Scientific Computing & Storage at
The Francis Crick Institute

Michael Holliday
Senior HPC & Research Data Systems Engineer
Michael.holliday@crick.ac.uk
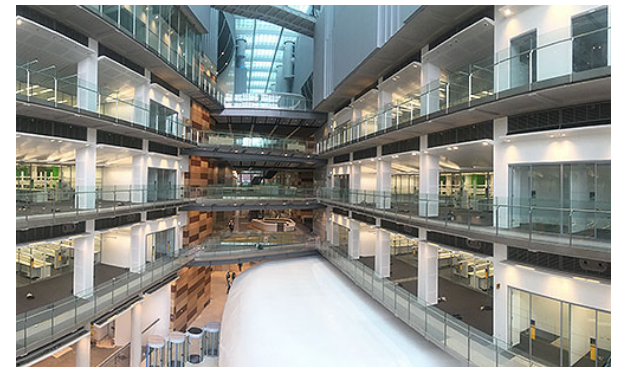
THE FRANCIS CRICK INSTITUTE

The Francis Crick Institute

- The Francis Crick Institute is a biomedical discovery institute dedicated to understanding the fundamental biology underlying health and disease.

- Founded in 2015 when the MRC's National Institute for Medical Research and CRUK's London Research Institute joined the crick along side our founding partners:
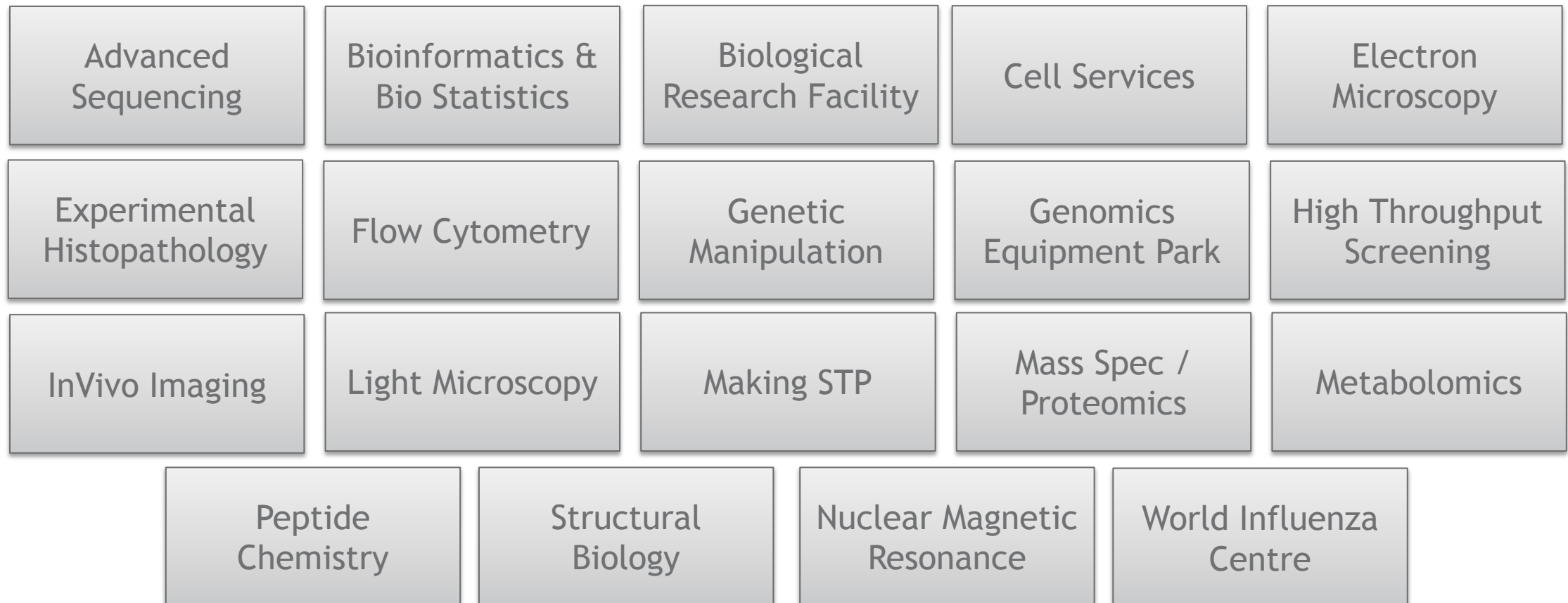
The Francis Crick Institute in numbers

