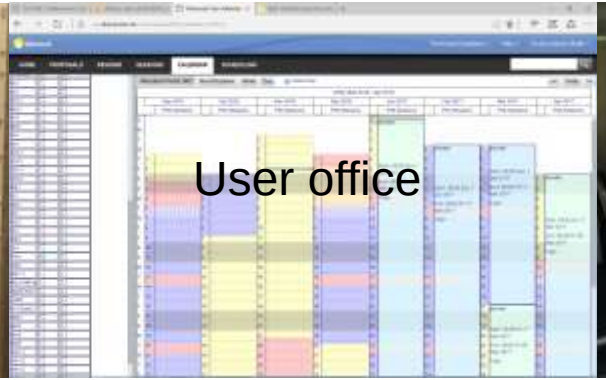
# Processing at eBIC/Diamond Light Source

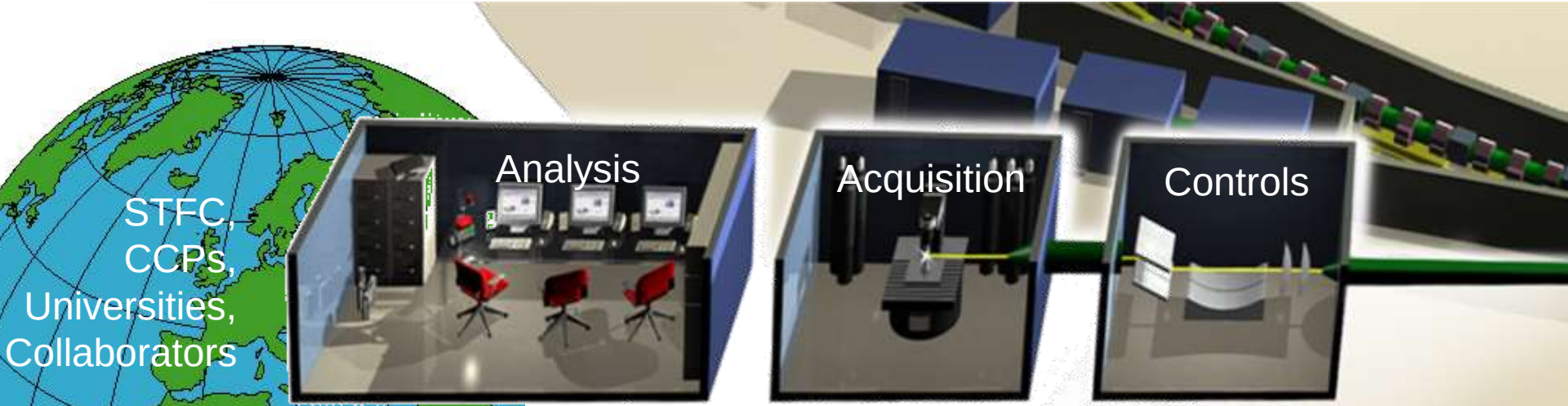
Alun Ashton

diamond

# Computing/Software Support Groups





# Data Analysis – Experimental Challenges

| **BEFORE** (FROM DLS/USER'S INSTITUTION) | **IMMEDIATE** (DURING EXPERIMENTS) | **SHORT TERM** (BEFORE THE USER GOES HOME) | **LONG TERM** (FROM DLS/USER'S INSTITUTION) |
|---|---|---|---|
| Simulations<br><br>Processing of older datasets | "Real time" data processing, analysis and visualisation – to make experimental decisions | Data reduction and processing – Users go home with clean data free of instrument artefacts.<br><br>Preliminary data analysis – helpful, but may require significant processing power and know-how | Detailed analysis – from data to information.<br><br>Incorporating results from other techniques.<br><br>Experiments:<br><br>➢Provide parameters for a model.<br><br>➢Test/verify a model or theory.<br><br>➢Show where a new theory or model is required. |

mond

Database ADM

Core Software

Core Analysis

MX,X

Scattering & Spectroscopy

Imaging

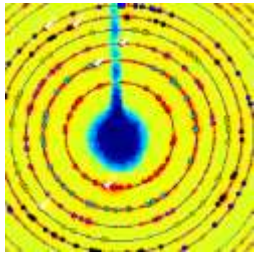
Software Engineers | Software Scientists | Other Funding

# Imaging Team

Savu

Multi-dimensional and Multi-Modal tomographic reconstruction pipeline

ptypy



Extensions to Savu for Multimodal data collections

SuRVoS

Machine learning assisted annotation and segmentation for volumetric data.

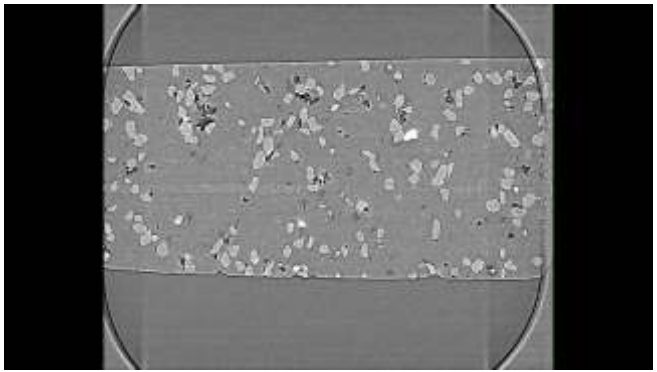
SCIPION RELION CCP-EM

FEI

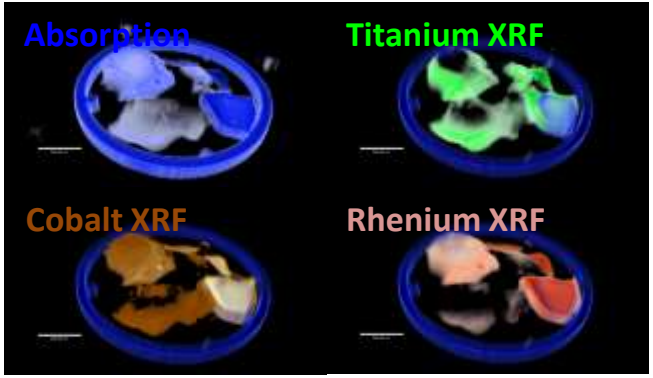
part of Thermo Fisher Scientific

Automation of data processing for electron tomography and single particle analysis.

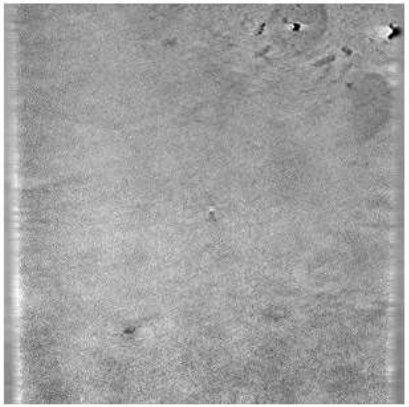
# Imaging Team



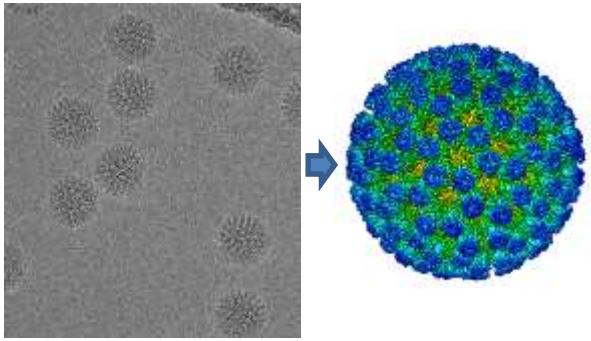
Thanks to Robert Atwood




Thanks to Stephen Price



Thanks to Michele Darrow



# LiMS:ISPyB/SynchWeb



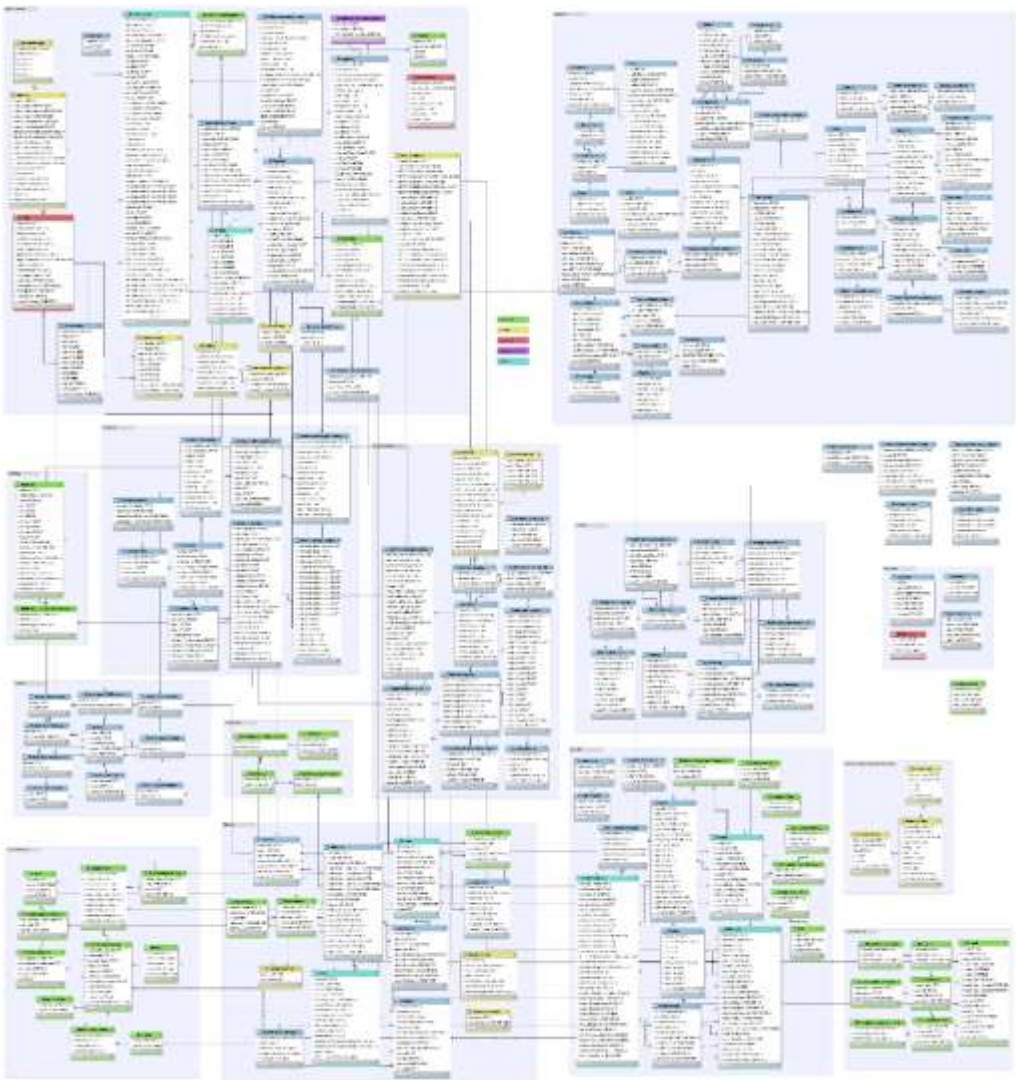
- MariaDB/MySQL-aware intelligent proxy
- Load balancing
- Monitoring mechanisms
- Logging, Filtering

