THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

A Cache-based Data Movement Infrastructure for On-demand Scientific Cloud Computing

David Abramson, Jake Carroll, Chao Jin, Michael Mallon, Zane van Iperen, Hoang Nguyen, Allan McRae

University of Queensland

Liang Ming

Huawei

Data-Intensive Scientific Computing

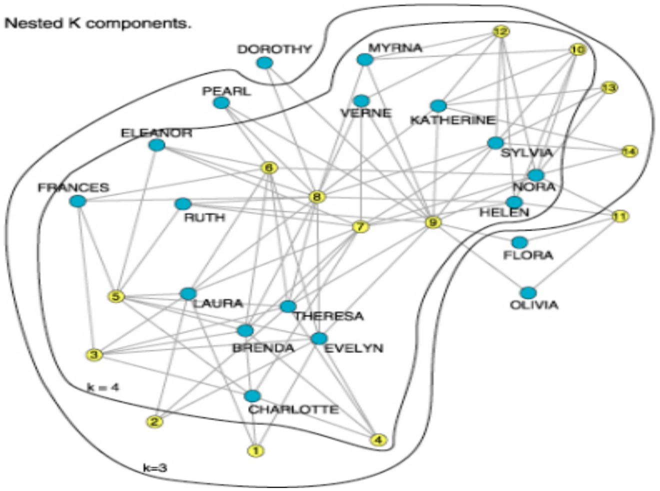
Very large data-sets or very large input-output requirements

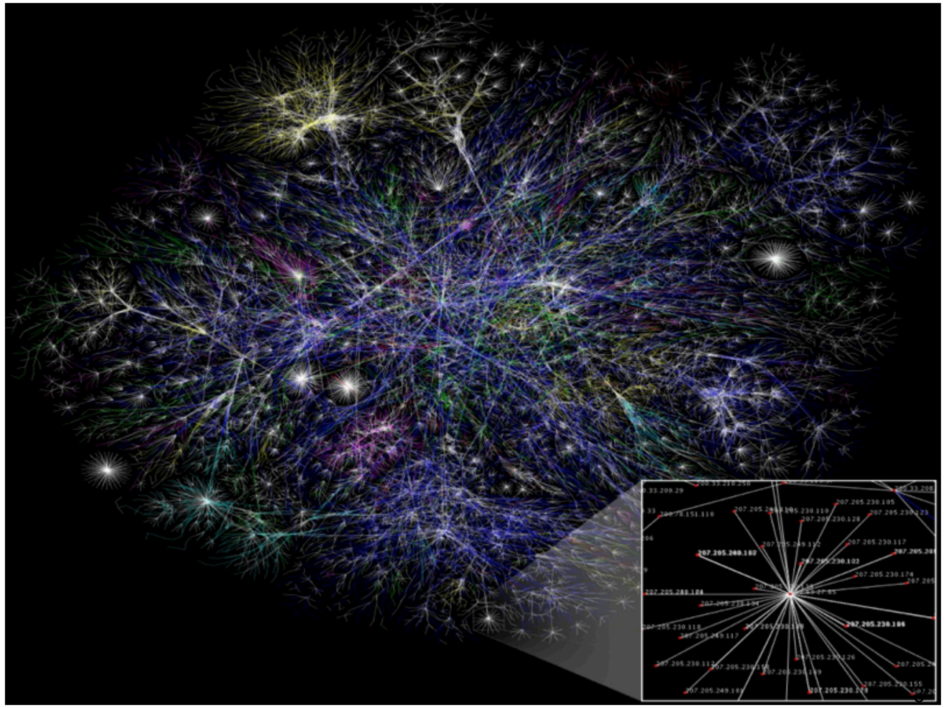
Two data-intensive application classes are important and growing

Data-Intensive Computing

Examples Applications:

- Genome sequence assembly
- Climate simulation analysis
- Social network analysis





Infrastructure for Data Intensive Computing

Computation

- Large amounts of main memory
- Parallel processors
- Smooth out memory pyramid

Storage

- Significant long term storage
- Smooth out the memory pyramid
- Many views of same data
  - Parallel File System
  - Local access (POSIX)
  - Remote collaboration and sharing (Object store)
  - Sync-and-share
  - Web
  - Cloud

