Scalable Distributed Infrastructure for Data Intensive Science

David Abramson

Director, Research Computing Centre
Professor of Computer Science
University of Queensland
david.abramson@uq.edu.au

Data Intensive Science

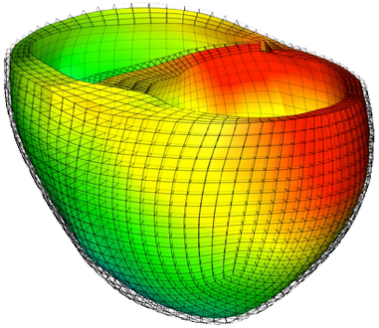
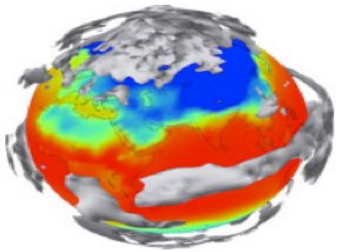
The University of Queensland Australia | CREATE CHANGE

Data-Intensive Computing

- Very large data-sets or very large input-output requirements
- Two data-intensive application classes are important and growing

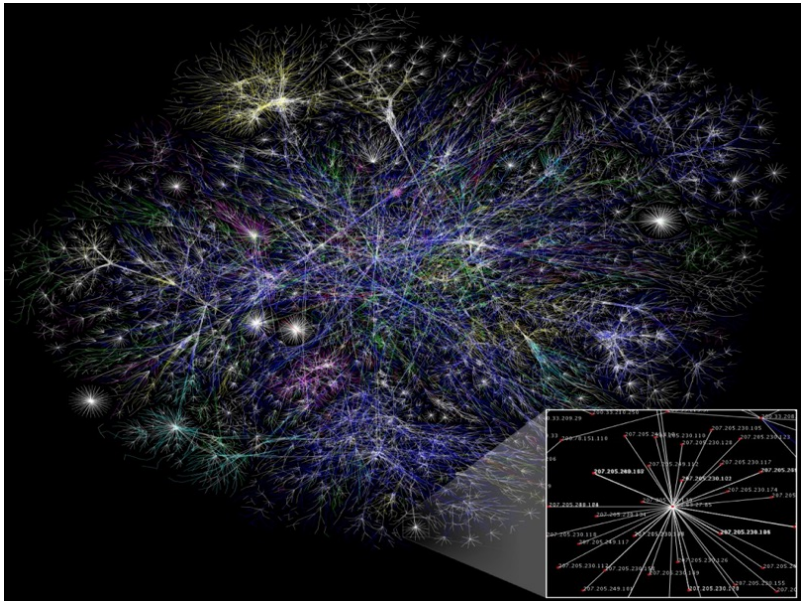


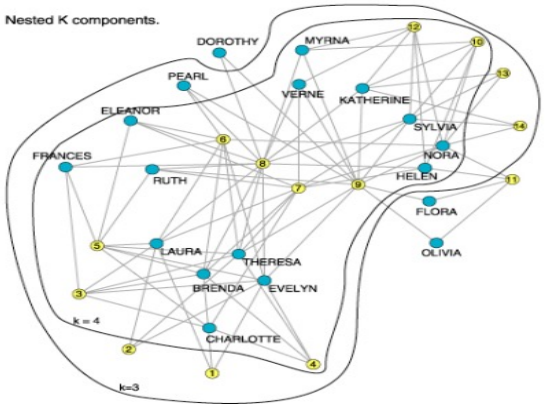

Data Mining &
Data Analytics

Data-Intensive Computing

- Examples Applications:
- Genome sequence assembly
- Climate simulation analysis
- Social network analysis
- Imaging





Infrastructure for Data Intensive Computing

- Computation
  - Large amounts of main memory
  - Parallel processors
- Storage
  - Short and long term storage
  - Many views of same data
    - Parallel File System
    - Local access (POSIX)
    - Remote collaboration and sharing (Object store)
    - Sync-and-share
    - Web
    - Cloud
- Can we unify these into a single model?

amazon web services™

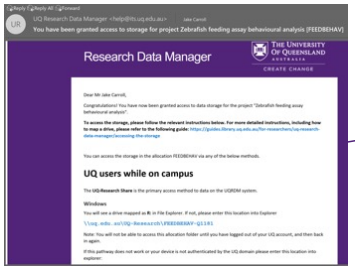
Then …

And now …

Researcher obtains a collection.

Research Data Manager

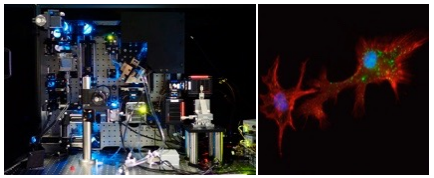
Designed at The University of Queensland

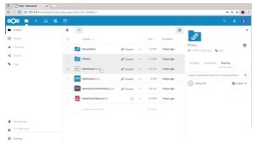


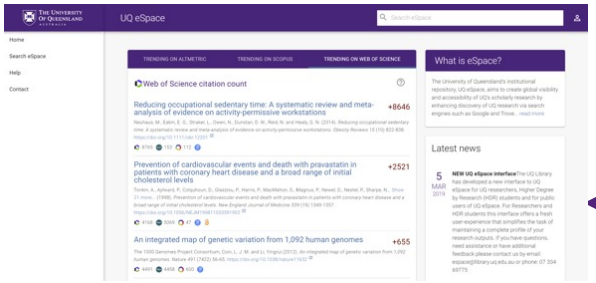
Researcher is sent an email explaining access instructions.

Collection then mounted on the fabric.

Researcher then acquires data and stores



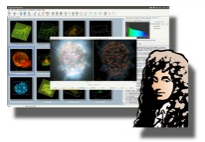




DOIs can be minted, published to UQ eSpace.
RDM can facilitate data linking back for durable URL.

PYTORCH

nectar



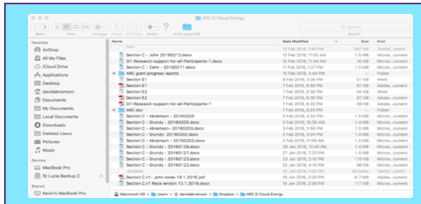
It is already on the fabric. Can be accessed via CVL, supercomputers, other.

