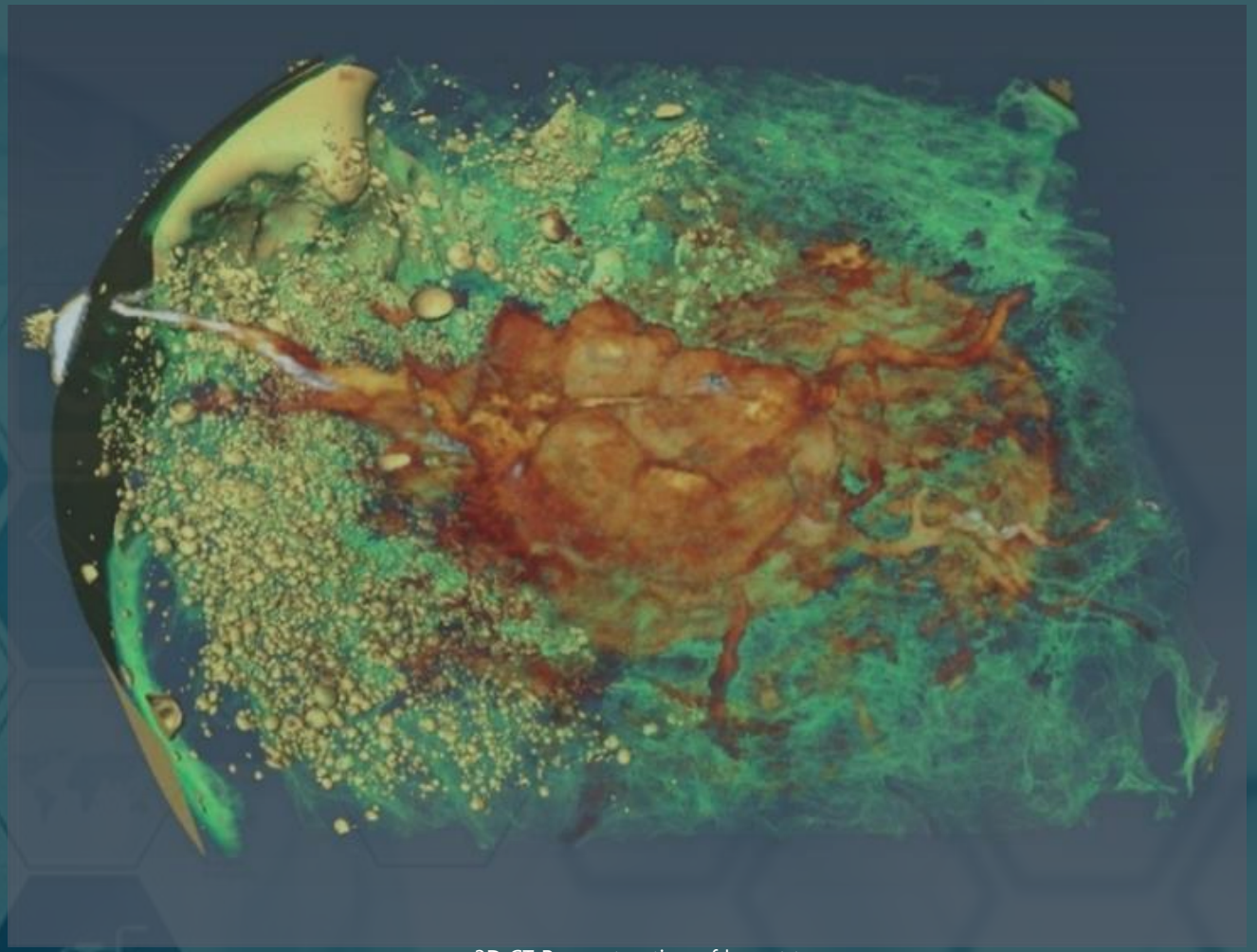


- GPU-based Tomographic Reconstruction
- Simulations of 5G Wireless and Beyond
- 3D Vegetation Mapping and Analysis
- Maia X-Ray Imaging

GPU-based Tomographic Reconstruction

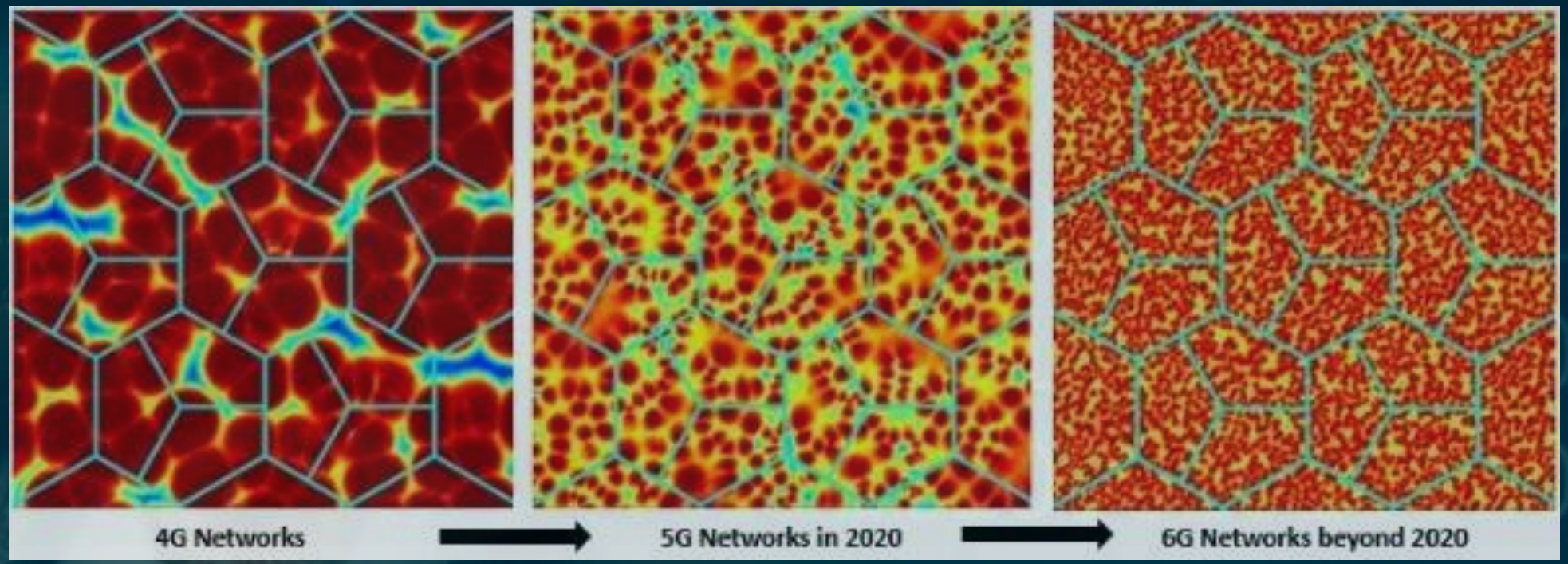
3D CT Reconstruction of an excised human breast containing a tumour (in red).

