Storage Systems Requirements for Massive Throughput Detectors at Light Sources

35th International Conference on Massive Storage Systems and Technology (MSST 2019)
May 21st 2019

Amedeo Perazzo
SLAC National Accelerator Laboratory
LCLS Controls & Data Systems Division Director

U.S. DEPARTMENT OF ENERGY
Office of Science

SLAC NATIONAL ACCELERATOR LABORATORY

Outline

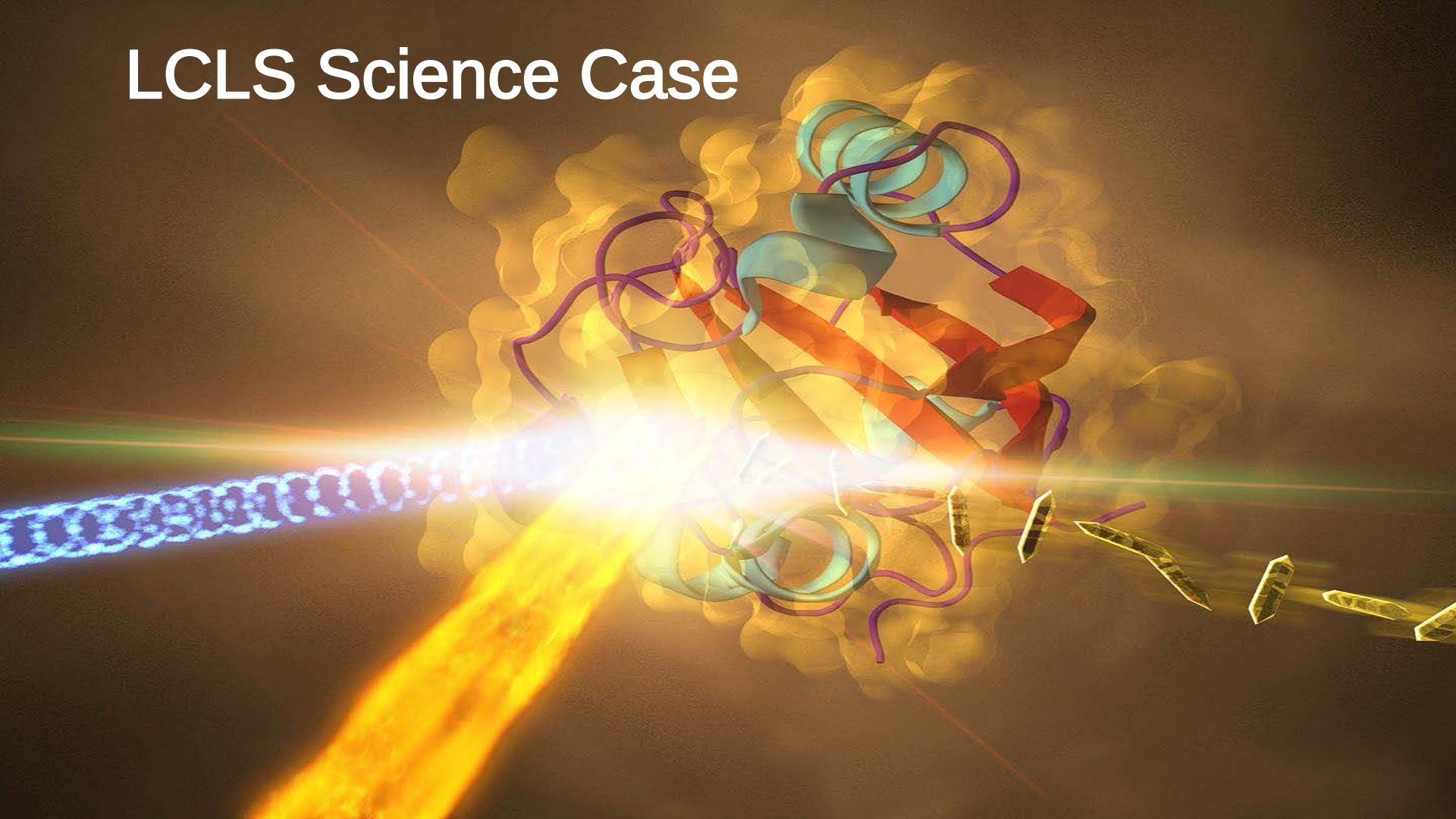
LCLS science case, requirements

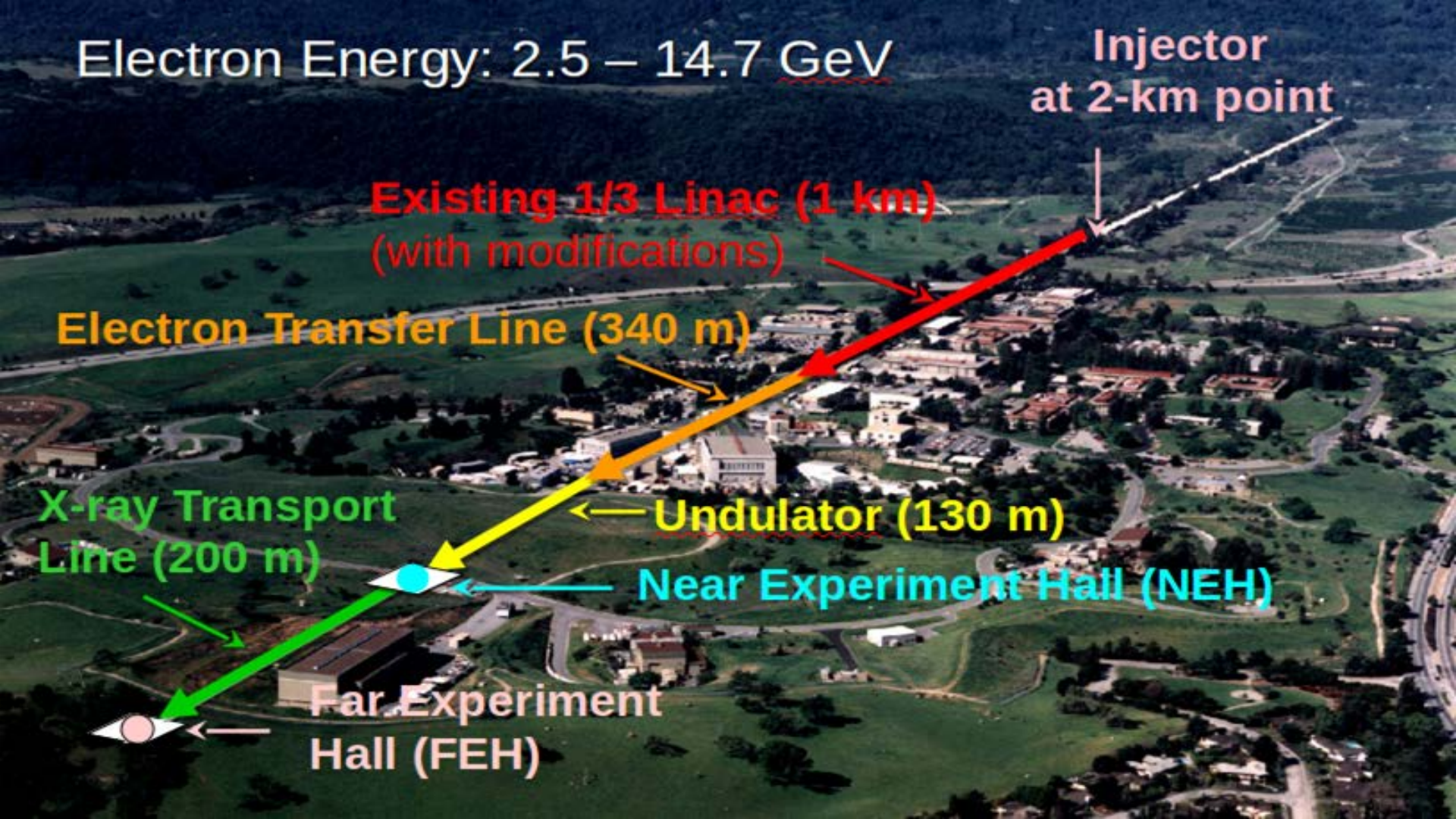
Storage and throughput projections

Current design

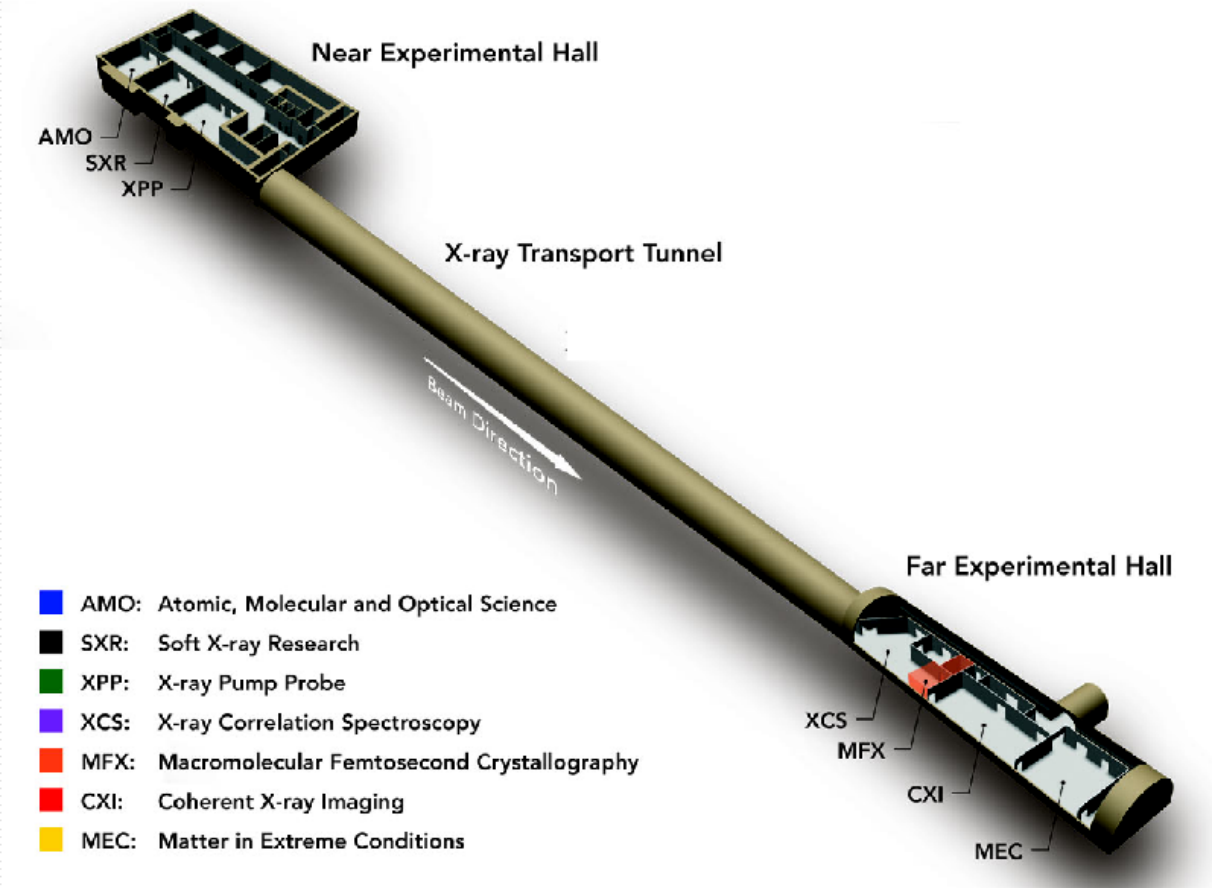
Possible storage innovations that could benefit the LCLS upgrade

LCLS Science Case




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)
X-ray Transport Line (200 m)
Undulator (130 m)
Near Experiment Hall (NEH)
Far Experiment Hall (FEH)

LCLS Instruments



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

Data Analytics for high repetition rate Free Electron Lasers

FEL data challenge:

- **Ultrafast X-ray pulses** from LCLS are used like flashes from a high-speed strobe light, producing stop-action movies of atoms and molecules
