DataDirect™ NETWORKS

INFORMATION IN MOTION®

iRODS in complying with Public Research Policy

Vic Cornell

Senior Storage Consultant

Overview

- Compliance overview
- UK examples
- Imperial College MedBio
  - Requirements
  - Architecture
  - iRODS integration
- iRODS capabilities
- Proposed Workflow
- Challenges and Unknowns

From: EPSRC Data Management Policy

- Research organisations will ensure that appropriately structured metadata describing the research data they hold is published (normally within 12 months of the data being generated) and made freely accessible on the internet.
- Where the research data referred to in the metadata is a digital object it is expected that the metadata will include use of a robust digital object identifier (For example as available through the DataCite organisation).
- Research organisations will ensure that EPSRC-funded research data is securely preserved for a minimum of 10 years from the date that any researcher 'privileged access' period expires or, if others have accessed the data, from last date on which access to the data was requested by a third party

MRC (Medical Research Council UK)

- From MRC Research data policy:
  - All research must have a “Data Management Plan”
  - Speaks of:
    - ○ **“Managing, storing and curating data. “**
    - ○ **“Metadata standards and data documentation “**
    - ○ **“Data preservation strategy and standards”**
  - These must all be specified *and adhered to.*

Imperial College MedBio

- The Imperial College Bioinformatics Support Service is a part of the Imperial College Centre for Integrative Systems Biology and Bioinformatics.
- The mission of the Imperial College Centre for Bioinformatics is to promote and co-ordinate world-class research and training in Bioinformatics within Imperial College and to provide state-of-the-art Bioinformatics support to members of Imperial College for their research.

Imperial College MedBio

- Conduct a large number of studies with respect to Systems Biology and Bioinformatics
- Range of data sources from
  - Internal
    - Next Generation Sequencers
    - Very high resolution microscopes.
  - “Big Data”
    - Phenome study systems produce 7GB data every 15 minutes.
    - They have 10 of them and they run for 2 weeks/month.
    - Maybe ½ PB Year?
  - External Datasets
- Staff!
  - Current staff are overloaded with IT tasks and don’t have time to embrace new methods. Too many workflows to follow.
  - More staff being recruited but its hard to find people with the right skills

