

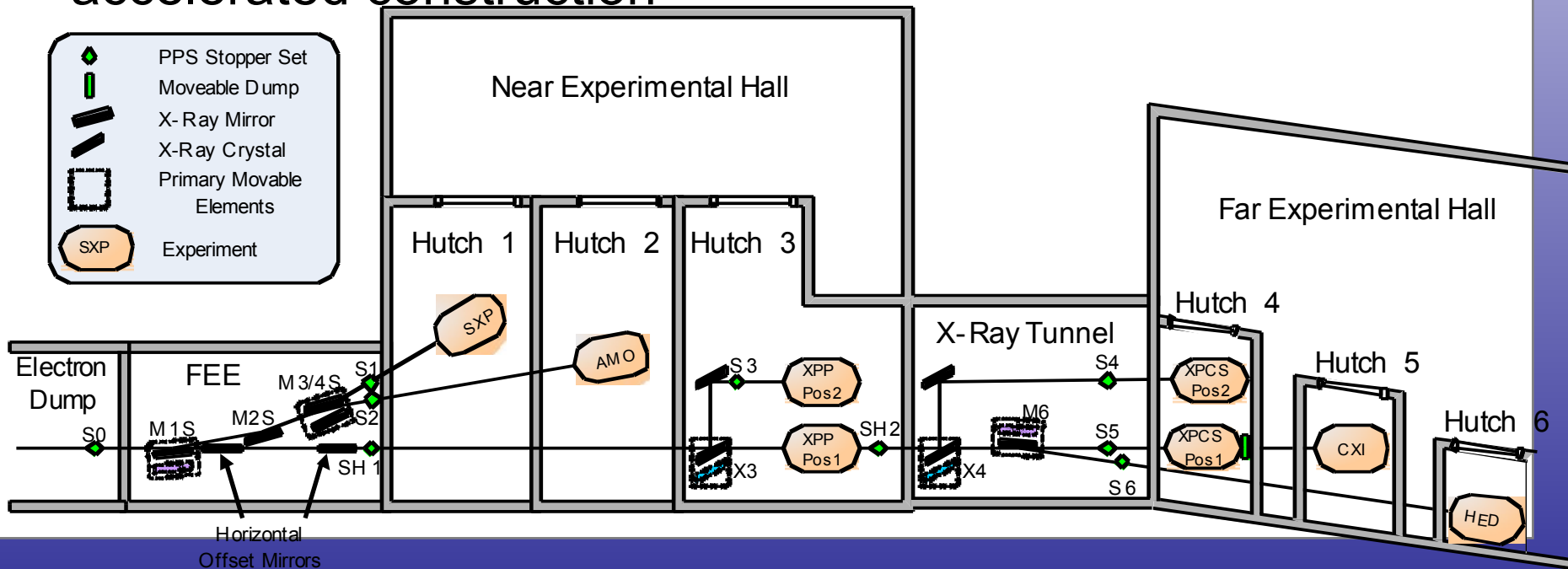
LCLS Online and Offline Computing

Alf Wachsmann
SLAC National Accelerator Laboratory

alfw@slac.stanford.edu



- Linac Coherent Light Source (LCLS) is being built right now
- It had its first LASER light on April 10!
- LCLS Ultrafast Science instrument (LUSI) will build instruments (detectors) for LCLS
- Funding for all 6 hutches has been granted which allows accelerated construction



■ Near-Experimental Hall (NEH):

- Atomic, Molecular, and Optical Science (AMOS)

- LUSI:

 - X-ray Pump Probe (XPP)

 - Soft X-ray Research (SXR)

■ Far-Experimental Hall (FEH):

- LUSI:

