Practical approach to privacy, access, and processing for biological synchrotron experiments

Keith Brister
Northwestern University

MICHIGAN STATE UNIVERSITY

THE UNIVERSITY OF MICHIGAN 1817

Van Andel Institute

WAYNE STATE UNIVERSITY

THE UNIVERSITY of WISCONSIN

ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

NORTHWESTERN UNIVERSITY

VANDERBILT UNIVERSITY

Life Sciences Collaborative Access Team

GRAND VALLEY STATE UNIVERSITY

MEDICAL COLLEGE OF WISCONSIN

Argonne NATIONAL LABORATORY

gsk GlaxoSmithKline

deCODE biostructures

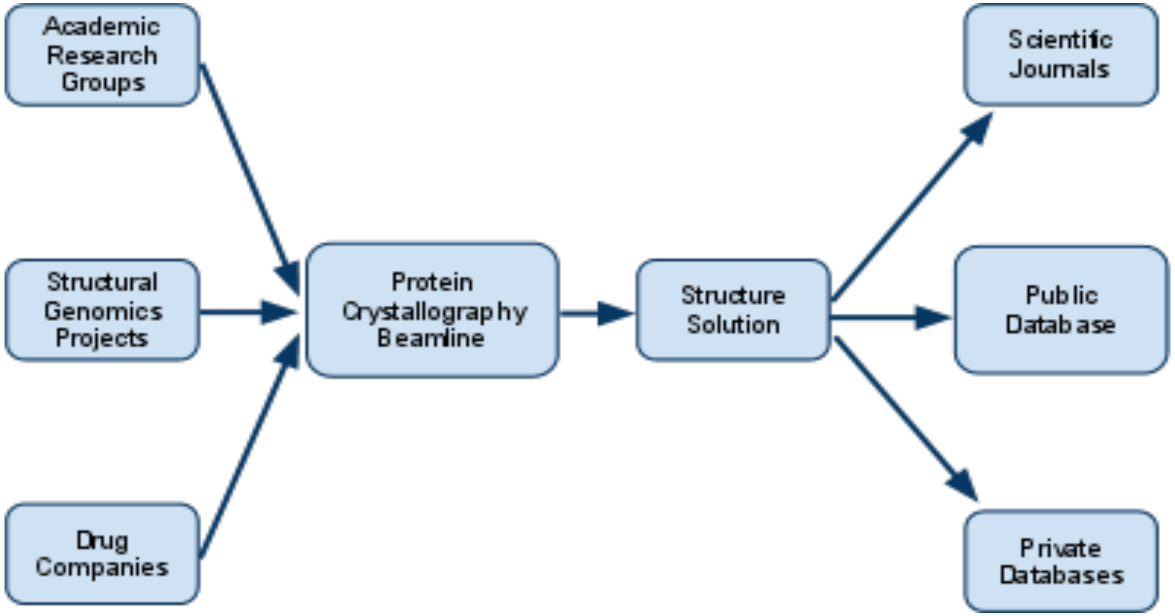
Summary

- Introduction to Protein Crystallography Beamlines
- LS-CAT approach to privacy, access, and processing
- What grid computing might bring to the table
- What a grid computing proposal might look like

Protein crystallography beamlines

Websites:
http://biosync.rutgers.edu
http://www.pdb.org

27 synchrotron radiation facilities world wide with 143 beamlines run by 63 different organizations



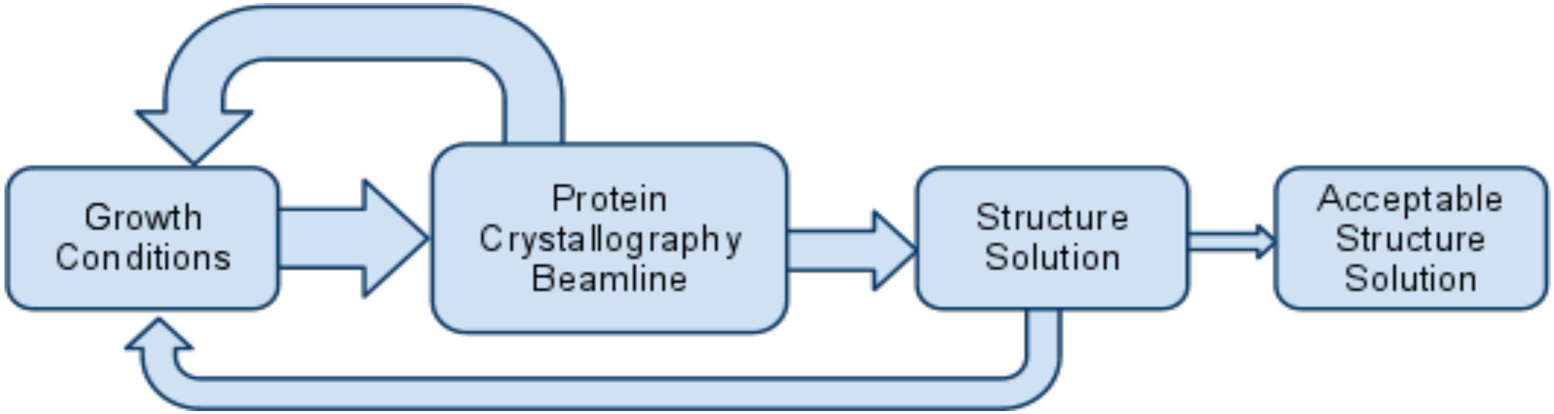
>63,000 entries
>6,000 new/year

Protein structures: an iterative process

Most (~99%) of the data collected do not contribute directly to a structure determination. Instead, the quality of the crystals found is used to alter the crystal growth conditions to attain better crystals.

