

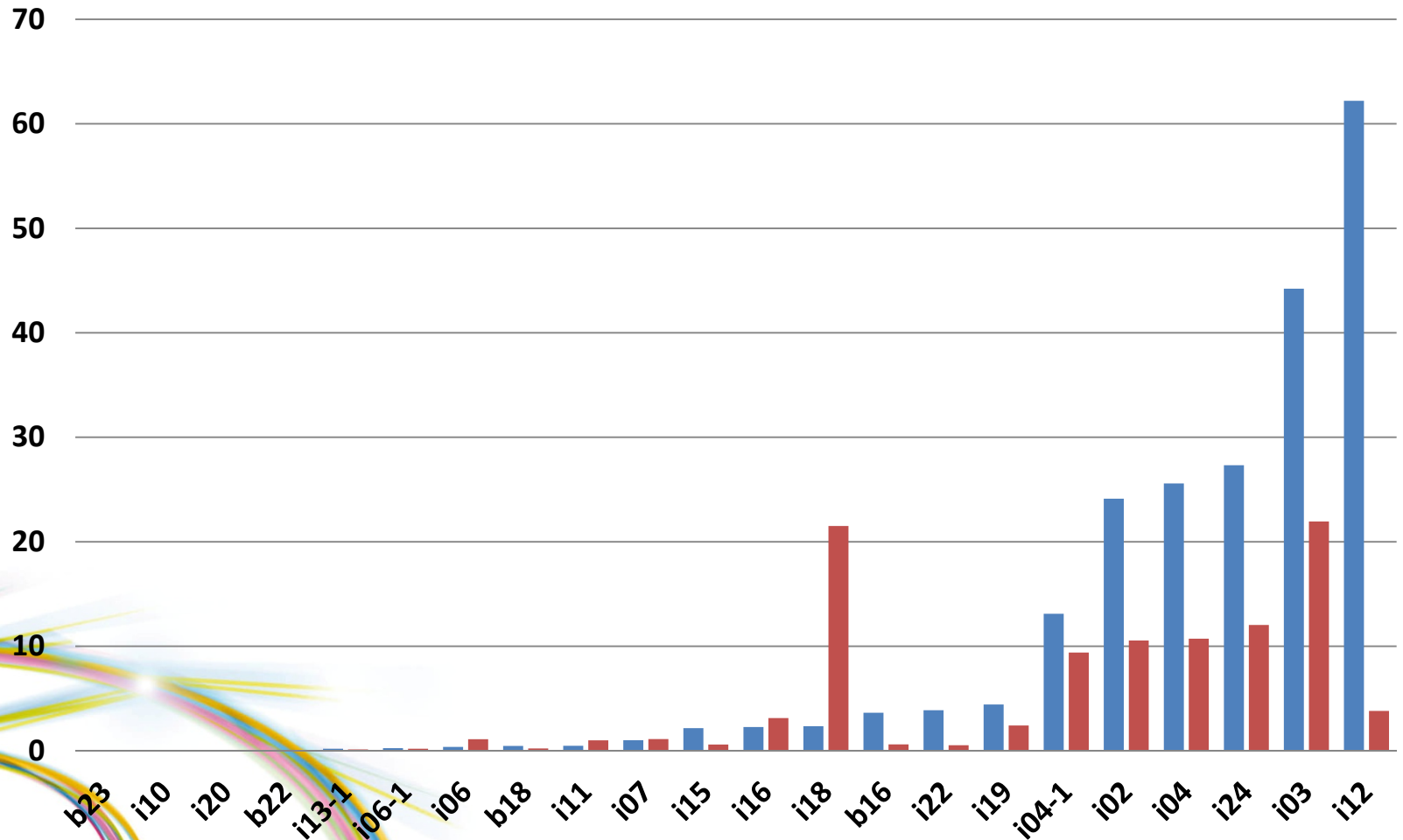
Crystallography and Parallel Data Management

Alun Ashton
Scientific Software Team
Diamond.

With thanks to many groups and collaborators.



Total number of files and data volume per beamline

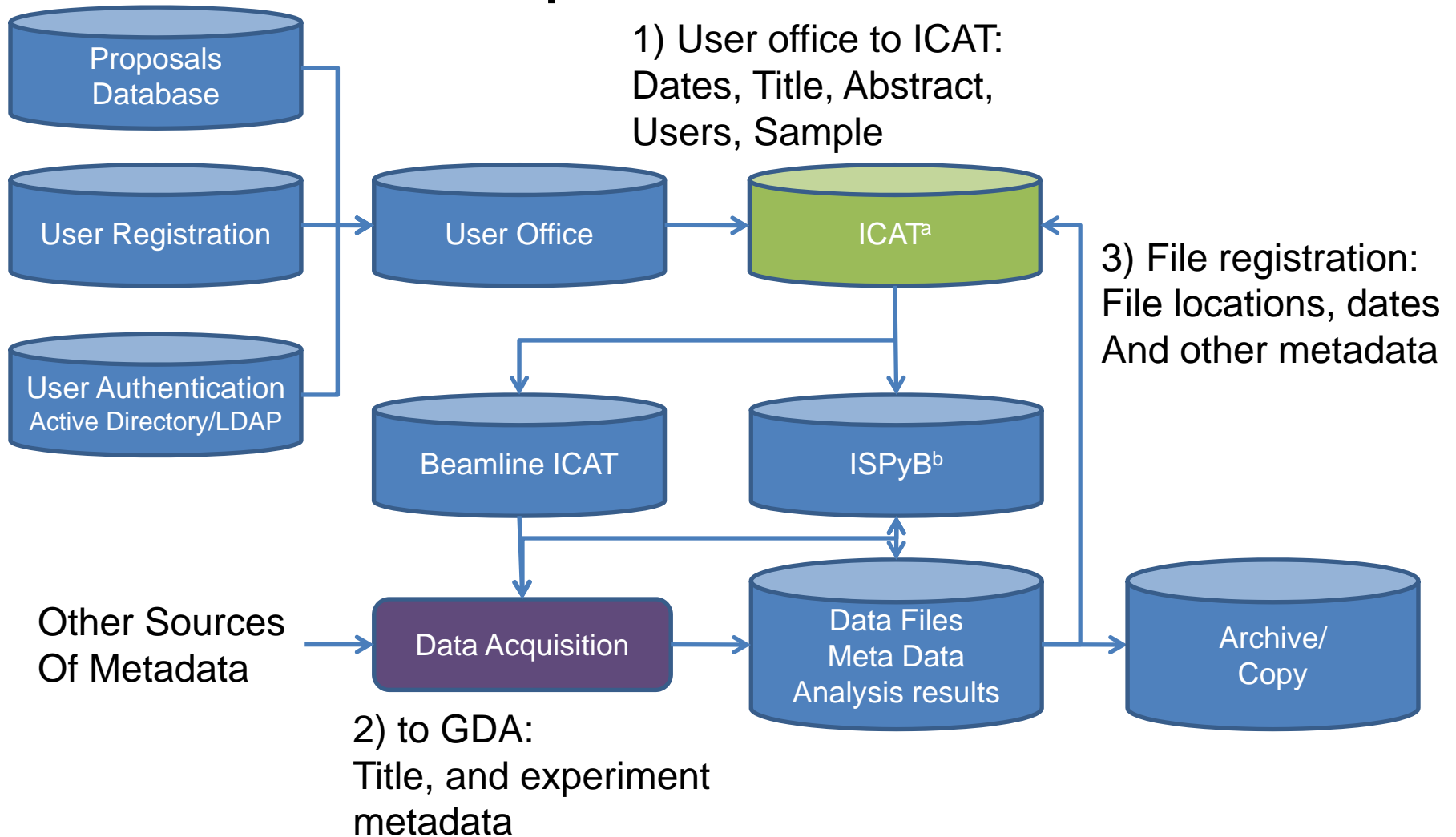


■ Total Size (TB)

■ Number of Files (x1,000,000)



Metadata capture and data archive.



Data analysis and visualisation

IMMEDIATE

(DURING EXPERIMENTS)

“Real time” data processing, analysis and visualisation – to make experimental decisions

Remote access/control

SHORT TERM

(BEFORE THE USER GOES HOME)

Data reduction and processing – Users go home with clean data free of instrument artefacts.