Storage Solution

South Kensington

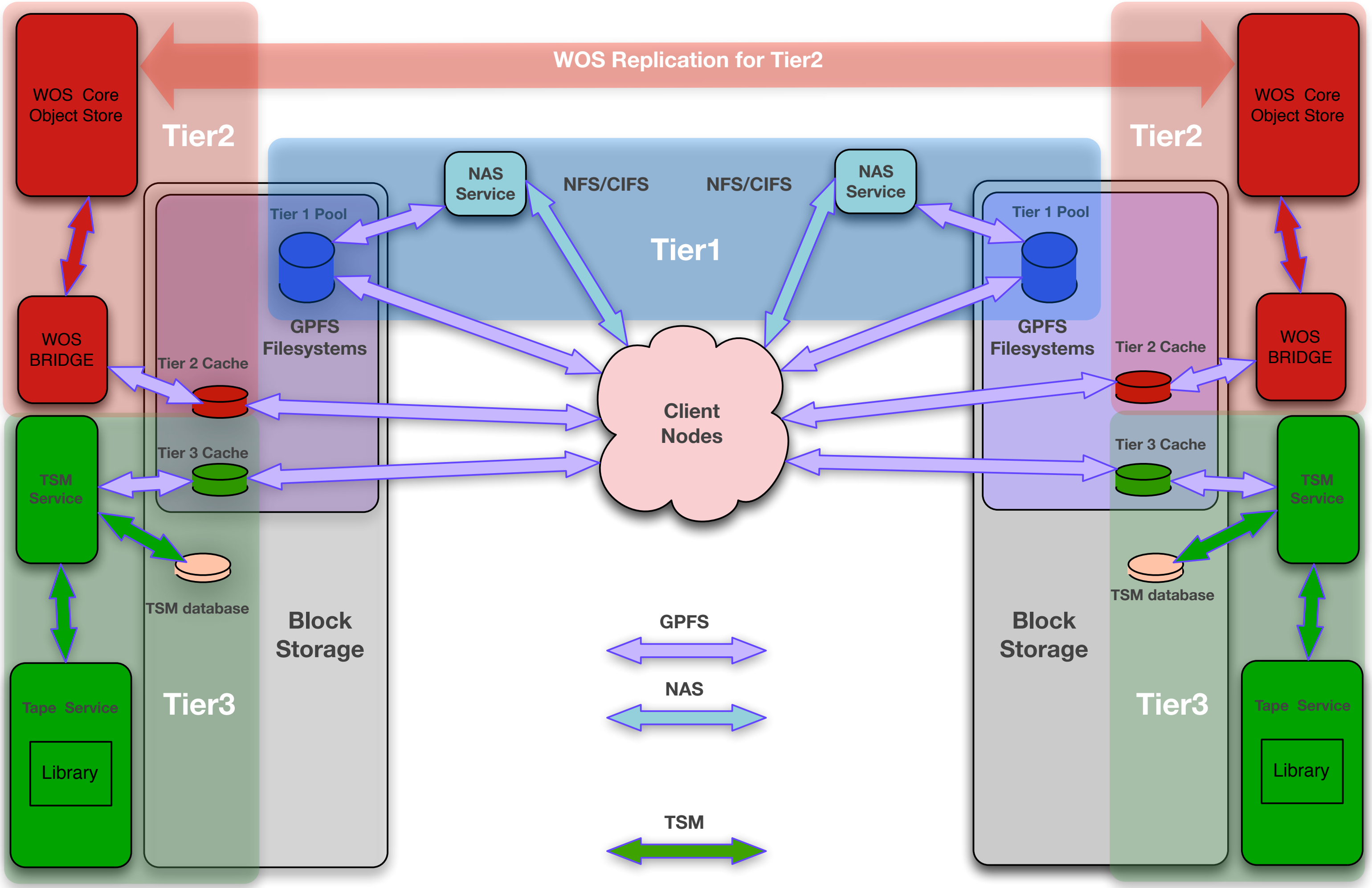
- DDN SFA 12K20 – 180 4TB Disks
- 7 x WOS 7000 – 420 4TB Disks
- 1 x Tape Library
- DDN WOS Bridge
- GridScaler (GPFS)
- GridNAS
- TSM with Space Manager for GPFS