- Multi-master – 3 nodes
- Each node has the full dataset
- Synchronous replication

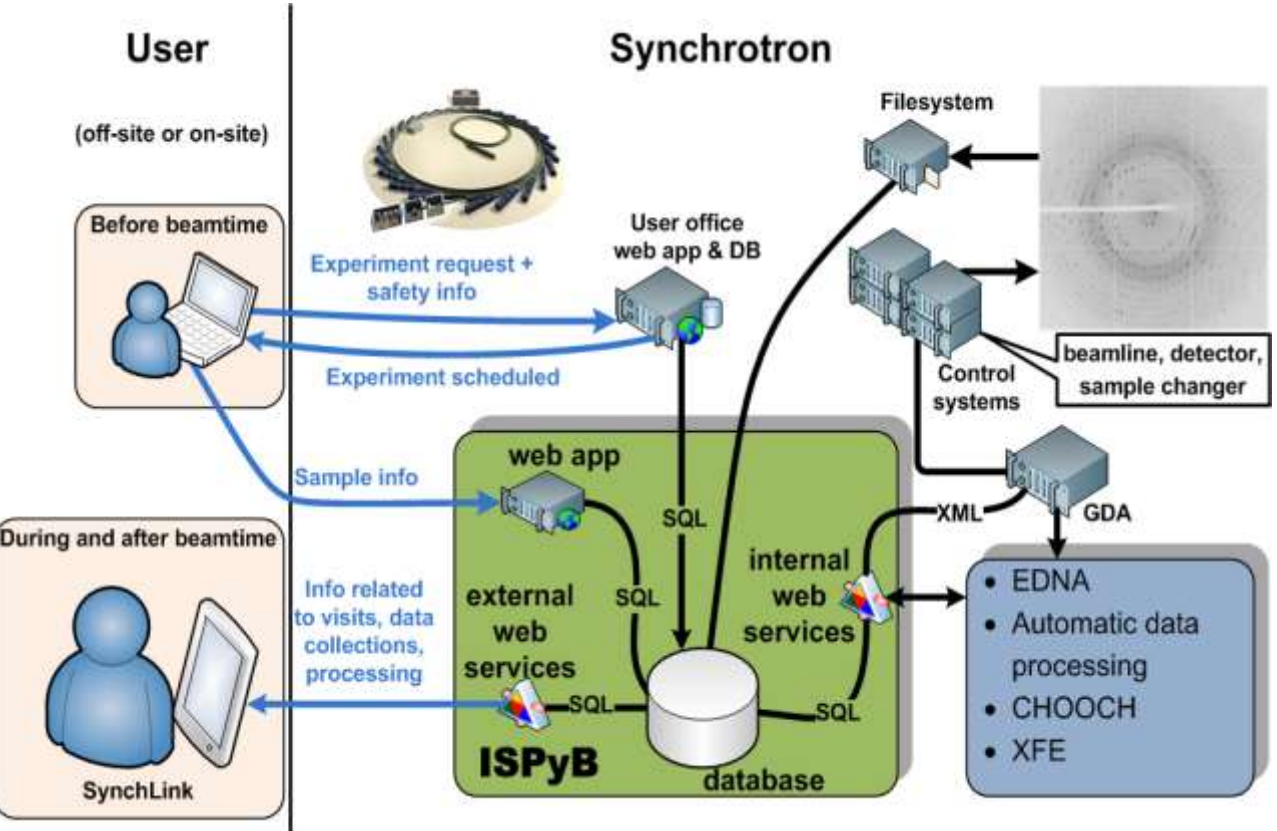
# LiMS:ISPyB/SynchWeb


- MariaDB/MySQL-aware intelligent proxy
- Load balancing
- Monitoring mechanisms
- Logging, Filtering

- Multi-master – 3 nodes
- Each node has the full dataset
- Synchronous replication

# LiMS:ISPyB/SynchWeb

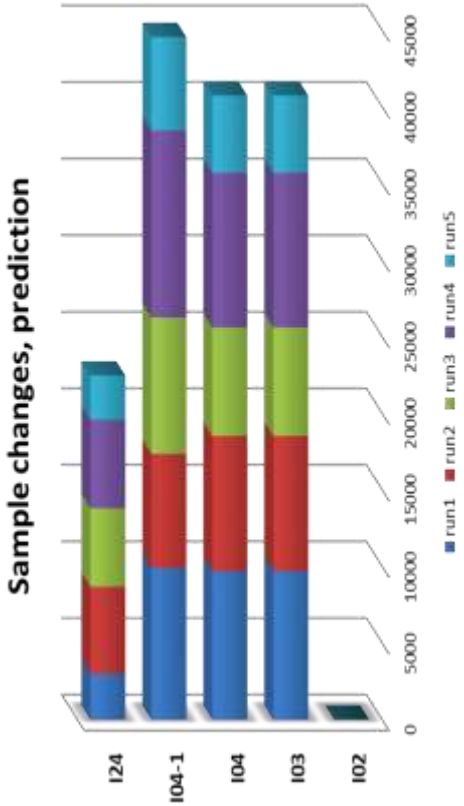




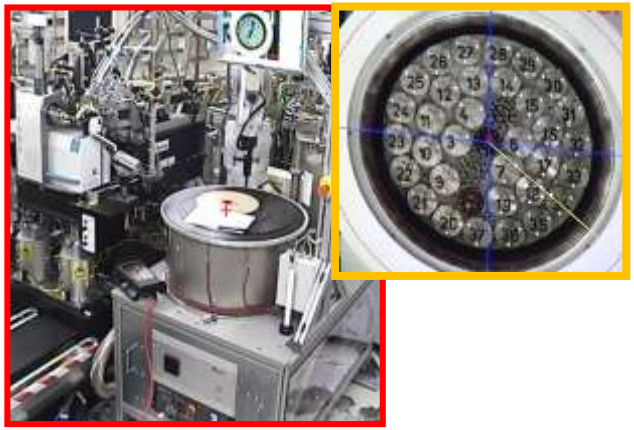



# LiMS:ISPyB/SynchWeb

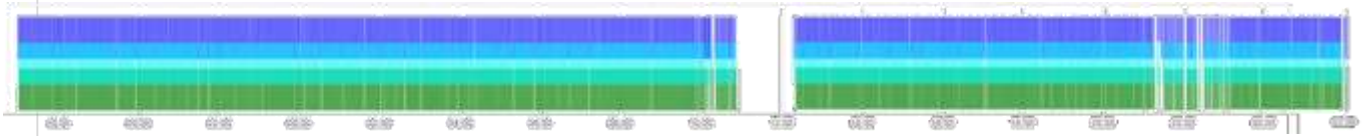
# Macromolecular Crystallography Automation




- Use case: fragment screening
  - Crystals mounted at synchrotron (100s/day)
  - Auto-collect (visual centring)
  - Critical: Capacity, speed, robustness
  - Not critical: small footprint…



*"67 pucks, 978 data collections, zero problems - in about 40 hours"*



Before: *~20 datasets / hr* After: **~31/hr**

Time / s

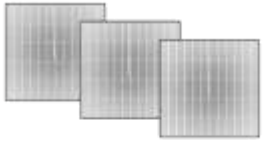
0

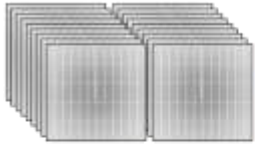
60

120

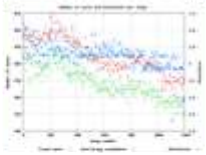
180

## Data Collection: characterisation or full set

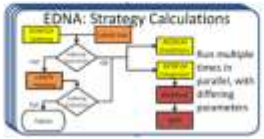




## Per-image Analysis



## Strategy calculations – EDNA, Mosflm & XDSME




EDNA:

- Standard native & anomalous
- High redundancy (SAD)
- Gentle

Mosflm

- Standard native & anomalous (no exposure times)

## Fast Data Processing: Fast_DP

Fast data processing employing XDS and substantial parallelism – delivers merging statistics and scaled data in around 2 minutes