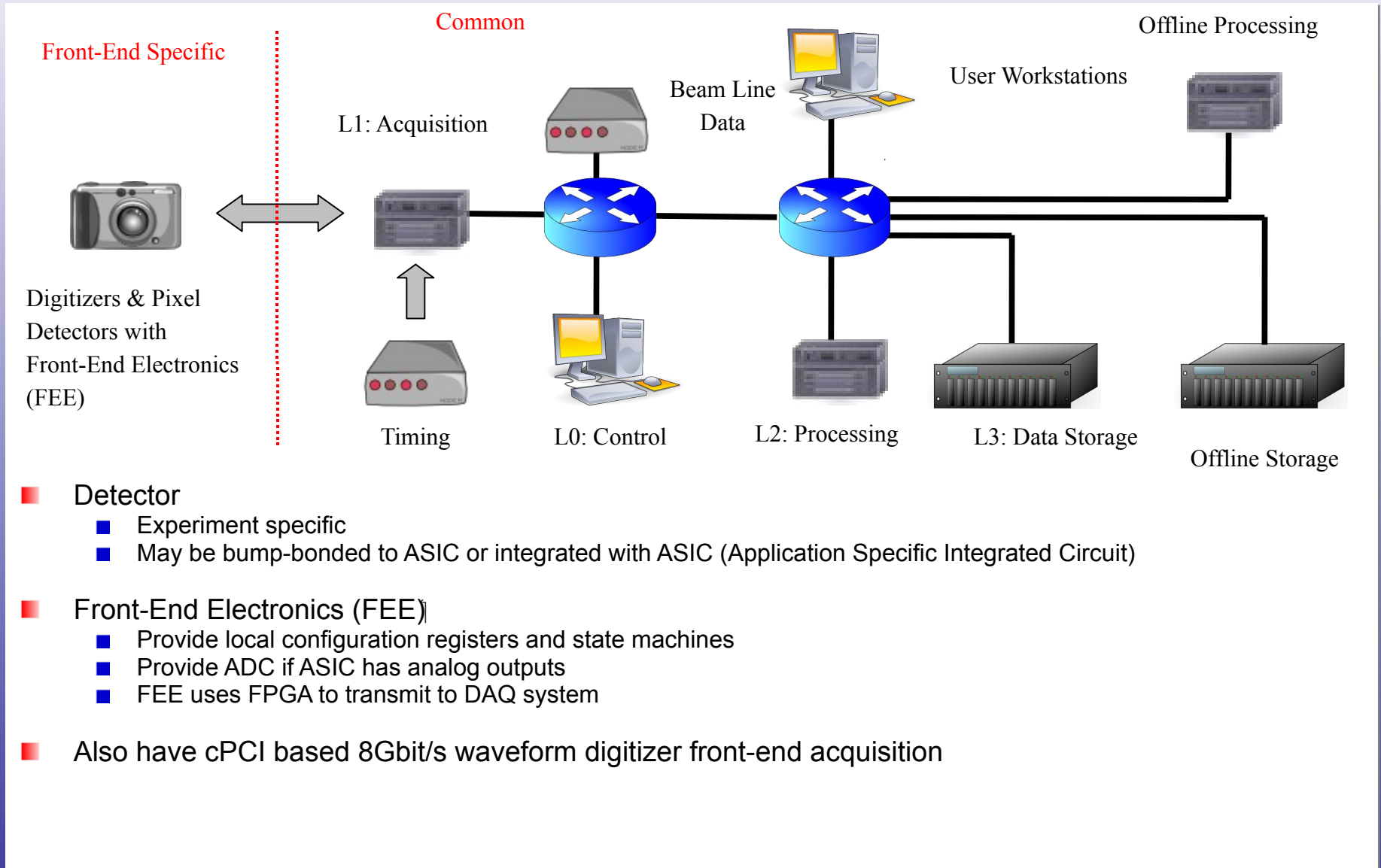
 - Coherent X-ray Imaging (CXI)

 - X-ray Correlation Spectroscopy (XCS)

 - High-Energy Density Science (HEDS)

■ All experiments are fully funded since January 2008

- Specifications (some of them only need to be met by 2015)
 - 120 Hz per-pulse data collection
 - < 100 fsec second timing
 - Multi-Gigabit/sec peak rate
 - Multi-Terabyte daily data volume (2011)
 - Hundreds of Terabyte yearly accumulation (2012)
 - Real-time analysis
 - Online and offline data rendering
 - Offline computation (CXI, tbd) (2011-12)



- **Detector**
 - Experiment specific
 - May be bump-bonded to ASIC or integrated with ASIC (Application Specific Integrated Circuit)
- **Front-End Electronics (FEE)**
 - Provide local configuration registers and state machines
 - Provide ADC if ASIC has analog outputs
 - FEE uses FPGA to transmit to DAQ system
- Also have cPCI based 8Gbit/s waveform digitizer front-end acquisition

- Level 0: Control
 - DAQ operator consoles
- Provide different functionalities:
 - Run control
 - Partition management, data-flow
 - Detector control
 - Configuration (modes, biases, thresholds, etc)
 - Run monitoring
 - Data quality
 - Telemetry monitoring
 - Temperatures, currents, voltages, etc
- Manage all L1, L2 and L3 nodes in a given partition (i.e. the set of DAQ nodes used by a specific experiment or test-stand)

- Level 1: Acquisition, first level processing
 - Receive 120 Hz timing signals, send trigger to FEE (Front-End Electronics), acquire FEE data
 - Error detection and recovery of the FEE data
 - Control FEE parameters
 - Calibration & Correction
 - Event-build FEE science data with beam-line data
 - Beam-Line Data is 120 Hz real-time data received from accelerator, femto-second laser timing system, etc.
 - Image processing
 - Partial data reduction
 - Rejection using 120 Hz beam-line data
 - Processing both in software and firmware (VHDL)
 - Send collected data to Level 2 nodes over 10 Gb/s Ethernet

