Joint Facilities User Forum on Data Intensive Computing, June 16th 2014

Data Needs for LCLS-II

Amedeo Perazzo
SLAC

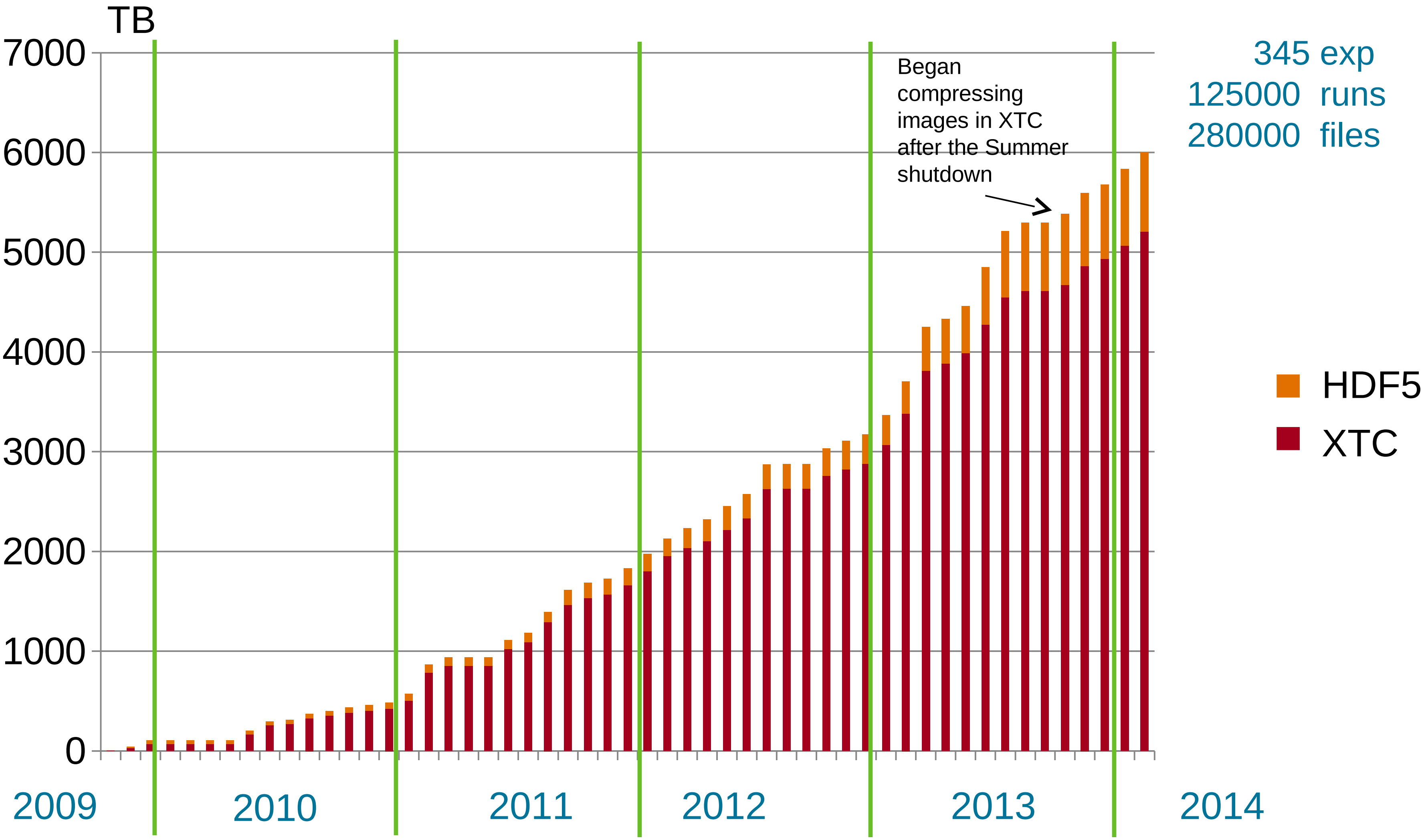
U.S. DEPARTMENT OF ENERGY
Office of Science