Preliminary data analysis – helpful, but may require significant processing power and know-how

LONG TERM

(AT DLS/USER'S INSTITUTION)

Detailed analysis – from data to information.

Incorporating results from other techniques.

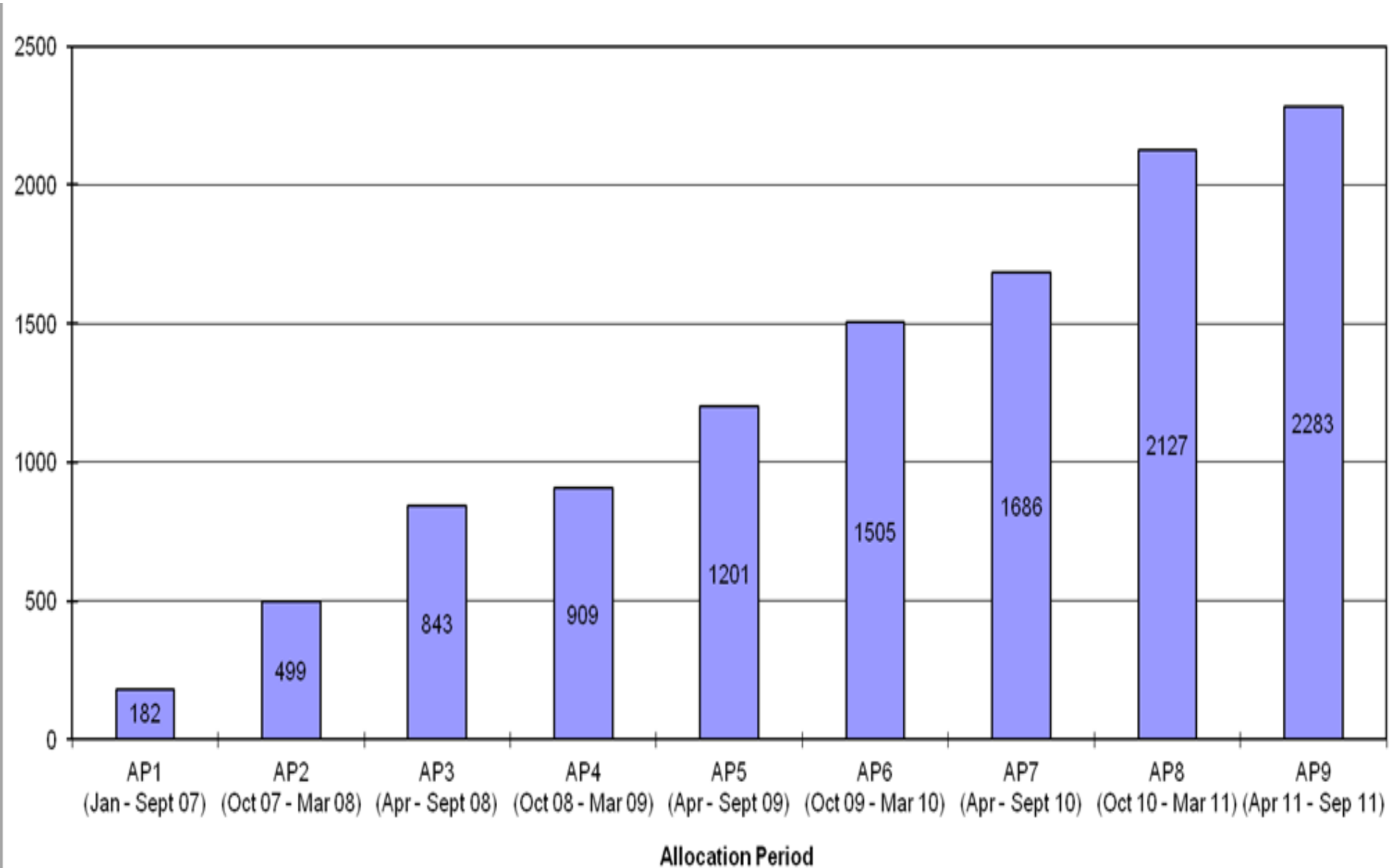
Experiments:

➤ Provide parameters for a model.

➤ Test/verify a model or theory.

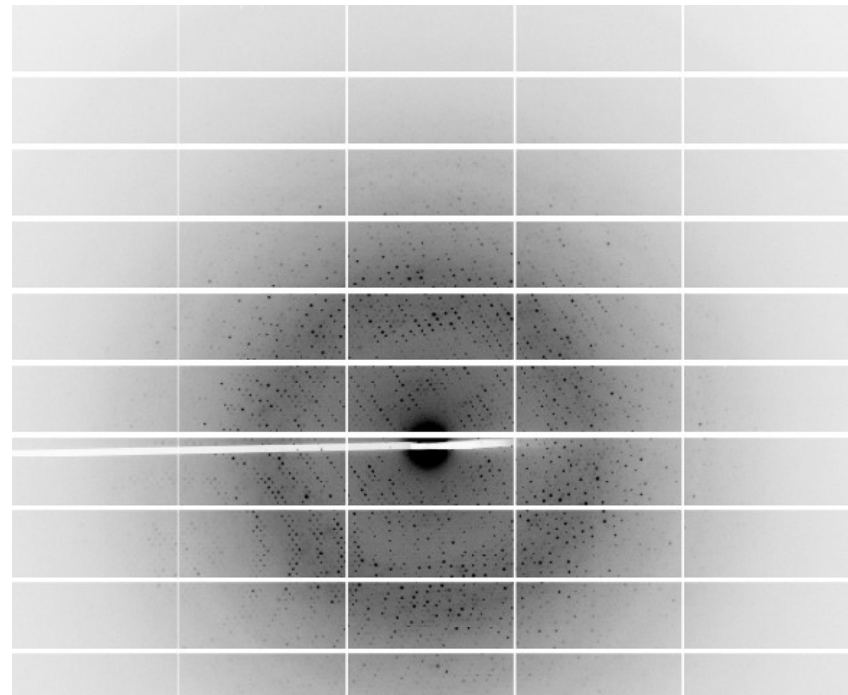
➤ Show where a new theory or model is required.

Number of External Users on Experiments



```
&SRS
SRSRUN=5810,SRSDAT=2012319,SRSTIM=54614,
SRSSTN='ws00',SRSPRJ='i03
',SRSEXP='commissi',
SRSTLE='
',
SRSCN1='          ',SRSCN2='          ',SRSCN3='          '
',
&END
```

SLYCENTRE	Time	qbpm1
-0.91050	0.20	0.24961
-0.80950	0.20	0.24907
-0.70950	0.20	0.25041
-0.60950	0.20	0.25021
-0.50950	0.20	0.24987
-0.40900	0.20	0.25822
-0.31000	0.20	0.26743
0.090500	0.20	0.29894



Analysis and visualisation tools

The screenshot displays a software interface for X-ray diffraction data analysis and visualization. The interface is divided into several panels:

- File Navigator:** Shows a hierarchical tree structure of files and folders. The 'DC' folder is expanded, listing files such as DC_0001.cbf, DC_0002.cbf, DC_0003.cbf, DC_0004.cbf, DC_0005.cbf, DC_0006.cbf, DC_0007.cbf, DC_0008.cbf, and DC_0009.cbf.
- Image Explorer View:** Displays a grid of diffraction images. A blue square highlights a specific region in the top-left image.
- Dataset Plot:** Shows a grid of diffraction images. A green square highlights a specific region in the bottom-left image.
- Header Dataset Plot:** Displays a table of key parameters and their values.
- Side: Dataset:** Contains tabs for 'Dataset Inspect', 'Colour Mappin', 'Line Profile', 'Resolution Ring', 'Image Metadata', 'Peak Profile', and 'Spot V'. The 'Image Metadata' tab is active, showing a 3D visualization of the data.

