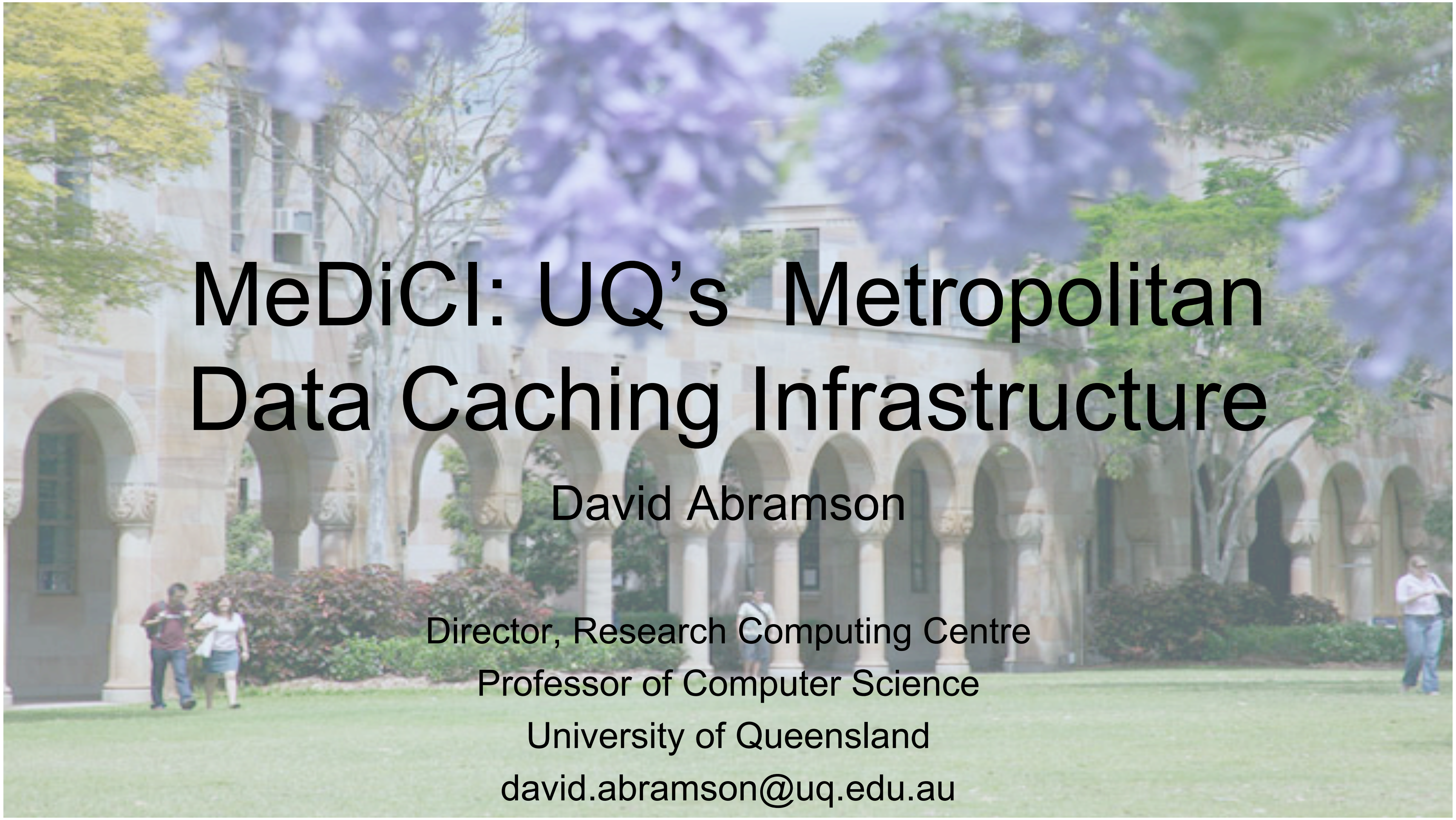


MeDiCI: UQ's Metropolitan Data Caching Infrastructure

David Abramson

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

~~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.

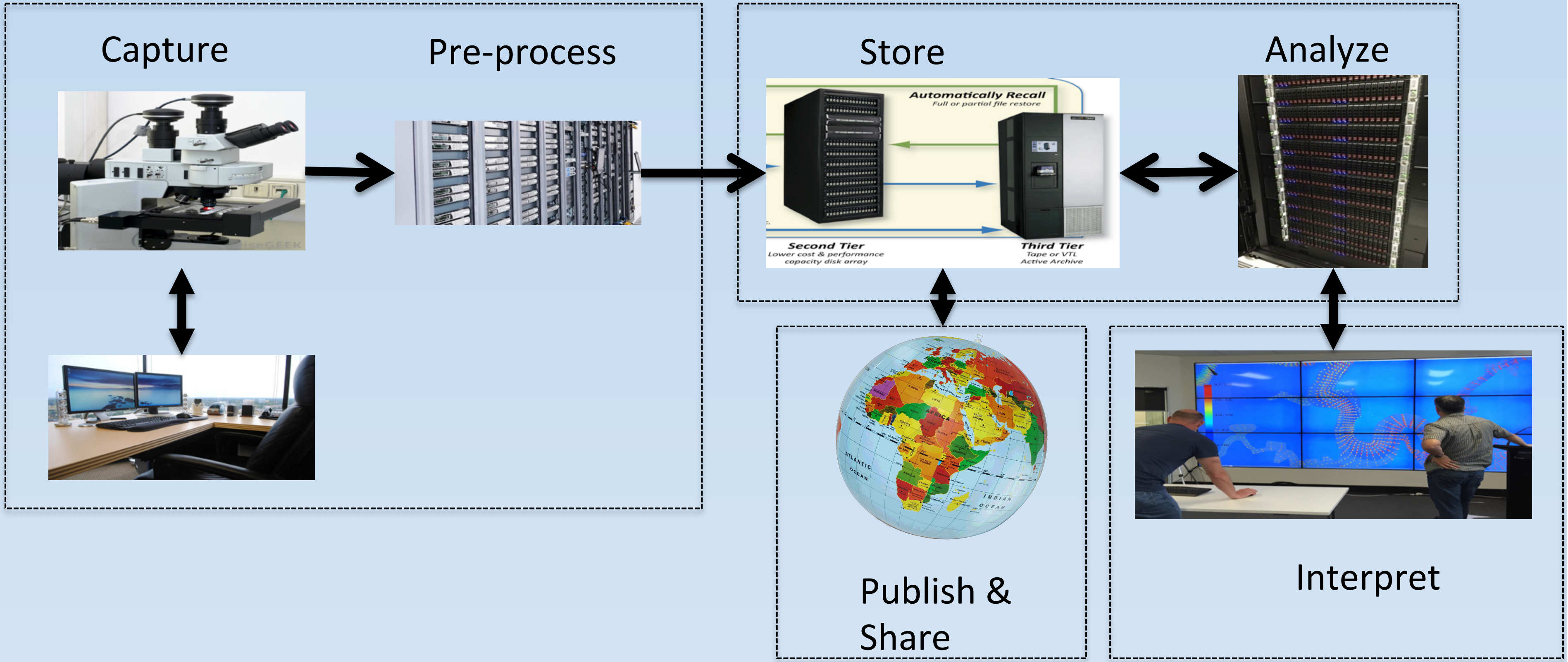
Questioning what the final turtle might be standing on, the anecdote humorously concludes that it is turtles all the way down””

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

Why do we need to do anything special?