SLAC NATIONAL ACCELERATOR LABORATORY

LCLS Data Throughputs

- **Current LCLS data system can handle fast feedback and offline analysis requirements for most LCLS experiments**
  - DAQ throughput ranges 0.1 – 10 GB/s, typically 1GB/s
    - CSPAD detector: 2 x 2.3 Mpixel @ 120Hz = 1.1 GB/s
- **Predictions for future LCLS data throughput are not obvious**
  - Dictated by project cost, more than physics requirements
  - My guess:
    - One order of magnitude in 4 years time scale
      - 2 x 16Mpixel @ 120Hz (larger CSPAD detectors)
    - Two orders of magnitude in 8 years time scale
      - 100K points @ 100KHz (1D detectors @ LCLS-II data rates)
      - 2 x 4 Mpixel @ 4KHz (ePix detector family)

LCLS Data Volumes



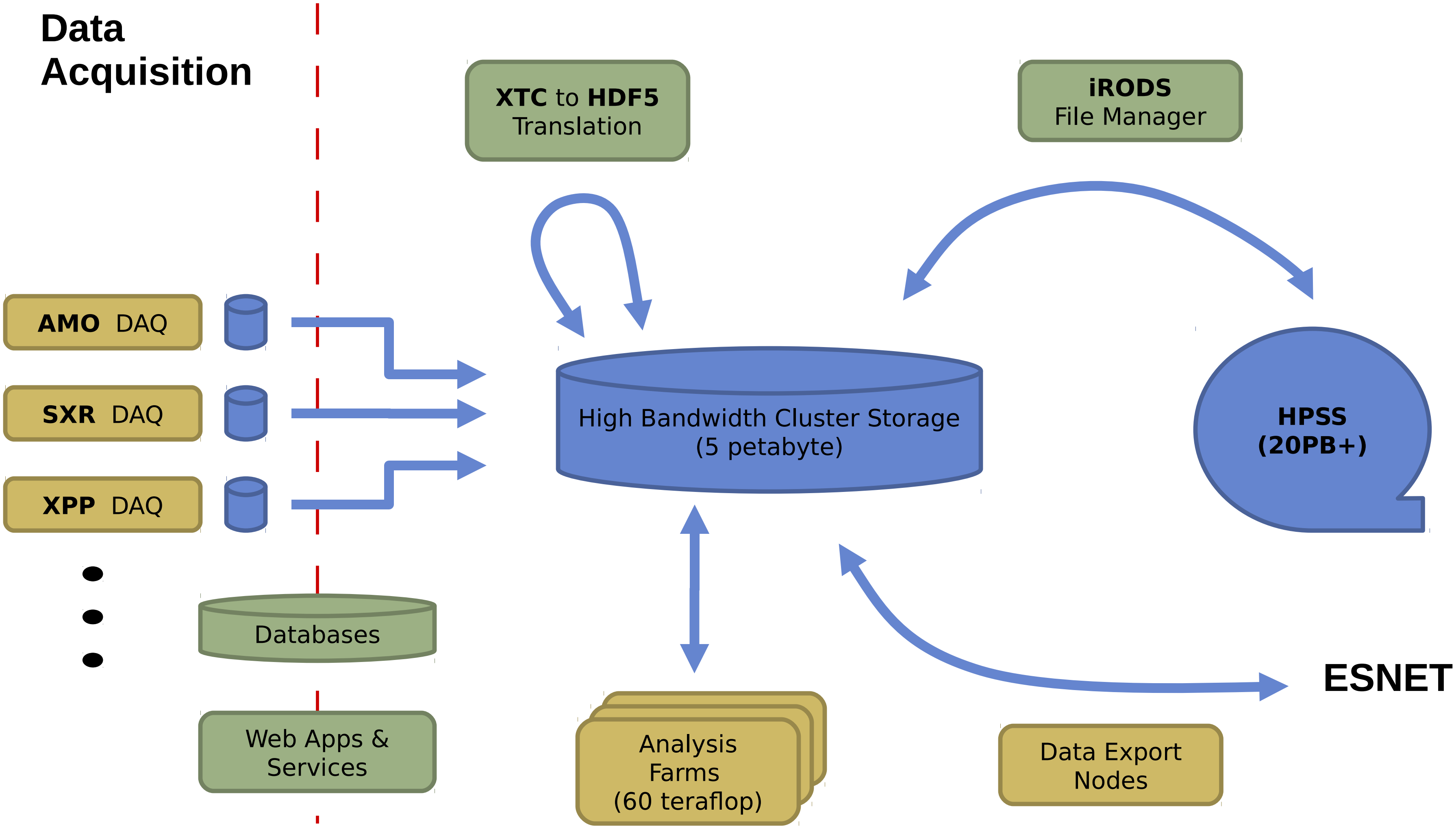
Began compressing images in XTC after the Summer shutdown