That is, only about 1 in 10 crystals yield diffraction worth mentioning and only a small percentage of these lead to publishable structure factors.

Most of the job is in finding the good crystals with the advantage going to the groups that have frequent, inexpensive, access to PX beamlines.


Growth Conditions
Protein Crystallography Beamline
Structure Solution
Acceptable Structure Solution

Evolution of data collection

Film (until late 80's / early 90's)

Crystals characterized at home source and aligned optically
PI leads the 10-12 people needed at the beamline

Image Plates (through the 90's)

Crystals characterized at home source
PI or Postdoc leads group of 4 at beamline
Data rate: 200 MB/hour limited by detector readout

CCD Detectors (mid 90's to present)

No home source
Postdoc or senior grad leads group of 2 or 3 (or remote only)
Data rate: 20 GB/hour limited by detector readout

Pixel Array Detectors (Starting to be used now)

Beamline technician only at beamline
Data rate: 200 GB/hour: limited by sample changers

Evolution (continued)

The trend is fewer people actually coming to the beamlines and those that do have less experience.

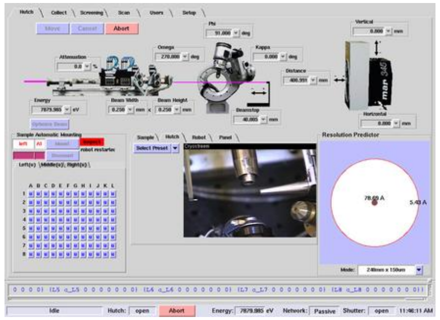
As x-ray experiments have become routine, the user groups concentrate on harder biological problems. The result is that more crystallographic expertise is expected to reside with the beamlines or with collaborators. This trend is likely to continue.

Remote Control of Experiments

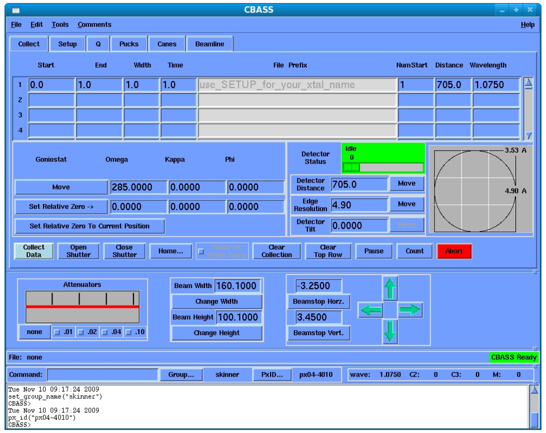
Due to the large differences in control systems and philosophies, developments by any one of the 60+ beamline operators are rarely usable directly by any of the others. However, concepts and ideas are shared: there is a fair bit of innovation as beamlines implement capabilities originating with another beamline.

A number of methods of remote access have been tried by various beamlines:

- Interface to existing beamline controls
  - VNC: too slow
  - Access Grid: too complicated and too slow
  - NX: acceptable performance, used by many beamlines



- Web based controls
  - Java client (NSLS)
  - Adobe Flex client (LS-CAT)



	Start	End	Width	Time	File Prefix	NumStart	Distance	Wavelength
1	0.0	1.0	1.0	1.0	use_SETUP_for_your_xtal_name	1	705.0	1.0750
2								
3								
4								

Goniostat	Omega	Kappa	Phi
Move	285.0000	0.0000	0.0000
Set Relative Zero ->	0.0000	0.0000	0.0000

Set Relative Zero To Current Position

Detector Status: Idle 0
Detector Distance: 705.0
Edge Resolution: 4.90
Detector Tilt: 0.0000

Beam Width: 160.1000
Beam Height: 100.1000
-3.2500 Beamstop Horz.
3.4500 Beamstop Vert.

File: none
CBASS Ready

Command: skinner px04-4010 wave: 1.0750 C2: 0 C3: 0 M: 0

```
Tue Nov 10 09:17:24 2009
set_group_name("skinner")
CBASS>
Tue Nov 10 09:17:24 2009
px_id("px04-4010")
CBASS>
```

Practical Considerations

Beamlines work, typically, with hundreds of researchers from dozens of institutions on a shoestring budget. Neither the beamlines nor the researchers can afford complicated or time consuming steps to gain access to a remote interface.

X-Ray diffraction is a small part of the entire experiment. We cannot insist that users jump through too many hoops.

In particular:

- SSL client certificate administration is too complicated and too time consuming to be of use. (No Access Grid)
- IT support can not be assumed. (No special ports)
- Software installations are going to be done by non-experts, perhaps with no root access. (Youtube video instructions needed)

Privacy

Although we'd like to extend our Lustre file system to our member labs, the current Lustre distribution does not support the required security features. We'd like to see Kerberos support for both users and clients.

In the mean time, we require:

- No direct file system access by publicly accessible machines
- https connections for the user interface, ssh for data
- No shells or user submitted scripts allowed
  - We cannot trust the person at the other end of the connection
- All data are treated as confidential: access control maintained though UNIX uid/gid management.
- We are not storing personal information beyond phone numbers and email addresses.
- Data, even proprietary data do not mean much without knowledge of the protein (which is not stored on our systems): Confidential does not mean top secret.