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



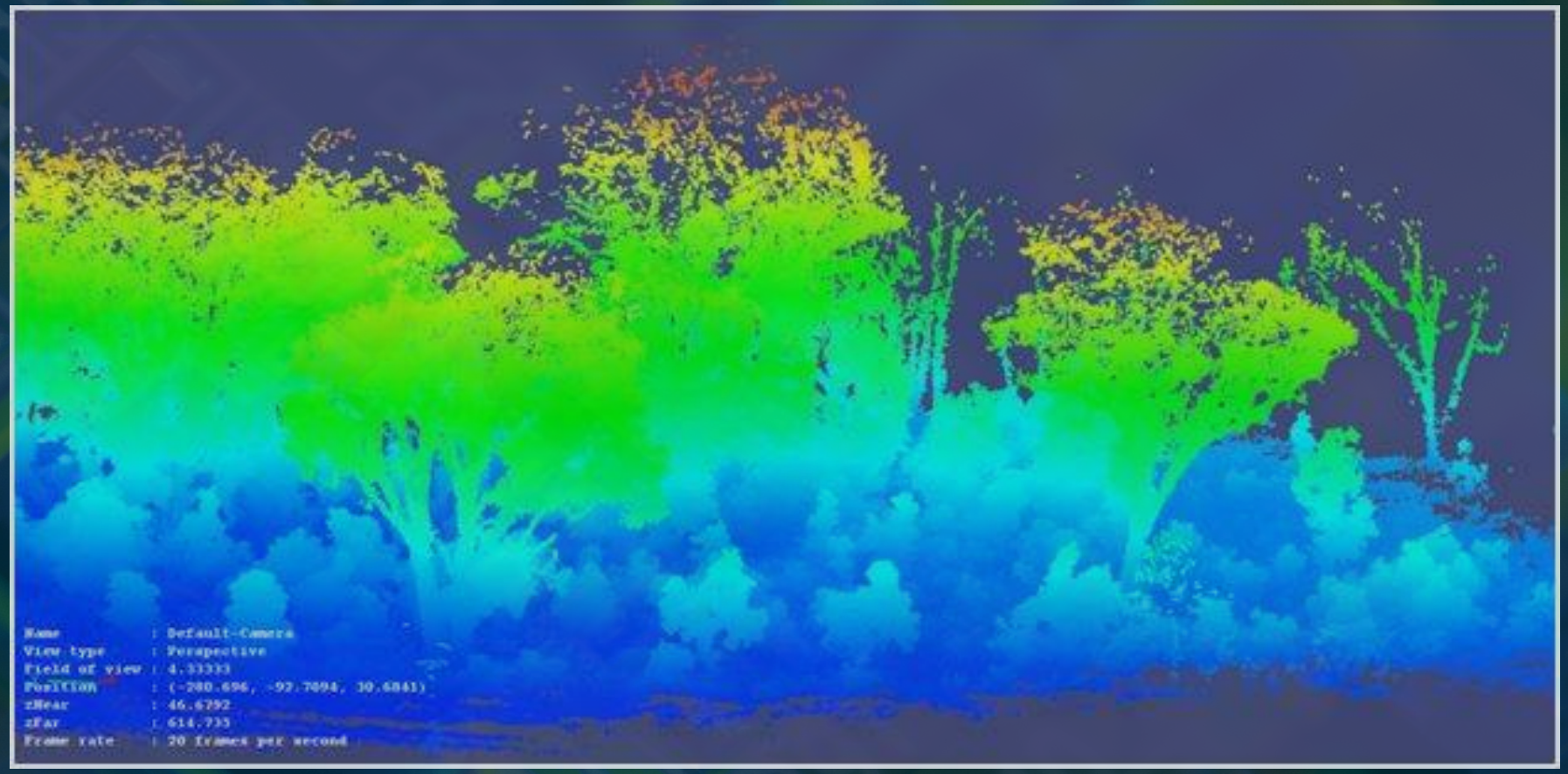
3D CT Reconstruction of breast tumour
Imaging and Medical Beamline, Australian Synchrotron

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



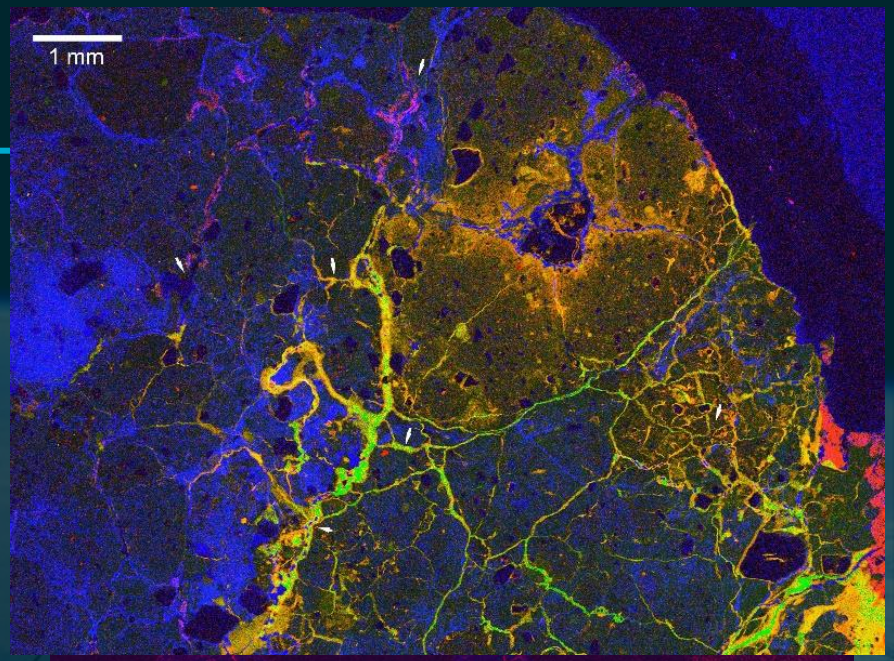
Generating vegetation cover maps in 3D from data acquired via a Zebedee handheld laser scanner

