ADC Facilities Jamboree 2014

Filtrato e scolato

S. A. Tupputi

DDM restyling

- Rucio
  - in production (whisper it quietly) since 1/12
  - Support for protocols alternative to SRM for transfers and storage management
  - Gradually space tokens will be dropped
- FAX
  - All data accessible from any location
    - Jobs can run wherever there's free cpu (overflow)
  - WAN modes
    - Failover (SE fails and another one is contacted)
      - Default in PanDA since 3/2014
    - Overflow (JEDI decides to contact another SE)
    - Explicit overflow (user explicitly decides for another SE)
      - wansinklimit
      - wansourcelimit

Prodsys-2

- Powered by JEDI/PanDA as in production
- Resource (dynamical) scheduling optimization
- Merging at T2s minimizes data traffic
- Monitoring tools and rucio-integrated

Event Service

- Job granularity from file to events: don't stage in whole input files
  - Panda/JEDI, AthenaMP, Data Federation, Object Store
- To be deployed firstly for simulation (Geant4):
  - Validation starting on Amazon before going into production
- Opportunistic approach: quick reassignment of WN for role changes/draining
  - E.g. prod/analy, SCORE/MCORE
- Potential benefits for analysis
  - Dynamic resource management and allocation for particular cases
  - Output asynchronously streamd off WN to achieve prompter retrieval
  - Asynchrony can be fully exploited to ease up I/O intensive tasks

Multicores

- https://twiki.cern.ch/twiki/bin/view/LCG/DeployMultiCore#Batch_system_related_information
- 23% of prod wct has gone multicore
- Issues
  - Non-steady submission
  - Jobs need to be longer for making any sense
    - Merging is SCORE and weighs over shorter tasks
  - Requests for large mem empties slots
- Switching approach: parameters to be passed to batch system rather than relying on queue settings
  - GlueCEPolicyMaxCPUTime == ncores x maxtime
  - GlueCEPolicyMaxWallClockTime == maxtime
  - GlueHostMainMemoryRAMSize == maxrss
  - GlueHostMainMemoryVirtualSize == maxrss + maxswap
  - For mem cgroups are being enabled (link)
    - HTCondor, SLURM, UGE

Resource requirements: going dynamically

- ATLAS VO Card
- Thanks to Rucio/Prodsys2 each job type will be sampled and the report registered in Panda
    - Rss and swap profile
    - Cpu usage profile
  - Tasks will have resource requirements bettere defined
- Multicore: need to pass job reqs to batch system
  - OK for ARC-CE
  - CREAM-CE need Blah
  - OSG-CE to investigate
- Pilot and multiple site queues don't let dynamic brokerage
  - AGIS: single site queue with site (global) properties
  - Task queue dynamically created by PanDA/AGIS
  - PF sends pilots with ad hoc resource configuration to CE/BS
- For ARC-CE possible direct payload submission model (skipped)
  - Data input/output done outside batch job by direct interaction with PanDA
  - Per-single-job customization
  - Proposal to take over CREAM-CEs
- In general: towards a model where no longer panda queues job types are the handles: (few) parameters to be passed to BS is the preference

Resource requirements: memory, disk, network, batch system

- Standard WN (2GB/core) may get killed by large memory jobs
  - Limit total rss and swap usage
    - Cgroups (SLURM, Condor)
- Input/output files to be reduce to max 5 GB to reduce merge jobs
  - CPU-intensive jobs will go mcore, disk will benefit
- 1 Gb/s on fat nodes (32-64 cores) is a bottleneck
  - Go for 10 or 40 Gb/s
- Improve SE-WN bandwidth
- Drop Maui, Torque (no cgroups!); recommended:
  - SLURM, HTCondor, SGE, …
  - Universal queues need be supported by BS (mostly true)
  - No hard partitioning for SCORE/MCORE, analy/prod

Storage stuff

- Tool for storage consistency developed at MTW2
- Rucio dumps: verify storage contents with Rucio db
  - RSE, dataset locks, DIDs, replicas,
- LOCALGROUPDISK management tool (US)
  - By disk usage by user (<= 3 TB or 7 replicas)
  - By dataset (1 year)

Storage protocol overview

- Specetokens to be dropped (short)
- Move to xrootd/https for upload/download (short)
- http/WebDAv for deletion (short)
- Decommission SRM from non-tape (medium)
- Move to xrootd for all files IO (medium) and decommission other protocols
- Xrootd/WebDAV for 3rd party transfers (instead of gridFTP) (medium/long)
- Consolidate davix and evaluate davix/xrootd for directIO (long)
  - Drop one of the two or keep both

Webdav, davix

- Analysis jobs
  - aria2c to fetch source code and libs
  - Davix for ROOT file access
  - List in agis
  - Used by HC
- Debug/tuning work in progress to raise webdav/davix to I/O standards of protocols already consolidated

FTS

- 3 is the magic number of instances:
  - CERN, RAL, BNL
- Working towards a full-range (server) failover functionality
  - Working for job submission
  - Not working for jobs already submitted
  - Failover configuration for IT cloud:
    - BNL – CERN – RAL
- Centrally configured

Run2 (I)

- Network improvements provide full mesh of sites
  - Many T2s provide T1-level services of computing, storage and WAN
  - SE/CE are no longer two different planets
  - T1,T2,T3 classification is obsolete
- ATLAS Storage Pool
  - TAPE:
  - STABLE disk-storage: T1 + reliable T2
    - Custodial & primary; prod tasks; prod final outputs storage
    - Space doubled!
    - Less need for data migration
  - UNSTABLE disk-storage: less reliable T2
    - Secondary data; not relied upon as transfer sources
  - VOLATILE disk-storage: unreliable T2, T3, opportunistic
    - LOCALGROUPDISK, Rucio cache, …

Run2 (II)

- Prod/analysis 90/10 %
  - Derivation analysis being moved to groups
- Merging to be reduced
- Bigger output by jobs
  - Tape
  - Less overall transfer latencies
- T2 classification: ASAP
  - Analysis tests do all relevant checks on CE/SE availability
    - STABLE (>90%); UNSTABLE; VOLATILE (<80%)
    - ICB will be reported on ASAP outcomes and funding agencies informed if it's the case
  - Proposals: link1, link2

Run 2 : new concepts

- Flat hierarchy
  - Rucio supports datasets replicas to be distributed over many sites
  - Primary replicas will be consolidated
  - Secondary replicas will stay distributed
  - Jobs brokering handles
    - Data proximity
    - Tranfers cost matrix
    - Dynamic evaluation of transfer times (recent story of past activities)
- Global cloud
  - Tasks not assigned to specific site: a global task
  - Less task to manage, globally distributed
    - Better prioritization
  - Sites will be able to dynamically specialize for jobs with given features based on their recent history