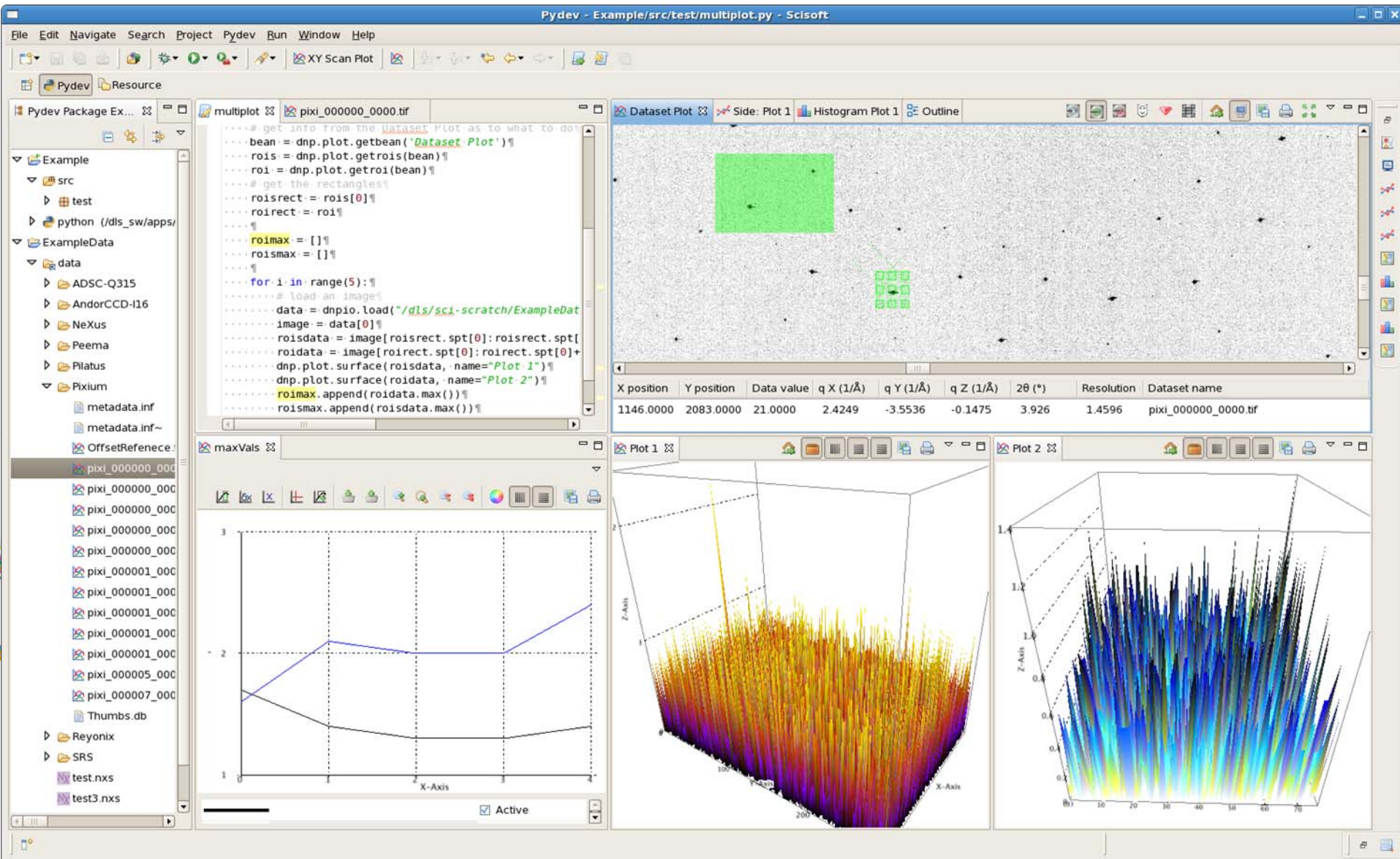
Header Dataset Plot Table:

Key	Value
# Count_cutoff	1220583 counts
# Angle_increment	0.3000 deg.
# Filter_transmission	1.0000
# Silicon sensor, thickness	0.000320 m
numPixels_y	1679
numPixels_x	1475
# Flat_field:	(nil)
# N_excluded_pixels =	255
# Detector_distance	0.22270 m
Unknown 1	# 2011-04-19T04:51:31.047
Unknown 0	# Detector: PILATUS 2M, S/N 24-01...
# Polarization	0.990
Unknown 5	# Image_path: /dls/i04-1/data/2011...
Unknown 4	# Trim_file: p2m0107_E14000_T700...
Unknown 3	# Gain_setting: low gain (vrf = -0.300)
Unknown 2	# Threshold_setting: 7000 eV
# Flux:	0.0000
# Tau =	124.0e-09 s
# Start_angle	90.0000 deg.
# Wavelength	0.91730 Å
# Excluded_pixels:	badpix_mask.tif
# Pixel_size	172e-6 m x 172e-6 m
# Exposure_time	0.9964000 s

Dataset Plot Table:

X position	Y position	Data value	q X (1/Å)	q Y (1/Å)	q Z (1/Å)	2θ (°)	Re
680.5729	738.2269	45.0000	0.2802	0.5349	-0.0267	5.058	10

Analysis and visualisation tools



Analysis and visualisation tools

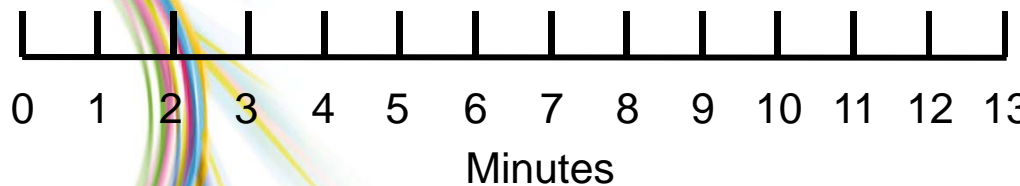
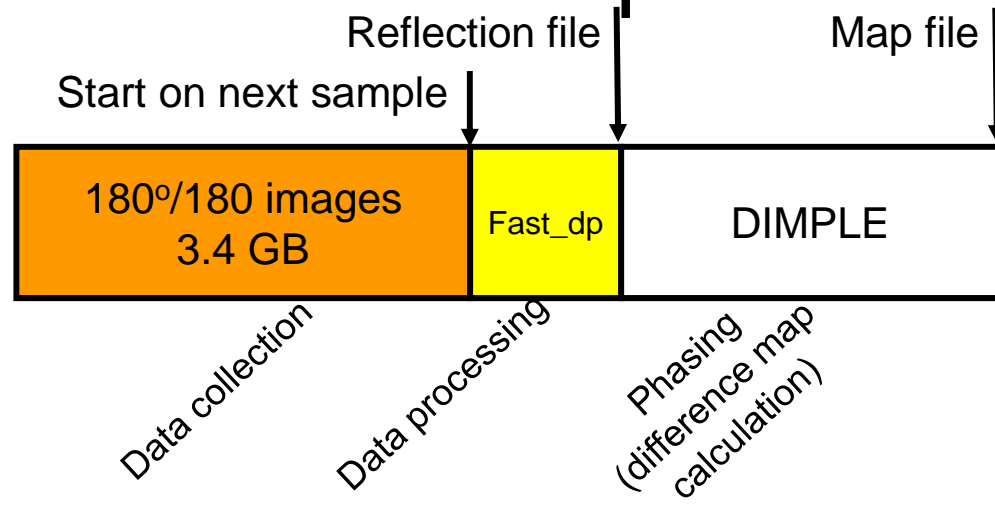
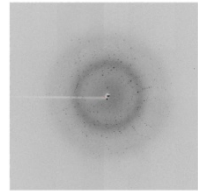
The screenshot displays a workflow editor interface for a file named 'folder_example.moml'. The main workspace shows a data processing pipeline with the following components:

- Input Data:** Three data sources labeled 'dark_0001', 'dark_0002', and 'flat_0001' feed into a 'Median' node.
- Processing:** The 'Median' node's output feeds into a 'Subtract' node. The 'Subtract' node also receives input from a 'results' node.
- Output:** The 'Subtract' node's output feeds into a 'Divide' node, which also receives input from a 'flat_0002' data source.
- Export:** The 'Divide' node's output feeds into a 'Data Export' node.
- Monitoring:** The 'Data Export' node's output feeds into a 'Monitor Directory' node.