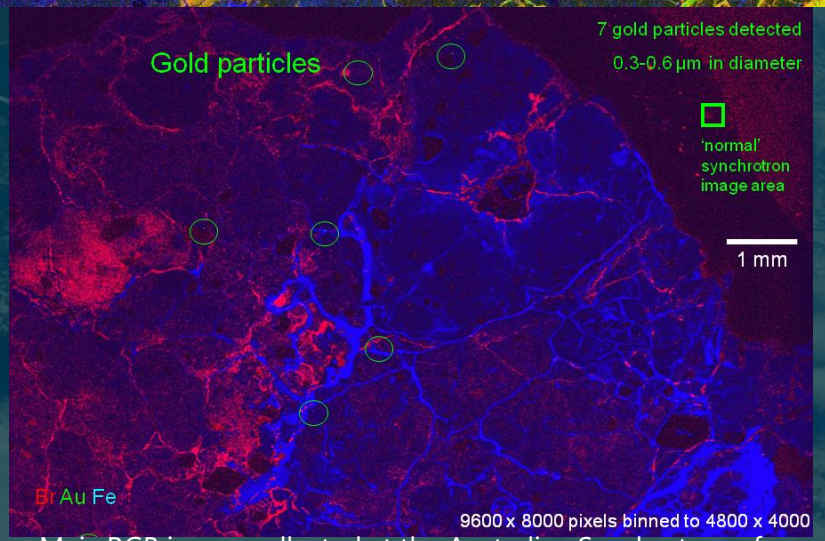
Maia X-Ray Imaging

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



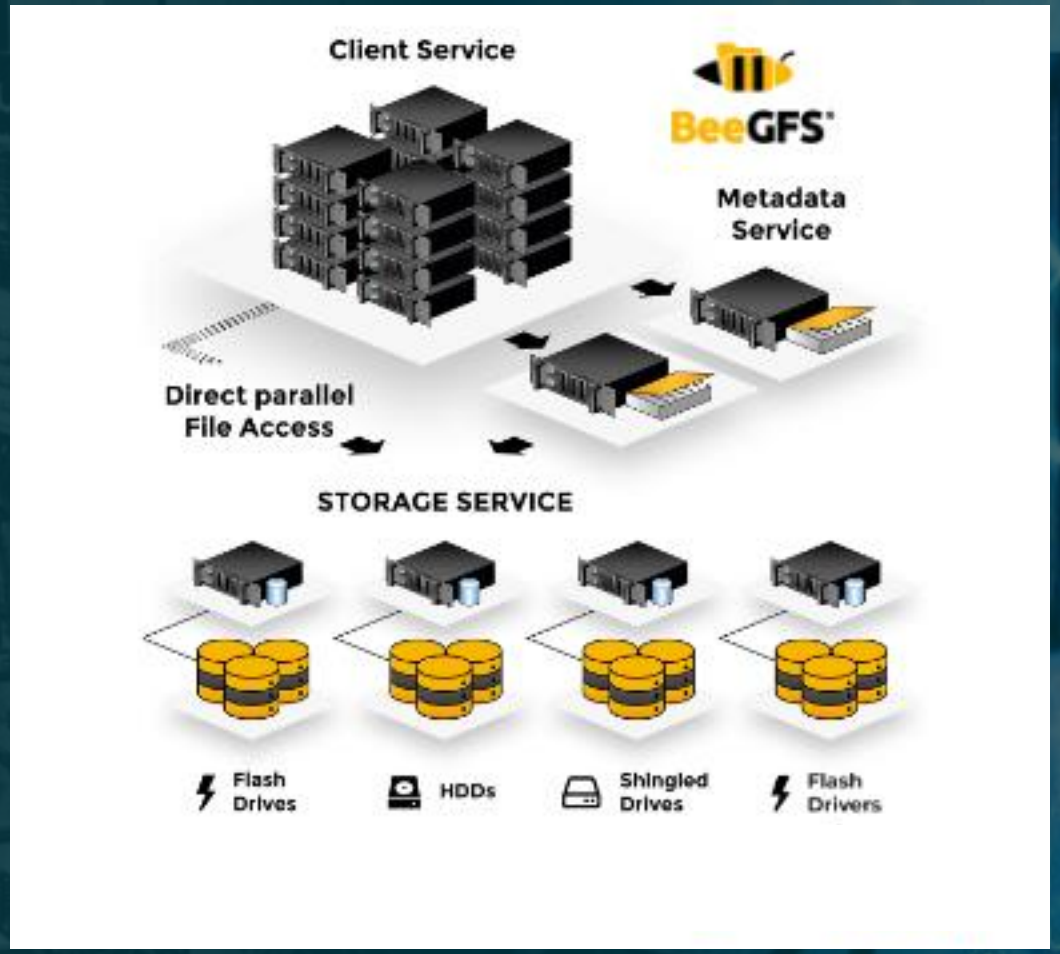


Maia RGB image collected at the Australian Synchrotron of a clay sample from the Mt Gibson gold deposit in Western Australia (green = iron, blue = bromine, red = arsenic).

