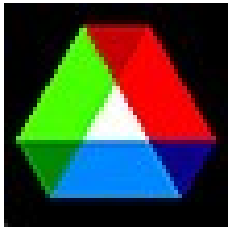
ATLAS Data Acquisition

Jinlong Zhang
Argonne National Laboratory
on behalf of
the ATLAS Collaboration

IEEE NPSS Real Time Conference 2009
Beijing, 10-15 May 2009



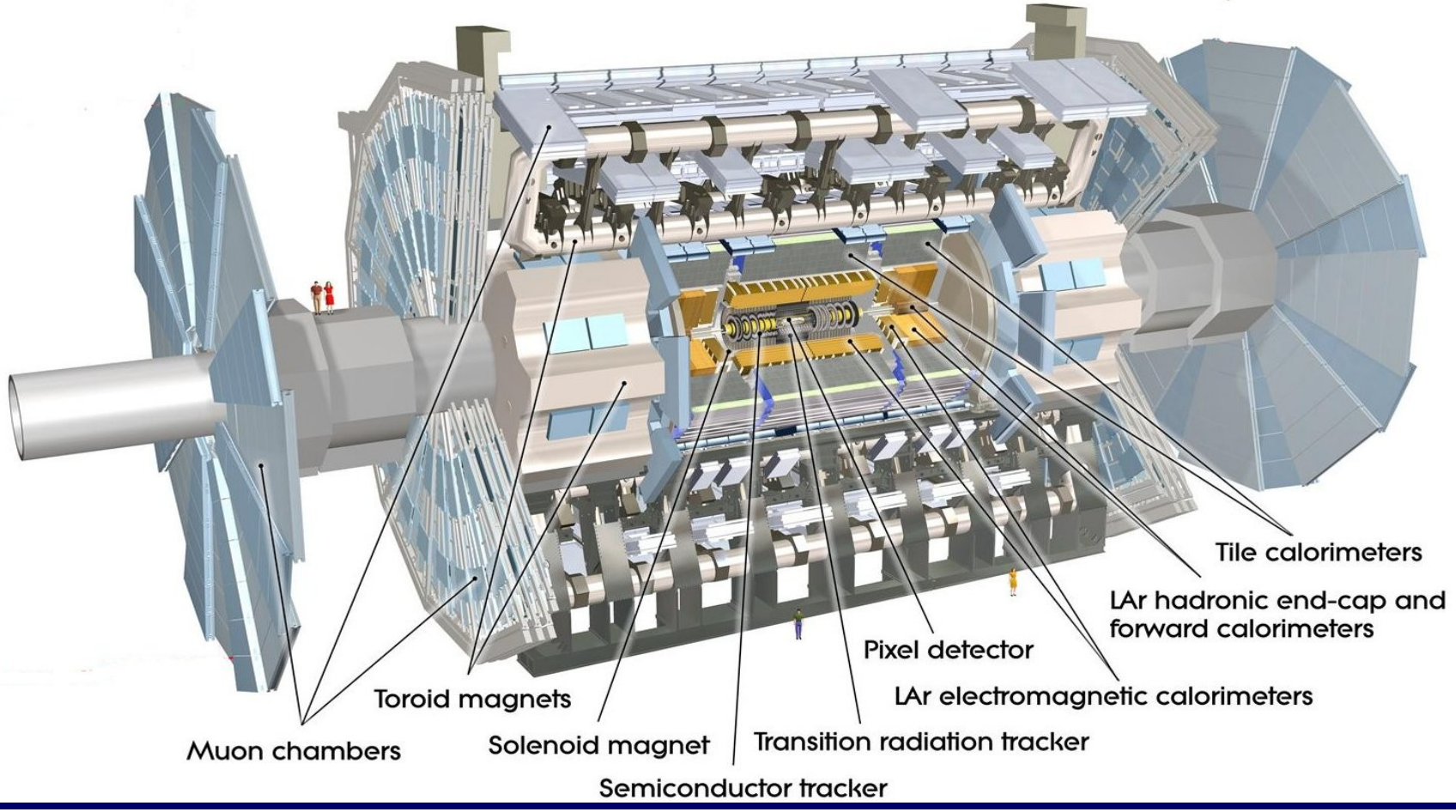
Outline

- Trigger/DAQ (TDAQ) system overview
  - Trigger status by C. Padilla and F. Winklmeier
- Dataflow system
- Online Configuration and Control
- Monitoring
- Integration, commissioning and early operation
- Summary

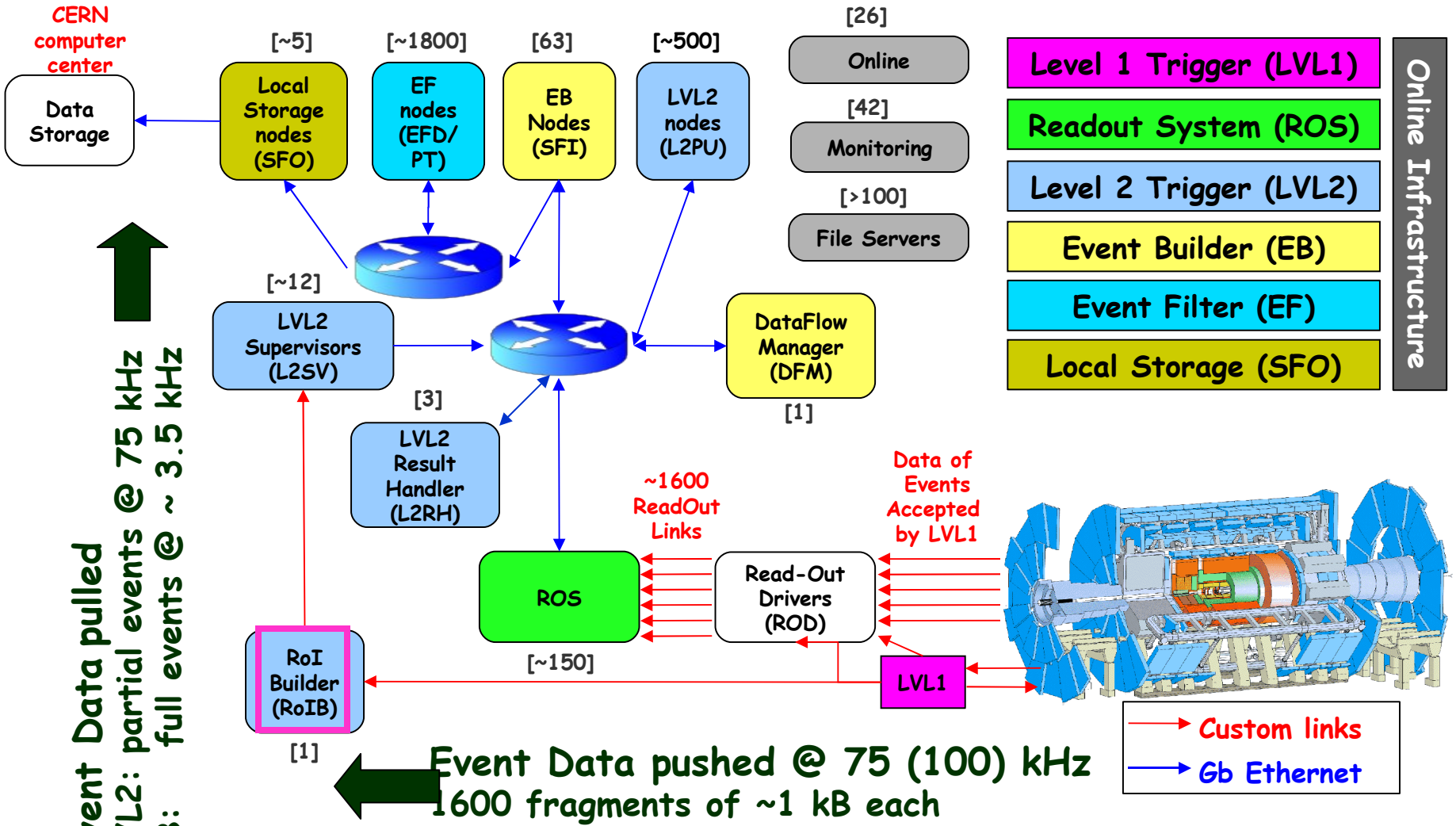
Trigger/DAQ (TDAQ) System Overview

ATLAS Detector

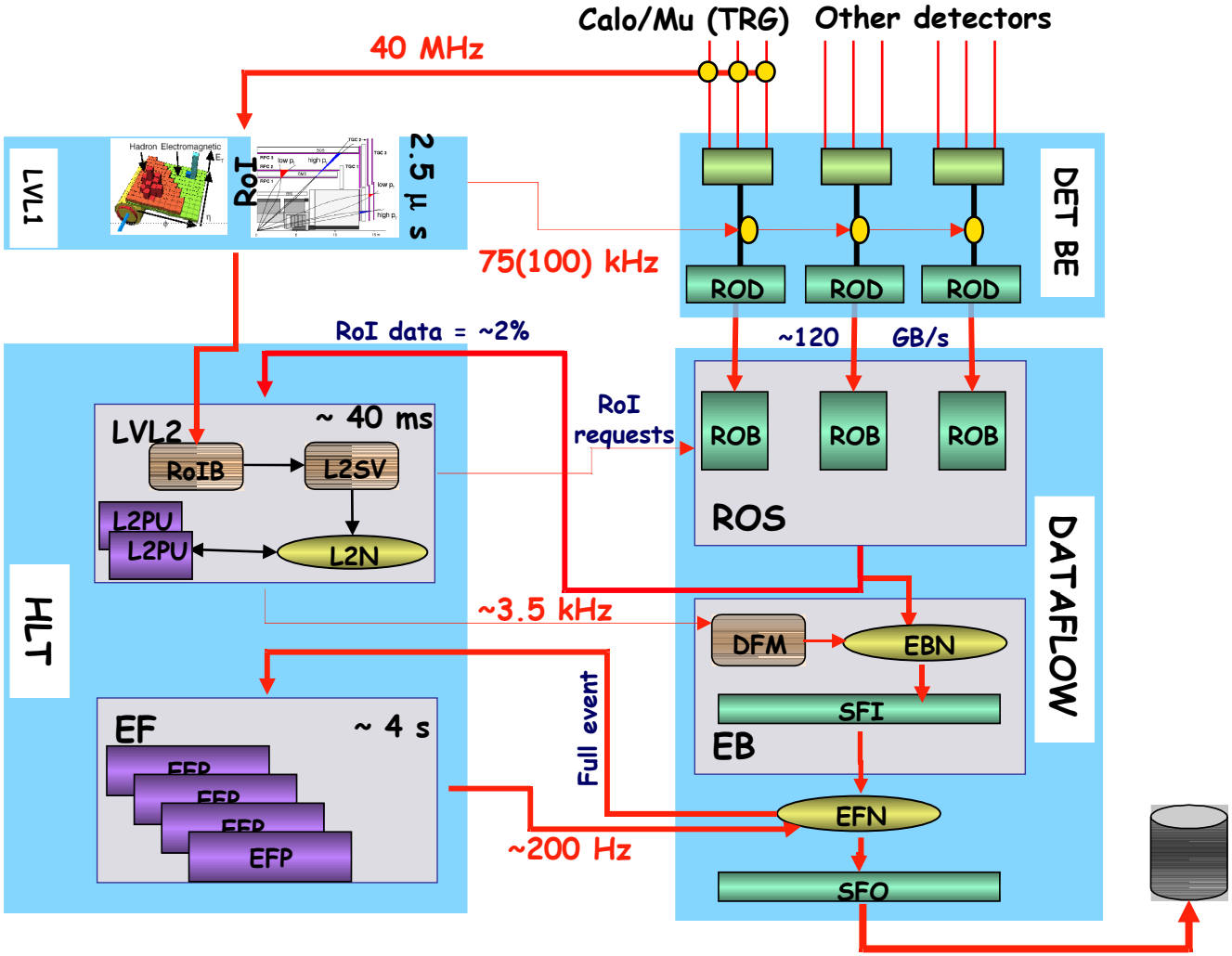
pp collisions @ 14 TeV and $L=10^{34}\ cm^{-2}s^{-1}$



