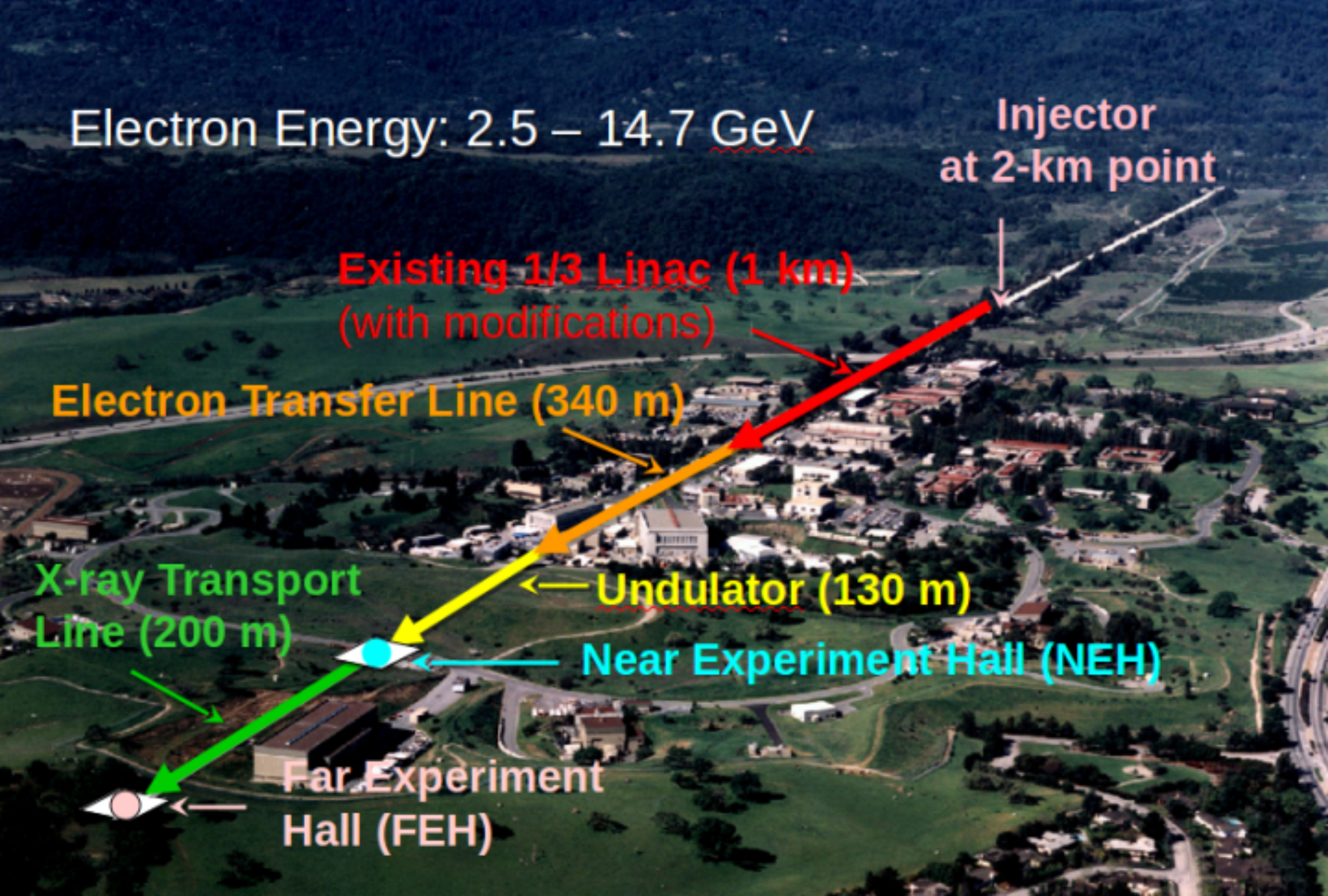
Current and Future Data Systems for the LINAC Coherent Light Source

SC16
Riccardo Veraldi for the LCLS IT Team

U.S. DEPARTMENT OF ENERGY
Office of Science

SLAC NATIONAL ACCELERATOR LABORATORY

Linac Coherent Light Source



Electron Energy: 2.5 – 14.7 GeV

Injector
at 2-km point

Existing 1/3 Linac (1 km)
(with modifications)

Electron Transfer Line (340 m)

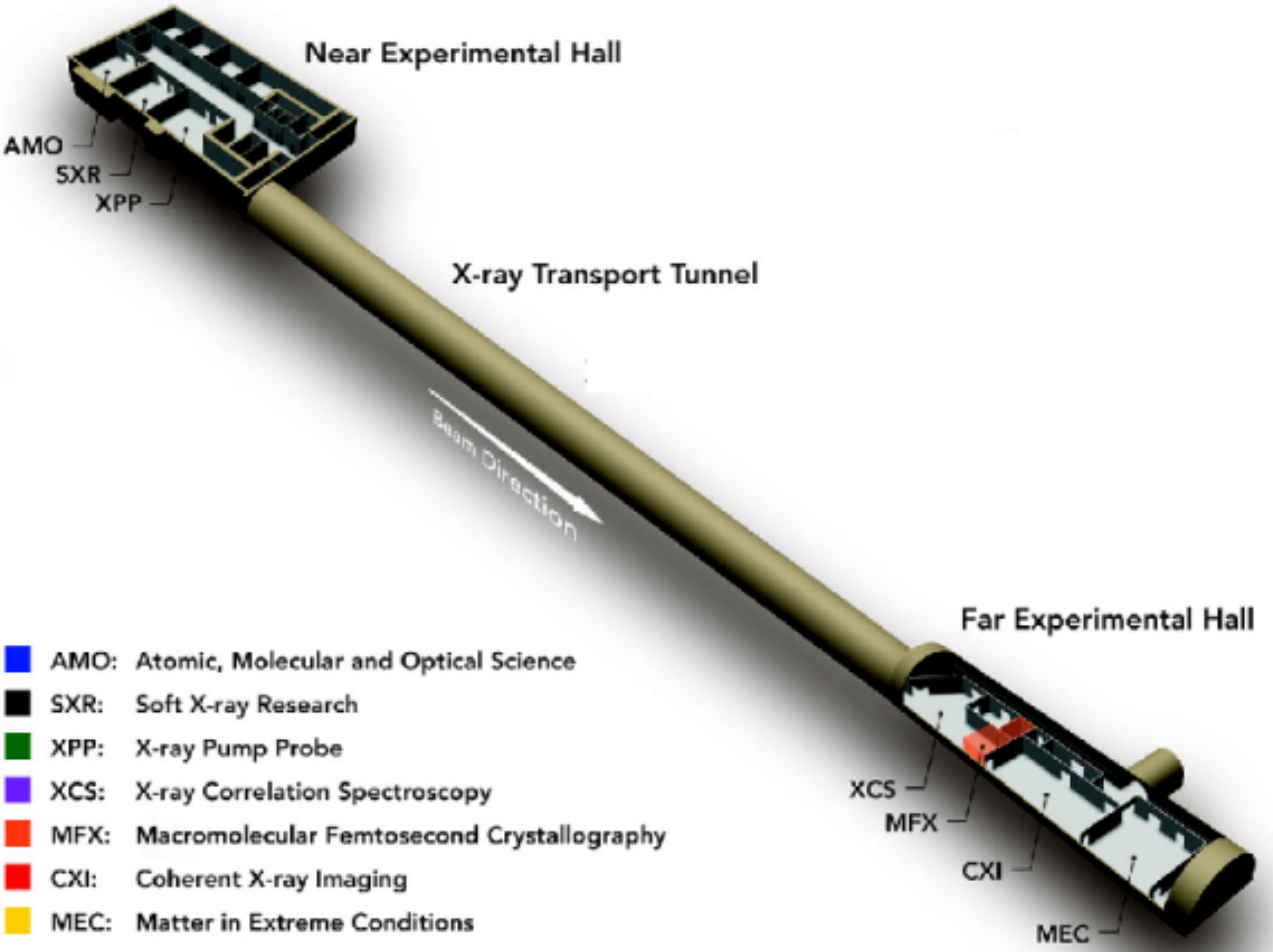
Undulator (130 m)