Storage Drivers

- 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

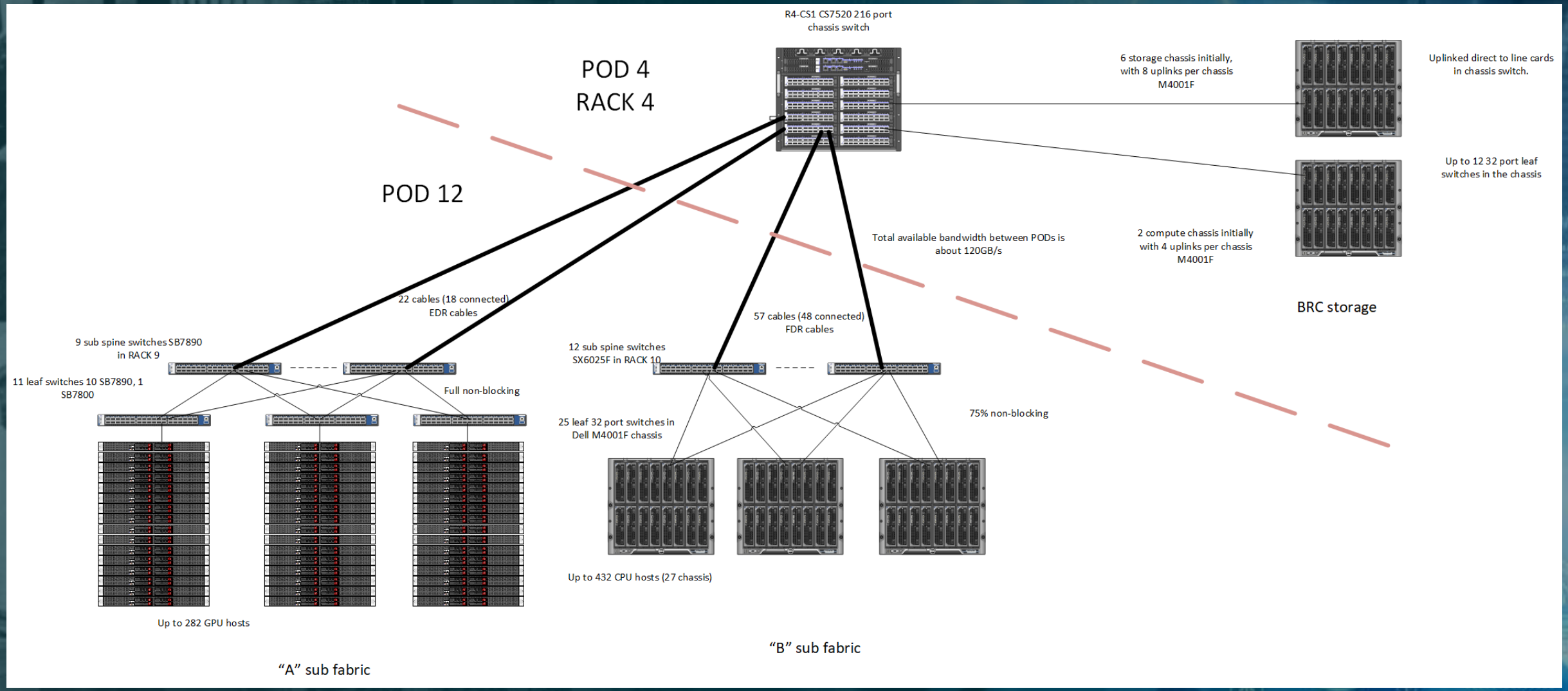
Storage Drivers



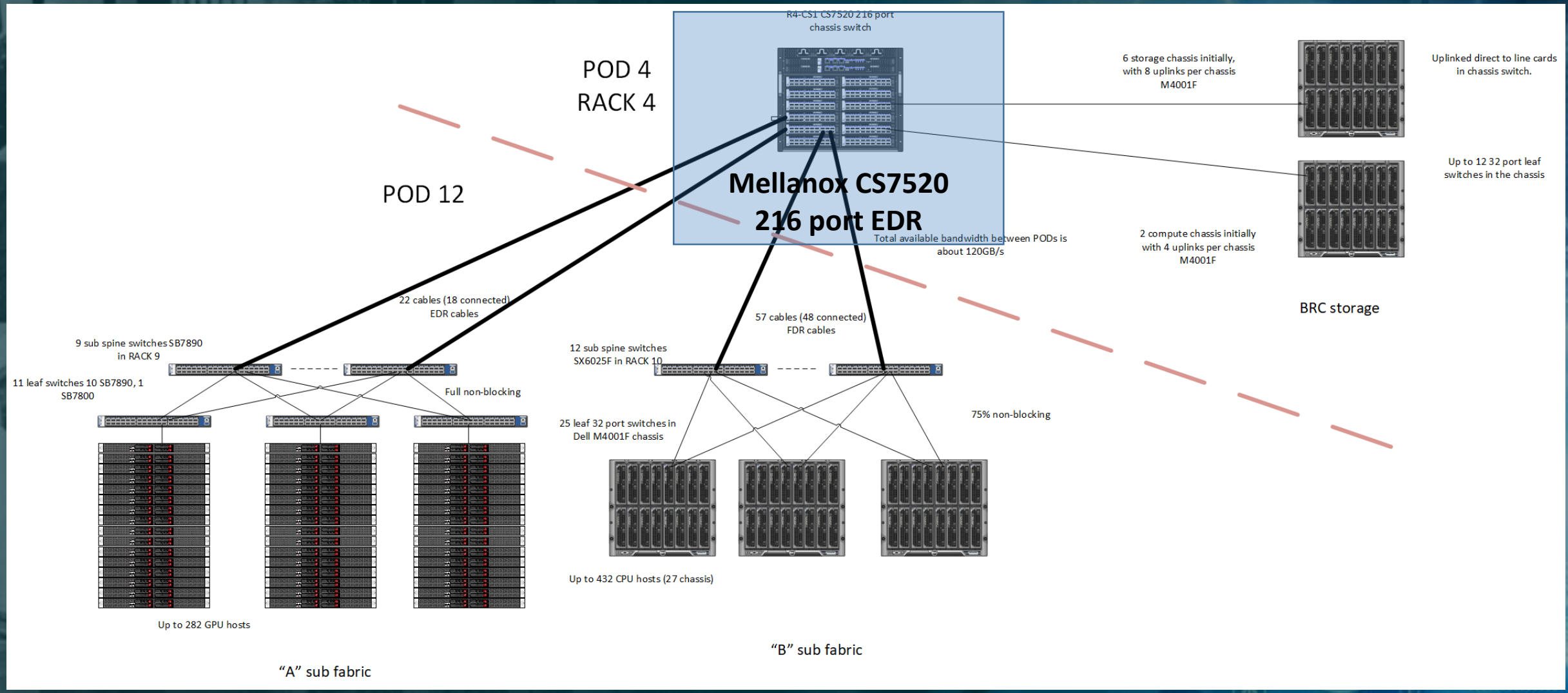
Hardware Building Blocks

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

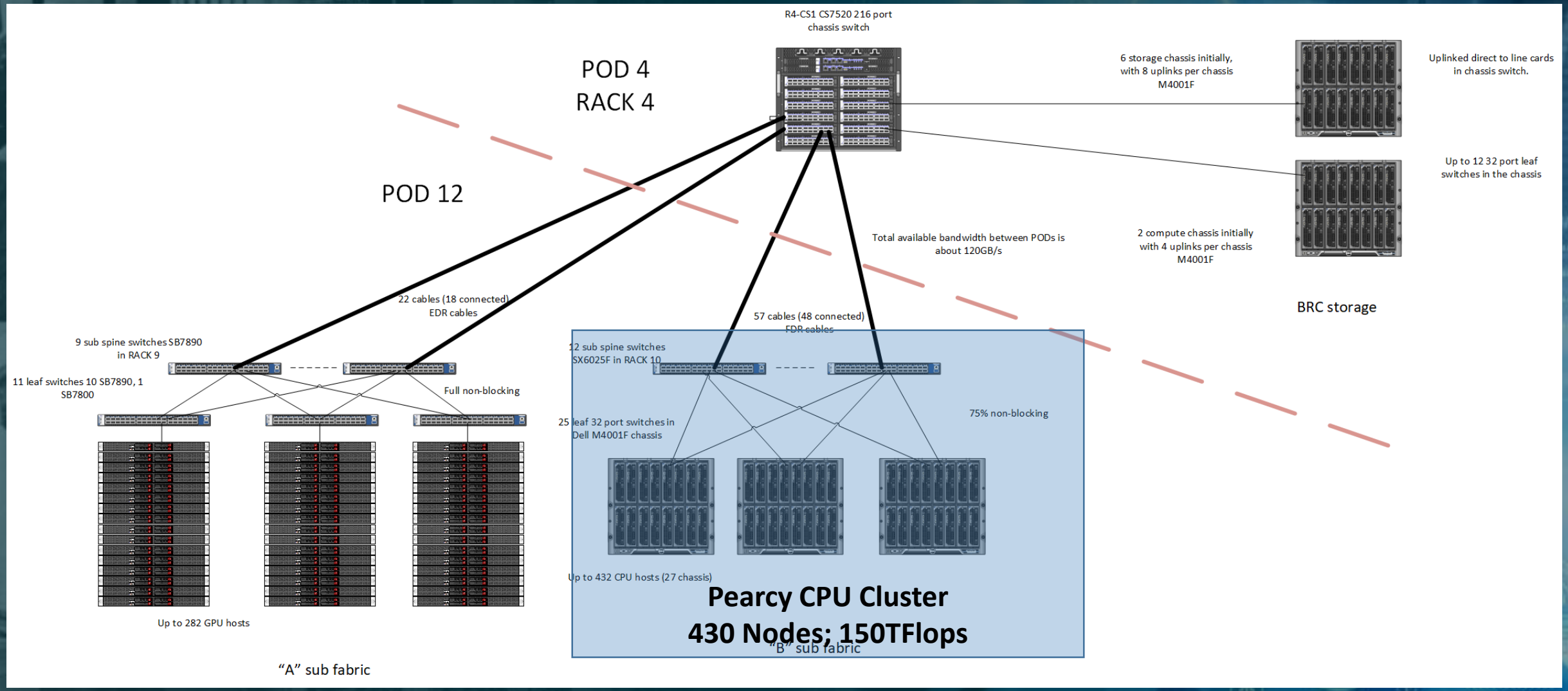
Switch Centric View of Compute and Storage Clusters



