Data Analysis Infrastructure for DLS MX/CX

M. Gerstel  R. Gildea  K. Levik  A. Ashton  G. Winter

Diamond Light Source, Harwell, Oxfordshire, UK

NOBUGS 2016
GDA grid scan
Grid scan processing

- Data collection
- Grid scan?
  - Yes: per image analysis (all images)
  - No: per image analysis (≤ 250 images)
GDA grid scan runs a script

```bash
#!/bin/sh

# script that is run at the end of every collect in HK
# original author of this script is Alan Ashton 24/5/08
# LI0831 added archiving to the end - awa
# input to script:
# automaticProcessing_Yes 4315 /disks/08/data/2008/01/0 test2_HIS1_4_4_4_4.png /disks/08/data/2008/01/0 test2_HIS1_4_4_4_4.png
# Create # Full and proper log/script for re running reproprocessing. # $3
# echo "* date >> $3/spool/RunEndOFCollectLog.sh
# processing_dir=/disks/sw/apps/cr-scripts/hisco/RtnProcessingDir.sh $1($)
# echo "Test == $["processing_dir"] >> $3/spool/RunEndOFCollectLog.sh
# echo "$0 $*" >> $3/spool/RunEndOFCollectLog.sh
# echo "sleep 120" >> $3/spool/RunEndOFCollectLog.sh
#
# Pause for nfs

# umask 002

# Pause=60

# Visits: basename $3
# ImageDirectories($5)
# FileImageDirectory: $" /> ImageDirectory="$ImageDirectory"
# echo RfEnc ImageDirectory $ImageDirectory
# exit SHEFILE env or set
# set SHEFILE: then
# LINE: echo "# $: i cut $/" >$
# file=/disks/sw/apps/cr-scripts/hisco/isInImageCollection $2
# if [ "$file" == "True" ] : then
# then
# Tm appear in a separate job for monitoring reasons
# /disks/sw/apps/histo_all_images.sh $1 $2 $3 $4 $5 ImageDirectory $6
#
# Run XHE2

if [[ "$3DataCollection == "True" ]]: then
  if [[ "$SHEFILE == "123" ]]: then
    /disks/sw/apps/cr-scripts/autorun2/run-fast-directory.sh $1 $2 $3 $4 $5 ImageDirectory $6 /tmp/date +S$ 24:1 1 #/dev/null
  fi
  if [[ "$1" == "automaticProcessing_Yes" ]]: then
    rmdir -A $1/disks/sw/apps/cr-scripts/autorun2/RunDirectFromDir.sh
    rmdir -A $1/disks/sw/apps/cr-scripts/autorun2/RunDirectFromDir.sh
    eval $1/disks/sw/apps/cr-scripts/autorun2/RunDirectory.sh $2 $3 $4 $5 ImageDirectory $6 #/dev/null 24:1 1 #/dev/null
  fi
  echo "date $3" $1/disks/sw/apps/cr-scripts/autorun2/RunEndOFCollectLog.sh $2 $3 $4 $5 ImageDirectory $6" >> $3/sw/SHEFILE/Logs/RunEndOFCollectCommand.log
fi

# Auto Run EDMA

if [[ "$SHEFILE == "123" ]]: then
```
Scripts run more scripts
Scripts run more scripts

```
#!/bin/sh
*/

script that is ran at the end of every collect in M6
* original author of this script is Alan Ashton 24/000
*
*
+ input to script:
+ automaticProcessing:Yes 4315 /dis/105/data/2008/0/test2_M6S_4_**.img /dis/105/data/2008/0/0-0
*
* Create a full and proper log/script file running reprocessing. 4H 130/02
* echo 'A date' >> $(/s/pool/RunEndDFCollectLog.sh)
* processing_dir=/dis/sw/app/scripts/kiscoReprocessingDir.sh $(2)
* echo 'Processing Dir: $processing_dir' >> $(2)/pool/RunEndDFCollectLog.sh
* echo 'test6' >> $(2)/pool/RunEndDFCollectLog.sh
* echo 'sleep 120' >> $(2)/pool/RunEndDFCollectLog.sh
*
*
+ Pause for nfs
*
+

qsub -cwd

qsub -cwd

qsub -cwd
```
Data set processing

- **data collection**
  - Rotation scan?
    - yes
      - Related data collections?
        - yes
          - autoprocessing: "multi-xia2"
        - no
          - Screening images?
            - yes
              - Strategy calculation
                - EDNA (5×)
                - MOSFLM
            - no
              - autoprocessing: FastDP
                - xia2 (3×)
                - autoPROC
Data set processing II