X-ray Transport
Line (200 m)

Near Experiment Hall (NEH)

Far Experiment
Hall (FEH)

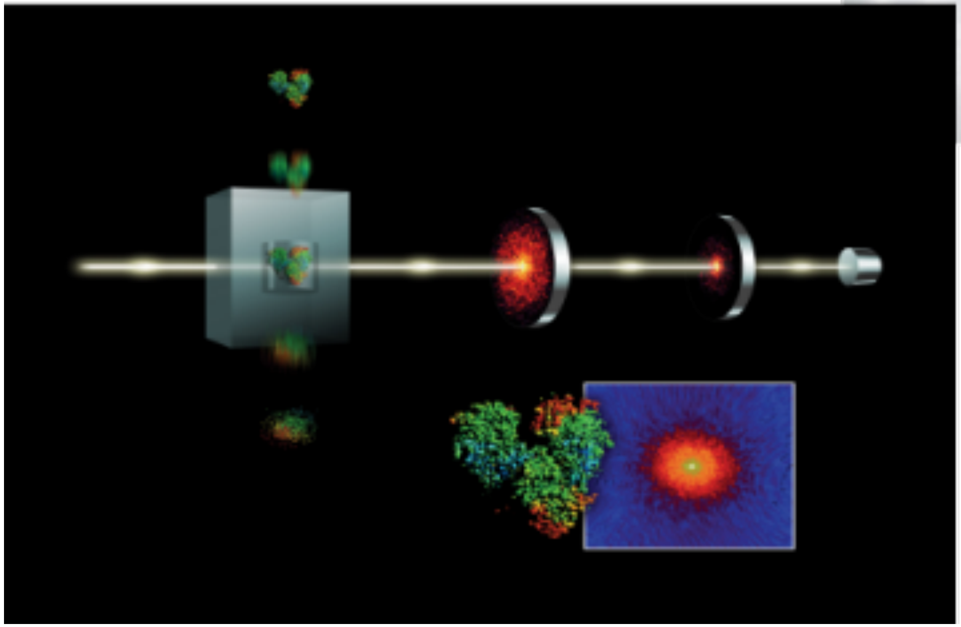
LCLS Experimental Floor



🟦	AMO:	Atomic, Molecular and Optical Science
⬛	SXR:	Soft X-ray Research
🟩	XPP:	X-ray Pump Probe
🟪	XCS:	X-ray Correlation Spectroscopy
🟧	MFX:	Macromolecular Femtosecond Crystallography
🟥	CXI:	Coherent X-ray Imaging
🟨	MEC:	Matter in Extreme Conditions

LCLS Parameters

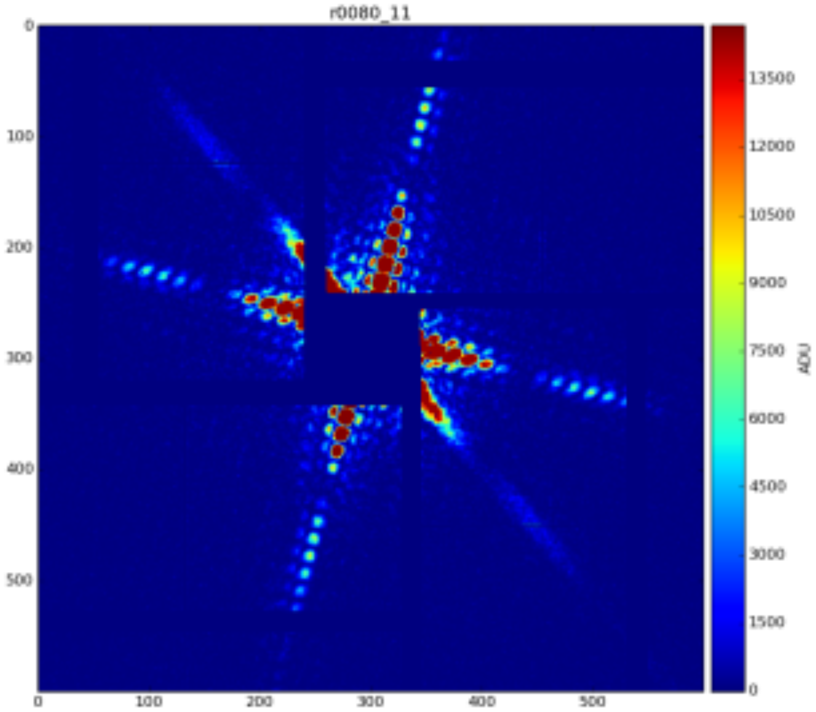
X-Ray range	250 to 11,300 eV
Pulse length	< 5 - 500 fs
Pulse energy	~ 4 mJ
Repetition Rate	120 Hz

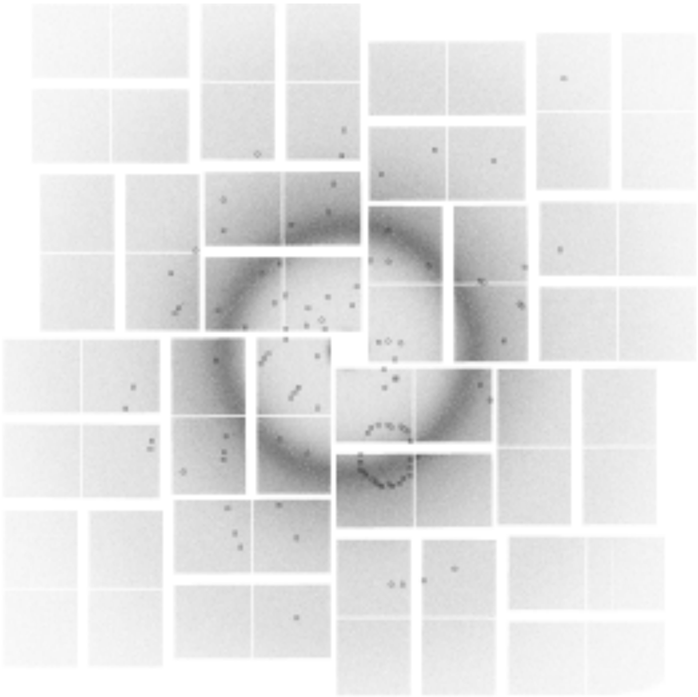


LCLS Physics

LCLS has already had a significant impact on many areas of science, including:

- Resolving the structures of macromolecular protein complexes that were previously inaccessible;
- Capturing bond formation in the elusive transition-state of a chemical reaction;
- Revealing the behavior of atoms and molecules in the presence of strong fields;
- Probing extreme states of matter





LCLS Data Challenges

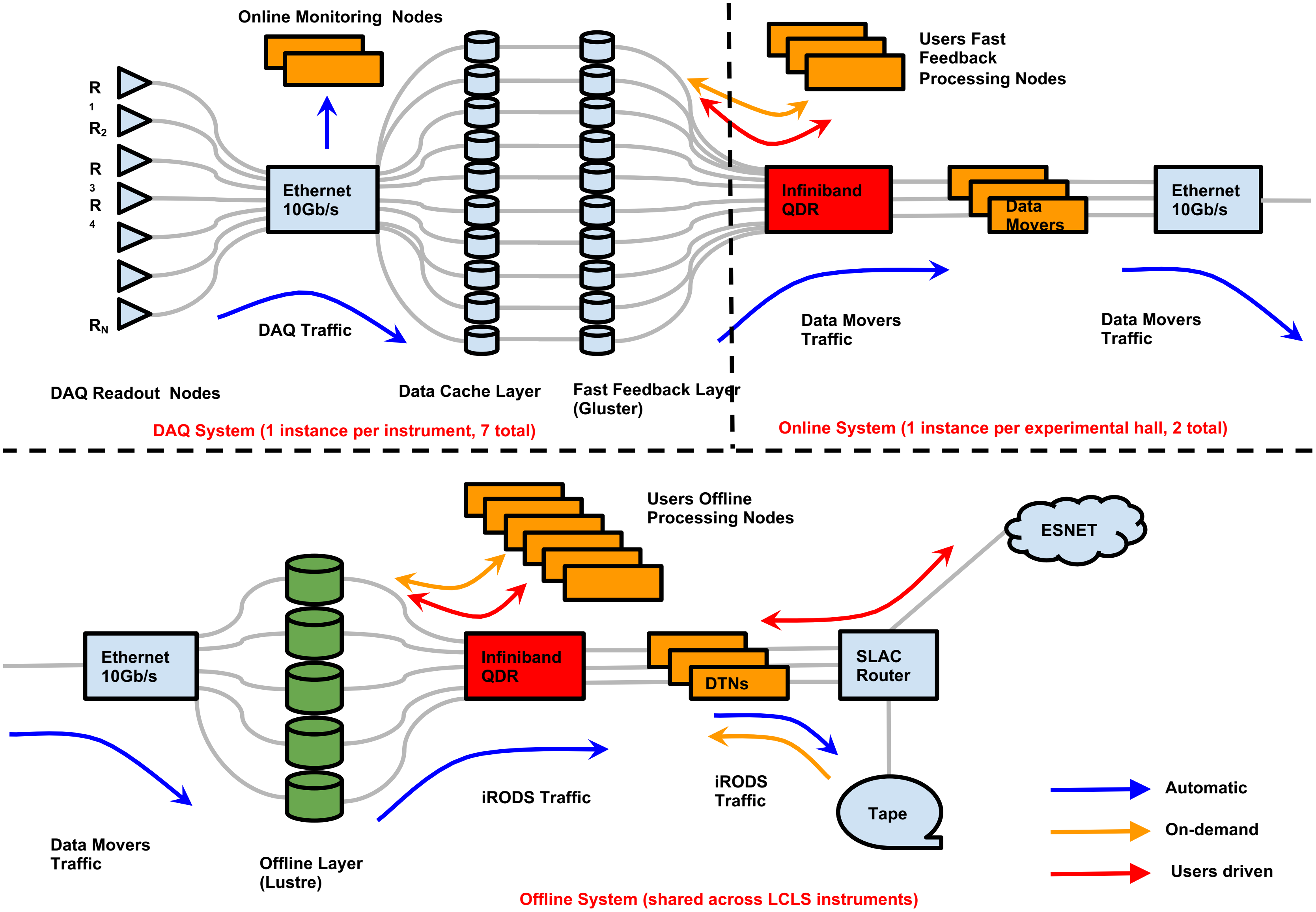
From the beginning LCLS data systems group faced these challenges:

1. Ability to readout, event build and store multi GB/s data streams
2. Capability for experimenters to analyze data on-the-fly (real-time)
3. Flexibility to accommodate new user supplied equipment
4. Capacity to store and analyze PB scale data sets
5. Changing analysis software/algorithms implemented by non-expert users (weekly!)

LCLS currently handles well the first two of the three Vs of data challenges: volume, velocity, variety

- The variety of analysis requirements due to the heterogeneity of experiments is by far the main challenge in the LCLS data analysis arena right now
  - Each experiment requires its own setup (detectors, geometry, etc) and its own intelligence in the analysis
  - Must provide easier, more powerful tools to lower the threshold and provide all groups equal opportunity

LCLS Data Flow: Current



Online Monitoring Nodes

R_1

R_2

R_3

R_4

R_N

Ethernet 10Gb/s

DAQ Traffic

DAQ Readout Nodes

Data Cache Layer

Fast Feedback Layer (Gluster)

DAQ System (1 instance per instrument, 7 total)

Users Fast Feedback Processing Nodes

Infiniband QDR

Data Movers

Ethernet 10Gb/s

Data Movers Traffic

Data Movers Traffic

Online System (1 instance per experimental hall, 2 total)

Users Offline Processing Nodes

ESNET

Ethernet 10Gb/s

Infiniband QDR

DTNs

SLAC Router

Data Movers Traffic

Offline Layer (Lustre)