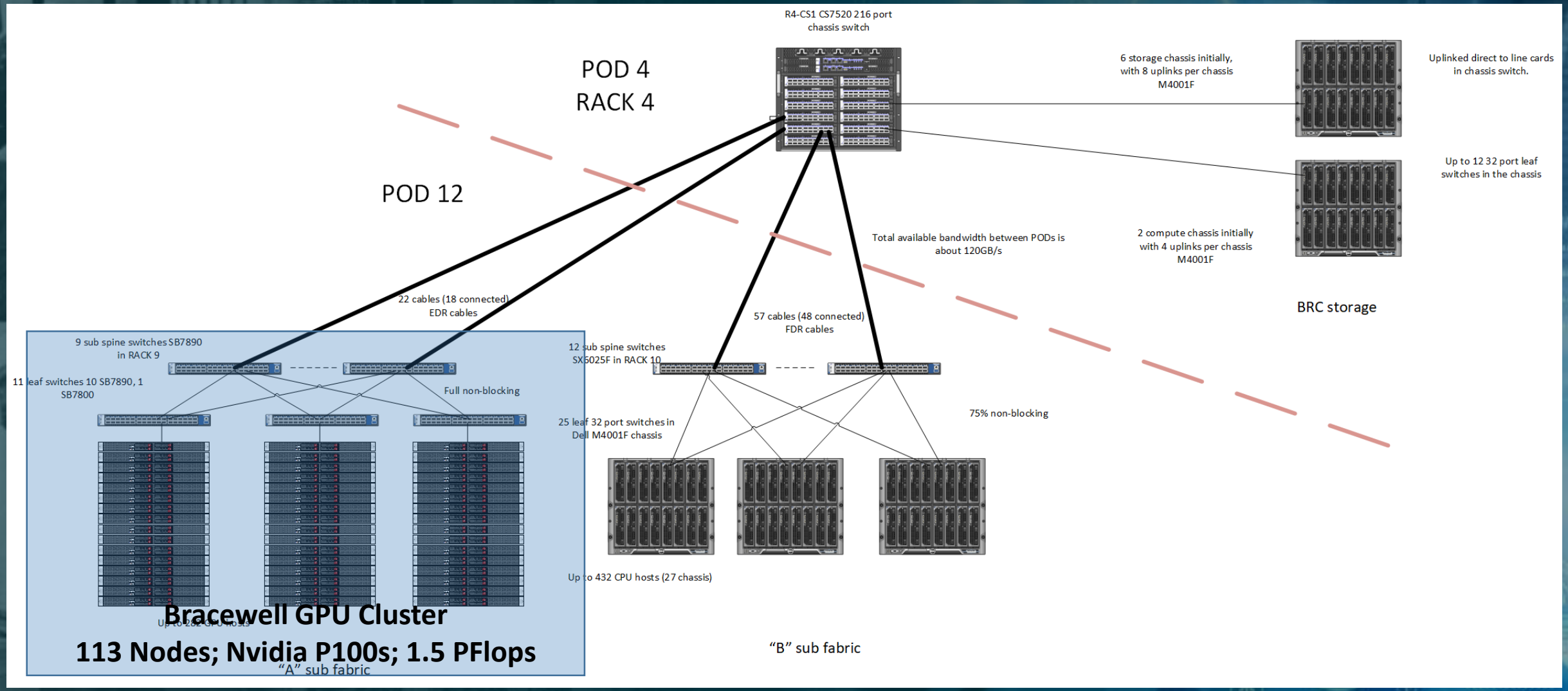
Switch Centric View of Compute and Storage Clusters



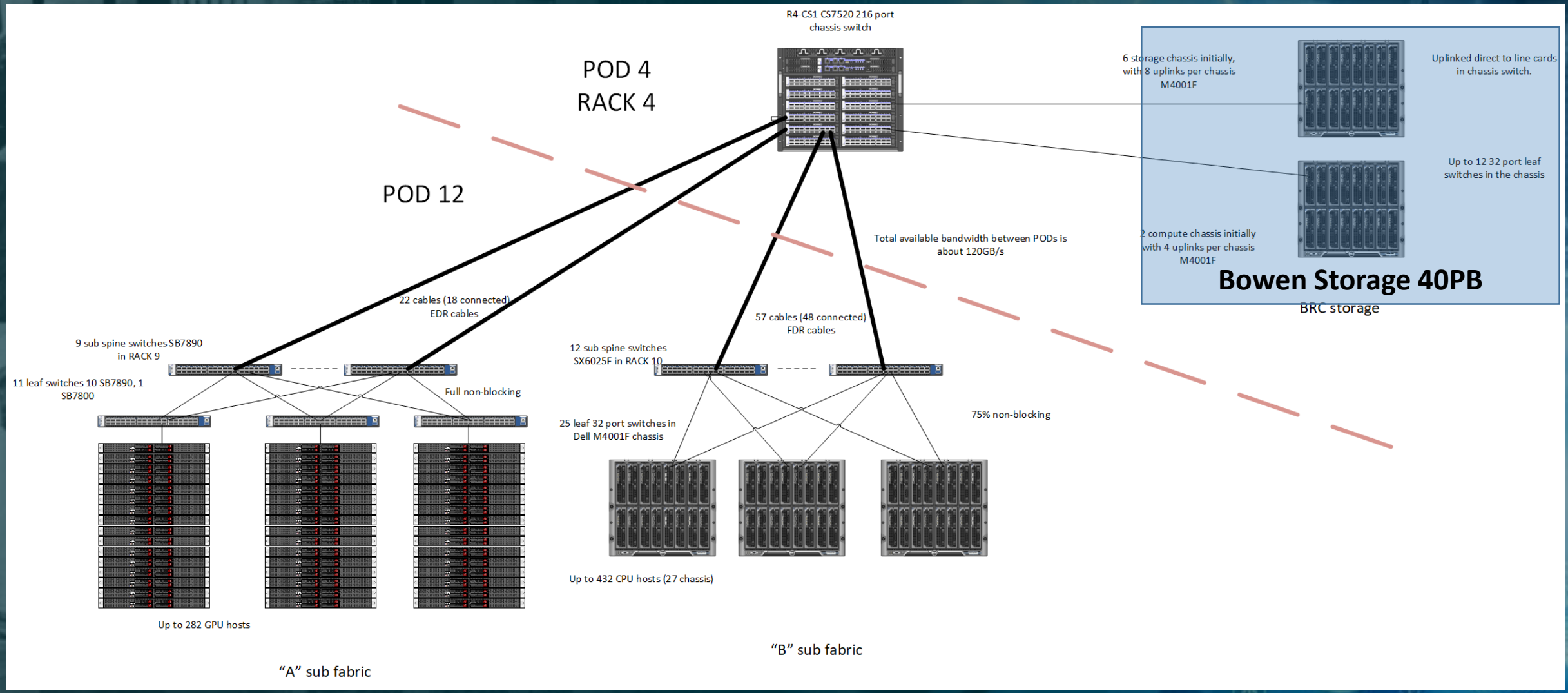
Switch Centric View of Compute and Storage Clusters