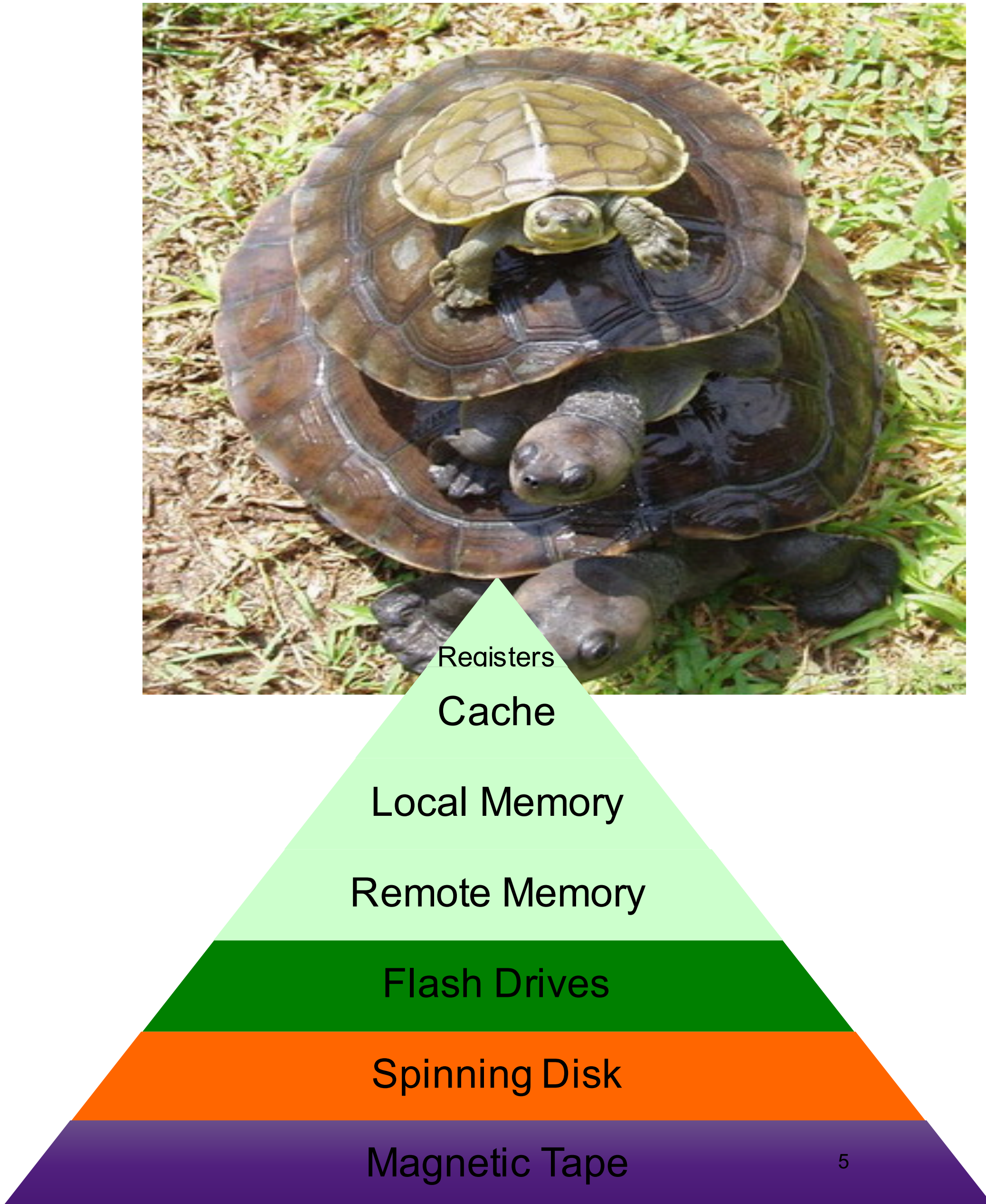
amazon web services™

~~Turtles~~ Caches all the way down

“a jocular expression of the infinite regress problem in cosmology posed by the "unmoved mover" paradox.

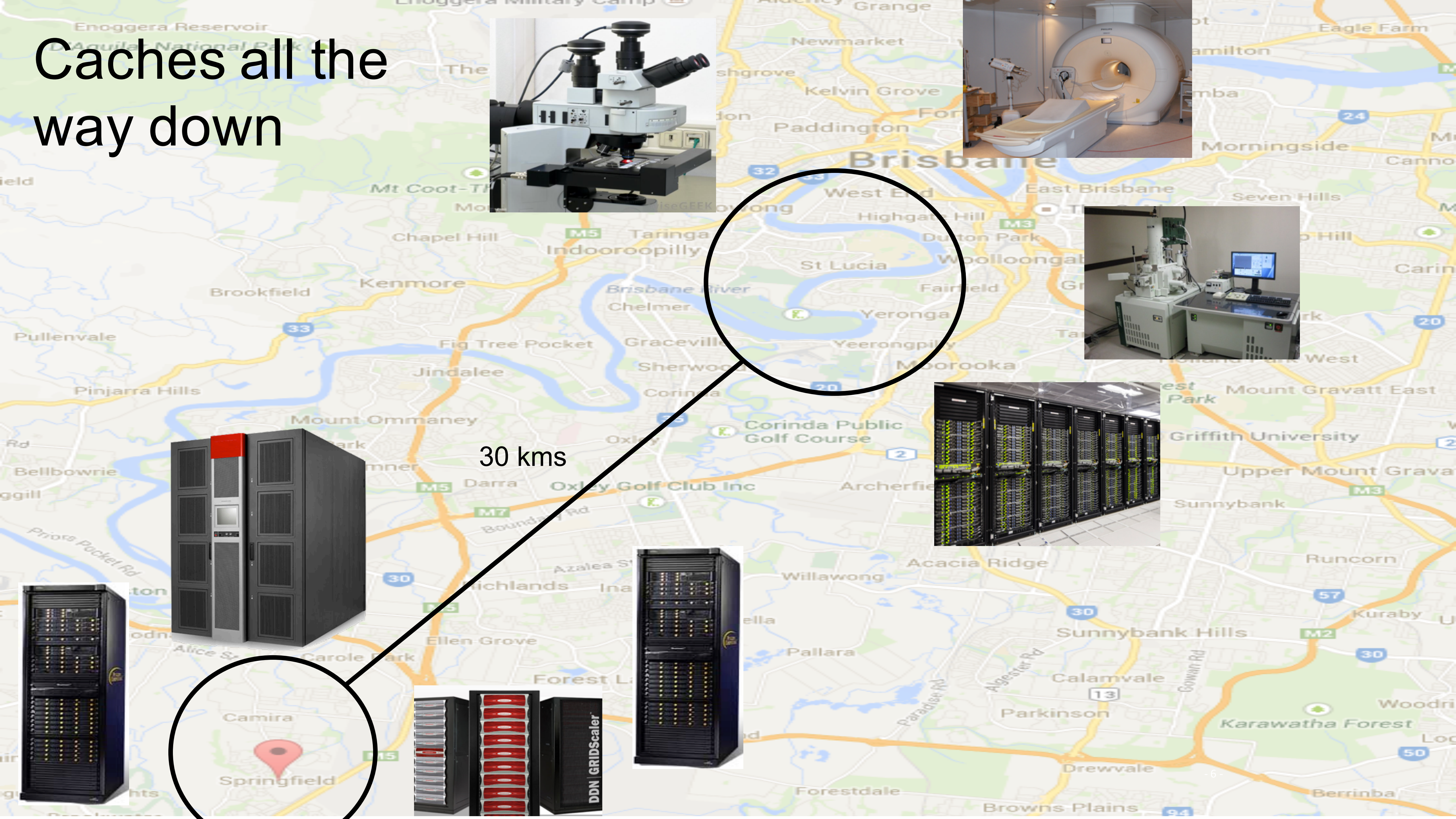
The metaphor in the anecdote represents a popular notion of the theory that Earth is actually flat and is supported on the back of a World Turtle, which itself is propped up by a chain of larger and larger turtles.

Questioning what the final turtle might be standing on, the anecdote humorously concludes that it is turtles all the way down””