iRODS Traffic

iRODS Traffic

Tape

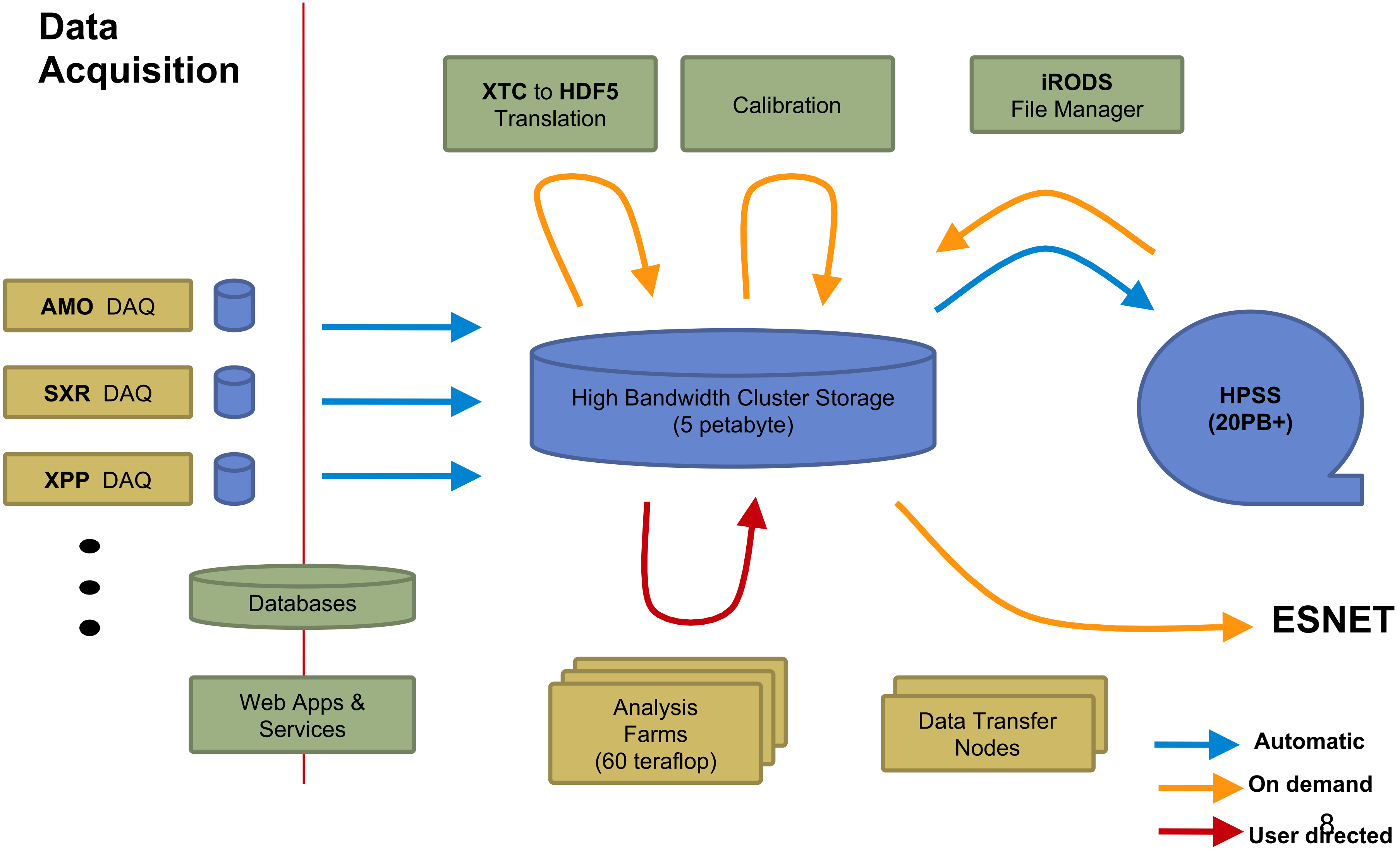
Automatic

On-demand

Users driven

Offline System (shared across LCLS instruments)

Current LCLS Data Systems Architecture



Data Acquisition

XTC to HDF5 Translation

Calibration

iRODS File Manager

AMO DAQ

SXR DAQ

XPP DAQ

High Bandwidth Cluster Storage (5 petabyte)

HPSS (20PB+)

Databases

Web Apps & Services

Analysis Farms (60 teraflop)