- Both **data processing** and **scientific interpretation** demand intensive computational analysis

LCLS-II will increase **data throughput by three orders of magnitude** by 2025, creating an exceptional scientific computing challenge

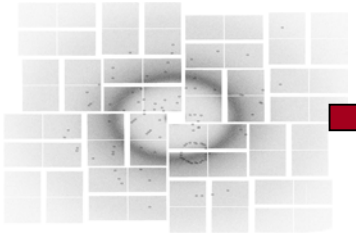
LCLS-II represents SLAC's largest data challenge by far

Example of LCLS Data Analytics: The Nanocrystallography Pipeline

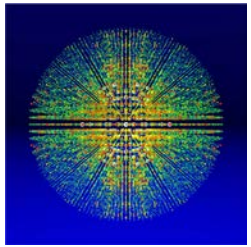
SLAC

Serial Femtosecond Crystallography (SFX, or nanocrystallography): huge benefits to the study of **biological macromolecules**, including the availability of femtosecond time resolution and the avoidance of radiation damage under physiological conditions (**"diffraction-before-destruction"**)

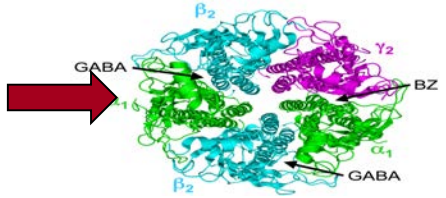
Megapixel detector



X-Ray Diffraction Image



Intensity map from multiple pulses



Electron density (3D) of the macromolecule

Well understood computing requirements

Significant fraction of LCLS experiments (~90%) use large area imaging detectors

Easy to scale: processing needs are linear with the number of frames

Must extrapolate from 120Hz (today) to 5-10 kHz (2022) to >50 kHz (2026)