XNAT

myTardis

OMERO

Managed Data through Repositories



MeDiCI

ownCloud

Synchronous

Asynchronous

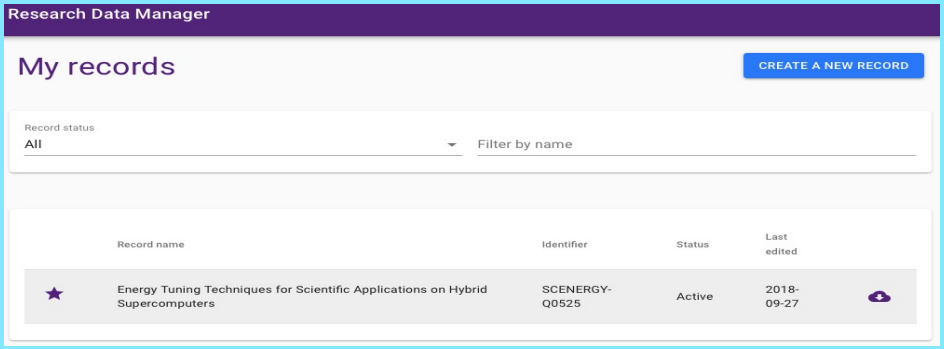
Unmanaged Data

HPC Access

S3, Swift

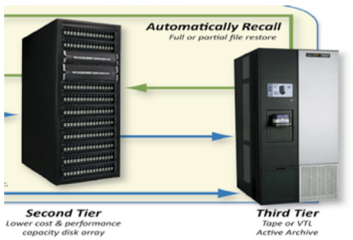
Cloud Access

MeDiCI



RDM

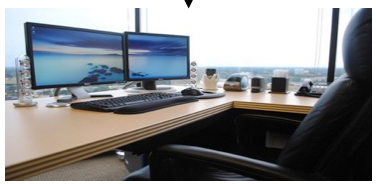
aws



Use cases

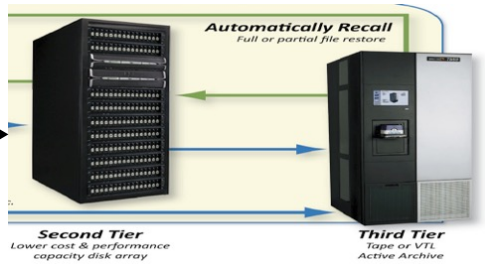
Use Case: Microscopy

Capture

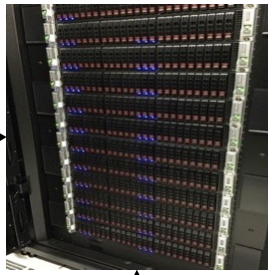


Pre-process

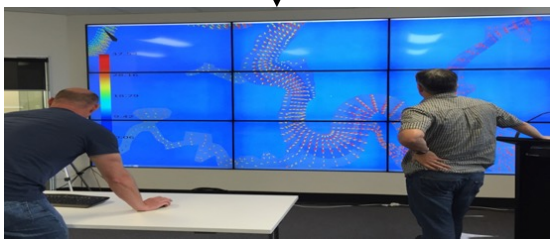
Store



Analyze

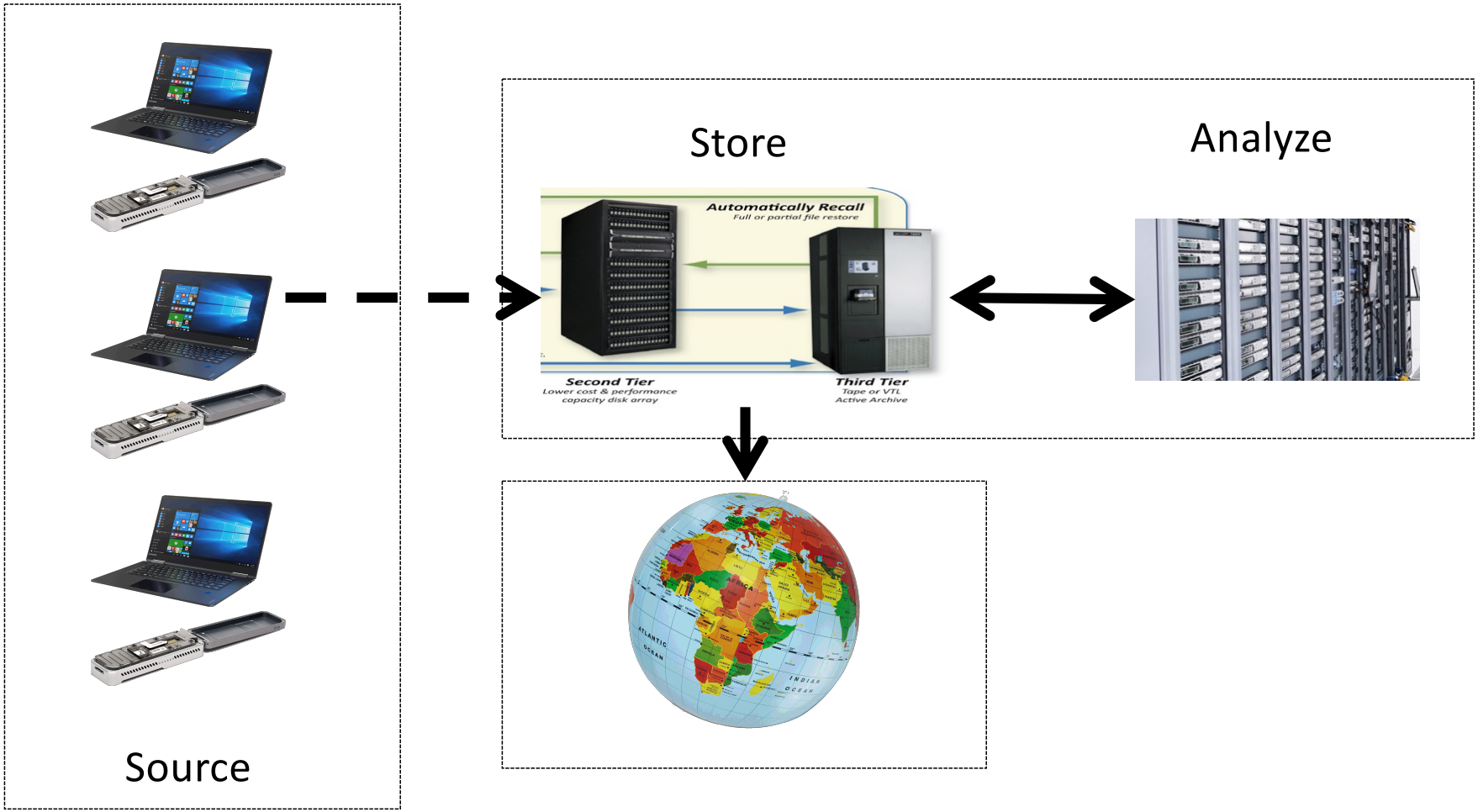


Publish & Share

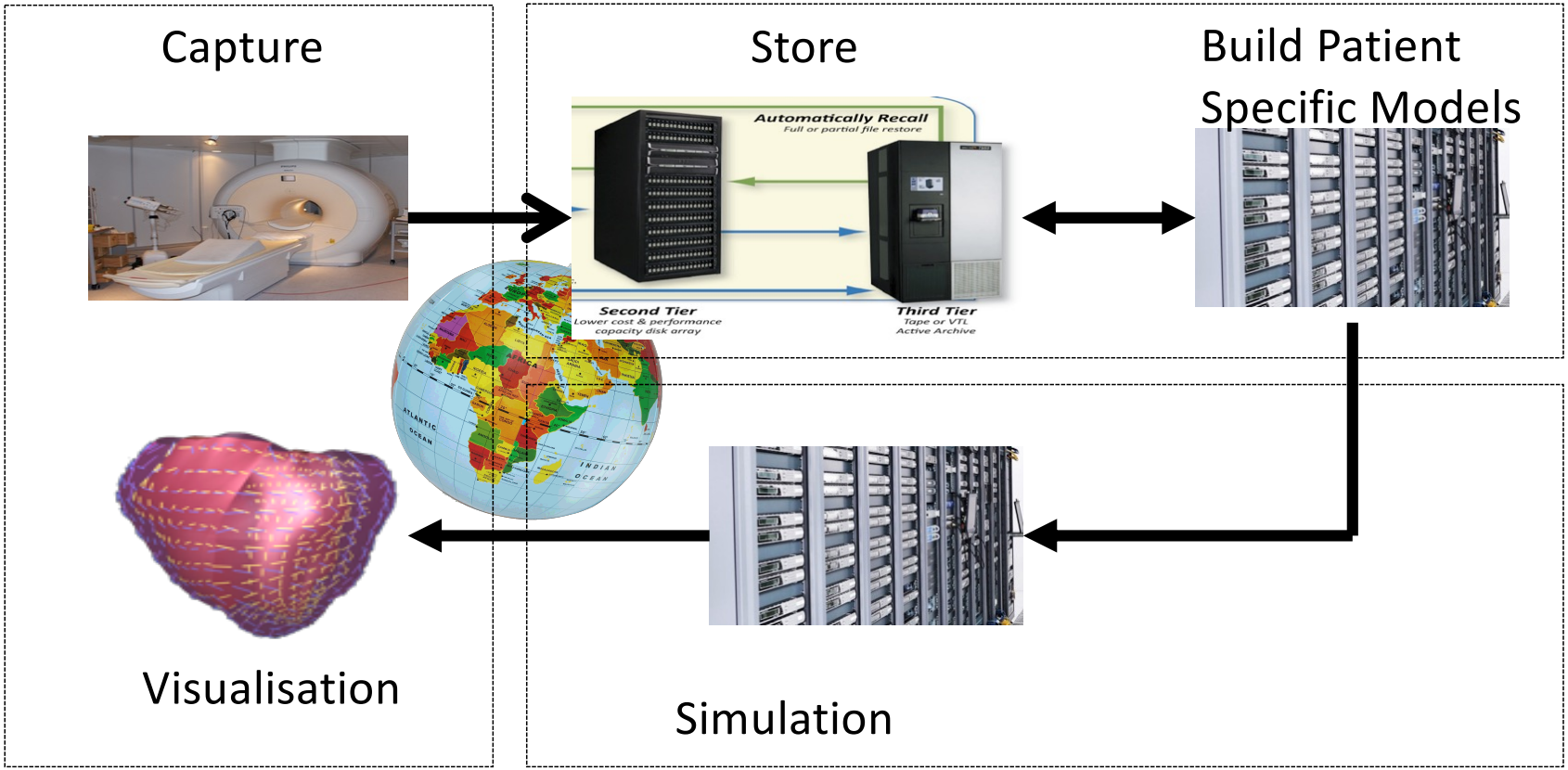


Interpret

Use Case: Personal Genomics

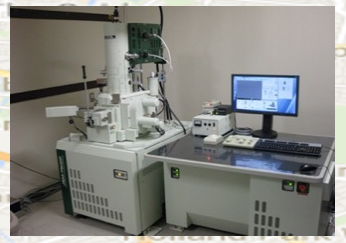


Use Case: Cardiac Science

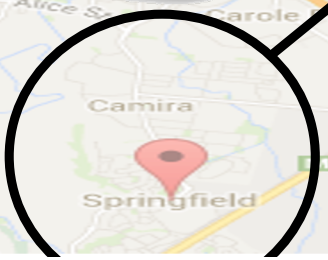


MeDiCI

UQ Landscape: An inconvenient truth

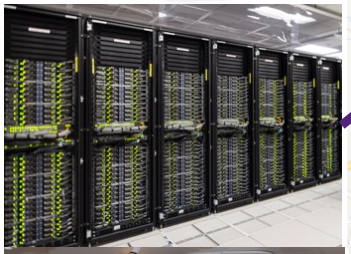


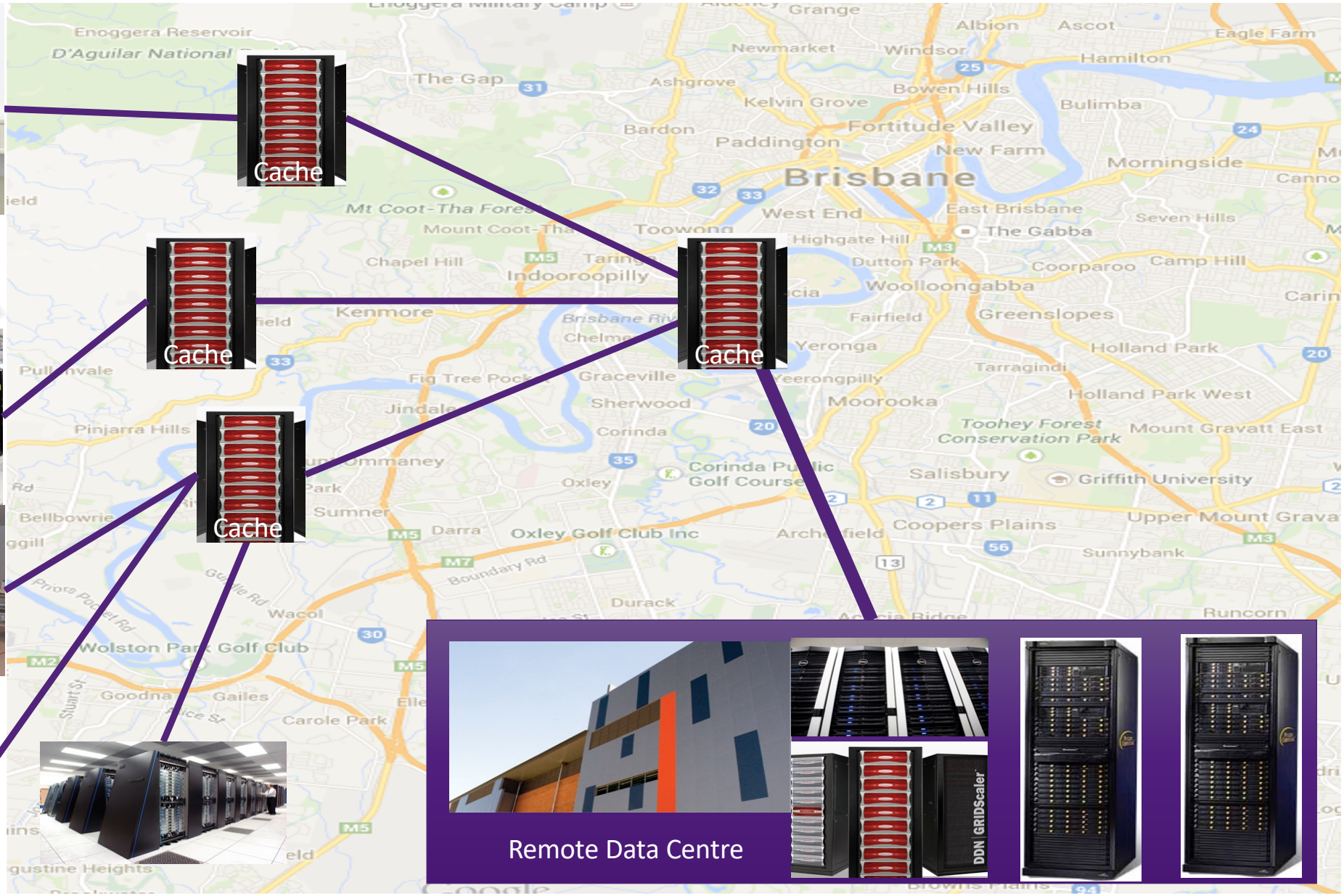
30 kms

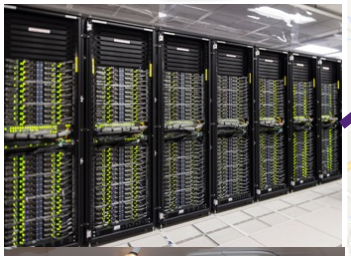


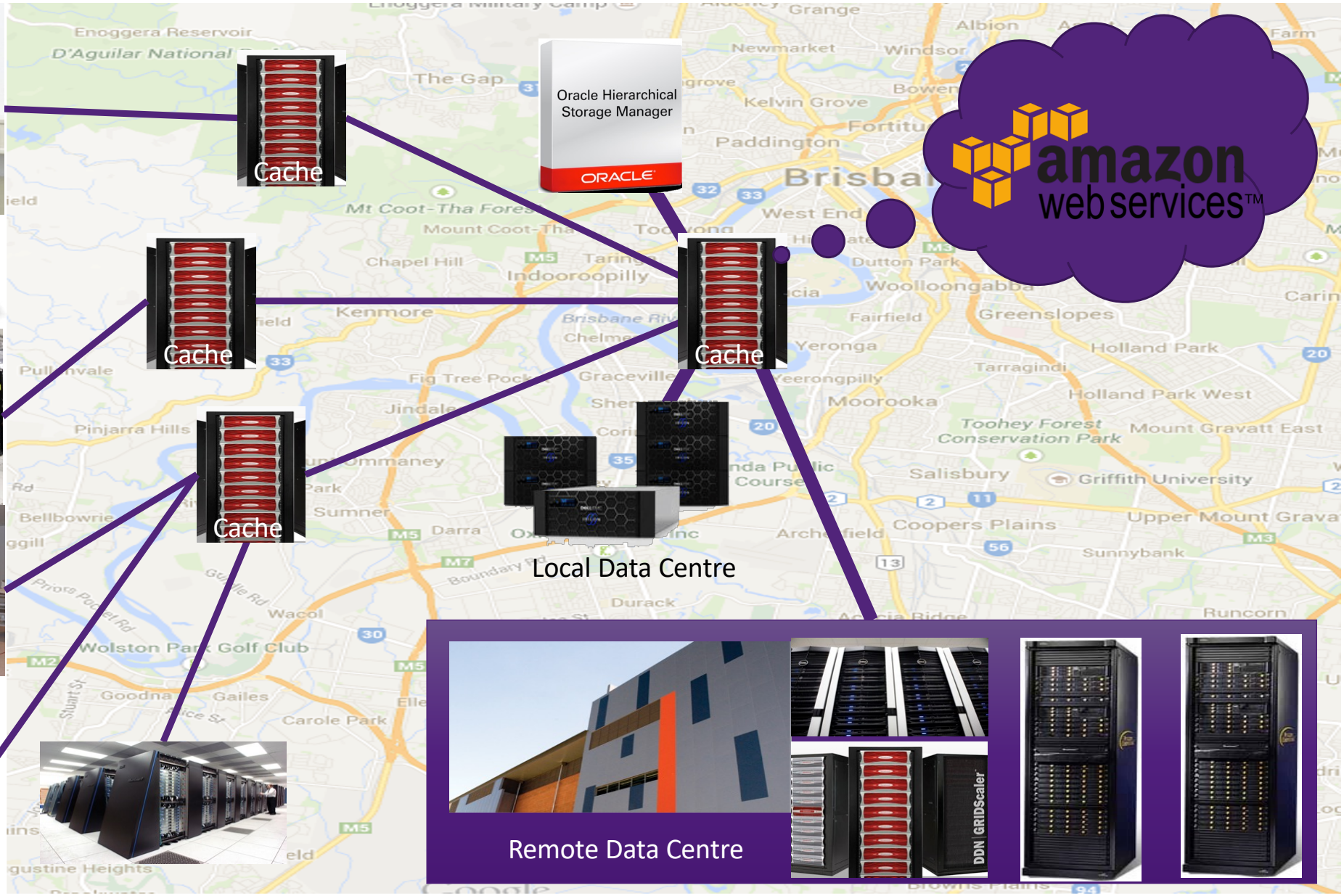
MeDiCI

- Centralising research data storage and computation