Data Transfer Nodes

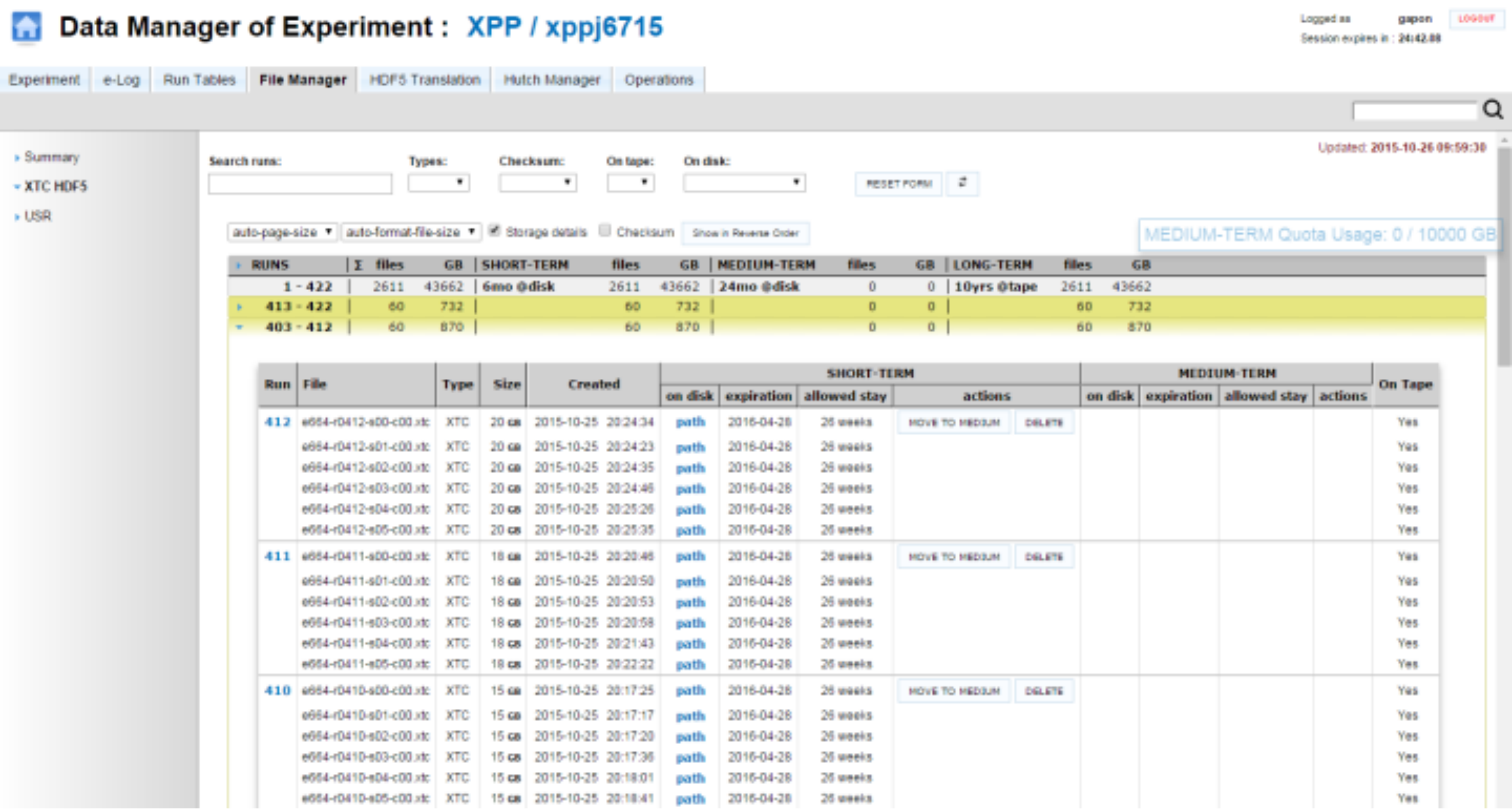
ESNET

Automatic

On demand

User directed

Web Portal Snapshot: File manager



Data Manager of Experiment : XPP / xppj6715

Logged as gapon
Session expires in : 24:42.08

Updated: 2015-10-26 09:59:30

☑ Storage details □ Checksum

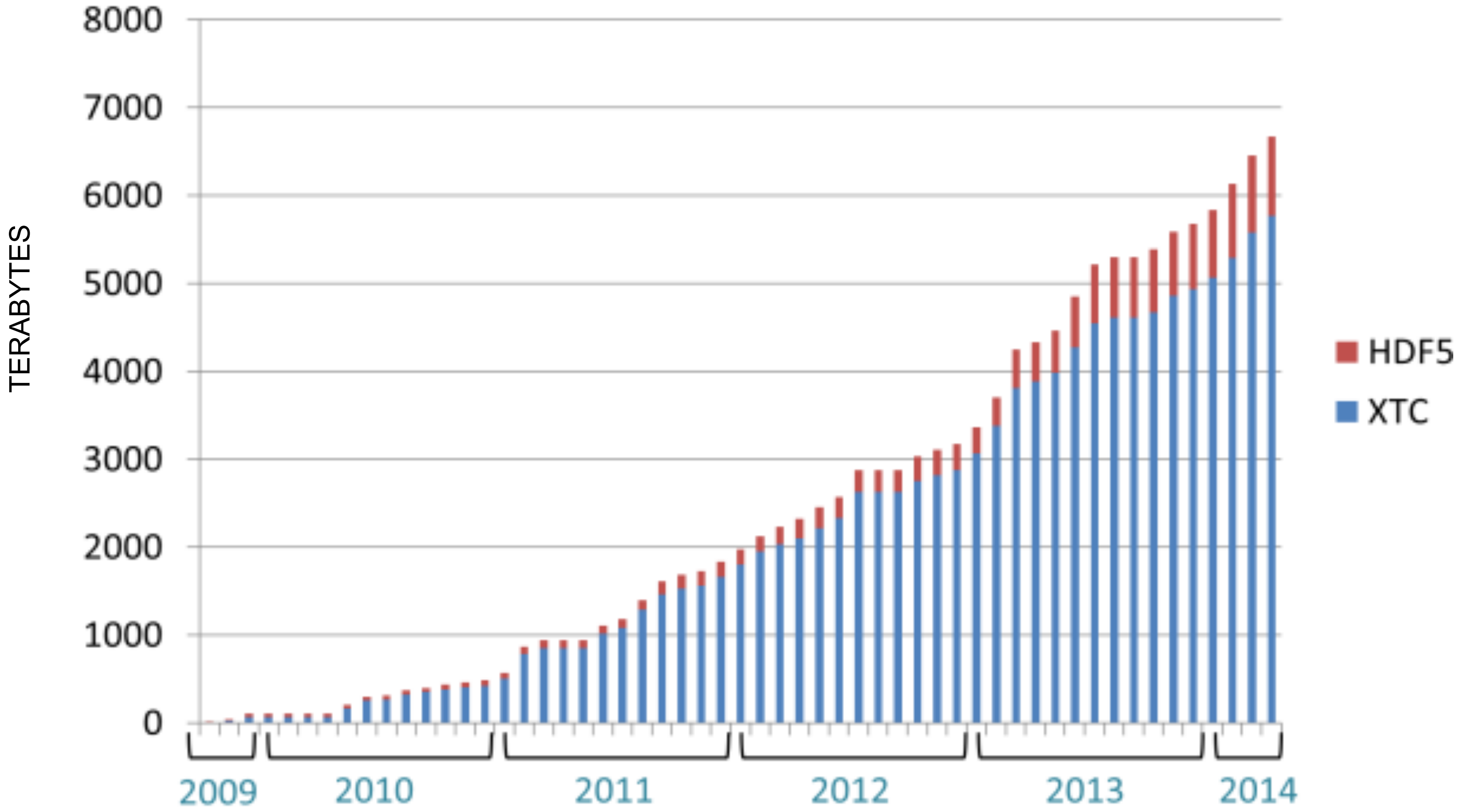
MEDIUM-TERM Quota Usage: 0 / 10000 GB

RUNS	Σ files	GB	SHORT-TERM	files	GB	MEDIUM-TERM	files	GB	LONG-TERM	files	GB
1 - 422	2611	43662	6mo @disk	2611	43662	24mo @disk	0	0	10yrs @tape	2611	43662
413 - 422	60	732		60	732		0	0		60	732
403 - 412	60	870		60	870		0	0		60	870

Run	File	Type	Size	Created	SHORT-TERM				MEDIUM-TERM				On Tape
					on disk	expiration	allowed stay	actions	on disk	expiration	allowed stay	actions	
412	e664-r0412-s00-c00.xtc	XTC	20 GB	2015-10-25 20:24:34	path	2016-04-28	26 weeks	MOVE TO MEDIUM DELETE					Yes
	e664-r0412-s01-c00.xtc	XTC	20 GB	2015-10-25 20:24:23	path	2016-04-28	26 weeks						Yes
	e664-r0412-s02-c00.xtc	XTC	20 GB	2015-10-25 20:24:35	path	2016-04-28	26 weeks						Yes
	e664-r0412-s03-c00.xtc	XTC	20 GB	2015-10-25 20:24:46	path	2016-04-28	26 weeks						Yes
	e664-r0412-s04-c00.xtc	XTC	20 GB	2015-10-25 20:25:26	path	2016-04-28	26 weeks						Yes
	e664-r0412-s05-c00.xtc	XTC	20 GB	2015-10-25 20:25:35	path	2016-04-28	26 weeks						Yes
411	e664-r0411-s00-c00.xtc	XTC	18 GB	2015-10-25 20:20:46	path	2016-04-28	26 weeks	MOVE TO MEDIUM DELETE					Yes
	e664-r0411-s01-c00.xtc	XTC	18 GB	2015-10-25 20:20:50	path	2016-04-28	26 weeks						Yes
	e664-r0411-s02-c00.xtc	XTC	18 GB	2015-10-25 20:20:53	path	2016-04-28	26 weeks						Yes
	e664-r0411-s03-c00.xtc	XTC	18 GB	2015-10-25 20:20:58	path	2016-04-28	26 weeks						Yes
	e664-r0411-s04-c00.xtc	XTC	18 GB	2015-10-25 20:21:43	path	2016-04-28	26 weeks						Yes
	e664-r0411-s05-c00.xtc	XTC	18 GB	2015-10-25 20:22:22	path	2016-04-28	26 weeks						Yes
410	e664-r0410-s00-c00.xtc	XTC	15 GB	2015-10-25 20:17:25	path	2016-04-28	26 weeks	MOVE TO MEDIUM DELETE					Yes
	e664-r0410-s01-c00.xtc	XTC	15 GB	2015-10-25 20:17:17	path	2016-04-28	26 weeks						Yes
	e664-r0410-s02-c00.xtc	XTC	15 GB	2015-10-25 20:17:20	path	2016-04-28	26 weeks						Yes
	e664-r0410-s03-c00.xtc	XTC	15 GB	2015-10-25 20:17:36	path	2016-04-28	26 weeks						Yes
	e664-r0410-s04-c00.xtc	XTC	15 GB	2015-10-25 20:18:01	path	2016-04-28	26 weeks						Yes
	e664-r0410-s05-c00.xtc	XTC	15 GB	2015-10-25 20:18:41	path	2016-04-28	26 weeks						Yes

Data Collection Statistics (Oct 2009 - May 2014)