“Infinity” Slough

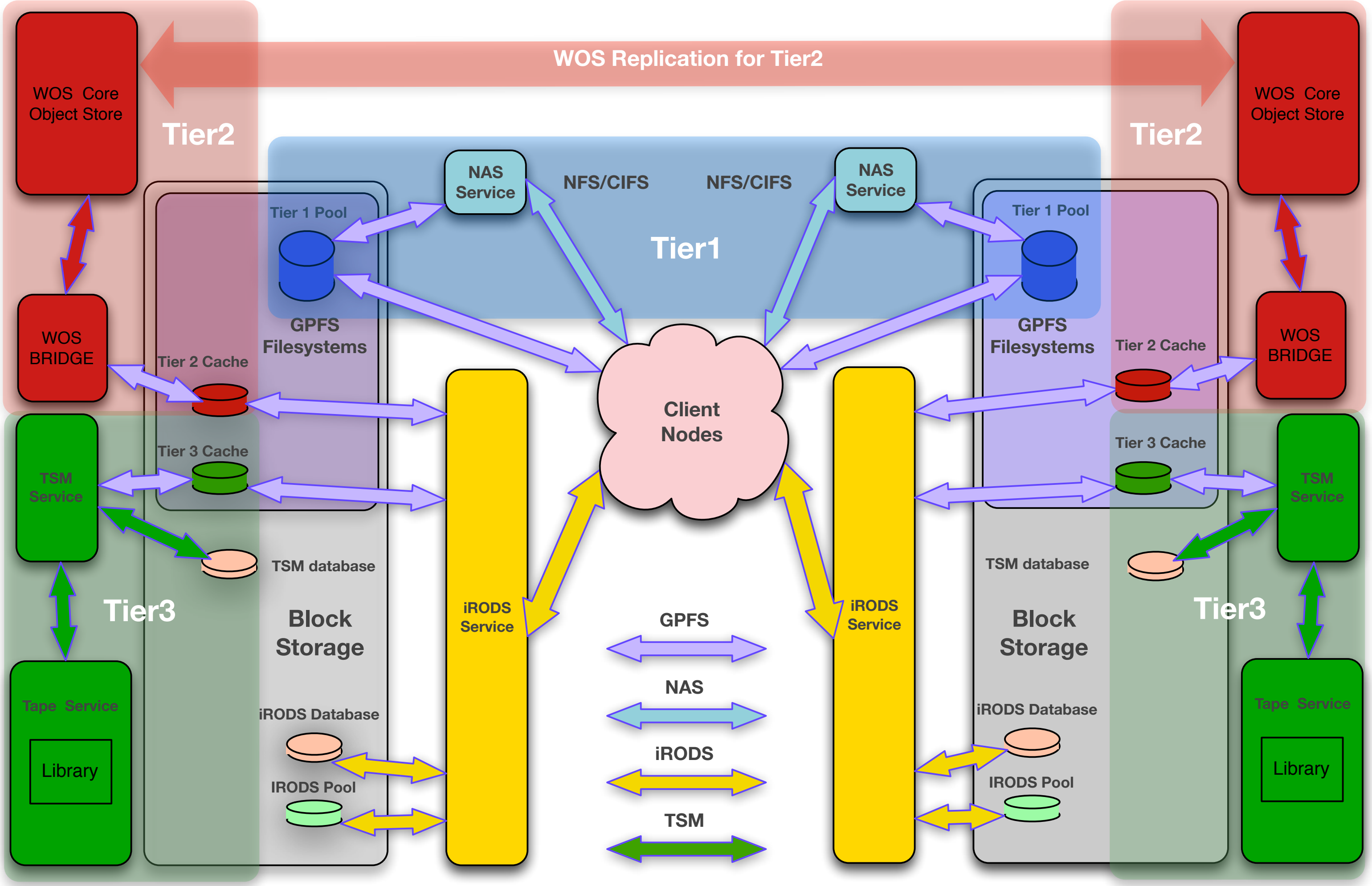
- DDN SFA SFA 7700– 180 4TB Disks
- 7 x WOS 7000 – 420 4TB Disks
- 1 x Tape Library
- DDN WOS Bridge
- GridScaler (GPFS)
- GridNAS
- TSM with Space Manager for GPFS