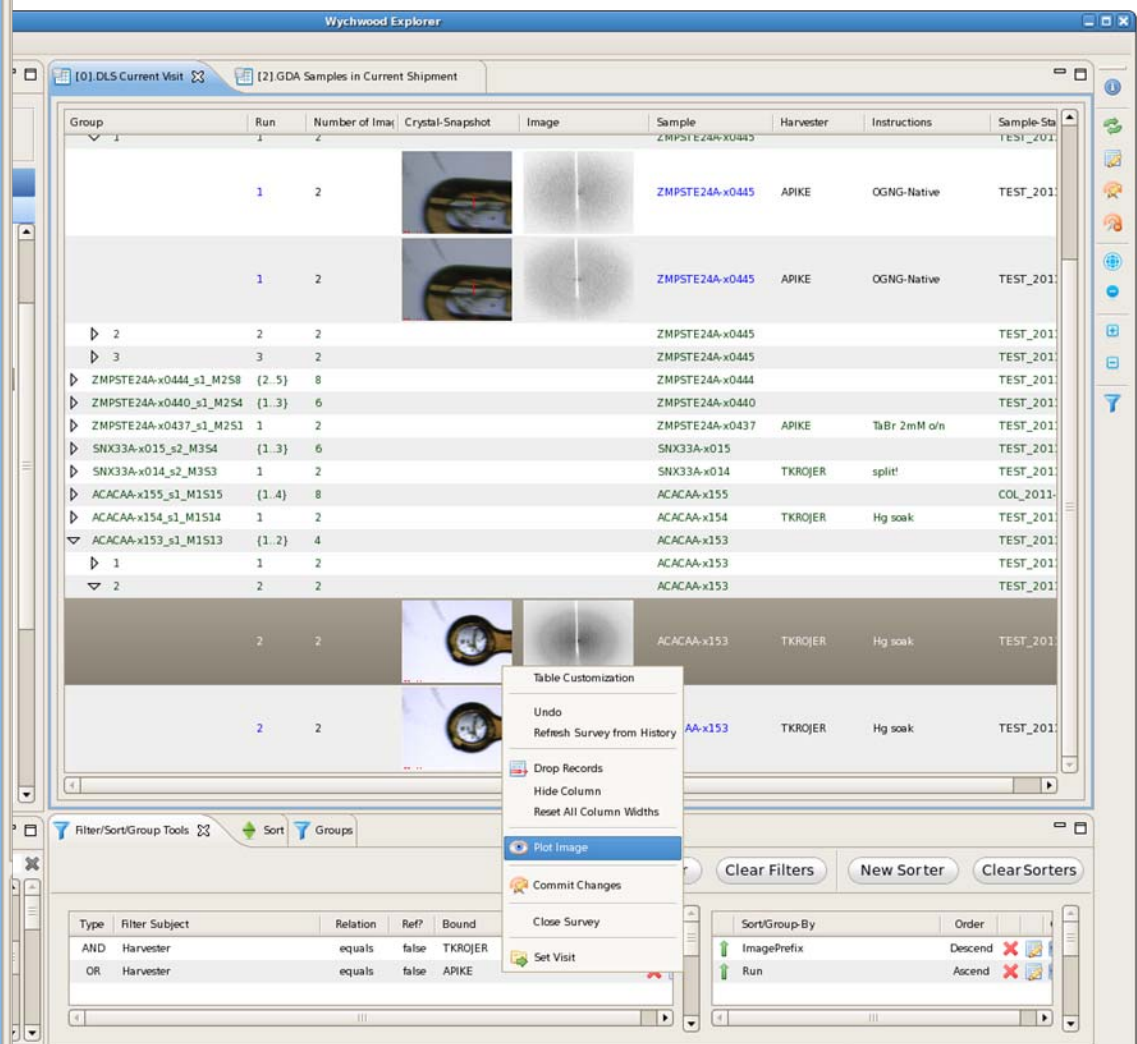
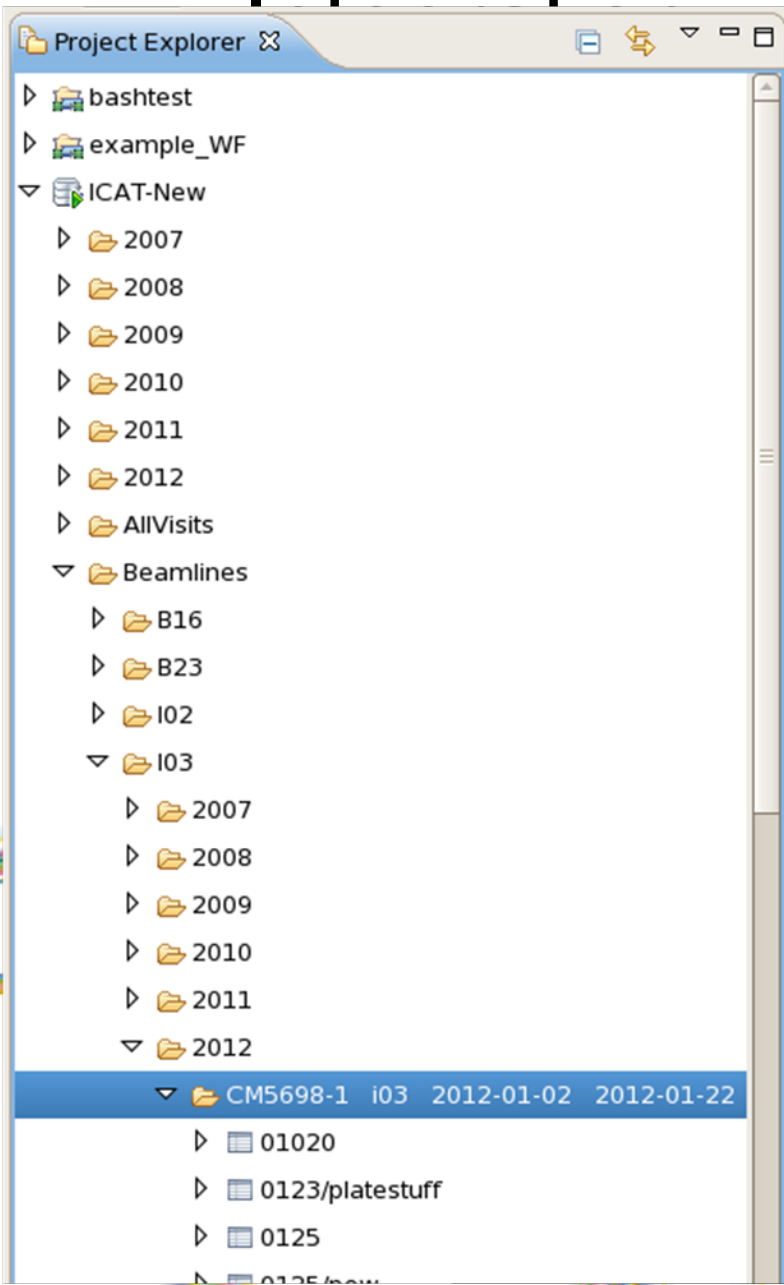
The interface includes a 'Project Explorer' on the left showing a tree of workflow files, a 'Palette' on the left with various tool categories (General, Input/Output Ports, etc.), and a 'Pall' on the right with a search filter and a list of available operations (Add, Subtract, Multiply, etc.). At the bottom, there is a 'Run Edit XML' bar and a 'Console' window displaying the properties of the selected 'Data Export' node.