As of the beginning of May 2014
Nearly **400** experiments and **500** individual users
Over **6.7 PB** of data, **140,000** DAQ runs, and **332,000** experimental files



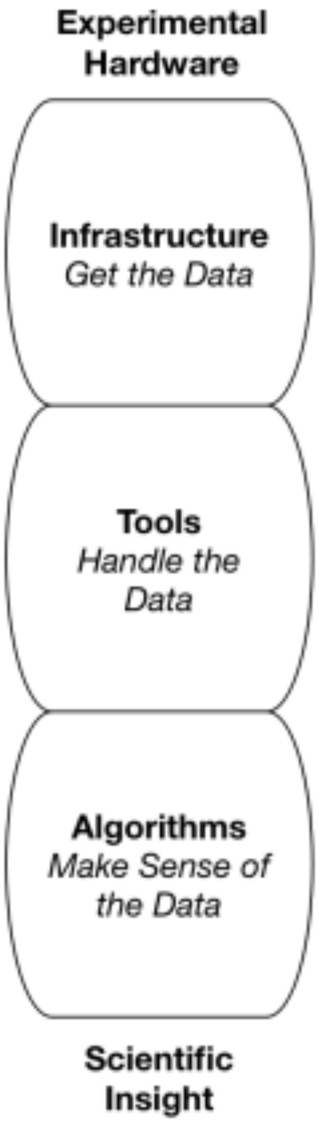
LCLS Data Strategy: Drivers

- **LCLS-II Upgrade**
  - The high repetition rate (1-MHz) and, above all, the potentially very high data throughput (100GB/s) generated by LCLS-II will require a major upgrade of the data acquisition system and increased data processing capabilities
- **Fast feedback**
  - Experience has shown that a capable real-time analysis is critical to the users' ability to take informed decisions during an LCLS experiment
    - Powerful fast feedback (~ minute or faster timescales) capabilities reduce the time required to complete the experiment, improve the overall quality of the data, and increase the success rate of experimentals
- **Time to science**
  - Sophisticated analysis frameworks can reduce significantly the time between experiment and publication, improving productivity LCLS science community
- **No user left behind**
  - Most of the advanced algorithms for analysis of the LCLS science data have been developed by external groups with enough resources to dedicate to a leading edge computing effort
    - Smaller groups with good ideas may be hindered in their ability to conduct science by not having access to these advanced algorithms
    - LCLS support for externally developed algorithms and, possibly, development of in-house algorithms for some specific science domains, would alleviate this problem

LCLS Data Strategy: Approach

Strategy organized into a three-pronged approach aligned with the following areas:

- **Infrastructure**
  - Includes the systems for processing and managing the LCLS data: computing farm, disk and tape storage, data movers, experiment portal
- **Tools**
  - Includes all the core software needed by the LCLS users to access and analyse their data: build tools, documentation, version control, visualization, calibration, data persistency and basic data analysis algorithms like fitting and filtering
- **Algorithms**
  - Includes the adoption, development and support of advanced algorithms specific to the various LCLS scientific areas: better crystallography pipelines, diffuse scattering, single particle imaging, etc



Infrastructure Challenges (1)

- **Data Acquisition**
  - Current online, network based, event builder will stop working at high rates
  - Reading out images at full rate will not be feasible
- **Real Time Analysis**
  - LCLS experience has shown that the most effective way to perform real time analysis is allowing users to run their code against the data on disk (fast feedback storage layer)
  - Existing spindle-based storage technologies are too slow for the LCLS-II fast feedback layer
- **Data Storage**
  - SLAC tape archive system is approaching limits in overall storage capacity and throughput
    - Such limits are already observed at LCLS when archiving data from on-going experiments while serving concurrent user requests to restore files from tape
- **Data Management**
  - Some aspects of the current system, such as checksum calculations, HPSS interface, and lack of prioritization, will become limitations at higher data volumes

Infrastructure Challenges (2)

- **Data Processing**
  - We expect that LCLS-II will require peta to exascale HPC
  - Deploying and maintaining very large processing capacity at SLAC would require a significant increase in the capabilities of the existing LCLS and/or SLAC IT groups
- **Data Network**
  - SLAC recently upgraded its connection to ESNET from 10Gb/s to 100Gb/s
    - Primary reason for upgrading this link is to gain the ability to offload part of the LCLS science data processing to NERSC while keeping up with the DAQ
  - The 100Gb/s link will need to be upgraded to 1 Tb/s for LCLS-II
- **Data Format**
  - The translation step from XTC (*DAQ format*) to HDF5 (*users format*) will become a bottleneck in the future and LCLS-II should adopt a single data format
  - HDF5 de-facto standard for storing science data at light source facilities
    - In order to effectively replace XTC in LCLS, a couple of critical features are required: ability to read while writing, ability to consolidate multiple writers into a consistent virtual data set

LCLS-II Data Throughput, Data Storage and Data Processing Estimates

- **Examples LCLS-II 2020:**
  - 1 x 16 Mpixel ePix @ 360 Hz = 12 GB/s
  - 100K points fast digitizers @ 100kHz = 20 GB/s
  - 2 x 4 Mpixel ePix @ 5 kHz = 80 GB/s
  - Distributed diagnostics 1-10 GB/s range
- **Example LCLS-II 2025**
  - 3 beamlines x 2 x 4 Mpixel ePix @ 100 kHz = 4.8 TB/s
- **Data parameters scaling between LCLS-I and LCLS-II**

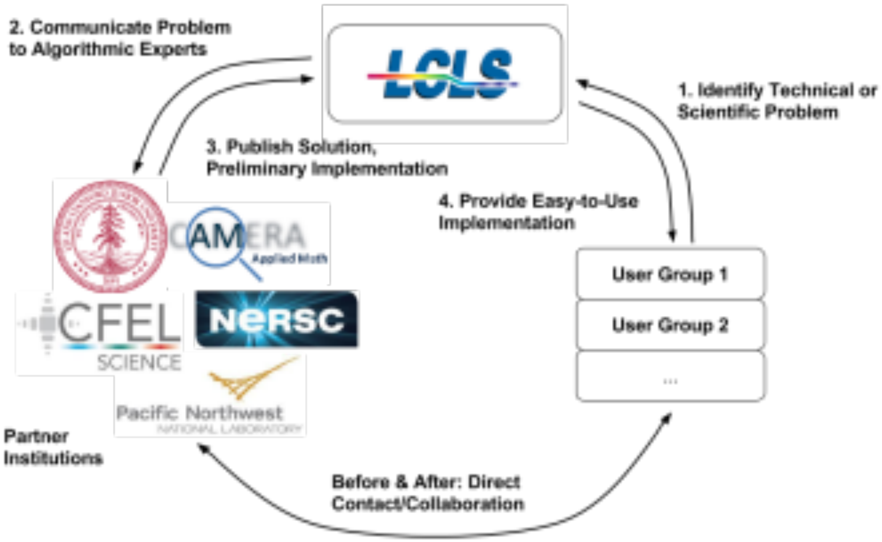
Parameter	**LCLS-I**	**LCLS-II 2020**	**LCLS-II 2025**
Average throughput	0.1 - 1 GB/s	2 - 20 GB/s	2 GB/s - 1.2 TB/s
Peak throughput	5 GB/s	100 GB/s	4.8 TB/s
Peak Processing	50 TFLOPS	1 PFLOPS	60 PFLOPS
Data Storage	5 PB	100 PB	6 EB