345 exp
125000 runs
280000 files

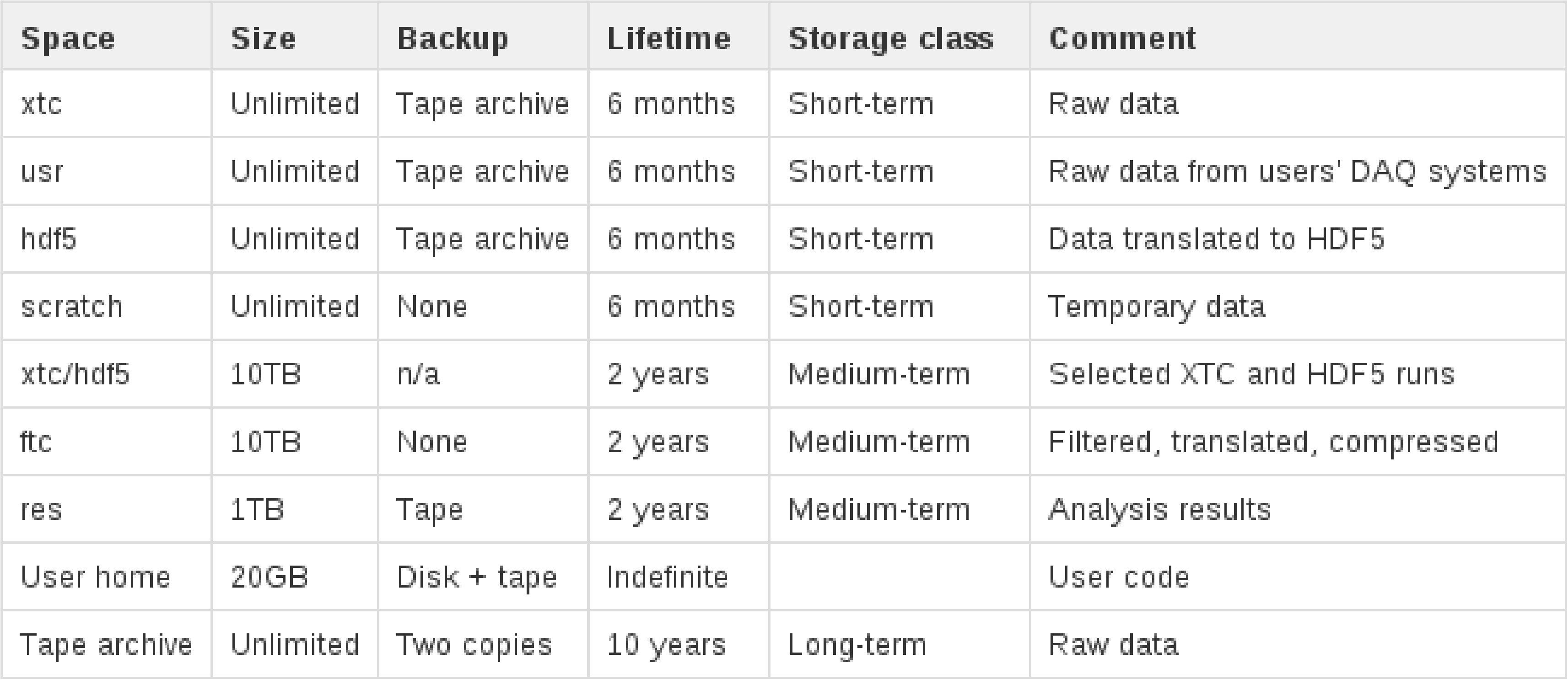
LCLS Data Analysis

- **Large variety of tools for analyzing LCLS science data**
  - Real-time, on-the-fly, network based monitoring framework
    - Augmented via modules implemented as shared libraries or shared memory for external framework analysis
  - Fast-feedback, 1-10s delay, disk based analysis
  - Offline analysis: psana (C++/Python), interactive psana, Matlab, CASS, etc
- **Fragmentation of analysis tools partially dictates data infrastructure**
  - Eg. POSIX file systems requirements