Property	Value
Type	Data Export
Name	Data Export
Calibration	None
Dataset Name	
Expression Mode	Evaluate on every data input
File Format	tiff (33-bit)
Output	/\${project_name}/output/
Writing Type	Create new file for each evaluation using \${file_name}

Automatic data processing

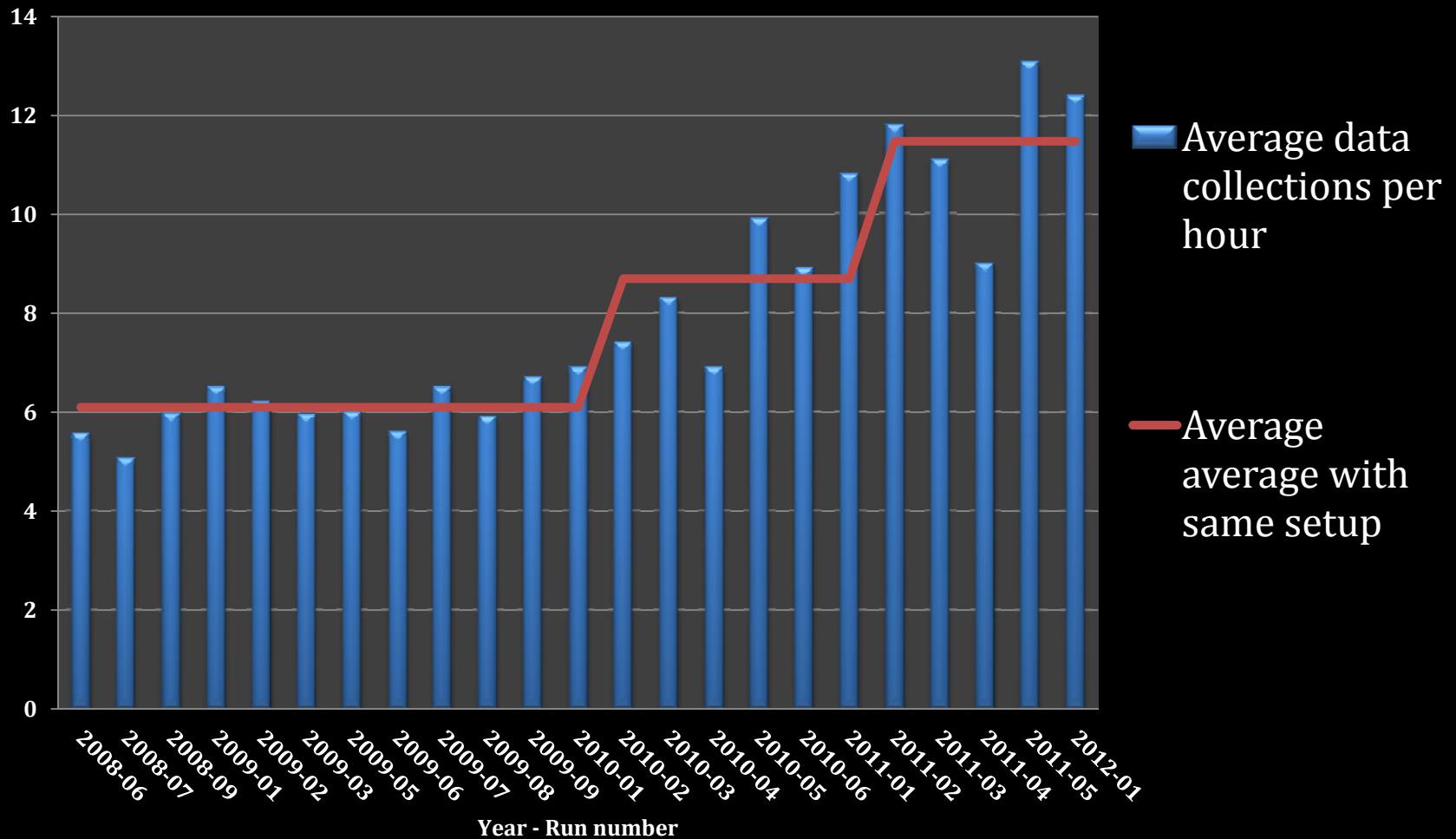


with the data archives.



Increasing beamline 'performance'

