ATLAS reconstruction jobs & conditions DB access

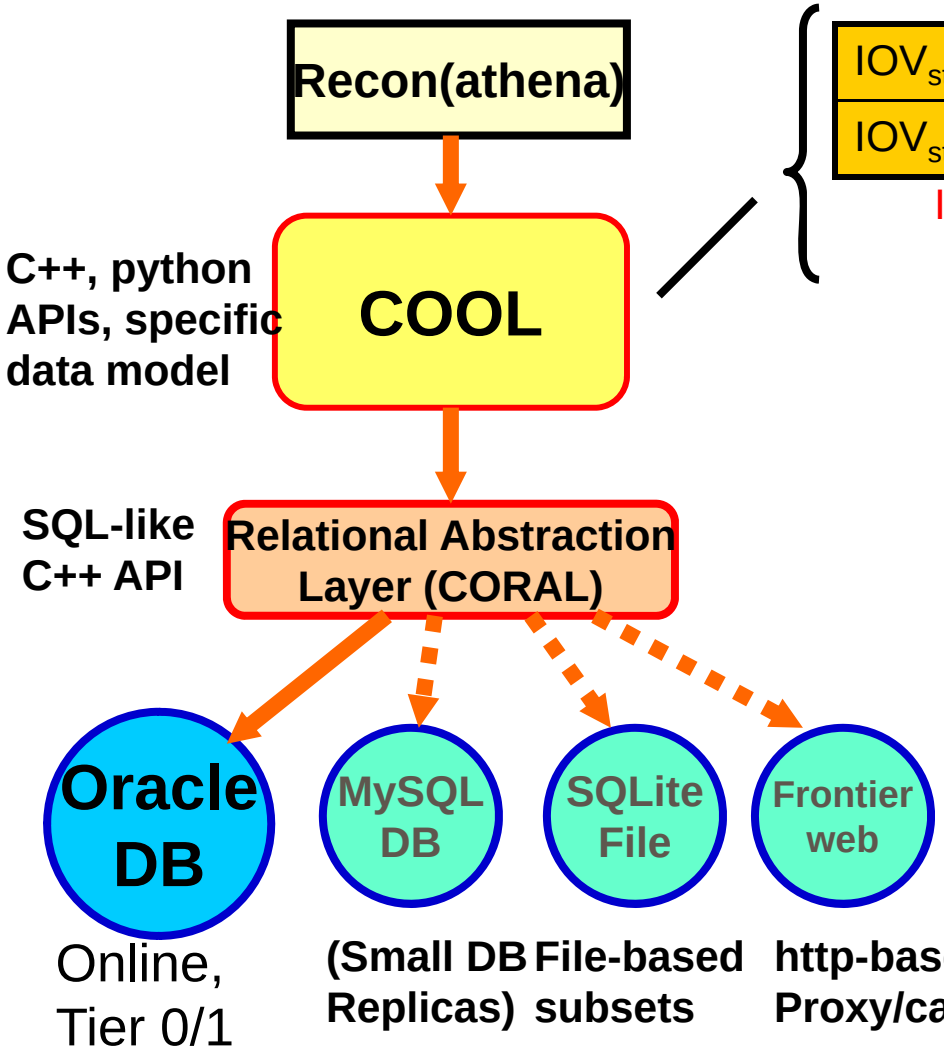
Richard Hawkings (CERN)

Distributed database operations workshop, CERN, 11/11/08

- Conditions database basic concepts
  - Types of data …
- Access in Athena
  - Athena reconstruction jobs
  - Conditions data access
  - Scaling and problems
    - Possible improvements

Conditions DB - basic concepts

- COOL-based database architecture - data with an **interval of validity (IOV)**



IOV_{start}	IOV_{stop}	channel1	(tag1)	payload1
IOV_{start}	IOV_{stop}	channel2	(tag2)	payload2

Indexed

- COOL IOV (63 bit) can be interpreted as:
  - Absolute timestamp range (e.g. for DCS)
  - Run/Lumiblock range (1 LB ~ 1 minute)
- COOL payload defined per 'folder'
  - Ttuple of simple types ≡ 1 DB table row
    - Can also be a reference to external data
  - Use channels for multiple instances of data in 1 folder (e.g. 1000 HV channels)
  - COOL tags allow multiple data versions
- COOL folders organised in a hierarchy
  - Tags in the hierarchy referenced from a global 'top-level' tag