https://en.m.wikipedia.org/wiki/Turtles_all_the_way_down

Caches all the way down



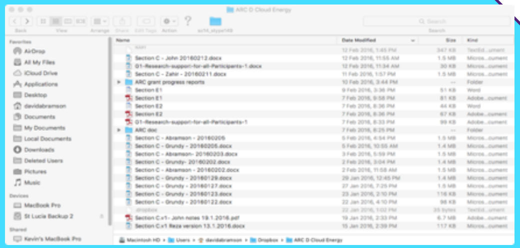
Data Data everywhere anytime

THE UNIVERSITY OF QUEENSLAND AUSTRALIA



ImageTrove | myTardis | OMERO

Managed Data



MeDiCI — Synchronous

ownCloud — Asynchronous

Unmanaged Data

OpenClinica

Clinical Data

S3, Swift

Cloud Access

MeDiCI





QRIScloud Compute and Storage Fabric

MeDiCI

Centralising research data storage and computation

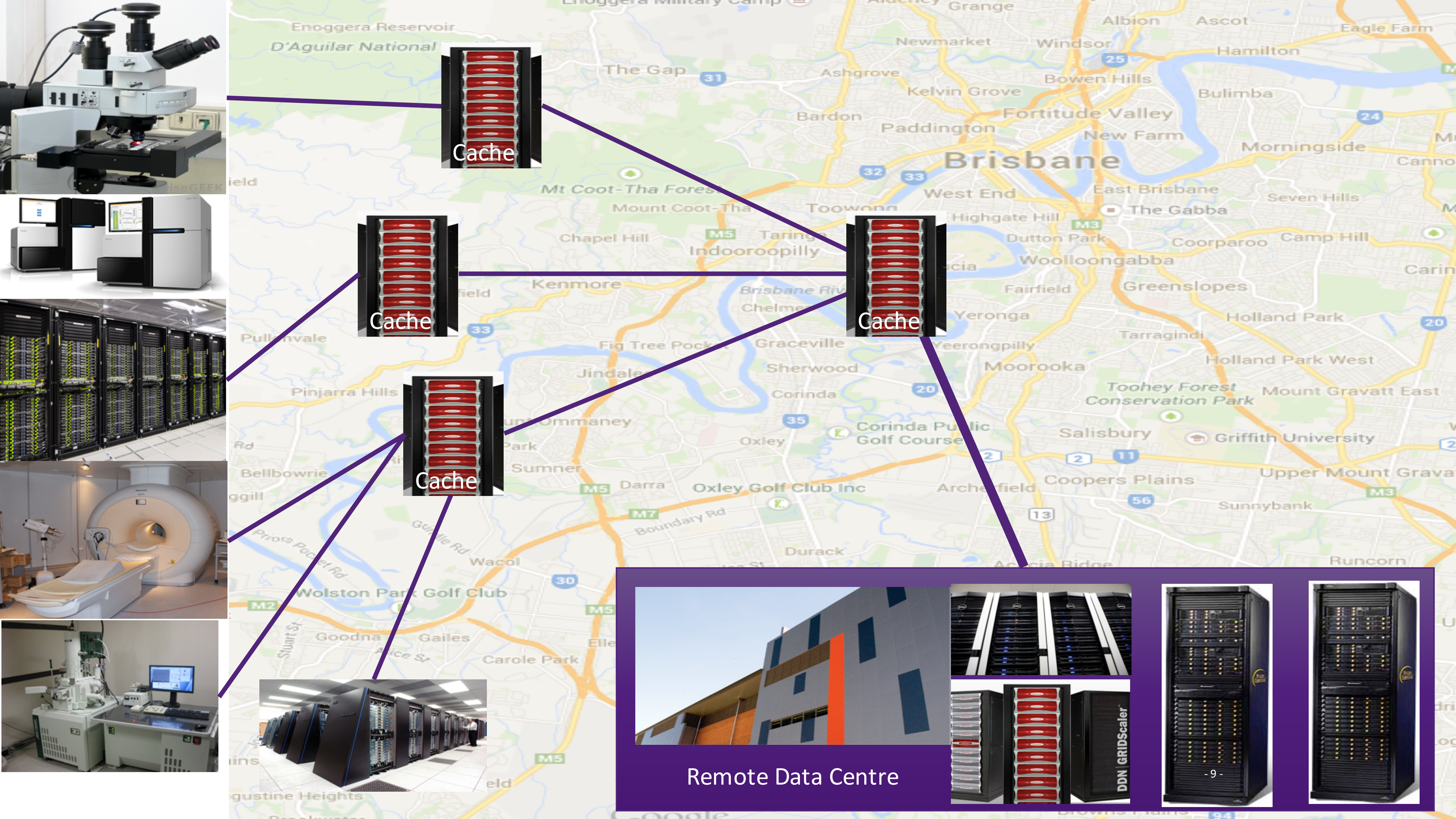
Distributed data is further from both the instruments that generate it, some of the computers that process it, and the researchers that interpret it.