Data Intensive Computing

Use Case: Microscopy



Capture

Pre-process

Store

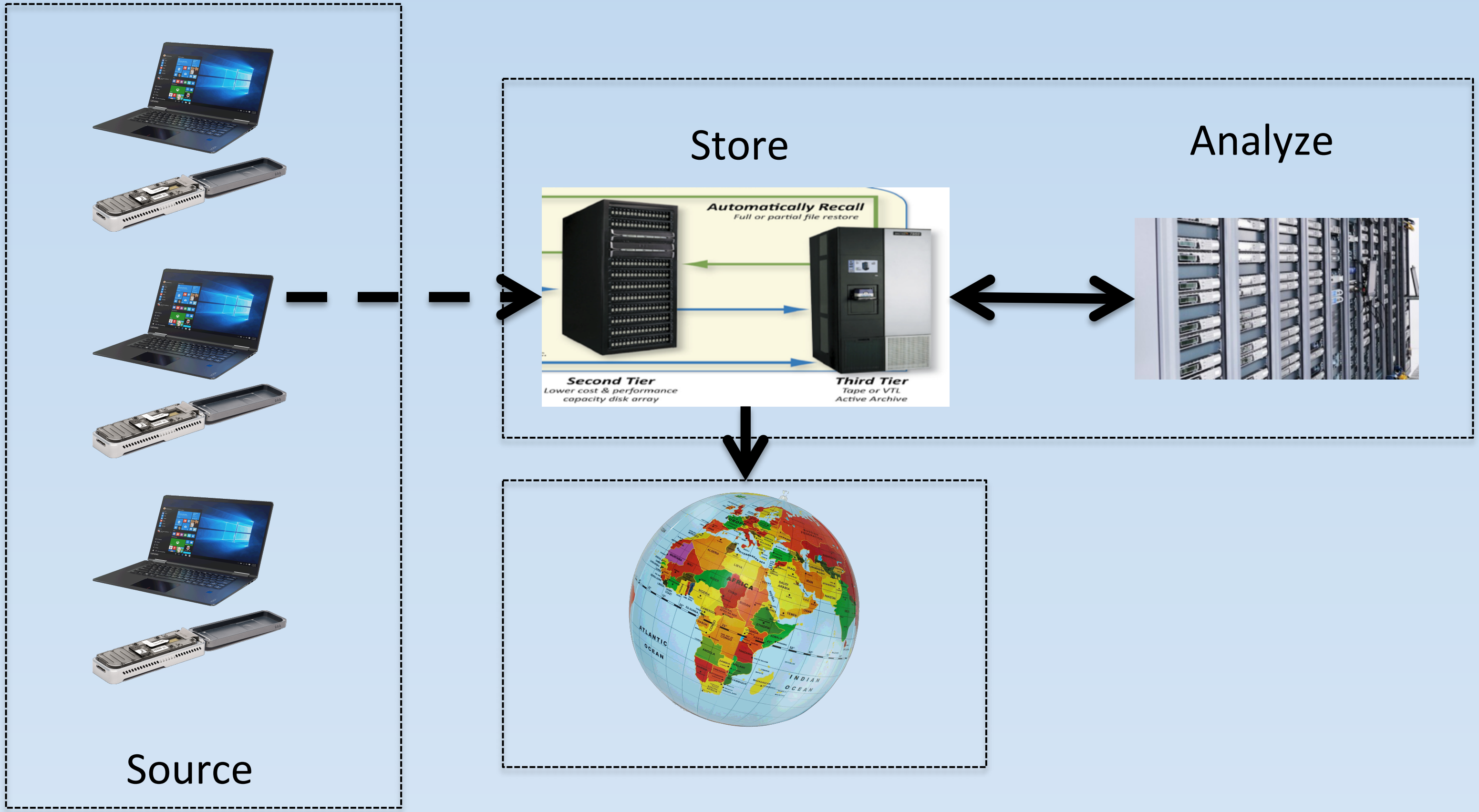
Analyze

Publish & Share

Interpret

Automatically Recall
Full or partial file restore
Second Tier
Lower cost & performance capacity disk array
Third Tier
Tape or VTL Active Archive
ATLANTIC OCEAN
INDIAN OCEAN

Use Case: Personal Genomics



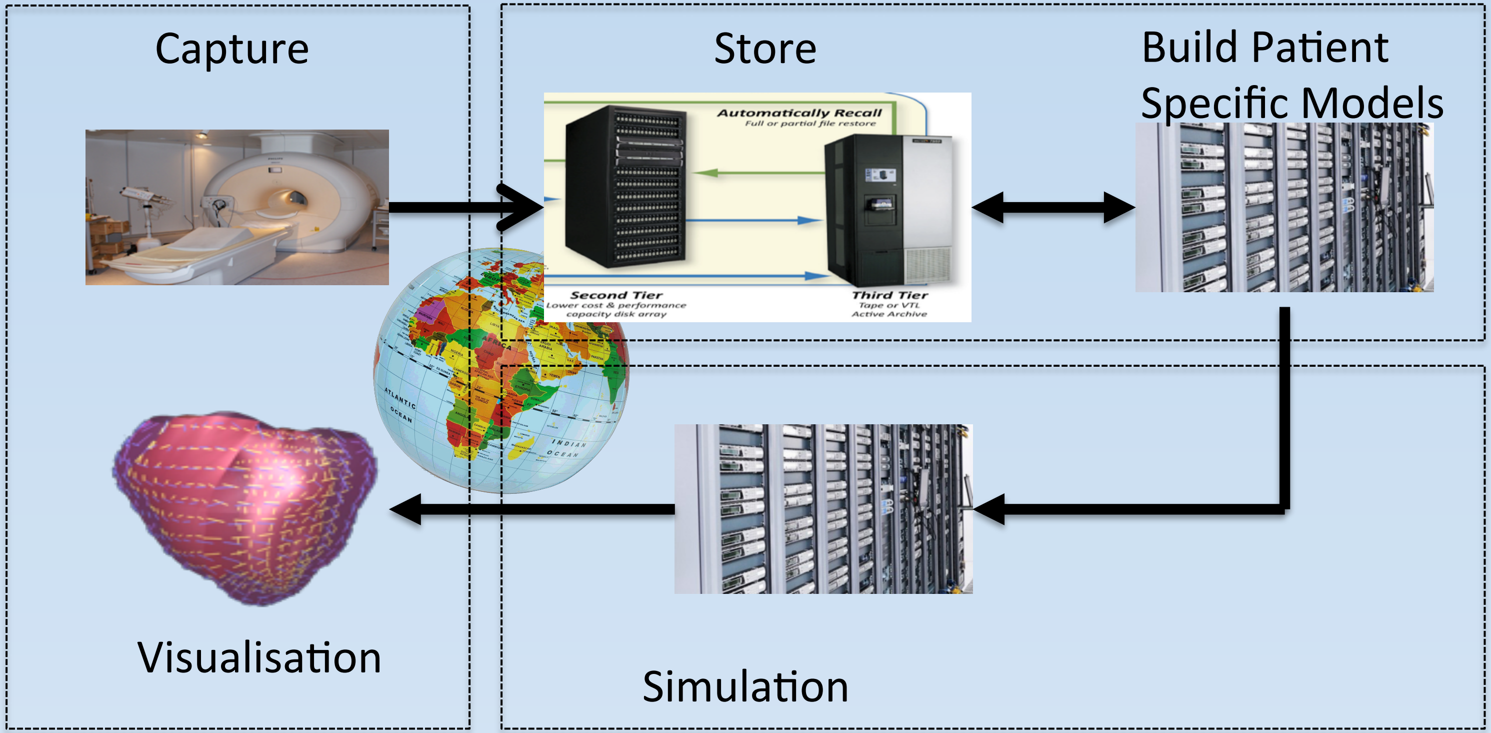
Store

Analyze

Source

Automatically Recall
Full or partial file restore
Second Tier
Lower cost & performance capacity disk array
Third Tier
Tape or VTL Active Archive
ATLANTIC OCEAN
INDIAN OCEAN

