The ATLAS Tier-0

Overview and Operational Experience

For the ATLAS Tier-0 Team:

Markus ELSING
Luc GOOSSENS
Armin NAIRZ
Guido NEGRI

CERN, Geneva/Switzerland

Overview

- Tier-0 system overview
  - Tier-0 functional requirements
  - Tier-0 quantitative requirements
  - Tier-0 architecture and components
- Operational experience in 2008
  - Computing exercises
  - Detector commissioning, cosmics and single-beam data taking
- Future plans and developments
- Summary

Tier-0 Functional Requirements

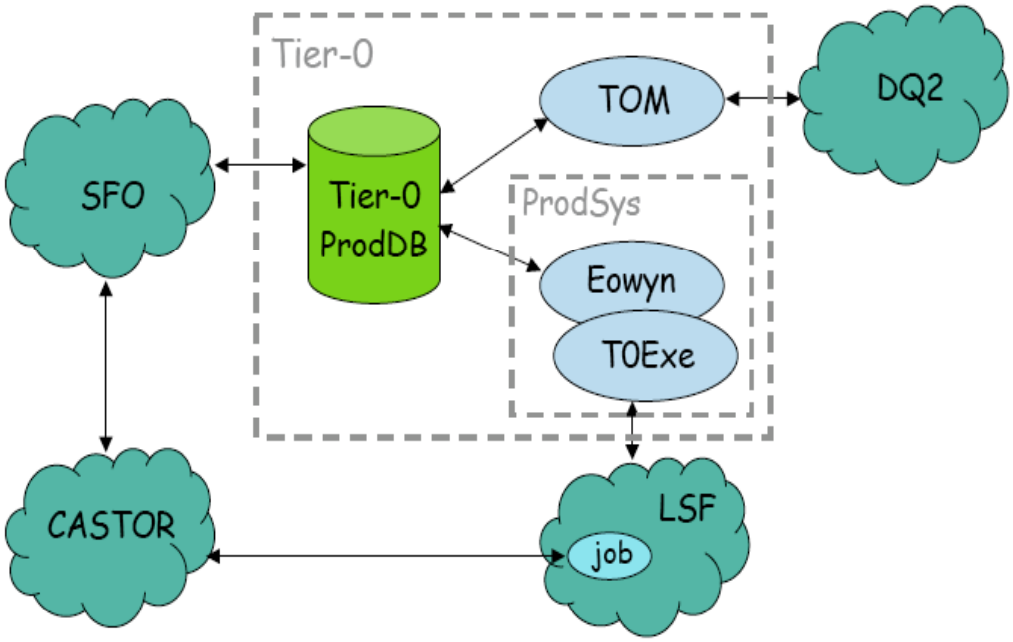
- First-pass ESD (Event Summary Data), AOD (Analysis Object Data), primary DPD (Derived Physics Data) and TAG production
- First-pass calibration and alignment processing
- Express Stream reconstruction (2 passes)
- Uploading of TAG files into TAG database(s)
- Registration of all data products with the ATLAS DDM (Distributed Data Management system) and AMI (ATLAS Metadata Interface)
- Archival of RAW and derived data products on tape
- Support of off-line data-quality monitoring (DQM)
- Merging of RAW files
- Replication of selected data to the CERN Analysis Facility (CAF)

Tier-0 Quantitative Requirements

- O(10k) jobs per day
- O(10k) permanent files per day
  - RAW, ESD, merged DPD and AOD, …
- O(10k) temporary files
  - unmerged RAW, unmerged AOD, …
- Disk writing 880 MB/s
- Disk reading 1900 MB/s
- Tape writing 540 MB/s
- Approx. 1500 reconstruction jobs in parallel
- TAG uploading rate 200 Hz
- Above figures don't include calibration and Express Stream processing (+25%)

Tier-0 Architecture

- The Tier-0 consists of two process entities and a database:
  - An instance of the Tier-0 Manager (TOM) process
  - An instance of a Production System (ProdSys) process
    - Supervisor ("Eowyn")
    - Executor ("T0Executor")
  - The database (ProdDB) persistifies the states of both process entities, in addition to storing logging and monitoring data