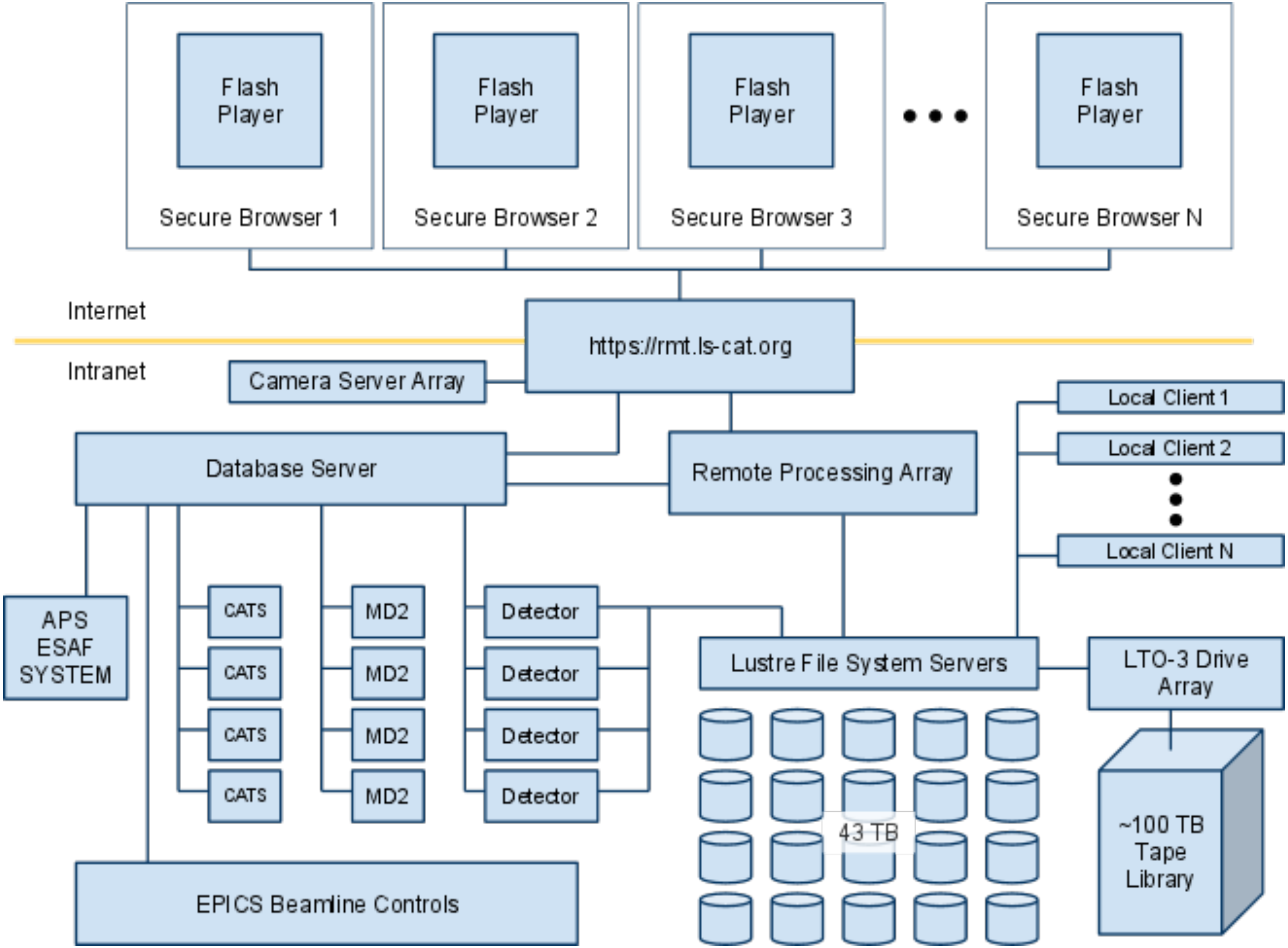
Access

- All experiments at the APS require an on-line Experimental Safety Assessment Form (ESAF). This form identifies all users involved with the experiment.

- The ESAFs for our beamlines must be approved by LS-CAT before the experiment can start. We use this approval to automatically generate computer accounts for the experiment and for each user involved with the experiment.
  - A user has access only to data for which he or she has been named on the corresponding ESAF