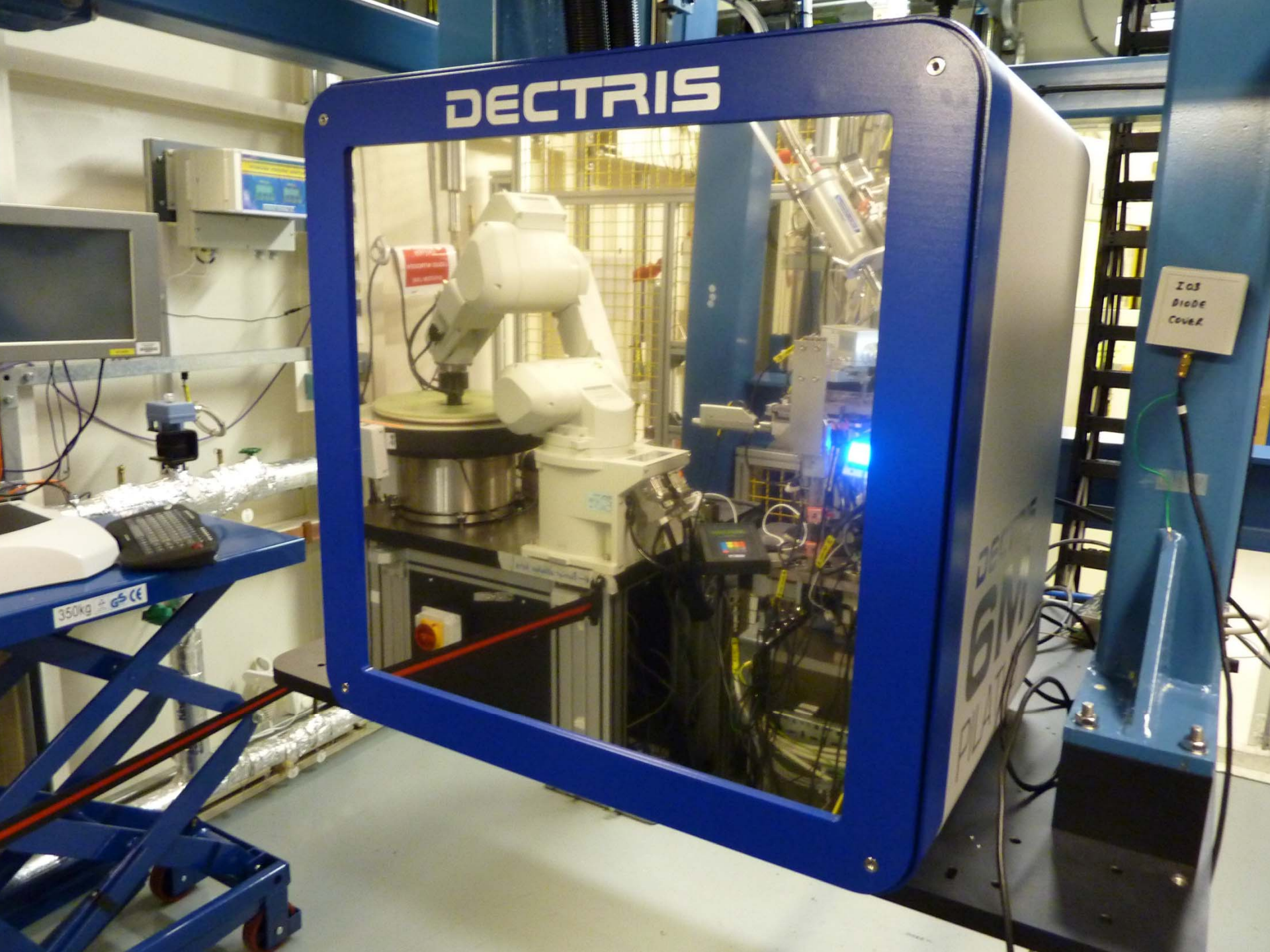
Productivity from 2008 to 2012



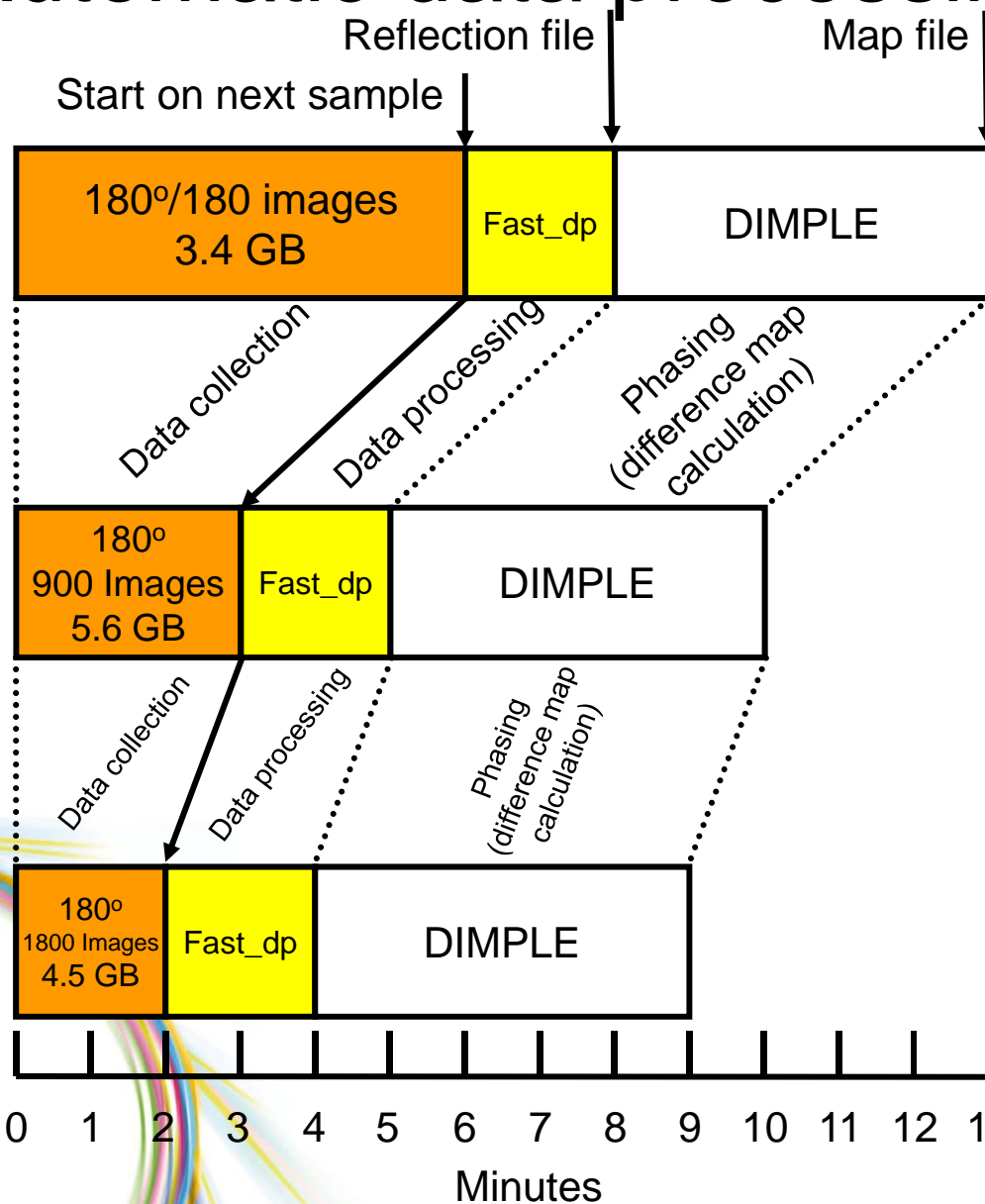
DECTRIS

IOS
BioDE
COVER

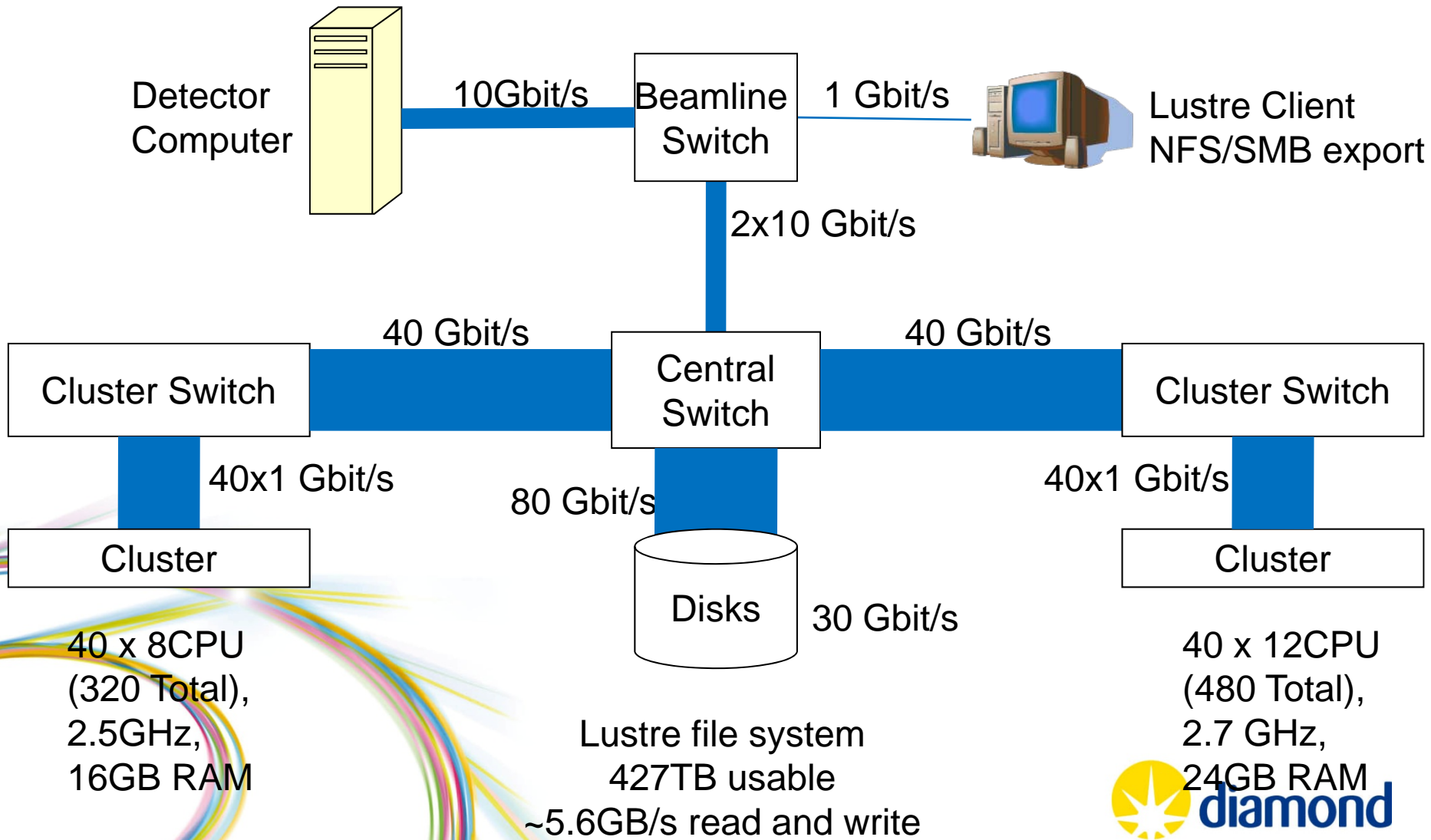
350kg GS CE



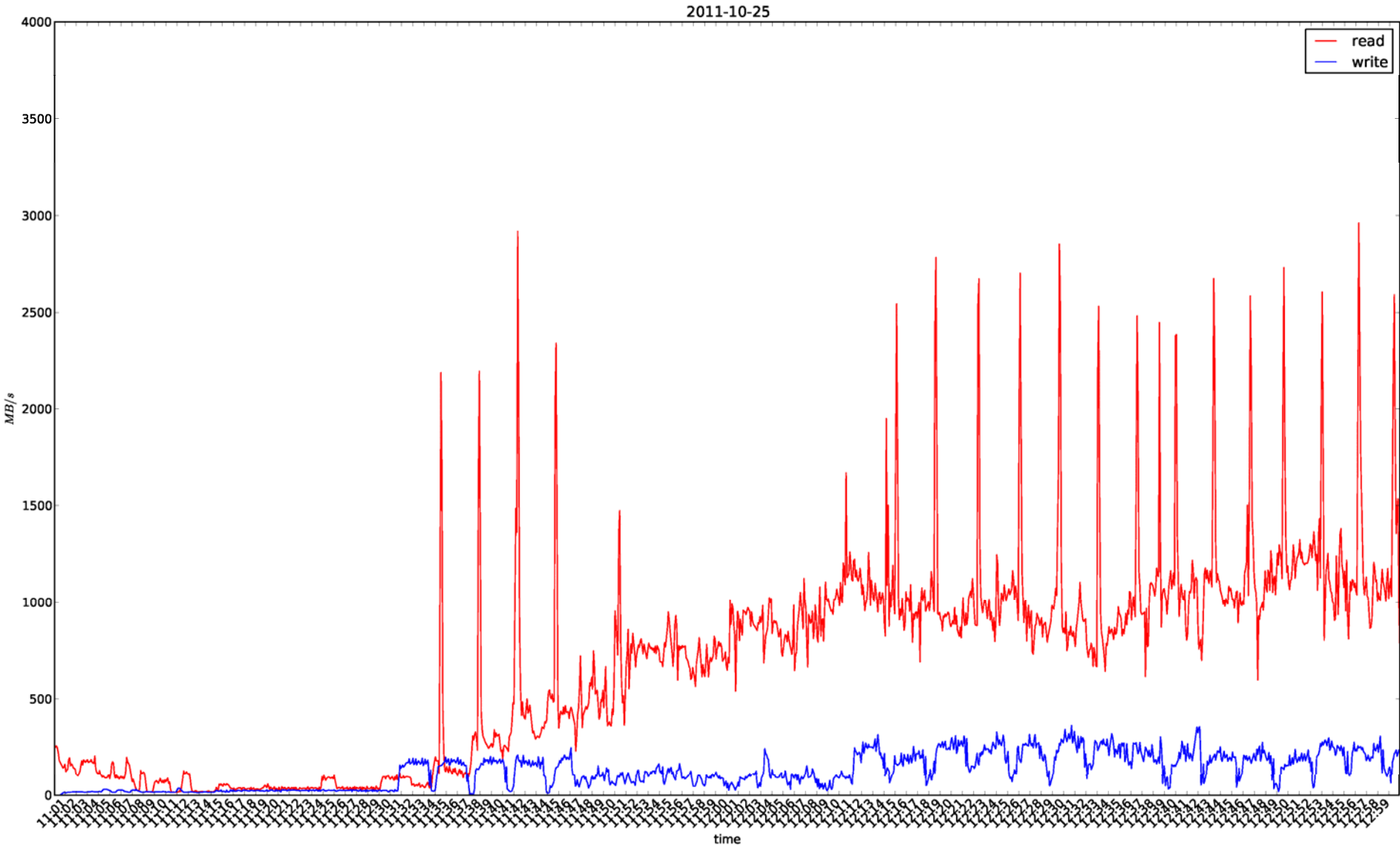
Automatic data processing



'Local' networking.



Data collection and processing data rates



A natural evolution...

The screenshot shows a Linux desktop environment with a terminal window and the CCP4 software interface. The terminal window displays the command `module load ccp4`. The CCP4 GUI shows a 'Data Reduction' section with a table of job status and a 'Directories&ProjectDir' panel on the right.

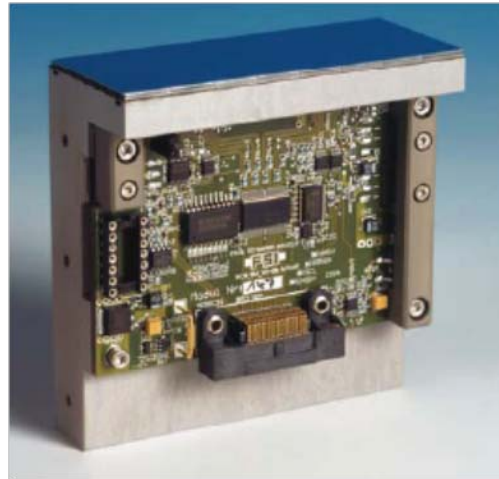
Job ID	Date	Status	Job Name	Parent Job	Title
4	30 Sep 10	FINISHED	edna-mxv1-characterisatio		
3	28 May 10	FAILED	dimple	[No title]	
2	27 May 10	FAILED	dimple	[No title]	
1	26 May 10	FAILED	dimple	[No title]	

The CCP4 GUI also shows a 'Directories&ProjectDir' panel with options like 'View Any File', 'View Files from Job', 'Search/Sort Database..', 'Graphical View of Project', 'Delete/Archive Files..', 'Kill Job', 'ReRun Job..', 'Edit Job Data', 'Preferences', and 'System Administration'. There are also buttons for 'Mail CCP4' and 'Exit'.