Use Case: Cardiac Science



Capture

Store

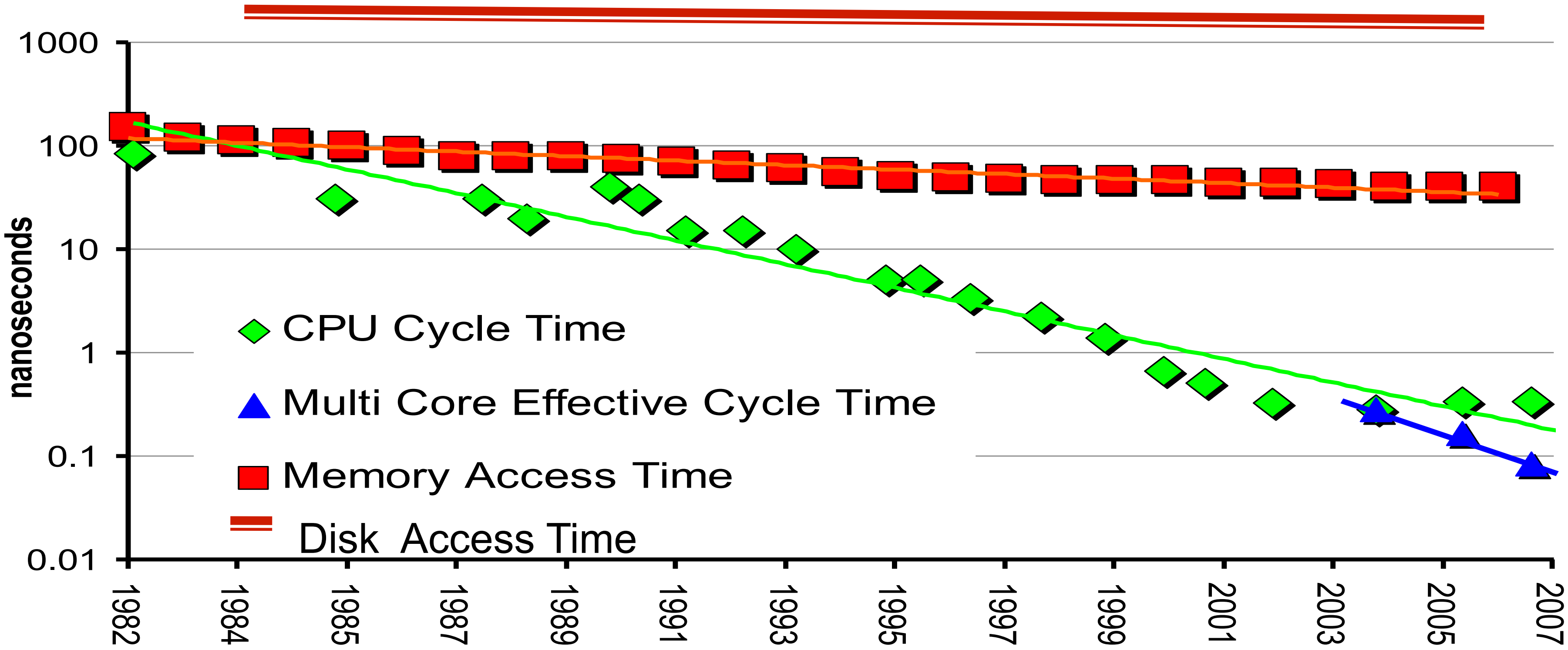
Build Patient Specific Models

Visualisation

Simulation

Infrastructure Challenges of Big Data

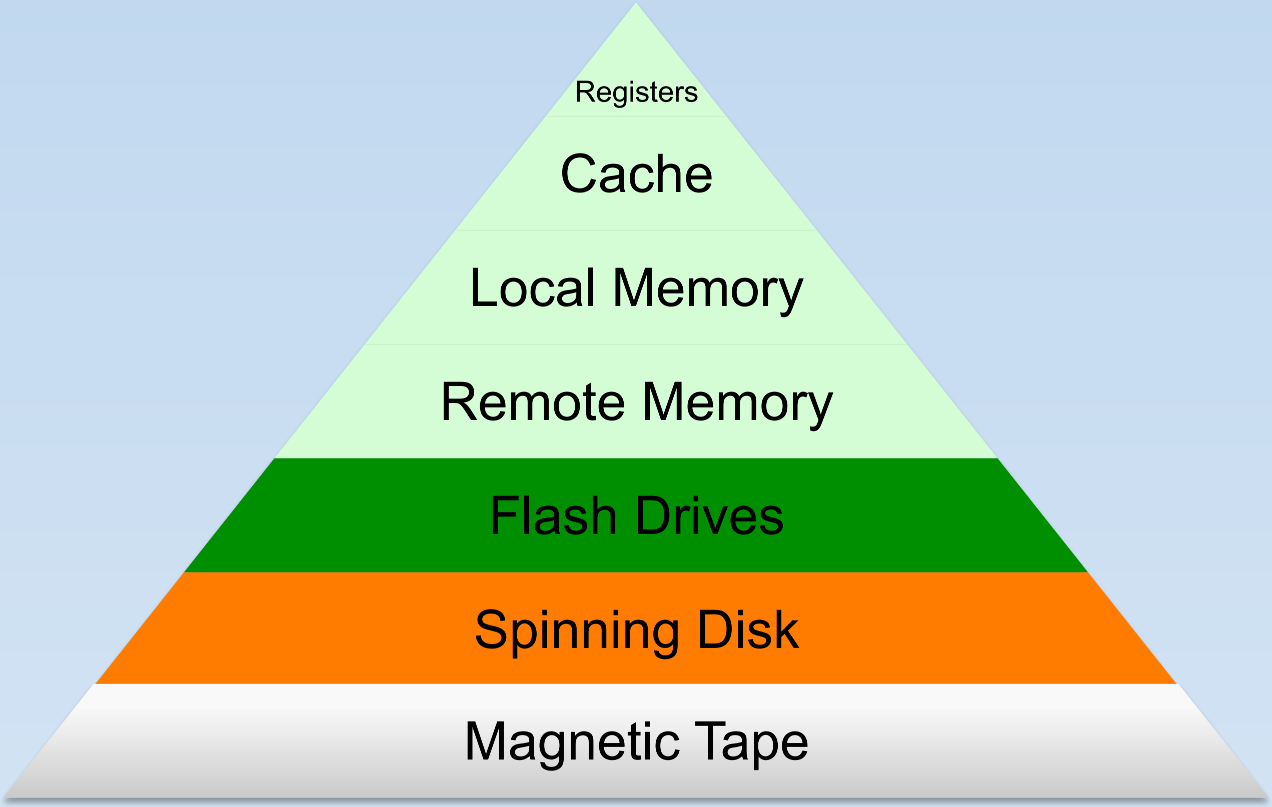
Red Shift: Data keeps moving further away from the CPU with every turn of Moore's Law



data due to Dean Klein of Micron

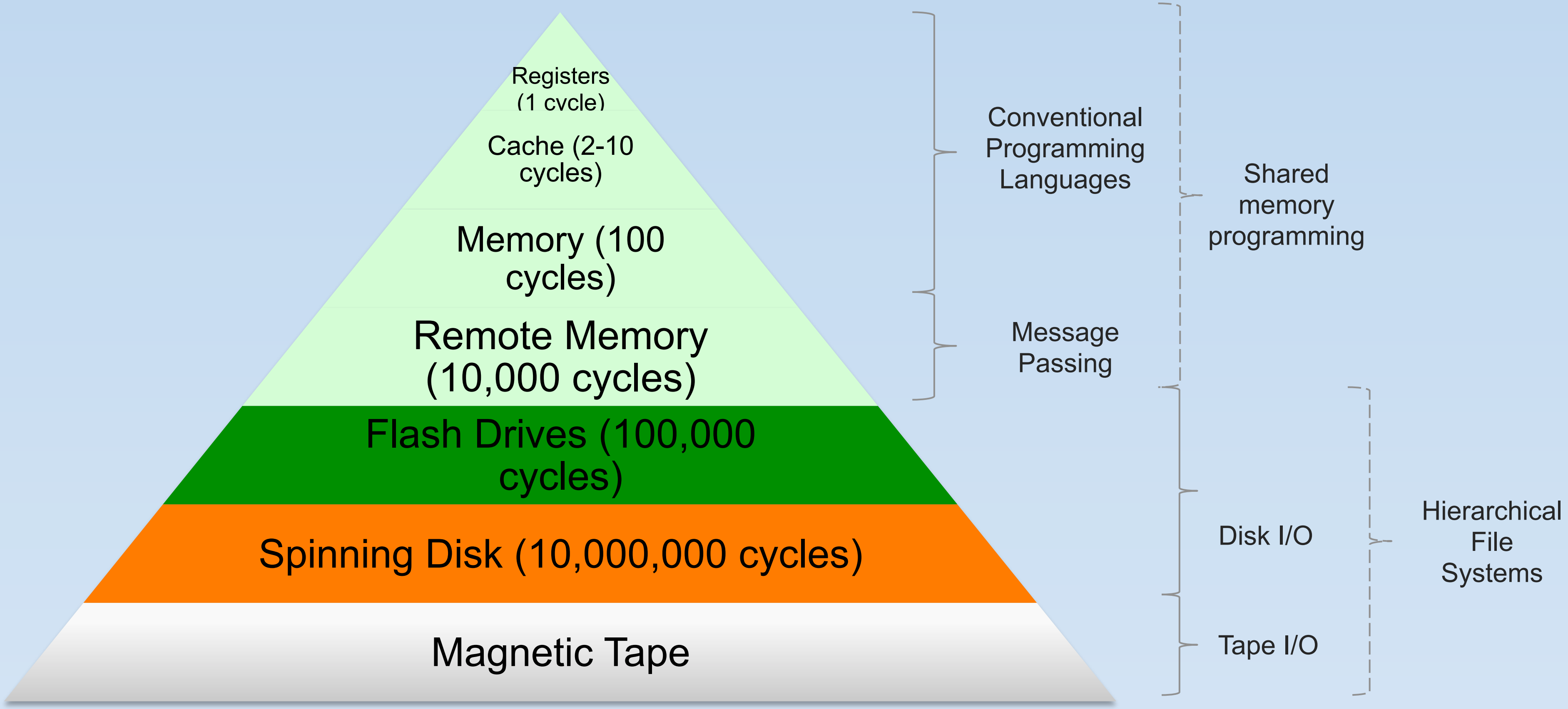
Slide courtesy Mike Norman, SDSC

It’s always been caches all the way down



Explicit vs Implicit management

Memory Hierarchy



Registers (1 cycle)

Cache (2-10 cycles)

Memory (100 cycles)

Remote Memory (10,000 cycles)

Flash Drives (100,000 cycles)

Spinning Disk (10,000,000 cycles)