- The Tier-0 interacts with four primary external entities:
  - DQ2, the ATLAS distributed data management system
  - SFO (Sub-Farm Output), the Event Filter (=level-3 trigger) output processes
  - CASTOR, the CERN mass-storage system
  - LSF, the CERN batch system

TOM and ProdSys Instances

- Cf. detailed presentation on the Tier-0 software suite at CHEP'07
- Tier-0 Manager (TOM)
  - Driven by RAW data arriving from the SFOs
  - Defines all subsequent datasets, tasks/jobs in the Tier-0 workflow
  - Very modular, easily configurable
    - Most of the processes derive from single template
    - Processes can be loaded at initialisation or run time
    - Process configuration can be dynamically changed at run time
- ProdSys consists of a "facility" neutral supervisor and a "facility" specific executor component
  - Runs the jobs defined by TOM
  - Easily configurable
  - Supervisor component "Eowyn"
    - Communicates with ProdDB
    - Picks up jobs to be run, submits jobs (via the executor plugin), follows the job status, attaches output files to corresponding datasets, …
  - Executor component "T0Executor" (LSF specific)

Monitoring and Shift System

- Extensive Tier-0 monitoring in place
  - Based on cron jobs (acrontab)
  - cron job collects information from various sources
    - Lemon (CERN IT), ProdDB, AFS, …
  - … computes sums, averages, etc. and stores result both in ProdDB and RRD archive (every 5 min)
  - … calls RRD tools to create graphs in web-readable AFS directory
- Tier-0 shift system and infrastructure in place
  - Necessary web interfaces for interventions, electronic logbook, etc.
  - About four months of experience gained during 2008 data taking
  - Lots of useful suggestions and feedback from shifters
    - Input to new developments
  - Team of about 25 experienced shifters so far
    - Enough to run two 8-hour shifts per day, more needed and expected to join for 24/7

Tier-0 Activities in 2008

- The Tier-0 has been integral part in all major data-taking exercises of the last (almost) two years
- Computing exercises in 2008
  - FDRs (Full Dress Rehearsals): goal to test as much as possible the full data processing and analysis chain, from DAQ to the end-user
    - FDR-1: week of Feb 4th 2008
      - » 10 hours of $L=10^{31}$ data (+ one hour of 10^{32}), about 0.4 pb^{-1} of data
    - FDR-2: week of Jun 2nd 2008
      - » A few hours of $L=10^{32}$ data (+ a few minutes of 10^{33}), about 1.5 pb^{-1} of data
      - » Re-processing of a few datasets with improved s/w and calibration (FDR follow-ups)
  - CCRC'08 (WLCG Common Computing Readiness Challenge)
    - Tier-0 mainly used as a mock-data generator
- Cosmics and single-beam data taking in 2008
  - "Milestone" weeks: cosmics data taking, with combined sub-detector systems
    - M6 (Mar 2008), M7 (May-Jun 2008), M8 (Jul-Aug 2008)
  - Detector weeks: commissioning of individual sub-detector systems with cosmics
  - Continuous cosmics data-taking period (Jun-Oct 2008)
  - Single beam data-taking period (Sep 2008)

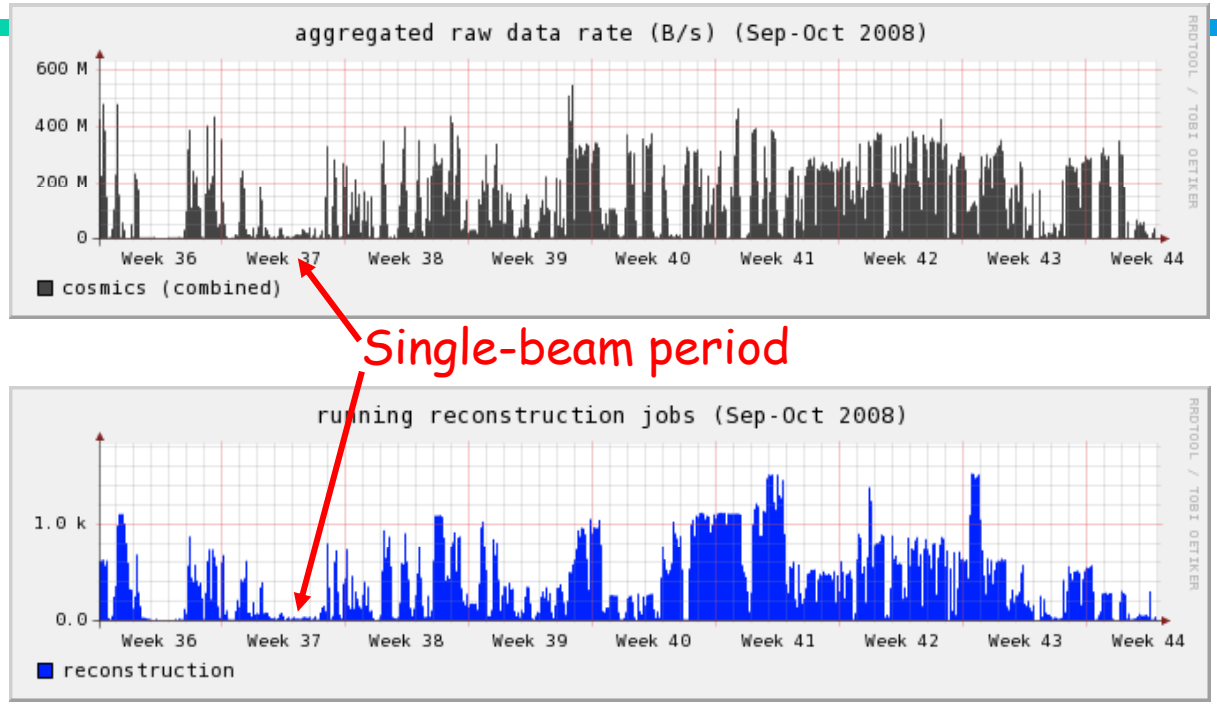
Important Achievements in 2008

- Implementation of an SFO "handshake" database and a "handshake" mechanism between SFOs and Tier-0
  - Contains all necessary information about runs, luminosity blocks, files
    - Completeness, status on SFOs, transfer status, file metadata, etc.
  - Database itself is protected by the firewall of the online network
    - Read-only offline replica in place (Oracle streams)
    - Information is pulled from there to the Tier-0 ProdDB
- Exercising of full processing chains/cycles for physics, express, and calibration/alignment streams (during FDR1&2)
- Development and establishment of extensive monitoring and a fully functional shift system/infrastructure
  - Related: web interfaces for changing the Tier-0 run configuration and to sign off datasets for processing
    - (To be) used by offline commissioning and data preparation groups
- Partial re-design of TOM, to achieve more modularity and flexibility
  - After experiences with FDR
    - Requirements became only clear in the course of the exercises

Tier-0 Readiness

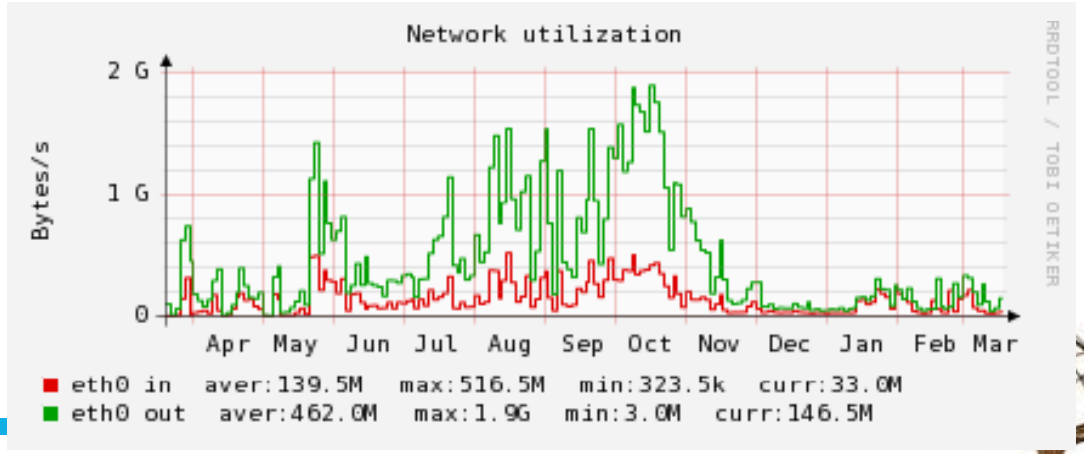
- Tier-0 has been working basically fine and reliably
  - Software and infrastructure in place to handle nominal data flows and carry out all necessary work flows
    - Tier-0 Management System (TOM), ProdSys instance (Eowyn, T0Executor)
    - Hardware infrastructure: CASTOR pools, LSF batch farm
  - Full processing chain in place
    - Interactions with online SFO DB (look-up of new RAW data)
    - Dataset definition, task/job configuration and definition
    - Running of express-stream, bulk, calibration (proof-of-principle) processing
    - Running of reconstruction, (all sorts of) merging, uploading jobs
    - Input for data quality monitoring (DQM) groups, creation of DQM web displays
    - Registration of data with DDM and AMI
    - Replication of data necessary for calibration/alignment to the CAF
    - Running of clean-up procedures (on CASTOR, AFS)
  - Extensive monitoring in place
  - Shift system in place

Tier-0 Activities: Data Taking



- Left: Tier-0 monitoring plots from cosmics and single-beam data taking in the period Sep-Oct 2008

- Right: I/O rates on the main Tier-0/CASTOR production pool in the period Mar 2008 to Mar 2009



~25% ~35% of nominal annual rate

Data Type	June 24th - October 28th, 2008			January 1st - June 24th, 2008		
	Files	Events	Tot. Size [MB]	Files	Events	Tot. Size [MB]
RAW	693,509	464,362,139	1,104,305,943	87,796	40,059,042	136,324,739
CALRAW	17,028	110,463,488	17,558,049	130	n/a	1,886
ESD	651,109	434,868,635	240,578,525	104,645	56,379,218	14,017,253
ESD_FILTERED	474,716	n/a	7,795,534	70,284	n/a	242,337
CBNT	651,109	434,868,635	404,906,622	104,645	56,379,218	21,286,696
AOD	2,572	79,046,922	1,894,468	---	---	---
TAG	2,572	79,046,922	4,407	---	---	---
TAG_COMM	4,623	325,676,278	27,558	1,914	17,289,650	388
HIST	5,378	n/a	183,594	2278	n/a	77,183
NTUP_PIXELCALIB	354	n/a	21,974	---	---	---
NTUP_MUONCALIB	5,022	n/a	2,285,878	---	---	---

Glossary:

- RAW: physics and debug streams from DAQ
- CALRAW: calibration streams from DAQ
- ESD_FILTERED: ESDs of events with at least one ID track
- CBNT: “Combined n-tuples” (plain ROOT, to be phased out in 2009)
- TAG_COMM: commissioning TAGs (produced in reconstruction step)
- TAG: physics TAGs (produced in AOD merging step)
- HIST: merged DQM histograms
- NTUP_*CALIB: merged n-tuples for Pixel and Muon detector calibration purposes

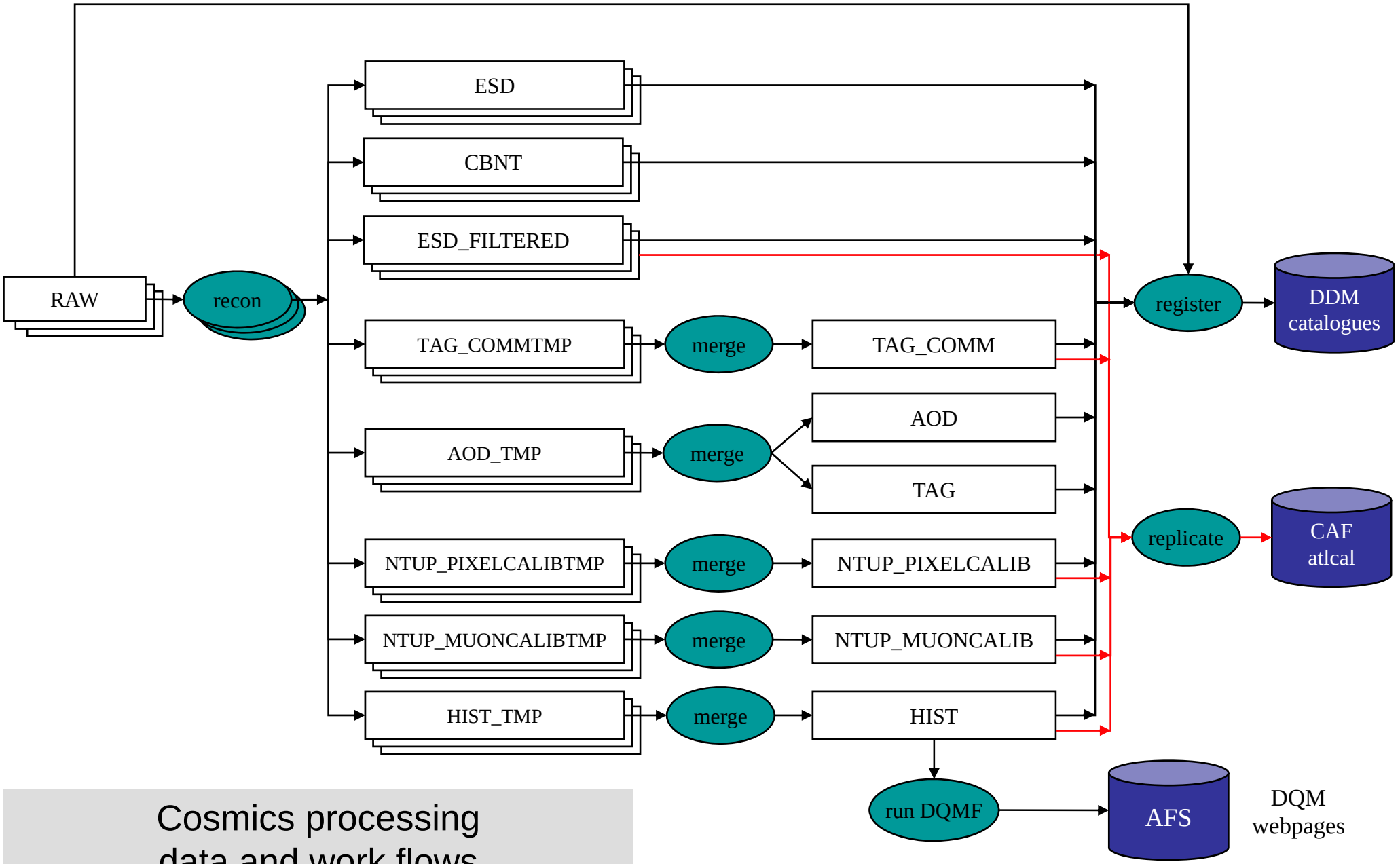
Cosmics data taken and processed Jan-Oct 2008

Task Type	June 24th - October 28th, 2008				January 1st - June 24th, 2008			
	Tasks	Total Jobs	Done Jobs	CPU Wall Time [d]	Tasks	Total Jobs	Done Jobs	CPU Wall Time [d]
recon	5,602	691,296	651,411	29,246	2,081	107,649	105,547	4,236
aodmerge	774	2,581	2,572	219	n/a	n/a	n/a	n/a
histmerge (1st pass)	5,469	30,381	30,322	298	2,004	3,907	3,872	101
histmerge (2nd pass)	5,414	5,400	5,392	15	1,122	1,112	1,110	2
dqmdisplay	5,349	5,349	5,345	24	1,988	2,190	1,879	8
tagmerge (1st pass)	4,658	9,768	9,666	16	829	1,460	1,460	<1
tagmerge (2nd pass)	4,632	4,628	4,625	7	829	820	820	<1
tagupload	n/a	n/a	n/a	n/a	400	26,820	26,820	1
ntupmerge	908	5,381	5,376	15	n/a	n/a	n/a	n/a

Glossary:

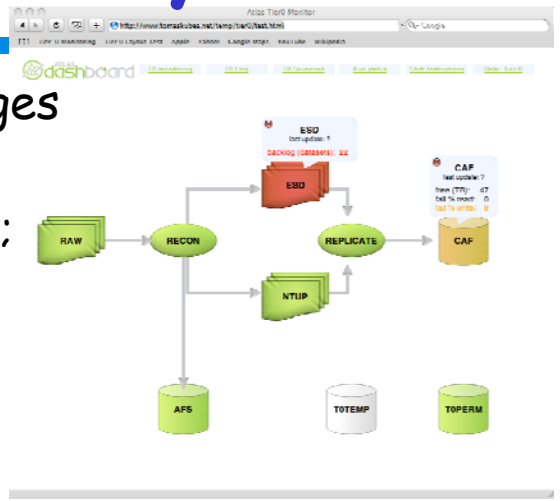
- recon: first-pass reconstruction (usually one job per RAW file)
- aodmerge: AOD merging and physics TAG production
- histmerge: DQM histogram merging, done in two steps
  (step 1: incremental merging every 15min;
  step 2: final merging, to get one file per run/stream)
- dqmdisplay: DQM webpage creation (based on merged run/stream histograms)
- tagmerge: commissioning TAG merging (done in two steps, similarly to histmerge)
- tagupload: uploading of TAGs to DB (done only for M8 tags)
- ntupmerge: merging of Pixel and Muon calibration n-tuples

Cosmics processing tasks and jobs run Jan-Oct 2008



Cosmics processing
data and work flows
(as of beginning of October; simplified)

Work Plan for 2009/10 and Beyond



- Development of new shifters' interface and monitoring pages
  - Based on experience from last year's running
  - Easier overview on running processes and their states, alarms; easier means of intervention
  - Prototypes expected in the coming 1-2 months
  - Fine-tuning during Spring/Summer cosmics data taking
  - Expected to be operational for collisions
- Development of an automated task/job management system for commissioning, calibration and alignment groups
  - Brainstorming meeting with involved groups end of Jan 2009
  - Prototypes expected in the coming 1-2 months
  - Expected to be operational for collisions
- Long term: focus on maintenance
  - Refinement of the system, extension/adjustment of its functionality
  - Based on experience from 2009/10 cosmics and collisions running
  - Limited experience with collisions data taking yet (only from FDRs)
    - Have to be prepared for possible (up to "drastic") changes to the requirements

Summary

- The Tier-0 has proven to work stably and reliably during all data-taking and computing exercises in 2008
- The system is mature and robust enough to handle the expected rates and throughput for 2009/10 cosmics and collisions data taking
  - Hardware, software, monitoring and shift infrastructure, …
- Despite experience limited only to FDR exercises, we believe that the Tier-0 software suite is prepared for handling collisions data taking workflows
  - And flexible enough to be adjusted quickly to new requirements
- New features/improvements are under development and expected to be ready (and tested) before collisions data taking
  - New monitoring pages and shifters' interface
  - Task management system for calibration and alignment activities

Back-Up Slides

Tier-0 Hardware Infrastructure

- Current Tier-0 hardware
  - CASTOR pools
    - "t0atlas": main production pool, for permanent data, with tape back-end
      - » 38 disk servers, 200 TB
      - » Used in common by SFOs, Tier-0, DDM
    - "t0merge": for transient data, disk only
      - » 15 disk serves, 80 TB
  - LSF batch farm
    - About 200 nodes, 1500 cores
  - Oracle production database (ProdDB) instance
  - Two server machines
    - One for production, one for development and as a spare