Computing Requirements for Data Analysis: a *Day in the Life of a User* Perspective

SLAC

- During **data taking**:
  - Must be able to get real time (~1 s) **feedback** about the **quality of data taking**, e.g.
    - Are we getting all the required detector contributions for each event?
    - Is the hit rate for the pulse-sample interaction high enough?
  - Must be able to get **feedback** about the **quality of the acquired data** with a latency lower than the typical lifetime of a measurement (~10 min) in order to optimize the experimental setup for the next measurement, e.g.
    - Are we collecting enough statistics? Is the S/N ratio as expected?
    - Is the resolution of the reconstructed electron density what we expected?
- During **off shifts**: must be able to run **multiple passes** (> 10) of the full analysis on the data acquired during the previous shift to optimize analysis parameters and, possibly, code in preparation for the next shift
- During **4 months** after the experiment: must be able analyze the raw and intermediate data on **fast access storage** in preparation for publication
- **After 4 months**: if needed, must be able to **restore** the archived data to test new ideas, new code or new parameters

The Challenging Characteristics of LCLS Computing

SLAC

1. **Fast feedback** is essential (seconds / minute timescale) to reduce the time to complete the experiment, improve data quality, and increase the success rate
2. **24/7 availability**
3. **Short burst** jobs, needing very short startup time
4. **Storage** represents significant fraction of the overall system
5. **Throughput** between storage and processing is critical
6. Speed and flexibility of the **development cycle** is critical - *wide variety of experiments, with rapid turnaround, and the need to modify data analysis during experiments*

Example data rate for LCLS-II (early science)

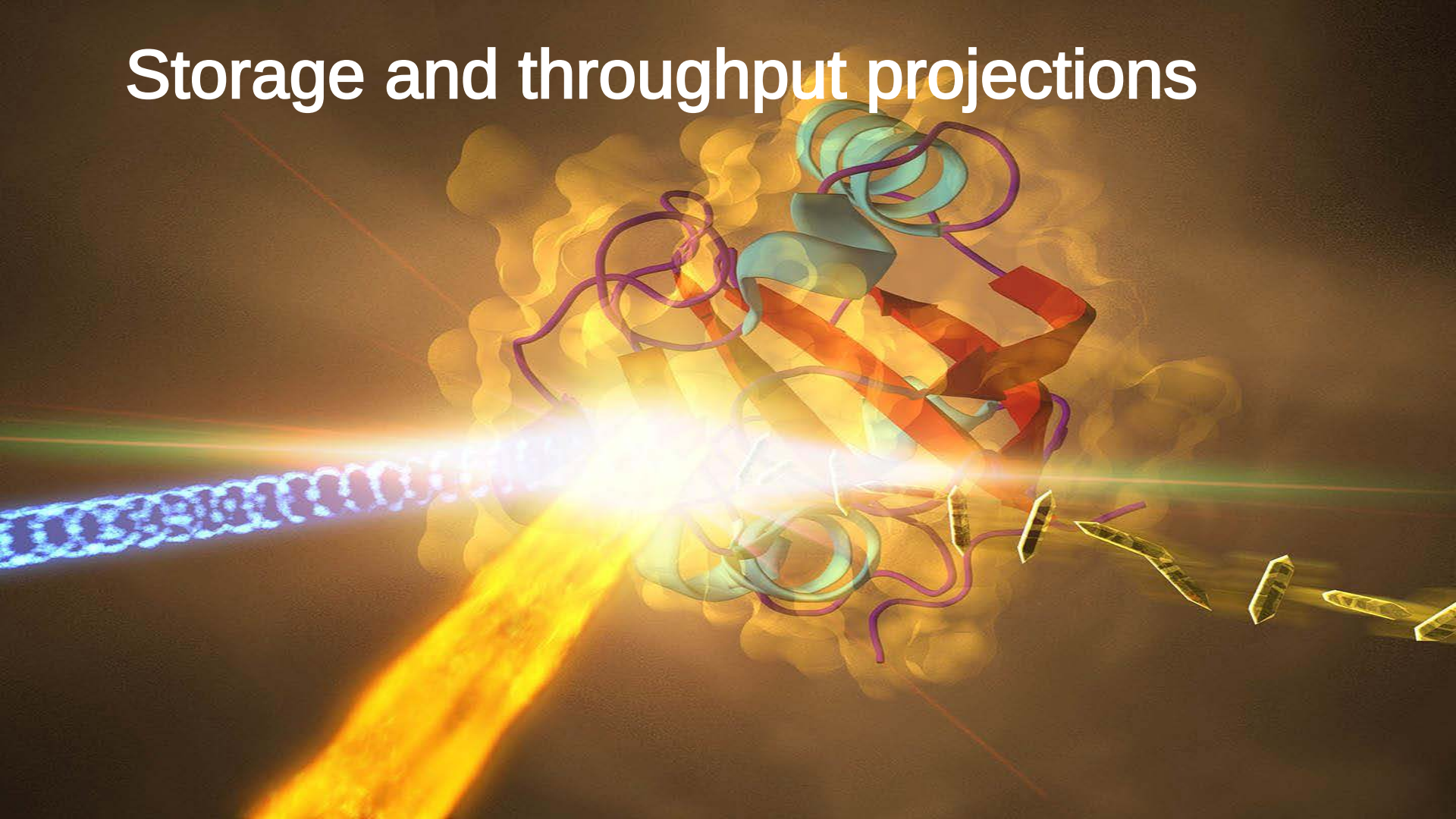
- 1 x 4 Mpixel detector @ 5 kHz = **40 GB/s**
- 100K points fast digitizers @ 100kHz = **20 GB/s**
- Distributed diagnostics 1-10 GB/s range

Example LCLS-II and LCLS-II-HE (mature facility)

- 2 planes x 4 Mpixel ePixUHR @ 100 kHz = **1.6 TB/s**

Sophisticated algorithms under development within ExaFEL (e.g., M-TIP for single particle imaging) will require exascale machines