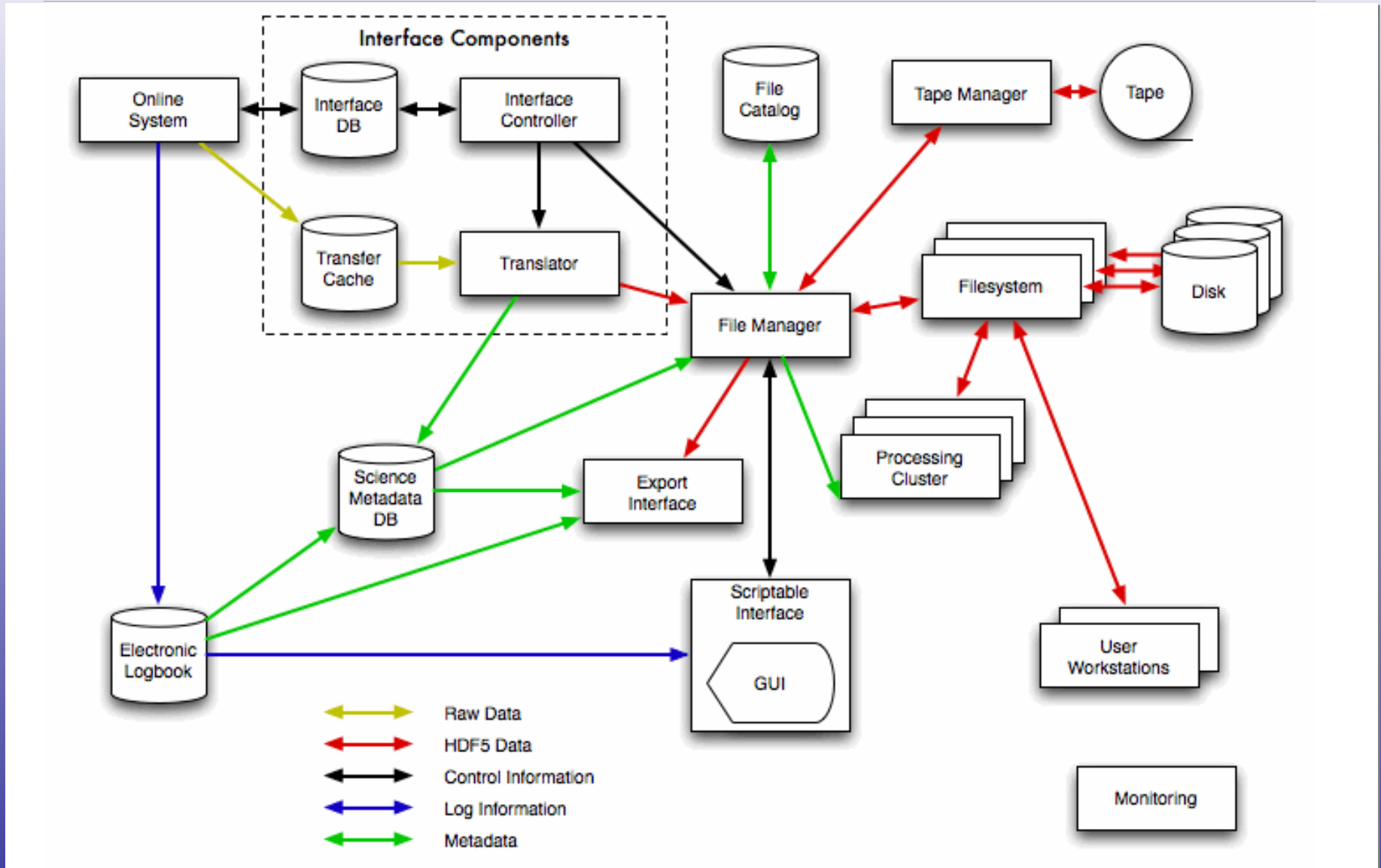
- Level 2: Processing
 - Vetoing, sorting, classification of images
 - Curvature correction
 - Histogramming
 - Feature extractions
 - Curve fitting
 - Lossless compression
 - Correlations
 - Statistical parameters (single shot and averaged)
 - High level data processing
 - Run monitoring
 - Data quality
 - Data analysis and visualization
 - Real-time feedback to experimenter
 - Error detection and recovery

■ Real-Time Display

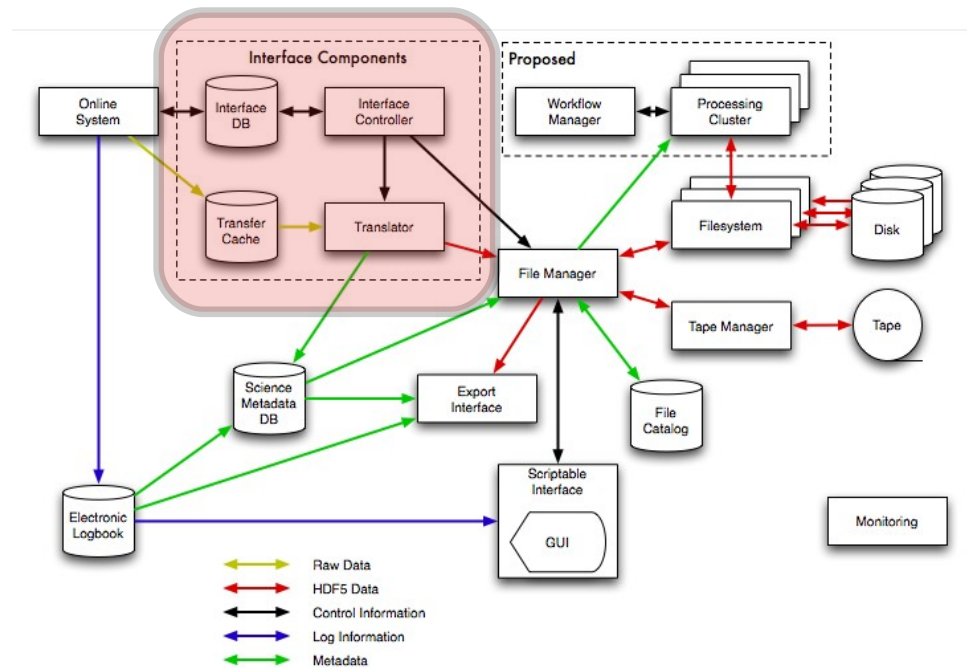
- Raw Data
- Meta Data
- Processed Data
- Radial Average
- Autocorrelation
- Peak Information
- Estimated Data Completeness