LCLS Data Systems Architecture



LCLS Data Policies



Space	Size	Backup	Lifetime	Storage class	Comment
xtc	Unlimited	Tape archive	6 months	Short-term	Raw data
usr	Unlimited	Tape archive	6 months	Short-term	Raw data from users' DAQ systems
hdf5	Unlimited	Tape archive	6 months	Short-term	Data translated to HDF5
scratch	Unlimited	None	6 months	Short-term	Temporary data
xtc/hdf5	10TB	n/a	2 years	Medium-term	Selected XTC and HDF5 runs
ftc	10TB	None	2 years	Medium-term	Filtered, translated, compressed
res	1TB	Tape	2 years	Medium-term	Analysis results
User home	20GB	Disk + tape	Indefinite		User code
Tape archive	Unlimited	Two copies	10 years	Long-term	Raw data

LCLS Data Infrastructure

- **DAQ systems dedicated per hutch, user analysis system shared across instruments**
- **Four storage layers**
  - Online cache (flash), fast-feedback (disk), medium term (disk), long term (tape)
  - Medium-term storage currently 5 petabytes
    - Each PB aggregated throughput of 12GB/sec
  - Long-term storage uses tape staging system in the SLAC central computing facilities
    - Can scale up to several petabytes
- **Processing: batch pool and interactive pool**
  - 60Tflop total
  - Most cycles are given out to other SLAC groups because of the bursty nature of LCLS experiments
- **Farms live in the experimental areas with fast (IB QDR) access to the science data files in medium-term storage**