LCLS-II Data Analysis: Onsite and Offsite Components

- Data analysis strategy motivated by (1) very high data throughput generated by LCLS-II, (2) need to provide powerful fast feedback capabilities, (3) desire to reduce time between experiment and publication, and (4) need for the facility to develop algorithms for XFEL science and support externally developed ones

- **Software** - In-house: core framework, tools and core algorithms
  Collaborations: advanced algorithms (e.g. IOTA, ray tracing, diffuse scattering, M-TIP, AI/ML) - Stanford CS, LBL/CAMERA, LBL/MBIB, SLAC/PULSE, PNNL, LANL

- **Infrastructure** - Onsite: data reduction pipeline and real time analysis, 1-10 PFLOPS processing, 100 petabyte storage
  Offsite: NERSC supercomputers (1 EFLOPS)

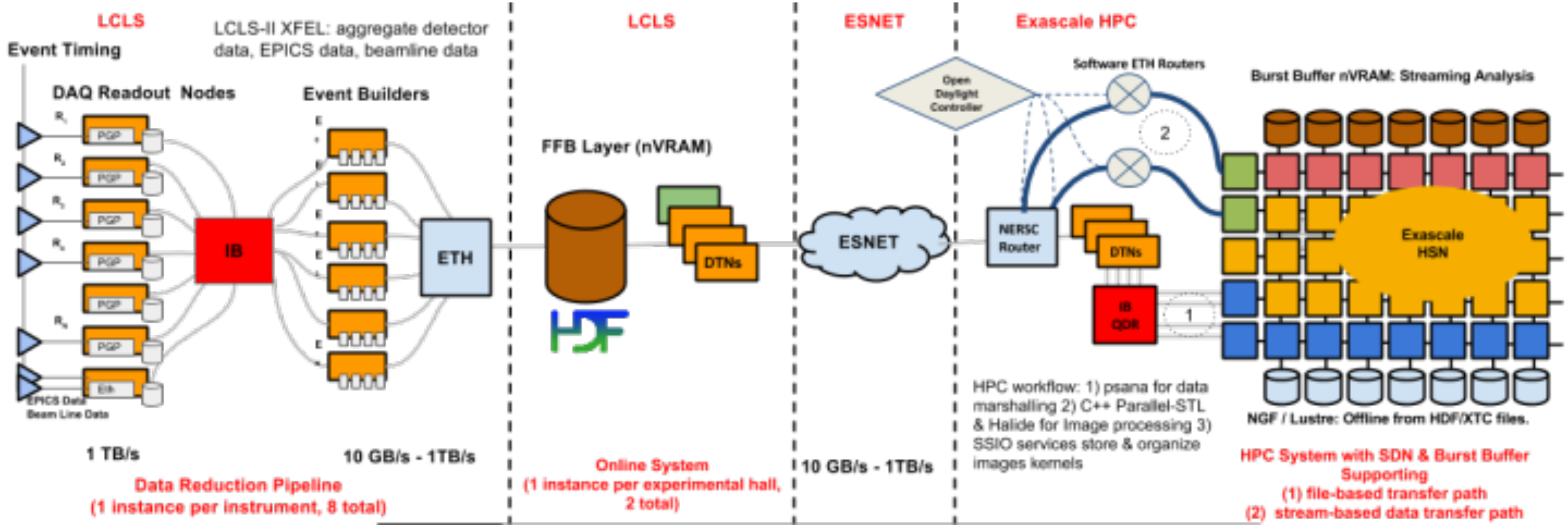

2. Communicate Problem to Algorithmic Experts
LCLS
1. Identify Technical or Scientific Problem
3. Publish Solution, Preliminary Implementation
4. Provide Easy-to-Use Implementation
CAMERA
Applied Math
CFEL SCIENCE
NERSC
Pacific Northwest NATIONAL LABORATORY
User Group 1
User Group 2
...
Partner Institutions
Before & After: Direct Contact/Collaboration

NERSC collaboration avoids need to scale the HPC needs to the highest demand (exascale) experiments while maintaining critical capabilities at SLAC

Data Analytics at the Exascale for Free Electron Lasers

$10M over 4 years: 40% SLAC (LCLS, CS), 20% LANL, 40% LBL (CAMERA, MBIB, NERSC)

High data throughput experiments	Single Particle Imaging	LCLS data analysis framework	Infrastructure
Algorithmic improvement with IOTA (Integration Optimization, Triage, and Analysis) and ray tracing - Use example test-case of Serial Femtosecond Crystallography	Algorithmic advances with M-TIP (Multi-Tiered Iterative Phasing)	Porting psana to supercomputer architecture, change parallelization technology to allow scaling from hundreds of cores (now) to hundred of thousands of cores	Data flow from SLAC to NERSC over ESnet
Sauter, Brewster - LBNL/MBIB	Zwart, Donatelli, Sethian - LBNL/CAMERA	Aiken - Stanford/SLAC CS, Shipman - LANL, O'Grady - SLAC/LCLS	Perazzo - SLAC/LCLS, Skinner - LBNL/NERSC, Guok - LBNL/ESnet



LCLS-II XFEL: aggregate detector data, EPICS data, beamline data

Data Reduction Pipeline
(1 instance per instrument, 8 total)

Online System
(1 instance per experimental hall, 2 total)