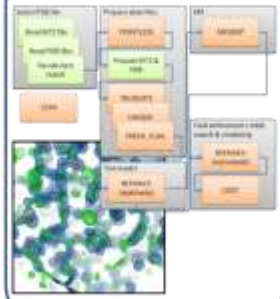
## Data Processing with xia2

- -2d: Mosflm / Aimless
- -3d: XDS / XSCALE
- -3dii: XDS / XSCALE all images for autoindex
- -dials: DIALS / Aimless

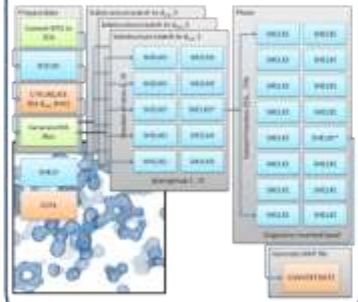
Also: multi-xia2
Also: Spacegroup from ISPyB

~ 10-30 minutes

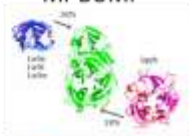
## Diff Map: DIMPLE



## Experimental Phasing: Fast_EP



## Mr BUMP



~ 10-30 minutes

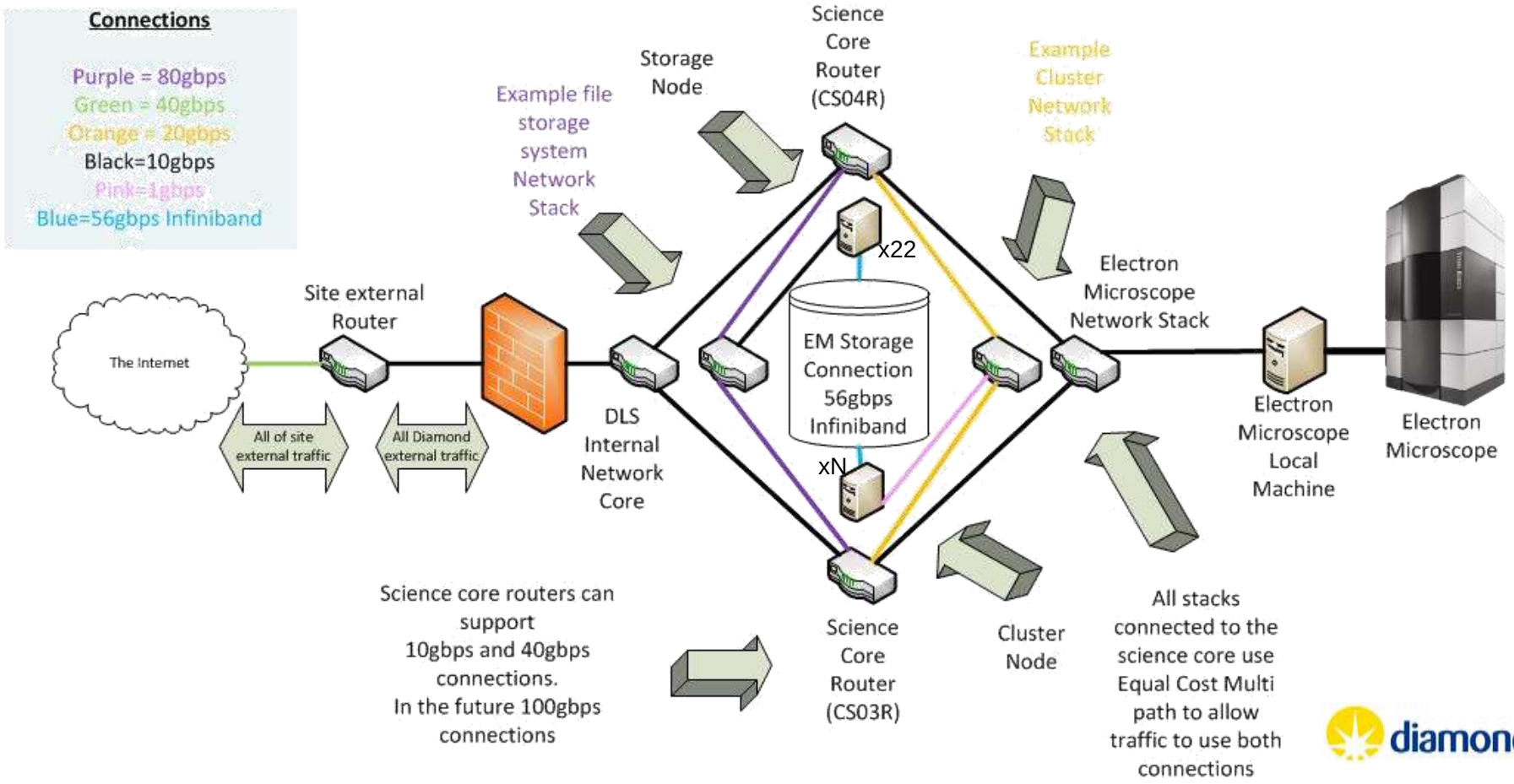

diamond

# For eBIC

- Building on what we already have

- Produce automated pipelines for
  - Single particle analysis
  - Tomography
- Prioritise feedback useful at time of data collection
- Collect data into ISPyB to facilitate interoperability of data between measurement types
- View data online from SynchWeb

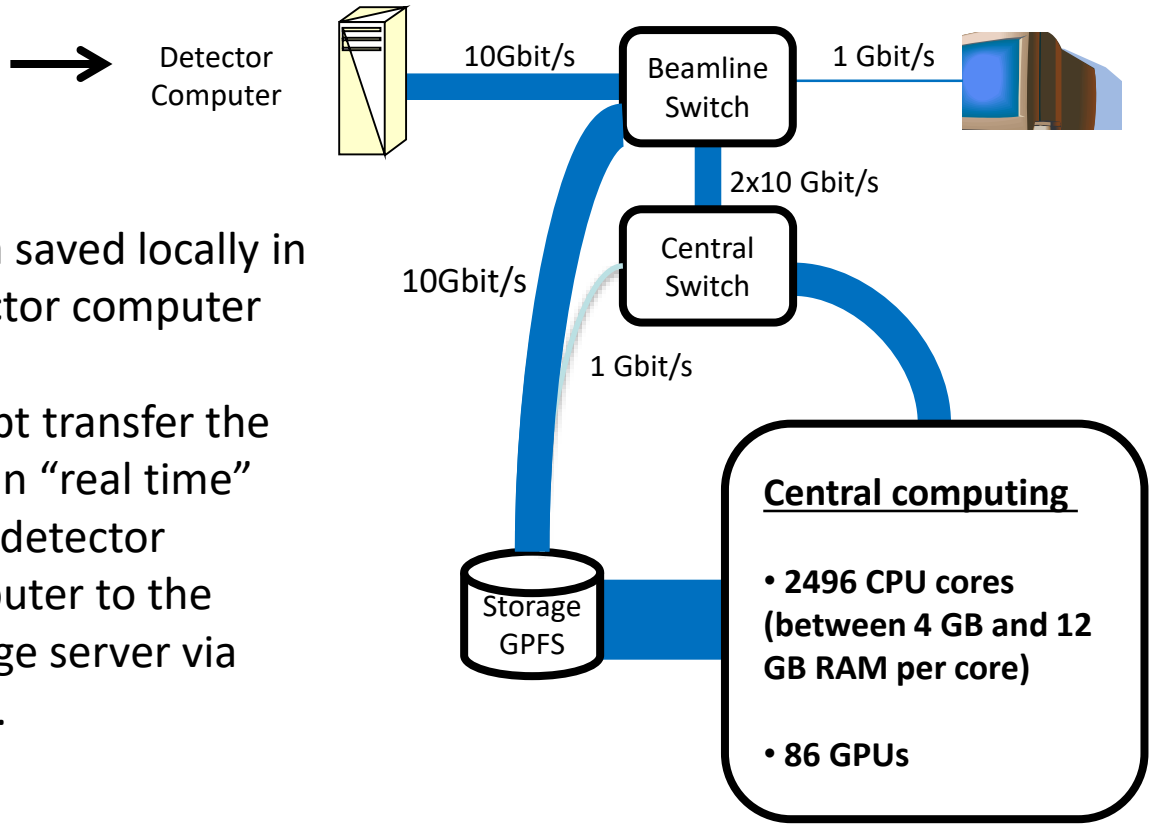
# Electron Microscope Connectivity



# Outline of current computing resources

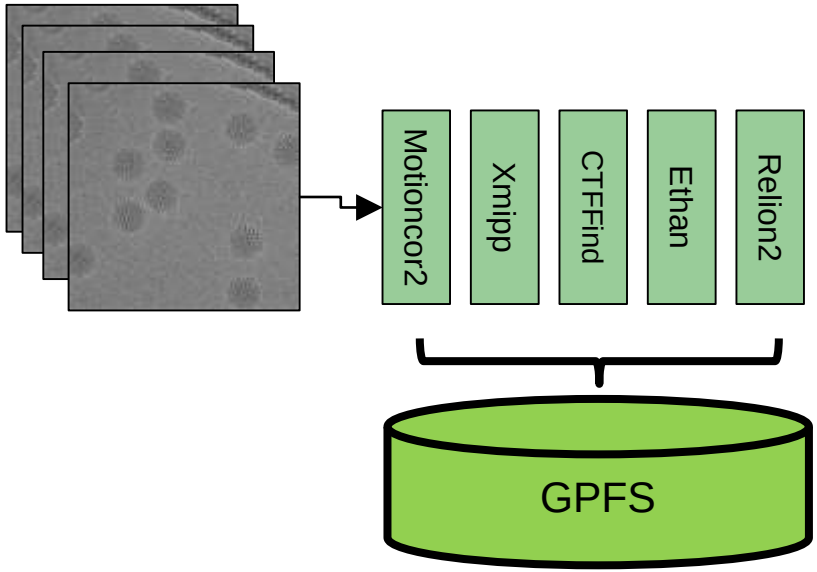
- Data saved locally in detector computer

- Script transfer the data in “real time” from detector computer to the storage server via rsync.