- Level 3: Short/Medium Term Data Storage
 - Provide data storage
 - Located in server room in experimental hall
 - Off-line system will transfer data from local storage to tape staging system
 - Tape staging system located in SLAC central computing facilities

- Translate data into a form for long-term scientific access
 - Translate to HDF5 format
 - Store data attributes in a database (science metadata)
- Store translated data and archive to tape
- Provide access for scientists to stored and archived data
 - Access by attributes (science metadata)
 - Manage data ownership and access restrictions
 - Data access for:
 - Processing clusters
 - User workstations
 - Offsite export (network and other means)
- Manage derived data products



- Interface Database
- Transfer Cache
- Translator
- Controller



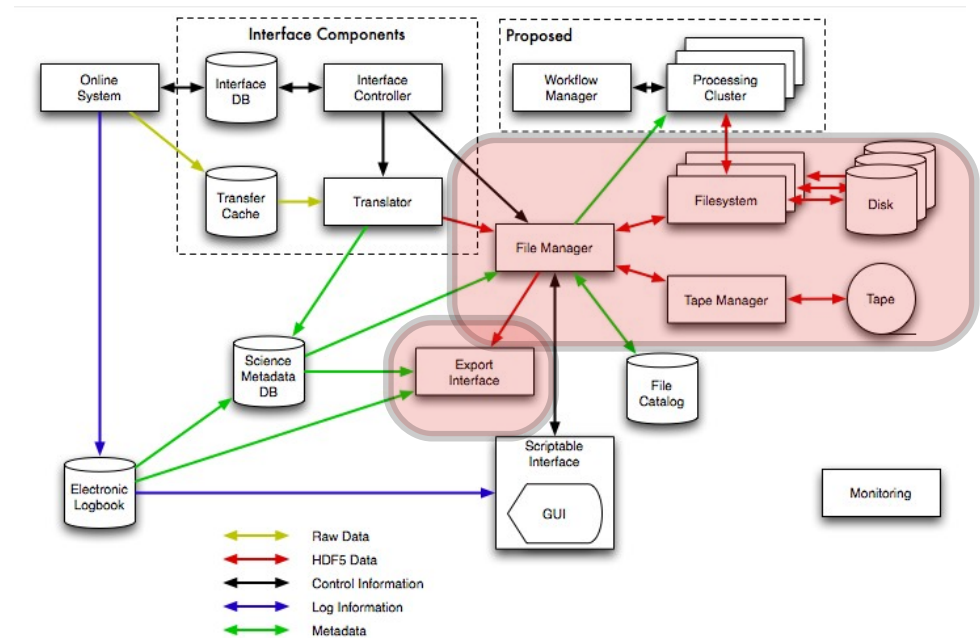
- Sole link for control communications between online and offline system
- Tables:
 - Status of each online file
 - Destinations in transfer cache
- Implemented using master/slave MySQL replicas

- Dedicated storage area for files being transferred and translated
- Avoids conflicts with scientific workloads
- Bandwidth:
 - 200 MB/sec write
 - 100 MB/sec read
- Implemented as cluster of servers with local disk (Lustre)

- Converts from online data format to HDF5 format
- Merges raw data and EPICS data streams
- Converted data passed to File Manager
- Attributes extracted from data and stored in science metadata database
- Implemented in custom C++ code.
Likely an iRODS microservice

- Monitors transfer cache and writes destination entries into database
- Monitors file transfers and triggers translator
- Removes files from transfer cache
- Replaces master database with slave if needed
- Controller restarted in case of failure
- Implemented in Python

- File Manager
- Filesystem
- Disk Hardware
- Tape Manager
- Tape Hardware
- Export Interface



- Routes files to appropriate destinations
- Maintains physical location information
- Interfaces with filesystem, tape manager, and export interface
- Implemented with iRODS

- Presents standard Unix filesystem interface
- May be multiple filesystems to facilitate resource management
- Implemented with Lustre