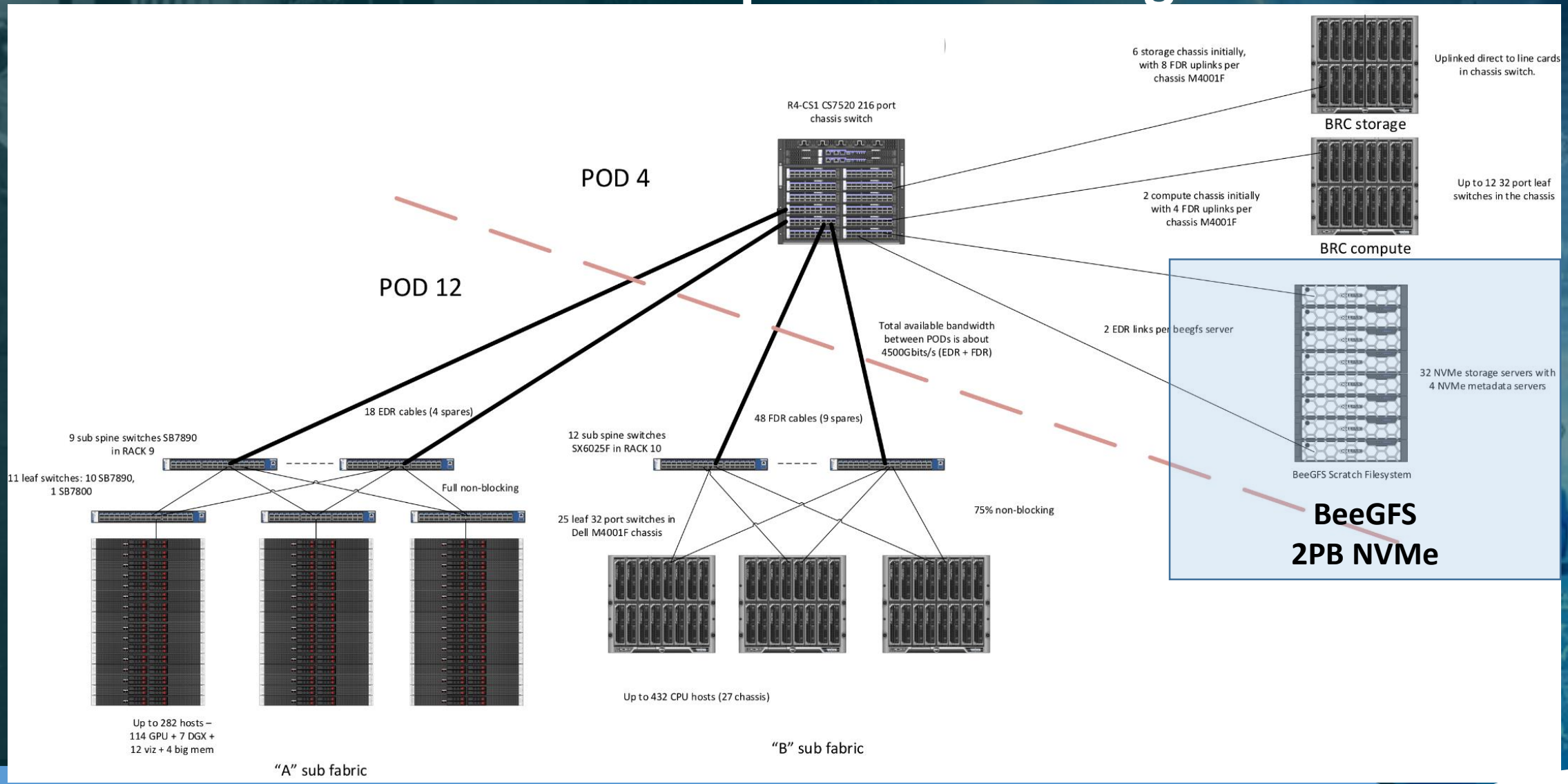
Switch Centric View of Compute and Storage Clusters



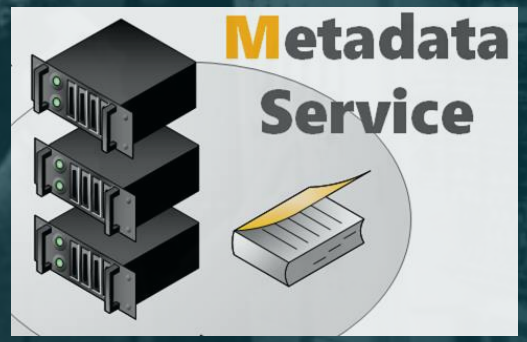
Switch Centric View of Compute and Storage Clusters



Switch Centric View of Compute and Storage Clusters



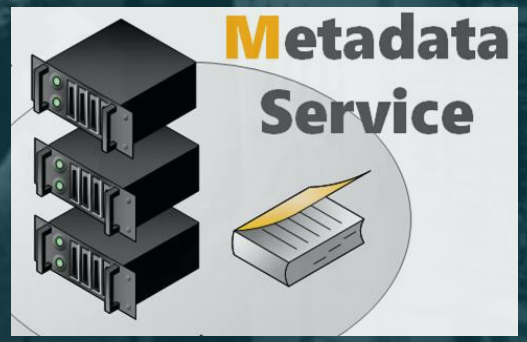
Metadata Service Building Blocks



4 Metadata servers

- DellEMC R740
- Dual Intel 6154
  - *3.0GHz 12 core, 384GB*
- Dual *ConnectX-5 EDR*

Metadata Service Building Blocks

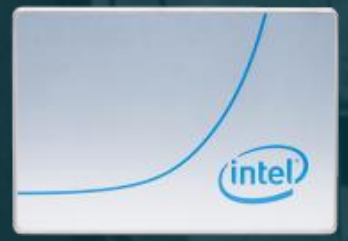


4 Metadata servers

- DellEMC R740
- 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



Metadata Service Building Blocks

32 Storage servers

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

Metadata Service Building Blocks

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

IO500 Benchmark

10 node challenge:

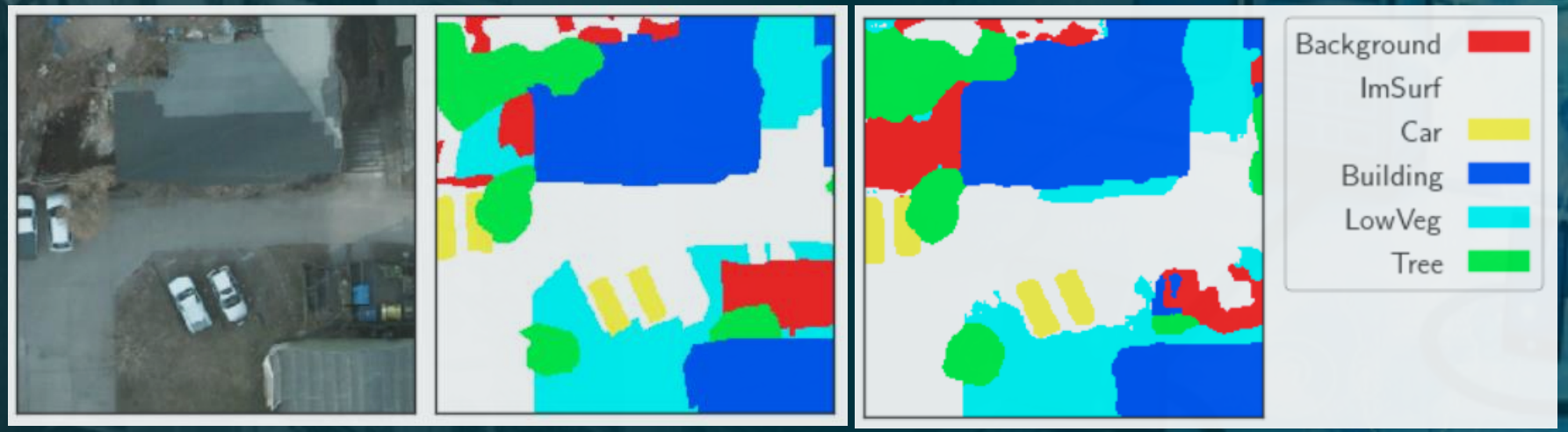
[Summary] Results files in /scratch1/leh015/io-500-dev/results/2019.05.21-23.18.14
[Summary] Data files in /scratch1/leh015/io-500-dev/datafiles/io500.2019.05.21-23.18.14

[RESULT] BW	phase 1	ior_easy_write	113.373 GB/s : time 310.48 seconds
[RESULT] BW	phase 2	ior_hard_write	2.453 GB/s : time 314.14 seconds
[RESULT] BW	phase 3	ior_easy_read	93.800 GB/s : time 375.27 seconds
[RESULT] BW	phase 4	ior_hard_read	56.562 GB/s : time 13.62 seconds
[RESULT] IOPS	phase 1	mdtest_easy_write	151.681 kiops : time 319.39 seconds
[RESULT] IOPS	phase 2	mdtest_hard_write	9.850 kiops : time 327.67 seconds
[RESULT] IOPS	phase 3	find	1502.420 kiops : time 34.09 seconds
[RESULT] IOPS	phase 4	mdtest_easy_stat	873.490 kiops : time 58.00 seconds
[RESULT] IOPS	phase 5	mdtest_hard_stat	136.769 kiops : time 26.38 seconds
[RESULT] IOPS	phase 6	mdtest_easy_delete	247.030 kiops : time 204.46 seconds
[RESULT] IOPS	phase 7	mdtest_hard_read	32.739 kiops : time 100.87 seconds
[RESULT] IOPS	phase 8	mdtest_hard_delete	14.666 kiops : time 224.25 seconds

[SCORE] Bandwidth [redacted] GB/s : IOPS [redacted] kiops : TOTAL [redacted]

User Story

- Deep learning models
- ~60 million to 260+M parameters
- large memory footprint
- ~1+TB for training of SAR data
- Using Bracewell Cluster

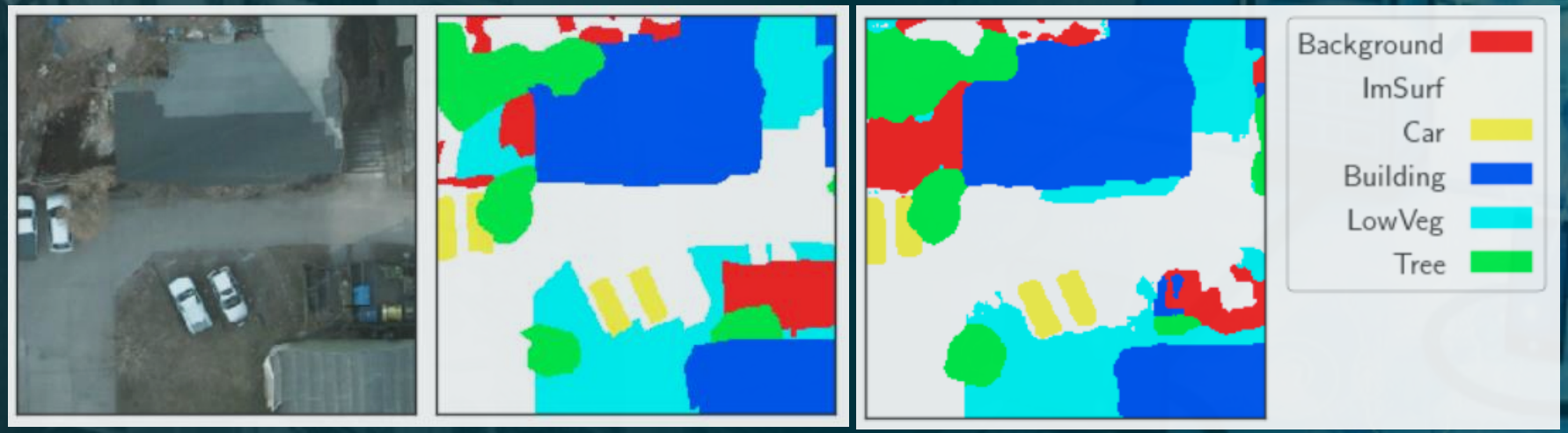


Foivos Diakogiannis, CSIRO Data Scientist

User Story

- Deep learning models
- ~60 million to 260+M parameters
- large memory footprint
- ~1+TB for training of SAR data
- Using Bracewell Cluster

- Performance boost is of the order x3 on a single node for training
- 3 weeks down to 1 week
- IO bounds parts of the job changed from 48 hours down to 3.5 hours