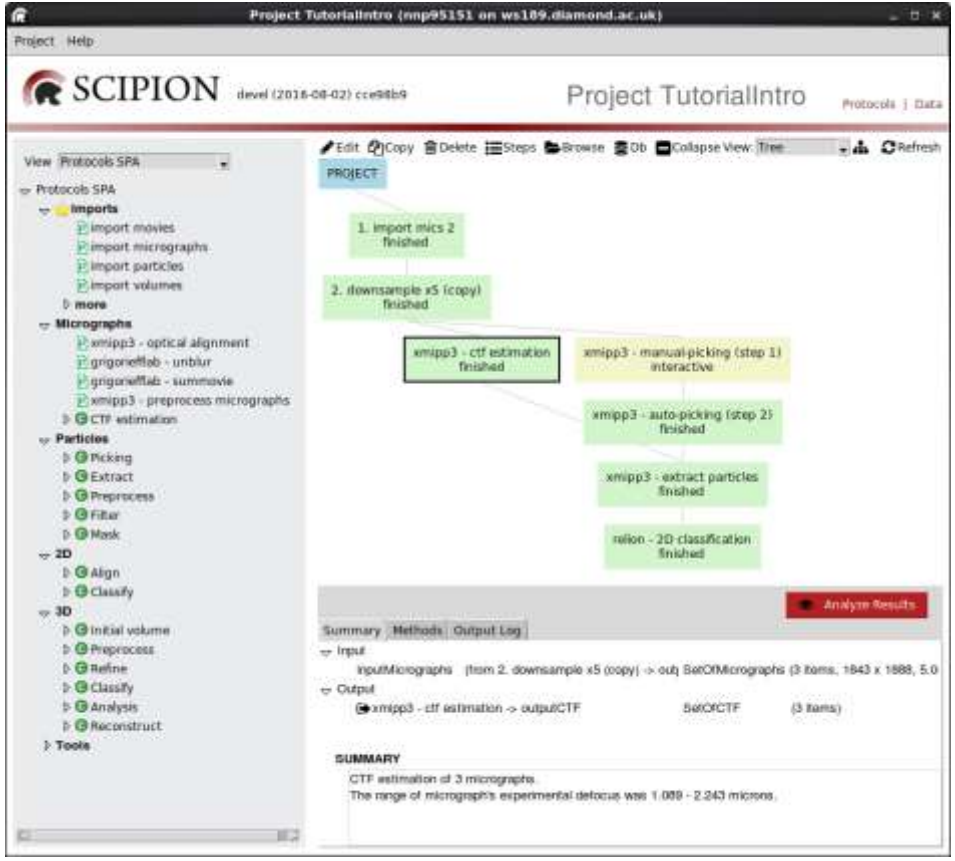


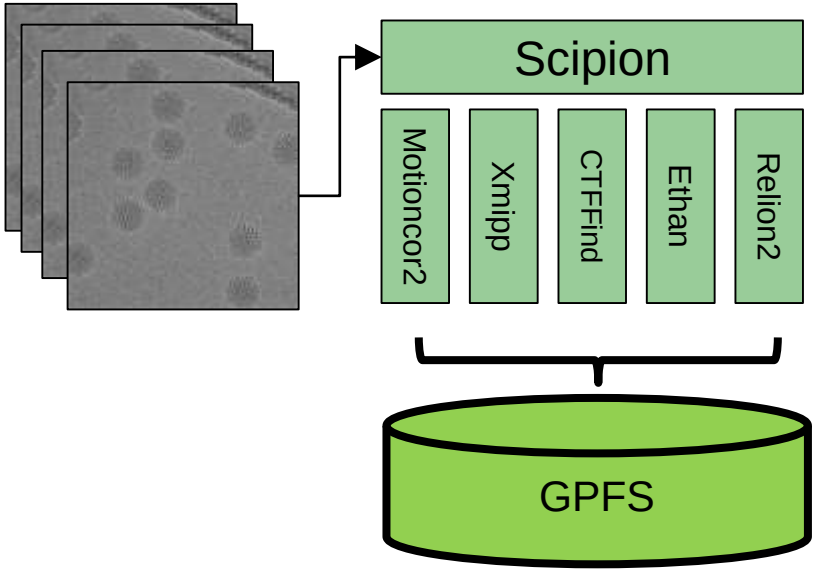

diamond

# Single Particle Analysis



# Single Particle Analysis

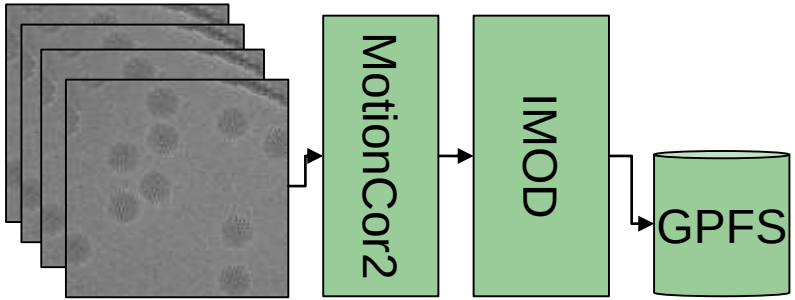


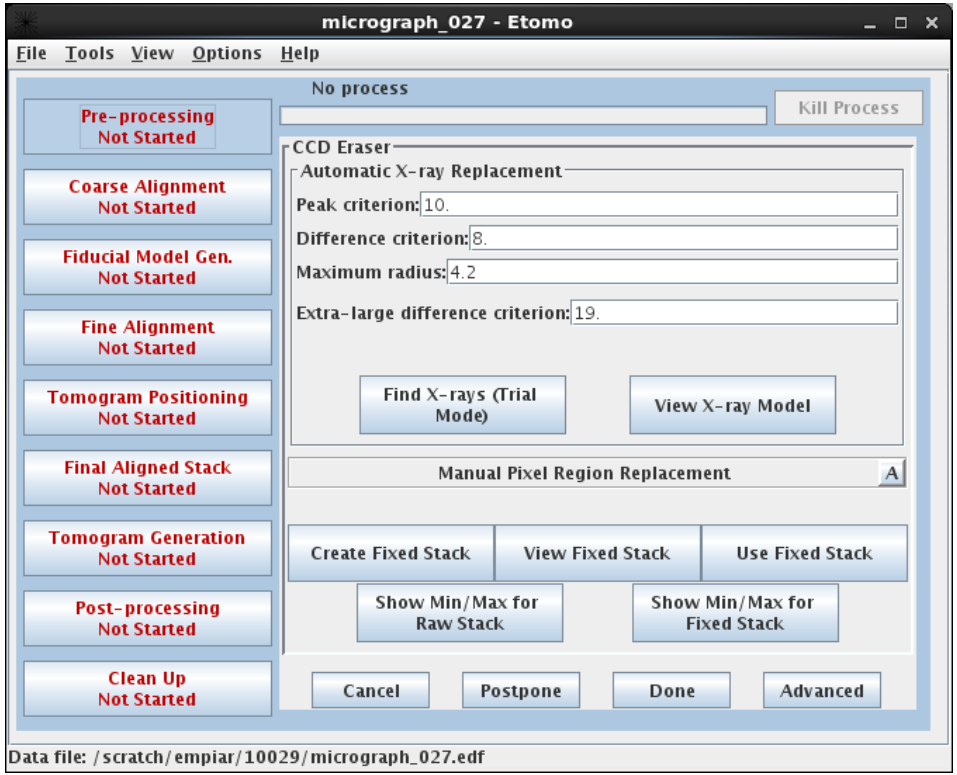

iNEXT

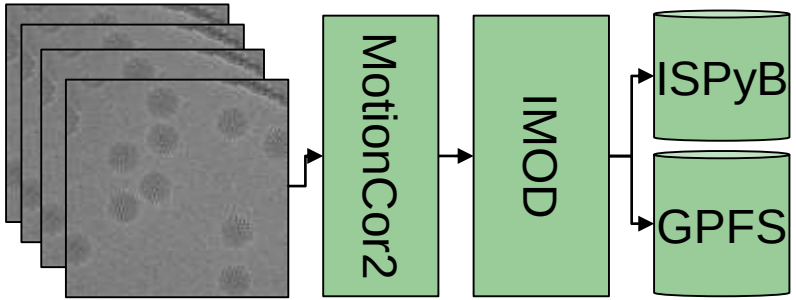
diamond

# Tomography

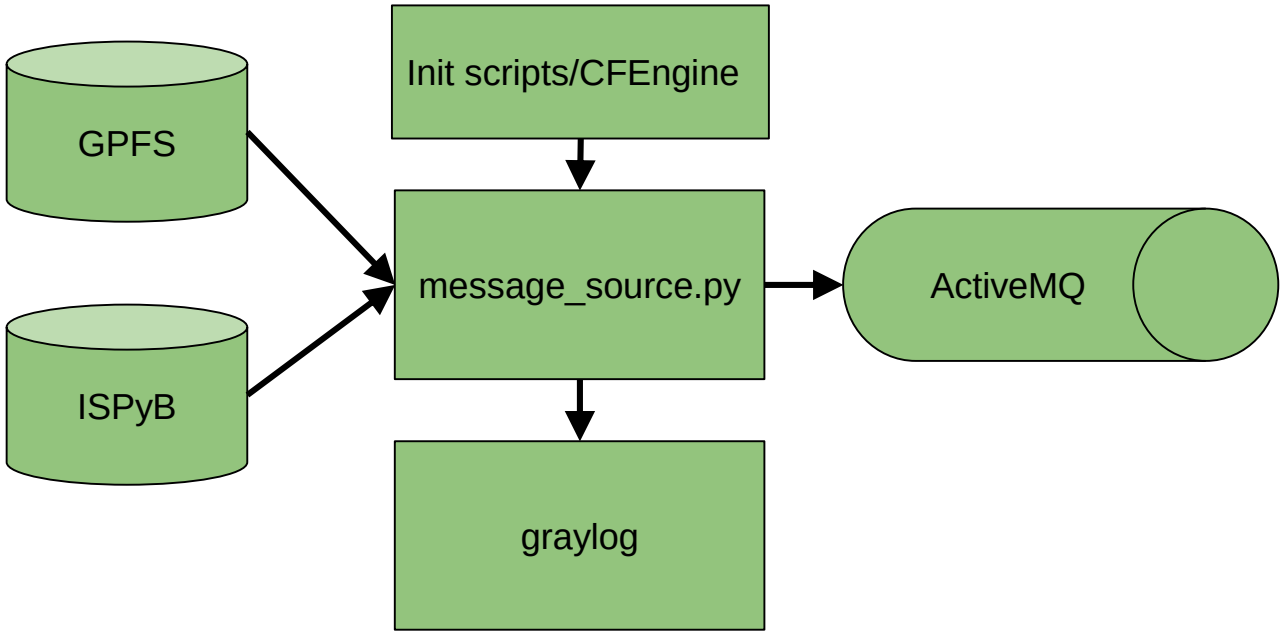


# Tomography: IMOD

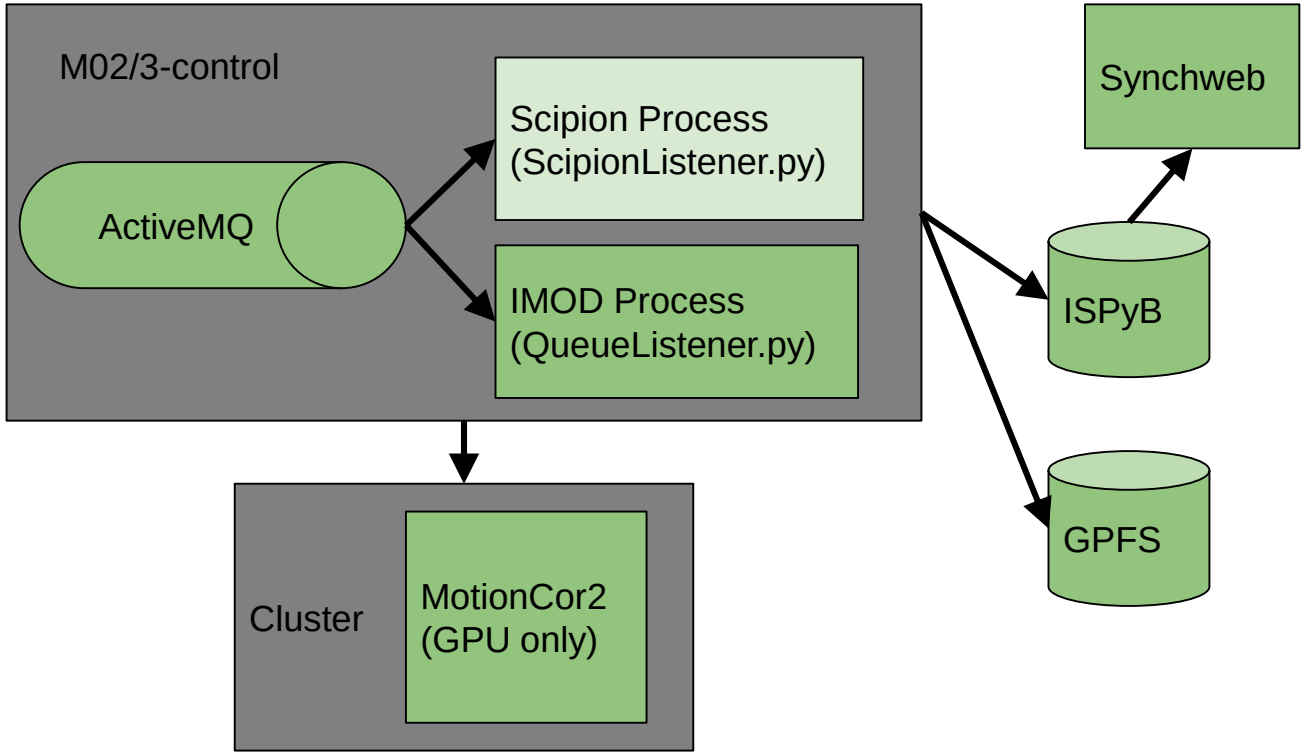