Storage and throughput projections



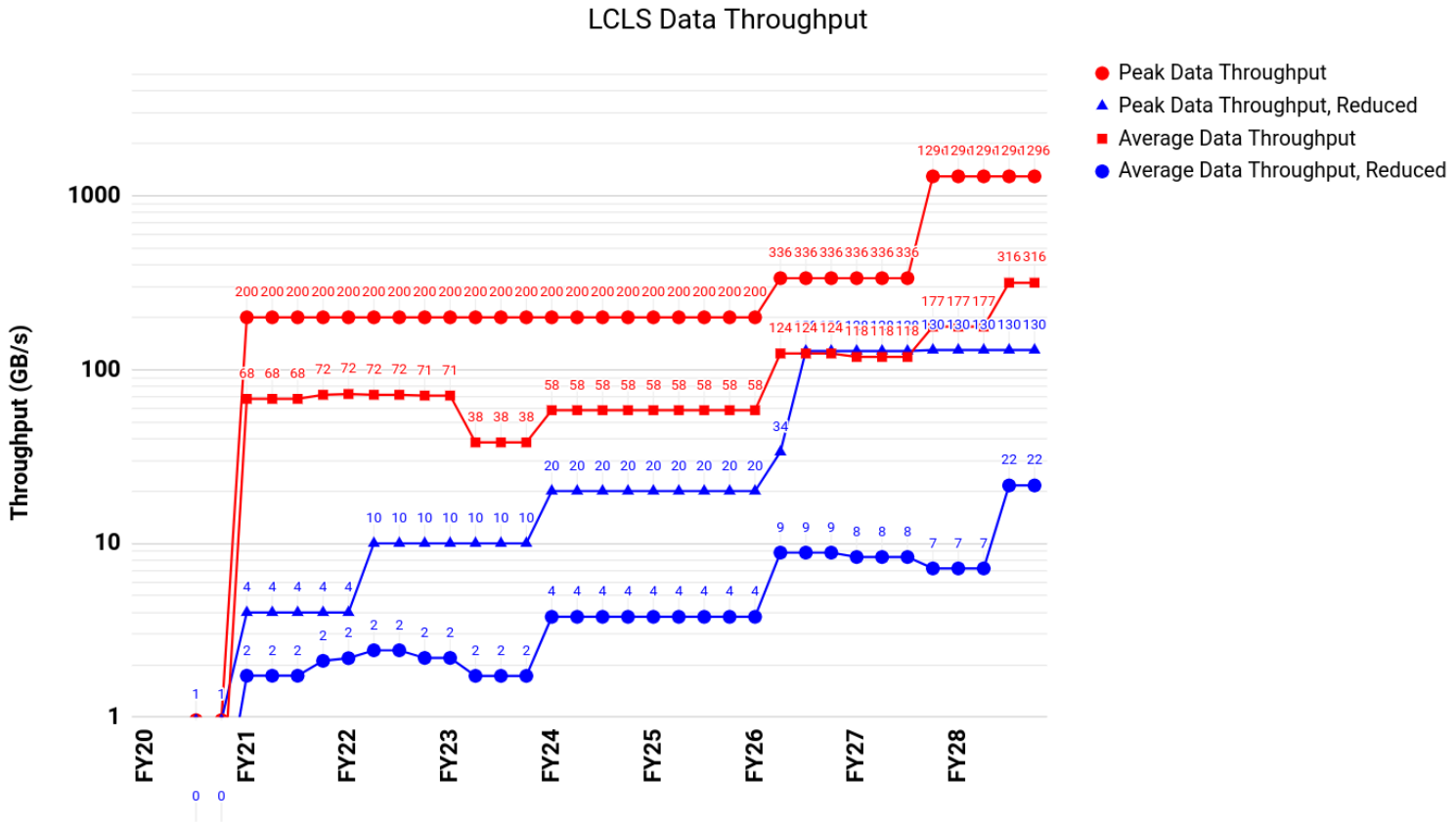
Process for determining future projections

Includes:

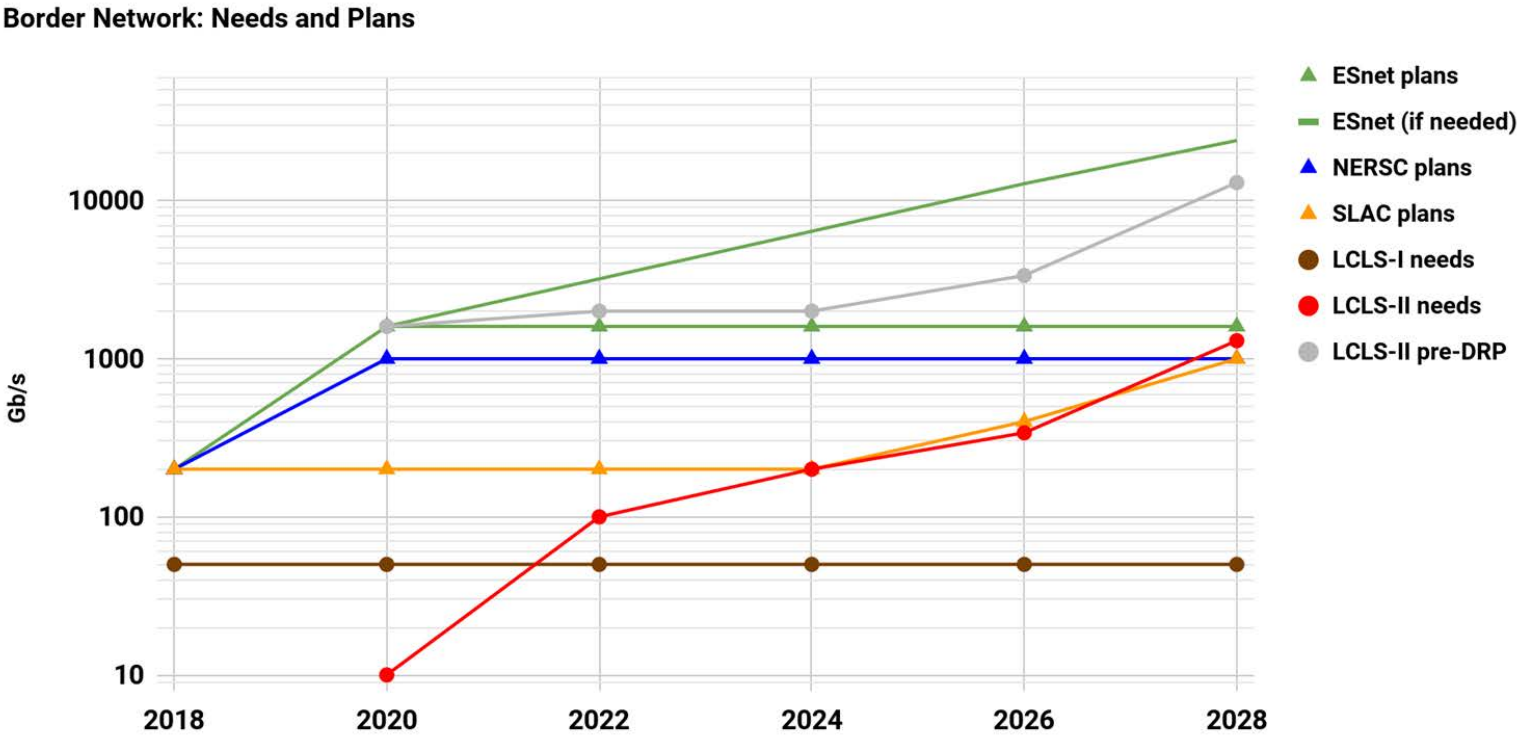
1. **Detector rates** for each instrument
2. **Distribution of experiments** across instruments (as function of time, ie as more instruments are commissioned)
3. Typical **uptimes** (by instruments)
4. **Data reduction** capabilities based on the experimental techniques
5. Algorithm **processing times** for each experimental technique