ATLAS TDAQ System



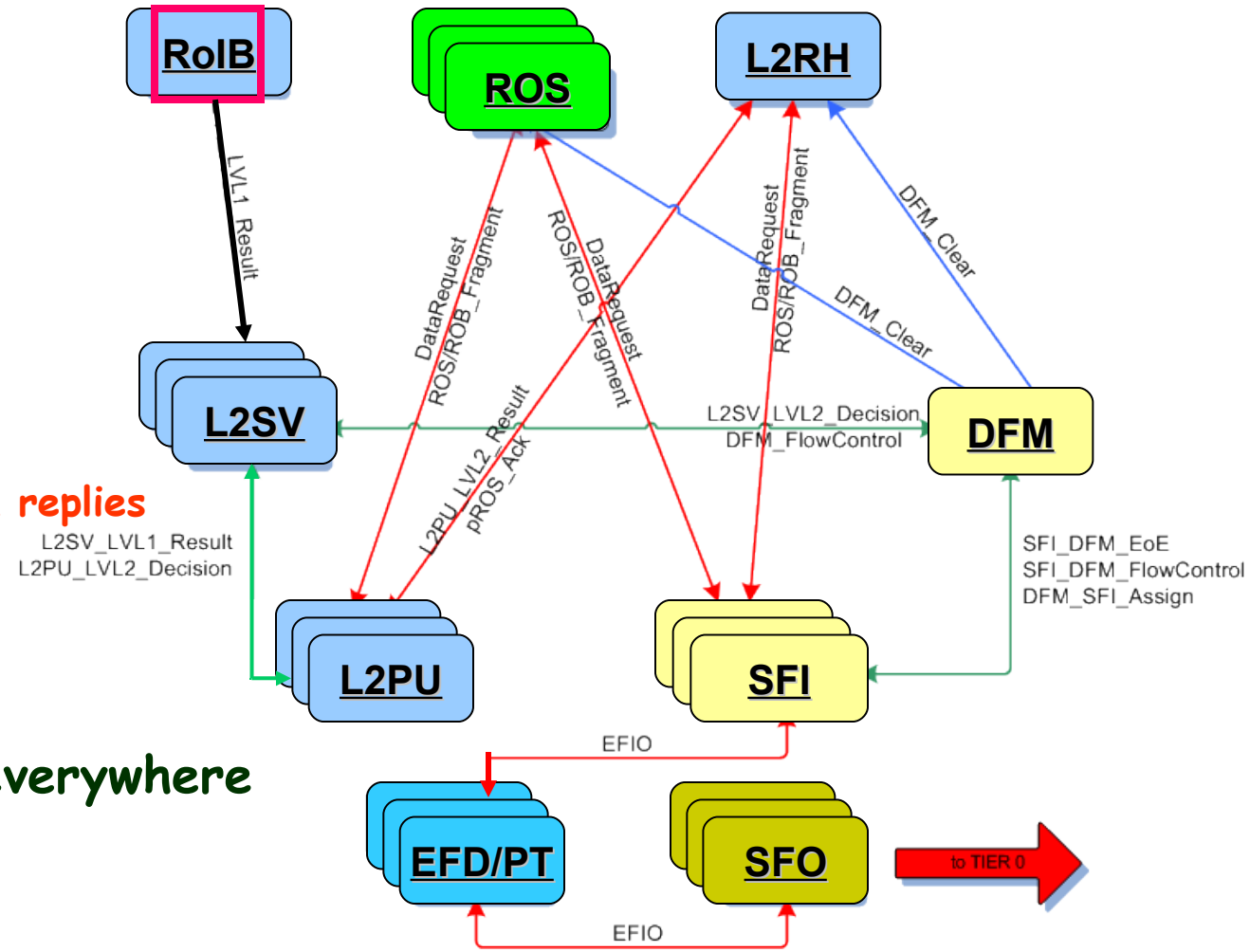
ATLAS TDAQ Strategy



Dataflow System

Dataflow Architecture

- UDP / IP
  - data requests and data replies
- UDP / IP multicast
  - DFM clear messages
- TCP / IP
  - data flow commands
- Possibly TCP / IP everywhere

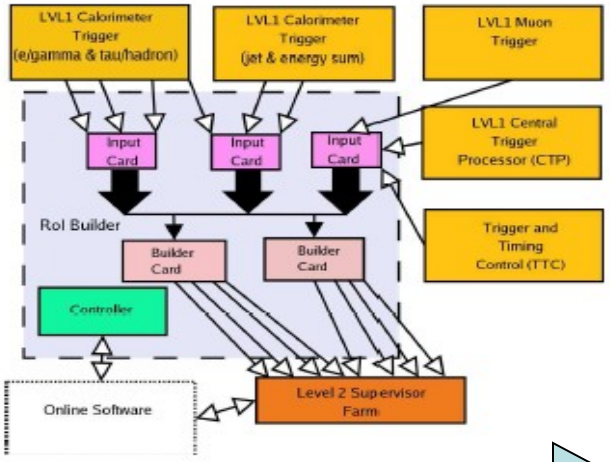


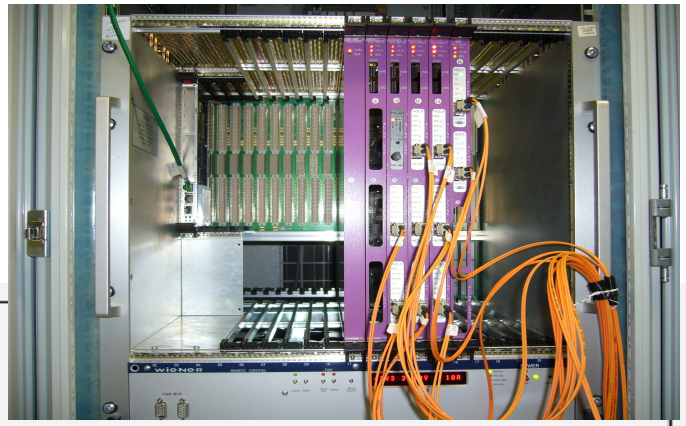
Hardware Installation Status

Component	Current	Final
RoIB	1	1
L2SV	2	12
L2PU node	See XPU	~500
DFM	12	12
SFI	63	63
SFO	5	5
EF node	See XPU	~1800
ROS	153	153
XPU	~850	~850

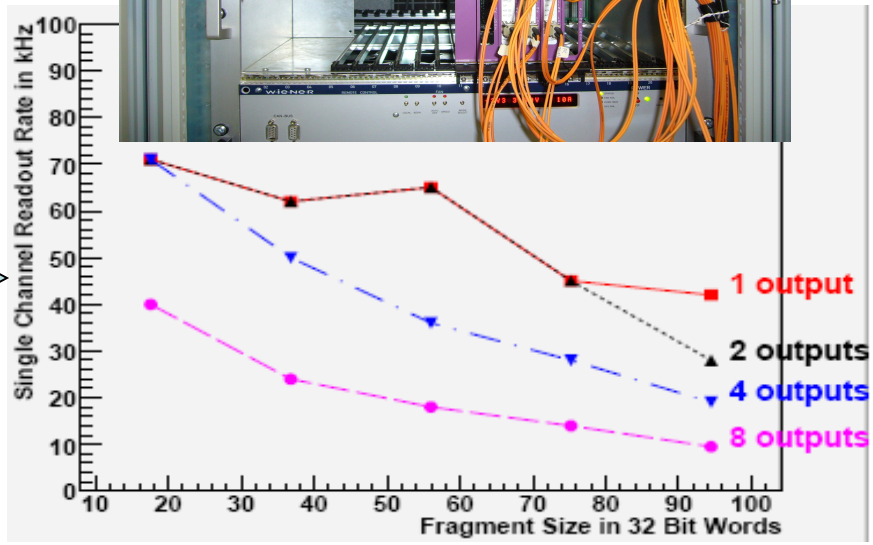
Region of Interest Builder (RoIB)

- Assembling Regions of Interest (RoI) information from LVL1 into complete data structure and delivering to L2SV
- A 9U VMEbus system with SBC, clock board, input boards and builder boards





- 8 inputs as final system
- 1 output with data checking



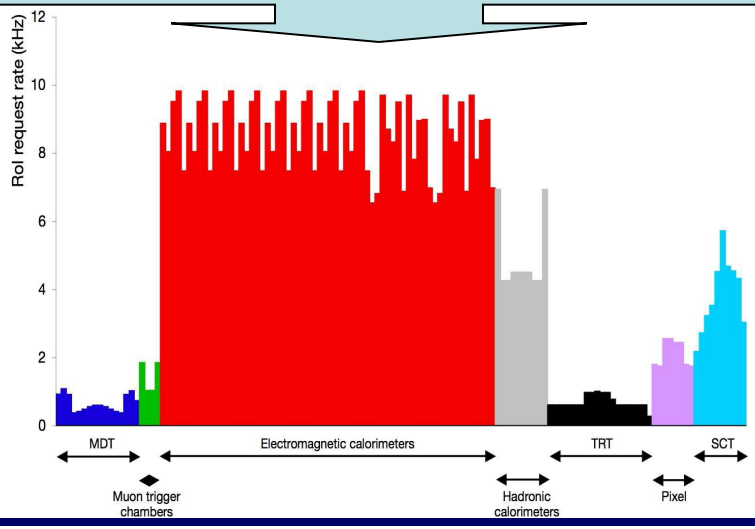
L2SV rate > 16 KHz

Readout System (ROS)