- Distributed data is further from both the instruments that ge
  the computers that process it, and the researchers that inte
- Existing mechanisms manually move data
- MeDiCI solves this by
  - Augmenting the existing infrastructure,
  - Implementing on campus caching
  - Automatic data movement
- Current implementation based on IBM Spectrum Scale (GPFS) and HPE DMF

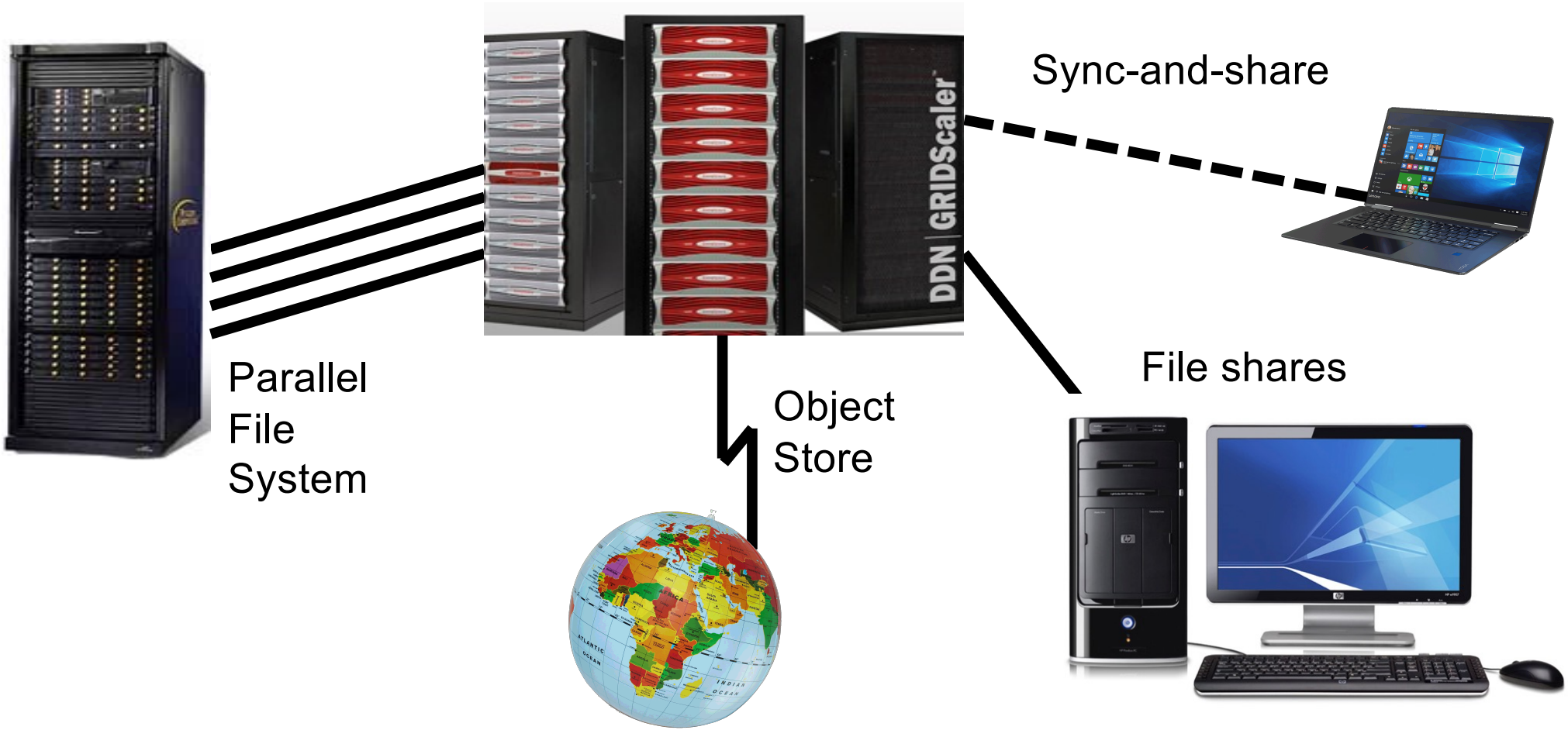


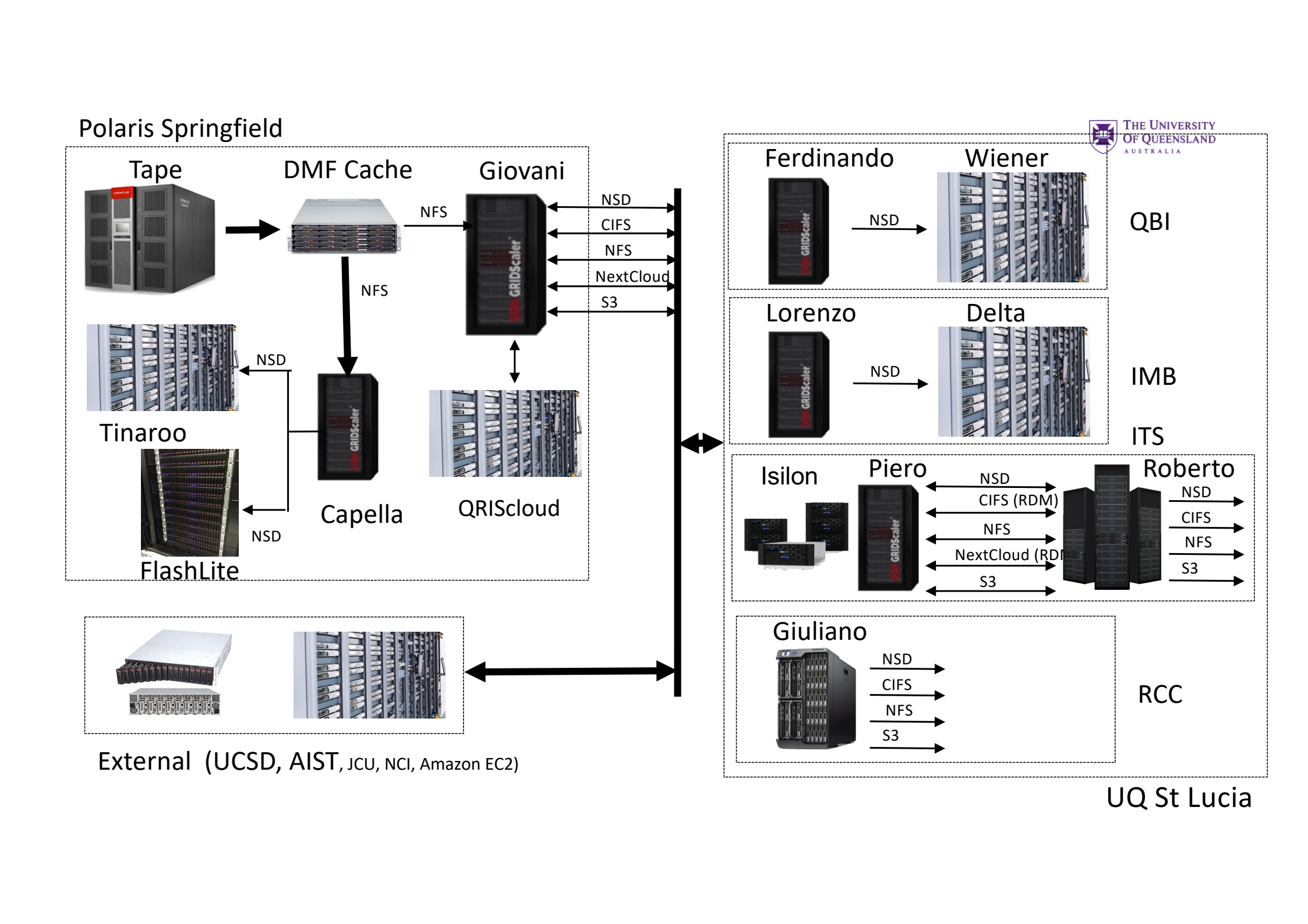

Cache
Cache
Cache
Cache
Brisbane
Remote Data Centre
DDN GRIDScaler




Cache
Oracle Hierarchical Storage Manager
ORACLE
amazon web services™
Cache
Cache
Cache
Local Data Centre
Remote Data Centre
DDN GRIDScaler

MeDiCI unifies data access




Polaris Springfield
Tape
DMF Cache
Giovani
NFS
NSD
CIFS
NFS
NextCloud
S3
NFS
NSD
Tinaroo
NSD
FlashLite
Capella
QRIScloud
External (UCSD, AIST, JCU, NCI, Amazon EC2)
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
Ferdinando
Wiener
NSD
QBI
Lorenzo
Delta
NSD
IMB
ITS
Isilon
Piero
Roberto
NSD
CIFS (RDM)
NFS
NextCloud (RDM
S3
NSD
CIFS
NFS
S3
Giuliano
NSD
CIFS
NFS
S3
RCC
UQ St Lucia