Undulator	Instrument	Endstation	Technique	Detector	Detector Size	Detector Rate (Hz)	Data Rate (aggregate) (GB/s)	Ultilization Factor (0-1)	Data Reduction Type (1st Cut)	DR Factor (1st cut)	Data Reduction Type (Optimistic))	DR Factor (Optimistic)	FY20 Q1	FY20 Q2	FY20 Q3	FY20 Q4	FY21 Q1	FY21 Q2	FY21 Q3	F
SXU	NEH 1.1	DREAM	COLTRIMS	Digitizer	800000	100000	160.0	0.75	Zero suppression	0.020	Peak Finding	0.0020		1.00	1.00	0.50	0.25	0.25	0.25	
SXU	NEH 1.1	DREAM	Time of Flight	Digitizer	1000000	100000	200.0	0.75	Zero suppression	0.020	Peak Finding	0.0020				0.13	0.13	0.13	0.06	
SXU	NEH 1.1	LAMP	Time of Flight	Digitizer	1000000	100000	200.0	0.75	Zero suppression	0.020	Peak Finding	0.0020				0.13	0.13	0.13	0.06	
SXU	NEH 1.1	LAMP	Imaging	SXR Imag. + Digi.	4000000	10000	82.0	0.45	Veto	0.100	N.A.	0.1000							0.13	
SXU	NEH 2.2	LJE	XAS / XES	TES	1000	100000	20.0	0.60	Zero suppression	0.100	Binning	0.0000								
SXU	NEH 2.2	LJE	XAS / XES	TES	10000	100000	200.0	0.60	Zero suppression	0.100	Binning	0.0000								
SXU	NEH 2.2	LJE	XAS / XES	RIXS-ccd	4096	1000	0.0	0.60	N.A.	1.000	Accumulating	0.0010				0.25	0.50	0.25	0.25	
SXU	NEH 2.2	RIXS	IXS / RIXS	RIXS-ccd	4096	1000	0.0	0.60	N.A.	1.000	Accumulating	0.0010						0.13	0.13	
SXU	NEH 2.2	RIXS	XRD / RXRD	SXR Imaging	1000000	10000	20.0	0.60	ROI	0.100	Accumulating	0.0001						0.06	0.06	
SXU	NEH 2.2	RIXS	XPCS	SXR Imaging	1000000	10000	20.0	0.60	Compression	0.500		0.1000						0.06	0.06	
SXU	NEH 1.2	---	X-ray/X-ray	SXR Imaging	1000000	10000	20.0	0.30	ROI	0.100	Binning	0.0001								
SXU	NEH 1.2	---	Imaging	epix100-HR + Digi.	4000000	5000	42.0	0.45	Veto	0.100	N.A.	0.1000								
SXU	NEH 1.2	---	XAS / XES	RIXS-ccd	4096	1000	0.0	0.60	N.A.	1.000	Accumulating	0.0010								