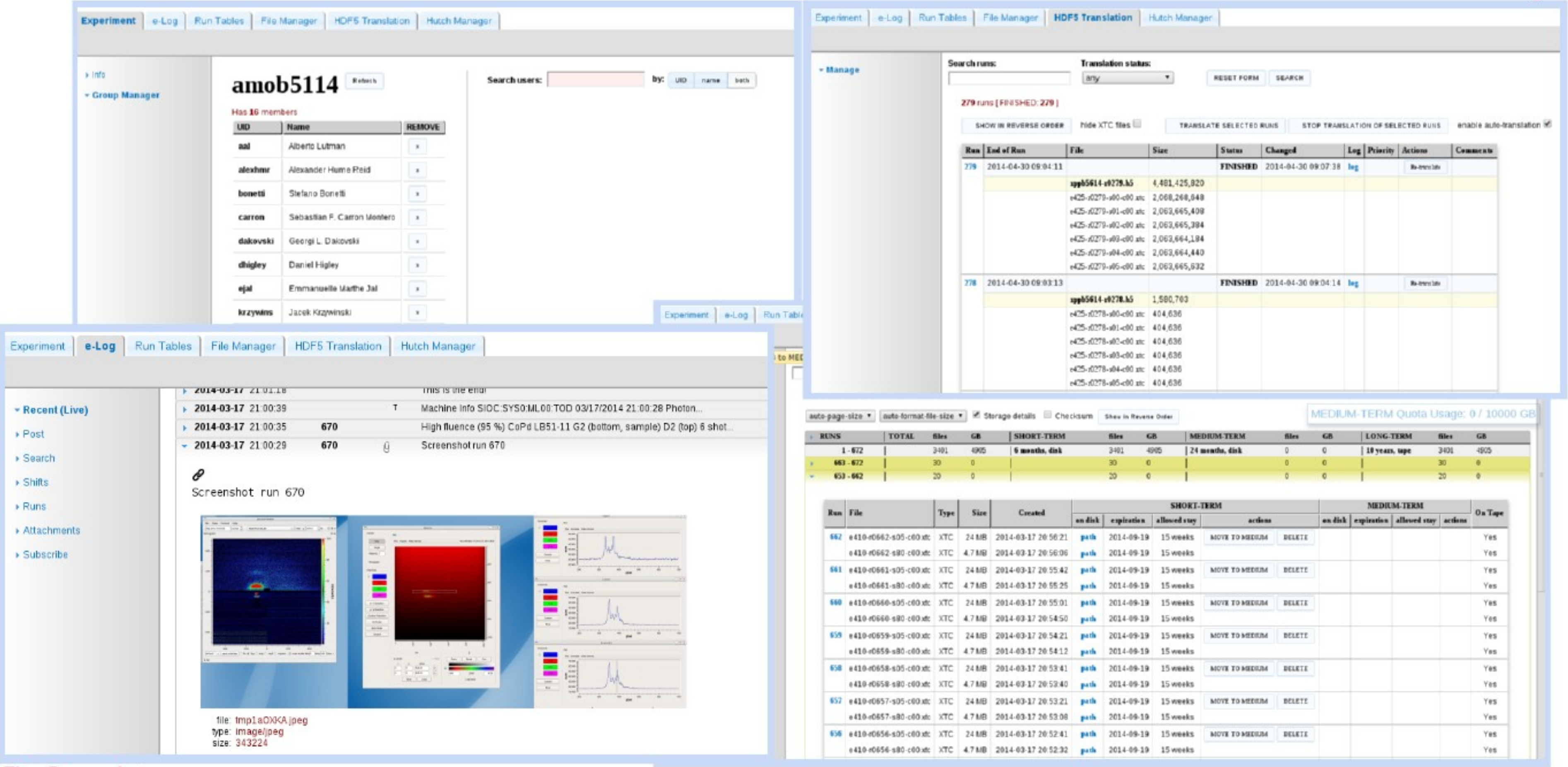
LCLS Data Management Framework

- **Data Management system handles all content-opaque operations**
  - Moves data across storage layers (online cache, fast-feedback, offline storage, tape)
  - User accessible through LCLS web-portal (electronic logbook)
  - Handles data policies (security, access, retention)
  - Handles DAQ generated data or data resulted from centralized processing (eg HDF5 translation, compression, filtering)
  - Archive to tape (HPSS) implemented as iRODS service
- **Currently handling 11PB LCLS data, raw and user generated**
  - 5PB on disk, 6PB on tape

LCLS Data Management Framework
Interface Examples

Group Management

Translation Manager



amob5114

Has 16 members

UID	Name	REMOVE
aal	Alberto Lutman	x
alexhmr	Alexander Hume Reid	x
bonetti	Stefano Bonetti	x
carron	Sebastian F. Carron Montero	x
dakovski	Georgi L. Dakovski	x
dhigley	Daniel Higley	x
ejal	Emmanuelle Marthe Jal	x
krzywins	Jacek Krzywinski	x

Search runs:

Translation status: any

279 runs [FINISHED: 279]

hide XTC files □ enable auto-translation ☑

Run	End of Run	File	Size	Status	Changed	Log	Priority	Actions	Comments
279	2014-04-30 09:04:11			FINISHED	2014-04-30 09:07:38	log		Re-translate	
		xppb5614-r0279.h5	4,481,425,820						
		e425-r0279-s00-c00.xtc	2,068,268,648						
		e425-r0279-s01-c00.xtc	2,063,665,408						
		e425-r0279-s02-c00.xtc	2,063,665,384						
		e425-r0279-s03-c00.xtc	2,063,664,184						
		e425-r0279-s04-c00.xtc	2,063,664,440						
		e425-r0279-s05-c00.xtc	2,063,665,632						
278	2014-04-30 09:03:13			FINISHED	2014-04-30 09:04:14	log		Re-translate	
		xppb5614-r0278.h5	1,580,703						
		e425-r0278-s00-c00.xtc	404,636						
		e425-r0278-s01-c00.xtc	404,636						
		e425-r0278-s02-c00.xtc	404,636						
		e425-r0278-s03-c00.xtc	404,636						
		e425-r0278-s04-c00.xtc	404,636						
		e425-r0278-s05-c00.xtc	404,636						

```
2014-03-17 21:01:18                This is the end!
2014-03-17 21:00:39        T       Machine Info SIOC:SYS0:ML00:TOD 03/17/2014 21:00:28 Photon...
2014-03-17 21:00:35  670           High fluence (95 %) CoPd LB51-11 G2 (bottom, sample) D2 (top) 6 shot...
2014-03-17 21:00:29  670           Screenshot run 670
```