Single module

8x2 chips

172 μm x 172 μm pixels



P6M

5x12 modules

~ 43x45 cm²



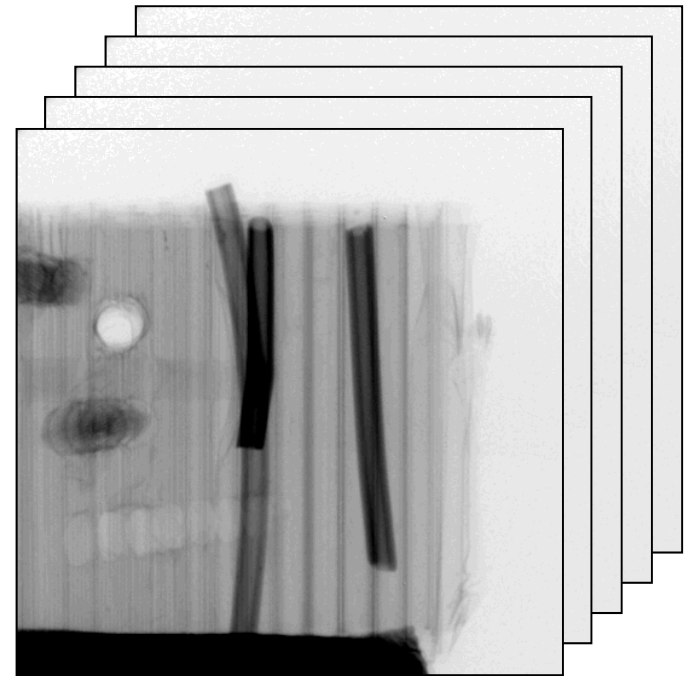
- **Hybrid detector, 60 monolithic silicon sensors,**
- **Pixel size:** 172 μm
- **No. of Pixels:** 2463 x 2527
- **Counter depth:** 20 bits
- **Frame rate:** 12 fps
- **Image size (Tif 32 bits):** ~23 MB
- **Compressed image:** ~ 6 MB
- **Data set:** ~ 1800 images
- **“Volume” of a data set:** 11 GB
- **P6M recently upgraded to 25 fps**
- **Future development P6M to 100 fps**

Future scalability problems with crystallographic analysis

- Sheer number of files and data rates
 - Changing the file formats and structure, no longer perceived as a series of ‘films’
- Increased deployment of ‘longer’ running processing jobs
 - Use more disparate computing and exploit new hardware
- Continued use of legacy code
 - New initiatives to write code and new algorithms

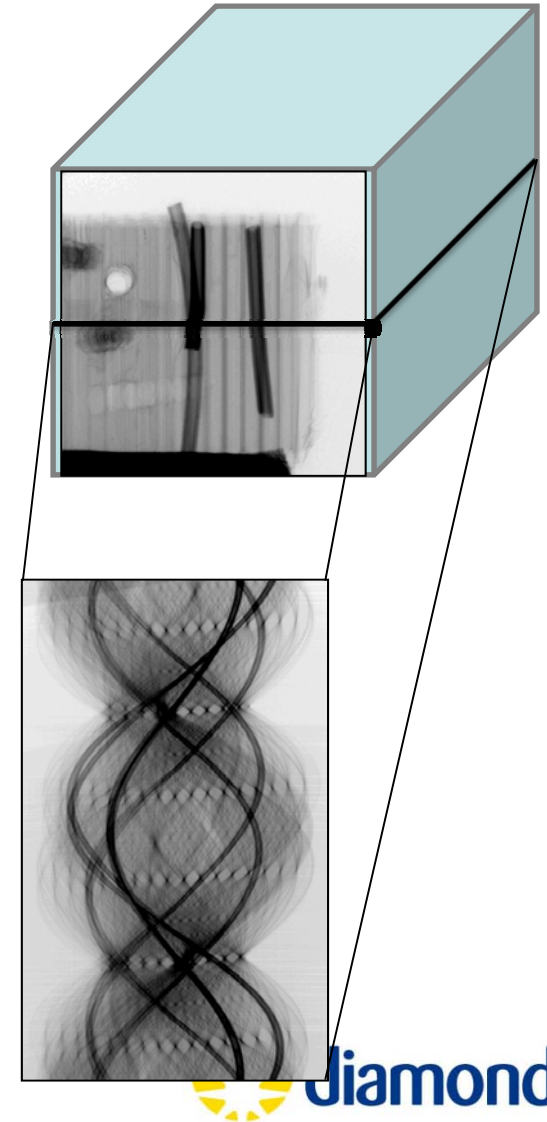
Tomography – Data Collection

- Data is collected straight to HDF5 3D data arrays
 - 4000x2600 pixel 16 bit greyscale image
 - Stored as an [4000,2600,6000] array
 - Totals around 120 GB of data
 - Collection Time can be as fast as 30 minutes.



