Data Acquisition

Amedeo Perazzo
Photon Controls and Data Systems (PCDS) Group

SLAC, June 9th 2009
FAC Review
**Data System Architecture**

**Level 0: Control**
- Run & configuration control
- Run & telemetry monitoring

**Level 1: Acquisition**
- Image acquisition, calibration
- Event-building with beam-line data
- Correction using calibration constants
- Data reduction (vetoing, compression)

**Level 2: Processing**
- Event building, pattern recognition, sort, classify, alignment, reconstruction

**Level 3: Online Archiving**
- NEH/FEH local data-cache
- Offline will transport data to tape staging area in SCCS Computer Center
Network Connections: AMO

CISCO 6509

2x1Gb

SCCS

CISCO 6509

2x1Gb

NEH Server Room

CISCO 6509

2x1Gb

Hutch 1

Fndry FWSX

2x1Gb

L0

Service Network

ATCA CIM

nx10Gb

nx10Gb

L2

2x10Gb

L3

2x10Gb

L1

nx10Gb

Fndry FWSX

nx1Gb

L0

DAQ Network
Network Connections: XPP

- SCCS
  - Cisco 6509
  - 2x1Gb

- NEH Server Room
  - Cisco 6509
  - 2x1Gb

- Hutch 3
  - Fndry FWSX

- Service Network
  - Cisco 6509
  - Cisco 6509
  - Cisco 6509
  - nx10Gb

- DAQ Network
  - ATCA CIM
  - L1
  - L2
  - L3
  - L0
  - nx10Gb
CGS Catalyst 6509

- 720 Gbps switch fabric backplane
- 9 slots (currently: 2 supervisors, 2 x 48 1Gb RJ45, 2 x 24 SFPs)

Foundry FWSX448

- 2x10Gb/s XFP, 4xSFPs, 48xRJ45
Network Devices (II)

- Cluster Interconnect Module (CIM) SLAC custom made ATCA switch
  - Based on two 24-port 10Gb Ethernet switch ASICs from Fulcrum
    - Up to 480 Gb/s total bandwidth
  - Fully managed layer-2, cut-through switch

- Fully configurable
  - Designed to optimize crates populated with RCE boards (see next slides)
  - At the same time may be configured to connect standard ATCA blades
    - L2 DAQ nodes
Reconfigurable Cluster Element (I)

- **SLAC custom made ATCA board**
  - Used as readout node (L1) for custom cameras

- **Based on System On Chip (SOC) Technology**
  - Currently implemented with Xilinx Virtex 4 devices, FX family
    - Targeting XC4VFX60
  - Xilinx devices provide
    - Reconfigurable FPGA fabric
    - DSPs (200 for XC4VFX60)
    - Generic CPU (2 PowerPCs 405 running at 450 MHz for XC4VFX60)
    - TEMAC: Xilinx TriMode Ethernet Hard Cores
    - MGT: Xilinx Multi-Gigabit Transceivers 622Mb/s to 6.5Gb/s (16 for XC4VFX60)
Reconfigurable Cluster Element (II)

- **FPGA fabric**
  - Interfaces to:
    - memory subsystems
    - JTAG debug port
    - custom multi-function display
    - various I/O channels
  - Generic DMA Interface (PIC) designed as set of VHDL IP cores
    - Up to 16 PIC channels
  - PIC in conjunction with Multi-Gigabit Transceivers and protocol cores, provide many channels of generic, high speed, serial I/O
  - 10Gb Ethernet for communication with L2 nodes
  - PGP for FEE communication
  - PIC in conjunction with TriMode Ethernet Hard Cores also provide commodity network interfaces
    - 1Gb Ethernet
Reconfigurable Cluster Element (III)

- **System Memory Subsystem**
  - 512 MB of RAM (currently 128 MB)
    - Memory controller provides 8 GB/s overall throughput
    - Uses Micron RLDRAM II

- **Platform Flash Memory Subsystem**
  - Stores firmware code for FPGA fabric

- **Configuration Flash Memory Subsystem**
  - 128 MB configuration flash
  - Dedicated file system for storing software code and configuration parameters (up to 16 selectable images)

- **Storage Flash Memory Subsystem (optional)**
  - Up to 1TB per RCE persistent storage flash (currently 256GB per RCE)
    - Low latency/high bandwidth access through I/O channels using PGP
    - Uses Samsung K9NBG08 (32 Gb per chip)
RCE Board with RTM
Concurrent PP512

- Used as readout node (L1) for commercial cameras (eg OPAL1000) and digitizers (Acqiris DC282)
- Dual 2.5 GHz Core Duo processors
- Up to 8 GB DDR2 SDRAM
- Compact flash site
- Two PMC/XMC slots
- Three 1Gb/s Ethernet interfaces
- 6U PXI/CompactPCI standard, 64 bit, 66 MHz PCI bus
DAQ User Interfaces