Existing mechanisms manually move data

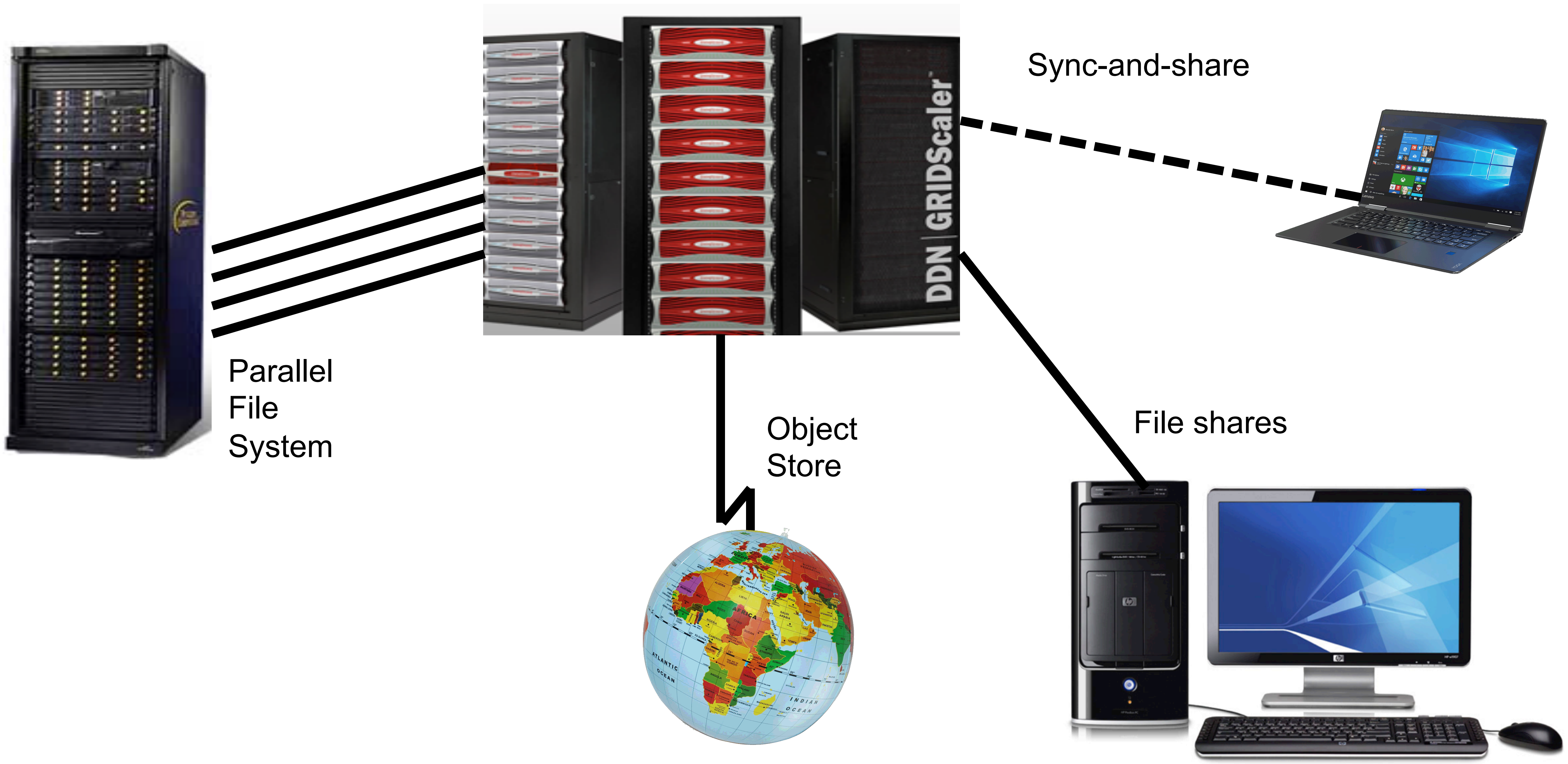
MeDiCI solves this by

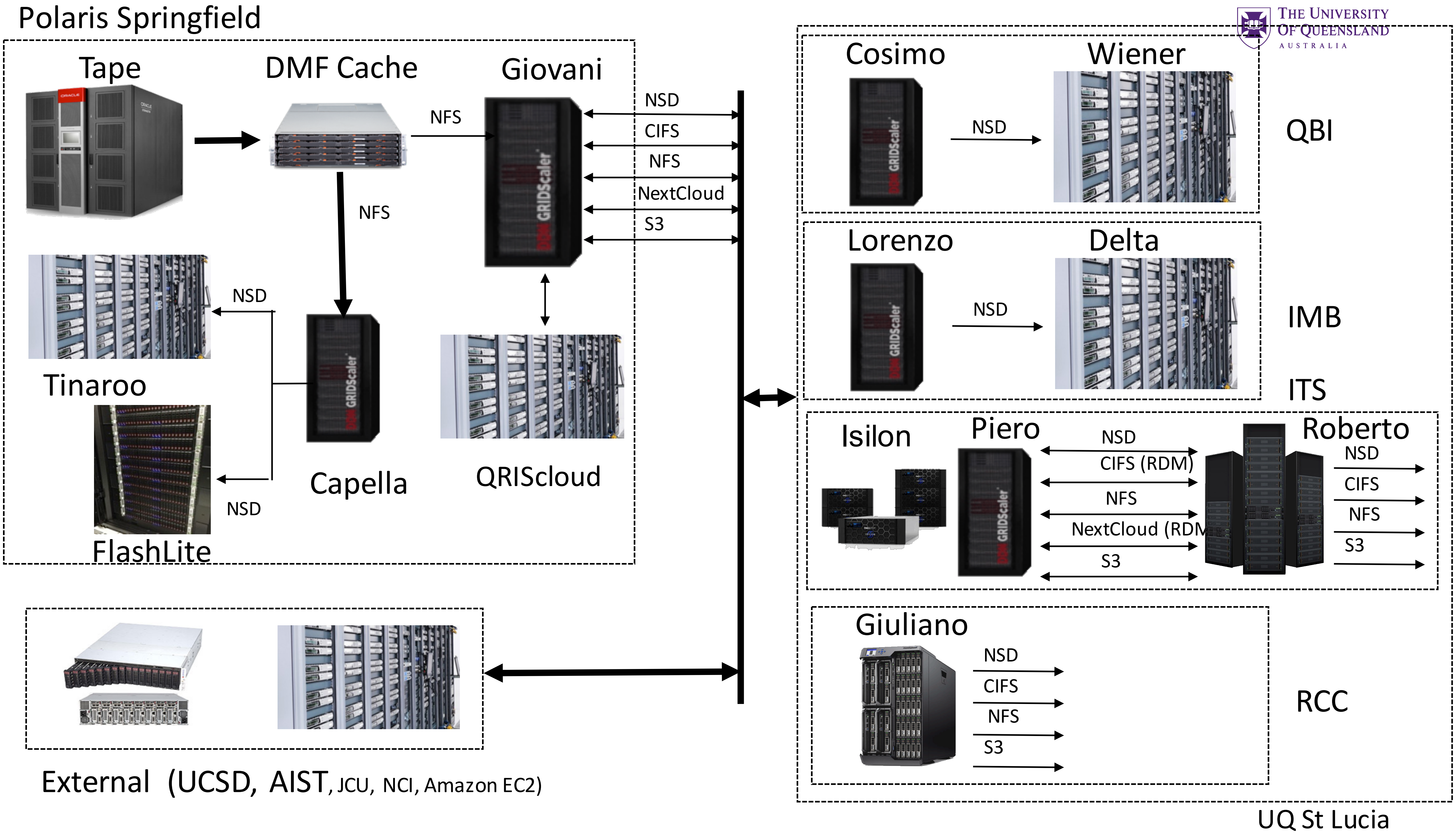
- Augmenting the existing infrastructure,
- Implementing on campus caching
- Automatic data movement