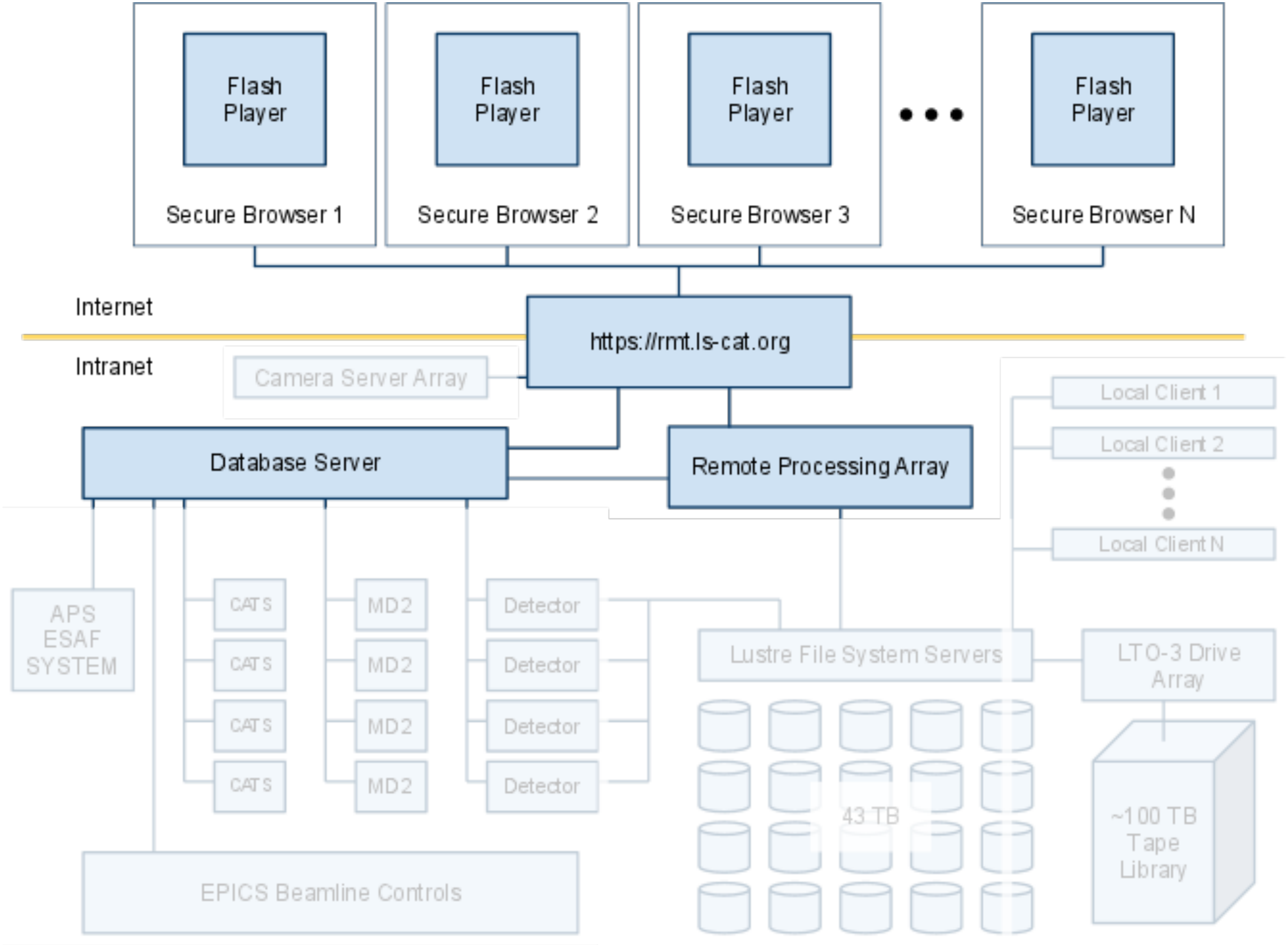
- Although not fool proof, this method combines automatic system administration with staff oversight (the approval).

- New users can gain access within minutes of making themselves known to the system.



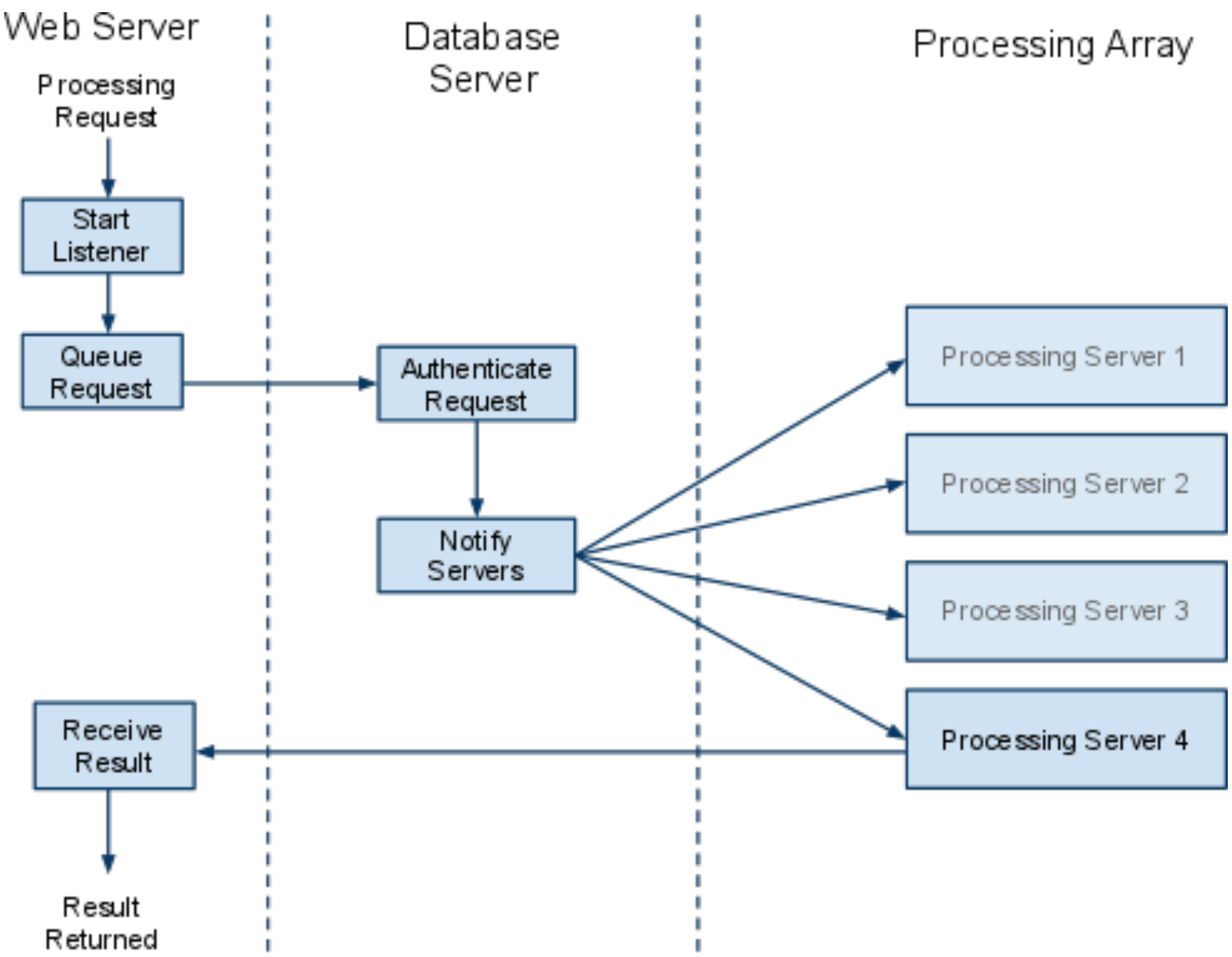
LS-CAT Computing Architecture

Keith Brister OSG All Hands March 9, 2010



LS-CAT Computing Architecture

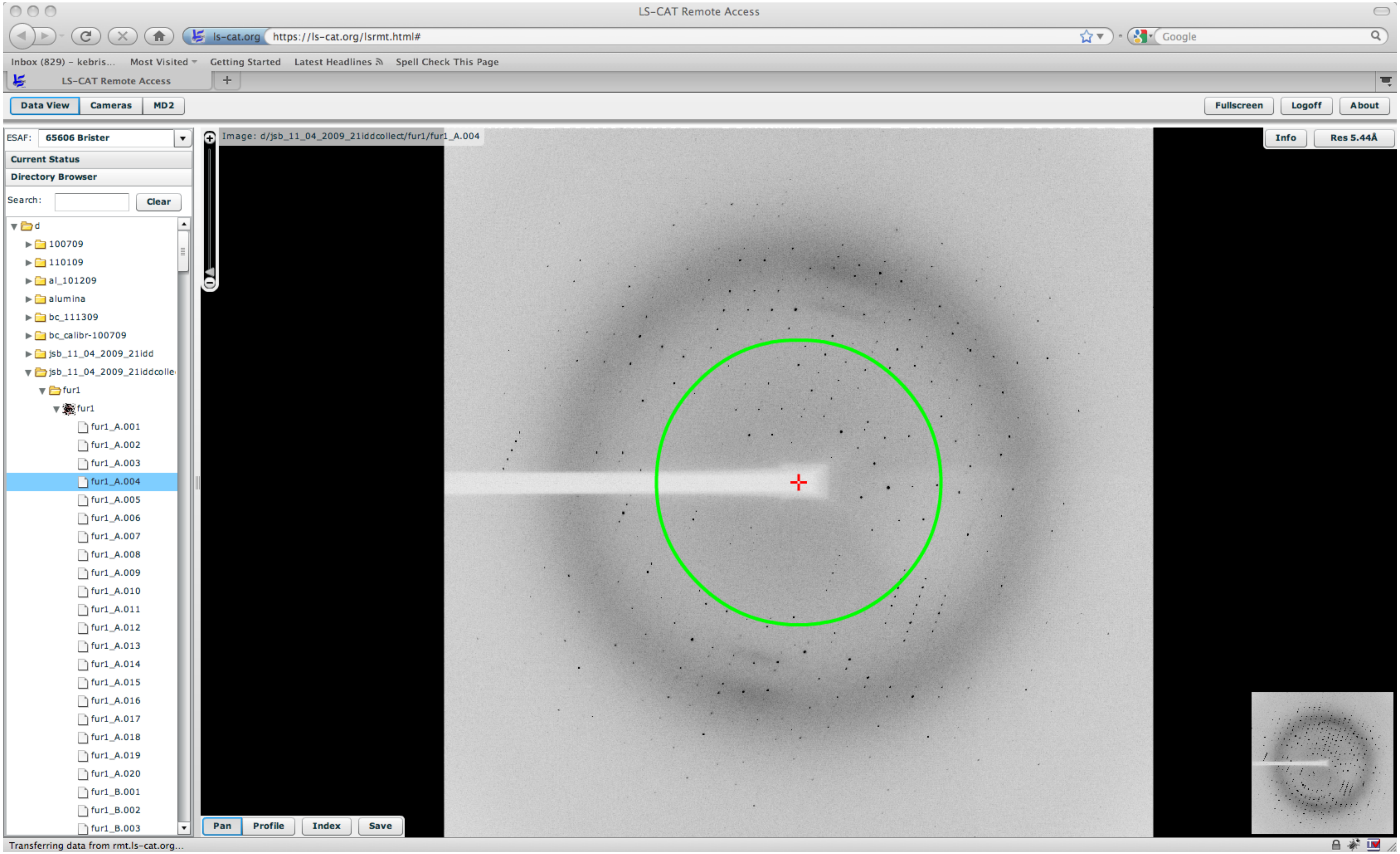
Keith Brister OSG All Hands March 9, 2010



Requests run predefined scripts requiring access to the raw images. Results are returned as a single stream of data.

All processing servers are notified of every request, the first to respond gets the work.

Since all processing servers share the same file system, it's possible to implement predictive processing.



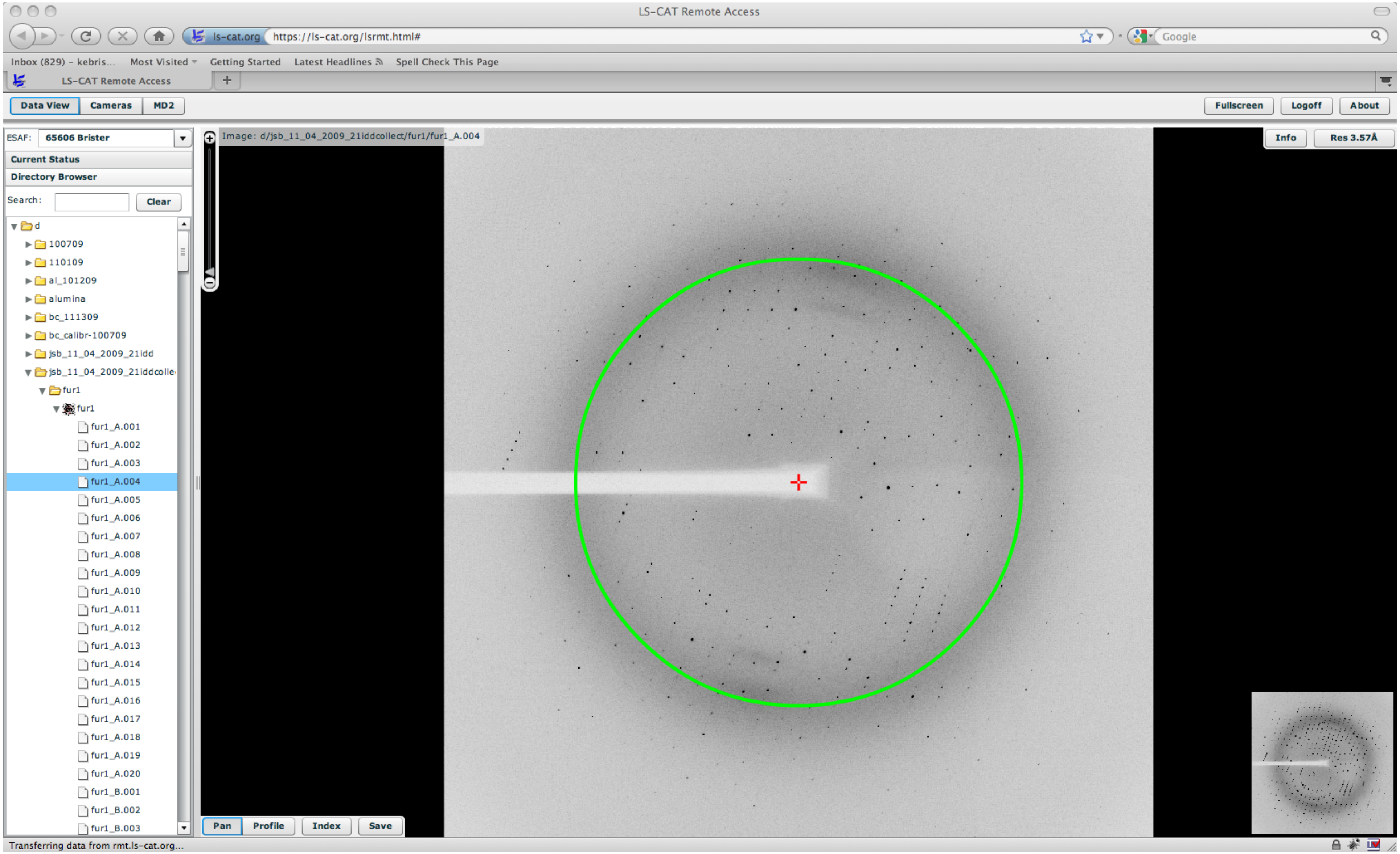
https://ls-cat.org/lsrmt.html#

ESAF: 65606 Brister

Image: d/jsb_11_04_2009_21iddcollect/fur1/fur1_A.004

Res 5.44Å

Transferring data from rmt.ls-cat.org...



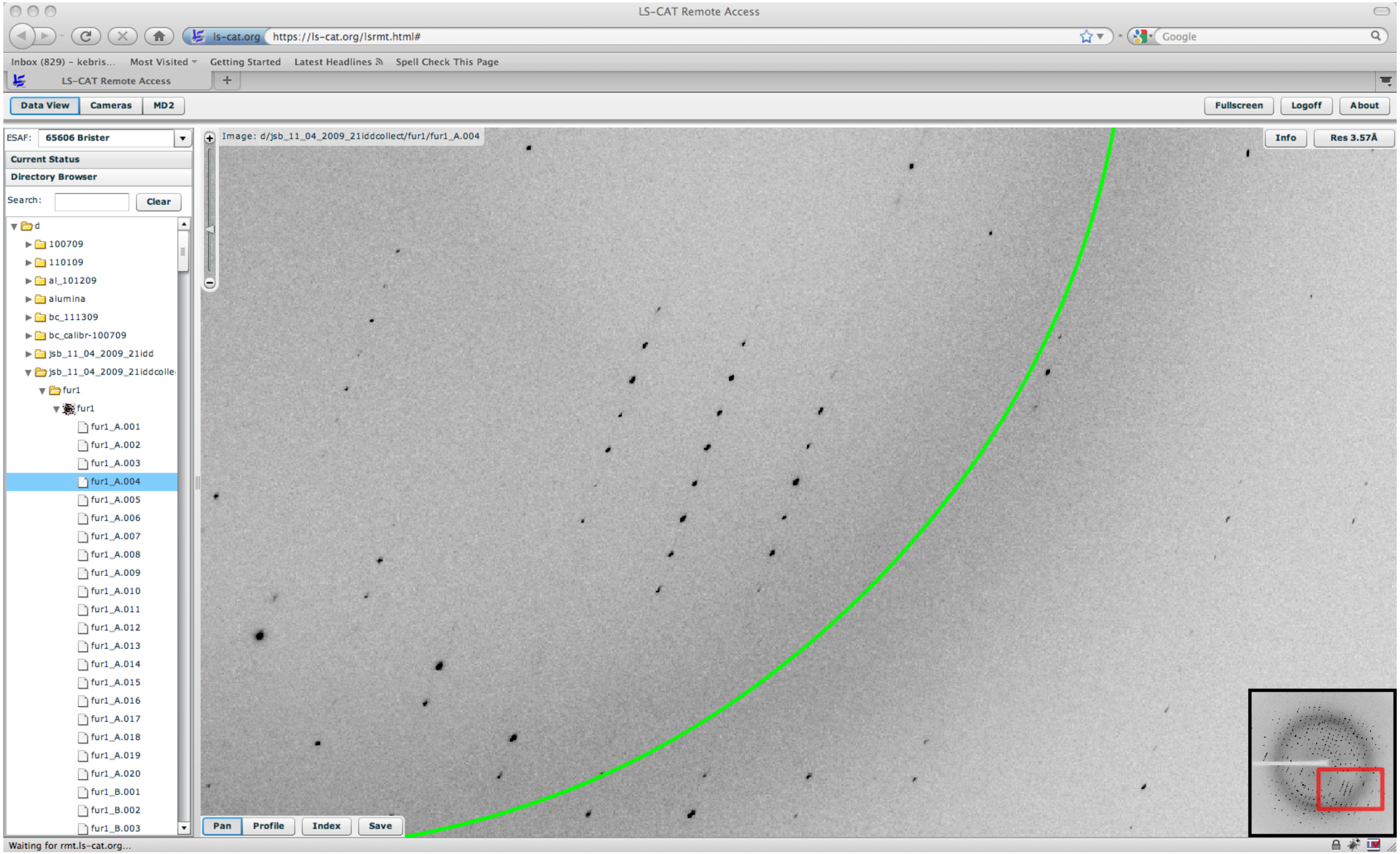
ESAF: 65606 Brister

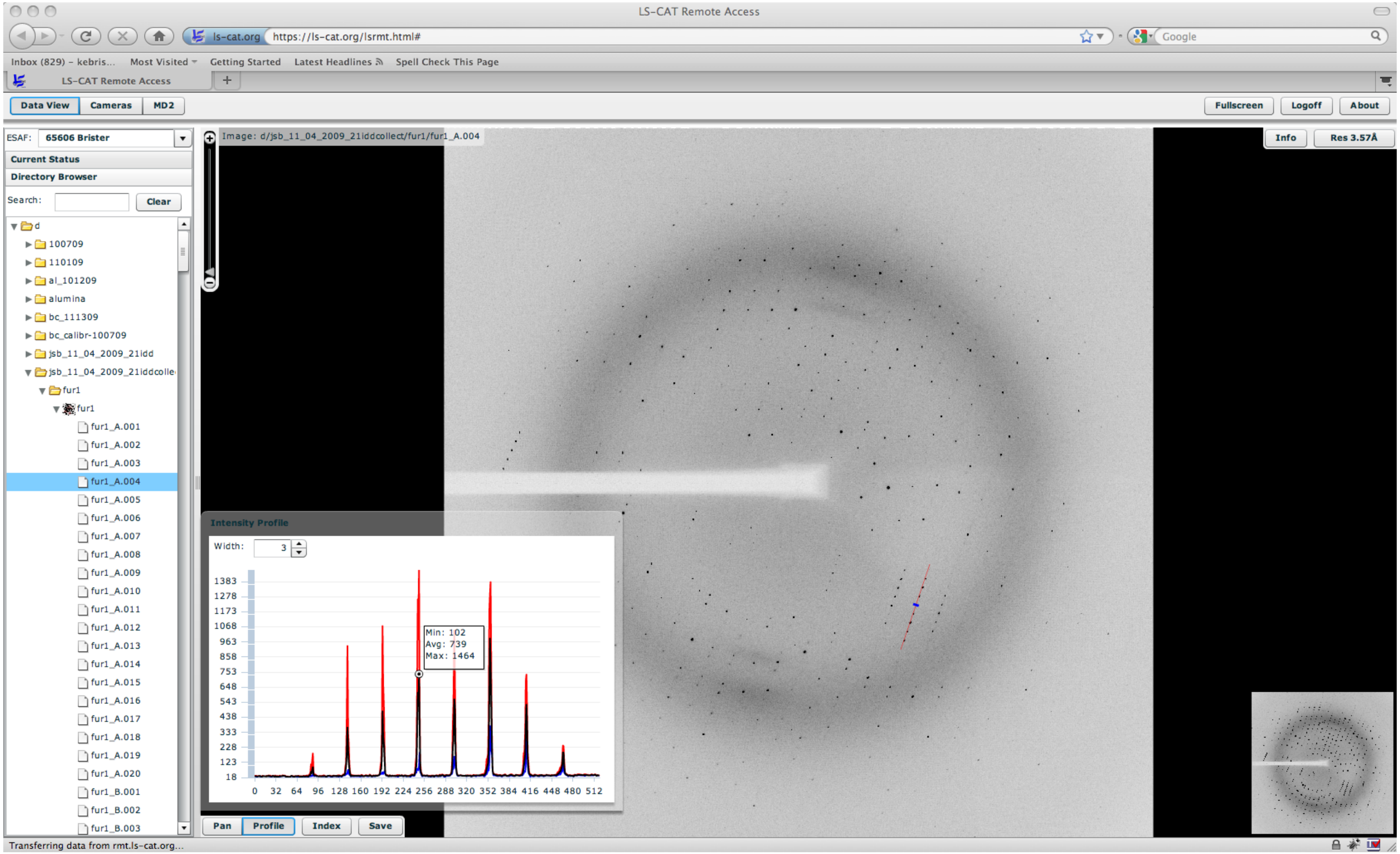
Image: d/jsb_11_04_2009_21iddcollect/fur1/fur1_A.004