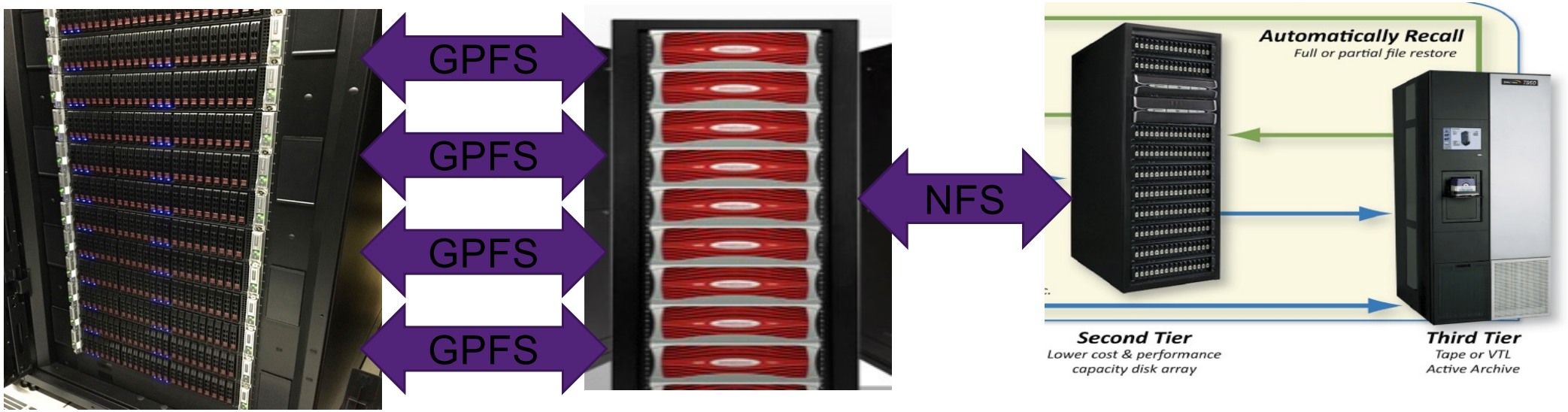
MeDiCI as a parallel file system

DDN SFA12KXE

FlashLite

Parallel file system

Accessing long term collections



DDN SFA12KXE

HPE DMF Disk/Tape

FlashLite

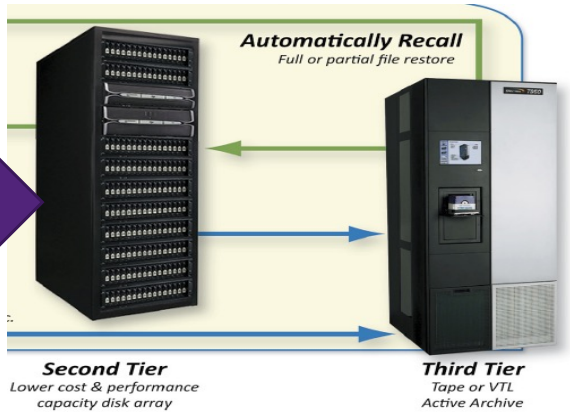
Parallel file system

Long term data collections

MeDiCI Wide Area Architecture

NFS

HPE DMF Disk/Tape



Machine Room Network

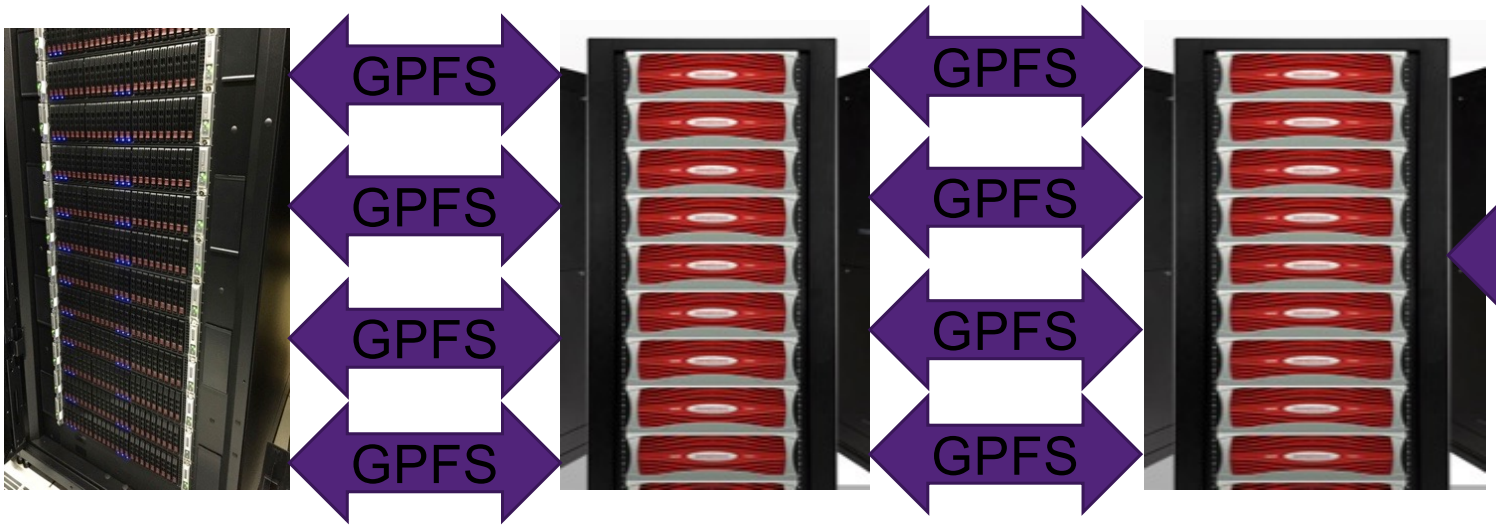
Compute

Parallel File System

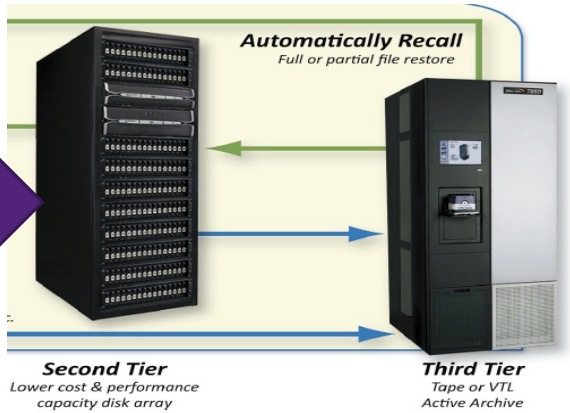
Long term data collections

Cache

MeDiCI Wide Area Architecture



HPE DMF Disk/Tape



Machine Room Network

Wide Area Network

Compute

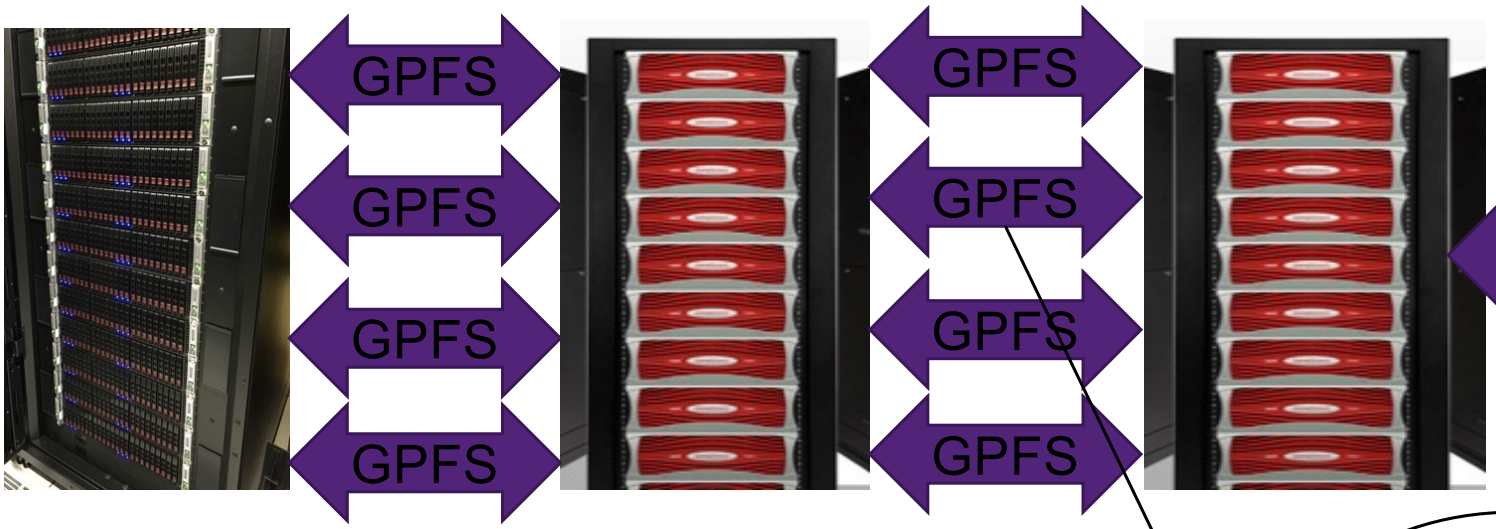
Parallel File System

Cache

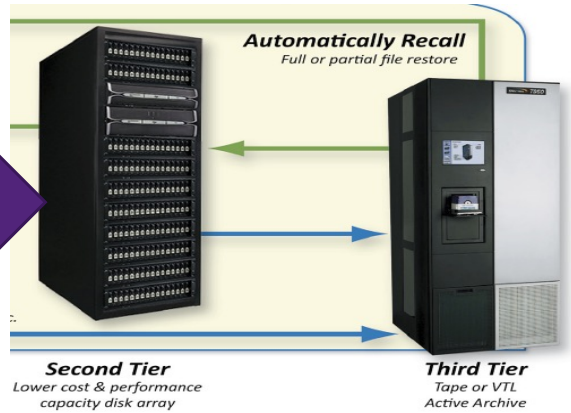
Cache

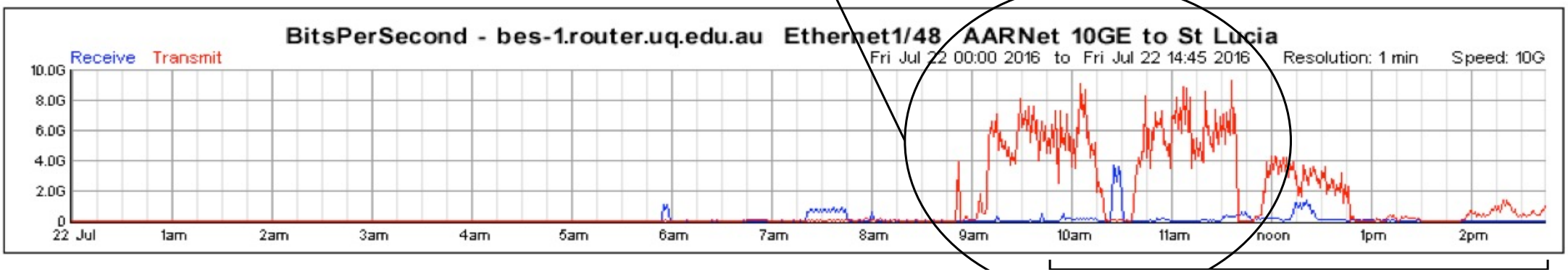
Long term data collections

MeDiCI Wide Area Architecture



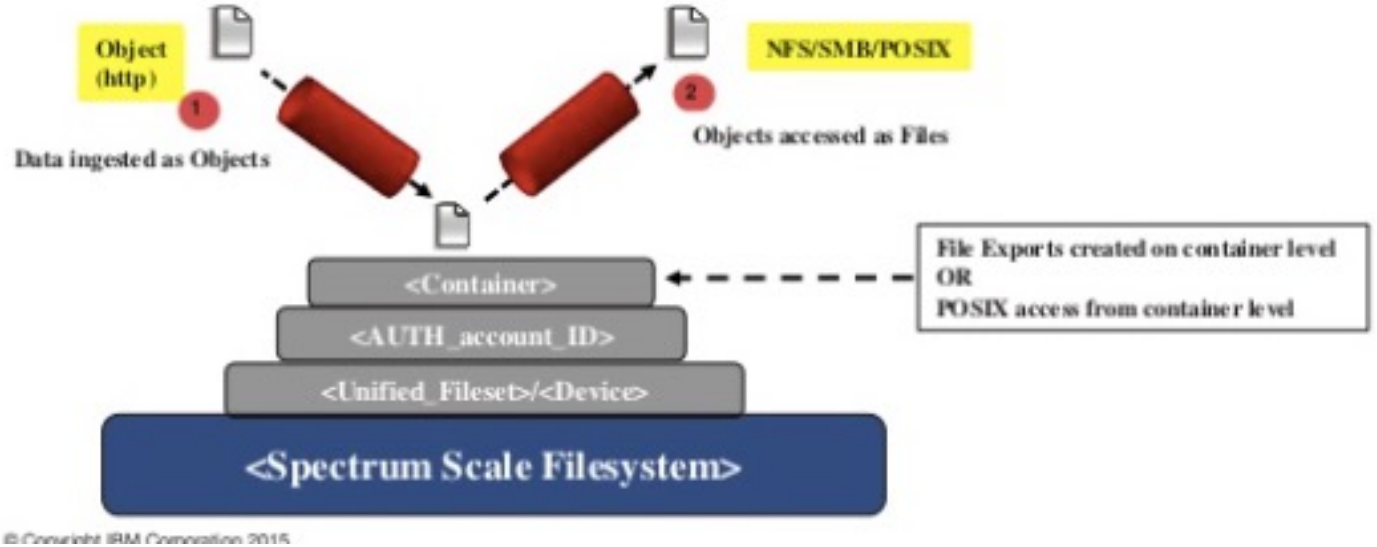
SGI DMF Disk/Tape