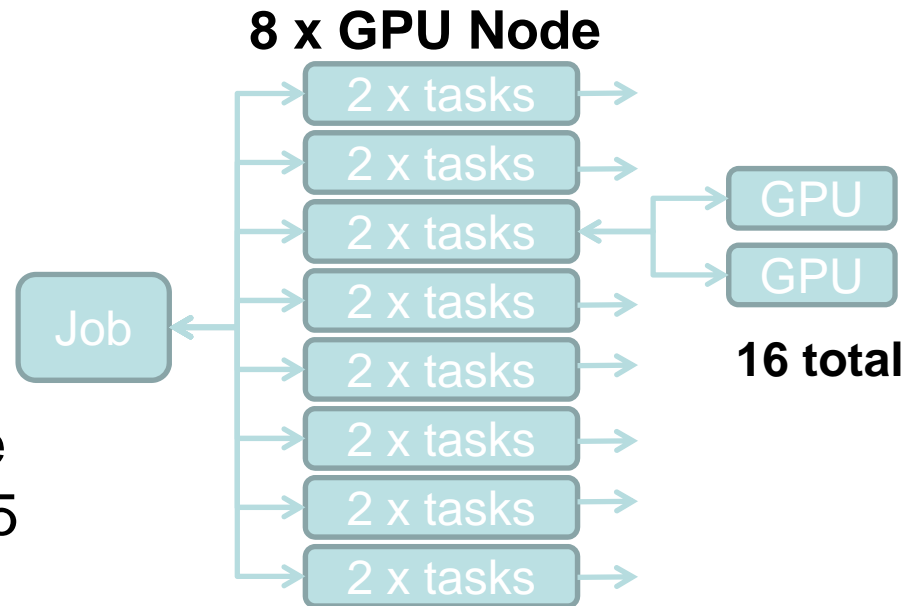
Tomography – Data Slicing

- The data is collected as [4000,2600,1] images
- The Camera software caches images though so that it can be written in [4000,16,8] chunks
- This is so that the data can be read out quickly as [4000,1,6000] sinograms, directly from the file.



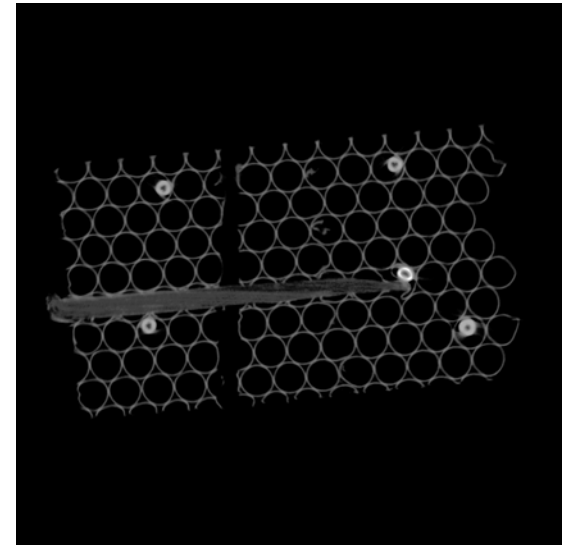
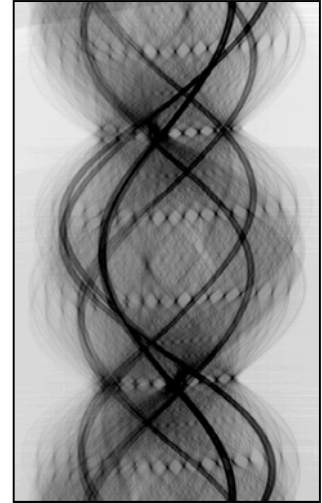
Tomography – Data Extraction

- Data is read out as a stack of sinograms using parallel hdf5 across 8 dedicated GPU cluster nodes, running 2 tasks per machine.
- The bandwidth from the Lustre file system using parallel HDF5 is 25MB/s per Job, meaning a total parallel read speed of about 400MB/s
- We are hoping to push this closer to line speed (50MB/s) with further optimisations.



Tomography – Reconstruction

- Each of the 16 jobs then processes their set of sinograms using one of our GPUs
- Currently takes about 17 seconds per Sinogram, so this totals a reconstruction time of about 45 minutes
- We are currently looking to treble the number of GPU's in this older cluster, and buy a new GPU Cluster to decrease times, and deal with more tomography beamlines coming on line.



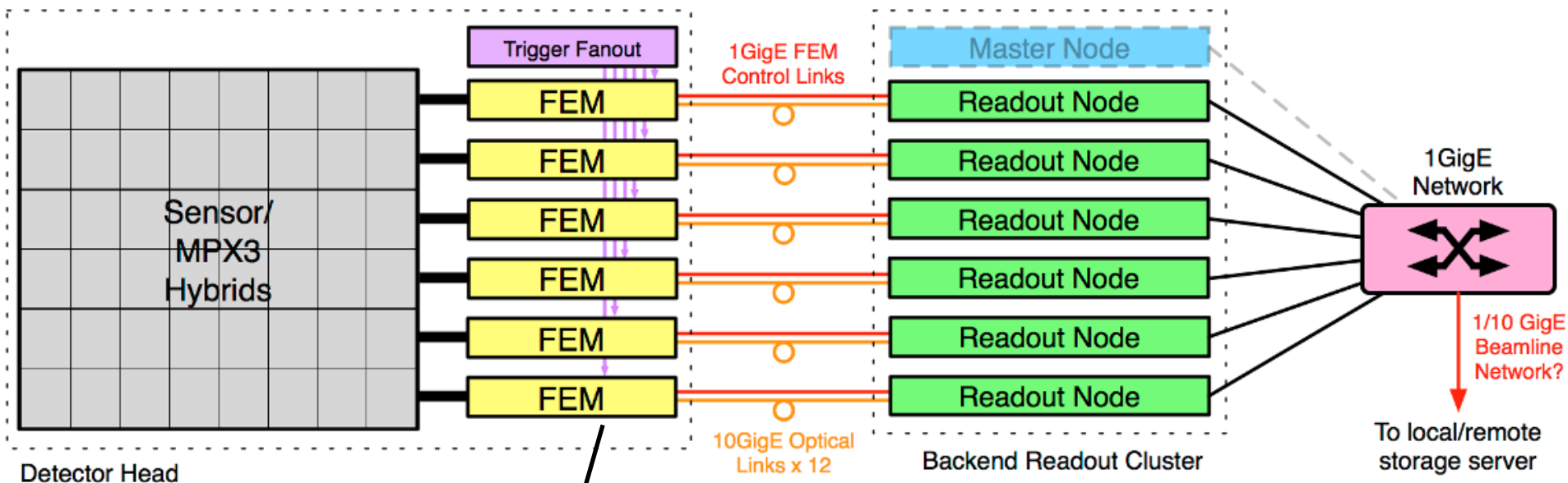
EXCALIBUR

3 modules

3M pixels (11 x 10 cm)

8x2 MPX3/module

1 FPGA card / row of 8 MPX3s



Front-End Module (FEM) FPGA card



Readout backend: 6 Linux nodes

- Buffering, local storage & processing of image data
- Interface to EPICS for DAQ & control

Conclusions.

Both during and after the 'beamtime':

- Data rates are increasing from larger or faster detectors.
- Data rates are increasing from beamline stability and performance.
- Data rates are increasing from ever growing automated processing, analysis, visualisation expectations.
- Structured hierarchical parallel well R/W file formats and structures can help to some extent with the data avalanche.
- Multiple data streams and keeping data in memory will help.
- Investment in new algorithms and exploitation of new hardware.