Current implementation based on IBM Spectrum Scale (GPFS) and SGI DMF


Cache
Cache
Cache
Cache
Remote Data Centre
DDN GRIDScaler

MeDiCI unifies data access

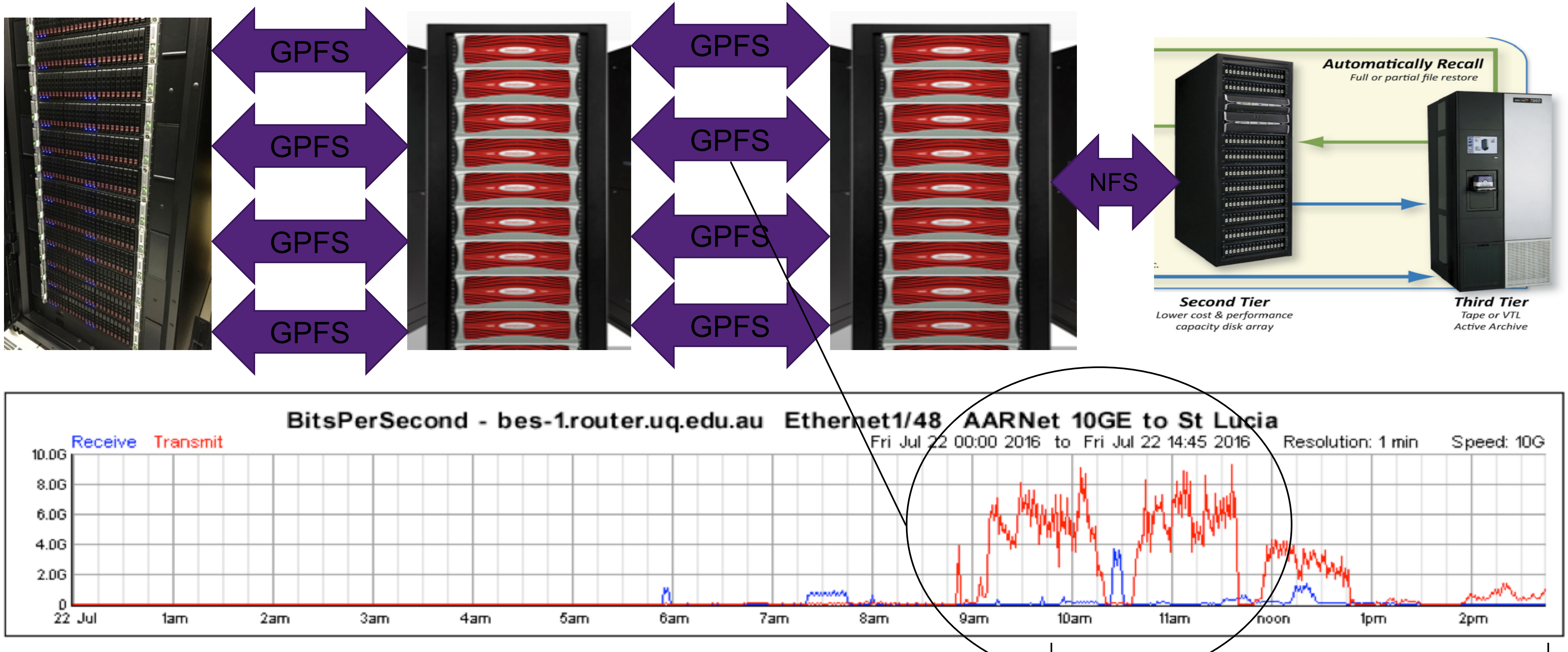




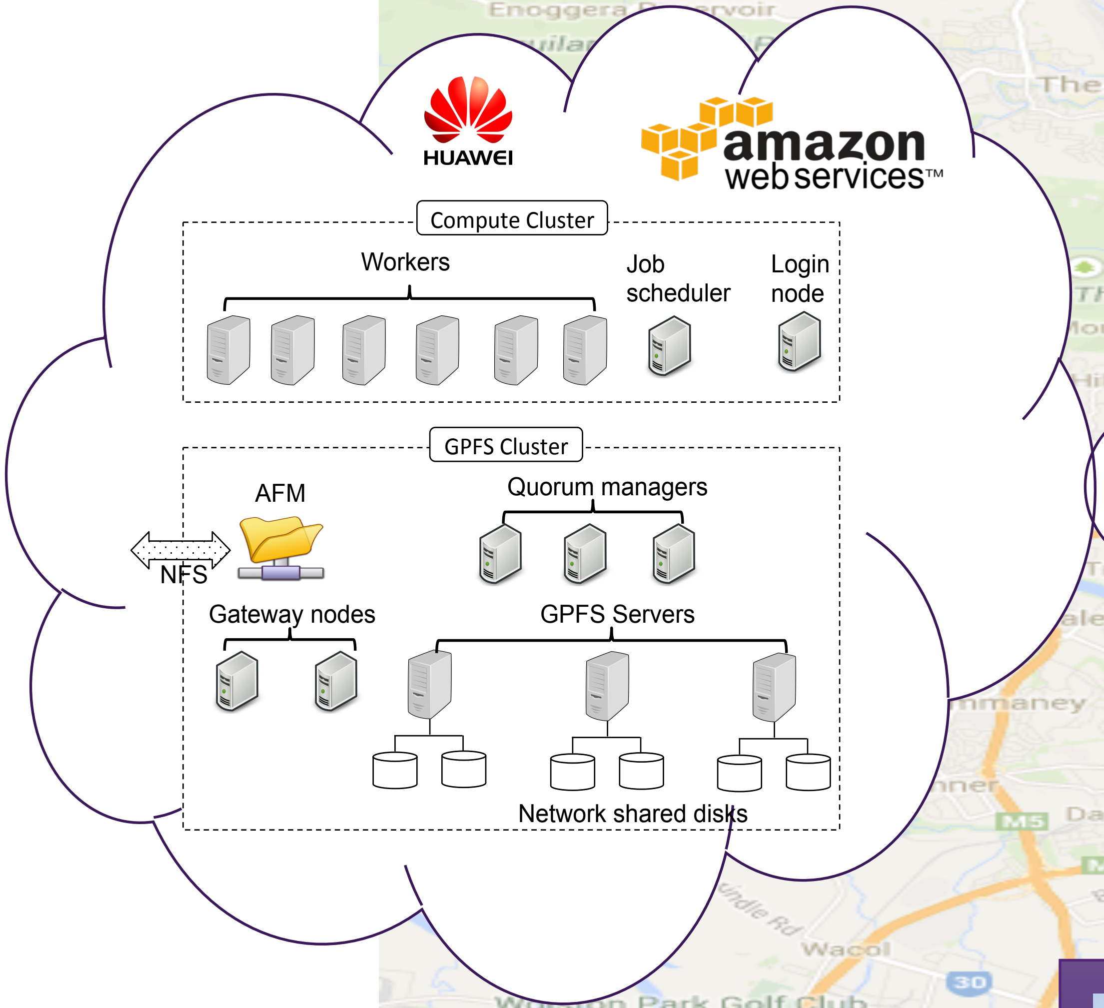
Polaris Springfield
Tape
DMF Cache
Giovani
NFS
NSD
CIFS
NFS
NextCloud
S3
NFS
NSD
Tinaroo
NSD
FlashLite
Capella
QRIScloud
Cosimo
Wiener
NSD
QBI
Lorenzo
Delta
NSD
IMB
ITS
Isilon
Piero
Roberto
NSD
CIFS (RDM)
NFS
NextCloud (RDM
S3
NSD
CIFS
NFS
S3
Giuliano
NSD
CIFS
NFS
S3
RCC
External (UCSD, AIST, JCU, NCI,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