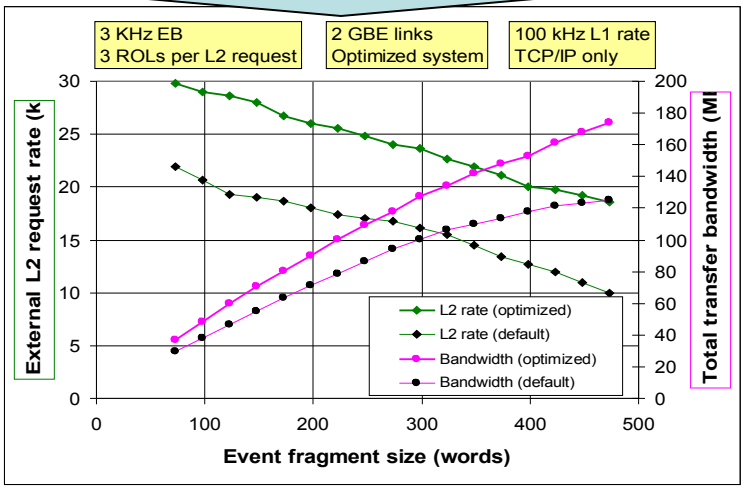
ROBIN

- CPU: Uni-proc 3.4 GHz Xeon
- RAM: 512 MB
- Network: 2 of 6 GbE for data collection
- 4 Readout Buffer Input (ROBIN) cards

RoI request rate per ROS at 10^{33} $cm^{-2}s^{-1}$



Current ROS performance with optimization



Level 2 (LVL2)

- L2SV distributing RoI (LVL1 results) to L2PUs
- L2PU running LVL2 trigger algorithms seeded by RoI
- L2RH providing LVL2 results to EB process

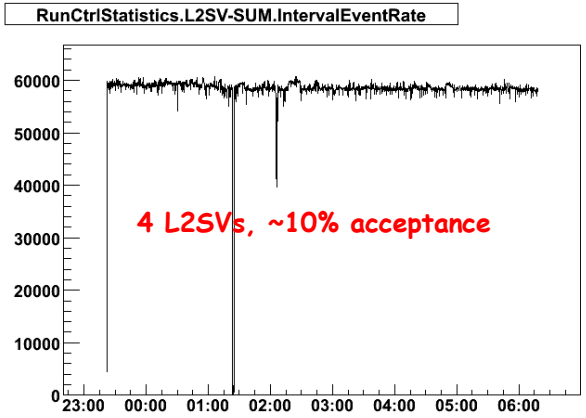
L2SV/L2RH

Filar

PCI interface with 4 custom link inputs

HLT Farms (L2PU, EF)

- CPU: AMD Opteron 252 2.6 GHz
- RAM: 2 GB
- Network: 1 of 2 GbE for data collection
- 1 Filar card in L2SV



- 31 nodes/rack
- CPU: 2 x Intel quad-core 2.50 GHz
- RAM: 2 GB / core
- Network: 1 of 2 GbE for data collection

Event Builder (EB)

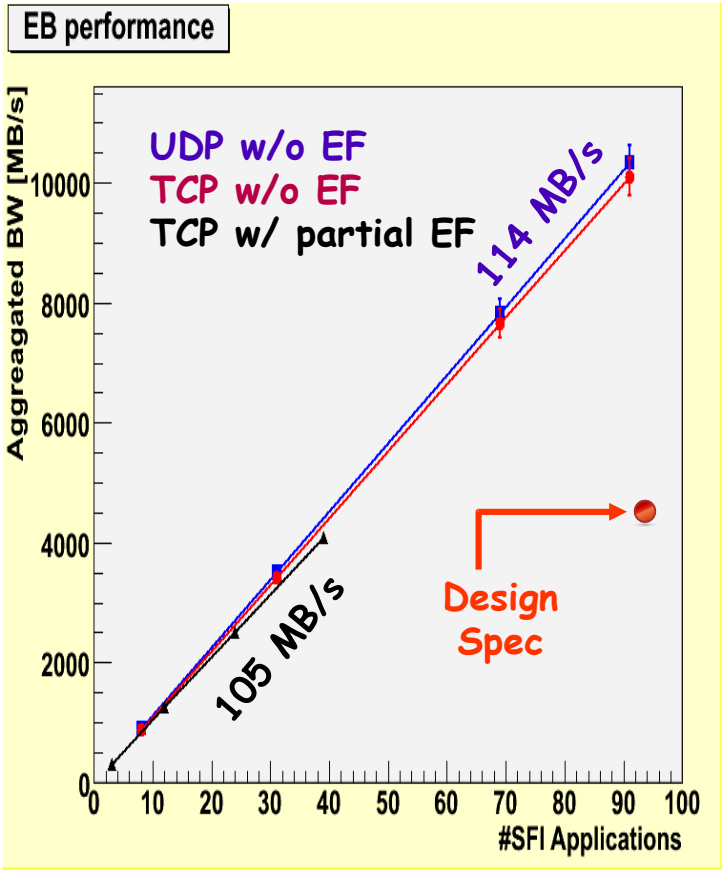
- DFM assigning events to SFIs and sending clear messages to ROSs
- SFI requesting data from ROSs and building the full event

DFM same hardware as L2SV

SFI

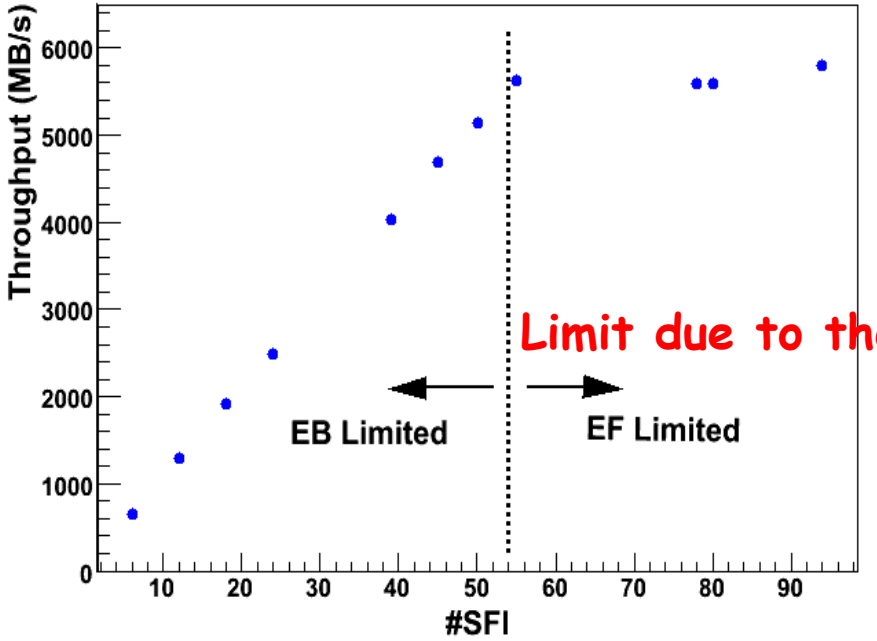
- CPU: AMD Opteron 252 2.6 GHz *single*
- RAM: 2 GB
- Network: 2 of 3 GbE for data collection
- CPU: 2 X Intel Xeon dual-core 3.0 GHz *double*
- RAM: 8 GB
- Network: 4 of 6 GbE for data collection

- 94 SFIs
- 1.6 MB event size



Event Filter (EF)

- Same hardware as L2PU
- Event filter algorithms accessing full event



Limit due to the installed bandwidth

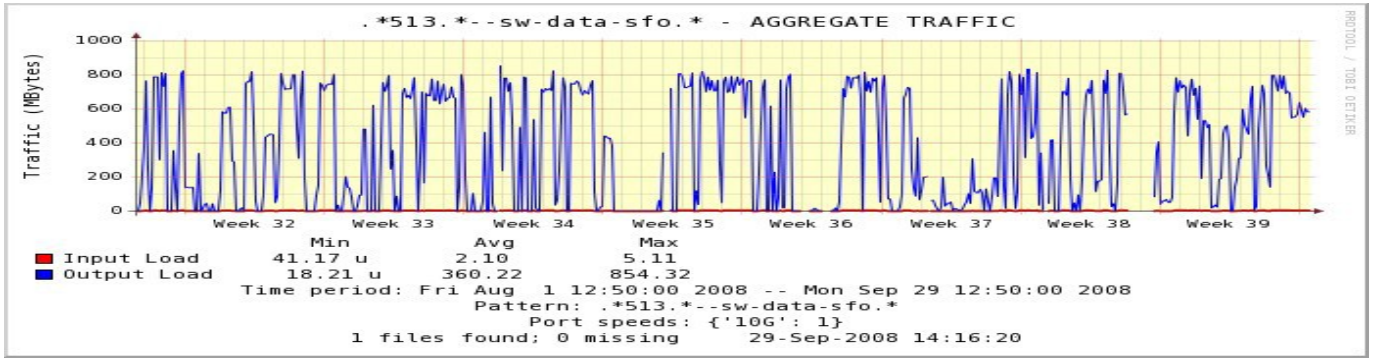
- 27 EF racks, 2 GbE links/rack, 30*8=240 processing units/rack
- Limits referring network limit
- EB/EF performance scaling with the number of racks

Data storage

- Data (passed EF) written into files on SFO and asynchronously transferred to the mass storage at Tier-0
- SFO reading/writing in three different filesystems to maximize the throughput
- SFO providing sustained I/O rate of 550MB/s (target 300MB/s) with peak rate >700 MB/s