- 1300 Scientists across 130 Labs, Research Groups & Science Technology Platforms

- 300 Operations Staff

- Over 30000 pieces of Scientific Equipment

- 13 Floors, 1500 Rooms

- 2 National / International facilities

The Francis Crick Institute – Science Technology Platforms

Advanced Sequencing	Bioinformatics & Bio Statistics	Biological Research Facility	Cell Services	Electron Microscopy
Experimental Histopathology	Flow Cytometry	Genetic Manipulation	Genomics Equipment Park	High Throughput Screening
InVivo Imaging	Light Microscopy	Making STP	Mass Spec / Proteomics	Metabolomics
	Peptide Chemistry	Structural Biology	Nuclear Magnetic Resonance	World Influenza Centre

The Francis Crick Institute – Instruments

- Titan Cryo Electron Microscopes

- Sequencers

- CT Scanners

- UltraSound Machines

- NMR Spectrometers

- Light Microscopes

- Mass Spectrometers

CAMP – Crick Data Analysis and Management Platform –
Where did we come from:

- 4 Isilon Clusters with around 4PB of data going back 40 years

- Numerous NAS Boxes, Hard drives, USBs etc (we're still finding out about them now…)

- Storage from CRUK and NIMR were managed in very different ways, and had their own Domains and Permissions

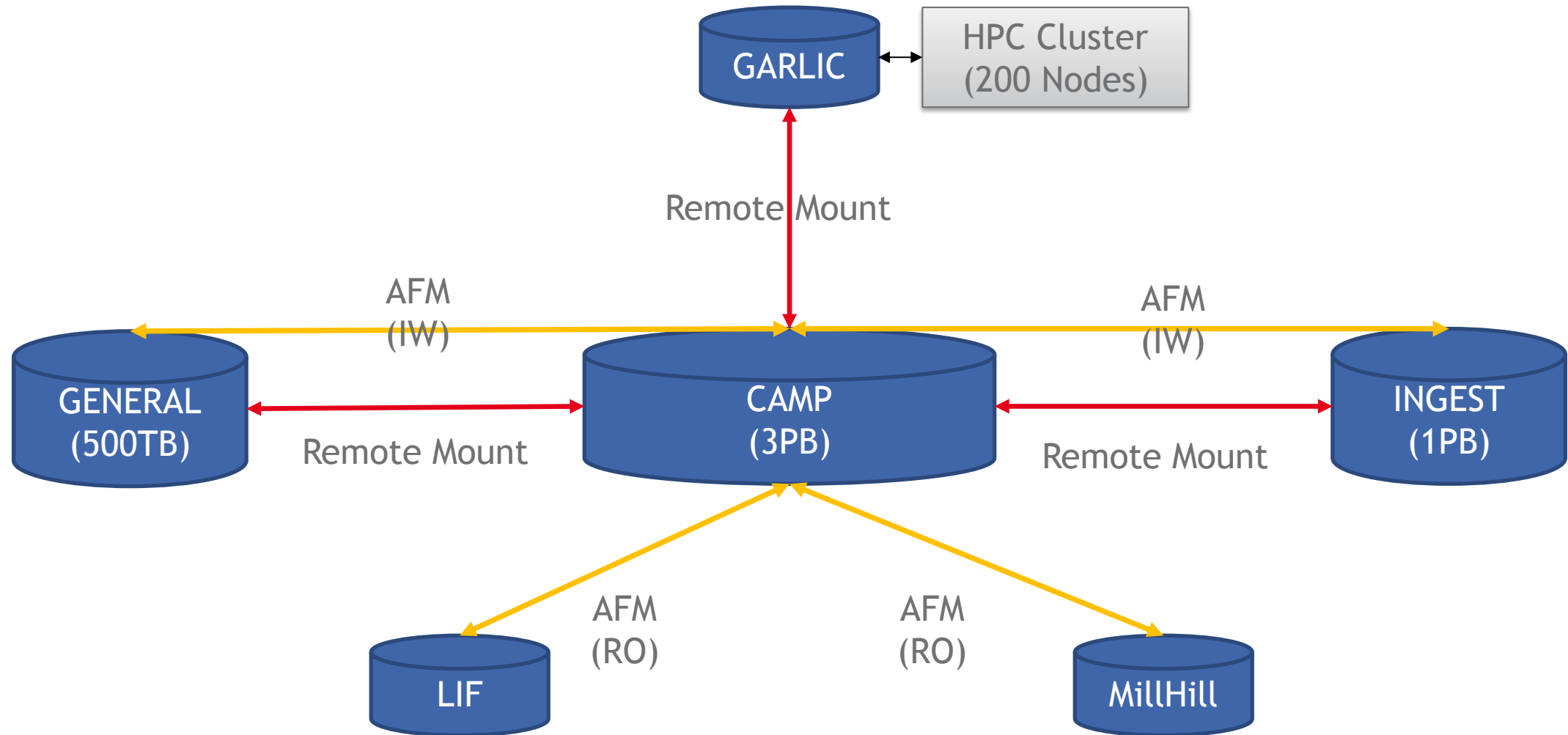- Data replicated in multiple locations but it wasn't always clear