Res 3.57Å

- d
  - 100709
  - 110109
  - al_101209
  - alumina
  - bc_111309
  - bc_calibr-100709
  - jsb_11_04_2009_21idd
  - jsb_11_04_2009_21iddcolle
    - fur1
      - fur1
        - fur1_A.001
        - fur1_A.002
        - fur1_A.003
        - fur1_A.004
        - fur1_A.005
        - fur1_A.006
        - fur1_A.007
        - fur1_A.008
        - fur1_A.009
        - fur1_A.010
        - fur1_A.011
        - fur1_A.012
        - fur1_A.013
        - fur1_A.014
        - fur1_A.015
        - fur1_A.016
        - fur1_A.017
        - fur1_A.018
        - fur1_A.019
        - fur1_A.020
        - fur1_B.001
        - fur1_B.002
        - fur1_B.003

Transferring data from rmt.ls-cat.org...


LS-CAT Remote Access
ls-cat.org https://ls-cat.org/lsrmt.html#
Google
Inbox (829) – kebris...
Most Visited
Getting Started
Latest Headlines
Spell Check This Page
LS-CAT Remote Access
Data View
Cameras
MD2
Fullscreen
Logoff
About
ESAF: 65606 Brister
Current Status
Directory Browser
Search:
Clear
d
100709
110109
al_101209
alumina
bc_111309
bc_calibr-100709
jsb_11_04_2009_21idd
jsb_11_04_2009_21iddcolle
fur1
fur1
fur1_A.001
fur1_A.002
fur1_A.003
fur1_A.004
fur1_A.005
fur1_A.006
fur1_A.007
fur1_A.008
fur1_A.009
fur1_A.010
fur1_A.011
fur1_A.012
fur1_A.013
fur1_A.014
fur1_A.015
fur1_A.016
fur1_A.017
fur1_A.018
fur1_A.019
fur1_A.020
fur1_B.001
fur1_B.002
fur1_B.003
Image: d/jsb_11_04_2009_21iddcollect/fur1/fur1_A.004
Info
Res 3.57Å
Pan
Profile
Index
Save
Waiting for rmt.ls-cat.org...



https://ls-cat.org/lsrmt.html#

ESAF: 65606 Brister

Image: d/jsb_11_04_2009_21iddcollect/fur1/fur1_A.004

Width: 3

Min: 102
Avg: 739
Max: 1464

Transferring data from rmt.ls-cat.org...

PX data and Grid computing

Here are two possibilities for grid computing:

1) Make all diffraction images available to speed processing

- Once a good crystal is found, grid resources could be marshaled to find the best possible structure.
- Ideally the decision of which crystal to mount next could be made during the few minutes the crystal is mounted on the diffractometer.
- Needs a great deal of bandwidth from the beamline or a small grid near the beamline.
- Effort would be needed to implement a usable framework and to develop the processes that would actually run
- This is not needed today, it will probably be useful in ~ 5 years as Pixel Array Detectors display CCD detectors.

PX data and Grid computing

2) Make diffraction images part of the deposition to the PDB.

- Automate structure re-refinement based using newer techniques. (Is the robot better than the author?)
- Would allow others to process the data and see if the structure proposed by the author is the best possible. (Does a competing author do a better job than the original author?)
- Would allow systematic comparisons between beamlines leading to improvements of existing facilities. (Why is that data so lousy anyway?)
- Would need something like 500 TB to start with (Assuming data from existing structures could be found and uploaded)
- Would need something like 50-100TB/year to keep up
- Development costs would include converting existing data files into a uniform format.
- Would need staff to visit sites and help with the uploads and track down correct meta-data.
- Cost would approximately be that of building and operating a new beamline

Take away messages

- For this user community, analyzing diffraction data is just one of many steps in a long process. Simplicity is good.
- We can expect that beamline users will need more and more crystallographic support but beamlines will not find increased funding needed to provide it. Perhaps grid computing could allow beamline crystallographers to be more efficient and/or facilitate collaborations with other crystallographers.
- Making raw data available for deposited structures is not that expensive (~1% of existing beamline infrastructure) although finding someone to fund it may be difficult.
- Finding funding may be easier than getting people to use deposit raw data unless funding agencies require this as a condition of receiving a grant.

Progress on the LS-CAT remote interface possible only through the support from our member institutions and by our staff:

Lead UI programer:
Max Brister (New Mexico Tech)

Alpha/Beta Testers:
Janson Ackley
Michael Bolbat
Joseph Brunzelle
Shashank Sharma
Jay VonOsinski
Spencer Anderson
Nancy Brennan
Elena Kondrashkina
David Smith
Zdzislaw Wawrzak