Magnetic Tape

Conventional Programming Languages

Message Passing

Shared memory programming

Disk I/O

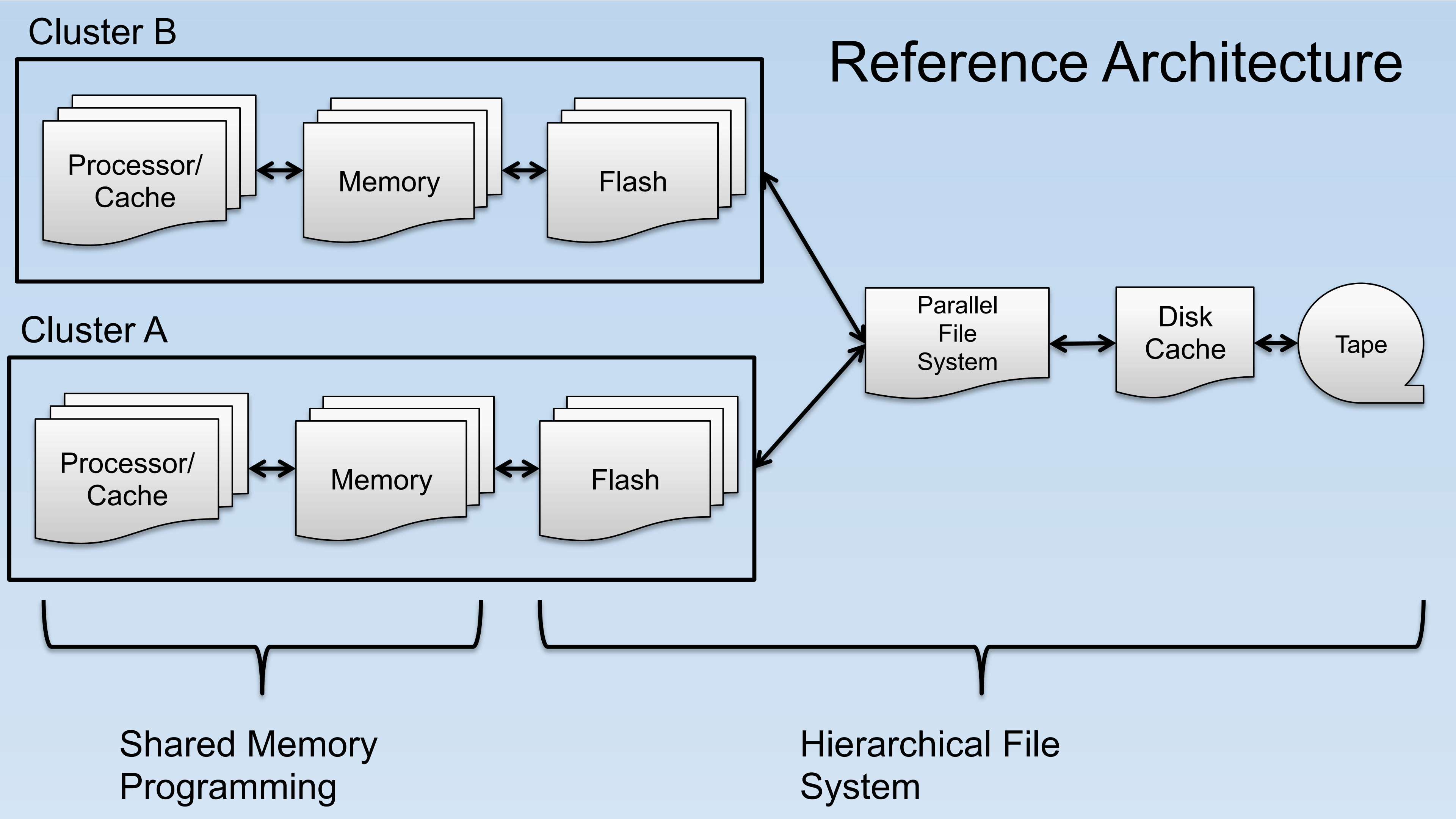
Tape I/O

Hierarchical File Systems

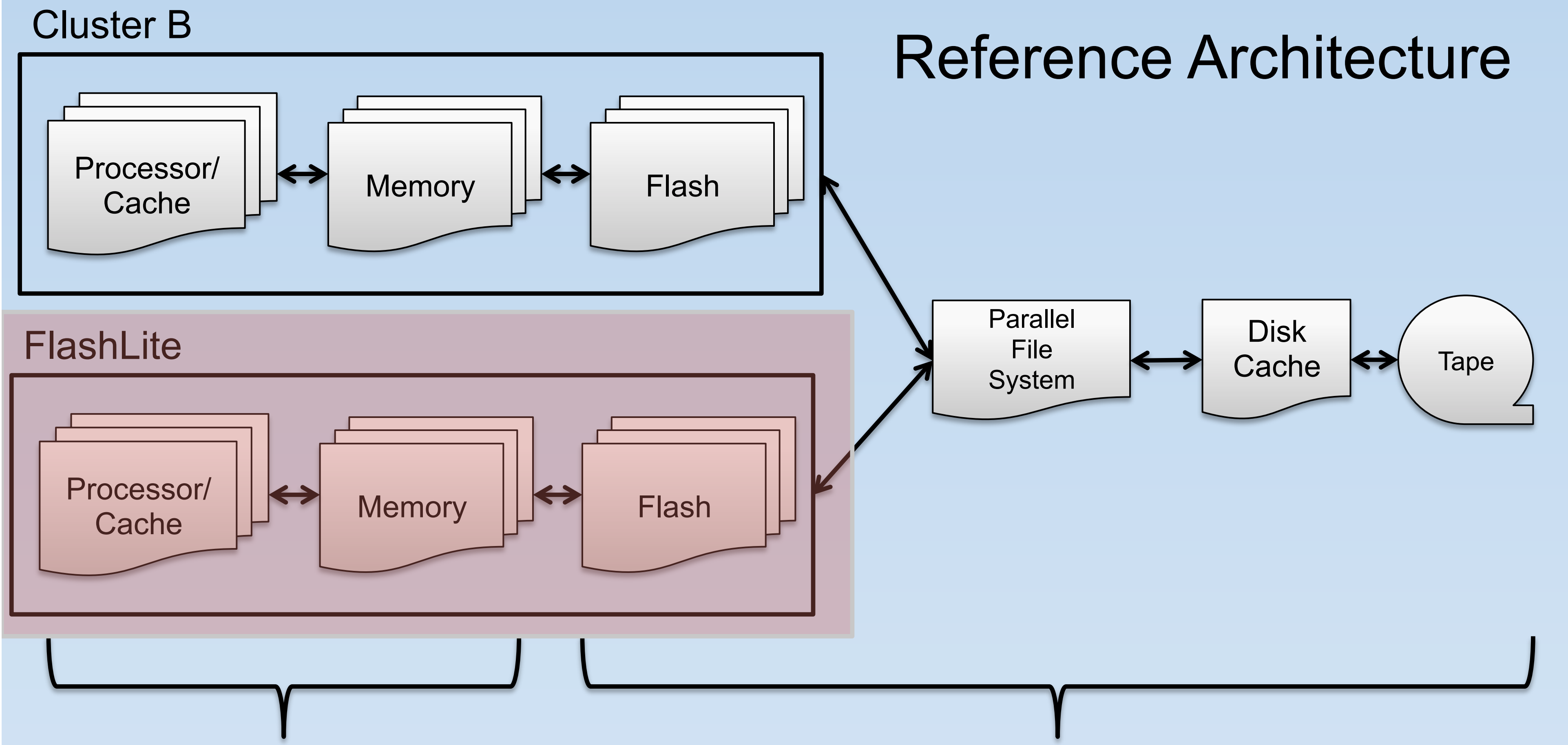
Infrastructure for Data Intensive Computing

- 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™


Cluster B
Reference Architecture
Processor/ Cache
Memory
Flash
Cluster A
Processor/ Cache
Memory
Flash
Parallel File System
Disk Cache
Tape
Shared Memory Programming
Hierarchical File System

Reference Architecture



Cluster B

Processor/ Cache

Memory

Flash

FlashLite

Processor/ Cache

Memory

Flash

Parallel File System

Disk Cache

Tape

Shared Memory Programming

Hierarchical File System

Data Intensive Computation Engine

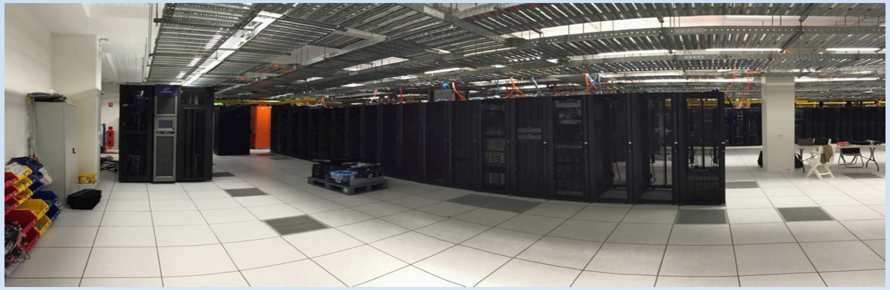
- Parallel
  - High performance network
  - Good numeric performance
- Massive memory
  - Ability to hold whole data sets or data bases in memory
- High IO throughput

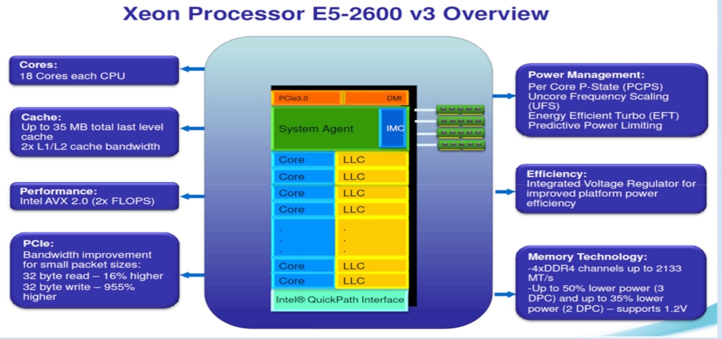
FlashLite

- High throughput solid state disk
- Large amounts of main memory
- Software shared memory
- Inspired by SDSC Gordon

What is FlashLite?