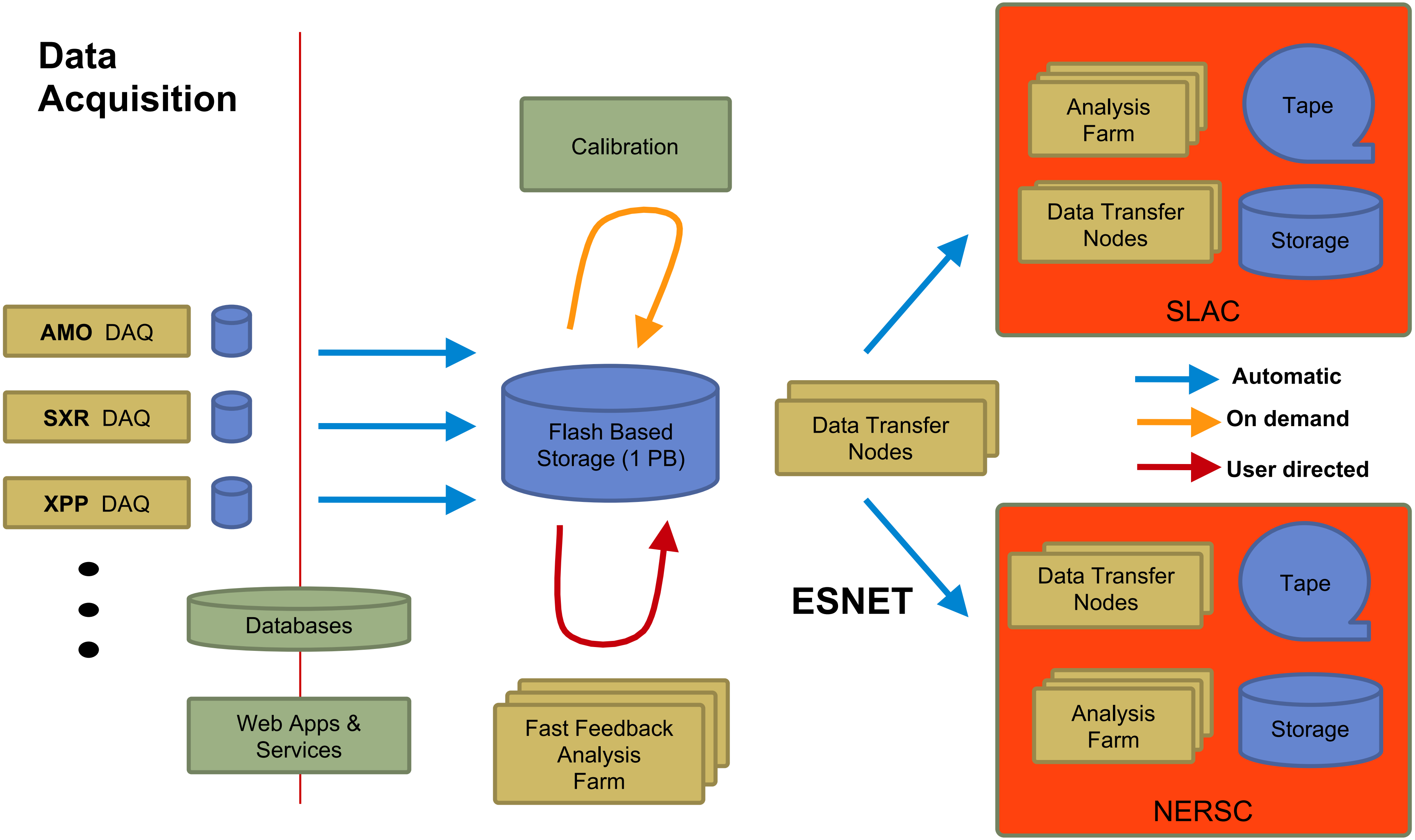
HPC workflow: 1) psana for data marshalling 2) C++ Parallel-STL & Halide for Image processing 3) SSIO services store & organize images kernels

HPC System with SDN & Burst Buffer Supporting
(1) file-based transfer path
(2) stream-based data transfer path

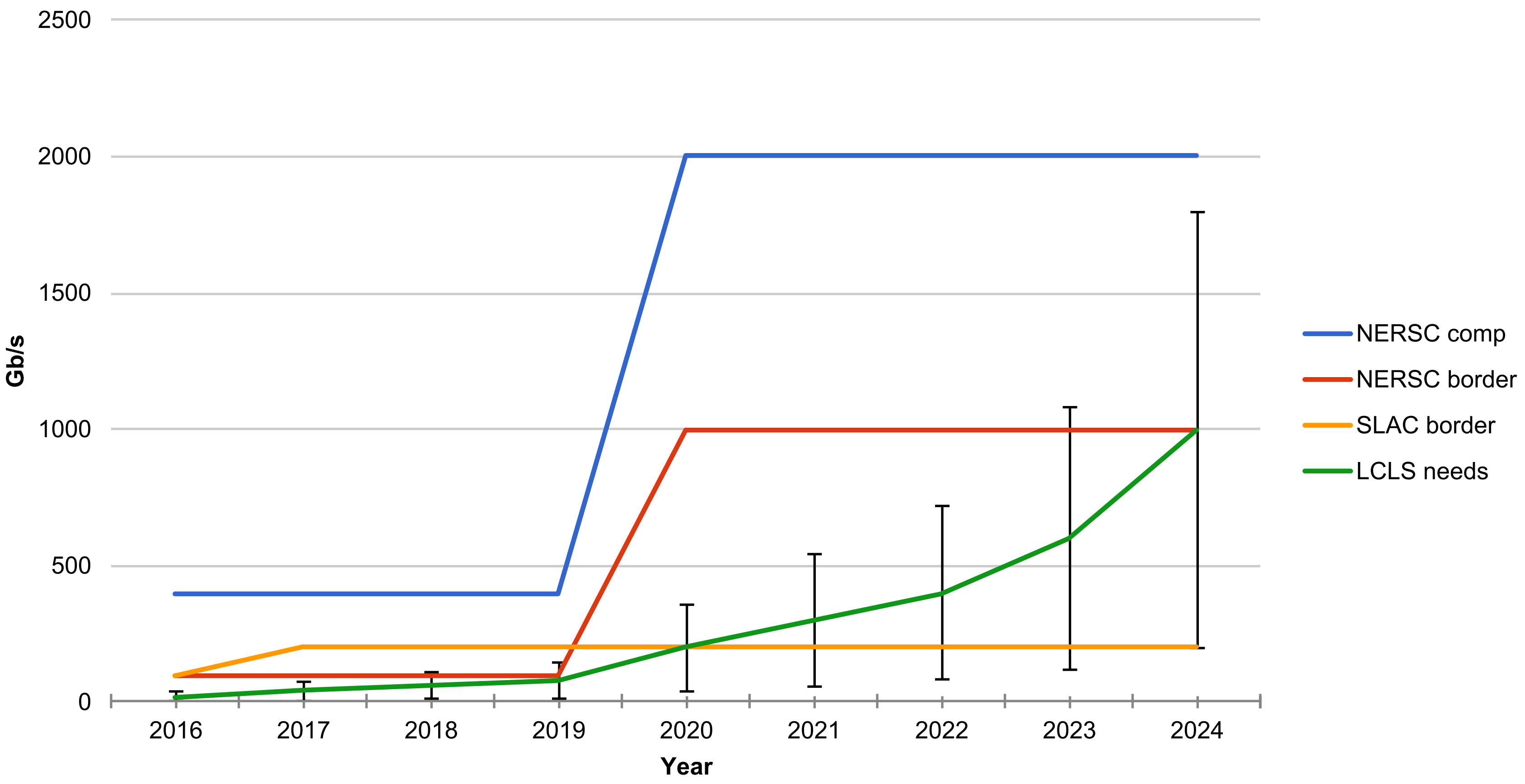
Why It’s Important

- Powerful fast feedback during high throughput experiments
- Enable broader user base, and short time to publication
- Step change in advanced algorithms development capabilities for SFX, diffuse scattering, SPI
- Provide seed funding for the new Computer Science division at SLAC
- Strengthen the collaborations between LCLS and CAMERA, MBIB, NERSC (previously carved out of operations funding), ESnet and Stanford
- Collaboration with supercomputer facilities avoids need to scale LCLS HPC to the highest demand (exascale) experiments, while maintaining critical capabilities at SLAC

Data Systems Architecture: Evolution



LCLS Network needs and border links



Core Technologies

Infiniband wherever possible (i.e. on short distances)
NVRAM devices and NVMe over fabric
100 Gb/s and 400 Gb/s Ethernet between experimental halls and data center(s)
Many cores CPUs (KNL) - see NERSC slides for future exascale architectures
HDF5 for data format
SDN
Python as programming language with C/C++ kernels
Main open question: file system technology - Lustre? Object storage? Other?