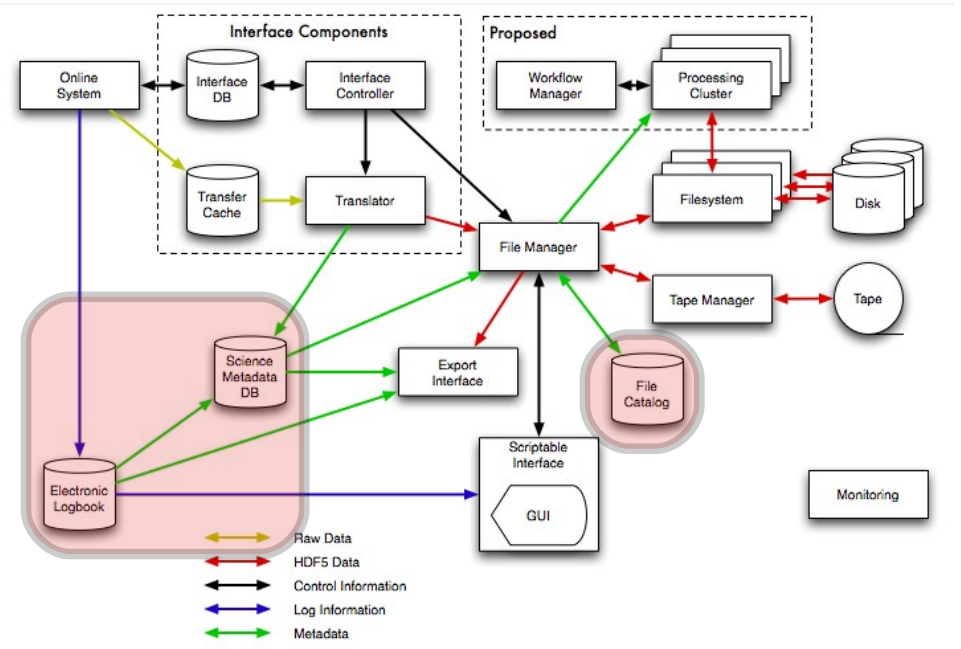
- High-bandwidth: 200 MB/sec read and write bandwidth
- Scalable to petabytes, with terabytes per file
- Average file size: 100s of MB to GBs

- Writes files to tape, recording locations
- Implemented using HPSS under iRODS

- Similar to existing tape robots

- HDF5 files plus metadata from science metadata database and electronic logbook
- Network transport:
 - Implemented using GridFTP, scp, bbcp
- Disk transport:
 - Implemented using e-SATA or USB 2.0

- File Catalog
- Science Metadata Database
- Electronic Logbook



- Records locations of files
- Maintains access control lists
- Implemented using iRODS internal metadata

- Stores user, run, pulse, and other scientifically interesting attributes
- From online system and electronic logbook
- Implemented using MySQL

■ Run attributes:

- Experiment name (e.g. AMOS, XPP, XCS, CXI)
- Project identifier
- Run identifier
- Run type
- Date/time range
- Experiment configuration
- Sample information

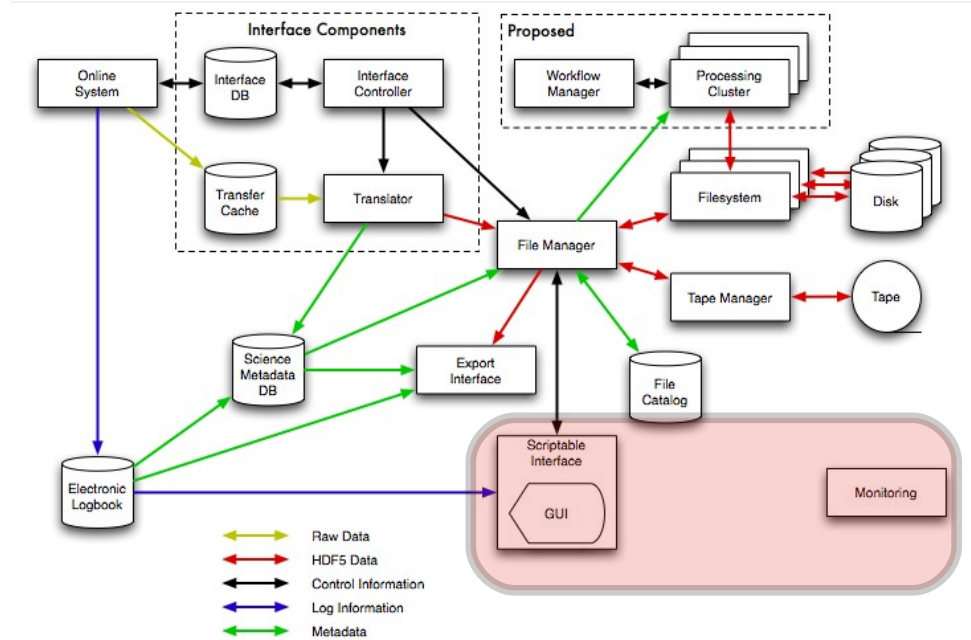
■ Photon beam attributes:

- Intensity
- Pulse length
- Repetition rate

- Free-form entry mode:
 - Keep log of experiment
 - Text entry, screen shots, attachments
- Data acquisition-driven entry mode:
 - Record per run
 - Pre-filled with defaults, completed by operator
- Master lives in online system
- Implemented using MySQL

- Web-based GUI for operator read/write access
- Replication to offline science metadata database
- Exports in PDF format