- FlashLite
  - ~ 70 compute nodes (~1600 cores)
    - Dual socket Intel E5-2680v3 2.5GHz (Haswell)
    - 512 GB DDR-2
    - 4.8 TB NVMe SSD
  - ScaleMP vSMP virtual shared memory
    - 4TB RAM aggregate(s)





Xeon Processor E5-2600 v3 Overview

Cores:
18 Cores each CPU

Cache:
Up to 35 MB total last level cache
2x L1/L2 cache bandwidth

Performance:
Intel AVX 2.0 (2x FLOPS)

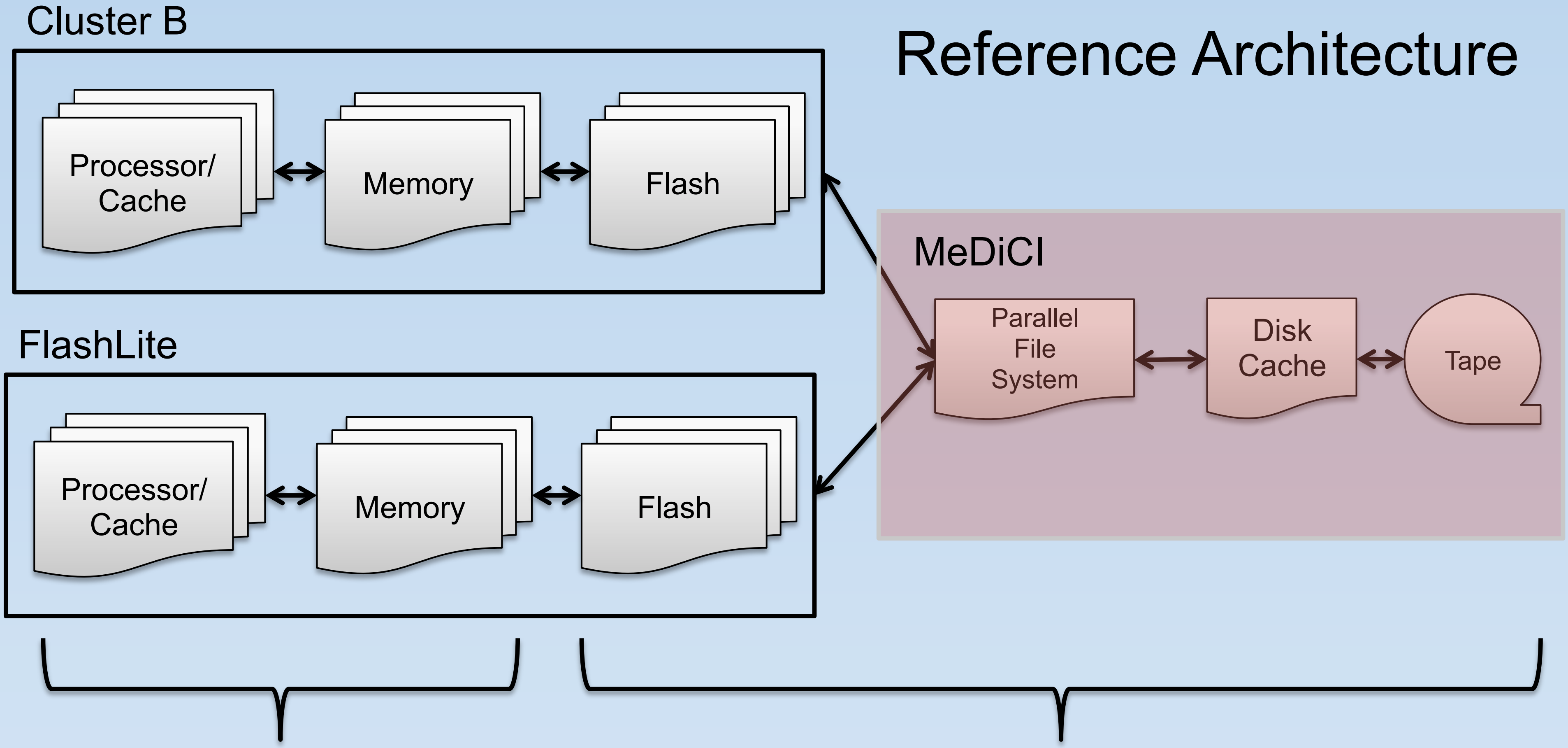
PCIe:
Bandwidth improvement for small packet sizes:
32 byte read – 16% higher
32 byte write – 955% higher

Power Management:
Per Core P-State (PCPS)
Uncore Frequency Scaling (UFS)
Energy Efficient Turbo (EFT)
Predictive Power Limiting

Efficiency:
Integrated Voltage Regulator for improved platform power efficiency

Memory Technology:
-4xDDR4 channels up to 2133 MT/s
-Up to 50% lower power (3 DPC) and up to 35% lower power (2 DPC) – supports 1.2V