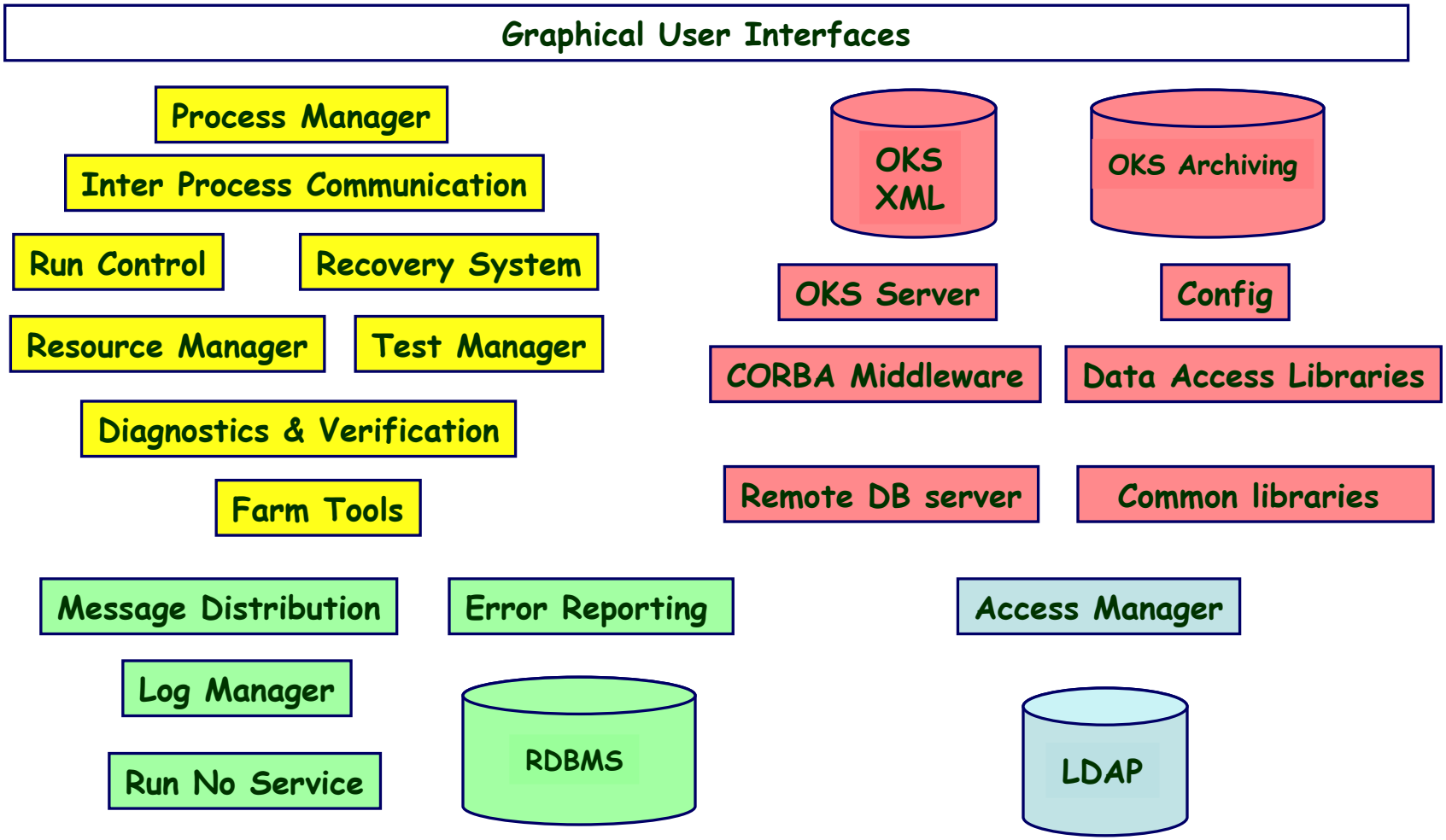
SFO

- CPU: 2 x Intel dual-core 2.0 GHz
- RAM: 4 GB
- Network: 4 of 5 GbE for data collection
- 3 RAID controllers
- 500 GB X 24 disks



Online Configuration and Control

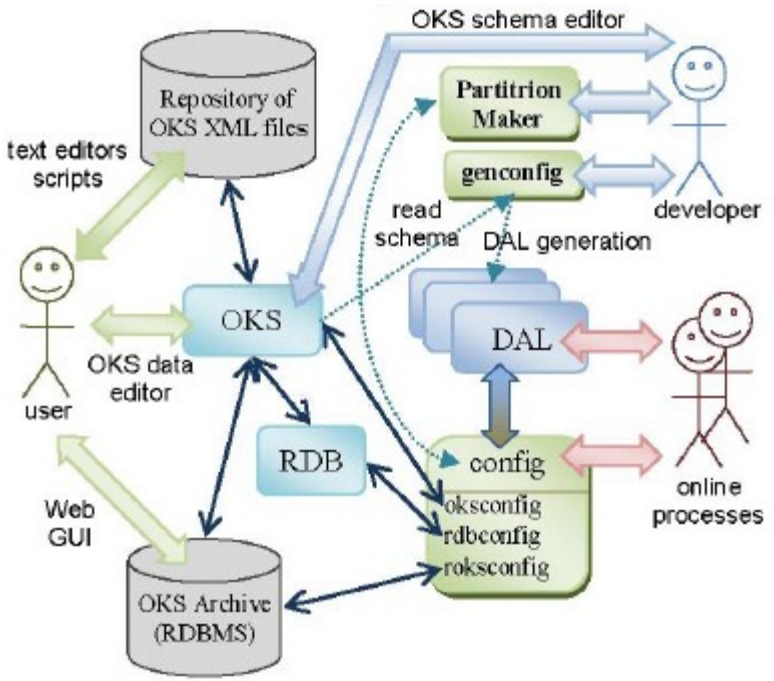
Online System



Online Configuration

- Providing configurations for TDAQ and detector descriptions
- Accessed by ~25K online processes
- Archived for offline access and analysis

Persistent object manager (OKS)
- Object data model
- XML files as persistent storage
- GUI tools for changes
- Remote access (CORBA based)
- Abstract API layer with plugins

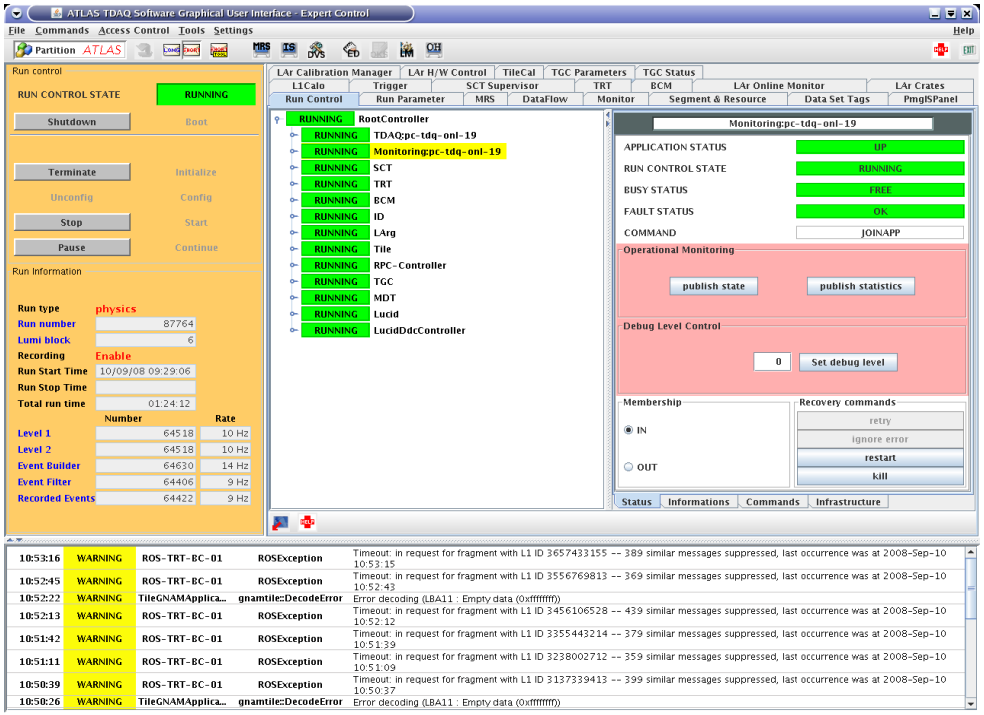


Online Control

- Coordinating for applications (~25K) in data-taking session
- Software operating over dedicated network

Control software facilities

- Process manager
- Run Control
- Diagnostics & verification
- Error recovery system
- IGUI



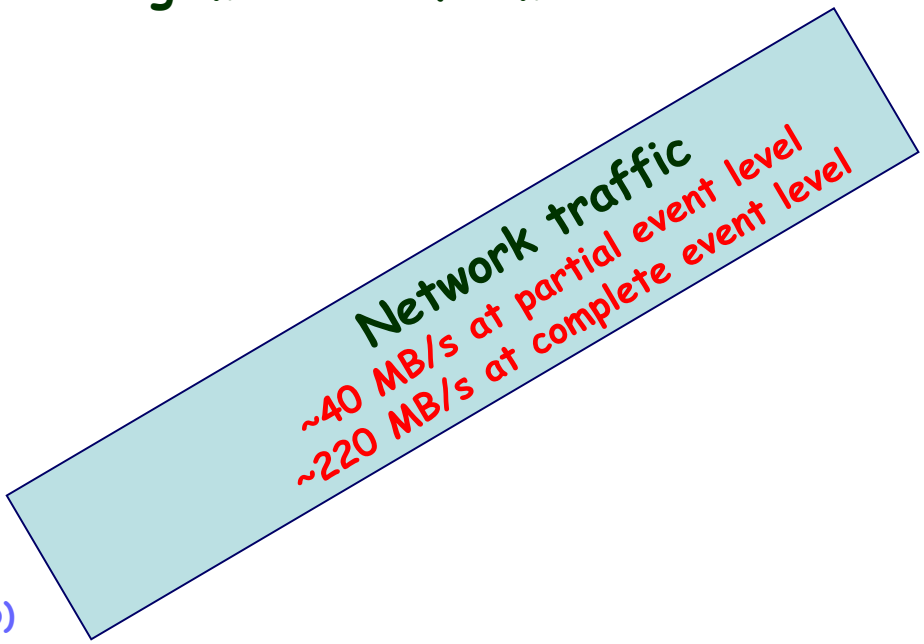
Online System Evolvement

- Audit
  - Formatting, archiving and analyzing messages
- Security
  - Access control and role assignment
- Fault Tolerance
  - Problematic components handling and recovery
- Scalability
  - Re-implementing or optimizing components
- Operability
  - Shifter friendly interfaces and tools