- Run Control
- Monitoring
- Electronic Logbook
- Data Export

- Data export described in this presentation refers to online system only
  - In the beginning data export will provided by online
  - Later on by offline
    - Offline system expected end 2009
Run Control & Monitoring

- **Run Control**
  - Configuration
  - Calibration
  - Partition management
    - nodes, executables
  - DAQ Finite State Machine

- **Status:**
  - Configuration, partitioning, FSM done
  - To do: calibration framework

- **Monitoring**
  - Implemented by *observer* nodes (same as L2 nodes) which send processed data to user consoles

- **Status:**
  - Core monitoring ready
  - To do: provide QT and ROOT demos for user analysis (see slides end this talk)
Electronic Logbook

- **Free-form entry mode:**
  - Keep log of experiment
  - Text entry, screen shots, attachments

- **Data acquisition-driven entry mode:**
  - Record per run
    - Filled by DAQ system, completed by operator

- **Implemented against MySQL**

- **Interfaces**
  - Web-based GUI for operator read/write access
  - Science metadata DB

- **Status:**
  - Underlining database schema done
  - Python library/API to access MySQL done
  - PHP based web interface under development
  - Demo will be shown to LCLS scientists end this week
Data files include:

- Experimental Data
  - Each event stamped with time and pulse ID

- Instrument Definition: describes each part of the experimental instrument
  - e.g. positions, distances, settings, calibrations

- Beam Line Data: set of electron and photon beam parameters taken every pulse
  - e.g. electron beam energy and spread, photon beam pulse length and energy

- EPICS Data: selected subset EPICS data

- Science Metadata
  - e.g. project identification, operator identification
Data Export (II)

- Online data export provided as filesystem
- High bandwidth LUSTRE access from inside PCDS enclave
  - 100MB/s-1GB/s bandwidth
    - processing farm, user workstations
- General SAMBA access outside of PCDS enclave
  - 10-100MB/s bandwidth
    - desktop office machines, laptops
  - The node providing SAMBA is client of the LUSTRE server on one side and exports the data as a SAMBA server on the other
- Status:
  - LUSTRE server ready
    - 100TB currently available in NEH and 24TB in test-stand lab
  - To do: setup SAMBA
Storage

- **Short term storage**
  - Disk based
    - Storage cluster located in NEH server room
  - Provides ready access to science data
    - For analysis on processing farm
    - For transfer to home institution
    - 100 TB available now in LUSTRE filesystem
    - More will be added when needed
  - At least 1 year retention policy

- **Long term storage**
  - Tape based
    - Staging system located in SLAC computing facility
  - Essentially unlimited retention period
  - Allows offline system to restore the science data to short term storage if data needed after 1 year
Two requests emerged from the AMO/CAMP users workshop

- Ability for the users to plug-in their applications to both online and offline systems
  - Need to envision way to integrate user code in the online monitoring system
- Ability for the users to run their analysis code on their own machines without the need to export the data
  - Need to envision way for non-SLAC machines to mount short term storage filesystem

User plug-in to online

- Subset of online monitoring nodes allocated for user applications
- User code built against online core code to allow:
  - User monitoring nodes to receive multicast data from L1
  - Event build data (same as L2)
  - Callback user code with pointer to event
  - User code parse event using online data format

User mount short term storage

- Implemented as samba server which is also read-only client to short term storage
CAMP detector commissioning
- Detector people want to run their ROOT based analysis code with < 2s latency
  - Needed to observe on the fly detector response to the modification of different parameters: change knob, observe behavior

CAMP users
- Ability to perform arithmetic operation on $n$ consecutive images, eg display:
  - $I = (I_1 - I_3) - (I_2 - I_3)$
    - $I_1$: LCLS beam + laser + molecular beam
    - $I_2$: LCLS beam + molecular beam
    - $I_3$: LCLS beam
- Other formulas to be expected
  - Must be able to change operations on the fly
Conclusions

- **Basic server room functionality ready**
  - Main service switch installed and configured
  - Connections to SLAC domain, accelerator domain, laser room and hutch 1 established
  - Accelerator timing distribution chassis installed and configured
  - AAA, NFS, DNS, NTP, LDAP servers setup using Linux high availability (HA) tools
  - Allow automatic fail-over in case master servers fail

- **Started building AMO production system in lab in bldg 84**
  - All L1 CPUs (cPCI blades), cameras (OPAL1000) and digitizers (Acqiris DC282) installed
  - L2 farm installed
    - Twelve 8-cores 10Gb/s ATCA blades
  - L3 data cache configured
    - 24 TB LUSTRE server
  - Ready to test the whole AMO dataflow chain