- Scriptable Interface
- GUI
- Monitoring

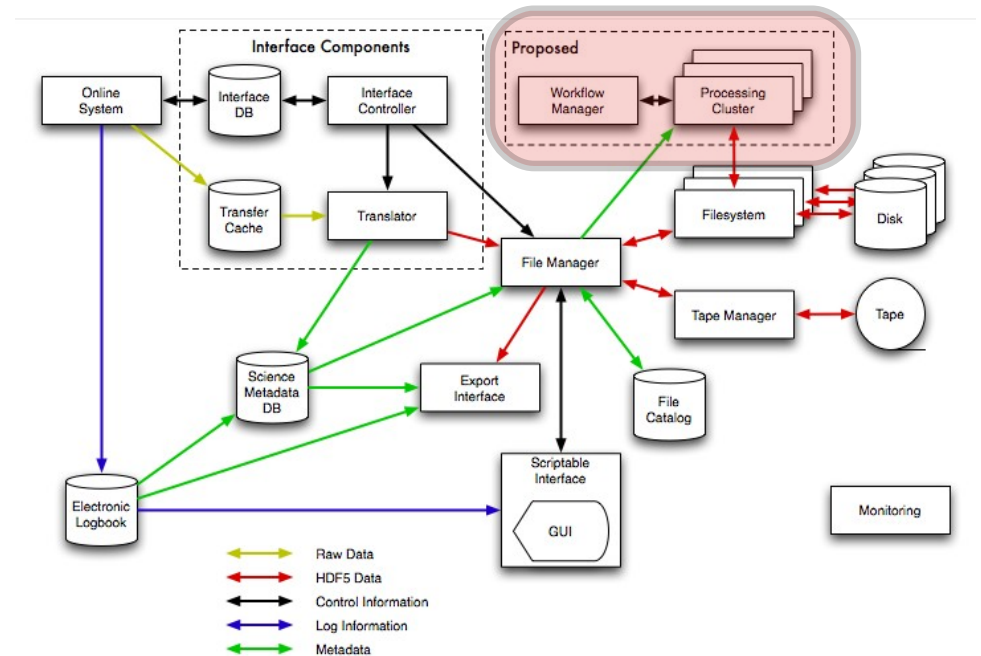


- Command-line interface
- Query catalog and science metadata
- Retrieve files or portions thereof
- Based on iRODS `imeta` and `iget`

- Web-based interface
- Same capabilities as scriptable interface
- Perform administrative tasks
- Based on PRODS PHP API for iRODS

- Monitored through central Web-based console
- Each machine and service produces status metrics
- Implemented using Ganglia and Nagios