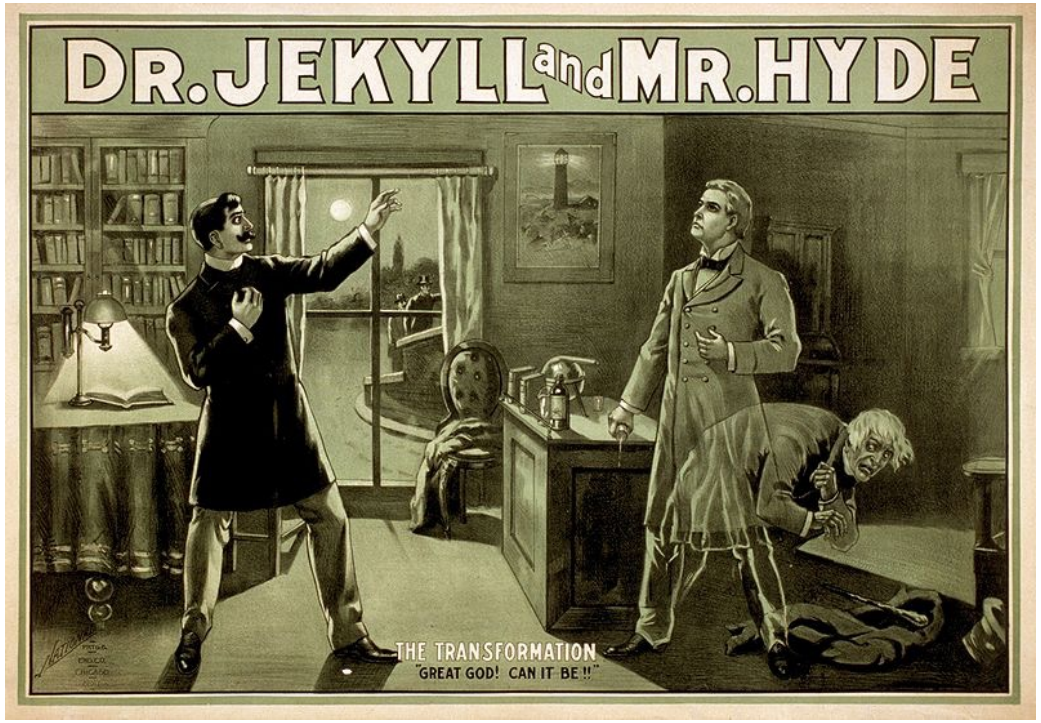
Object Storage

- S3 style objects becoming defacto standard for distributing data
- http put/get protocol
- Swift over GPFS
  - Unified Object/file interfaces



Identity!

- No single UID space across UQ/QCIF users
- Need to map UID space between UQ and Polaris
- GPFS 4.2
  - mmname2uid/mmuid2name

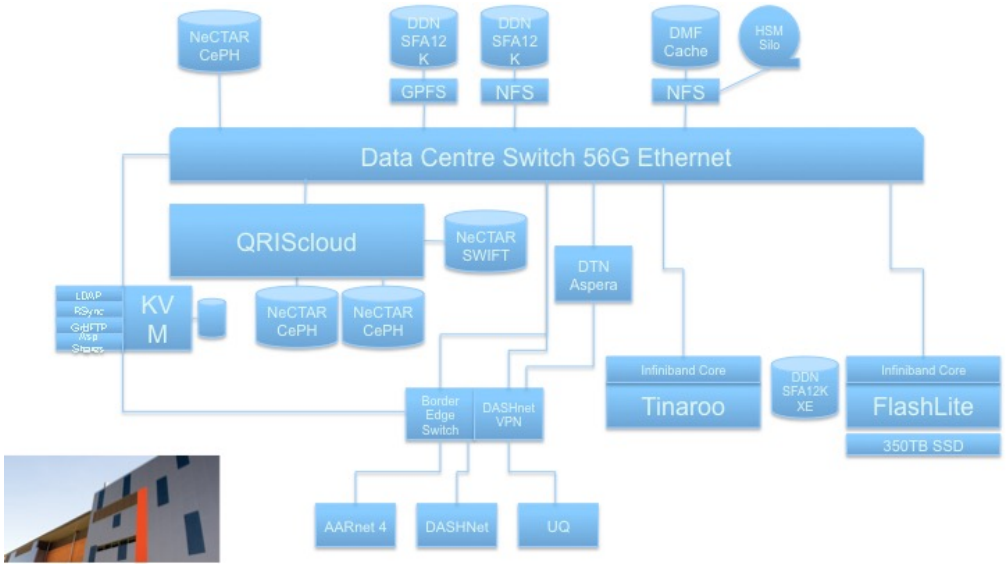


Building on basic architecture

- A Declarative Machine Room
- Leveraging Cloud Storage
- Very Very Wide Area File Systems
- Supporting repository stacks
- Orchestrating Workflows

A Declarative Machine Room?

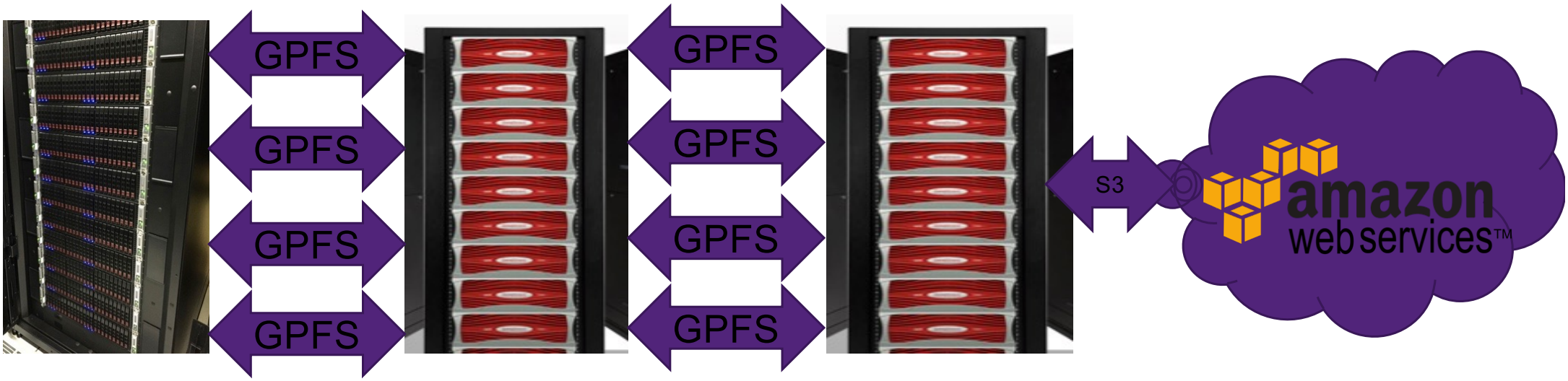
- Static allocation of disk and tape
- Policy driven allocation



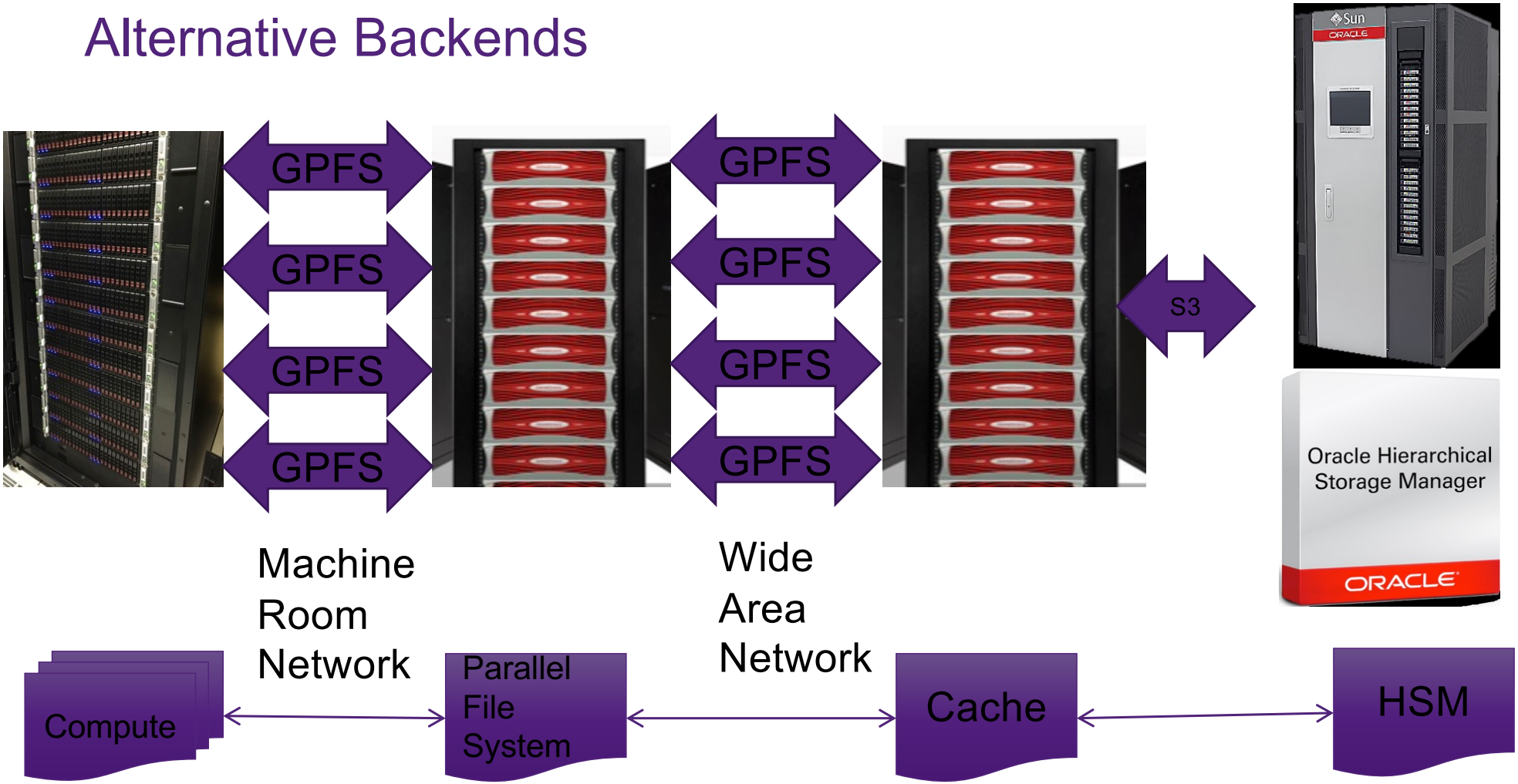

```
RULE 'prefetch-list'
LIST 'toevict'

WHERE  CURRENT_TIMESTAMP - ACCESS_TIME >
INTERVAL  '7' DAYS
  AND  REGEX(misc_attributes,'[P]') /* only list AFM managed files
*/
```