Screenshot run 670

file: tmp1aOXKA.jpeg
type: image/jpeg
size: 343224

Elog Screenshot

☑ Storage details □ Checksum

MEDIUM-TERM Quota Usage: 0 / 10000 GB

RUNS	TOTAL	files	GB	SHORT-TERM	files	GB	MEDIUM-TERM	files	GB	LONG-TERM	files	GB
1 - 672		3401	4905	6 months, disk	3401	4905	24 months, disk	0	0	10 years, tape	3401	4905
663 - 672		30	0		30	0		0	0		30	0
653 - 662		20	0		20	0		0	0		20	0

Run	File	Type	Size	Created	SHORT-TERM				MEDIUM-TERM				On Tape
					on disk	expiration	allowed stay	actions	on disk	expiration	allowed stay	actions	
662	e410-r0662-s05-c00.xtc	XTC	24 MB	2014-03-17 20:56:21	path	2014-09-19	15 weeks	MOVE TO MEDIUM DELETE					Yes
	e410-r0662-s80-c00.xtc	XTC	4.7 MB	2014-03-17 20:56:06	path	2014-09-19	15 weeks						Yes
661	e410-r0661-s05-c00.xtc	XTC	24 MB	2014-03-17 20:55:42	path	2014-09-19	15 weeks	MOVE TO MEDIUM DELETE					Yes
	e410-r0661-s80-c00.xtc	XTC	4.7 MB	2014-03-17 20:55:25	path	2014-09-19	15 weeks						Yes
660	e410-r0660-s05-c00.xtc	XTC	24 MB	2014-03-17 20:55:01	path	2014-09-19	15 weeks	MOVE TO MEDIUM DELETE					Yes
	e410-r0660-s80-c00.xtc	XTC	4.7 MB	2014-03-17 20:54:50	path	2014-09-19	15 weeks						Yes
659	e410-r0659-s05-c00.xtc	XTC	24 MB	2014-03-17 20:54:21	path	2014-09-19	15 weeks	MOVE TO MEDIUM DELETE					Yes
	e410-r0659-s80-c00.xtc	XTC	4.7 MB	2014-03-17 20:54:12	path	2014-09-19	15 weeks						Yes
658	e410-r0658-s05-c00.xtc	XTC	24 MB	2014-03-17 20:53:41	path	2014-09-19	15 weeks	MOVE TO MEDIUM DELETE					Yes
	e410-r0658-s80-c00.xtc	XTC	4.7 MB	2014-03-17 20:53:40	path	2014-09-19	15 weeks						Yes
657	e410-r0657-s05-c00.xtc	XTC	24 MB	2014-03-17 20:53:21	path	2014-09-19	15 weeks	MOVE TO MEDIUM DELETE					Yes
	e410-r0657-s80-c00.xtc	XTC	4.7 MB	2014-03-17 20:53:08	path	2014-09-19	15 weeks						Yes
656	e410-r0656-s05-c00.xtc	XTC	24 MB	2014-03-17 20:52:41	path	2014-09-19	15 weeks	MOVE TO MEDIUM DELETE					Yes
	e410-r0656-s80-c00.xtc	XTC	4.7 MB	2014-03-17 20:52:32	path	2014-09-19	15 weeks						Yes