Reference Architecture



Cluster B

Processor/ Cache ↔ Memory ↔ Flash

FlashLite

Processor/ Cache ↔ Memory ↔ Flash

MeDiCI

Parallel File System ↔ Disk Cache ↔ Tape

Shared Memory Programming

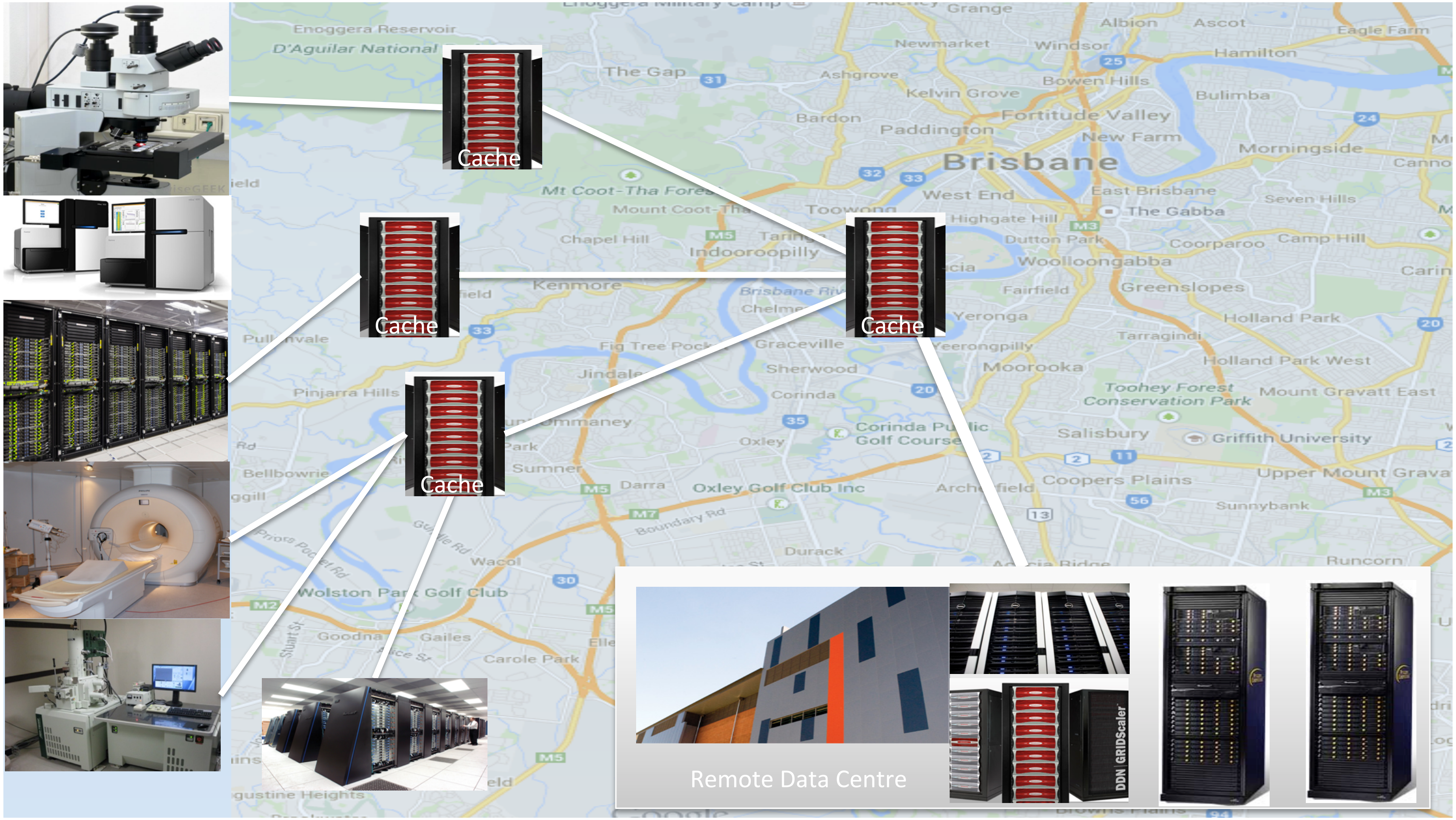
Hierarchical File System

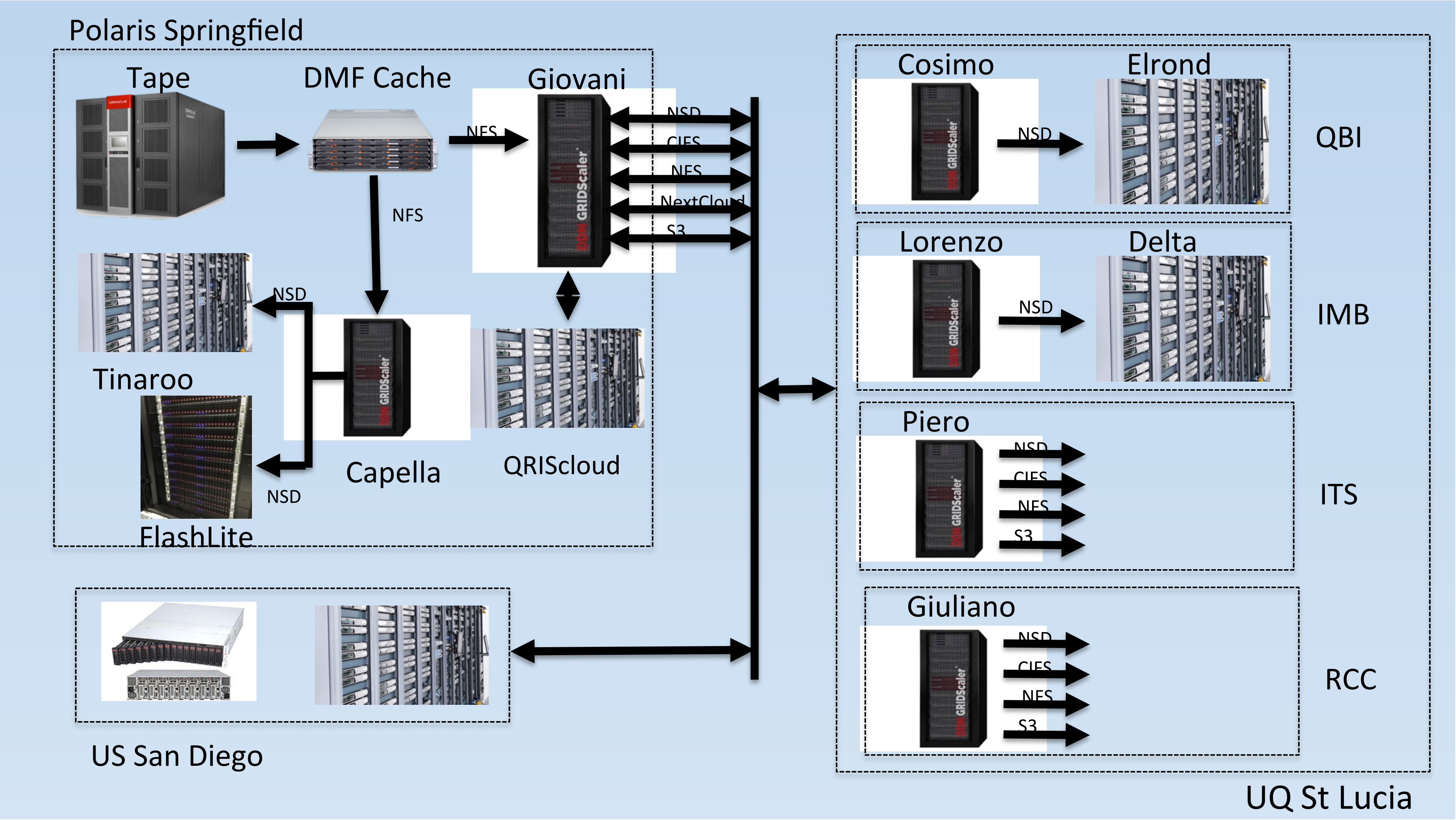
The caches continue …

MeDiCI

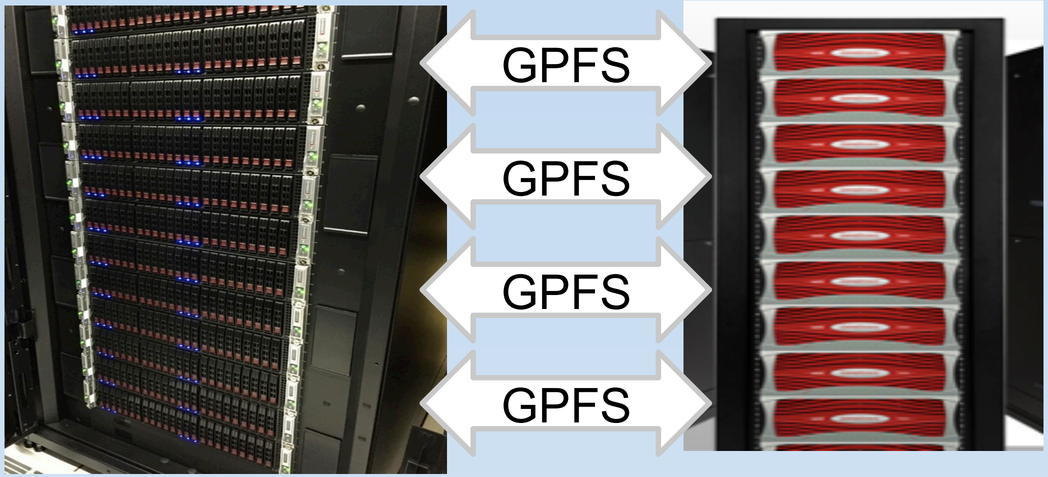
MeDiCI

- 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)


Cache
Cache
Cache
Cache
Remote Data Centre
DDN GRIDScaler
Brisbane
The Gap
Kenmore
Indooroopilly
Toowong
Fortitude Valley
Paddington
West End
Woolloongabba
Coorparoo
Greenslopes
Holland Park
Moorooka
Salisbury
Griffith University
Coopers Plains
Sunnybank
Jindalee
Darra
Oxley
Sherwood
Graceville
Wacol
Goodna
Gailes
Carole Park
Pullenvale
Pinjarra Hills
Bellbowrie
Chapel Hill
Fig Tree Pocket
Mt Coot-Tha Forest
Enoggera Reservoir
D'Aguilar National
Bardon
Ashgrove
Kelvin Grove
Newmarket
Windsor
Albion
Ascot
Hamilton
Eagle Farm
Bulimba
Morningside
Seven Hills
Camp Hill
New Farm
Bowen Hills
East Brisbane
The Gabba
Dutton Park
Highgate Hill
Fairfield
Yeronga
Yeerongpilly
Tarragindi
Holland Park West
Mount Gravatt East
Upper Mount Gravatt
Toohey Forest Conservation Park
Corinda
Corinda Public Golf Course
Oxley Golf Club Inc
Archerfield
Durack
Runcorn
Sumner
Wolston Park Golf Club
Boundary Rd
Grange


Polaris Springfield
Tape
DMF Cache
Giovani
NFS
NFS
NSD
CIFS
NFS
NextCloud
S3
Tinaroo
NSD
NSD
FlashLite
Capella
QRIScloud
US San Diego
Cosimo
Elrond
NSD
QBI
Lorenzo
Delta
NSD
IMB
Piero
NSD
CIFS
NFS
S3
ITS
Giuliano
NSD
CIFS
NFS
S3
RCC
UQ St Lucia