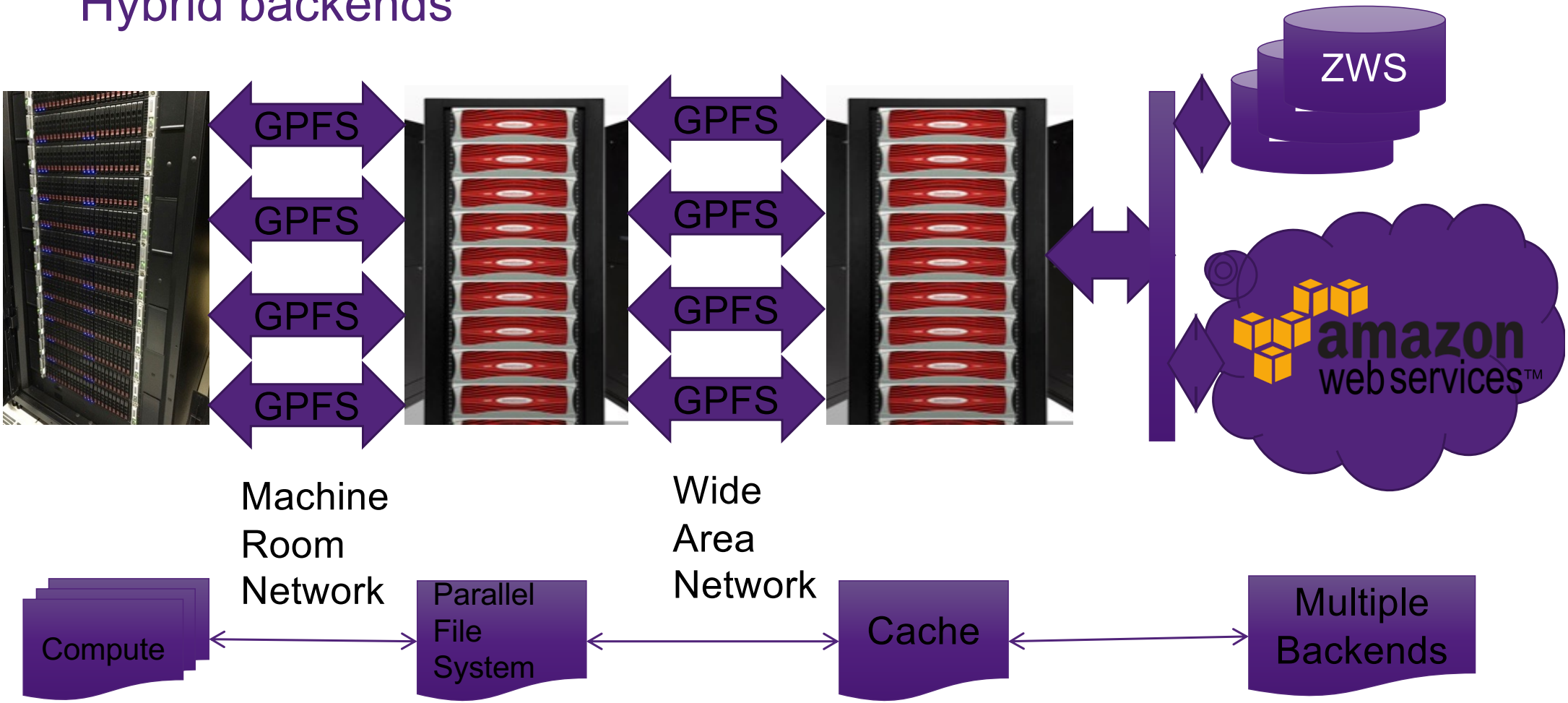
MeDiCI Very Wide Area Architecture



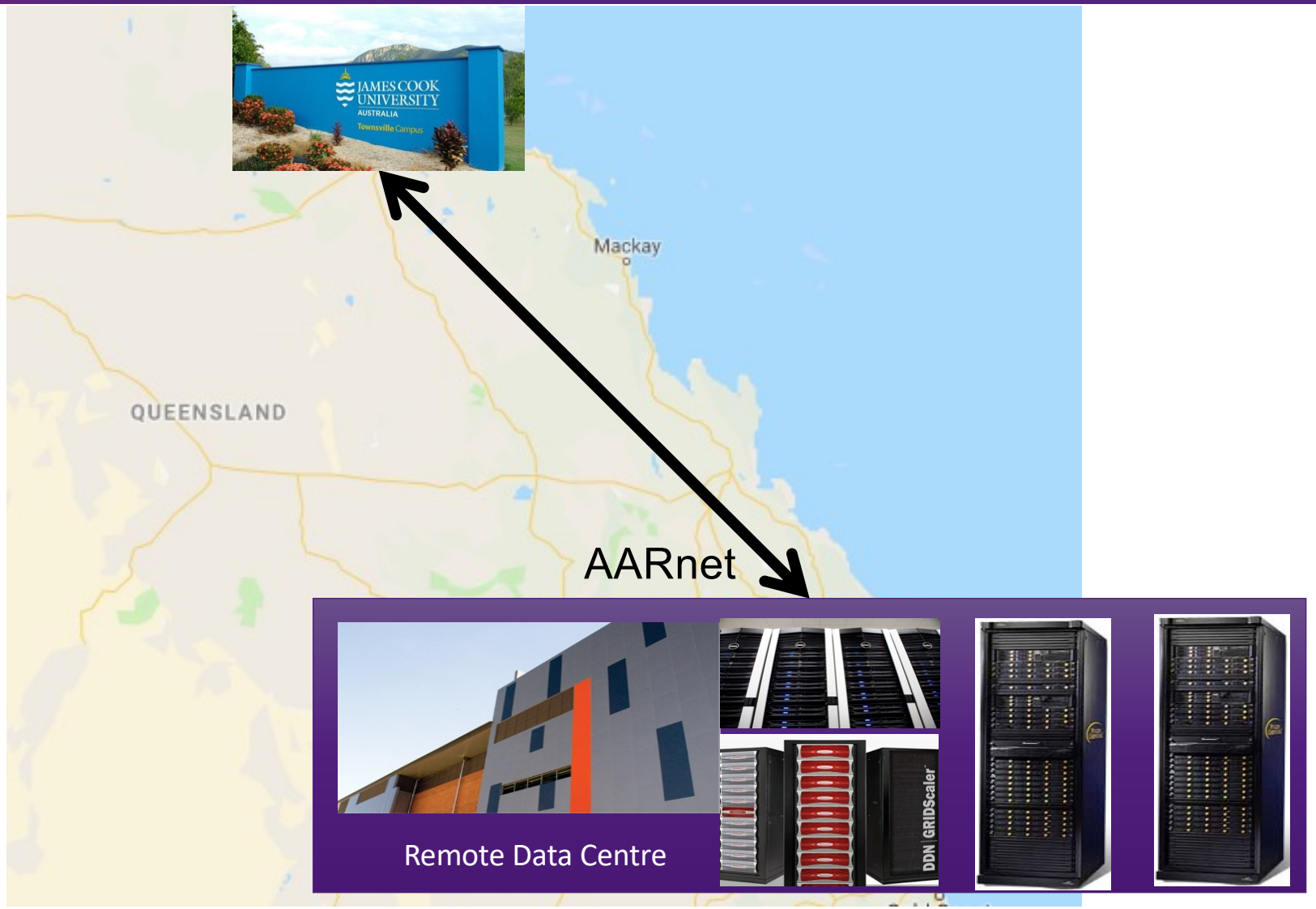
Alternative Backends



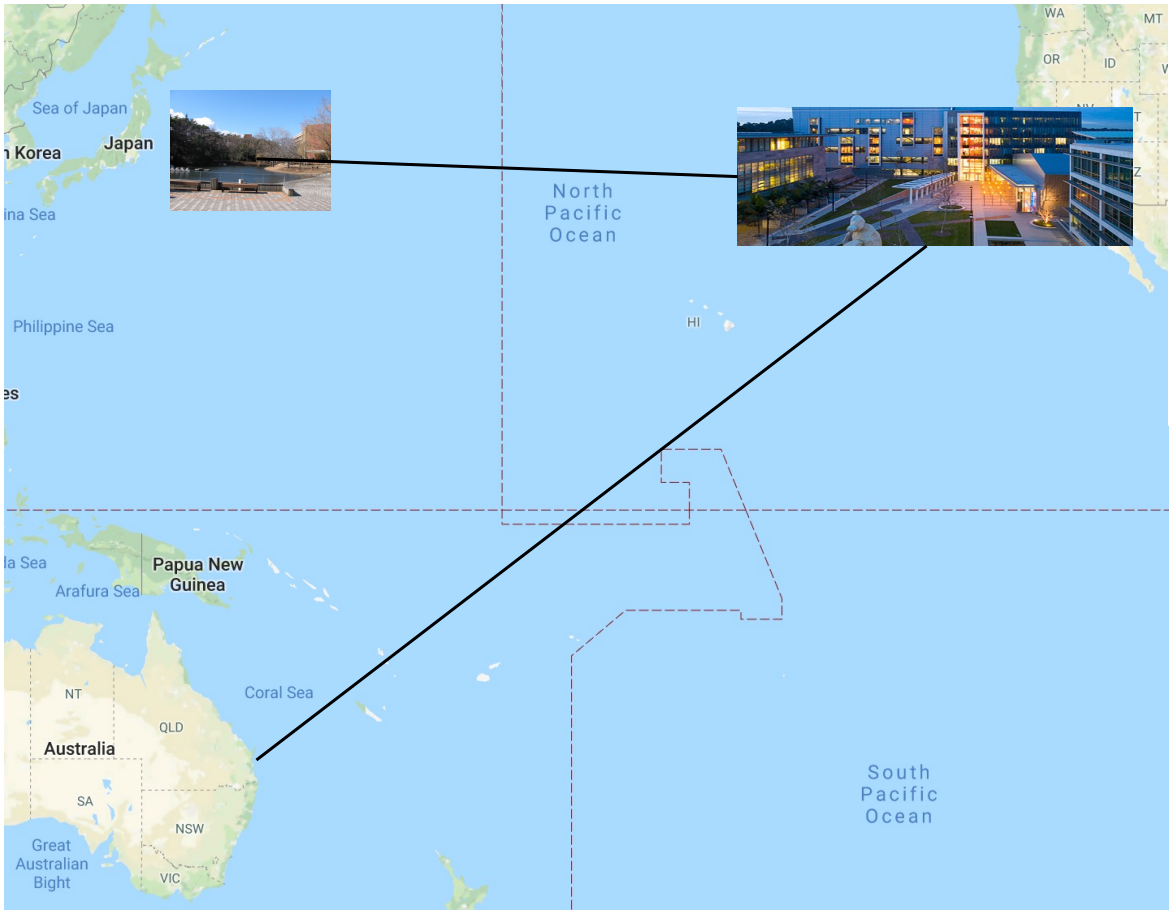
Hybrid backends

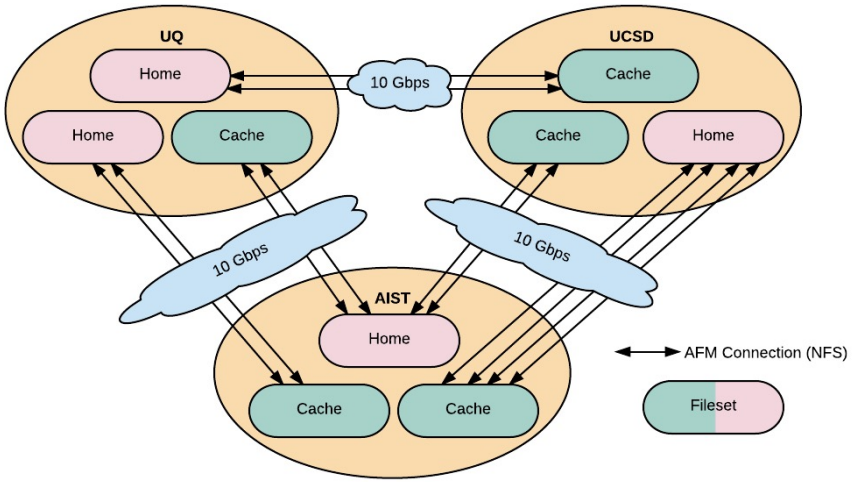


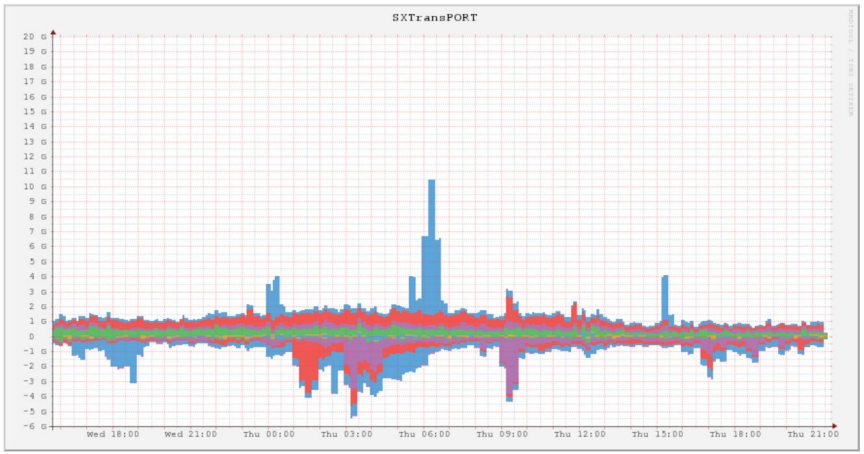
MeDiCI goes North



MeDiCI goes East



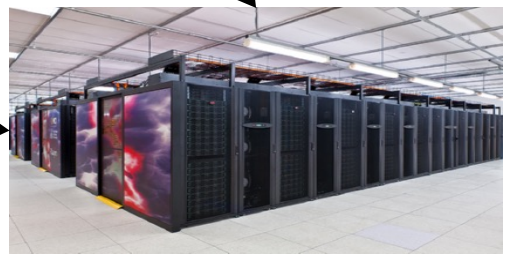




MeDiCI goes South and West

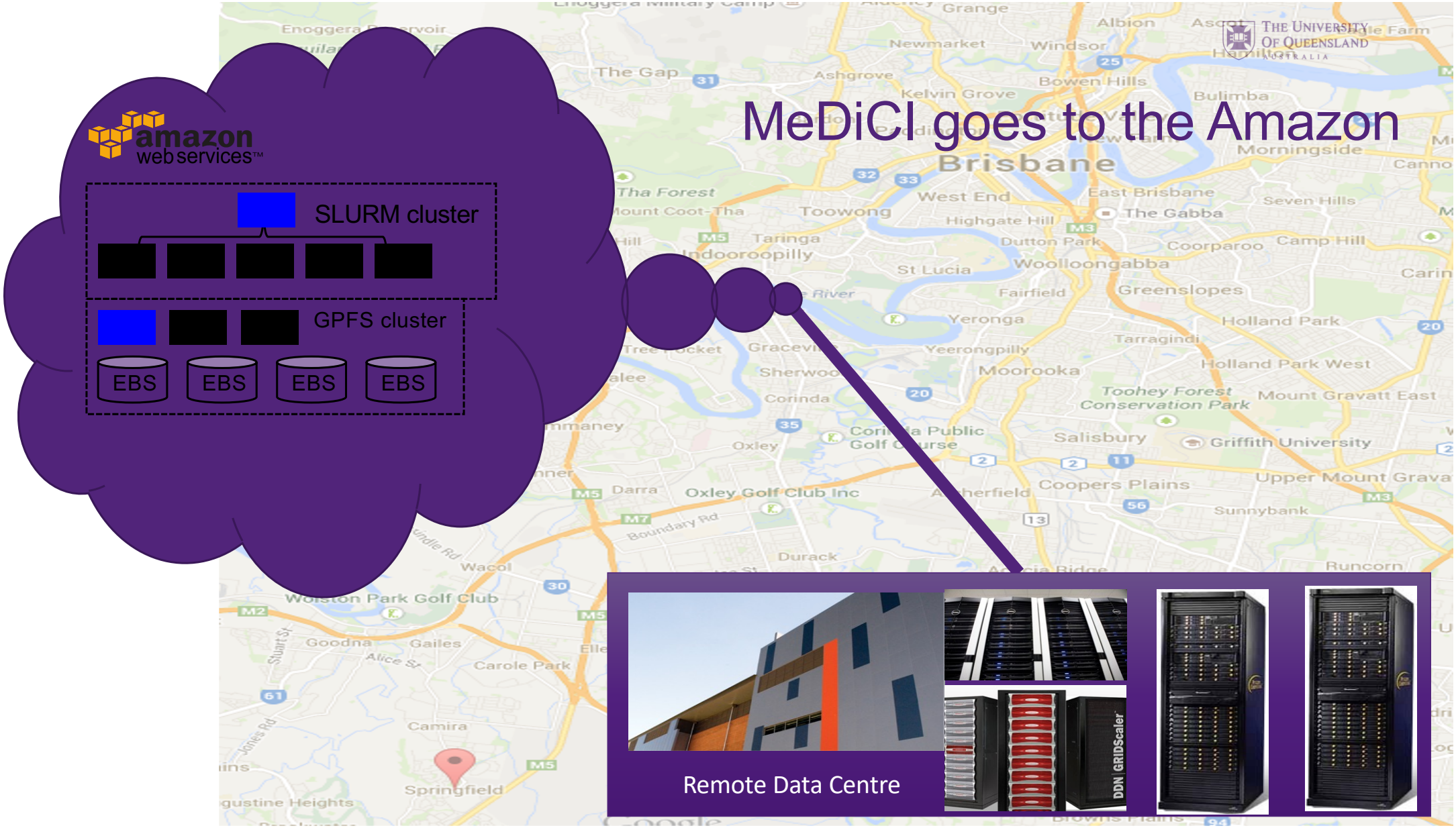
AARnet X

Magnus, Pawsey

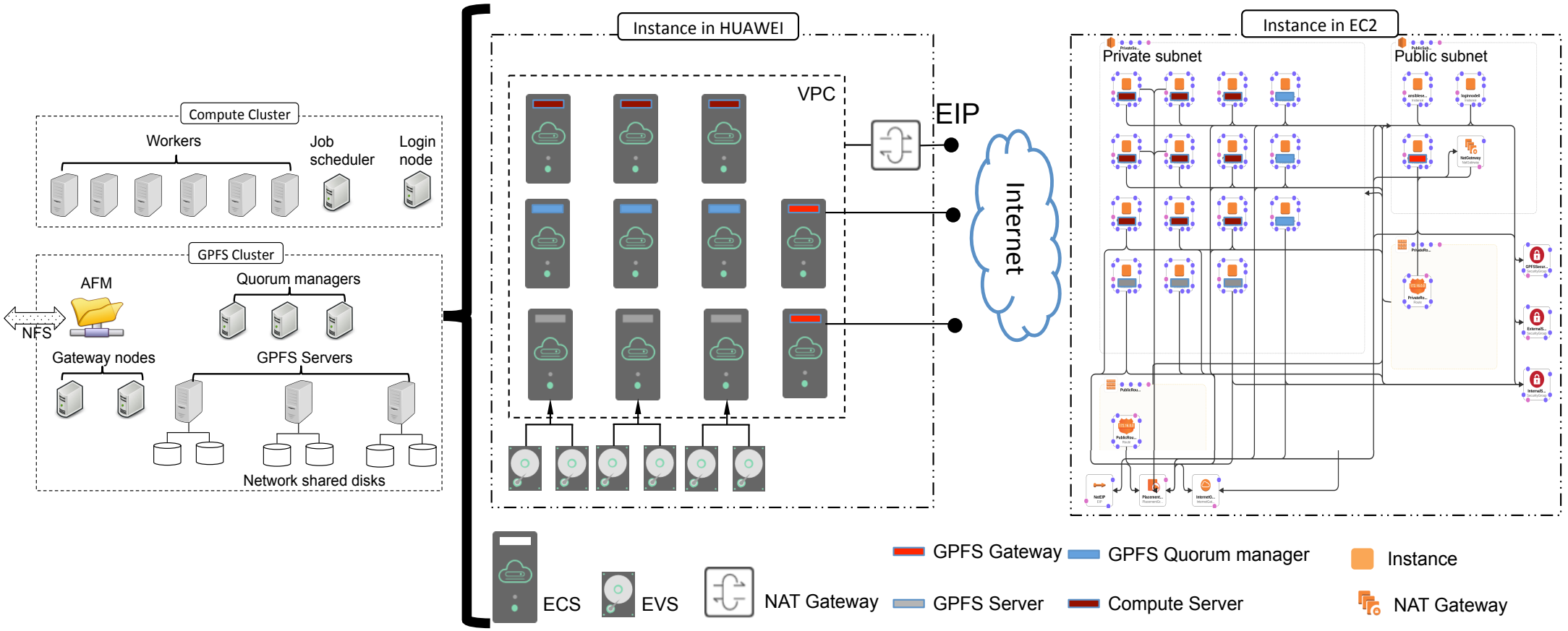


Raijin, NCI

MeDiCI goes to the Amazon



Deployment in Amazon and Huawei clouds

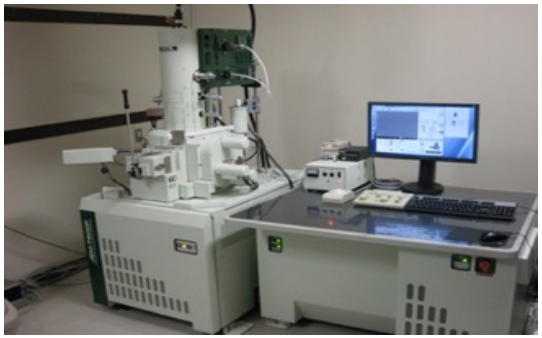


CAMERA

Unmanaged data workflow

Research Data Manager

Designed at The University of Queensland



RDM creates collection

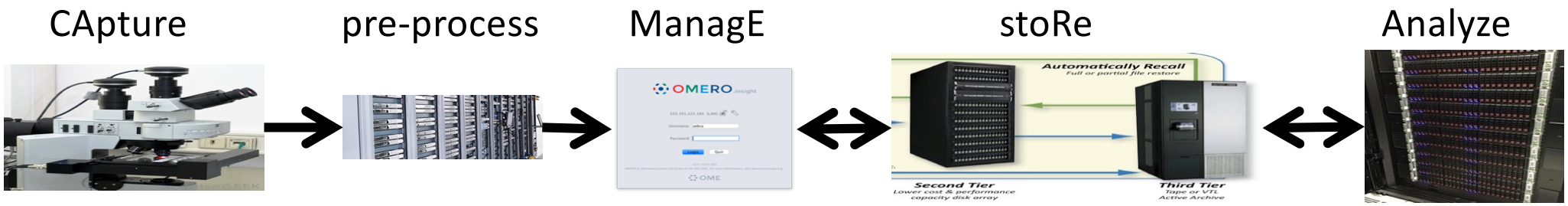
Collection mounted on scope

Data copied to collection

Collection mounted on HPC for analysis