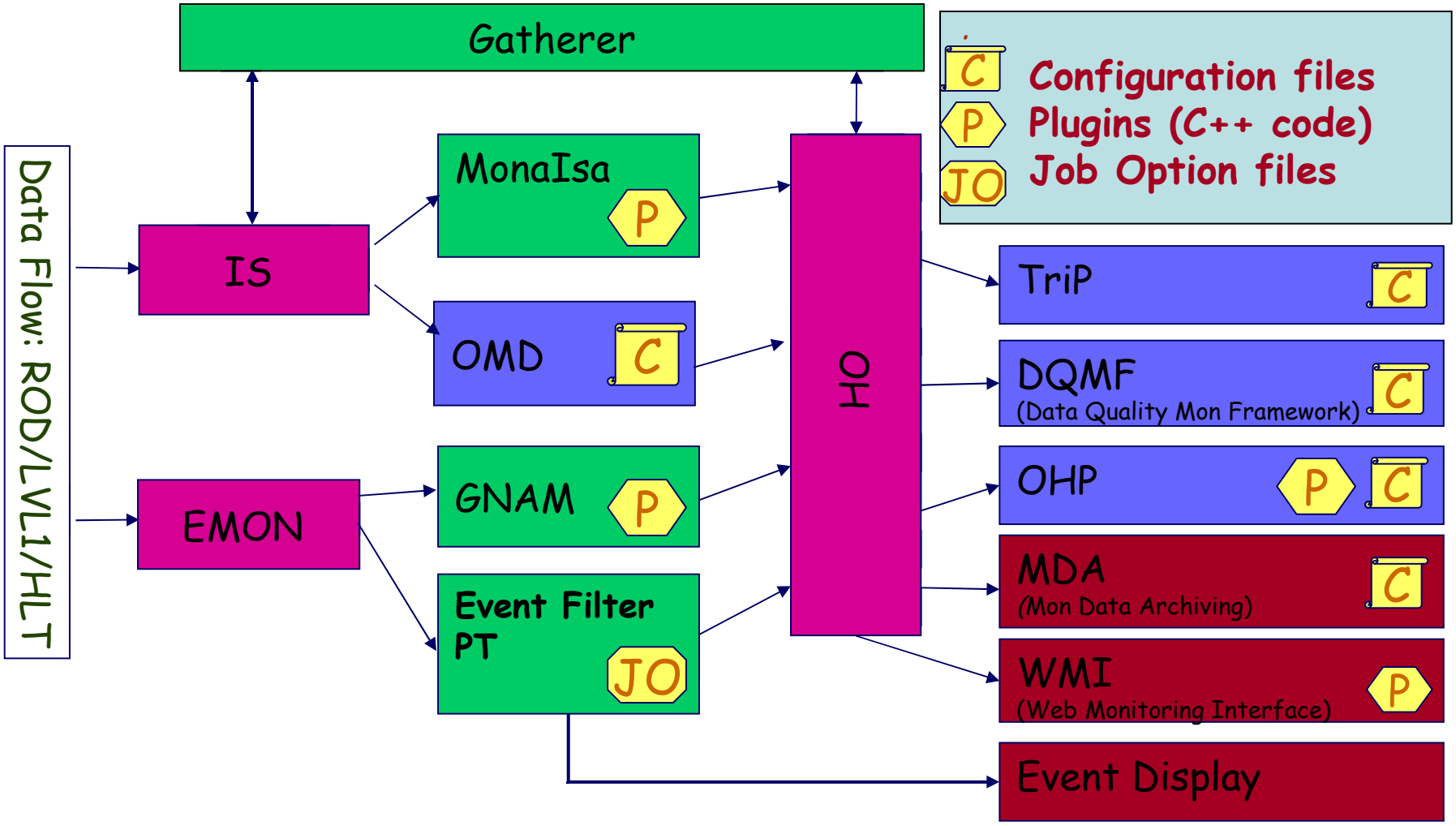
Monitoring

Online Monitoring

- Complex and diverse requirements on data quality and system performance monitoring
  - Analyzing event content
  - Analyzing operational conditions of hardware and software for TDAQ and detector
- Software framework and tools handling massive information
  - Online monitoring services
    - Information Service (IS)
    - Online Histogramming Service (OH)
    - Event Monitoring Service (EMON)
    - … …
  - Monitoring facilities
    - Gatherer
    - GNAM
    - … …
  - Monitoring presenters
    - Online Histogram Presenter (OHP)
    - Trigger Presenter (TriP)
    - Operational Monitoring Display (OMD)
    - … …
  - Histogram archiving
  - … …

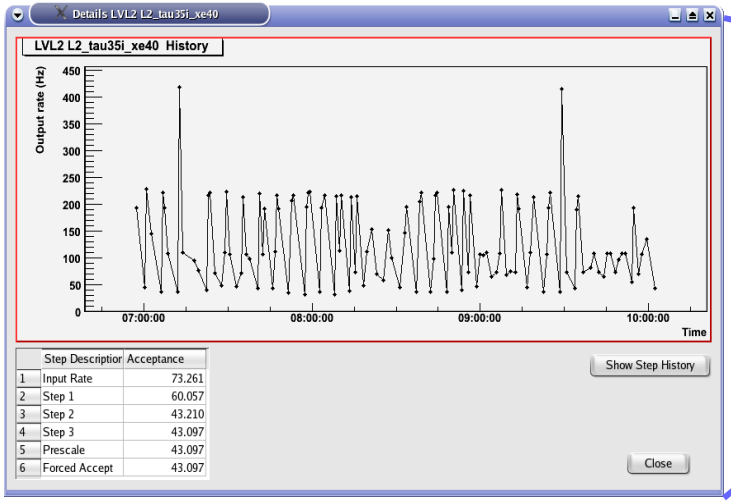


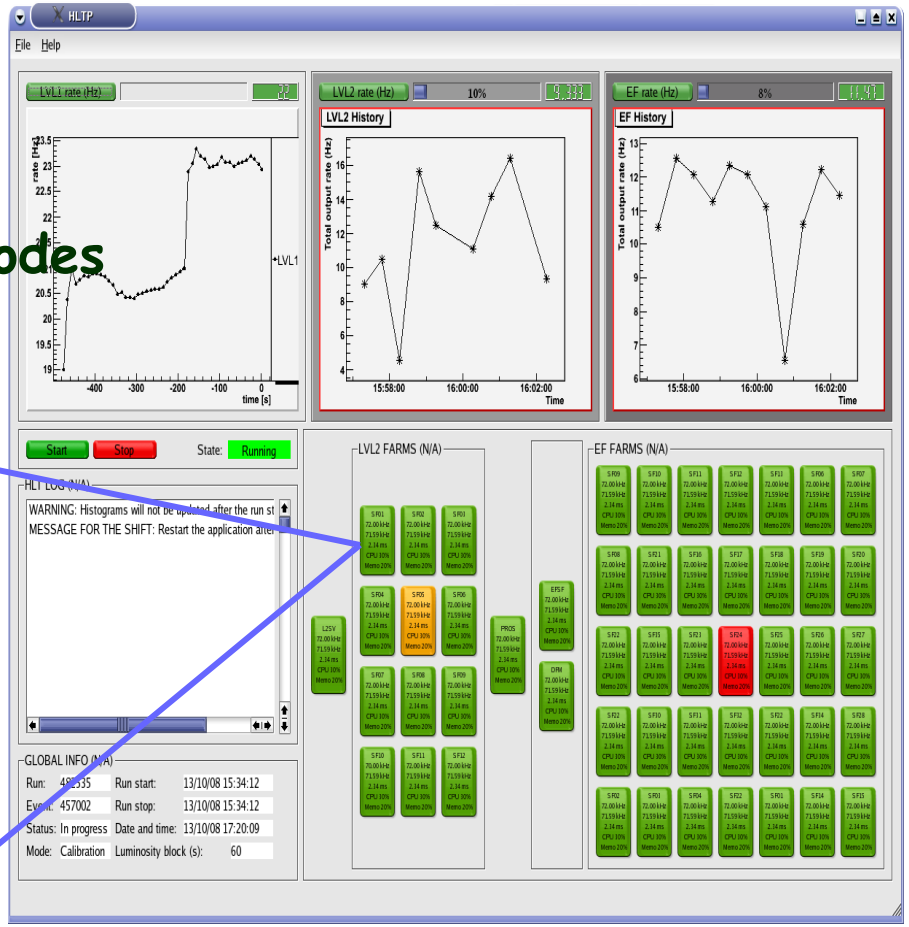
Monitoring software components



Trigger Presenter (TriP)

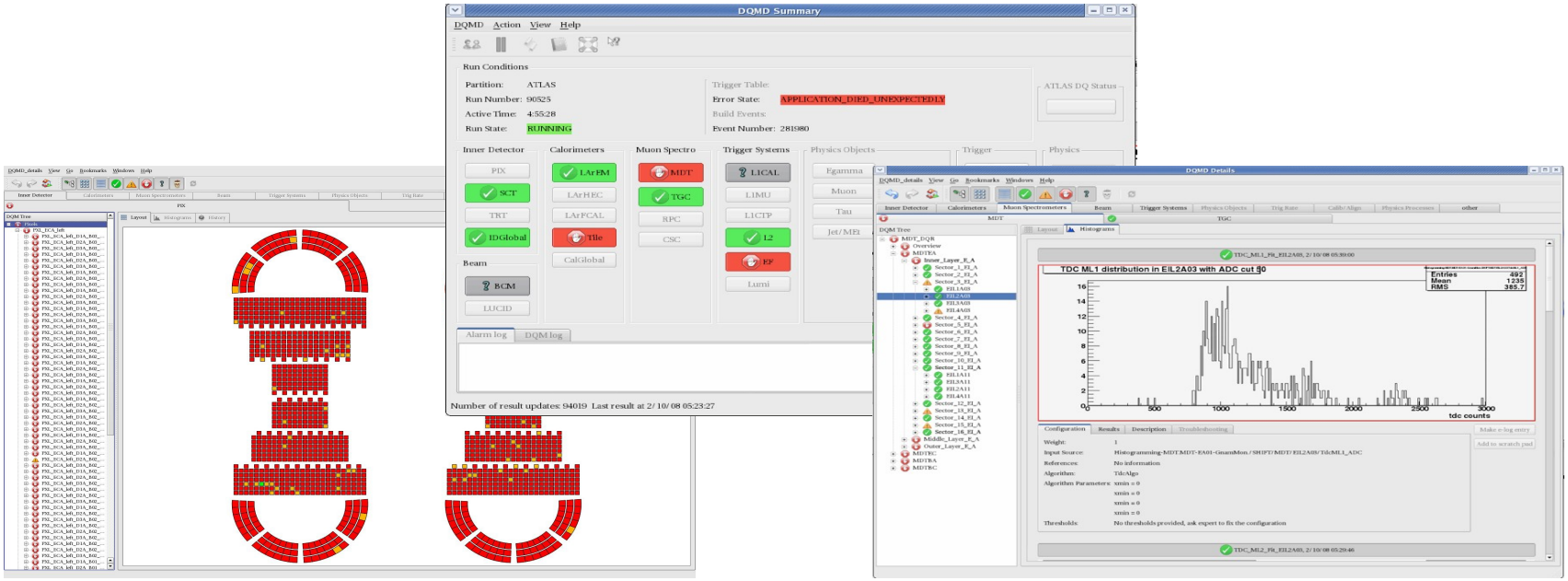
- Presenting trigger information
  - Trigger rates
  - Algorithm chain information
  - Farm status
- Reflecting status with color codes





DQMF (Display)

- Producing DQ status by automatic checking histograms with predefined algorithms
- Storing DQ status in condition database
- Showing overall DQ status for subsystems
- Showing DQ status in geographical view of detector components
- Showing detailed monitoring information