FlashLite in the Data Centre



GPFS

GPFS

GPFS

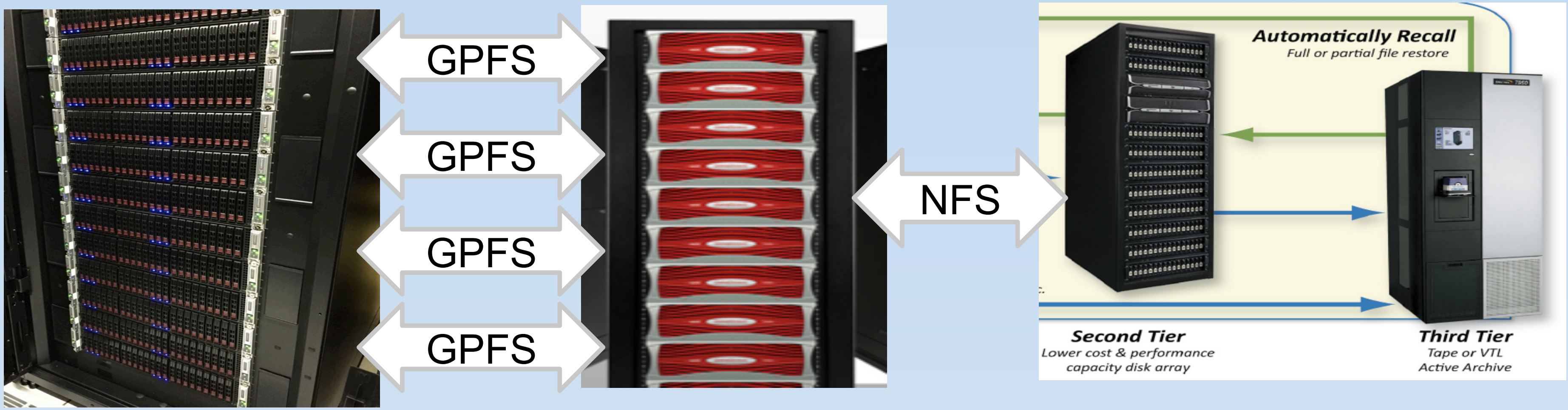
GPFS

DDN SFA12KXE

FlashLite

Parallel file system

FlashLite in the Data Centre



GPFS

GPFS

GPFS

GPFS

NFS

DDN SFA12KXE

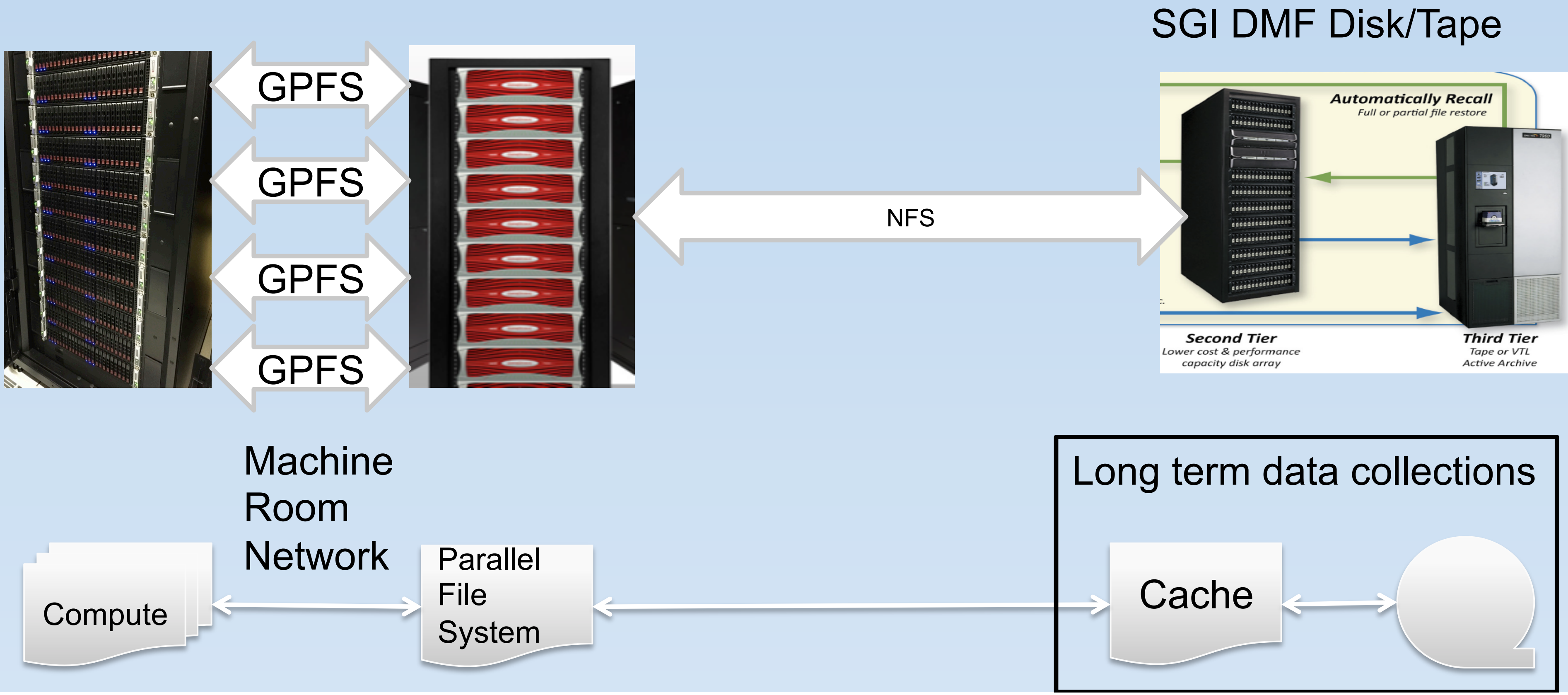
SGI DMF Disk/Tape

FlashLite

Parallel file system

Long term data collections

MeDiCI Wide Area Architecture



SGI DMF Disk/Tape

GPFS

GPFS

GPFS

GPFS

NFS

Automatically Recall
Full or partial file restore
Second Tier
Lower cost & performance capacity disk array
Third Tier
Tape or VTL
Active Archive

Machine Room Network

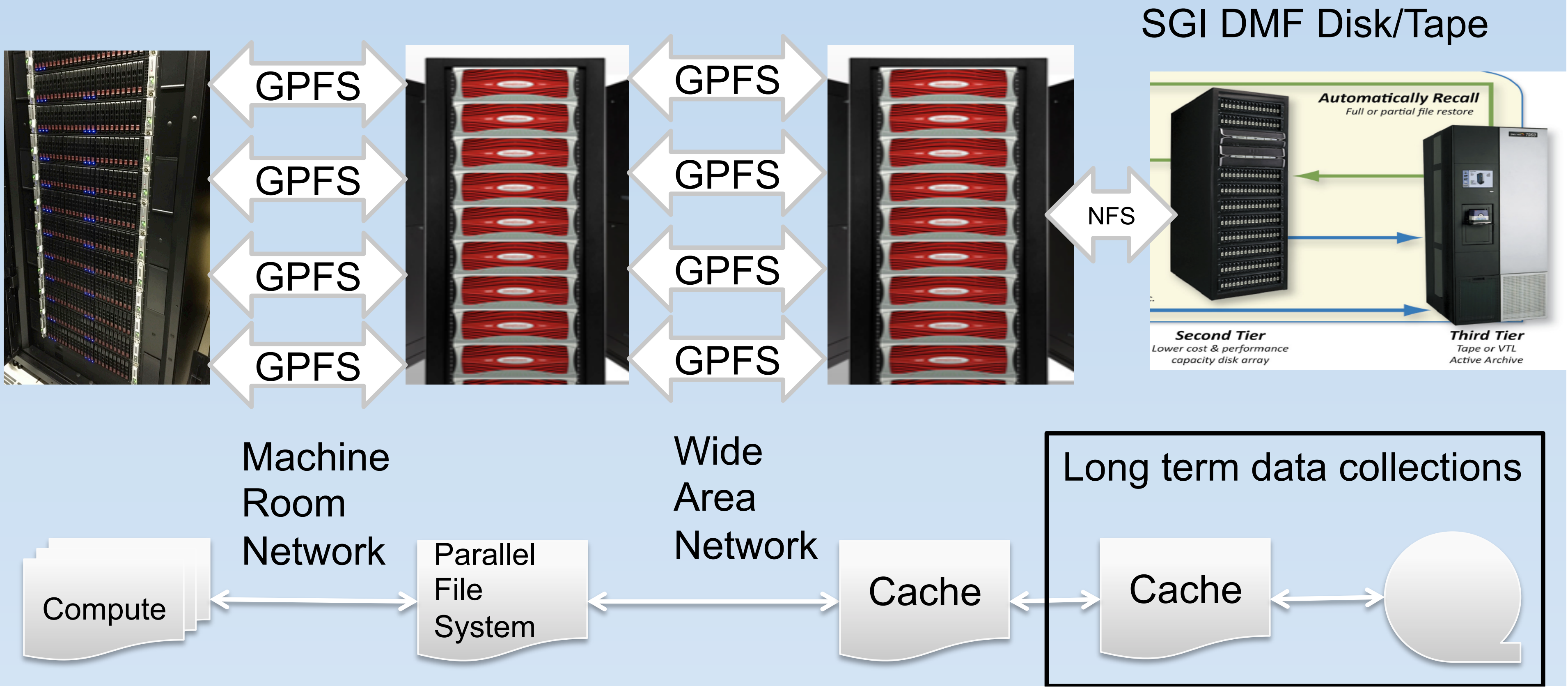
Compute

Parallel File System

Long term data collections

Cache

MeDiCI Wide Area Architecture



SGI DMF Disk/Tape

GPFS

GPFS

GPFS

GPFS

GPFS

GPFS

GPFS

GPFS

NFS

Automatically Recall
Full or partial file restore
Second Tier
Lower cost & performance capacity disk array
Third Tier
Tape or VTL Active Archive