Overview - types of conditions data

- Detector calibration, configuration and status data stored directly in COOL
  - Volume dominated by large CLOB objects ~O(20 MB) for muon calibration
  - Also many smaller objects
    - Typically valid for at least one run - update frequency ~ 1 per day or less
- Detector control system (DCS/'Slow controls') data stored directly in COOL
  - Expect O(1-2 GB/day) DCS data, distributed over <50 COOL folders
    - IOV valid from a few seconds to a few hours - depends on stability, filtering etc
    - Data is a subset of that available in PVSS Oracle - what is needed for Athena recon
    - Fit 1-2 GB/day size constraints (set by storage @ Tier-1 and Oracle streams bandwidth)
    - This dominates the data volume in Oracle
- Calibration data stored in referenced POOL files (COOL+POOL)
  - Volume in COOL is small (only references to POOL files via GUID/offset)
- All data volumes are approximate
  - Data volume is still ramping up as subdetectors increase storage of conditions data and use in reconstruction jobs
    - Requested space for 1 TB new data (table+index) in 2009, includes safety factor

Athena reconstruction job - what is it?

- Takes a RAW data file written by online, reconstructs to ESD and AOD files used by physics analysis
  - Characteristics depend on type of data taking (cosmics, colliding beams)
  - 1 RAW file is O(few 100 MB)-O(few GB) - takes a few hours CPU to reconstruct
    - E.g. file with 1000 1.6 MB events, ~15 seconds/event - 4 hour job
    - Trade-off between file length (storage likes few big files) and recon job wall-time
  - File contains event data with a timespan of ~30 seconds to several minutes
    - DAQ system divides data into 'data streams', which fill at rates from a few Hz to 50 Hz
    - Hope for 'balanced' streams in collision data taking
  - One 'run' can last several hours (divided into ~1 min lumiblocks), 1000s files/run
- What conditions data is required by reconstruction job?
  - Detector calibration and configuration data, indexed by run/LB
    - But typically changes only once per run - only one set required per job
  - DCS data, indexed by timestamp
    - Potentially changes every second … many sets of data per job
    - Athena can bulk-read a 10-minute 'chunk' of data in one go, cache it for later use - avoid recontacting database every few events

Athena job database access

- Configuration for October 2008 cosmics data processing
  - Data from around 70 COOL folders read, < 10 with DCS-type data (more coming)
  - Data organised in about 12 COOL **schema** (one per detector)
    - In COOL, each schema requires a separate Oracle **session** to read it
- Phases of data access within Athena
  - Job starts… reading configuraton, loading shared libraries etc
  - After ~1 minute: IOVDbSvc initialise
    - Go through configured list of COOL folders, check existence, read meta-data, tags
    - Sessions opened ‘on demand’, not ordered by schema, so gradual open of more sessions up to max of 12; then all are **closed** at end of initialise
  - Athena intialisation continues, then start first event processing
    - For each folder, read the actual data requested (COOL browseObjects calls)
    - Again happens ‘on demand’ (on callbacks and explicit data requests) - not ordered by schema, gradual increase of sessions up to max of 12, over a few minutes
    - All sessions/connections closed at end of first event
  - Can take ~ 10 minutes wall-clock time from job start to end of first event
    - Two ‘bursts’ of connections/sessions up to maximum of ~12, only one active at a time
  - Normally, no more database activity for rest of job… but Athena will transparently reopen connections if required (e.g. a file with a long timespan, multi-run file)

Scaling and problems

- Expect O(2000) jobs running in parallel at Tier-0, O(1000) at Tier-1s
  - Critical parameter is ‘job start’ rate, not nunber of parallel jobs
    - E.g. Tier-0 throttles to 10-20 jobs launched/minute (8/min required for 2000 4 hour jobs)
    - Harder to do in Tier-1 production environment - want to fill an empty farm, fluctuating load due to other production work - can we introduce load-based throttling?
  - Indirect measurements indicate reading O(40 MB) from database for each job
- Problems with current Athena access scheme
  - Originally designed for ‘on-demand’ access, connection kept throughout job
    - Various ‘sticking plaster’ improvements (close connections at end of initialise and event 1)
    - Trade off between open-close cycles and having idle connections open
- Possible improvements
  - Rewrite Athena IOVDbSvc to order and group data requests by schema
    - Keep only one connection open at a time, grab all needed data ‘speculatively’ and cache it locally on client - major reimplementation of IOVDbSvc needed
    - Take opportunity to also ‘align’ query IOVs, to improve performance with Frontier
  - Stuck with 12 connections per job unless COOL can share sessions across schema
    - Timescale for implementation with OracleAccess? / or wait for CORAL server?