Argonne NATIONAL LABORATORY

... for a brighter future

U.S. Department of Energy

UChicago ▶ Argonne LLC

Office of Science U.S. DEPARTMENT OF ENERGY

A U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

ATLAS experience, plans and resource request update for T0 and T1

WLCG Collaboration Workshop (Tier0/Tier1/Tier2)

CERN, Geneva, Switzerland

January 22-26, 2007

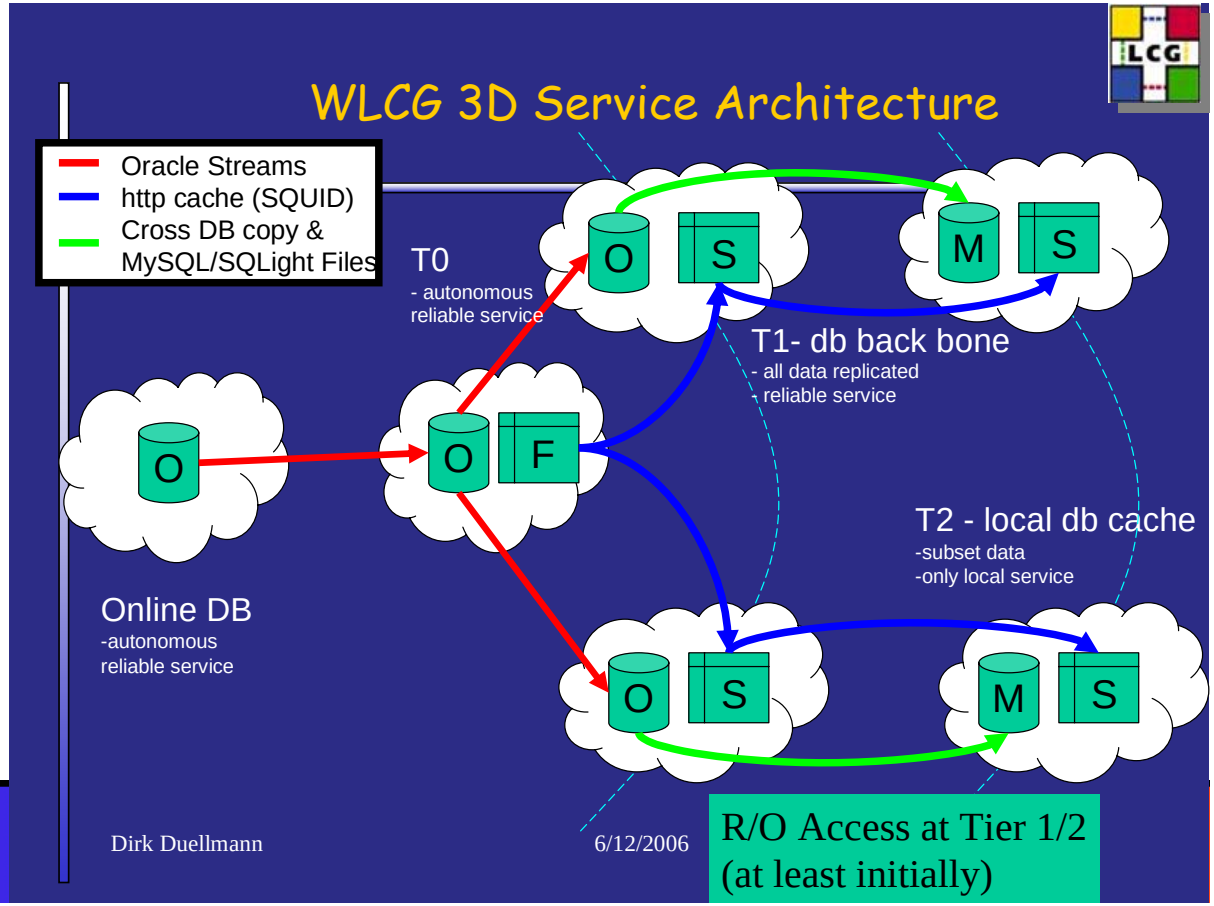
Alexandre Vaniachine (Argonne)

ATLAS Experience

- Main focus is on Oracle streaming to T1
  - For fast replication
- Detailed report by Gancho Dimitrov and Florbela Viegas at the 3D BOF Session on Monday:

http://indico.cern.ch/contributionDisplay.py?contribId=60&sessionId=23&confId=3738

- For 3D architecture extension beyond T1:
  - ATLAS choosing distribution of SQLite files to T2 (via T1)
    - *For slow replication*



Oracle Scalability Limit

- Most of current efforts in 3D concentrate on streams throughput
  - But the main reason to replicate to T1 is to run massive (second-pass) reconstruction (a lag of about six months)
- Can Oracle handle reconstruction job queries each 6 seconds?
- What is the hardware limit?
  - for the current multi-process mode is expected to be limited by total memory [Gancho]
    - *Do we have to switch to a multi-threaded mode for Oracle?*
      - Should be easy to do but less secure [Florbela]
    - *or just double the server memory at all T1s?*
- To address these issues ATLAS is working on Oracle scalability test with Athena COOL job
  - David, Sanjay, DDM operations team et al.