Integration, Commissioning and Early Operation

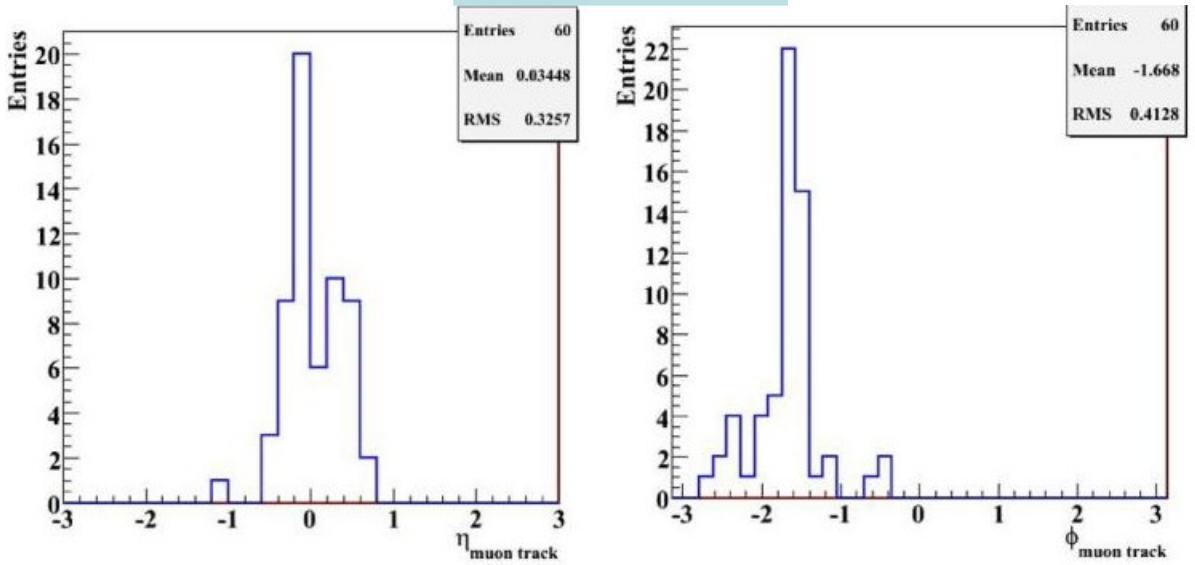
Strategy

- Growing the TDAQ system
  - Exercising on the testbed continuously for years
  - Installing and testing the final system incrementally
- Commissioning the TDAQ system
  - Emulating data source
    - Generating events by subsystem internally
    - Preloading simulated events
    - Replaying cosmic events taken with detectors
  - Performing technical runs
    - Deploying subsystem functionalities
    - Improving subsystem performance
    - Performing stress tests
    - Accumulating operation experience
- Integration with detectors
  - Continuous support for detector installation and commissioning
  - Invaluable feedback on the functionality and operational aspects of the system

1st Full Chain Integration

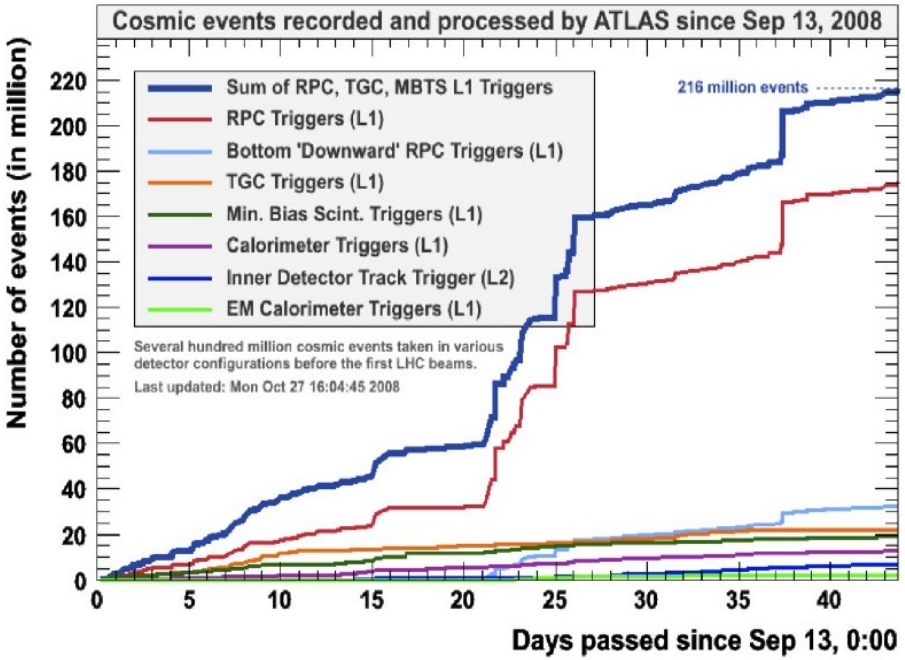
- Ran the full trigger chain with RPC+ LVL1 (muon) + RoIB + LVL2 + EB
- Triggered cosmic ray at 30 Hz (RPC setting) in LVL1 and ran pass-through mode in LVL2 to record all events
- Selected downward muons with LVL2 algorithm
- Stored accepted events after event building

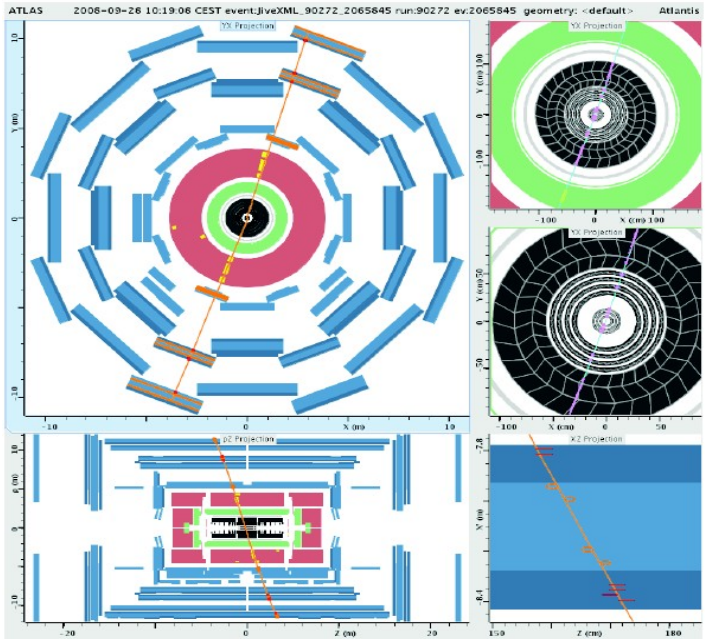
Feb 2007



Integration and Commissioning

- Regularly used for cosmic data taking and integration runs with detectors
  - Improvement on functionality, stability and reliability
- Continuous cosmic data taking (Aug 1 – Oct 31 2008 except beam time)
  - ~ 550 million events, ~ 1 TB, > 600 k files

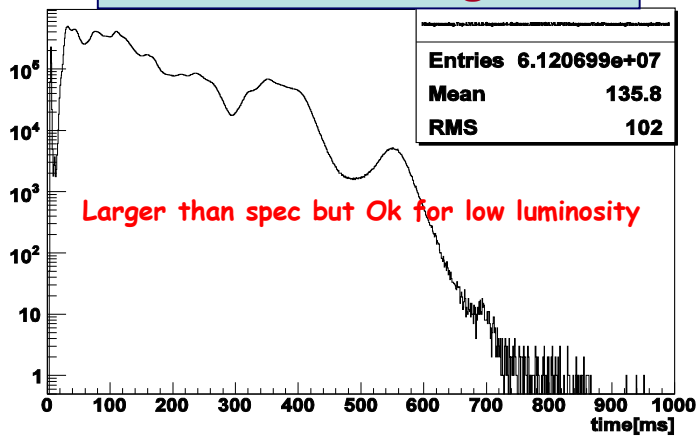




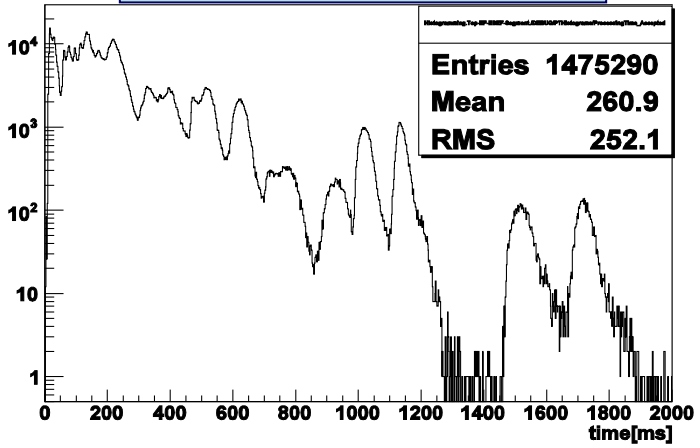
TDAQ at Low Luminosity

- LHC startup luminosity expected to be ~10^{31} cm^{-2} s^{-1} with less bunches
- ATLAS commissioning the trigger and detector systems, and studying the basic Standard Model physics signatures
- A trigger menu (10^{31} menu) deployed by applying low thresholds, loose selections and pass-through mode wherever possible
- The 10^{31} menu continuously exercised in the installed TDAQ infrastructure with simulated data (TDAQ technical run)
  - Full system with preloading simulated data into ROSs
  - Validation of all aspects including trigger algorithms

