European Physical Society
HEP 2007

ATLAS Computing System Commissioning: Real-Time Data Processing and Distribution Tests

Armin NAIRZ (CERN)

on behalf of

Luc GOOSSENS (CERN), Armin NAIRZ (CERN)
ATLAS Tier-0 Team

Miguel BRANCO (CERN), David CAMERON (Oslo), Pedro SALGADO (UTA, CERN)
ATLAS Distributed Data Management

Dario BARBERIS (Genoa, CERN), Kors BOS (NIKHEF), Gilbert POULARD (CERN)
ATLAS Computing Management

Overview of this Talk

- Introduction
  - ATLAS Offline Data Flow
  - ATLAS Tier Structure and Tier Functions
  - Tier-0 Scale
- Tier-0 Internal Tests
- Tier-0 → Tier-1 Export Tests
- The Full Dress Rehearsal (FDR)
  - Scope
  - Preparatory Exercises (Stream Test)
- Plans for 2007/2008
- Summary and Conclusions

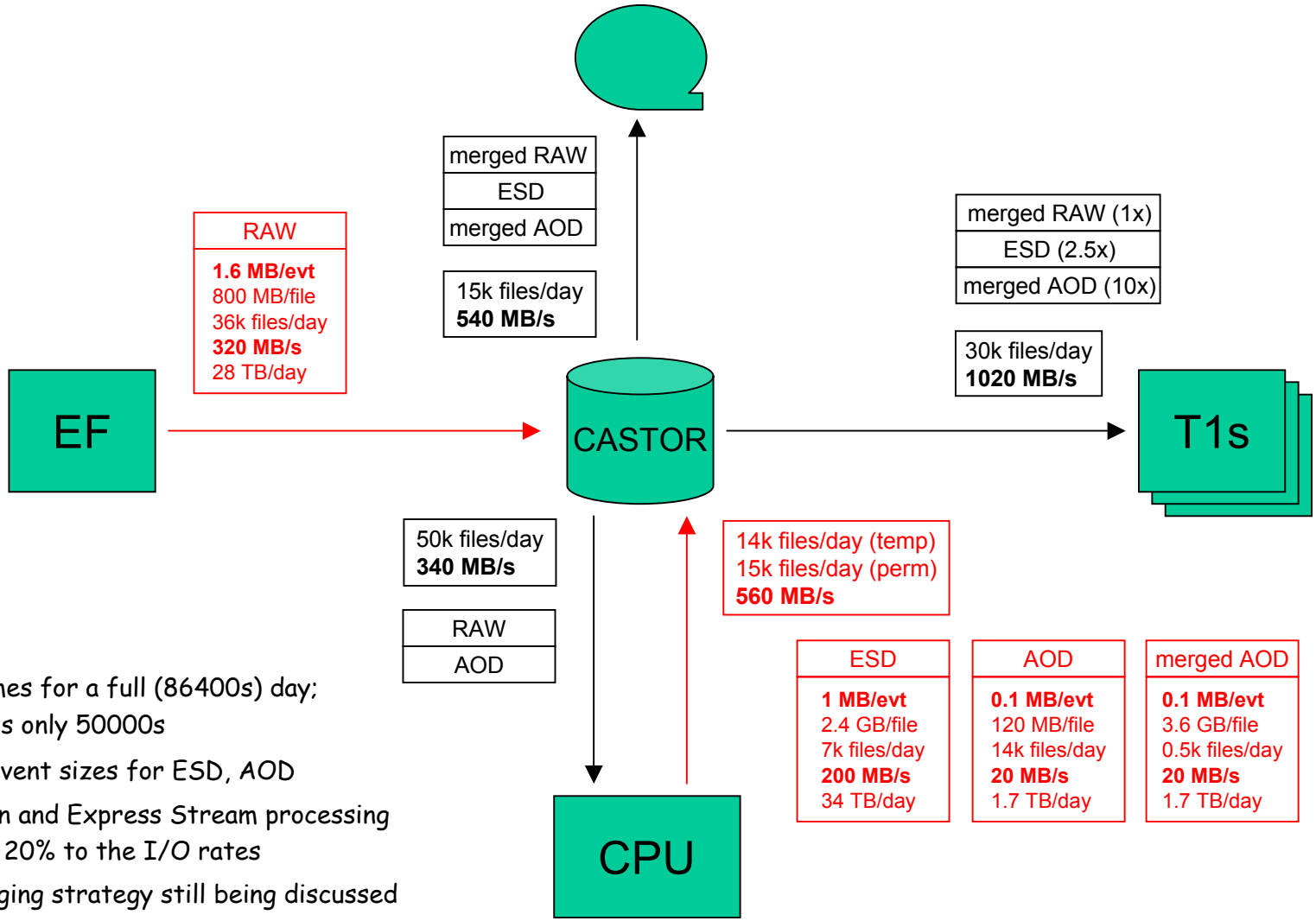
ATLAS Offline Data Flow

- Data flow described as anticipated in the ATLAS Computing Model
- High-level trigger ("Event Filter") processors send outputs to 5-10 Sub-Farm Output managers (SFOs)
  - Trigger rate: 200 Hz
- SFOs assemble files and transfer them to the Tier-0 centre
  - File closure at "luminosity block" (~1min) and run boundaries
  - Data streaming: 3-4 calibration streams; "Express Stream"; 5-6 physics streams
- Tier-0 processes the data
  - Calibration and alignment processing
  - Prompt Express Stream processing, using "best possible" calibration
  - First-pass reconstruction of physics streams
    - Latency up to 24h, waiting until calib/align processing for the run has finished
    - Produces Event Summary Data (ESD), Analysis Object Data (AOD), TAGs
