ATLAS & Service Challenges 3

SC3 workshop
CERN
June 13-15 2005

Gilbert Poulard (CERN PH-ATC)

ATLAS & SC3

- ATLAS is highly interested to participate to SC3
  - Continue to test and validate its Computing Model
  - Pursue the integration of
    - Production System (ProdSys)
    - ATLAS Data Management system (DonQuijote)
- Also interested in using COOL for the conditions data (calibration and alignment)

ATLAS & SC3

- Intends to participate to:
  - Preparation phase
  - Throughput phase
- Run a Tier-0 exercise
  - Reconstruction
  - ESD, AOD, TAG production
  - With distribution of data from Tier-0 to Tier-1's
- Test Distributed Monte Carlo production
  - With concentration of data in Tier-1's

ATLAS & SC3: preparation phase

- "gLite FTS" has just been released
  - We intend to test it asap
  - We will still keep an eye on CMS PhEDEX
- Need to understand the SRM issues
  - We said already that SRMCopy is important for us
- Plans (until July)
  1) Test CERN -> BNL transfer with "gLite FTS" (no Don Quijote)
  2) Test with FTS "within" DQ
     1) More file transfers; more sites
  3) More integration within DDM ready for July

ATLAS & SC3: preparation phase

- We see FTS as a service
- We want to use it (now) as it is, even if is fairly simple and even if it has less functionality than DQ
  - Because it uses GridFtp and SRM
    - And we don't need to handle the GridFtp and SRM errors
- We don't foresee major integration issues
  - Both DQ and FTS have "transfer queues"
  - DQ will ask FTS to do the transfer
  - Later on we will handle the "priority" issues
- We want to test all kind of transfers
  - Disk -> Disk; Tape -> Disk; Disk-> tape; Tape-> Tape
  - With all kind of data (Dummy; small files as AOD; large files)
- At BNL
  - Somebody will take care of the installation; maintenance and testing
- For testing we don't need too much hardware resources (1 machine is probably enough)

ATLAS and SC3

- In July
  - Scalability test
  - With "commissioning" (real) data if available
  - With "Rome Physics workshop" data
    - Small and large files (~100 MB; ~500 MB; ~1.8 GB)

ATLAS and SC3

- September
  - ATLAS release 11 (mid September)
    - Will include use of conditions data base and COOL
  - We intend to use COOL for several sub-detectors
    - Not clear how many sub-detectors will be ready
    - Not clear as well how we will use COOL
      - Central COOL database or COOL distributed database
  - Debug scaling for distributed conditions data access calibration/alignment, DDM, event data distribution and discovery
  - T0 exercise testing
  - A dedicated server is requested for the initial ATLAS COOL service
  - Issues on FroNtier are still under discussion and ATLAS is interested
  - Data can be thrown away
    -

ATLAS and SC3

- > Mid-October
  - Service phase: becomes a production facility
  - Scalability test
  - Using more data and more sites
  - Operations at SC3 scale producing and distributing useful data
  - New DDM system deployed and operating
  - Conduct distributed calibration/alignment scaling test
  - Progressively integrate new tier centers into DDM system.

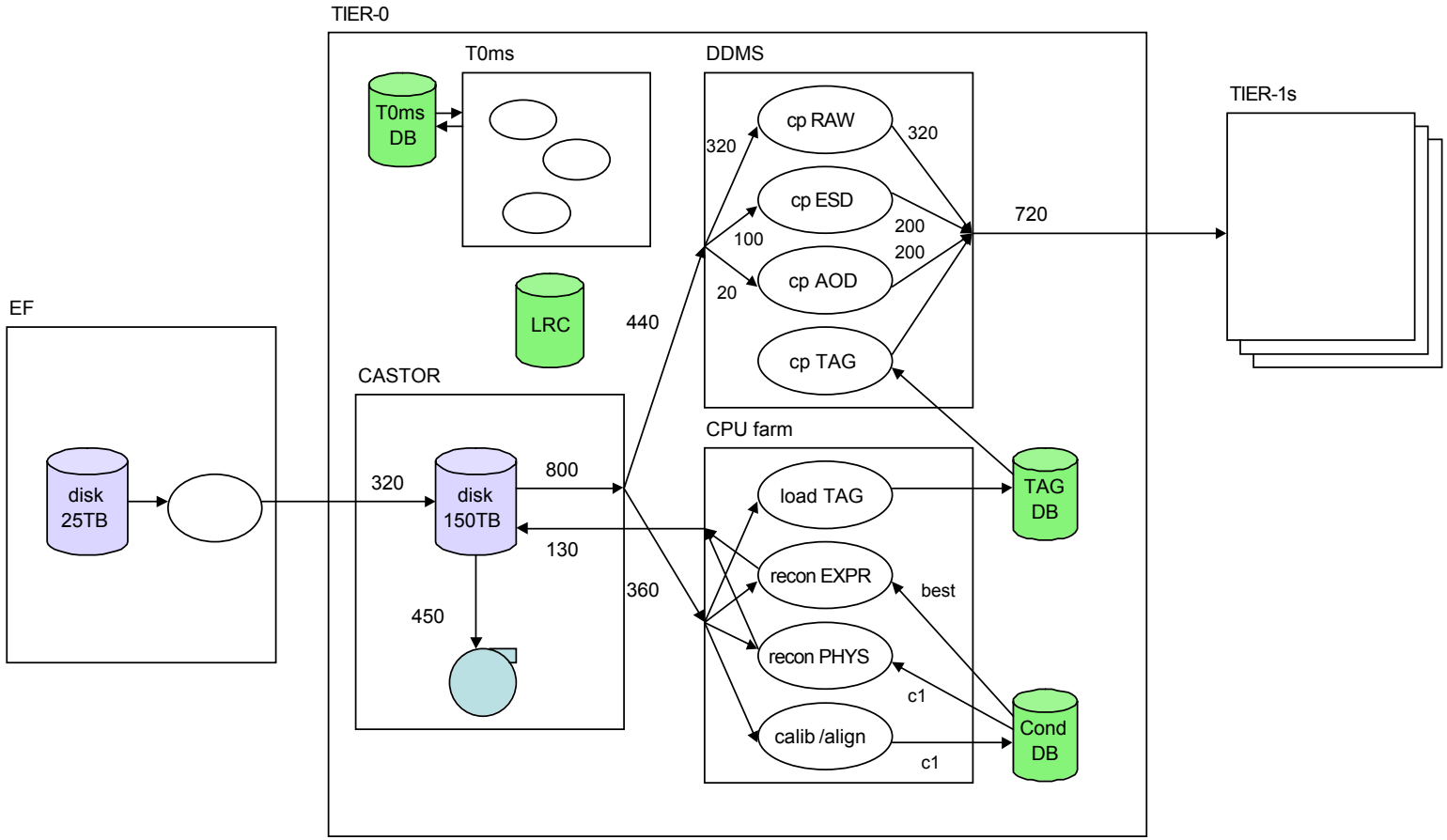
ATLAS and SC3

- > Mid-October; we intend to run:
  - Tier-0 exercise
    - Reconstruction at Tier-0 et production of ESD; AOD; Event collections
    - Data distributed from Tier-0 to Tier-1s then Tier-2s
  - Distributed Monte Carlo production
    - Data generated on Tier-2s, Tier-1s and are stored on Tier-1s for permanent storage
  - Use of conditions database will be part of the "game"
  - Reprocessing
    - Run at Tier-1s, "where the data is".
  - But this will be done in the last months of 2005
  - For DC3 we need to produce "few" 10 Million events
- We don't forget analysis!

Tier-0

- Responsibilities (from Computing Model)
  - calibration and alignment
  - first-pass ESD production
  - first pass AOD production
  - TAG production
  - archival of primary RAW and ESD/AOD/TAG
  - distribution of primary RAW and ESD/AOD/TAG

Tier-0


TIER-0
T0ms
DDMS
T0ms DB
cp RAW
cp ESD
cp AOD
cp TAG
320
320
100
200
200
20
720
TIER-1s
LRC
440
EF
CASTOR
CPU farm
disk 25TB
320
disk 150TB
800
130
360
450
load TAG
recon EXPR
recon PHYS
calib /align
best
c1
c1
TAG DB
Cond DB

Tier-0

- In from Event Filter
  - 320 MB/s
    - physics
    - calibration/alignment (45 MB/s)
    - express (6 MB/s)
    - pathological (<<<)
- Out to Tier-1s
  - 720 MB/s
    - RAW/ESD/AOD/TAG

Tier-0

- Tier-0 components
  - CASTOR Mass Storage System and catalog
  - CPU farm
  - conditions DB
  - TAG DB
  - T0 management system (+prod DB)
  - data management system

Tier-0: CASTOR

- CASTOR
  - input buffer (125 TB) = 5 days
  - output buffer (25 TB) = 2 days
  - write to tape at 440 MB/s
    - 320 RAW
    - 100 ESD
    - 20 AOD
    - 0.2 TAG
  - no recall from tape in normal operation

Tier-0: CPU

- CPU farm
  - recon of physics and express streams
  - calibration/alignment stream -> conditions DB
  - loading of TAG DB with TAG files

Tier-0: Data Management System

- Data Management System
  - will use (dedicated) ATLAS DDMS
  - is filled with requests to replicate datasets out of T0 to T1s
  - expects some catalogs to be filled with information about the files on CASTOR and the datasets

Tier-0: Management System

- T0 management system
  - orchestrates jobs on CPU farm and transfers out to T1s
    - will fill the DDMS catalogs
    - will launch jobs and monitor them
  - very similar to ATLAS ProdSys
    - will build upon ProdSys or use variant of
  - important difference is that system is data-driven rather than task-driven
  - state is persistified in production DB

Tier-0: Reconstruction

- First pass ESD production
  - starts after "green light" on conditions data
  - CPU farm matches EF data rate
    - 15kSI2Ks x 200 events/s = 3000 kSI2K
  - expect to take data only 50k secs/day (60%)
    - data taking periods are bursty
    - some margin to process backlog
  - excess capacity is not lost
  - 2GB in, 625 MB out

Tier-0: AOD production

- First pass AOD production
  - light process in CPU
  - streamed into 10 maximally balanced streams
    - compromise
    - 10 x more and 10 x smaller files
  - Computing Model proposes to split AOD from ESD
    - to process 50 AODs into a bigger ESD
    - adds another 100 MB/s to CASTOR reading
    - grouping (50 x) same for all streams
  - alternative is to combine ESD and AOD and group AOD per stream in separate jobs
    - adds only 20 MB/s to CASOR reading
    - allows different group sizes per stream
    - does not increase to nr of jobs nor the complexity

Tier-0: TAG production

- First pass TAG production
  - combined with AOD production
  - for robustness no direct upload to TAG DB
    - first as POOL explicit collections
    - upload in DB is separate asynchronous process
    - many small TAG files can be tar'ed before archival on CASTOR

ATLAS & SC3: Summary

- Now-July: Preparation phase
  - Test of FTS ("gLite-SRM")
  - Integration of FTS with DDM
- July: Scalability tests (commissioning data; Rome Physics workshop data)
- September: test of new components and preparation for real use of the service
  - Intensive debugging of COOL and DDM
  - Prepare for "scalability" running
- Mid-October
  - Use of the Service
  - Scalability tests of all components (DDM)
  - Production of real data (MonteCarlo; Tier-0; …)
- Later
  - "continuous" production mode (data needed for ATLAS DC3)
  - Re-processing
  - Analysis