# Automating : Notification



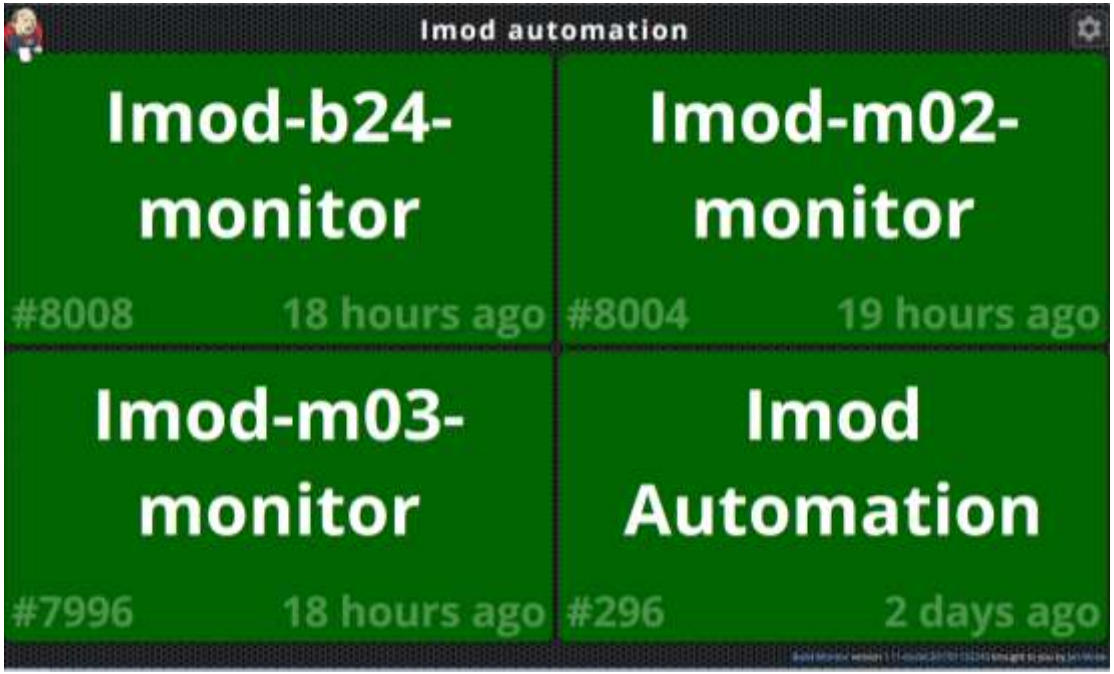
# Automating : Processing

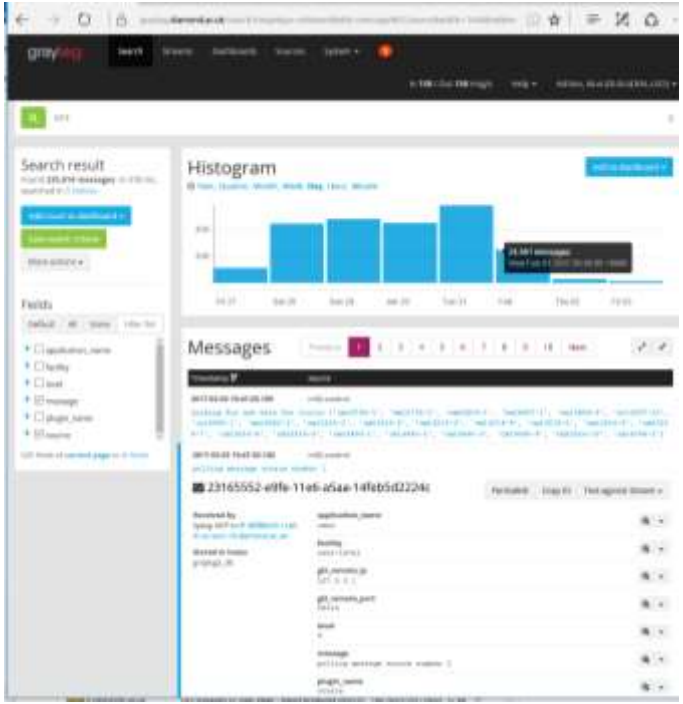


# Current Status

- Tomography running for some time
  - also running on full field cryo-Xray microscopy beamline
  - Scipion setup, modified and well used
- Automatically triggered Scipion went live 30th January 2017
- Some data in ISPyB