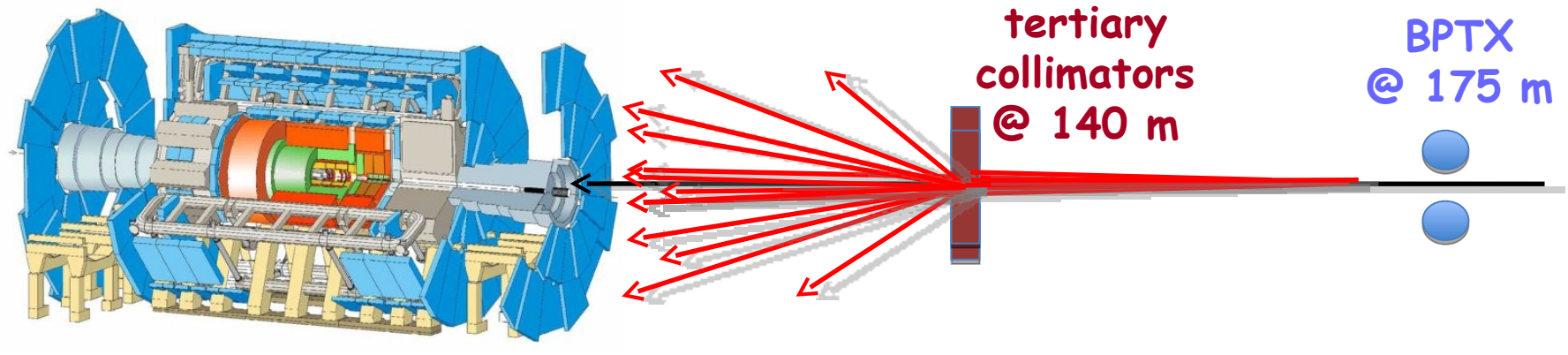
LVL2 Processing Time



EF Processing Time



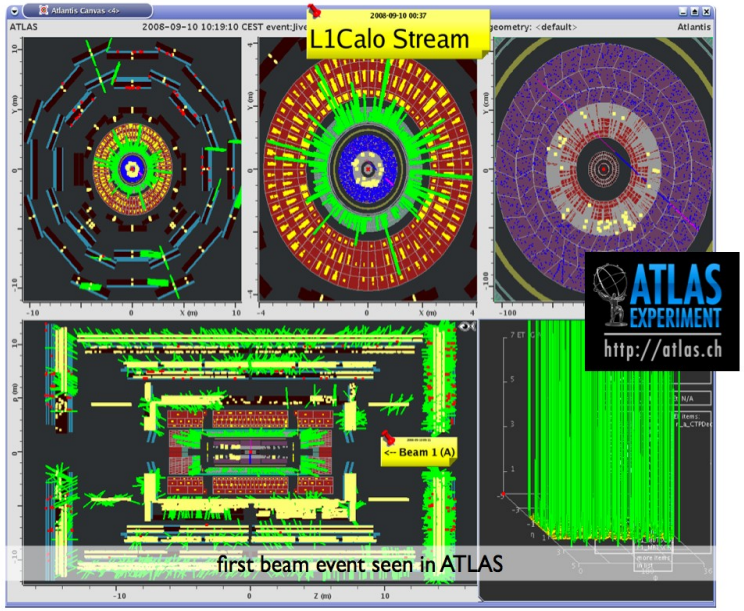
Beam on Sep 10th 2008

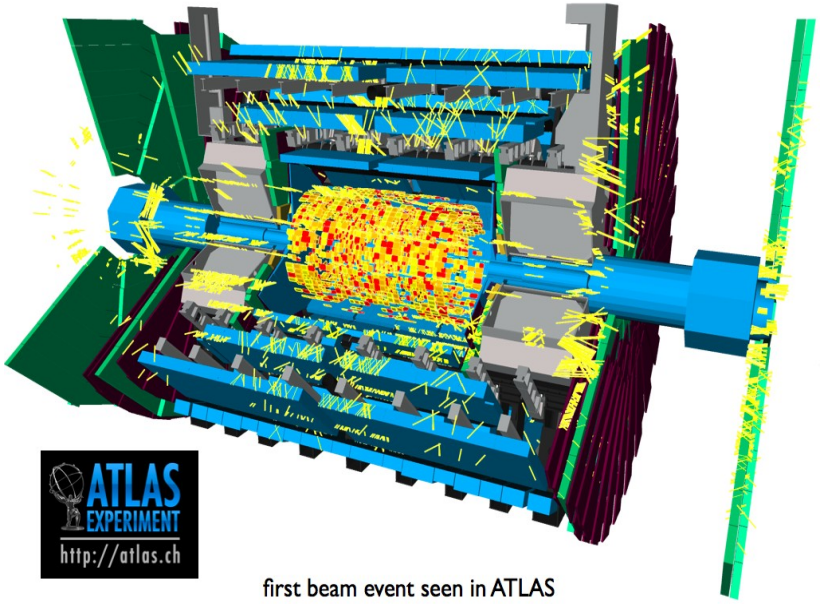


- LHC going step-by-step: stopping beam on collimators, re-aligning with centre, opening collimator, keeping going
  - Splash event from collimators for each beam shot
- ATLAS relying on small radius triggers with well defined cosmic timing (LVL1 calorimeter trigger and Minimum Bias Trigger Scintillator), starting timing in Beam Pickups (BPTX, the timing reference) rapidly to trigger on through-going beam

TDAQ for 1st Beam

- TDAQ system (was) ready for 1st beam
  - 130 ROSs, 4 L2SVs, 120 LVL2 nodes, 74 SFIs, 450 EF nodes, 5 SFOs
  - HLT running but no rejection





Summary

- ✓ The three-level ATLAS trigger architecture and highly distributed DAQ system have been deployed
- ✓ A large fraction of dataflow components has been installed and commissioned
- ✓ The online system has been implemented to meet the challenges
- ✓ The monitoring tools have been developed and are being widely used
- ✓ The TDAQ system is continuously being used for detector integration and commissioning
- ✓ The TDAQ system (was) ready for the 1st beam. We will have beam (again) in 2009

More Materials

Hardware Status

GROUPS	TOTAL	ONLINE	OFFLINE
Gateways	3	3	0
WebServer	1	1	0
CFS	3	3	0
ACR	20	19	1
⊞ SCR	29	24	5
⊟ TDQ	1291	1250	41
LFS	46	46	0
ONL	25	25	0
MON	31	31	0
⊞ ROS	154	154	0
SFI	63	63	0
SFO	5	5	0
DFM	12	12	0
L2SV	2	2	0
XPU	842	807	35
PRESERIES	111	105	6
⊞ SBC	157	149	8
PUB	12	11	1
⊞ DCS	19	18	1
⊞ SWITCH	108	104	4
OTHERS	57	55	2
TOTAL	1700	1637	63

show hosts

Comments
pc-tdq-xpu-0125,pc-tdq-xpu-0126,pc-tdq-xpu-0127,pc-tdq-xpu-0128 reserved by sysadmins for testing nagios/mysql configurations.
pc-tdq-xpu-0129 reserved by sysadmins for testing active directory
pc-tdq-xpu-0109 to pc-tdq-xpu-0114;pc-tdq-xpu-0098,117,121 are reserved [Memory upgrade].
pc-tdq-mon-29, 30, 32 are reserved for NFS 4 performance tests.

Refresh Interval: 3 minutes

Last Updated on: September 21, 2008, 7:12:43 pm

Site Feed

ATLAS TDAQ SYSADMINS
Phone: 164851
Email: atlas-tdaq-sysadmins
Ticket: atlas-tdaq-sysadmin-userticket

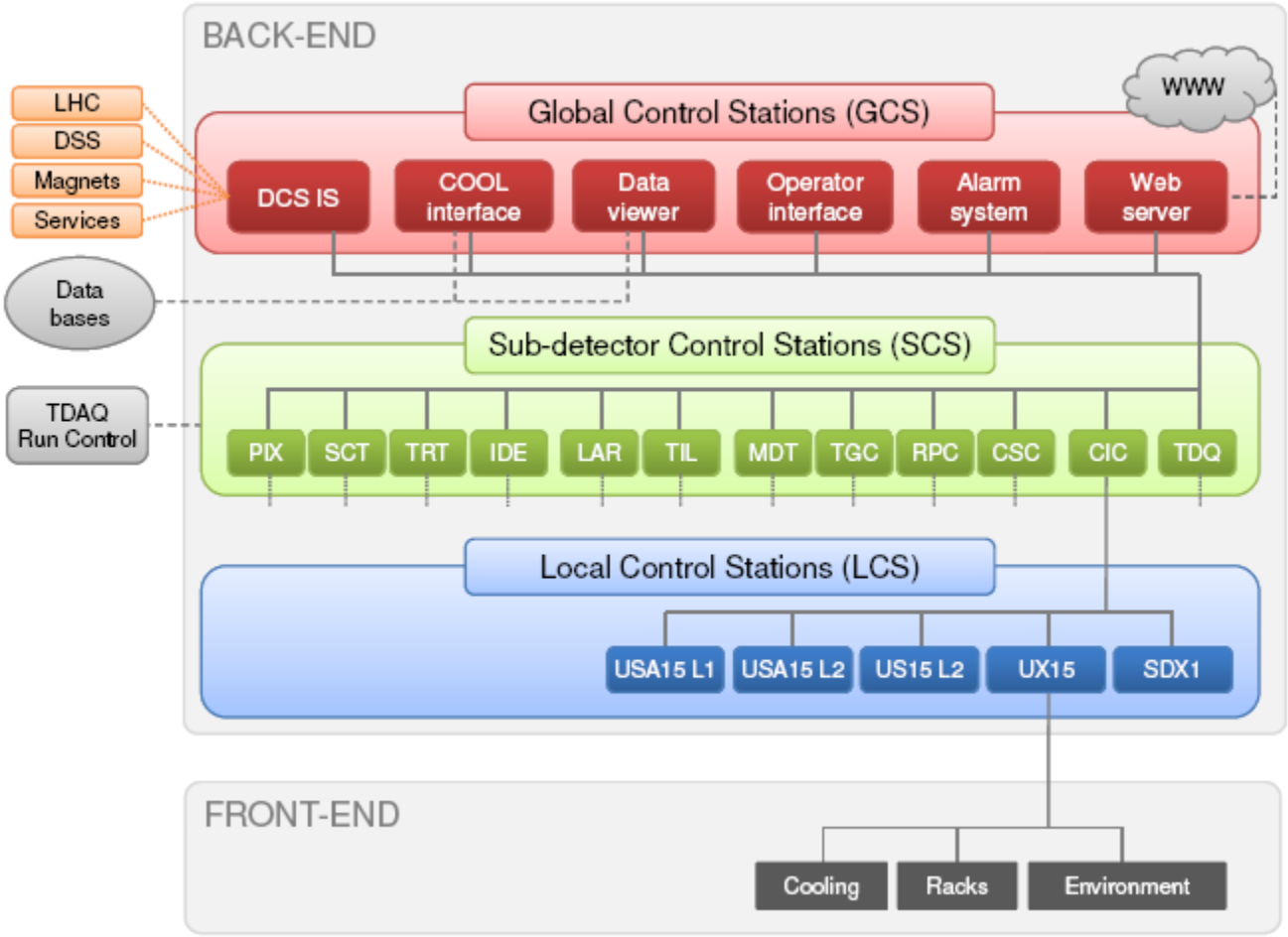
Network

- 2 x Force10 E600
  - Up to7 blades
    630 GbE ports total
    336 GbE ports @ line speed
- Control network
  - Run Control
  - Database
  - Monitoring sampler