MeDiCI Wide Area Architecture

HPE DMF Disk/Tape

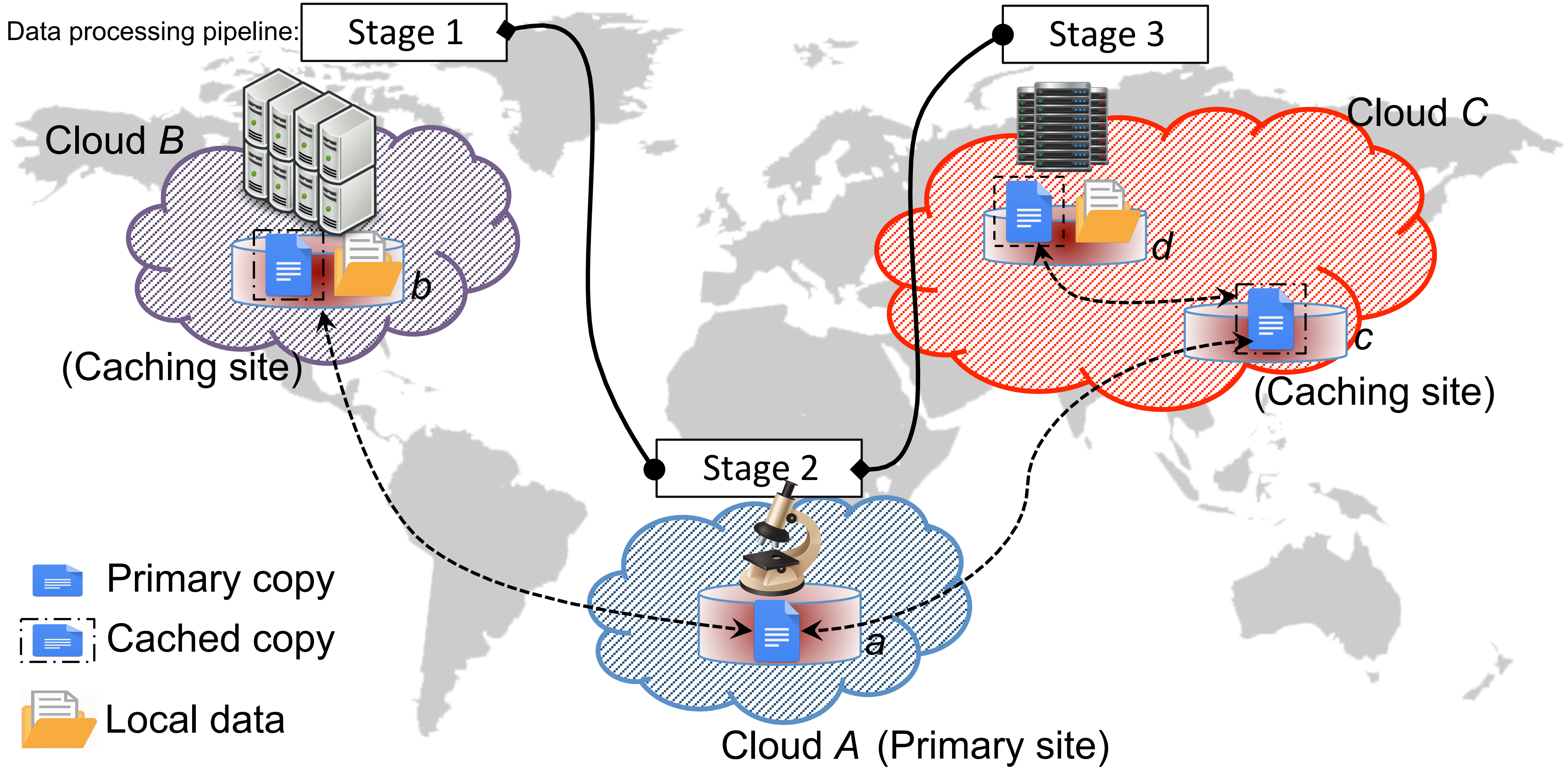


MeDiCI goes Cloudy

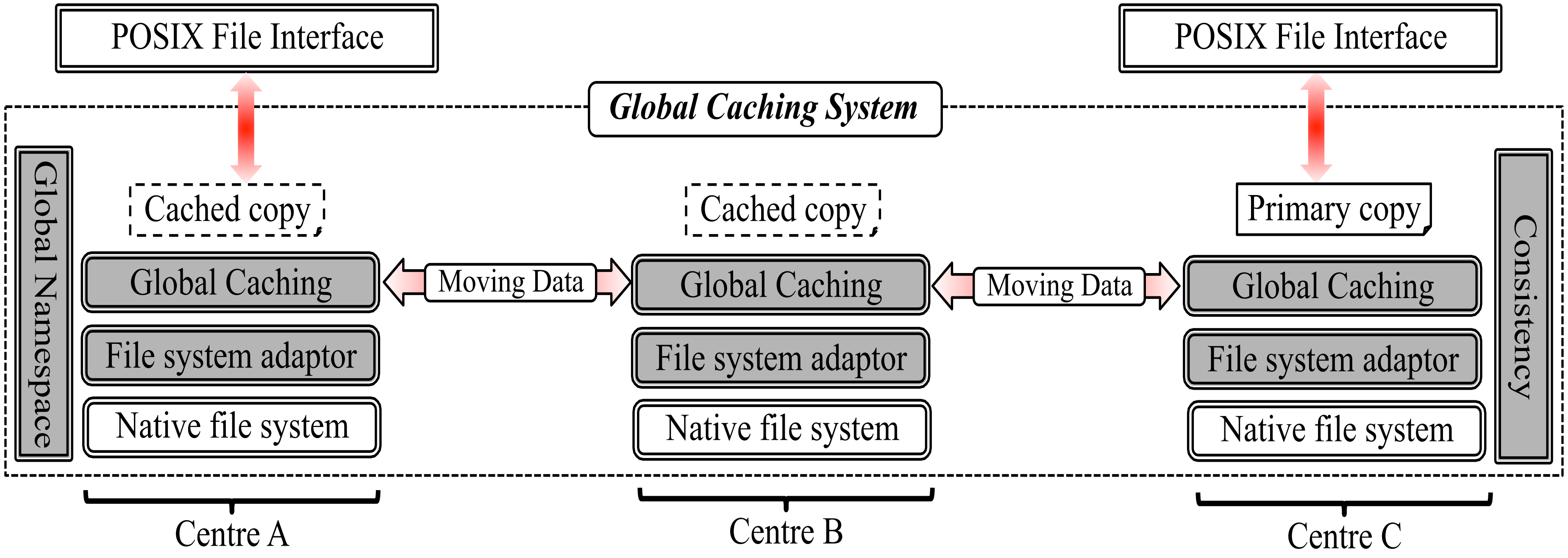


THE UNIVERSITY OF QUEENSLAND AUSTRALIA

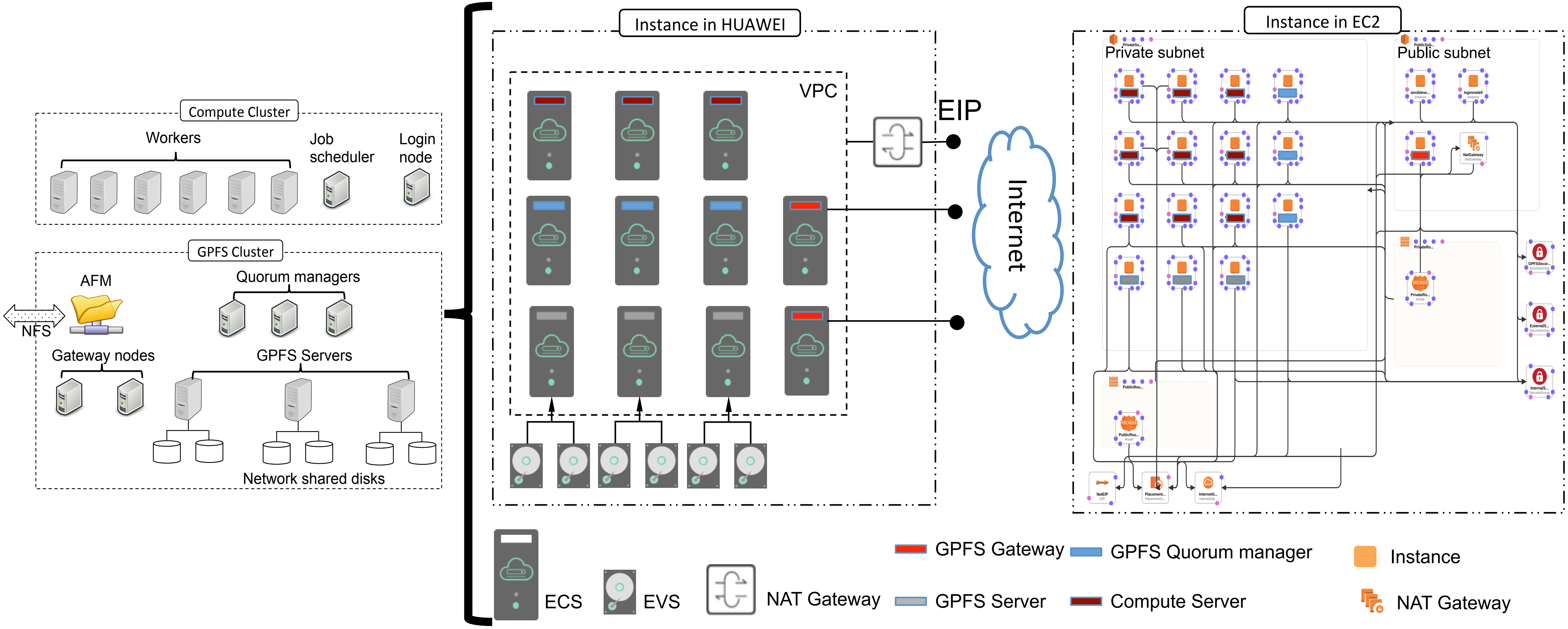
Across multiple cloud centres



Major components in the global caching architecture



Deployment in Amazon and Huawei clouds



Cloud parameters

Resources	Amazon EC2	HUAWEI Cloud	OpenStack
Virtual machine	Instance	Elastic Compute Server	Nova Instance
OS Images	AMI	Glance	Glance
Block Storage	EBS	Elastic Volume Service	Cinder
Private Network	VPC	VPC	Neutron network
Public IP	Public IP	Elastic IP	Floating IP