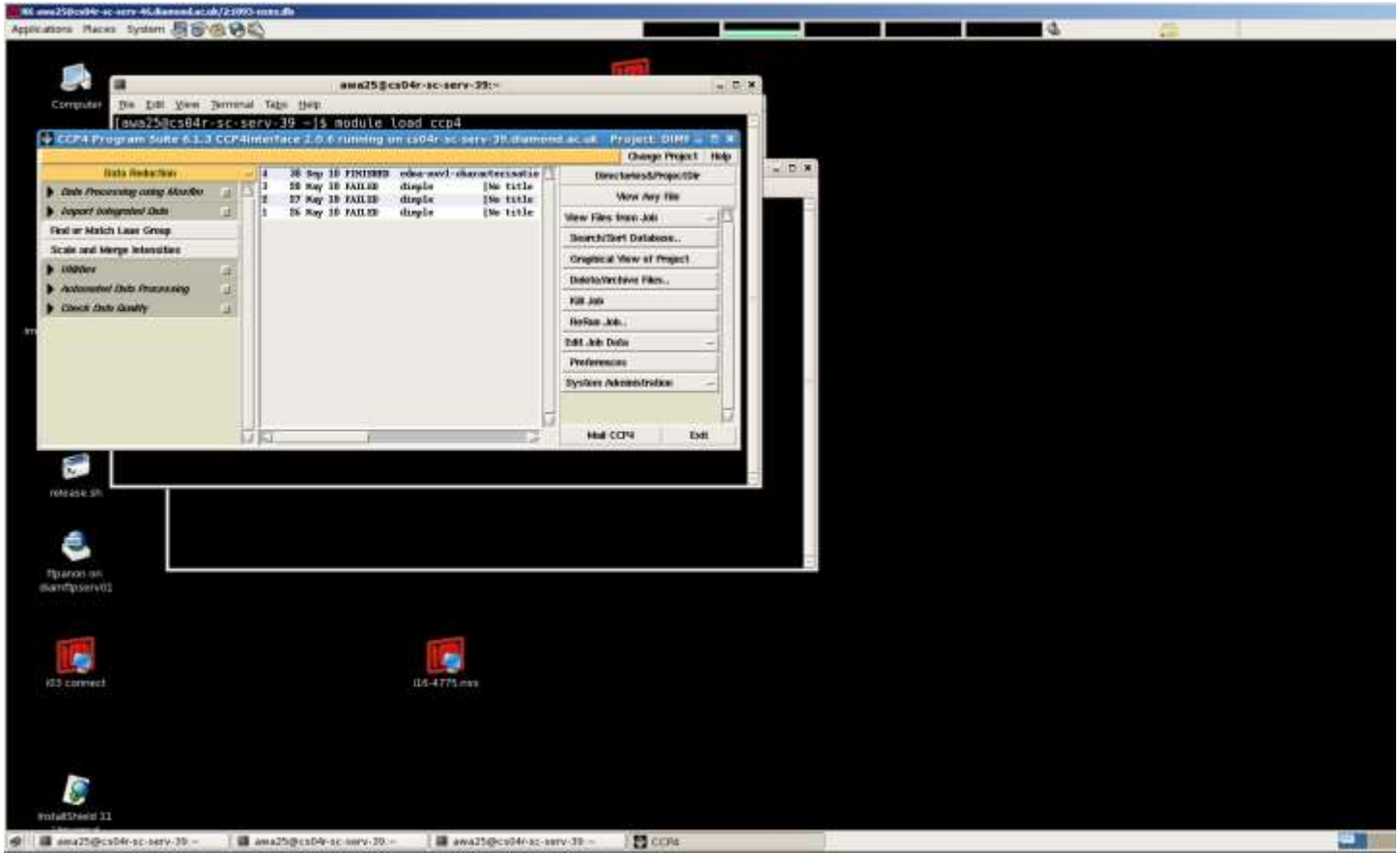
# Builds and Services




diamond

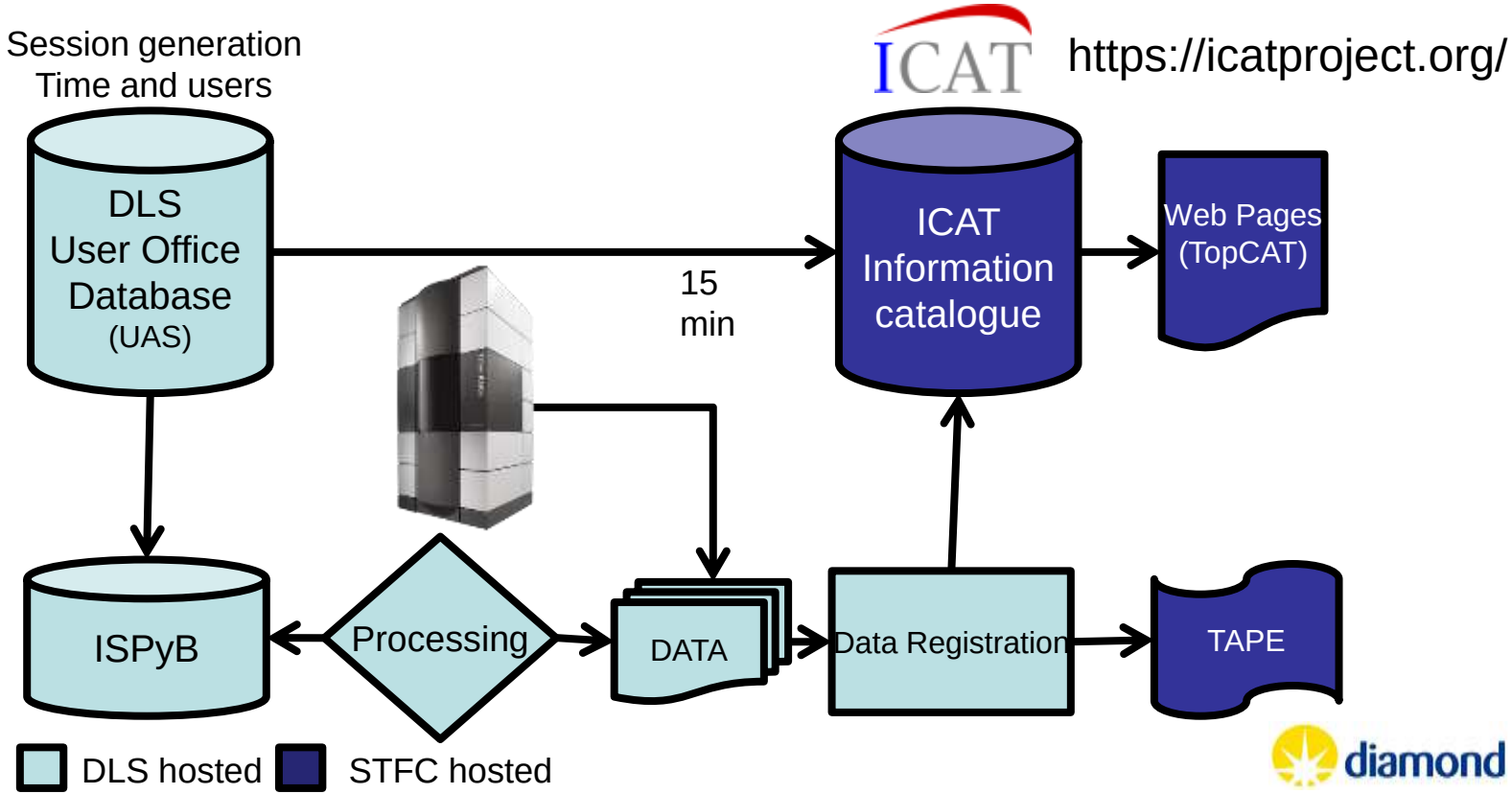
# Remote Reprocessing




Fair play policy.

diamond

# Data is archived and metadata captured from every stage of the process



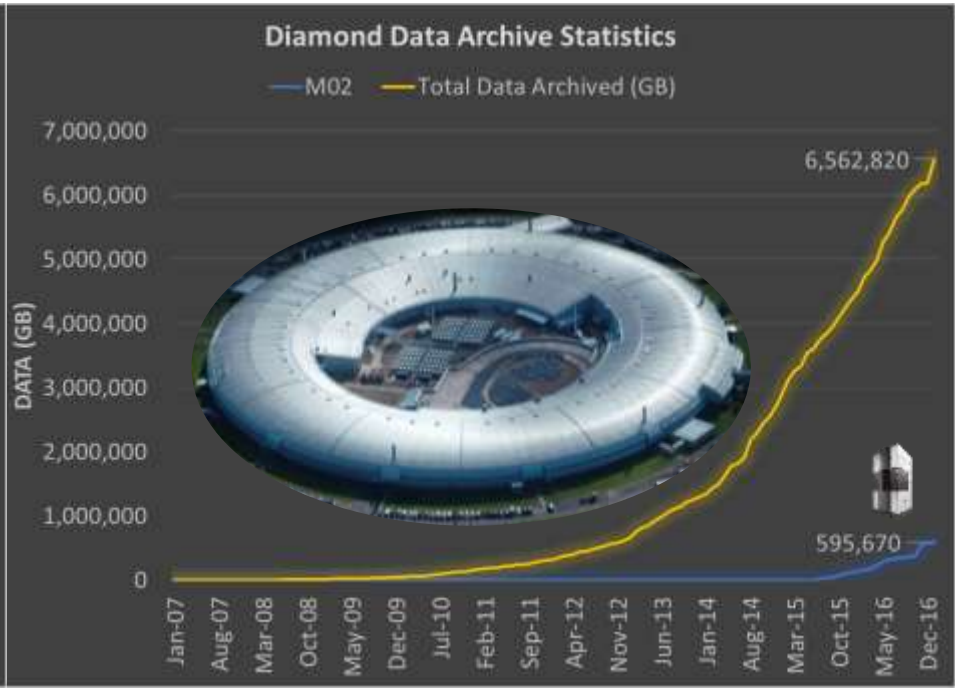
# EM data flow


- Data capture at full rate on both direct electron detectors (*e.g.* 17 fps for Falcon II or 40 fps for the K2).
- Falcon II data rate ~50-100 movies/hr
- K2 data rate ~25-75 movies/hr
- All data are directly written to high speed central computing/storage facility.
- All data are archived to tape and stored for the lifetime of the media.
- Diamond cluster available to external users particularly during beamline shut downs….

diamond

# EM data flow


M02
700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
Data (GB)
595,670
Jul-15
Aug-15
Sep-15
Oct-15
Nov-15
Dec-15
Jan-16
Feb-16
Mar-16
Apr-16
May-16
Jun-16
Jul-16
Aug-16
Sep-16
Oct-16
Nov-16
Dec-16
Jan-17