CAMP – Crick Data Analysis and Management Platform–

Where we started at the Crick:

CAMP – Crick Data Analysis and Management Platform –

- **FDR IB Fabric**
  - 2 Level Fabric

- **CAMP**
  - 2 DDN GS12K Systems
  - 20 Enclosures

- **INGEST, General, LIF, Millhill:**
  - Each has 1 DDN GS7K
  - 1 or 2 Enclosures

Problems we encountered...

- AFM
- Migration of legacy data
- Data, More Data & Even More Data....
- Instruments
- Labs adapting to the new systems

AFM

- INGEST and General were set up to be caches of CAMP storage

- At one point INGEST has 130 AFM links to CAMP

- Delays to syncing were noticeable to scientists, particularly those using the HPC cluster

- Permissions were not syncing correctly between the two

- Cache was out of sync with home and unable to recover

- Used Independent Writer Mode

Data Migration

- Moving Data from legacy sites as labs were physically migrating through a combination of AFM, GS7Ks, QNAPs

- AFM – was too slow for large data transfers, but was useful for small priority data

- QNAPS, GS7Ks allowed bulk transfers when physically transferring the systems between sites

- Labs need to continue work asap

- "Priority Data" meant different things to different labs….

Data, More Data & Even More Data....

- Our Labs and STPs are producing more data at faster and faster rates

- 1 Titan CryoEM Microscope can produce up to 1PB data a year (we have three of them).