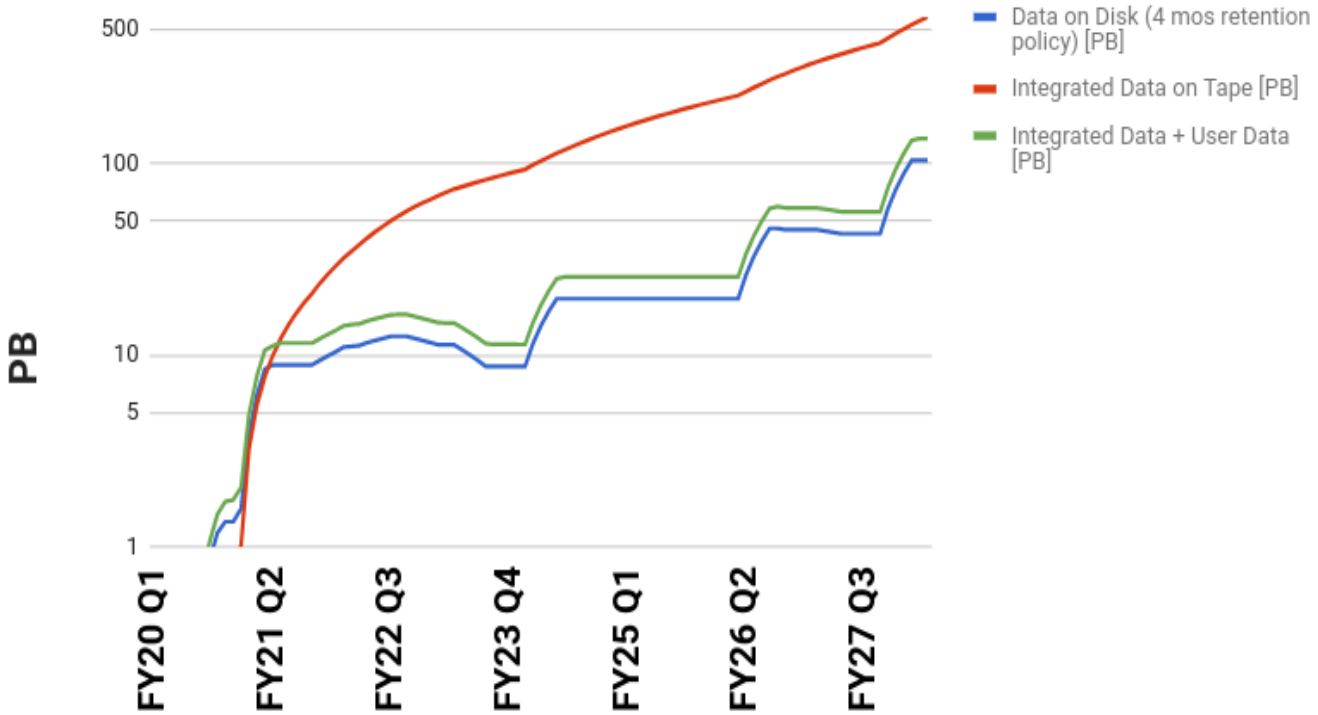
Data Throughput Projections



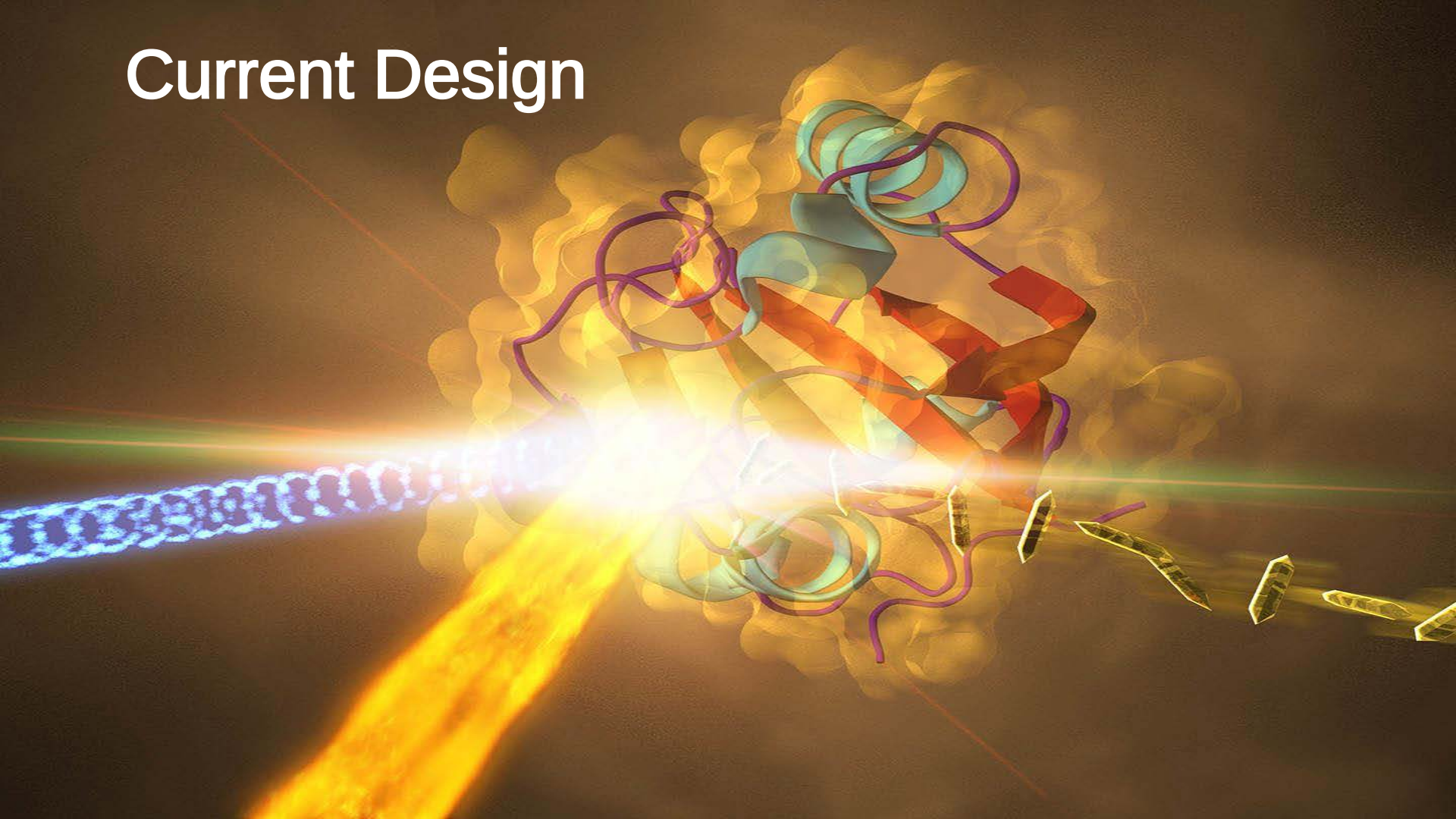
Offsite Data Transfer: Needs and Plans



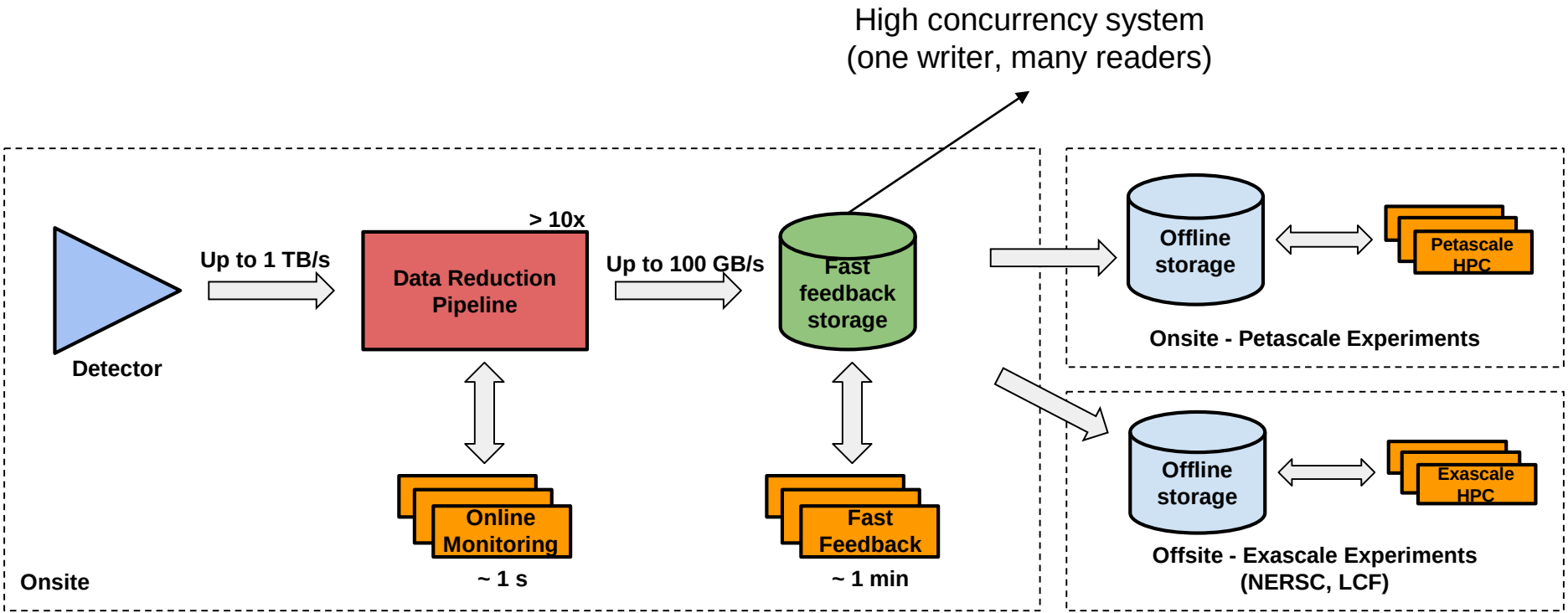
Storage and Archiving Projections



Current Design

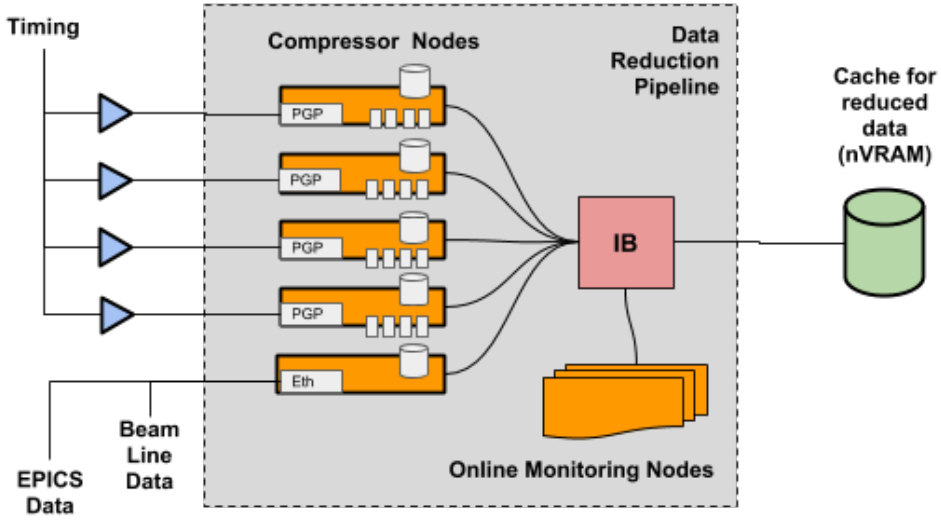


LCLS-II Data Flow



Data Reduction Pipeline

- Besides cost, there are **significant risks** by not adopting on-the-fly data reduction
  - Inability to move the data offsite**,** system complexity (robustness, intermittent failures)
- Developing toolbox of techniques (**compression, feature extraction, vetoing**) to run on a **Data Reduction Pipeline**
- Significant **R&D effort**, both engineering (throughput, heterogeneous architectures) and scientific (real time analysis)



Without on-the-fly data reduction we would face unsustainable hardware costs by 2026

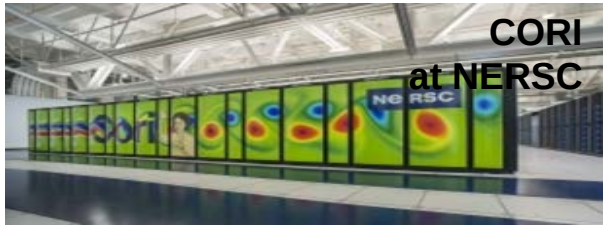
Make full use of national capabilities

SLAC

LCLS-II will require access to **High End Computing Facilities (NERSC and LCF)** for highest demand experiments (exascale)

ESnet Energy Sciences Network

NERSC

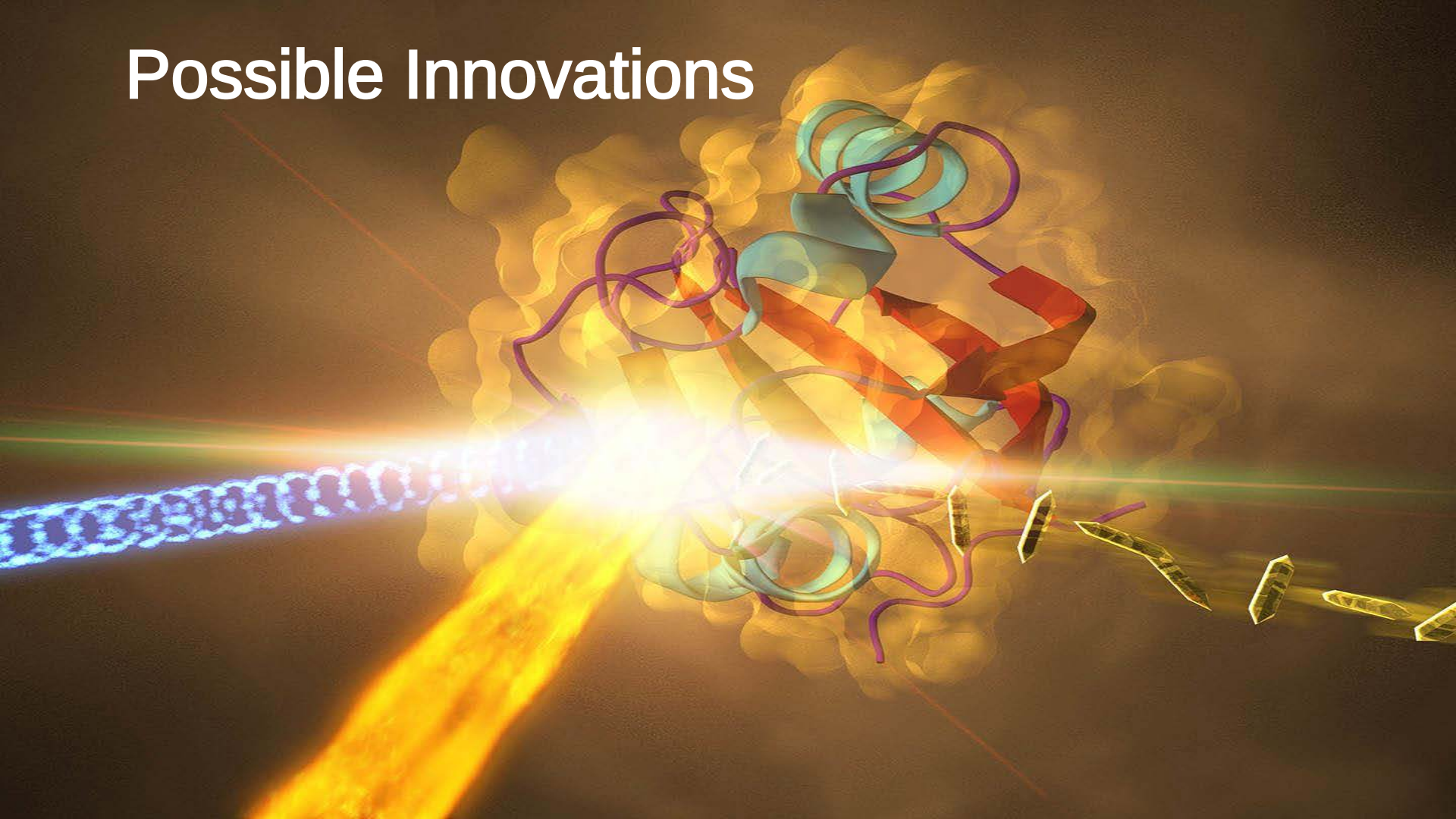


Photon Science Speedway

Stream science data files on-the-fly from the LCLS beamlines to the NERSC supercomputers via ESnet

Very positive partnership to date, informing our future strategy

Possible Innovations

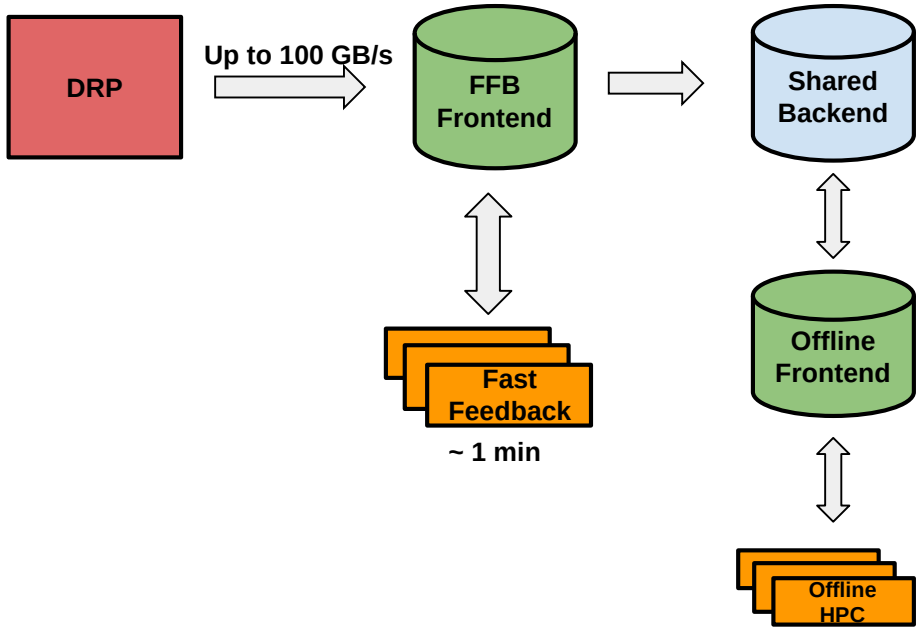


Shared backend between fast feedback (FFB) and offline storage layers

Potential of simplifying the data management system, improve robustness and performance

Key ingredients:

- Offline compute must **not affect FFB** performance
- File system transparently handles **data movement** and coherency between different frontends (cache) and the shared storage (as opposed to the data management system handling the data flow)

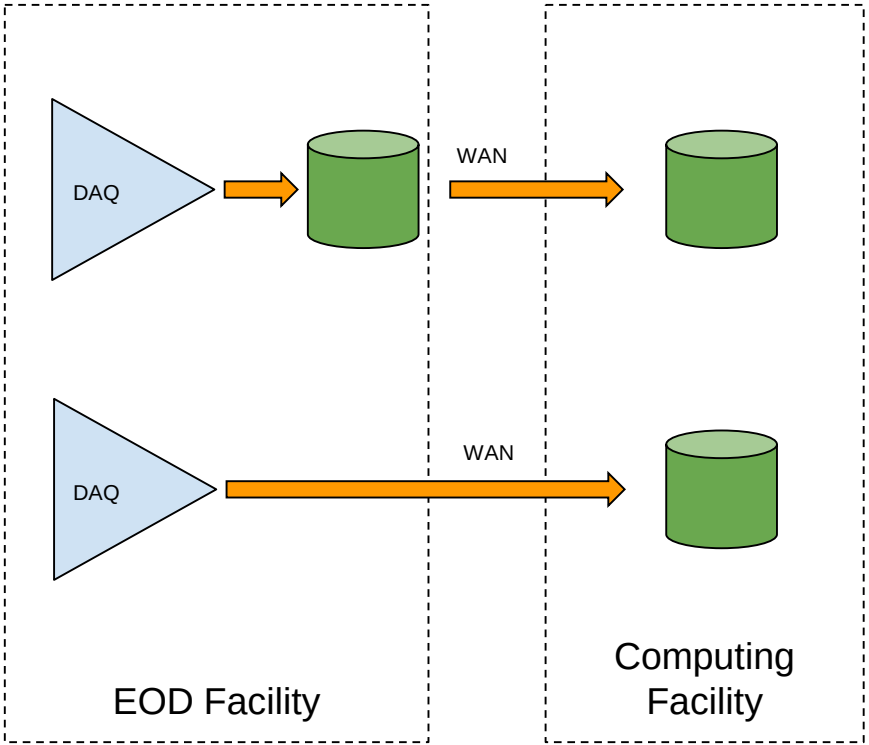


Remote mount over WAN

Ability to write directly from the data reduction pipeline to the remote computing facility

Potential of simplifying data management and reduce latency

Must handle throughput, network latency and network glitches

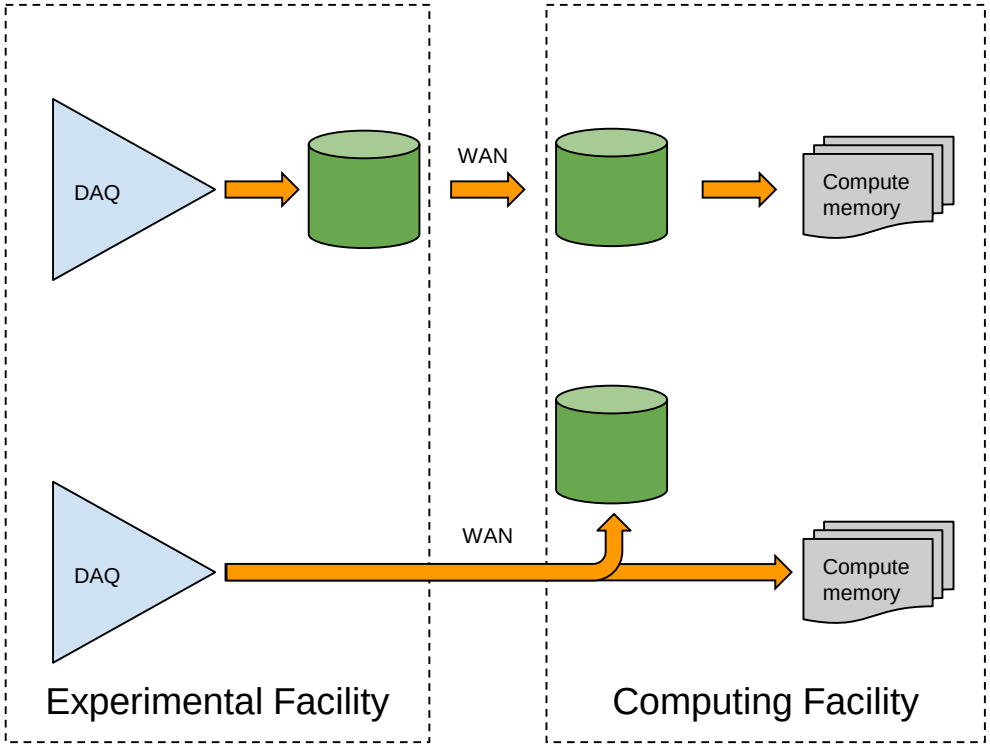


Zero-copy data streaming from front end electronics to computer memory

While data are being transferred to be analyzed, a copy of the same data must be made persistent for later analysis and archiving

This requires either:

- Persistent storage layer in the data path or
- the ability to send the data directly to the computer where it will be analyzed while replicating the data to persistent storage, without the need for an additional transfer ⇨ potential of significantly reducing latency



Conclusions

We have developed a base design for the LCLS storage system upgrades for LCLS-II by 2021, but…

we are looking into more advanced ways of handling storage in preparation for the further deluge of data (> 1 TB/s) expected after the 2026 LCLS-II-HE upgrade

Suggestions welcome!