File Manager

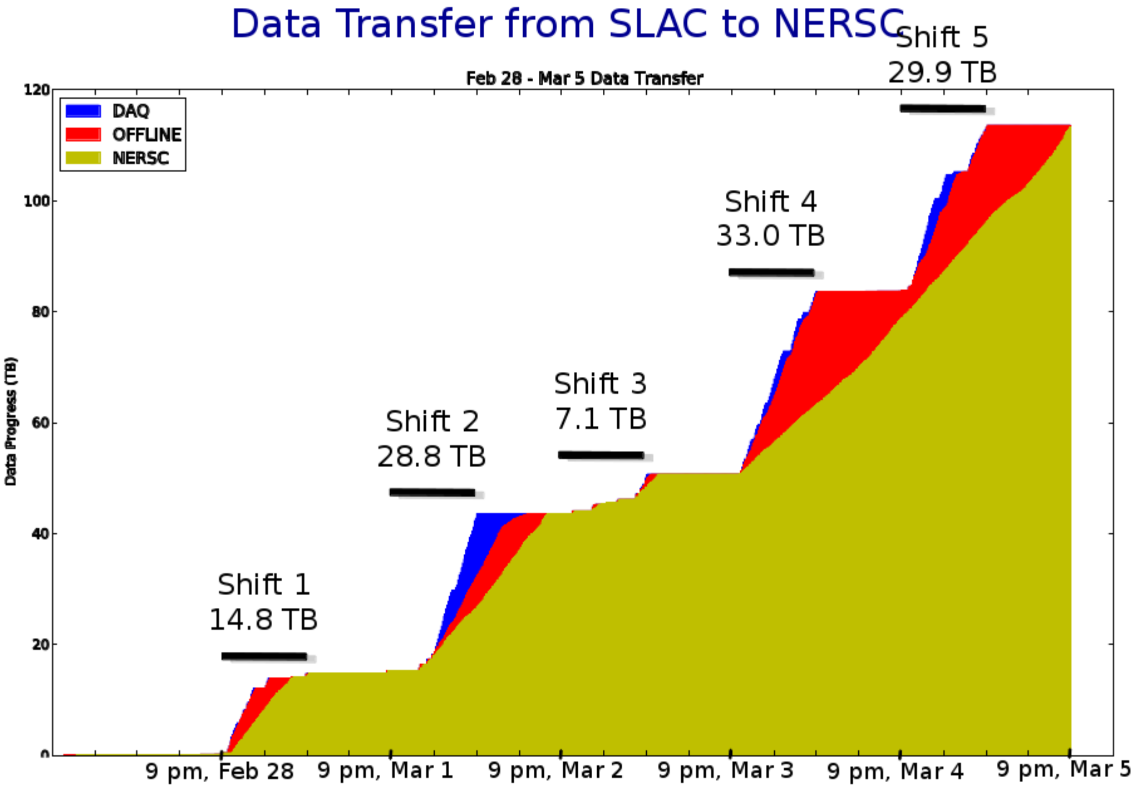
Vetoing Events for FEL Experiments Can Be Tricky

- **Very hard to implement effective trigger/veto system**
  - Not strictly a technical issue: the ability to veto events is already implemented in the system
  - Vetoing based on beam parameters not effective (most pulses are good)
  - Hard to get help from users in setting veto parameters which define event quality
    - Users themselves often don't know what these parameters or their thresholds should be
    - Users are usually very suspicious of anything which can filter data on-the-fly
    - Things may get better as algorithms mature
- **Benefit of vetoing events based on the event data is potentially very large for some experiments**
  - Factor 10-100 for some CXI imaging experiments
  - Many experiments, though, have hit rates close to 100%

LCLS/NERSC Data Pilot

- In 2012 PCDS requested and obtained a NERSC allocation under the “Data Intensive Computing Pilot Program”
- PCDS provided a data-mover script and web-based monitoring to automatically transfer the data for a CXI experiment to NERSC
  - Moved data from SLAC to NERSC at around 700MB/s (ie half of data taking rate)
- PCDS ported LCLS analysis framework to Carver (NERSC farm)

- This exercise showed that partnering with large computer centers like NERSC is part of the solution to LCLS data challenge but can't replace local midscale computing for fast feedback and initial analysis
- Collaborations beyond the data pilot would require 100Gb connection between SLAC and ESNET



Offloading LCLS Data Analysis Infrastructure

- **Data centers built towards data intensive systems could help offload the LCLS/SLAC offline computing system**
  - Based on expected data scaling, no modifications to data retention policies, general support for LCLS offline analysis in 2-3 years timescale would require:
    - ~50 PB tape storage, dedicated ~10 PB of disk storage, ~100 teraflop processing farm with an aggregate throughput to the storage above 10 GB/s per PB
- **Key requirements: ability for LCLS users to manage their data through the LCLS tools and workflows, ability to use their SLAC account (or a federated account)**