CAMERA: MeDiCI meets instruments

- MeDiCI abstracts storage location and protocol issues
- RDM captures experiment level meta-data and provisions storage
- Repository captures instrument level meta-data
- Collections can be attached to repository stacks



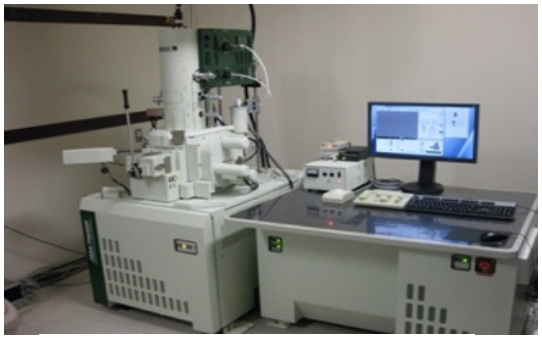
CAMERA …

- supports a complete life cycle for instrument-gathered data, seamlessly rendering it on a range of instruments, cloud systems, desktops and high-performance computers (HPCs).
- Leverages powerful underlying technologies, such as high-throughput networks and storage systems while hiding this complexity from the users.
- appears as a single platform, but its backend is actually an aggregate of the best-of-breed technologies, such as data repositories and metadata management systems.
- achieves seamless inter-operation with HPCs without the need to copy files in and out of a repository.