- Alternative: deploy FroNTier at T2
  - Needs extra manpower to resolve cache consistency issue

Oracle Monitoring Experience and Request

- ATLAS evaluated LCG Dashboard (by Julia Andreeva et al.) for use in 3D operations
  - Currently monitors Oracle streams data (by Zbigniew Baranowski)
- Conclusion
  - a very useful tool in support for operations

- ATLAS Request
  - Extend the LCG Dashboard monitoring of 3D operations:
    - *number of Oracle sessions*
    - *CPU load*
    - *Disk space:*
      - allocated for ATLAS use
      - actually used by ATLAS

ATLAS Experience with SQLite

- ATLAS uses database replicas in SQLite files
- SQLite is used for MC production (simulation/reconstruction) where the required CondDB content is small and can be determined in advance
  - SQLite cannot be used for real data, where data volumes will be much larger, and will change rapidly
    - *For real data we will need Oracle and/or FroNTier*
- For automated distribution of the Database Release files ATLAS uses its Distributed Data Management (DDM) operations team infrastructure

SQLite Database Access Pattern

- A side-effect of using DDM is that the SQLite file arrives to the site in the data area (the same area where the event data files are placed)

- But there is a difference in the Event Data file access pattern and the Database Release file access pattern:
  - Typically each job processes its own event data file or a portion of the event data file. Thus the event data files require simultaneous read-only access by one to fifty jobs (mostly by one job)

- In contrast most of the jobs access the same Database Release file
  - Thus, the database release file need read-only access by thousands of jobs running at the site

ATLAS Requirement for SRM

- It turns out that the current SRM release installed on LCG is capable of handling the requirements of Event Data file but does not scale well beyond that

- ATLAS needs the SRM to be able to handle the Database Release file requirements
  - We are looking forward to the new SRM release deployment
    - *Scalability for DB will be tested by the DDM operations team*

- Alternative 1: do not use SRM, use direct local file access
  - Needs extra manpower to implement

- Alternative 2: deploy FroNTier at T2
  - Needs policies to be implemented and strictly followed
    - *No easy way to request the 'very latest' data*

ATLAS Requirements for FroNTier

- We can deal with FroNTier cache consistency largely by policies
  - e,.g only accessing the data by frozen CondDB tag
    - *so the cache result for this query will not go stale)*
  - and/or by having the cached data expire after a certain time
    - *as shown by CMS in their FroNTier presentation this week*
      - which is quite encouraging

- If ATLAS choose the FroNTier alternative the major issue will be to get the T2 sites to deploy squid caches
  - in principle, this is straightforward
    - *as shown by the CMS deployment at ~20 T2 sites now*
      - but of course requires some hardware resources and some support at the T2s

ATLAS Schedule

- Preparation for ATLAS CDC in Spring:
  - Install SRM 2.2 and commission: March/mid-April
  - **3D services for CDC production: February**
- Readiness for the final phase of ATLAS FDR in September-October:
  - 3D services stress testing: June/July
  - FDR initial phase: July/August
  - **Max-out T1 3D capacities: August/September**
    - *most T1s now have only two nodes for 3D*
    - *ATLAS DB will need several TB per year*
  - 3D Production running: October-December

Request to max out T1 3D capacities

- We have to max out (fully utilize) the available resources:
- We expect to be memory bound, not CPU bound
  - Since mostly jobs request the same data

Max out request:

- There are six Oracle licenses per T1 (for ATLAS use)
  - If you have not used all your Oracle licenses – use it
    - *Deploy the third node*
- If your Oracle nodes do not have 4GB of memory
  - Upgrade to 4 GB of memory
- If your Oracle nodes do not have 64-bit Linux
  - Upgrade to 64-bit Linux
    - *In preparation for memory upgrade beyond 4GB*
- Since this is a rather modest upgrade request, the upgrade can be accomplished by August/September

ATLAS Disk Schedule

- Gancho and Florbela made detailed estimates for ATLAS running
  - COOL storage 0.2 TB / month
  - TAGS storage 1.5 TB / month

- We need to ramp T1 capacities to these numbers
  - The priority is storage for COOL
    - *ATLAS TAGS can be stored outside of Oracle, in ROOT files*

Ramping up to ATLAS running conditions:

- ATLAS will probably have maximum 6 months data taking per year
- Also, we may choose not to run at 200 Hz on average
  - Thus, a data-taking efficiency of 50% is assumed
    - *which is already included in the COOL numbers*
- at average 100Hz, 6 months running per year:
  - TAGS storage 4.5TB per year
  - COOL storage 0.6 TB per year

ATLAS Disk Storage Request

- Current capacities for ATLAS (collected by Gancho and Florbela)

T1 site	Disks (TB)	Nodes	CPU/ node	Memory (GB)	64-bit	Comment
IN2P3	0.5	2	4		yes	Shared with LHCb
CNAF	1.0	2	4			
RAL	0.4	2	2			
Gridka	0.6	2	4			
BNL	0.4	2	4			ATLAS site
ASGC	1.6	2	4		yes	Expandable to 10TB
TRIUMF	3.1	2	2		yes	
NDGF						3D Phase II T1 site
PIC						3D Phase II T1 site
SARA						3D Phase II T1 site

- T1 sites with Oracle disk storage less than 1 TB are requested to double their capacities by December

Summary of ATLAS Request for Resources

Readiness for Calibration Data Challenge (by March):

- Upgrade 3D services to production level

Readiness for Final Dress Rehearsal:

- Max out request **(**by August/September)**:**
  - Deploy the third node for ATLAS (for sites without third node)
  - Upgrade to 4 GB of memory (for sites with less than 4 GB of memory)
  - Upgrade to 64-bit Linux (for sites without 64-bit Linux)
    - In preparation for memory upgrade beyond 4GB

Readiness for Real Data:

- Disk space upgrade **(**by December)
  - T1 sites with Oracle disk storage less than 1 TB are requested to double their capacities