  - Uploading of TAGs to a central database
  - Variety of potential additional tasks/services
    - File merging, data-quality monitoring, meta-data (file, dataset) uploading, ...

ATLAS Offline Data Flow (cont.)

- Tier-0 archives RAW and derived data (ESD,AOD,TAG) on tape
- Tier-0 exports RAW,ESD,AOD,TAG data to the 10 Tier-1 centres, via the ATLAS Distributed Data Management (DDM) system
  - One copy of RAW
  - Two copies of ESD (plus one full copy to BNL)
  - 10 copies of AOD and TAG
- Tier-1 centres archive the data and distribute them to their associated Tier-2 centres
  - At least one full additional, shared AOD copy in each "Tier-1 cloud"
- Tier-1 centres are responsible for reprocessing of the RAW data
  - Once better calibration and alignment becomes available
- User analysis is done on the Tier-2 centres
  - User jobs are directed to the data at the Tier-2s, to avoid additional, "chaotic" data replication

Tier-0 Scale(*)



(*) Remarks:

- Data volumes for a full (86400s) day; "CM day" is only 50000s
- "Target" event sizes for ESD, AOD
- Calibration and Express Stream processing add about 20% to the I/O rates
- RAW merging strategy still being discussed

Overview of this Talk

- Introduction
  - ATLAS Dataflow
  - ATLAS Tier Organisation and Functions
  - Tier-0 Scale
- Tier-0 Internal Tests
- Tier-0 → Tier-1 Export Tests
- The Full Dress Rehearsal (FDR)
  - Scope
  - Preparatory Exercises (Stream Test)
- Plans for 2007/2008
- Summary and Conclusions

Tier-0 Internal Tests

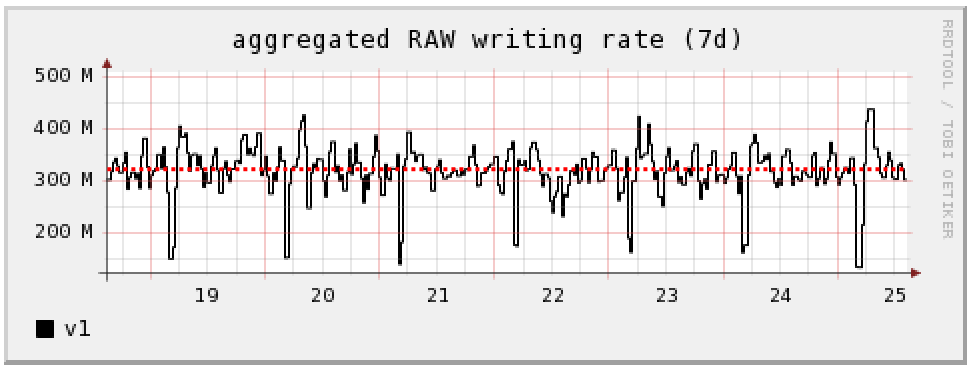
- Tier-0 test activities started in 2005
  - Test series since then:
    - Nov/Dec 2005
    - Jan 2006; Jun 2006; Sep/Oct 2006
    - Feb 2007; since May 2007 (in continuous "test mode")
  - Participation in ATLAS "Milestone-3" (M3) cosmics data taking run (Jun 2007)
- Tier-0 internal processing strategy
  - Focus on data transfer, realistic workflow
  - Run (semi-)dummy executables, use (semi-)dummy data
    - Due to limited CPU resources
    - Replace with more realistic executables once s/w and more resources become available
- Usage of dedicated CERN h/w resources
  - CASTOR: dedicated pool infrastructure
    - Currently 54 disk servers (~300 TB disk buffer), 16 tape drives
  - LSF (CERN batch farm): dedicated cluster, O(100-200) nodes
- Tier-0 internal "nominal" transfer rates were reached already in Jan 06

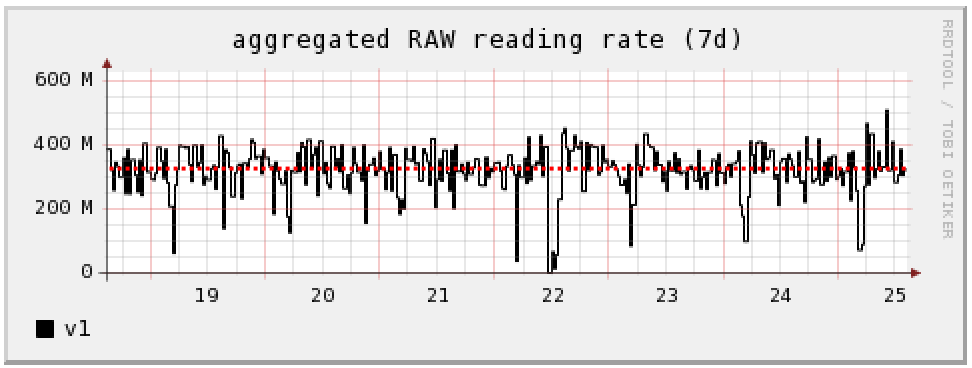
Tier-0 Internal Tests

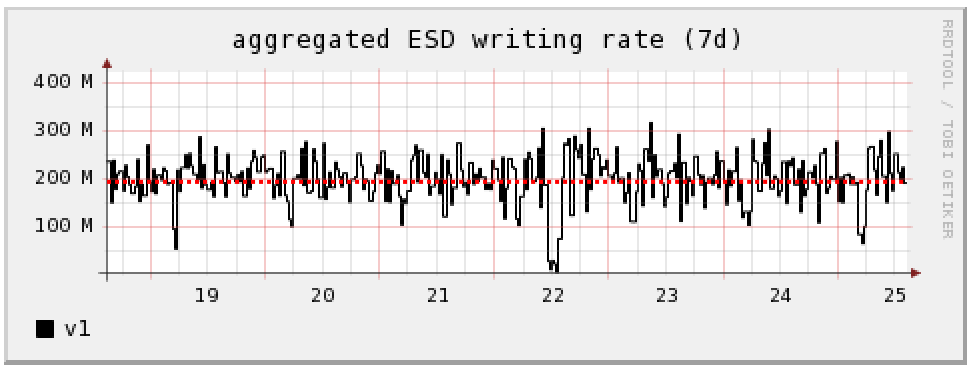
- Sep/Oct 2006 test: Tier-0 internal achievements (selection)
  - Detailed workflow for physics reconstruction and calib/align processing in place
  - Extensive monitoring (more than 90 variables)
    - Live monitoring page: http://atlas.web.cern.ch/Atlas/tier0/monitoring/
  - One week of stable running at 140% "nominal" rate (Oct 9-15)
  - 1.5 PB of data moved in total, 400k jobs executed in total, ...
- Feb 2007 test: "CASTOR crisis"
  - Common stager for whole of ATLAS overloaded, led to complete break-down
  - CASTOR/IT Task Force was put in place, to address problems quickly
  - Separate Tier-0 stager was set up in May 2007
    - Test-bed for newly developed CASTOR software
    - Tier-0 tests dedicated to providing fast feed-back to CASTOR developers
  - After successful test period, reverted to common ATLAS stager (Jun 2007)
- Since then the Tier-0 has been running in continuous "test mode"
  - Providing a "base load" on the stager, sustained over periods of weeks
  - CASTOR performance excellent (better than ever before !)

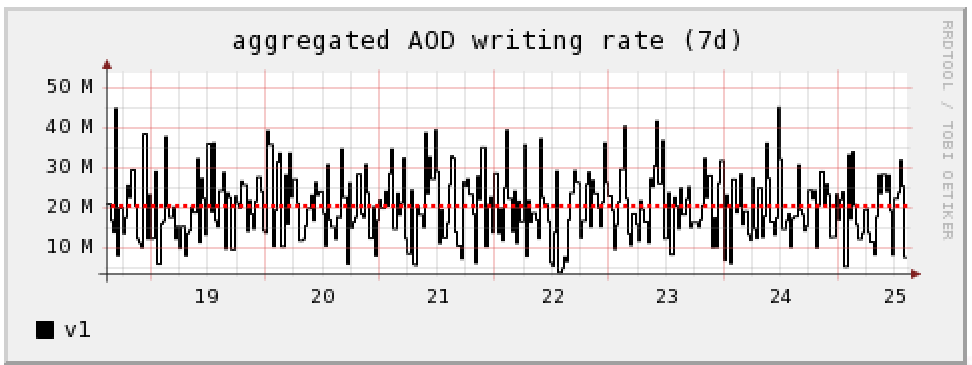
Tier-0 Monitoring (Examples)

- Aggregated rates for selected CASTOR transfers
  - Week of June 18-25, 2007
  - RAW (reading and writing), ESD (writing), merged AOD (writing)









---- Nominal rates

Overview of this Talk

- Introduction
  - ATLAS Dataflow
  - ATLAS Tier Organisation and Functions
  - Tier-0 Scale
- Tier-0 Internal Tests
- Tier-0 → Tier-1 Export Tests
- The Full Dress Rehearsal (FDR)
  - Scope
  - Preparatory Exercises (Stream Test)
- Plans for 2007/2008
- Summary and Conclusions

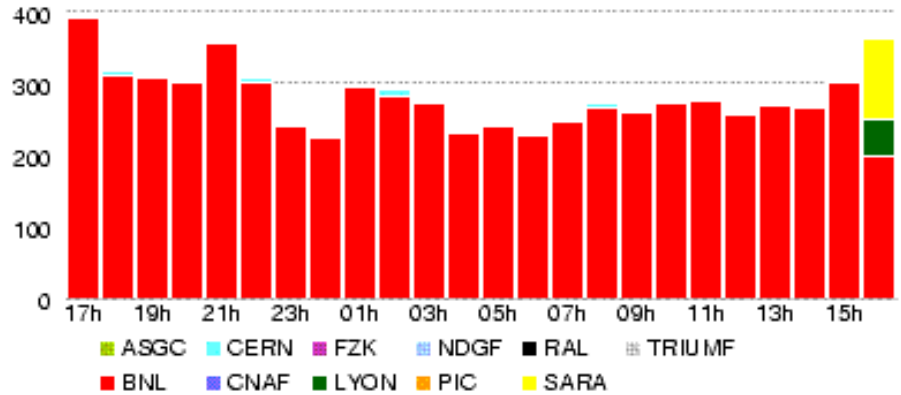
Tier-0 → Tier-1 Export Tests

- ATLAS Distributed Data Management (DDM) system is being developed at CERN
  - Underlying software: DQ2 (DQ = "Don Quijote")
  - Central DQ2 catalogues at CERN
  - External dependencies:
    - File transfer software: FTS, SRM (CERN)
    - Catalogues: LFC (CERN/LCG), LRC (US ATLAS), Globus RLS (Nordic Data Grid Facility)
- Export exercises are usually coupled to Tier-0 test series and activities
- Achievements during Jun 2006 Tier-0 test (selection)
  - Included all Tier-1 sites in the exercise from first day (except NDGF)
  - Included ~15 Tier-2 sites on LCG by the end of the second week
  - Maximum export rate ~700 MB/s, sustained over several hours
- Sep/Oct 2006 test
  - Problem: CMS data export running in parallel
    - Maximum achievable export rate only ~400 MB/s

Tier-0 → Tier-1 Export Tests

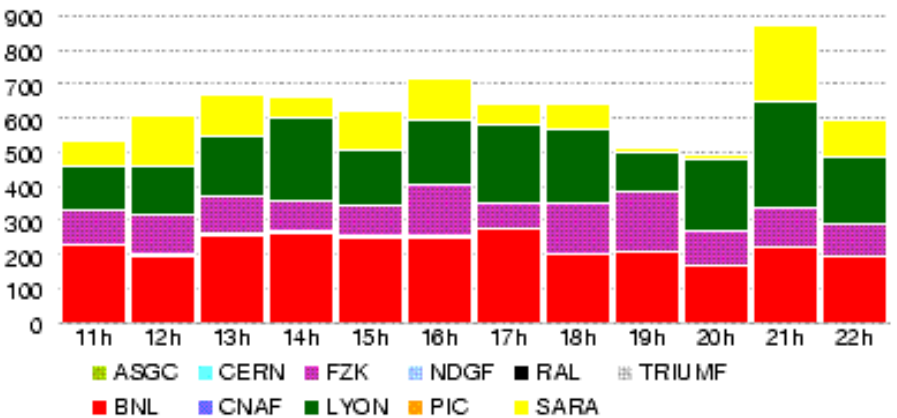
- DDM monitoring became responsibility of CERN/ARDA team
  - ARDA "dashboard": http://dashb-atlas-data.cern.ch/dashboard/request.py/site
  - Detailed, powerful monitoring tool
- During "CASTOR crisis" a systematic DDM debugging effort took place
  - Trying to understand, under controlled conditions:
    - Every single file transfer error
    - Complete data transfer flow across all layers (from DQ2 to the file systems)
    - Transfer throughput and its limitations
  - Led to elimination of some persistent problems
- CASTOR upgrades also substantially improved the Tier-0→Tier-1 export
- Recent achievements (selection):
  - Peak export rates of ~900 MB/s (but only for very short periods of time)
  - Daily peak export of ~600 MB/s
    - Sufficient to keep up with the nominal daily data rate (assuming 50k active seconds/day)
  - Stabilised, improved operation on many of the Tier-1s

Export Monitoring (Examples)

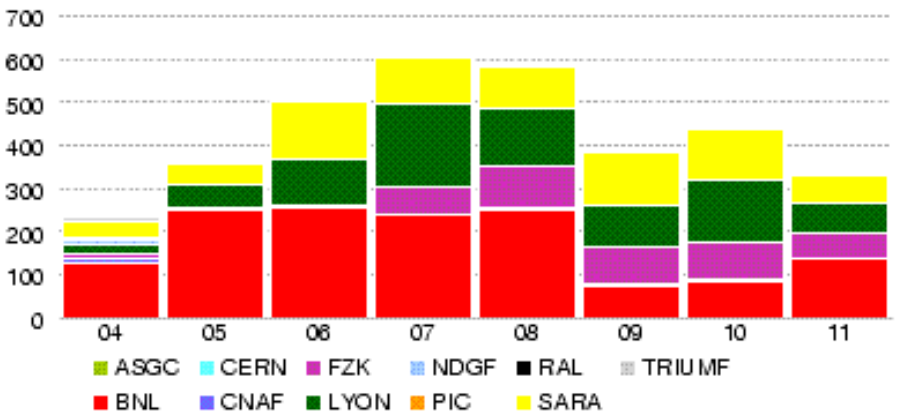


Export to BNL only, during 18 hours:

- From Jun 4, 17:00, to Jun 5, 15:00
- Stable running at 278 MB/s on average, with 98% efficiency



Export rates in MB/s on Jun 7, from 11:00 to 23:00



Daily export rates in MB/s, between Jun 4 and Jun 11

Overview of this Talk

- Introduction
  - ATLAS Dataflow
  - ATLAS Tier Organisation and Functions
  - Tier-0 Scale
- Tier-0 Internal Tests
- Tier-0 → Tier-1 Export Tests
- The Full Dress Rehearsal (FDR)
  - Scope
  - Preparatory Exercises (Stream Test)
- Plans for 2007/2008
- Summary and Conclusions

The "Full Dress Rehearsal" (FDR)

- Planned series of continuity tests of the complete offline chain SFOs → Tier-0 → Tier-1s → Tier-2s
  - "Mock data" fed into online output farm (SFO)
    - "Bytestream" input generated from simulated data, realistic file sizes
    - Realistic physics mix, realistic Trigger Tables
    - Organisation into O(6) trigger-based physics streams
  - Conditions database and in-file metadata access
  - Data Quality Monitoring and preprocessing steps (similar to Express Stream and calibration stream processing)
  - Real-time reconstruction at the Tier-0, producing ESD, AOD, TAG, ...
  - Export of RAW and reconstructed data to Tier-1s
    - Export of reconstructed data to Tier-2s
- Reprocessing test for Tier-1s
  - Reconstruction reprocessing from RAW, alternatively remaking AOD from ESD
- Test of Analysis Model
  - User analysis on Tier-2s

FDR Preparation: The Stream Test

- Done during first half of 2007
- Based on 18 pb^{-1}-equivalent of simulated data
  - Simplified 10^{33} trigger menu, including prescales
  - Data organised into trigger-based streams:
    - "**Inclusive streaming**" (5 streams)
      - » Events are written into any stream whose trigger selection they pass
    - "**Exclusive streaming**" (6 streams)
      - » Events are written only once
      - » Events passing more than one selection are written into dedicated "Overlap Stream"
    - Streams: jets, electrons, photons, muons, etmiss, (Overlap)
    - Luminosity variations implemented, some meta-data (e.g., "bad runs") added
  - Reconstruction run on the data
    - Using imperfect calibrations
    - Produced (among others) AOD, TAG
    - Populated TAG database
- Many unforeseen difficulties encountered
  - Data preparation and distribution; software readiness; trigger simulation; ...

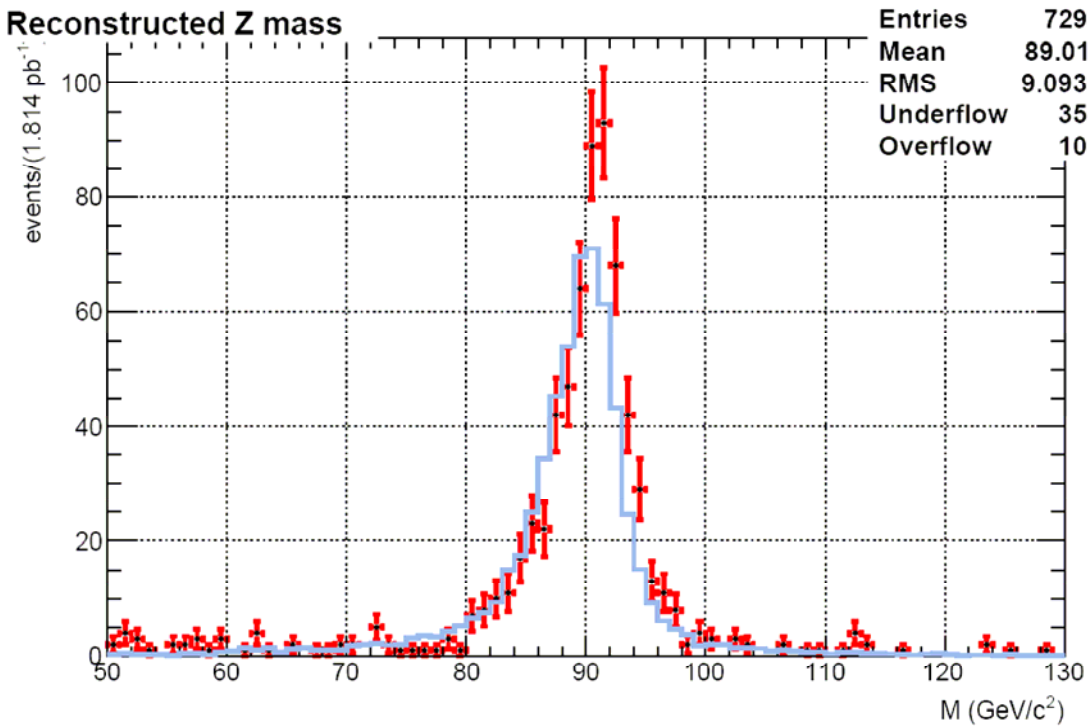
FDR Preparation: The Stream Test

- User analysis of Stream Test data has started
  - Results will be used to decide on final streaming model
  - Example plot from a recent meeting (Jul 5)
    http://indico.cern.ch/conferenceDisplay.py?confId=17090

Z^0 mass, reconstructed from $Z^0 \rightarrow e^+e^-$ events, selected from the Stream Test inclusive-electron sample

MC | 'data'
(normalized to **measured** cross section)

(LBL Group: A. Holloway *et al.*)



Overview of this Talk

- Introduction
  - ATLAS Dataflow
  - ATLAS Tier Organisation and Functions
  - Tier-0 Scale
- Tier-0 Internal Tests
- Tier-0 → Tier-1 Export Tests
- The Full Dress Rehearsal (FDR)
  - Scope
  - Preparatory Exercises (Stream Test)
- Plans for 2007/2008
- Summary and Conclusions

Plans for 2007/2008

- Continue autonomous Tier-0 internal and Tier-0 → Tier-1 export testing
  - Concentrating on throughput, improved functionality, stability of operation, achieving of required export rates
  - For Tier-0 internal processing: adding more realism w.r.t. used data and executables (available h/w resources permitting)
- Tier-0 will participate in FDR and cosmics data taking runs
  - E.g., "M4" cosmics run (Aug/Sep 2007)
    - Including data export to Tier-1s
    - Possibly first attempt of reprocessing at selected Tier-1s
  - During those periods the autonomous Tier-0 testing will be suspended
- FDR plans
  - First round in Autumn 2007
    - Based on one fill (10h) of emulated data
  - Second (final) round in Spring 2008
    - Using more recent simulation version
    - More statistics, richer physics mix, more complicated trigger menus

Overview of this Talk

- Introduction
  - ATLAS Dataflow
  - ATLAS Tier Organisation and Functions
  - Tier-0 Scale
- Tier-0 Internal Tests
- Tier-0 → Tier-1 Export Tests
- The Full Dress Rehearsal (FDR)
  - Scope
  - Preparatory Exercises (Stream Test)
- Plans for 2007/2008
- Summary and Conclusions

Summary and Conclusions

- Significant progress has been made in the last two years
  - The test programme carried out so far has turned out very fruitful
    - The parts of the Computing System covered in the tests so far are already in a very advanced state
  - In general, basic functionality seems to be in place
- Target for this year is to exercise the complete system at nominal rates (corresponding to 200 Hz data-taking rate), and for next year to reach even 1.5-2 times higher rates
  - To be able to cope with possible backlogs
- These are ambitious targets …
  - … and there is still a lot of work ahead of us
  - But we are confident that we are "almost there"
    - Don't expect fundamental show-stoppers any more