1. Autoprocessing
2. PDB?
   - Yes: Anomalous?
     - No: Dimple
3. Sequence?
   - Yes: MrBUMP
Limitations

- Maintainability
Limitations

- Maintainability
- Testing
Limitations

- Maintainability
- Testing
- Monitoring
Limitations

- Maintainability
- Testing
- Monitoring
- Data reprocessing
Limitations

- Maintainability
- Testing
- Monitoring
- Data reprocessing
- Efficient use of computing resources
DECTRIS EIGER X

100 Gbit/s
DECTRIS EIGER X

10 Gbit/s
ActiveMQ

Data
/dls/
network
file
system

Metadata
ISPyB
database
ActiveMQ Messaging Infrastructure

Data
/dls/
network file system

Events
ActiveMQ Messaging Server

Metadata
ISPyB database
ActiveMQ

- Fast
- Lightweight
- Not invented here
- Broad language support
- Already used by other groups
Broker: Central server
Terminology

- **Broker**: Central server
- **Queue**: Producer–Consumer
Terminology

- **Broker**: Central server
- **Queue**: Producer–Consumer
- **Topic**: Publish–Subscribe
Terminology

- **Broker**: Central server
- **Queue**: Producer–Consumer
- **Topic**: Publish–Subscribe
- **Transaction**: All or nothing
Delivery guarantee with transactions

Input queue

Service

Output queue

TXN

Data

ACK

process()

COMMIT

COMMIT
ActiveMQ

ActiveMQ Messaging Server

ISPyB database connector

GDA

Per-image-analysis service

Data reduction service
GDA grid scan announces data collection

$\{\text{DCID}\} + \text{do-this} \rightarrow /queue/data-collection
Data collection triggers processing

- `/queue/data-collection` ➔ `/queue/fast_dp`
- `/queue/data-collection` ➔ `/queue/xia2`
- `/queue/data-collection` ➔ `/queue/multi-xia2`
Supervised & encapsulated processing

/xqueue/xia2

drmaa submit

xia2 manager:
run task
monitor status
publish status & results

xia2 pipeline=dials...
Reporting results

xia2 manager:
run task
monitor status
publish status & results

xia2 pipeline=dials...

/queue/xia2-results
Reporting results

- `/queue/xia2-results`
- ISPyB
- GDA results panel
- QoS monitor

Results flow:
- `/queue/xia2-results` to ISPyB
- Results to GDA results panel
- Results/status to QoS monitor
Status monitor
## Status monitor

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>APP SERVER AVAILABILITY</strong></td>
<td>100.0%</td>
</tr>
<tr>
<td><strong>MEAN WEB RESPONSE TIME</strong></td>
<td>158ms</td>
</tr>
<tr>
<td><strong>MEAN API RESPONSE TIME</strong></td>
<td>33ms</td>
</tr>
<tr>
<td><strong>98TH PERC. WEB RESPONSE TIME</strong></td>
<td>700ms</td>
</tr>
<tr>
<td><strong>PAGES BUILDS FAILURE RATE</strong></td>
<td>0.1618%</td>
</tr>
<tr>
<td><strong>EXCEPTION PERCENTAGE</strong></td>
<td>0.0001%</td>
</tr>
<tr>
<td><strong>MEAN HOOK DELIVERY TIME</strong></td>
<td>0.85s</td>
</tr>
<tr>
<td><strong>98TH PERC. BROWSER TIME TO FIRST BYTE</strong></td>
<td>2458ms</td>
</tr>
</tbody>
</table>
Infrastructure Overview

- ISPyB database connector
- GDA
- Per-image-analysis service
- Data reduction service

ActiveMQ Messaging Server
Infrastructure Overview

- ActiveMQ Messaging Server
  - ISPyB database connector
  - DLS SciSoft status screen
  - GDA
  - Per-image-analysis service
  - Data reduction service
Infrastructure Overview

- ActiveMQ Messaging Server
- GDA
- File monitoring service
- Per-image-analysis service
- Spot finding service
- Indexing service
- Data reduction service
- Structure solution service
- Process monitoring service
- Live diffraction image viewer
- Live reciprocal lattice viewer
- Live processing status viewer
- DLS SciSoft status screen
- ISPyB database connector
Timeframe

- Diamond currently in shutdown for ring upgrade
- Short (2 week) run in December 2016
- New infrastructure to be deployed on 2-3 beamlines
- Running in parallel to existing scripts
- Begin replacing existing infrastructure from early 2017
Thank you