AAA	SSH key pairs	SSH key pairs	SSH key pairs

An experimental to measure performance

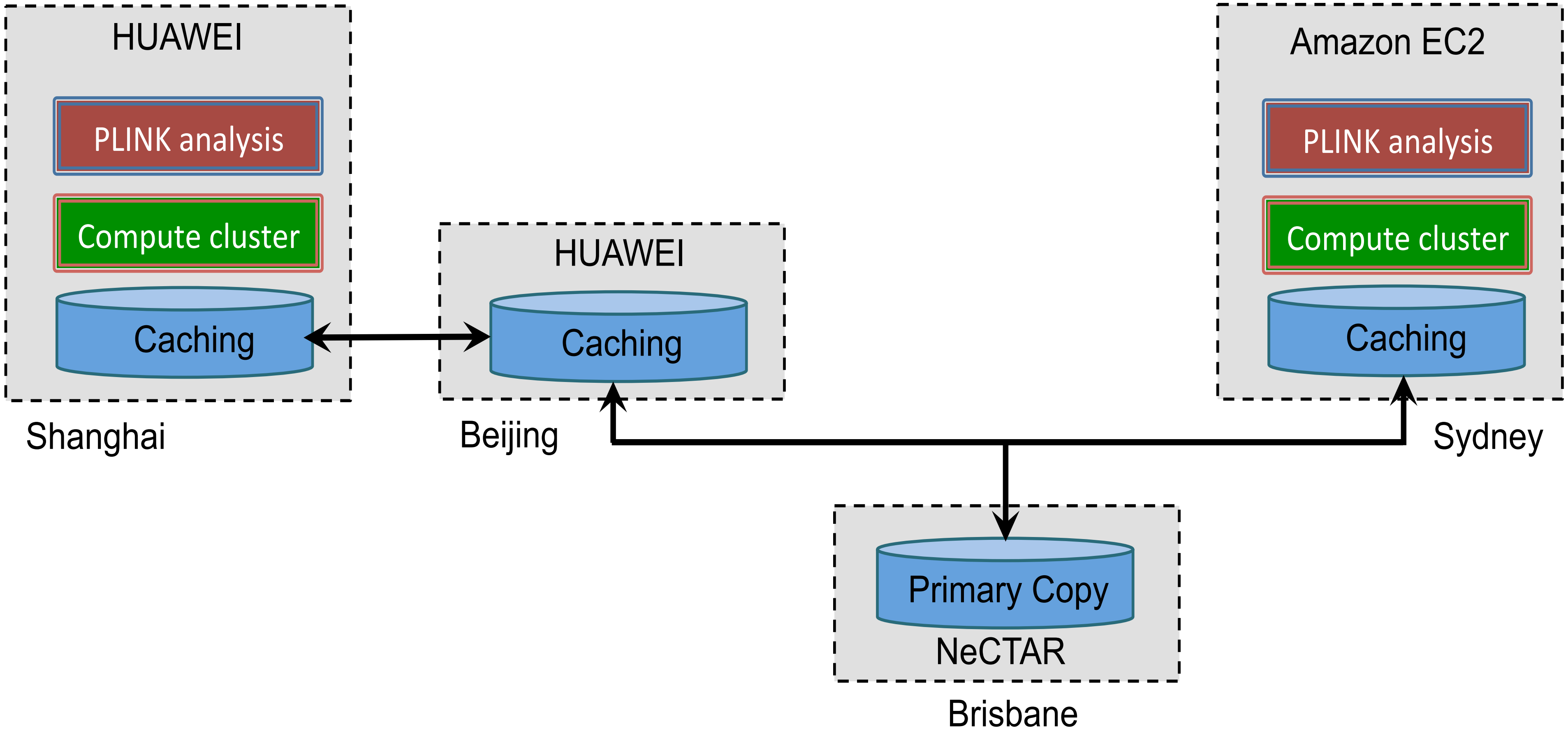
Genome Wide Association Study (GWAS)

- Hypothesis-free methods to identify associations between regions of the genome and complex traits and disease.
- Data from the Systems Genomics of Parkinson's Disease consortium, which has collected DNA methylation data on about 2,000 individuals.
- Test how genetic variation alters DNA methylation, an epigenetic modification that controls how genes are expressed
- Results are being used to understand the biological pathways through which genetic variation affects disease risk.