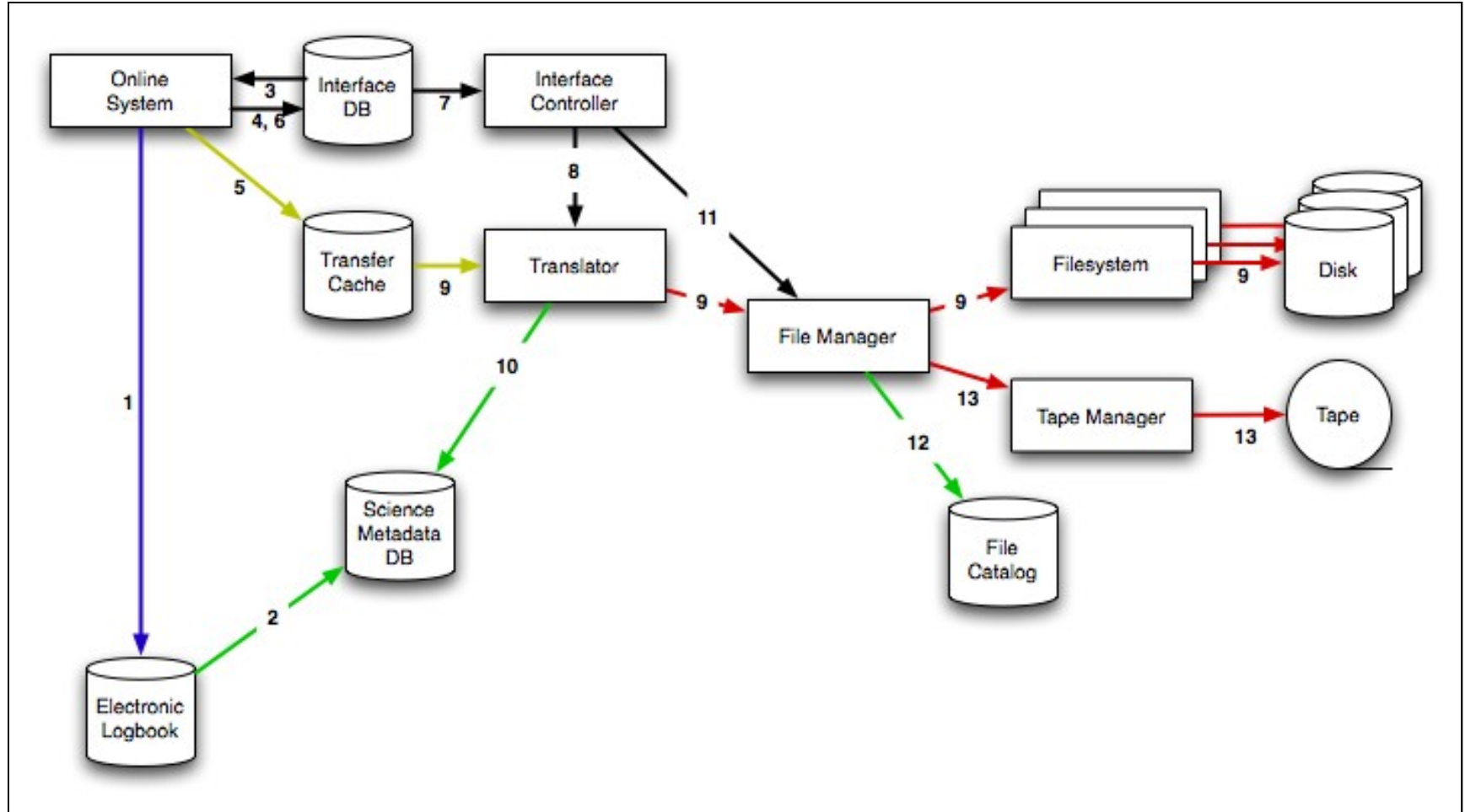
- Processing Cluster
- Workflow Manager

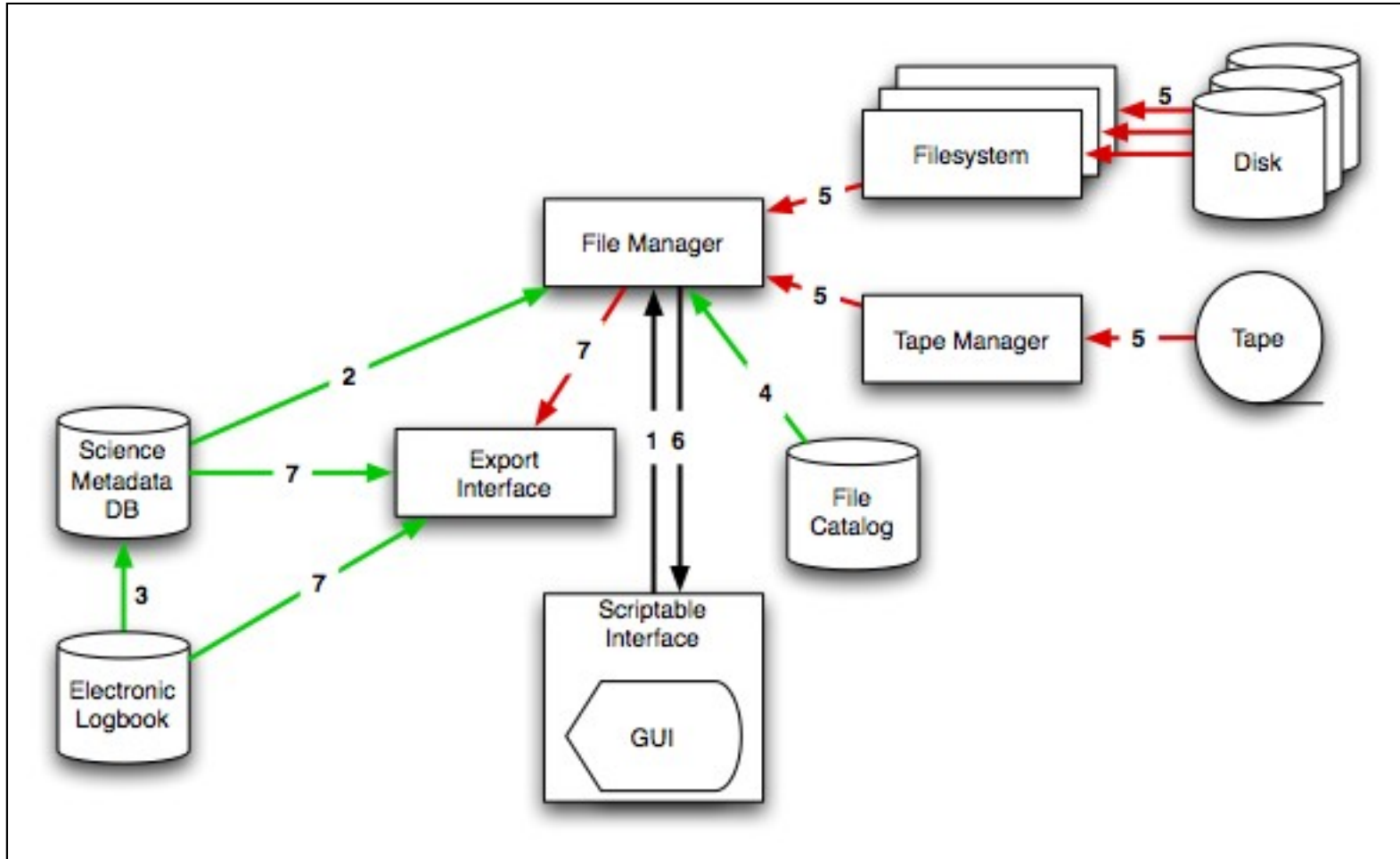


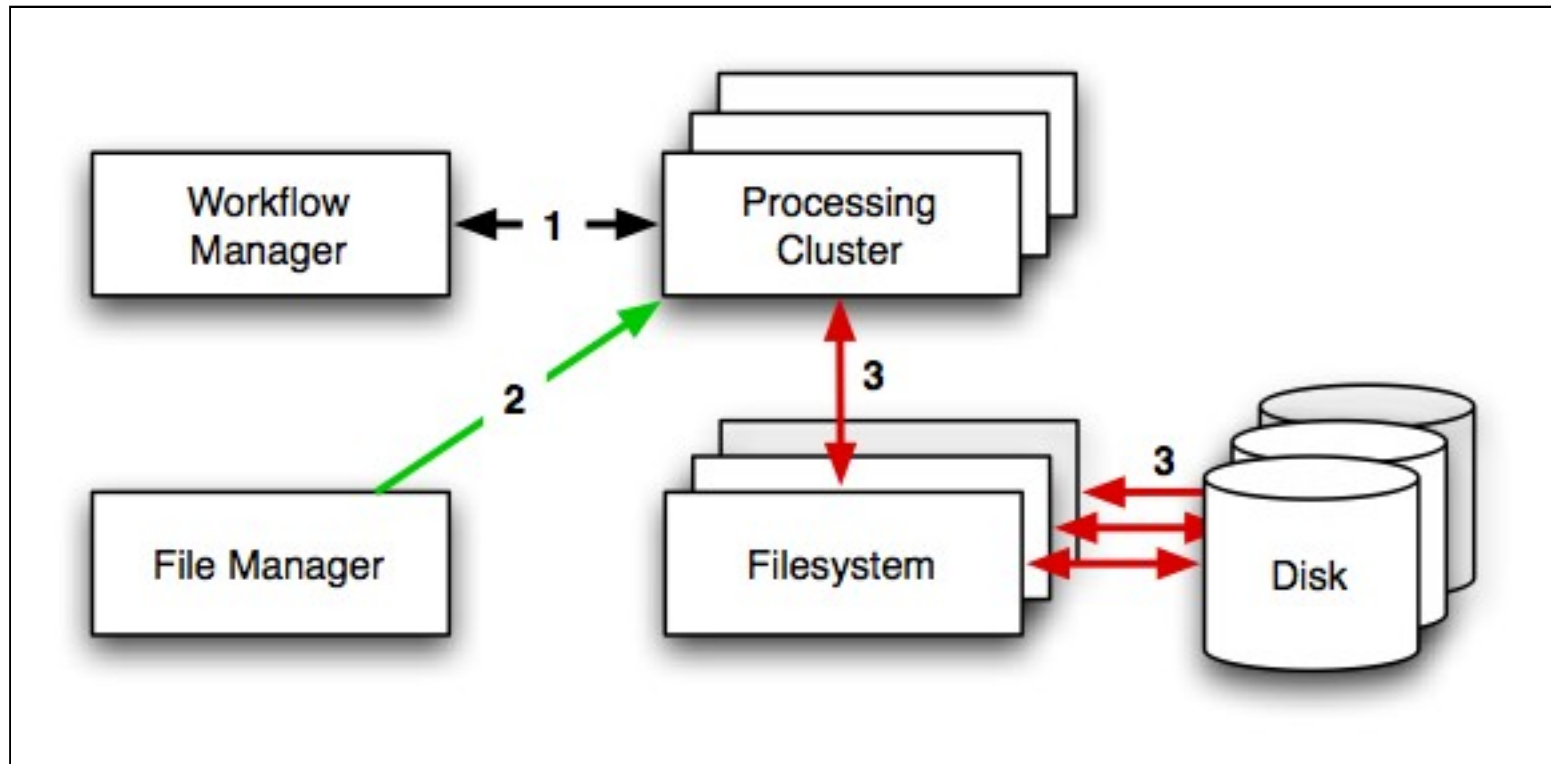
- Tightly integrated with filesystem machines
- Input locations obtained from file manager
- Output will go into separate “user dataset” filesystem

- Controls sequence and resource usage of processing jobs
- Implemented using LSF

- Security
- Data Ownership
- Identity Provisioning and Management
- Access Control Implementation







- Instrument Definition
- Experimental Data
- Beam Quality Data
- EPICS Provenance Data
- Science Metadata

■ Status

- Design completed and reviewed
- In the process of being implemented; Prototype mostly done
- Data challenge in June (one month later than planned)

■ Simulations

- Generate diffractive images
- Beam-sample interaction under different experimental conditions
- Apply detector and electronics simulation
- Instrumental will be the ability to superimpose realistic noise levels from machine and electronics
- Need to interface with detector scientists

■ Apply processing algorithm from online/offline chain

- Need to interface with online and offline groups

- Processing resources required for CXI
 - First need to determine algorithm details
 - Then estimate offline resources required
 - Add e.g blades, another option to run sections of algorithms (e.g. FFT's) on SLAC ATCA RCE Modules (> 400 DSP equivalent on each module)
 - Issue only for CXI experiment in Far Experimental Hall
 - Offline processing for experiments in other hutches not a driver and is being implemented
 - CXI analysis facility at SLAC?

- Most slides were taken from talks by Gunther Haller and K.-T. Lim