Diamond Data Archive Statistics
M02
Total Data Archived (GB)
7,000,000
6,000,000
5,000,000
4,000,000
3,000,000
2,000,000
1,000,000
0
DATA (GB)
6,562,820
595,670
Jan-07
Aug-07
Mar-08
Oct-08
May-09
Dec-09
Jul-10
Feb-11
Sep-11
Apr-12
Nov-12
Jun-13
Jan-14
Aug-14
Mar-15
Oct-15
May-16
Dec-16

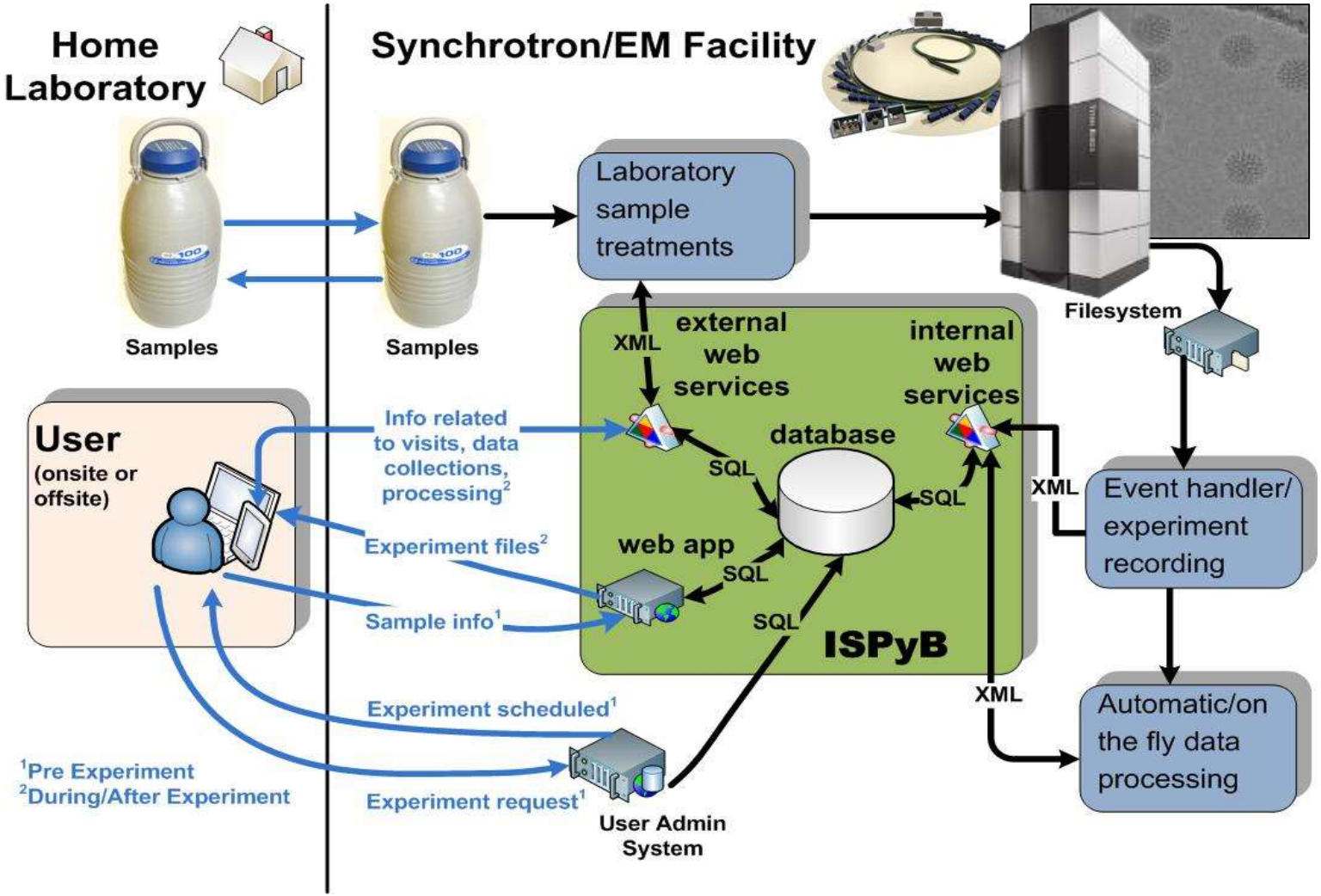
# Lessons learnt

- Persistent services : CFEngine
  - can make sure that a process is always running
  - Scientific computing helped setting it up
- Parallel file system : GPFS
  - Really fast at loading data
  - Does not like polling (we are going to change that bit)
- Workflow : Scipion
  - Designed as a GUI based tool
  - We were able to make modifications to Scipion to run “headless”
  - It’s nice on the inside!

diamond

# Current challenges

- Scipion headless is very ‘fresh’
  - Working on the actual workflow steps
- We would like ISPyB data to richer
- Processing goes wrong:
  - Fiducial Alignment
  - Particle Picking
- More microscopes
  - Data volumes


Home Laboratory
Synchrotron/EM Facility
Samples
Samples
Laboratory sample treatments
Filesystem
XML
external web services
internal web services
database
SQL
SQL
web app
SQL
SQL
ISPyB
XML
Event handler/ experiment recording
XML
Automatic/on the fly data processing
User
(onsite or offsite)
Info related to visits, data collections, processing[2]
Experiment files[2]
Sample info[1]
Experiment scheduled[1]
Experiment request[1]
User Admin System
[1]Pre Experiment
[2]During/After Experiment

# ISPyB Data Structure

Tomography is currently populating:

BLSAMPLEID, SESSIONID
startTime, endTime
ImageDirectory, fileTemplate
numberOfImages
XtalSnapshotFullPath[1-4] – converted to jpeg to display in SynchWeb
axisStart, axisEnd – angles

# ISPyB Data Structure

Single partcle will populate:

**Sample** : name, code, comments (i.e. no additional columns)

**DC group** : sample ID, session ID, experiment type, start & end time (i.e. no additional columns)

**DC - MX/code columns** :

dcg ID
session ID
sample ID
detector ID
imgdir
file_template
xtal_snapshot1
xtal_snapshot2
xtal_snapshot4

starttime
endtime
n_images
exp_time
run_status
resolution
comments

**DC - special EM columns** :

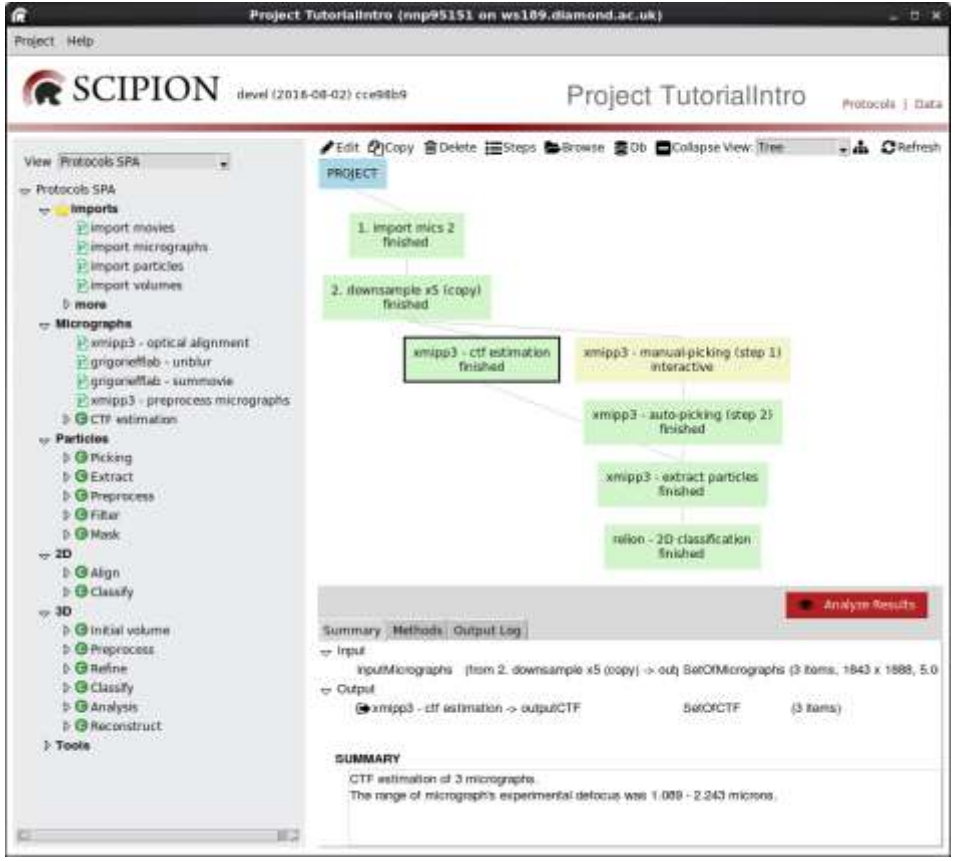
dat_file
magnification (unit: X)
total_absorbed_dose (Unit: e-/A^2 for EM)
binning (1 or 2. Number of pixels to process as 1. (Use mean value.)
particle_diameter (Unit: nm)
box_size_ctf (Unit: pixels)
min_resolution (Unit: A)
min_defocus (Unit: A)
max_defocus (Unit: A)
defocus_step_size (Unit: A)
amount_astigmatism (Unit: A)
extract_size (Unit: pixels)
bg_radius (Unit: nm)
voltage (Unit: kV)
obj_aperture (Unit: um)