Foivos Diakogiannis, CSIRO Data Scientist

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

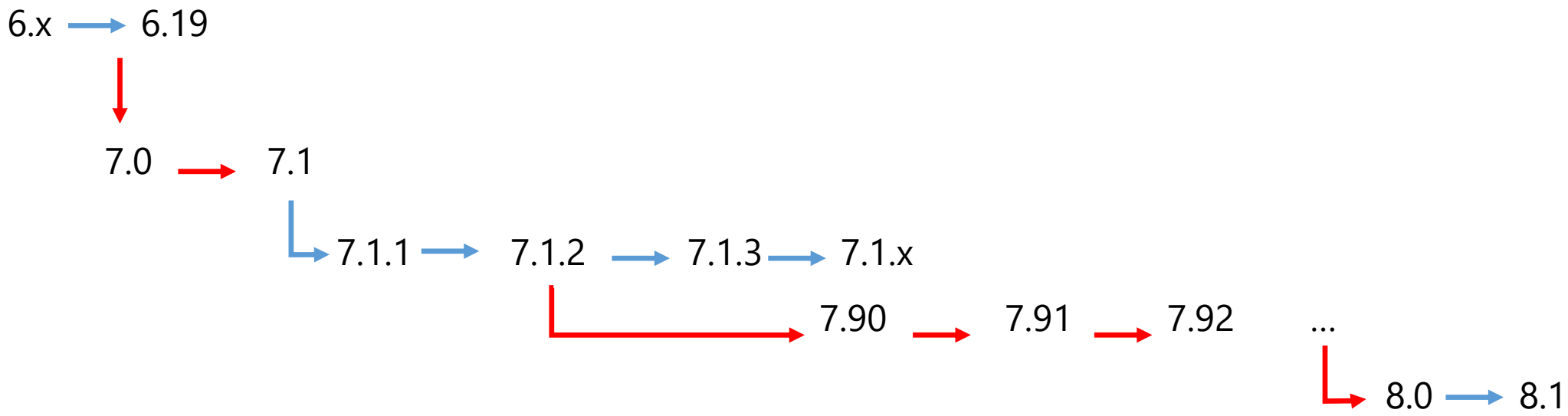
Thankyou

Overview of the BeeGFS Development Plans

Dr. Peter Rösch

BeeGFS Version Tree

- Currently, different BeeGFS 7.x versions include major changes, such as storage layout
- To cope with that in a better way, we decided to branch of a 7.90 version:



- The 7.9x versions lead us to 8.0 and then doesn't require any more storage layout changes
- Our long-term goal is to support semantic versioning

Roadmap Directions

- Refactorization and stabilization (7.9x; 8.x)
  - Code modularization (7.91)
  - UDP vs TCP
  - Standalone GUI Installer, based on ansible under the hood
  - New Implementation of internal wrapper for InfiniBand library (7.90)
  - Adaptions of meta and storage layout (8.0)
- Revised command-line interface ‘beegfs’ (7.91)
  - Based on a schema driven approach
  - Provides more consistency
  - Complete help pages
  - Basis for future administrative API and GUI interface
- fstab based mount for BeeGFS clients (7.91)
- Syslog support (7.91)

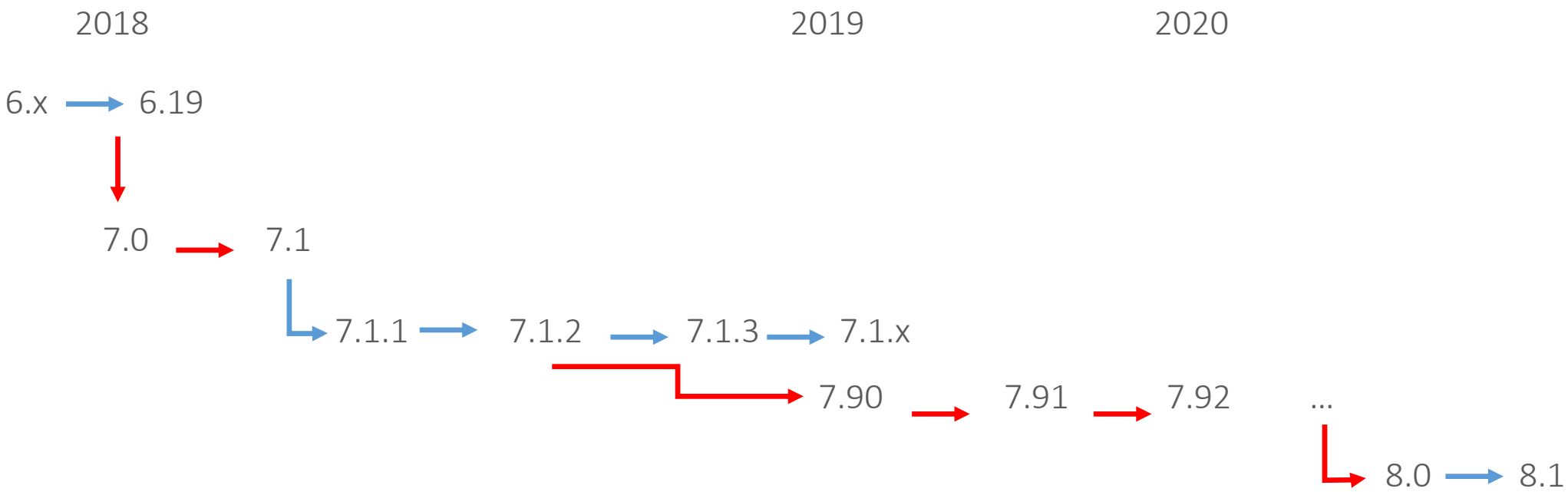
Roadmap Directions

- DKMS (Dynamic Kernel Module Support) BeeGFS client packages (7.92)
  - Enable prebuilt binary packages, and
  - Enable module build at user site (like now) at the same time

- Agent based monitoring (8.0)
  - Move away from BeeGFS specialised monitoring UI
  - Configurable data providers, enabling implementation of open monitoring protocols and usage with different monitoring front ends

- Centralized Configuration (8.1)
  - Allows better support of complex sites

Timeline



Survey

Survey

BeeGFS®

Please complete and drop off at the BeeGFS booth J-640

1. What are your favorite features of BeeGFS?

   _____

   _____

2. What is the most-awaited feature for you?

   _____

   _____

3. Is there any other operating system you would like to see supported?

   _____

   _____

4. What is the timing offset you would expect from us to support new kernel versions?

   - ☐ Immediate
   - ☐ Up to 3 months
   - ☐ Up to 6 months

5. Please rank the following in order of importance from 1 to 5, where 1 is most important to you and 5 least important to you.

   - ☐ Stability
   - ☐ Performance
   - ☐ Robustness
   - ☐ Flexibility
   - ☐ Ease of Use

6 . Any other comments? _____

_____

BeeGFS®

Close & Wrap-Up

Frank Herold

ISC Schedule

Monday 17th

- 5:00 PM
  - BeeGFS BoF (room Kontrast)
- 6:30 PM
  - Dell Solutions Overview (booth #J-640)

Tuesday 18th

- 10:30 AM
  - Excelero Solutions Overview (booth #J-640)
- 11:00 AM
  - BeeGFS & Inspur partner presentation (booth #F-940)
- 1:30 PM
  - BeeGFS Overview (booth #J-640)
- 3:30 PM
  - NetApp Solutions Overview (booth #J-640)

Wednesday 19th

- 1:30 PM
  - BeeGFS Overview (booth #J-640)
- 3:20 PM
  - BeeGFS & Bright Computing partner presentation (booth #J-632)

BeeGFS®

Thank You

Follow BeeGFS: