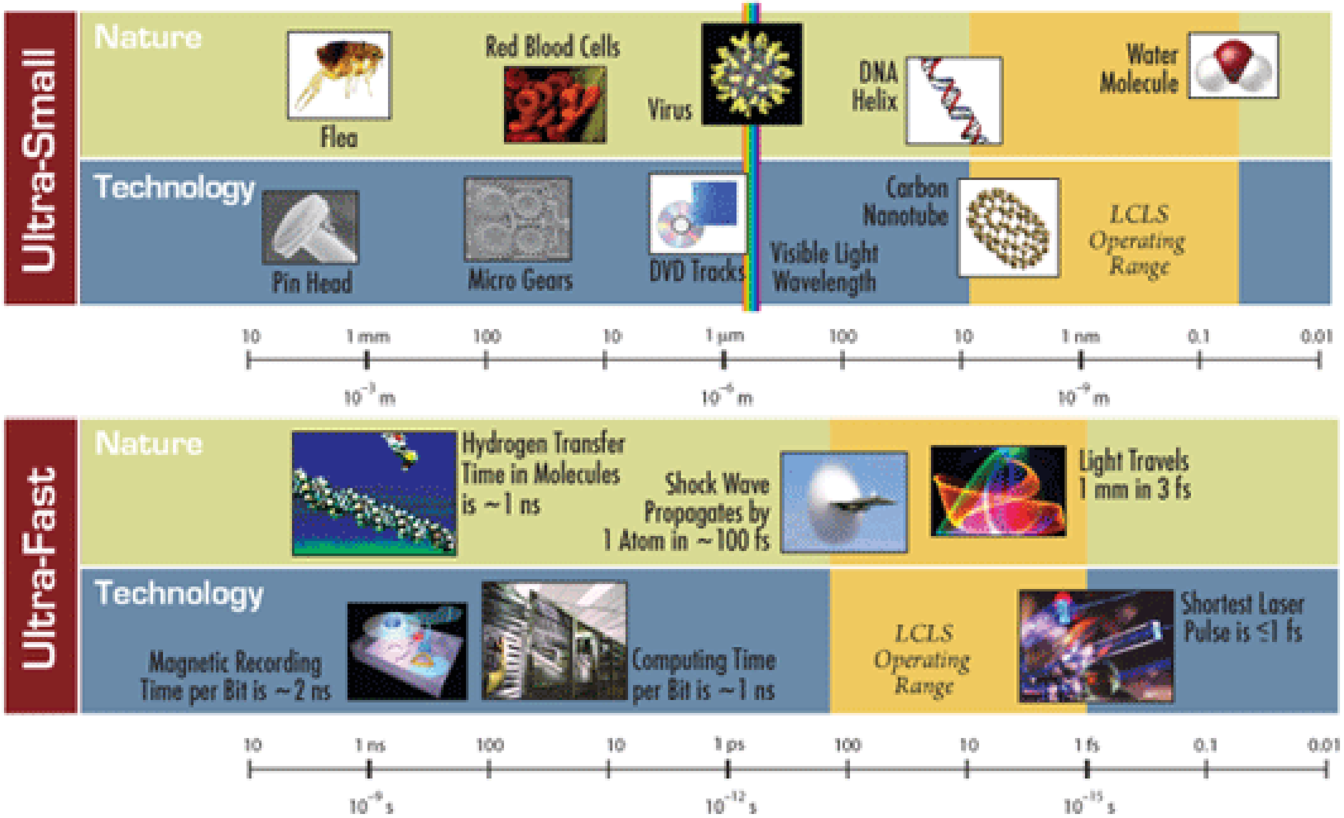
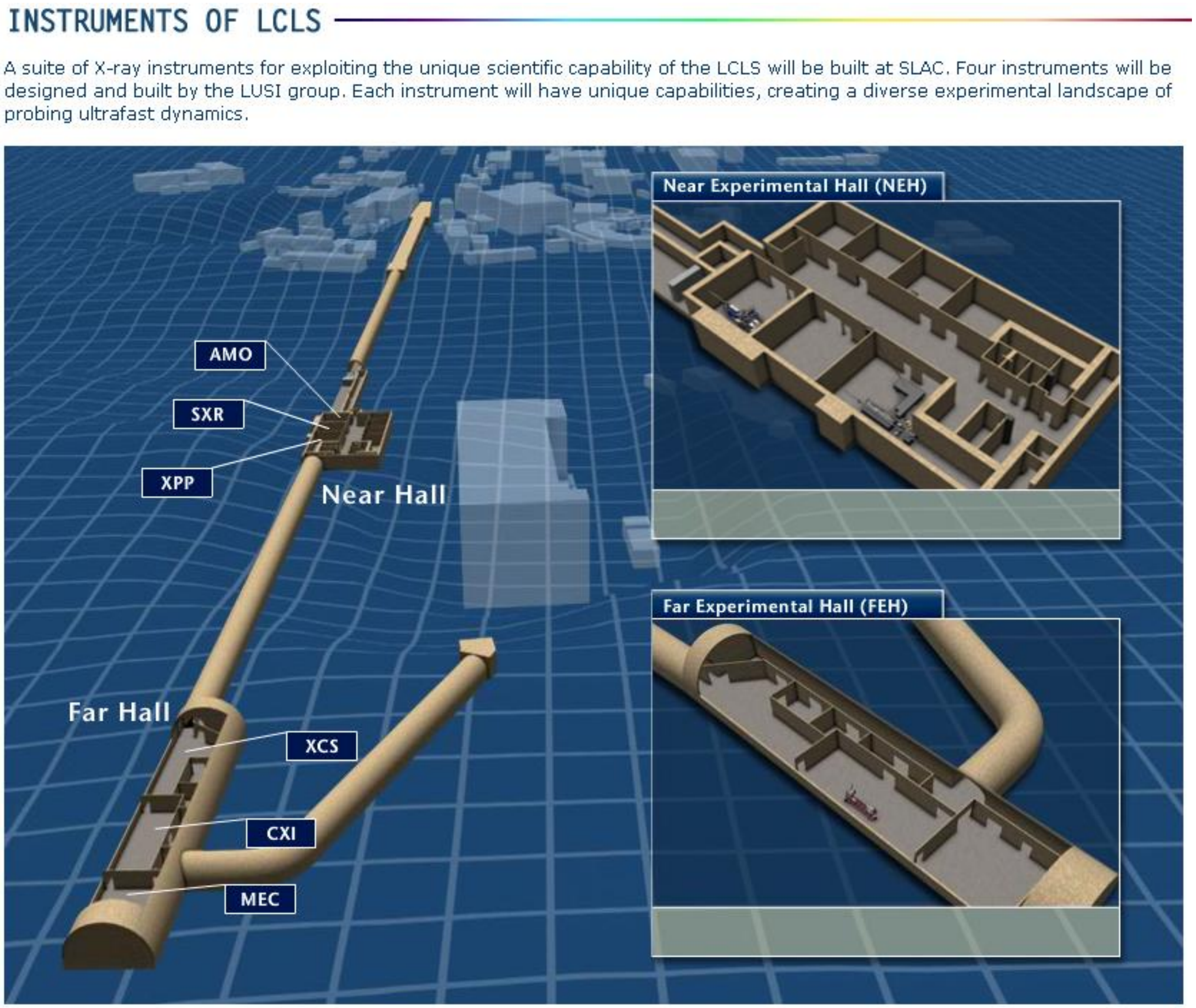
LCLS OFFLINE Computing

Igor Gaponenko, Andrei Salnikov, Marc Messerschmidt
SLAC National Accelerator Laboratory, USA

SLAC/PPA Computing Department Meeting
March 25, 2010

A New Kind of Laser - The LCLS will create X-ray pulses that can capture images of atoms and molecules in motion



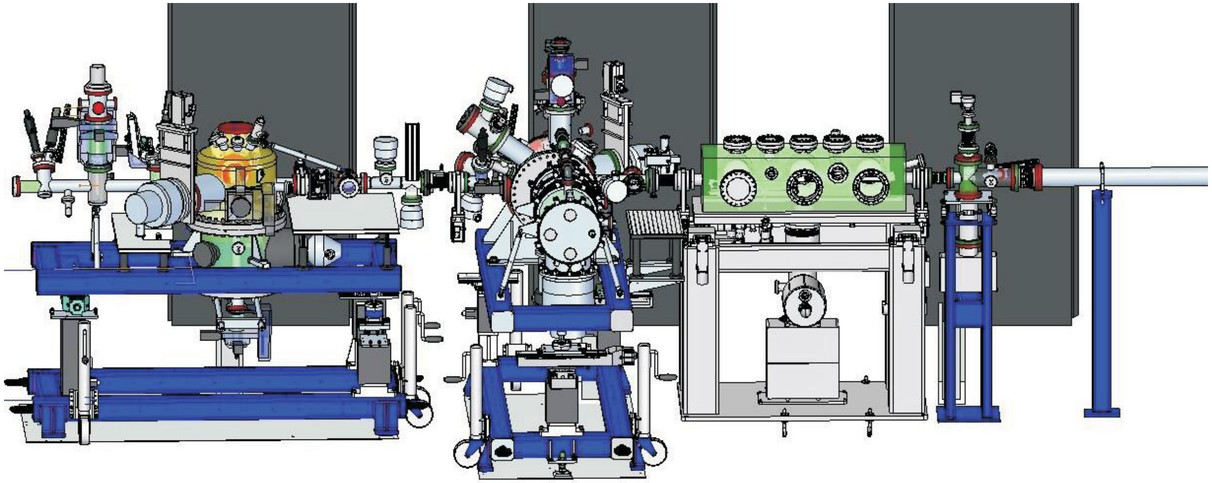


INSTRUMENTS OF LCLS

A suite of X-ray instruments for exploiting the unique scientific capability of the LCLS will be built at SLAC. Four instruments will be designed and built by the LUSI group. Each instrument will have unique capabilities, creating a diverse experimental landscape of probing ultrafast dynamics.

AMO (2009)

- **Atomic Molecular Optical Sciences (AMO/CAMP) (2009)**
- **Investigate multiphoton and high- field x-ray interaction with atoms, molecules and clusters**
  - Sequential and simultaneous multiphoton ionization/excitation
  - Accessible intensity on verge of high-field regime
- **Study time-resolved phenomena in atoms, molecules and clusters using ultrafast x-rays**
  - Inner-shell side band experiments
  - Photoionization of aligned molecules
  - Temporal evolution of state-prepared systems



SXR (2010)

- The beam line for Soft X-ray (SXR) Materials Science will enable the high brightness and timing capability of the LCLS to be applied to scattering and imaging experiments that require the use of soft x-rays.

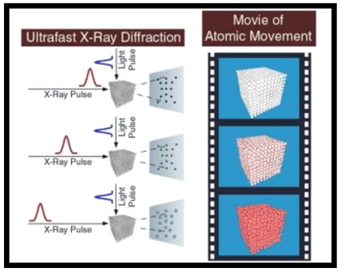
- Monochromator (500eV-2000ev) covers several important K- and L-edges of the second and third row elements for resonant excitation with a resolving power in the order of 5000, but the monochromator can also deliver beam in the non-monochromator mode where samples can be studied by XAS in transmission mode and detected in a single shot setup at the monochromator exit slit position.

- The SXR beam line is different in that it does not have a stationary end station. The consortium and collaborators will roll-up and connect different end stations and detectors for experiments by general users.

XPP (2010)

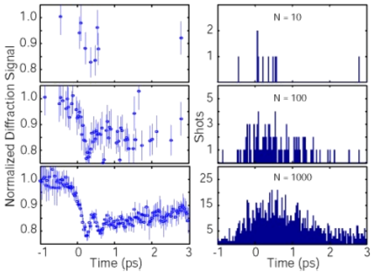
- **X-ray Pump Probe (XPP)**
  - The instrument will predominantly use a fast optical laser to generate transient states of matter, and the hard X-ray pulse from LCLS to probe the structural dynamics initiated by the laser excitation.
- **Scientific Applications**
  - Time resolved X-ray scattering
  - X-ray interactions with matter
- **Scattering Geometry**
  - Wide angle x-ray scattering
  - Diffraction

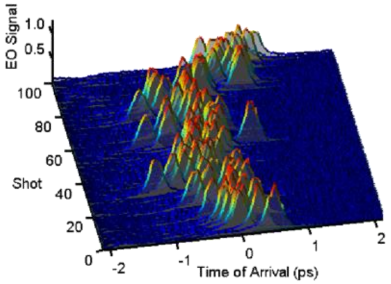
Traditional Pump Probing



- Time delay achieved by optical path length delay or RF phase shift
- Time resolution limited by LCLS/laser jitter
- ~ 100 fs resolution

Non-sequential Sampling





- Diagnostic measures LCLS/laser relative timing on a pulse-by-pulse basis
- Intrinsic jitter is used to sample various time delays
- Femtosecond temporal resolution

CXI (2011)

Computational intensive!
Large amount of data needed
By far the most computational demanding both in real-time or offline

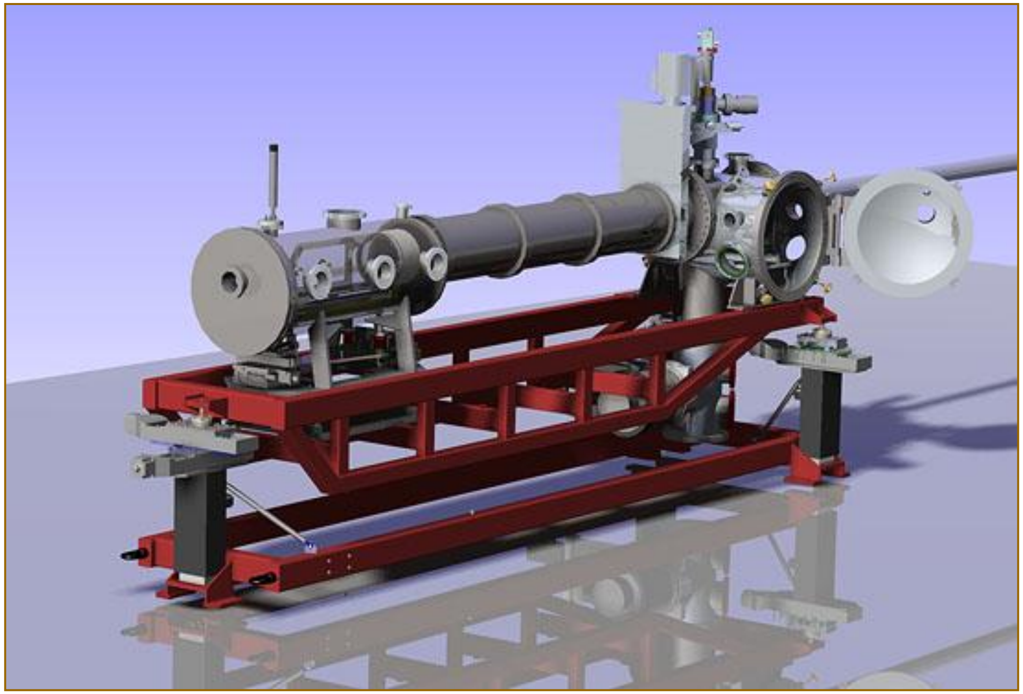
- The **Coherent X-Ray Imaging (CXI) instrument** will image single sub-micron particles. The full transverse coherence of the LCLS laser will allow single particles to be imaged at high resolution while the short pulse duration will limit radiation damage during the measurement.

- The XCI instrument will allow imaging of biological samples which will be introduced either fixed on targets or using a particle injector that will deliver free-standing particles to the beam.

- Two high quality focusing optics will generate two foci (1000 and 100 nm) and this will allow imaging of single nanoparticles of various sizes, pushing the limit down to single biomolecules.

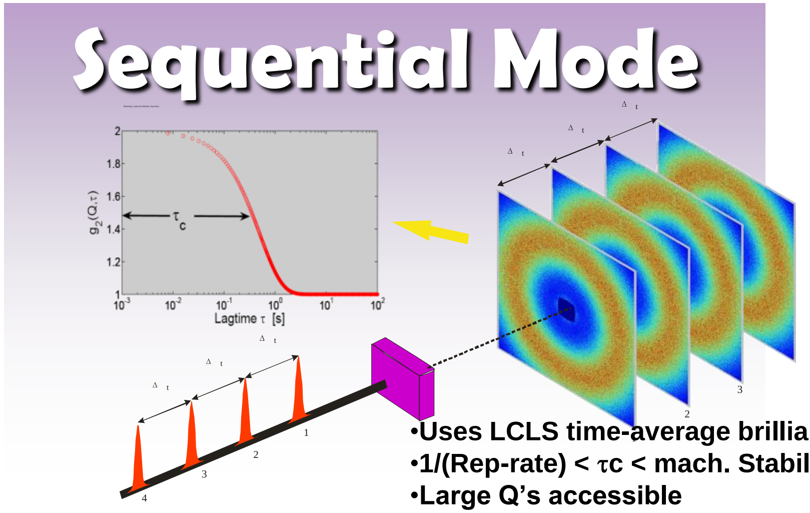
Scientific Program

- Imaging of reproducible biomolecules
- Nanocyrstallography of proteins
- Imaging of nanoparticles
- Imaging of hydrated living cells
- X-ray matter interactions
- Pump-probe imaging

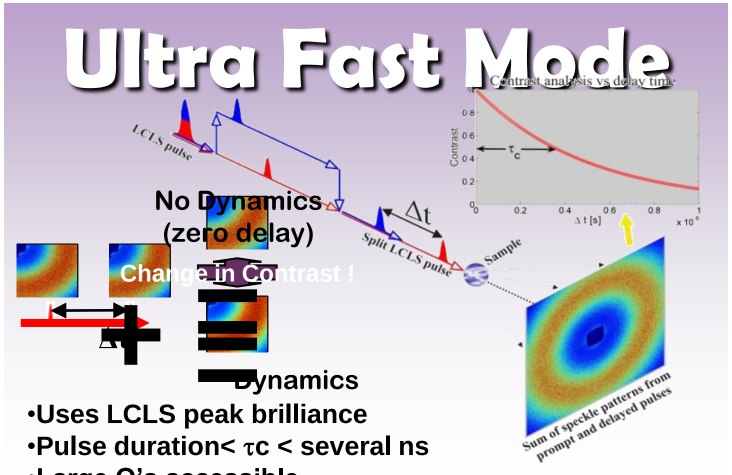


XCS (2011)

- The **X-ray Correlation Spectroscopy (XCS)** instrument will observe dynamical changes of large groups of atoms in condensed matter systems over a wide range of time scales.



- Uses LCLS time-average brilliance
- 1/(Rep-rate) < τc < mach. Stability
- Large Q's accessible



- Uses LCLS peak brilliance
- Pulse duration< τc < several ns
- Large Q's accessible

MEC (2012)

- The Matter in Extreme Conditions (MEC) endstation will be dedicated to studies of matter in extreme conditions with a focus on high energy density science.

- The MEC instrument will observe matter at temperatures exceeding 10,000 Kelvin and at pressures 10 million times the earth's atmospheric pressure at sea-level, enabling unprecedented understanding of exotic states of matter.

Data Rates & Volumes

Vary for different instruments/experiments

- **Three operation modes:**
  - **30 Hz** (first runs on LCLS, AMO/CAMP, Oct-Dec 2009)
  - **60 HZ** (second series of runs, May 2010 -> on)
  - **120 Hz** (Fall 2010?)
- **Instruments:**
  - **Fall 2009**: AMO/CAMP
  - **Summer 2010**: AMO/CAMP, XPP, SXR
  - **2011 (or 2012):** all 6 instruments
- **First runs at AMO/CAMP (30 Hz):**
  - up to **180 MB/s**, **3 TB/day**, **~100 TB** of raw data recorded
- **At full capacity (120 Hz):**
  - **Up to 1 GB/s** for some instruments
  - **10 TB/day** across entire system
  - **1 PB** raw data per year

Offline System “Philosophy“

- **Dealing with much less controlled users community:**
  - New experimental group comes each week
  - Experiments are very short (5 shifts x 12 hours = 60 hours)
  - Very little time (~1 month) to get scientists integrated into the environment
  - Need to deal with very diverse approaches to the data analysis
  - A variety of existing tools (MatLab, IDL, custom frameworks, etc.)
  - Etc.
- **=> Very different from HEP!**

- **Hence, the data (rather than work-flow) centric approach:**
  - Users deal directly with files stored on a huge POSIX file system
  - Communicate on data sources & formats rather than interfaces
  - Very light frameworks (under development)
    - Looking at C++, Python, probably MatLab

Offline System Functions

- **Translate raw data from XTC into HDF5 representation**
  - Also extract metadata and store data attributes in a database (science metadata)
- **Short/Medium term storage for data on a disk**
- **Long term archival on tapes**
- **Provide access for scientists to the data:**
  - Access by attributes (metadata)
  - Manage data ownership and access restrictions
  - Data access from:
    - Processing clusters
    - Offsite export
    - User workstations (investigating options)
- **Provide data analysis infrastructure (under development)**
- **Perform centralized data processing (under discussion)**

Data Types & Sources

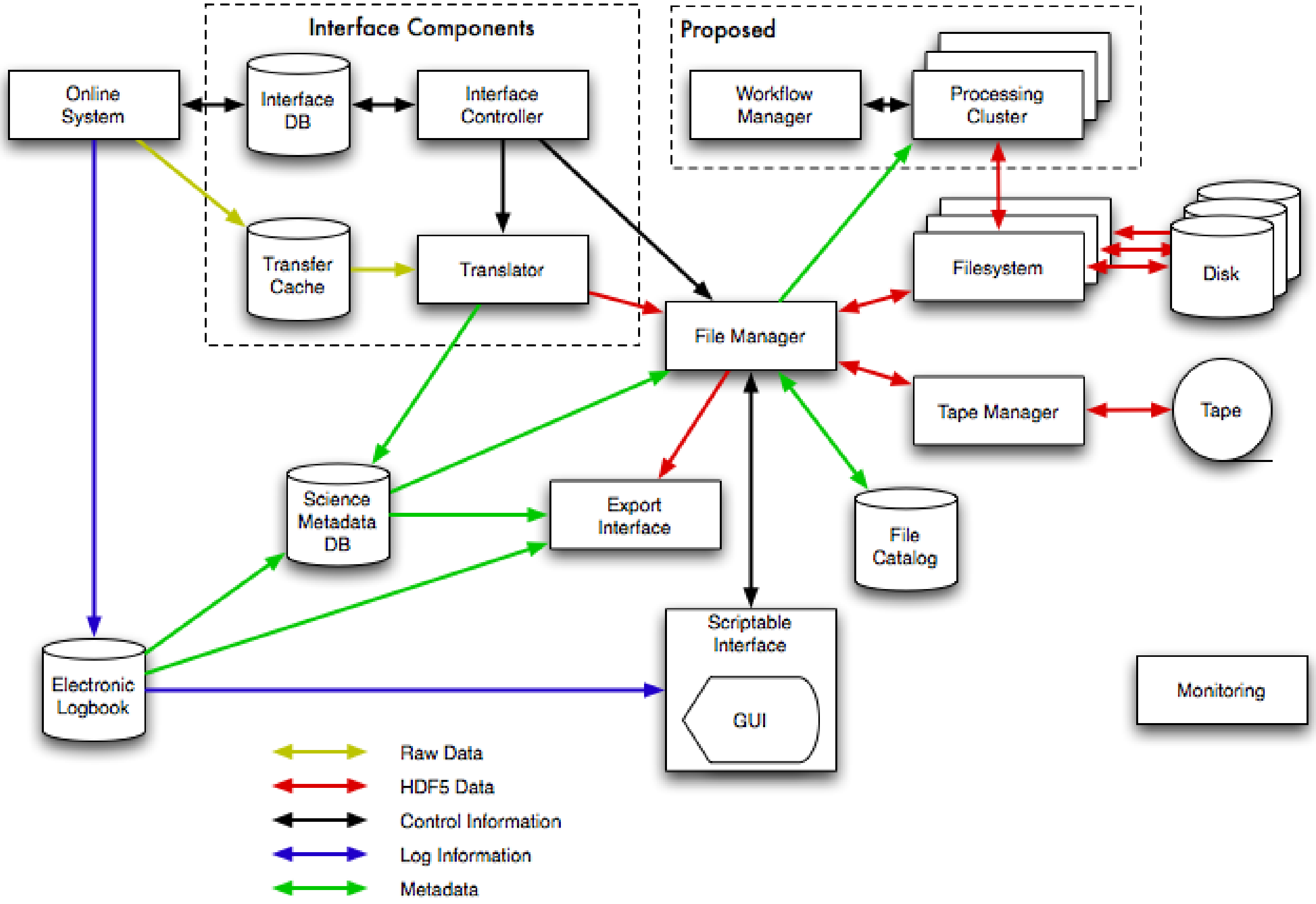
- **DAQ readout**
  - Raw data from detectors (CCD sensors, many waveforms for time of flight measurements, etc.)
- **EPICS**
  - A subset of the monitoring and environment information needed to analyze the raw data in OFFLINE
  - Also a source of metadata
- **Electronic Log Book**
  - Experiments annotation by operators and experts
  - https://pswww.slac.stanford.edu/apps/logbook/

- **Raw data formats:**
  - XTC files produced by ONLINE
  - HDF5 files produced by OFFLINE during the XTC-to-HDF5 translation

Offline System Architecture

Src: Engineering Specification Document

https://confluence.slac.stanford.edu/download/attachments/31293814/ESD-1.6-118-20080723.pdf?version=1



Disk Storage (Offline)

- **Technology choice:**
  - LUSTRE (Cluster File System by Sun)
  - Data Direct Network (DDN) **S2A9900**:
    - http://www.datadirectnet.com/9900
- **Capacity:**
  - **1 PB** May 2010
  - **2 PB** Fall 2010
  - **3+ PB** (probably next year)
- **I/O bandwidth:**
  - *Aggregate*: **5 GB/s** per 1 PB storage for read & write
  - *Single stream*: up to **1.2 GB/s** (limited by 10 Gbps Ethernet)
  - Various options exist in Lustre (striping, etc.)
- **Designed to scale in three dimensions:**
  - *Capacity*
  - *I/O bandwidth*
  - *Parallelism* (a number of streams/users)

Data Archival

Raw files are archived automatically by iRODS as they show up on the OFFLINE disk storage

- **HPSS**
  - 1 TB tapes
  - 6+ tape recorders
  - 10 Gbps network connection
- **2 copies of XTC files and 1 copy of HDF5 files at SLAC**
  - A possibility to ship 1 copy off site is being discussed
- **Data archival/retention policy (yet to be finalized):**
  - Indefinitely on tape
  - 1 year on disk
- **Data safety measures:**
  - Archive raw data files immediately as they're arriving to OFFLINE
  - Calculate MD5 sum on raw data files in ONLINE (under development)
  - Store the sums in the Data Management System (under development)
  - Read a few % of files back and compare MD5 (under development)

Data Exportation

- **No centralized data export!!!**
- **Users transfer data via a "bastion" host:**
  - Visible from ESNet (10 Gbps)
  - Connected to LUSTRE (10 Gbps)
- **Data transfer protocols:**
  - bbcp
  - sftp
  - Anything else
- **Web apps to locate files:**
  - A simplified interface to the File Manager
  - https://psdev.slac.stanford.edu/apps/explorer/
- **Client-side command line scripts (under development)**
  - To automate data location (experiments, runs, time intervals, metadata, etc.) and data transfer
  - To be integrated with Web services to access file catalogs and metadata

Data Processing Clusters

- **Three types of problems:**
  - CPU intensive:
    - High density blade clusters
    - 100+ nodes (< 1 k CPU cores)
    - 1 Gbps network per host
  - I/O intensive:
    - 10+ nodes (< 100 CPU cores)
    - 10 Gbps network connection for each node (dual CPU)
  - Data parallel:
    - GPU clusters (under investigation)
    - Many 10k GPU cores
- **Batch job scheduler**
  - LSF

Databases

- **An implementation of the system relies on a number of (MySQL) databases:**
  - Experiments Registry Database (also used in ONLINE)
  - Authorization Database
  - Electronic LogBook Database (shared with ONLINE)
  - Science Metadata Database
  - Interface Controller Database (to support XTC-to-HDF5 translation)
  - File Manager Catalog Database (iRODS)
- **Some database are also seen via Web apps and Web services**

Web Applications & Services

- **Technologies:** Apache, PHP5, Python, Web services, WebKDC (single signup user authentication), LDAP, MySQL, AJAX (JSON, a lot of Java Script!)

- **Major applications:**
  - Experiments Registry: https://psdev.slac.stanford.edu/apps/regdb/
  - Authorization Database: https://psdev.slac.stanford.edu/apps/authdb/
  - File Catalog: https://psdev.slac.stanford.edu/apps/explorer/
  - Electronic LogBook: https://psdev.slac.stanford.edu/apps/logbook/
  - Science Metadata: Under development

- **Services:**
  - Experiments Registry
  - Authorization
  - File Catalog

Software Release System

- **Simplifies and automates task of building software**
- **Built on top of popular SCons package written in Python (http://www.scons.org/)**
  - **SCons** is configured and extended with Python scripts
- **Release is a collection of packages, each package is a separate directory**
- **Package has internal structure (include/, src/, app/)**
  - Location of the file inside package determines what is done to the file
  - Package structure would make it easier to switch to different system (e.g. make) if needed
  - Each package has its own **SConscript** file which is identical for almost all packages

Software Release System

- **Automated discovery of dependencies**
  - Package-level dependencies are deduced from file dependencies gathered by SCons
  - Package dependency is used to build the list of libraries needed by particular executable or shared library
  - Package authors do not need to say explicitly which libraries are needed (but can specify additional libraries in some cases)

- **External software is represented by proxy packages**
  - Package with a special SConscript file in it (empty otherwise)
  - Usually creates symlinks to external libraries/includes but can do potentially anything

- **Build process is managed by BuildBot (http://buildbot.net)**
  - Python-based continuous build system
  - Both nightly and regular builds

Data Analysis Frameworks

- **The area is a bit challenging due to:**
  - Diverse user community (experience, background)
  - A variety of tools, languages, libraries and requirements
  - A little time to integrate users into our computing environment

- **Options are still under investigation**
- **"Light" framework approach**
- **Focusing at C++ and Python based implementations**
- **Need (actually have to) leverage a use of parallelism at various levels:**
  - Multi-core (same host)
  - Many-core (MPI?)
  - Emerging heterogeneous architectures (GPU, DSP, System-On-Chip)
- **A possible integration with the Data Management System**

Open, On-going Projects

- **Data Analysis**
  - Analysis frameworks
  - Software Releases Support for scientists
- **Data Management**
  - Data export interfaces & tools
  - System & Resources Monitoring (Nagios, HPSS, etc.)
- **Centralized data correction (processing)**
  - Apply calibrations (dark images, etc.) to raw data to get "usable" data in a first approximation
- **Feasibility studies for using GPUs to accelerate scientific applications**
  - Trying to get LDRD money
- **Fast data monitoring (ONLINE/OFFLINE)**
  - Main requirement: low latency
- **Web Applications & Services**
  - A number of issues & interesting ideas exists