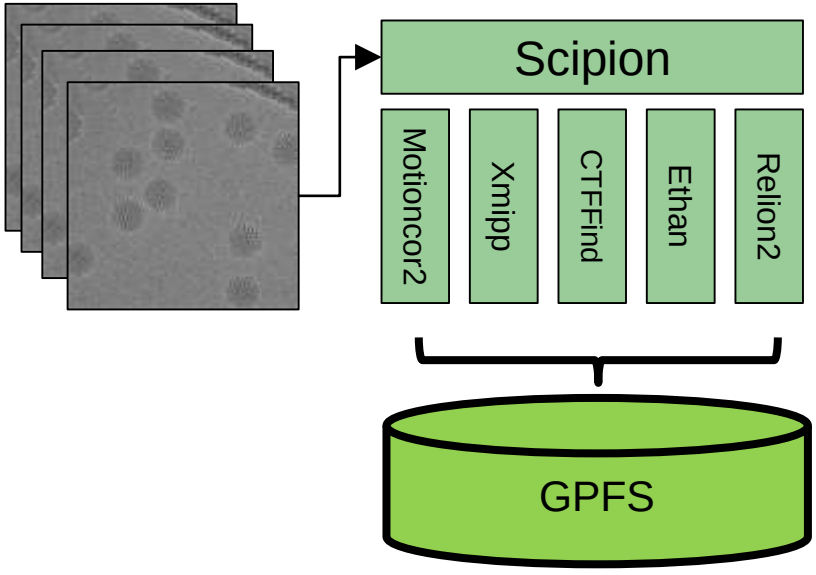
c1aperture (Unit: um)
c2aperture (Unit: um)
c3aperture (Unit: um)
c1lens (Unit: %)
c2lens (Unit: %)
c3lens (Unit: %)

# Potential areas for standardisation

- Metadata structures and API
- Sample and container IDs

# Single Particle Analysis




iNEXT

diamond

# Relion2

- 12 nodes (20 CPU, 256GB RAM, 2xK80 NVIDIA cards, infiniband connections)
  - Standard 3D classification tests:
    - 120 CPUs (6 x com12 nodes): 20:50 hours
    - 20 CPUs, 4 GPUs (1 x com10 node): 4:49 hours


- Currently planning future hardware resources

diamond

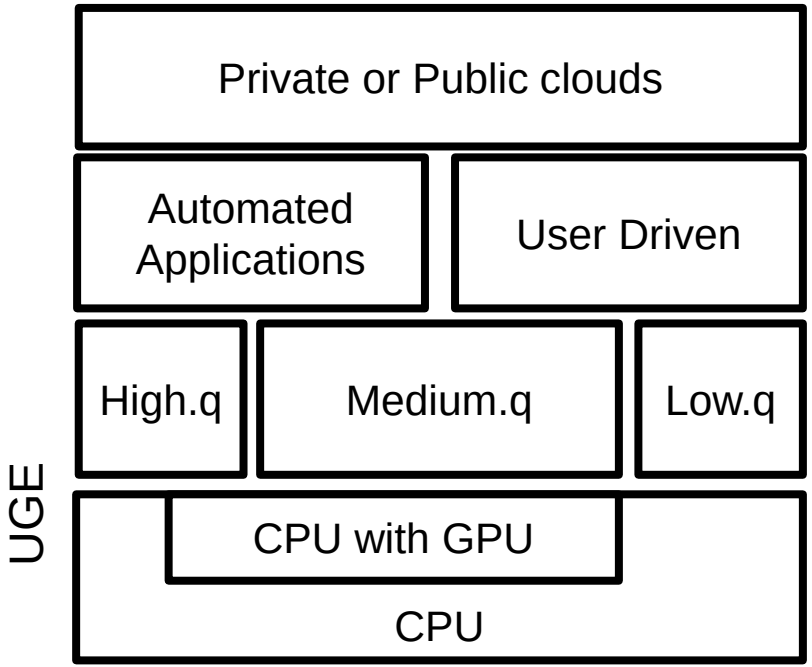
# Cluster Scheduling Software : Job burst out

UGE scalability

- Fair play policy per project
- Quota resources

Burst out tests with Azure

- Manual triggering IMOD reconstruction (incl data transfer) gives comparable results to local cloud




| Resource | Data transfer | Processing |
|---|---|---|
| Beamline Server | | 7m19.9s |
| Azure | 1m24s (1.7GB) | 4m56.9s |
| Cluster node | | 5m49.5s |

diamond

# Cluster Scheduling Software : Job burst out

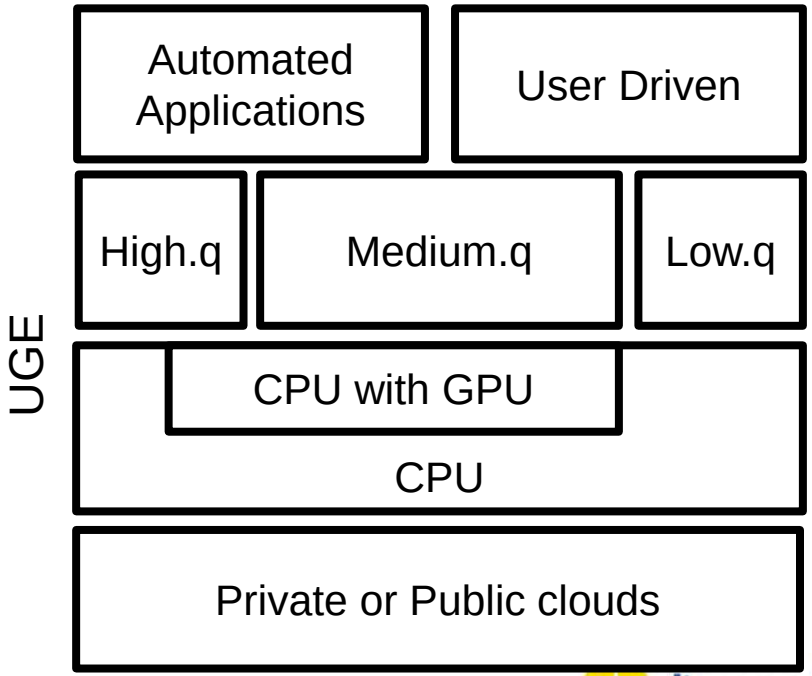
UGE scalability

- Fair play policy per project
- Quota resources

Burst out tests with Azure

- Manual triggering IMOD reconstruction (incl data transfer) gives comparable results to local cloud

| Resource | Data transfer | Processing |
|---|---|---|
| Beamline Server | | 7m19.9s |
| Azure | 1m24s (1.7GB) | 4m56.9s |
| Cluster node | | 5m49.5s |



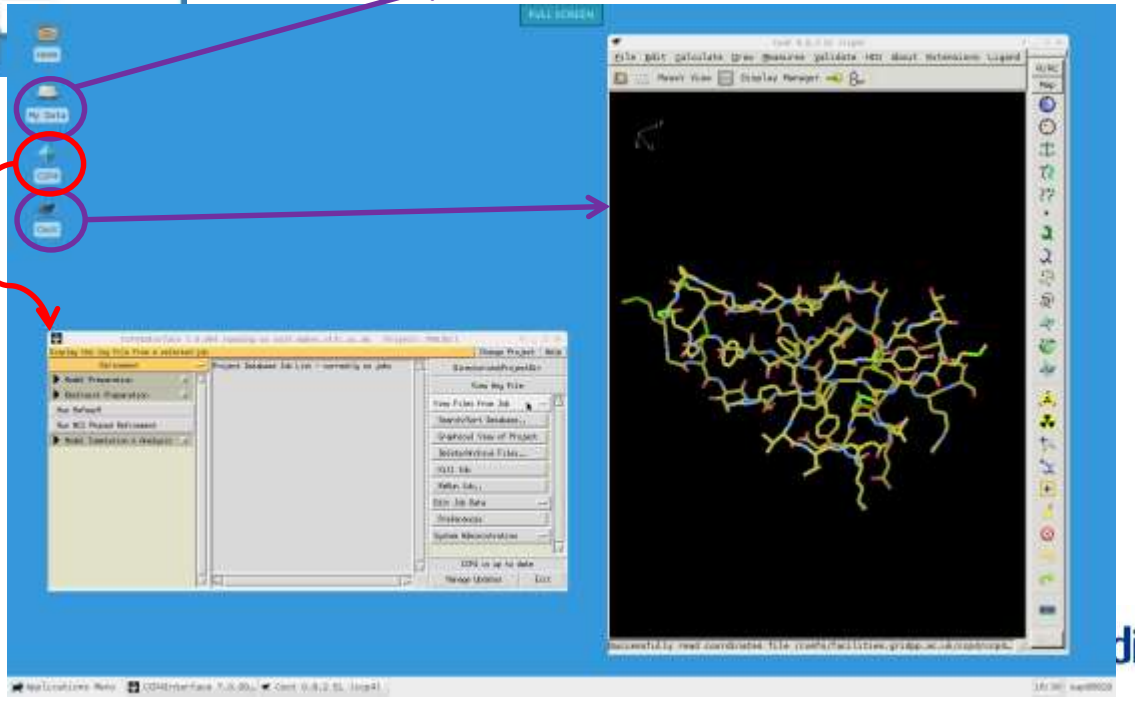

diamond


1. Log onto SCD cloud using fedid & password


4. Browser view of the user's persistent storage (mock-up)


2. Select work environment

3. VM has latest CCP4 & useful software pre-installed, pulled from CVMFS repo



diamond

# Acknowledgements


- Karl Levik, Tina Fredrich, Ala Al-Afeef, Alun Ashton, Dave Stuart &@ Diamond
- Dan Clare, Alistair Siebert, Corey Hecksel, Peijun Zhang at eBIC
- SCIPION team: Jose Miguel de la Rosa Trevin, Jose Maria Carazo from IIPC, Madrid
- Juha Huiskonen from Strubi, Oxford
- CCPEM management and developers: Colin Palmer, Tom Burnley and Martyn Winn


diamond