MedBio Architecture

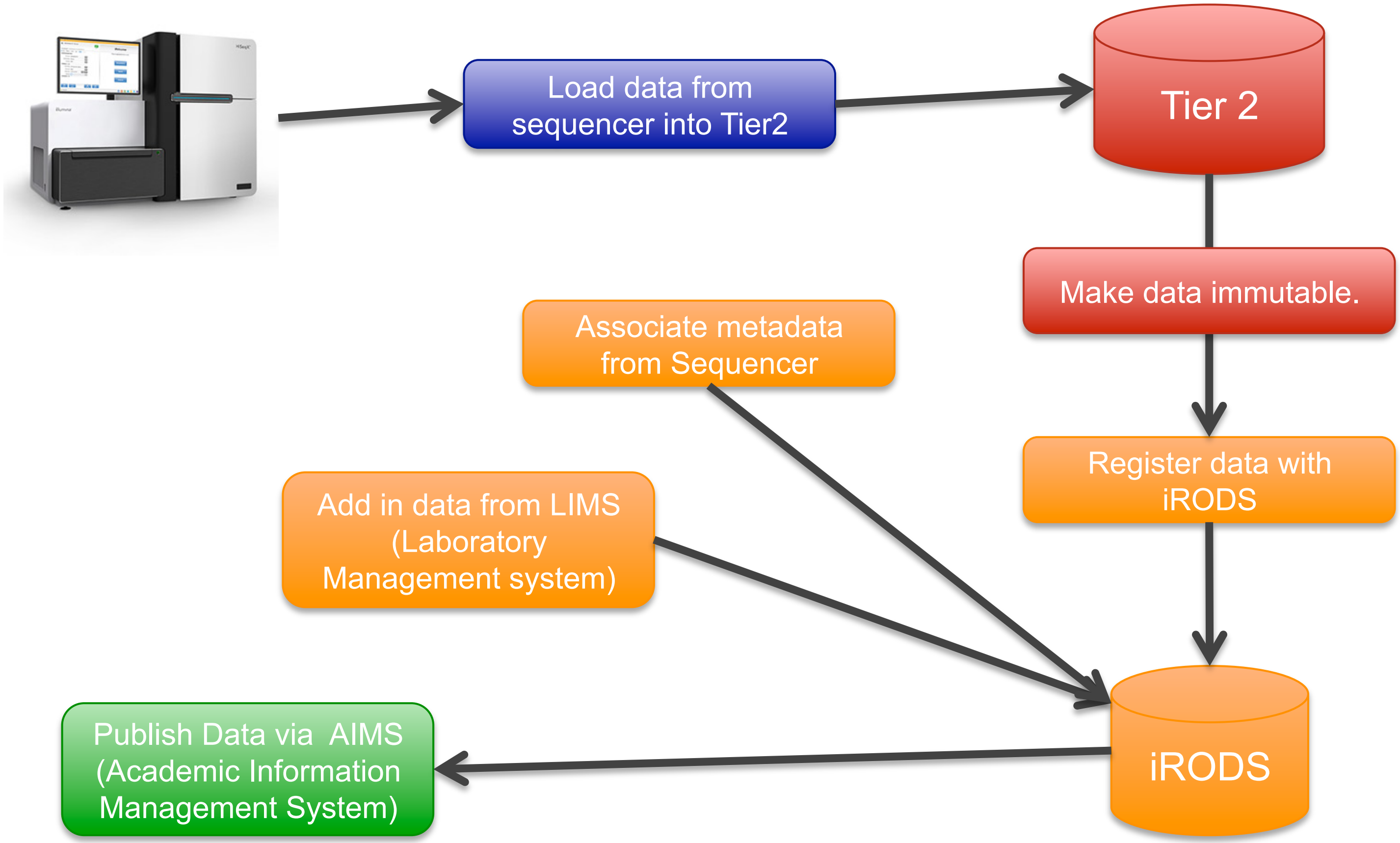
DataDirect NETWORKS INFORMATION IN MOTION



MedBio POC iRODS Architecture



Example Imperial MedBIO Workflow



Load data from sequencer into Tier2

Tier 2

Make data immutable.

Register data with iRODS

Associate metadata from Sequencer

Add in data from LIMS (Laboratory Management system)

iRODS

Publish Data via AIMS (Academic Information Management System)

iRODS for Compliance?

- Good:
  - Rules based engine which associates metadata with data.
  - Allows time based policies for data retention.
  - Can execute rules based on complex metadata queries to catch all required cases.
- Policy enforcement points
  - Can be used to implement data management and data retention policy.
  - Opportunities for data “harvesting”.
  - Once established can become a matter of record and “boilerplate” for subsequent projects
  - Still need to have resilient storage underneath.

iRODS for Compliance?

- Questions:
  - Can iRODS make data *really* immutable? At what level is this best done?
  - End-to-end metadata harvesting – can it manage an integration with LIMS.
    - Chain of custody
    - Chain of provenance.
  - Publishing: Integration with AIMS?
    - Which one?
  - Scale – this is a 7PB facility with file counts to scale – can Imperial do this in one zone?
  - If not – how will they manage federation so it works.
  - Database stability, availability and recoverability.
    - Who sets the standards?

DataDirect™
NETWORKS
INFORMATION IN MOTION™

Q&A

DataDirect™
NETWORKS
INFORMATION IN MOTION™

THANK YOU

Links

http://www.mrc.ac.uk/research/research-policy-ethics/data-sharing/data-management-plans/

https://documentation.tgac.ac.uk/display/COPO/COPO+Documents

http://www3.imperial.ac.uk/bioinfsupport