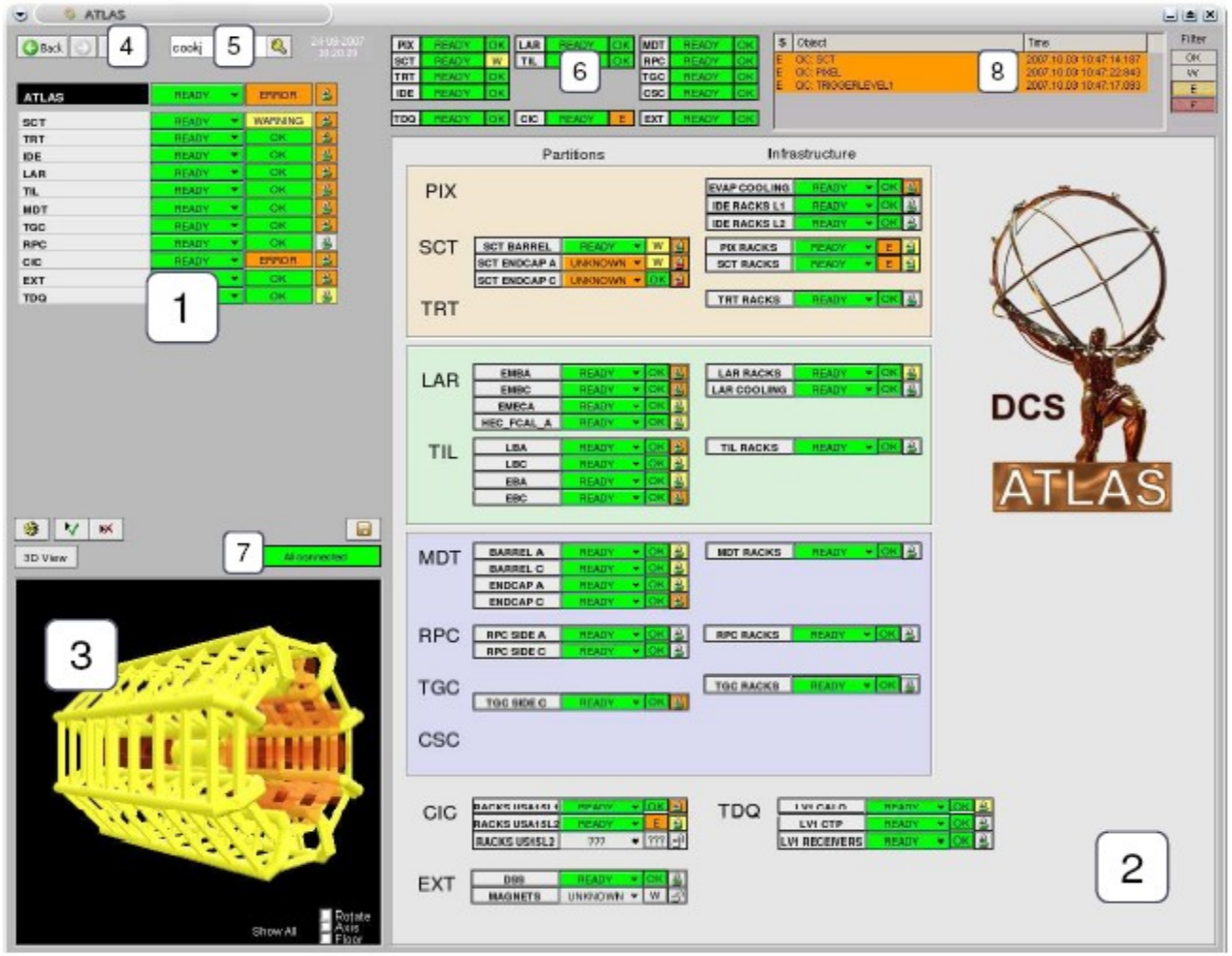
- 2 x Force10 E1200
  - 6 blades x 4 optical 10GbE ports
  - 2 blades x48 copper GbE ports
  - Up to 14 blades
    1260 GbE ports total
    672 GbE ports @ line speed
- Data network
  - LVL2 traffic
  - EB traffic

- Force10 E600
  - Up to 7 blades
    630 GbE ports total
    336 GbE ports @ line speed
- Data network
  - To EF

DCS Architecture



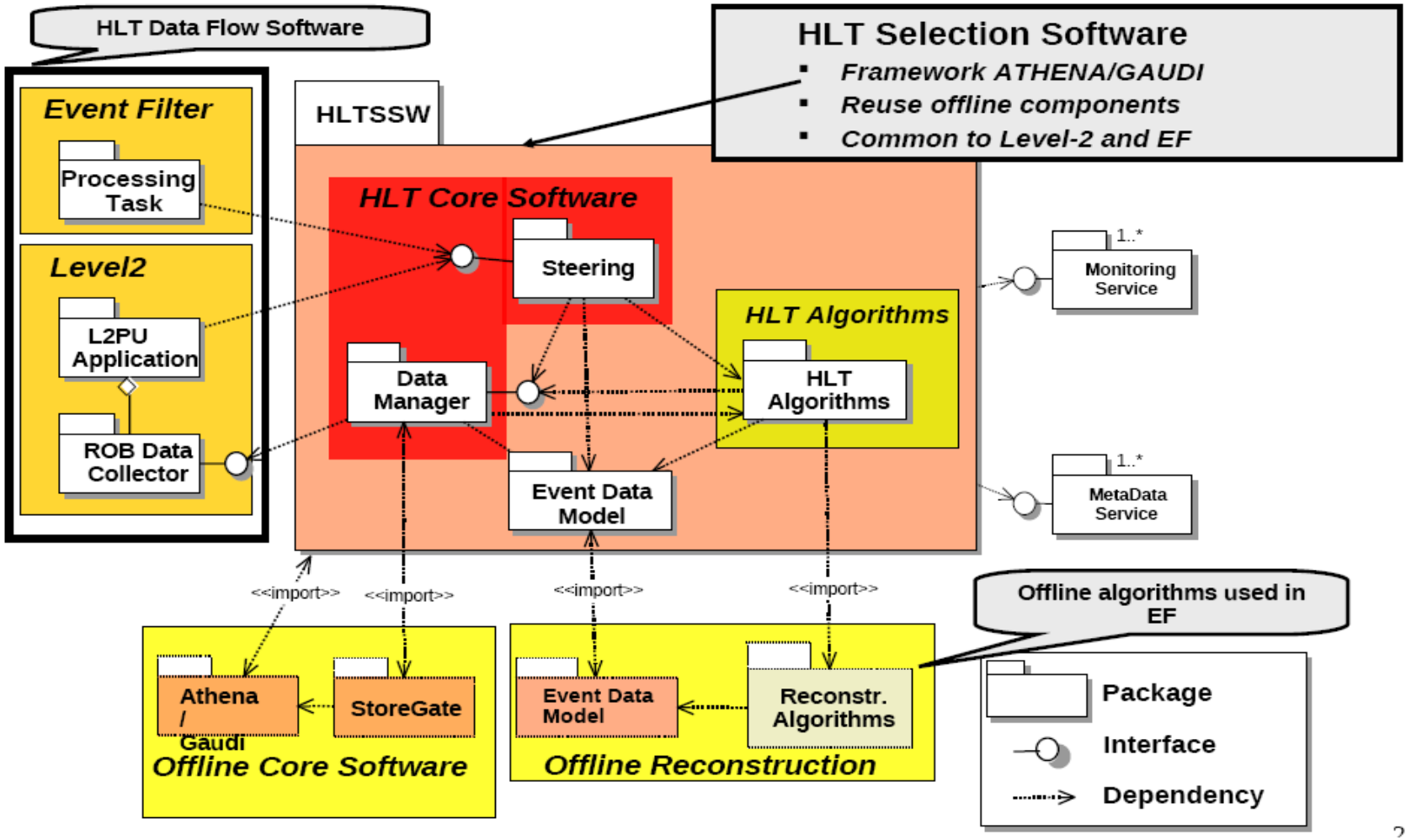
DCS FSM Operator Interface



HLT Requirement

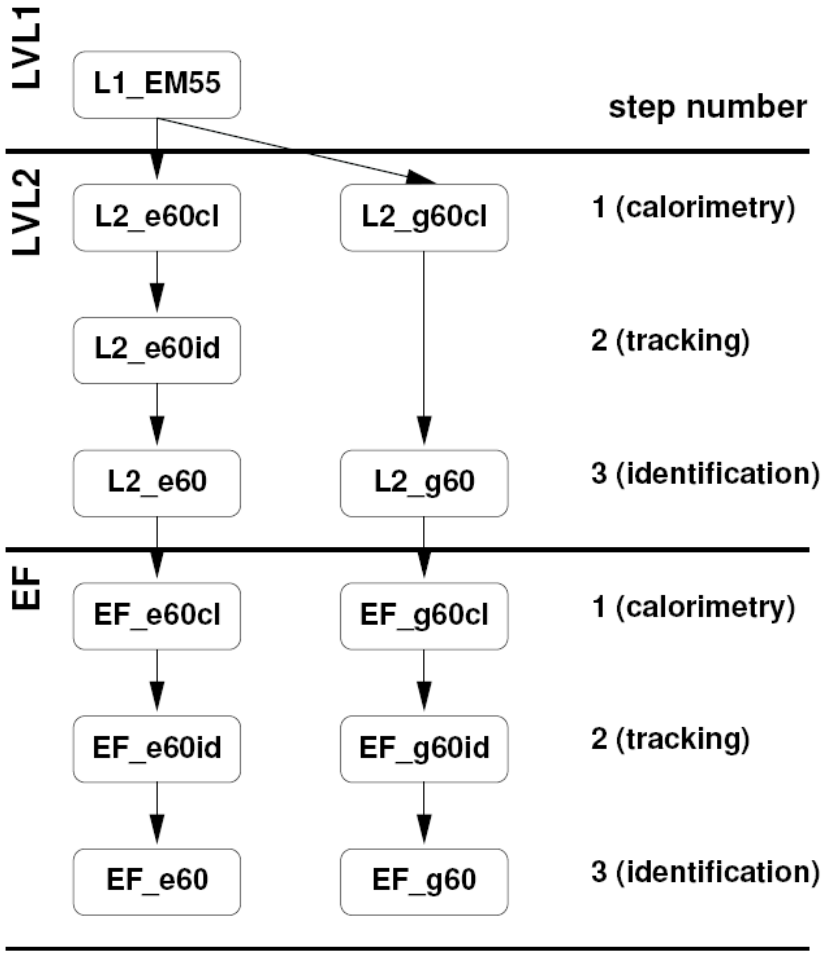
	LVL2	EF
Rates	Full System	Full System
	Input 100 KHz ~500 nodes 200 Hz per node	Input 3.5 KHz ~1800 nodes 2 Hz per node
	Current Hardware	Current Hardware
	8 L2PUs / node ~40 ms/event per L2PU (Data Collection+Processing time)	8 PTs / node ~4 s/event per PT
Data Size	~2% of data	Full Event ~1.5 MB

HLT and Offline Software



HLT Concept

- Algorithms are executed based on trigger chains
- Chains are activated based on result of previous level
- Each chain is divided in steps
- Each step executes an algorithm sequence (one or more algs)
- A step failed to produce an expected result ends the chain
- Any chain can pass the event



Region of Interest (RoI)

- LVL1 identifies Regions of Interest (RoI) defined by (η,φ)
- The average number of RoIs per event is ~1.6
- LVL2 performs selection with data along the RoI path (~2% of full detector)