Machine Room Network

Wide Area Network

Long term data collections

Compute

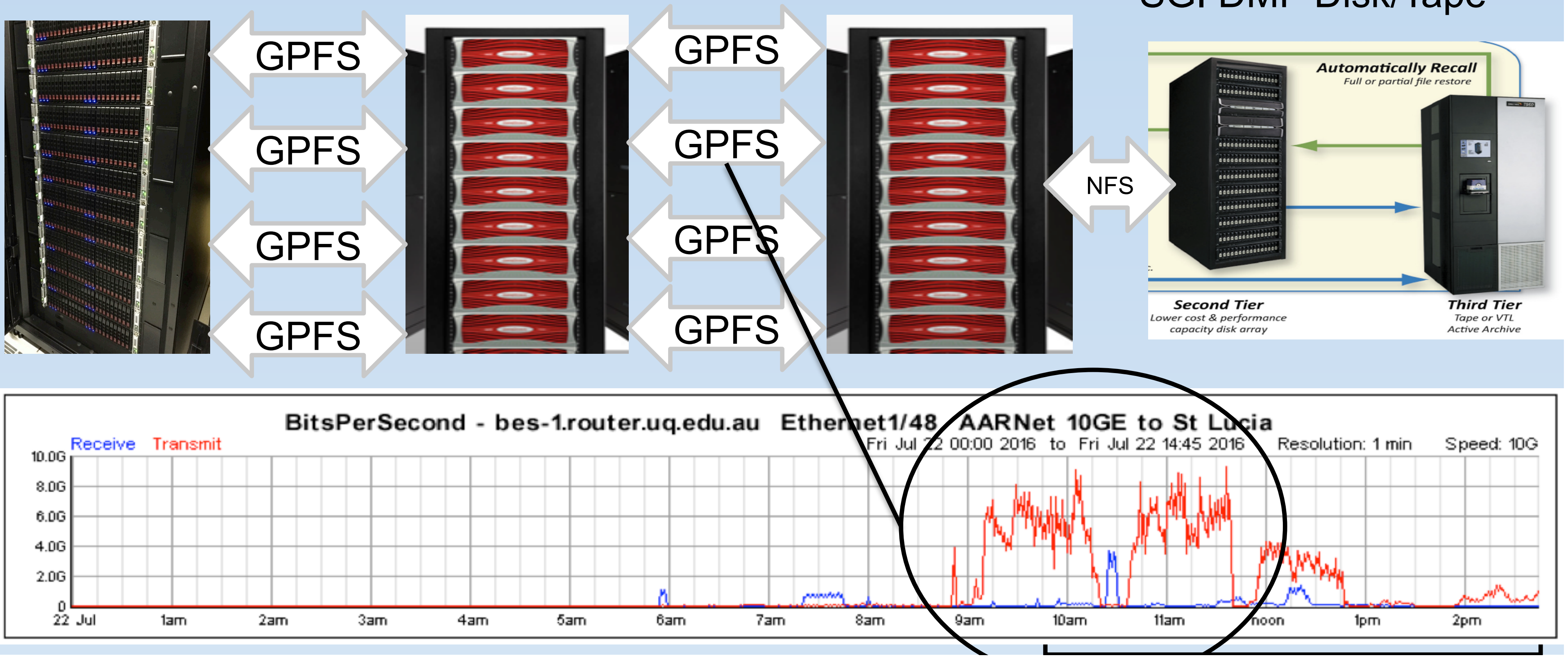
Parallel File System

Cache

Cache

MeDiCI Wide Area Architecture

SGI DMF Disk/Tape



GPFS

GPFS

GPFS

GPFS

GPFS

GPFS

GPFS

GPFS

NFS

Automatically Recall
Full or partial file restore
Second Tier
Lower cost & performance capacity disk array
Third Tier
Tape or VTL
Active Archive

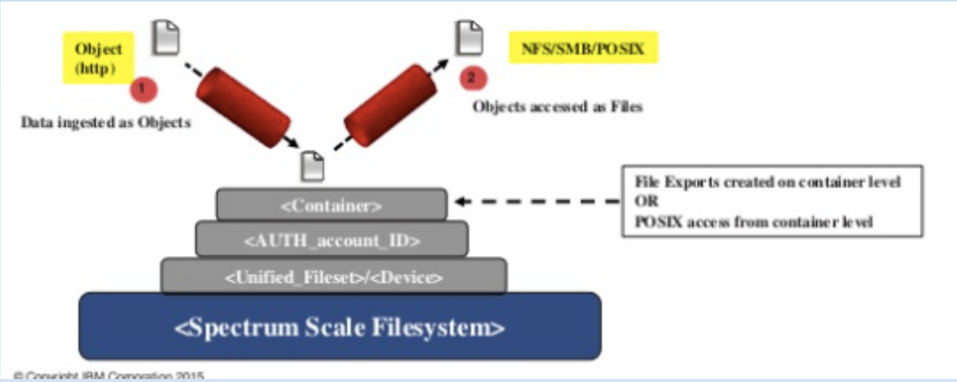
BitsPerSecond - bes-1.router.uq.edu.au Ethernet1/48 AARNet 10GE to St Lucia

Identity!

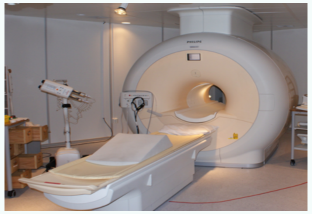
- No single UID space across UQ/QCIF users
- Need to map UID space between UQ and Polaris
- GPFS 4.2
  - mmname2uid/mmuid2name

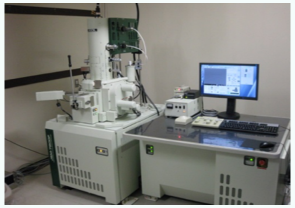
Object Storage

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


Object (http)
1
Data ingested as Objects
NFS/SMB/POSIX
2
Objects accessed as Files
File Exports created on container level
OR
POSIX access from container level
<Container>
<AUTH_account_ID>
<Unified_Fileset>/<Device>
<Spectrum Scale Filesystem>
© Copyright IBM Corporation 2015

Data Data everywhere anytime



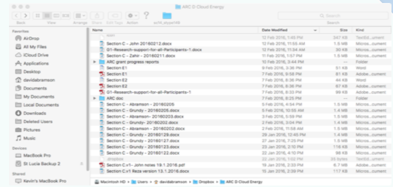


ImageTrove

myTardis

OMERO

Managed Data



Name	Date Modified	Size	Kind
[illegible]	12 Feb 2016, 1:45 PM	347 KB	TextEd...ument
Section C - John 20160212.docx	12 Feb 2016, 11:55 AM	1.5 MB	Micros...cument
G1-Research-support-for-all-Participants-1.docx	12 Feb 2016, 11:34 AM	30 KB	Micros...cument
Section C - Zahir - 20160211.docx	11 Feb 2016, 1:57 PM	1.5 MB	Micros...cument
ARC grant progress reports	10 Feb 2016, 3:44 PM	--	Folder
Section E1	9 Feb 2016, 3:36 PM	51 KB	Word
Section E1	7 Feb 2016, 9:58 PM	81 KB	Adobe...cument
Section E2	7 Feb 2016, 8:36 PM	44 KB	Word
Section E2	7 Feb 2016, 8:36 PM	67 KB	Adobe...cument
G1-Research-support-for-all-Participants-1	7 Feb 2016, 8:33 PM	99 KB	Adobe...cument
ARC doc	7 Feb 2016, 8:25 PM	--	Folder
Section C - Abramson - 20160205	5 Feb 2016, 4:54 PM	1.5 MB	Micros...cument
Section C - Grundy - 20160205.docx	5 Feb 2016, 10:55 AM	1.4 MB	Micros...cument
Section C - Abramson- 20160203.docx	3 Feb 2016, 5:59 PM	1.5 MB	Micros...cument
Section C - Grundy- 20160202.docx	2 Feb 2016, 3:04 PM	1.4 MB	Micros...cument
Section C - Abramson - 20160202.docx	2 Feb 2016, 11:58 AM	1.5 MB	Micros...cument
Section C - Grundy - 20160129.docx	29 Jan 2016, 12:45 PM	1.4 MB	Micros...cument
Section C - Grundy - 20160127.docx	27 Jan 2016, 7:25 PM	1.5 MB	Micros...cument
Section C - Grundy - 20160123.docx	23 Jan 2016, 2:10 PM	116 KB	Micros...cument
Section C - Grundy - 20160122.docx	22 Jan 2016, 4:10 PM	98 KB	Micros...cument
.dropbox	22 Jan 2016, 1:02 PM	35 bytes	TextEd...ument
Section C.v1- John notes 19.1.2016.pdf	19 Jan 2016, 2:33 PM	6.7 MB	Adobe...cument
Section C.v1 Reza version 13.1.2016.docx	15 Jan 2016, 2:39 PM	117 KB	Micros...cument

Macintosh HD > Users > davidabramson > Dropbox > ARC D Cloud Energy

MeDiCI

ownCloud

Synchronous

Asynchronous

Unmanaged Data

OpenClinica

Clinical Data

S3,
Swift

Cloud
Access

MeDiCI

QRIScloud Compute and Storage Fabric

Building on basic architecture

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

Conclusions

- FlashLite
  - Parallel computer
  - Very large amounts of local memory and Flash disk
  - Still learning what works
  - Need Burst Buffer s/w
- MeDiCI
  - Caches all the way down
  - IBM Spectrum Scale
  - AFM semantics

MeDiCI
UQ'S DATA FABRIC OF CHOICE
The University of Queensland's Metropolitan Data Caching Infrastructure (MeDiCI) spans urban Brisbane and provides seamless access to data regardless of where they are created, manipulated and archived.
CAMPUS COMPUTE & INSTRUMENTS
LOCAL CACHE STORAGE
LONG-TERM STORAGE
sgi
IBM

Acknowledgments

- Australian Research Council (FlashLite)
  - Zhou, Bernhardt, Zhang, Zhu, Tao, Chen, Drinkwater, Tomlinson, Coppel, Gu, Burrage, Griffiths, Turner, Mackey, Du, Mengersen, Edwards
- Queensland Cyber Infrastructure Foundation (QCIF)
- University of Queensland
  - Jake Carroll, Les Elliott, Michael Mallon, Irek Porebski, Yves St-Onge,
- UCSD
  - Ilkay Altintas, Tom DeFanti, John Graham, Mike Norman, Phil Papadopoulos Larry Smarr,
- AARnet
  - Peter Elford