An experimental to measure performance …

- 3.3 x 10^{12} statistical tests using the PLINK software
- Embarrassingly (pleasingly) parallel
- Does not require high performance communication across virtual machines within the cloud.
- Data to be analyzed, around 40GB in total, is stored in NeCTAR’s data collection storage site located in the campus of the University of Queensland (UQ) at Brisbane.
- The input data is moved to the virtualized clusters, acquired in Amazon EC2 and HUAWEI Cloud, as requested.
- Can control the size of the cloud resource, for both the compute and GPFS clusters, according to our testing requirements

Experiment Setup



AWS Instance types

Type of nodes	Instances	Count	Details
Nimrod worker	c5.9xlarge	25	750 Xeon Skylake cores in total.
AFM Gateway	i3.16xlarge	2	Each instance is equipped with
GPFS Quorum	i3.16xlarge	3	25 Gbit/sec network bandwidth
GPFS Server	i3.16xlarge	10	and 8 x 1.9TB NVME.

Performance evaluation

- During the 3 days of experiment, system utilization was on average about 85~92% on each node,
- I/O peaking at about
  - 420,000 writes/sec and
  - 25,000 reads/sec operations per second (IOPS).
- Total 500,000 tasks were launched in 5 batches sequentially.
  - Allowed us to optimize the system configuration while monitoring the progress of computing and expense used
  - The system was tuned in the first batch.
  - We only present the performance statistics for the last 4 batches.
  - We used the EC2 CloudWatch tools to monitor the performance. In particular, we captured CPU utilization, network traffic and IOPS for each instance
  - Overall, approximately 60 TBs of data were generated by the experiment and sent back to Brisbane for long-term storage and post-processing

Performance evaluation ..

- Although each PLINK task consists of similar computational complexity with almost same size of input data, we observed significant performance variation,
- Averaged execution time is 200 seconds with a long tail of outliers, and some special cases could take up to 1,000 seconds.
- Commonly, performance variability exists in a large scale of distributed system.
- Shared resources and system and network instability can lead to huge performance variation
- For our case, we observed significant variations of IO access for PLINK tasks

Performance evaluation …

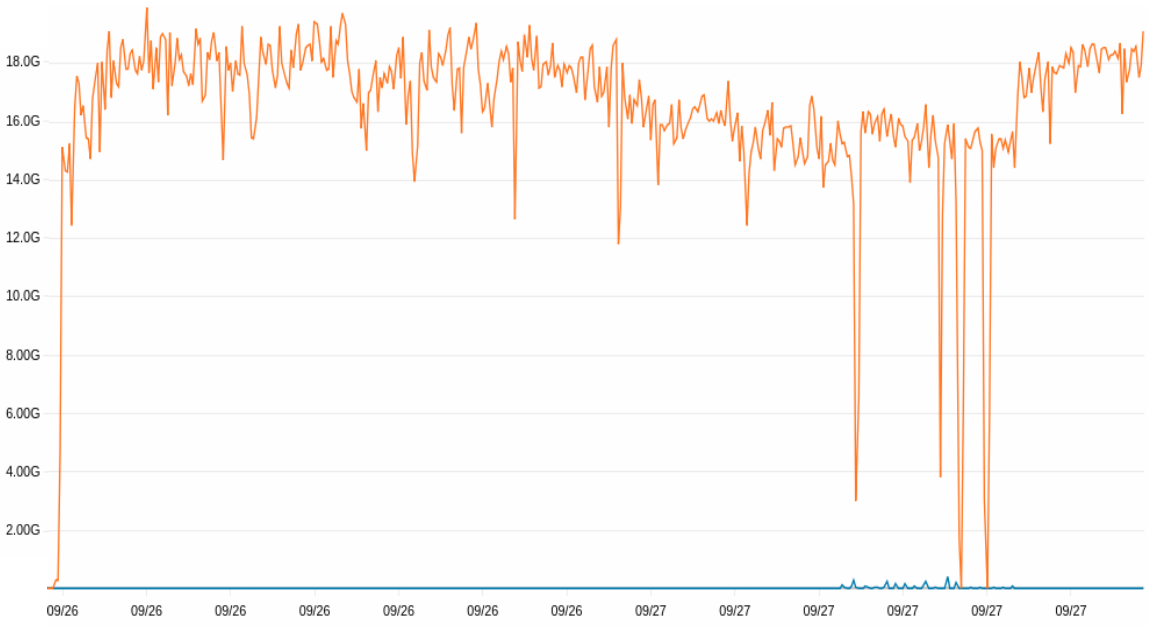
- Because of the PLINK workload, write IO is an order of magnitude higher than read performance.
- The metrics of different instances are correlated very well and it means the workload on each instance is pretty similar.
- Write performance was comparatively stable within the range of 200K and 400K.
  - the updates were first committed to local NVMe devices before being transferred to the home site through AFM gateway.
- In comparison, averaged read operations changes from around 22K to less 15K. This may be caused by unreliable long-haul network.

Disk read operations per second.

Disk write operations per second

Performance evaluation …

- Acceptable use of wide area networks (peaking at 18 Gbps)
  - Multiple AFM worker threads
- There are significant drops in the last day of experiment.
- We assume they were caused by shared bandwidth competition from other public users.
- This resource contention also impacts the PLINK execution time at the last day, especially the performance of read IO.



Outbound network traffic of AFM gateway nodes

Conclusions

- Able to spin up dedicated clusters for experiments that would otherwise execute on UQ resources
  - Can use same infrastructure to burst to commercial cloud
- Applications see their data collections back in UQ data centre
- Prototyped over three different clouds
- Existing storage software, including GPFS, AFM, and NFS.
- Cloud nodes efficient for parameter studies
- 500,000 GWAS tasks executed. Took 3 days on 25 worker nodes
- Economics are “cloudy”
  - CAPEX vs OPEX
  - Relatively poor network performance in commercial clouds can be compensated by faster storage, but this is expensive

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