- Currently ½ Billion Files most only kb's in size

- Covers pretty much every file type

- CAMP has been Expanded from 3PB to 10PB

- Data Needs to be kept for 10 years

- Archive is being planned

Data Usage Tracking
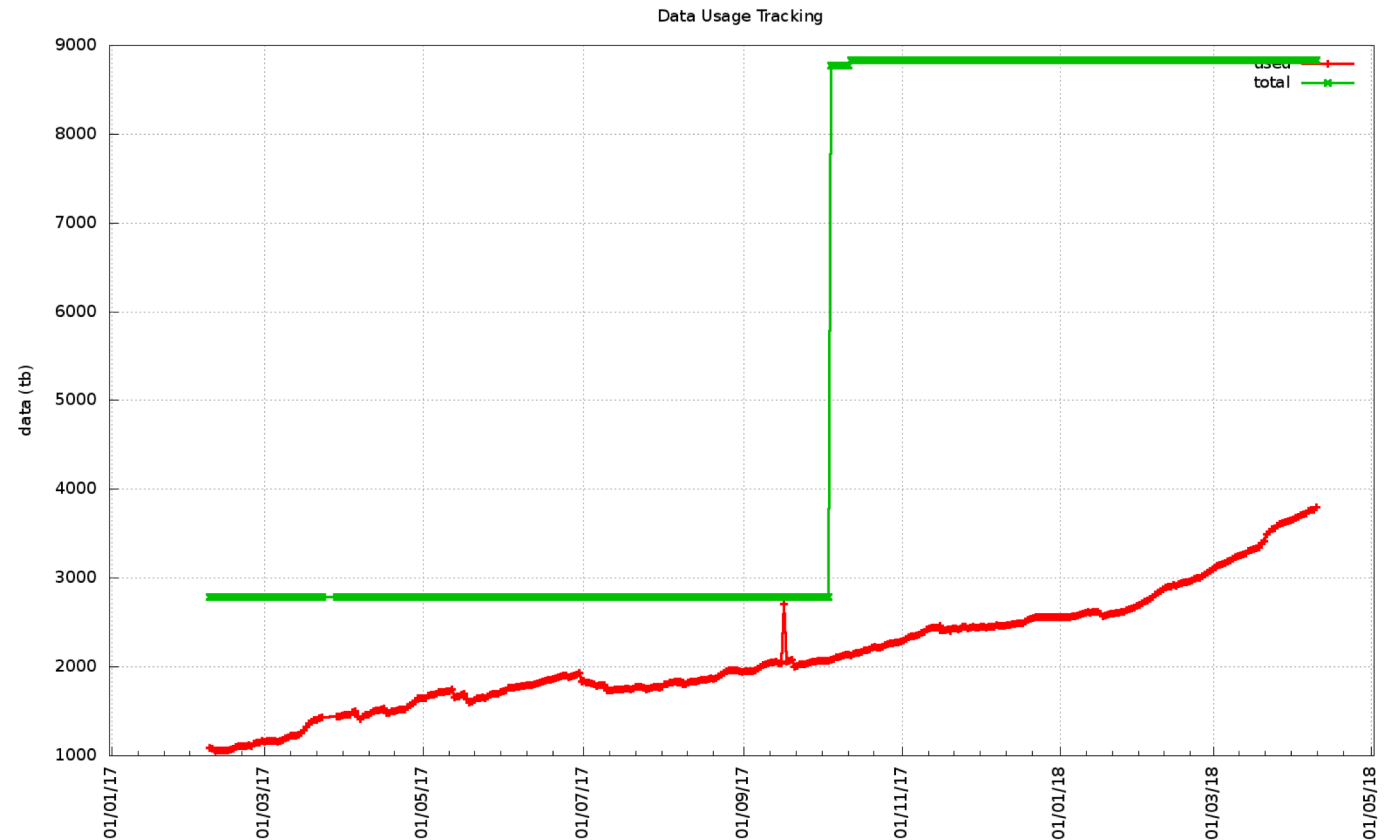
Connecting Instruments

- Instruments need to mount CAMP in a range of ways
    - NFS
    - SMB
    - Application Mounts

- Proprietary Operating Systems and Interfaces

- Unsupported Operating systems eg Windows XP, Centos 5

- Security and Access limitations – in terms of both logging in, and physically accessing the instruments

- Many had a local account shared between lab members

Adapting the Labs

- Each Lab has their own unique way of working

- Each of the old sites had differing levels of IT management

- Major Change moving to the Crick

- Different ways of using the system

- New structures, new limitations