Managed data workflow

Research Data Manager

Designed at The University of Queensland



OME

RDM creates collection

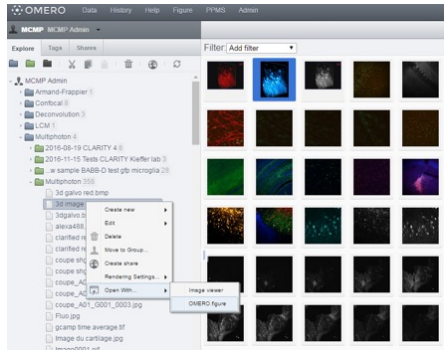
Collection mounted on repository

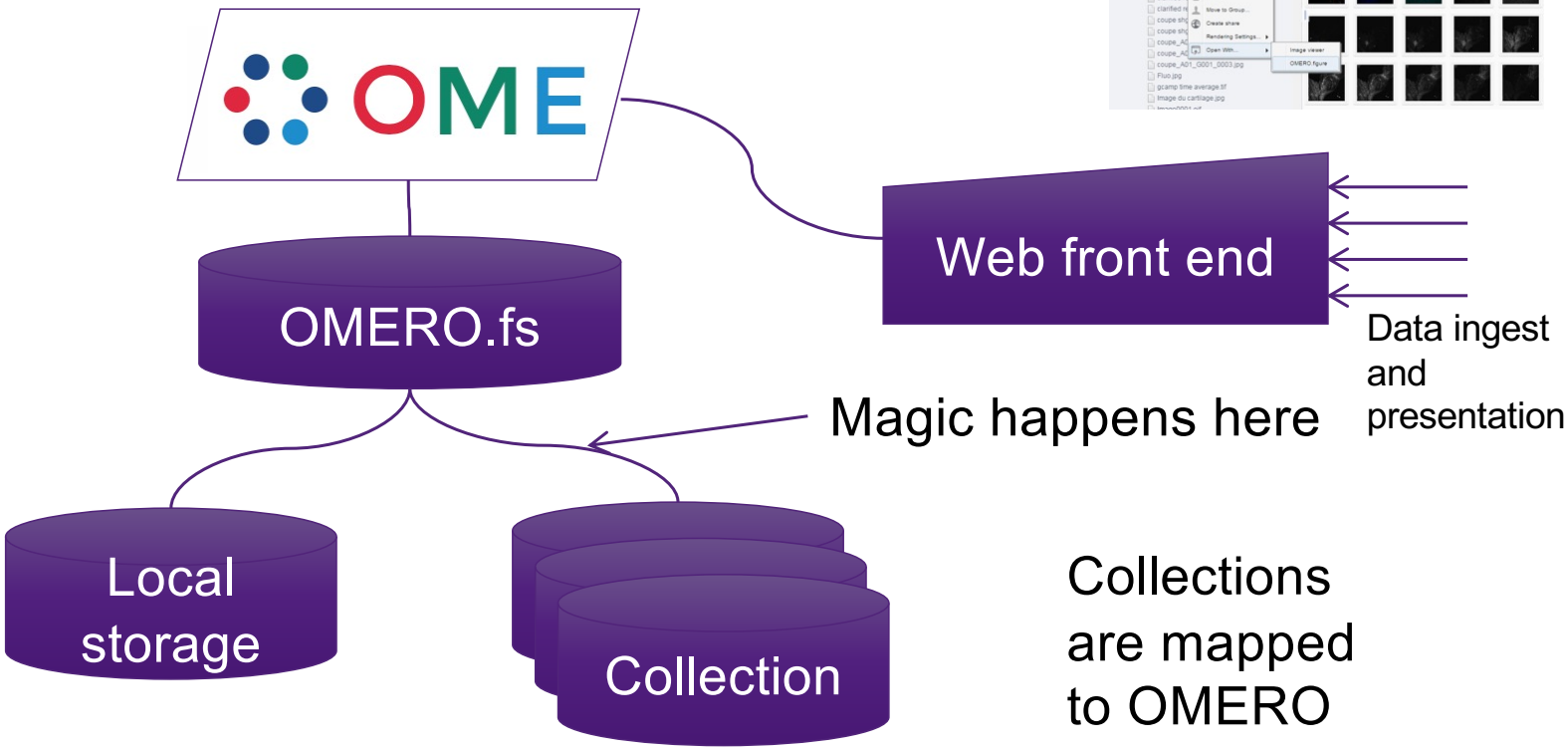
Data ingested into repository

Collection mounted on HPC for analysis

Managed Data Implementation

Cloud Hosted OMERO server

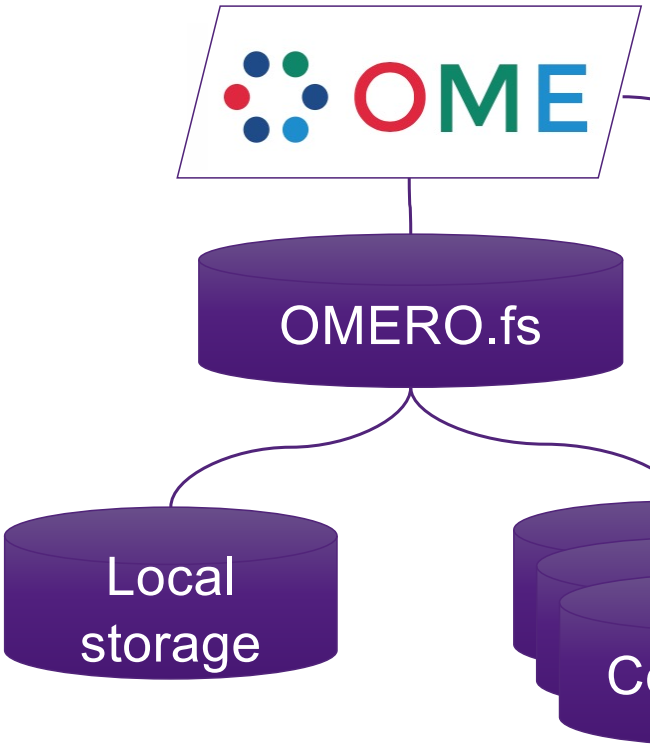


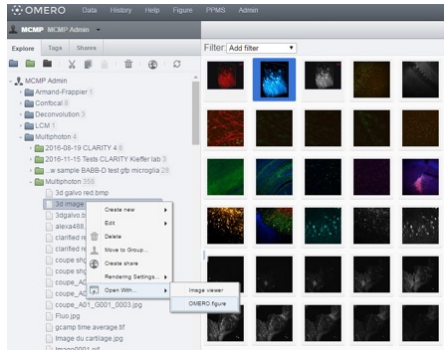


Collections are mapped to OMERO groups

Managed Data Implementation

Cloud Hosted OMERO server





Web front end

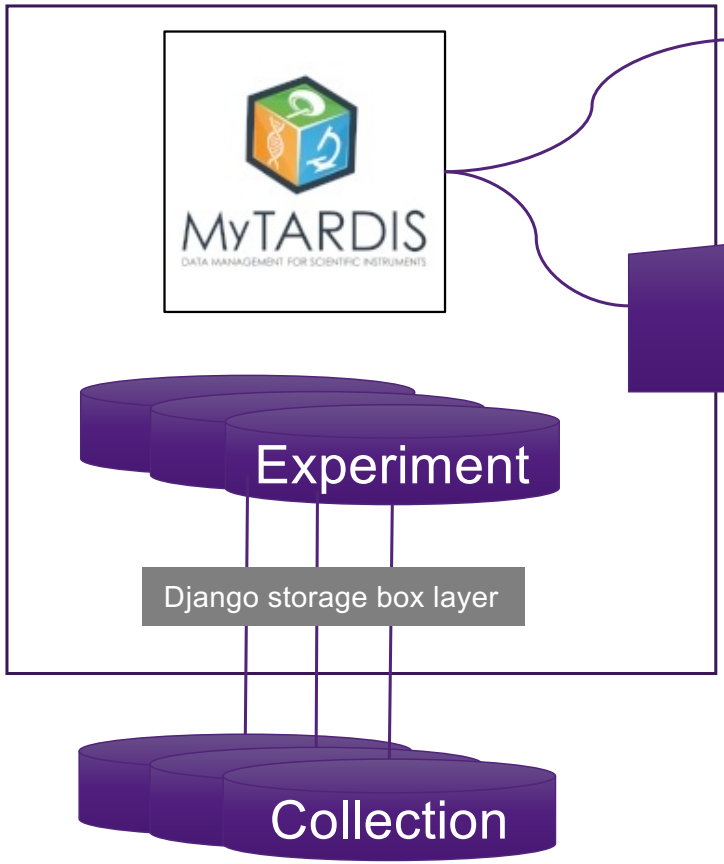
Data ingest and presentation

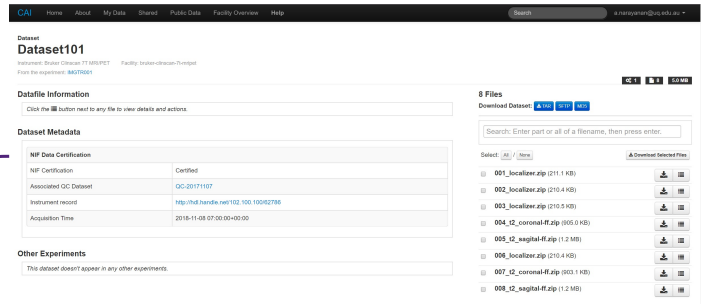
Analyse

Read Only

Managed Data Implementation

Cloud Hosted MyTardis server





Collections are mapped to MyTardis Experiments

Conclusions

- Data Intensive Science
  - Data meets computation seamlessly
- From cradle to grave
  - From RDM to processing
- MeDiCI Data Fabric
- CAMERA for managed instrument data

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Thank you

David Abramson| Professor and Director
Research Computing Centre
David.Abramson@uq.edu.au
07 0000 000

facebook.com/uniofqld

Instagram.com/uniofqld

twitter.com/RCCUQ

facebook.com/rccuq

rcc.uq.edu.au