Multicore Resources

- Rod Walker, LMU Munich, 28th Jan 2014

Motivation
ATLAS Request
Resource requirements
Scheduling

Motivation

- AthenaMP needed for RAM saving reasons for high-mu reconstruction
  - P1 and grid for real data reprocessing, but also MC pile/reco
  - to be tested in DC14
- Not absolutely needed for G4 but ..
  - still save RAM, maybe used for something else
    - e.g. HI RAM queues usually leave cores idle
  - HPC needs short(for backfill), whole-node jobs
  - fewer files(DDM), no merging (for normal length jobs)
- DC14 can usefully exercise multicore using G4
  - deploy at sites, panda config, monitoring
  - gain experience, for sites and ATLAS ProdSys

The Request

- Sites should provide multicore resources
  - dynamic provisioning strongly preferred
    - 8 core slots freed only when jobs in queue, otherwise used for serial jobs
    - all Batch systems support this
      - recipes to be shared via mcore TF
      - https://twiki.cern.ch/twiki/bin/viewauth/AtlasComputing/AtlasMulticore
  - being realistic, not all sites will be ready

Resoure Requirements

- Currently have ~10-15% available
  - cannot grow T1 MCORE share to 100%
    - special serial role: high-mem, merge
  - will need larger T2 contribution
- Need ~30% resources as MCORE in April
  - for DC14 pile/reco
- Will keep MCORE full of G4, and submit rest as serial jobs.

Proposal

- Use existing MCORE T2s, plus new volunteers
  - increase the MCORE share, by stopping/capping serial queue
- takes pressure off other sites
  - orderly testing of MCORE capability
  - Autumn deadline is then ok for ATLAS

Scheduling

- ATLAS consider this a site responsibility, as per classic HPC site
  - no plan to fill parallel pilot with serial jobs
- Open source, free, supported batch systems that support parallel jobs and schedule well
  - throttle draining, backfill
  - maybe TF concludes Maui not capable/optimal
- Needs accurate walltime from jobs
  - not helpful when job finishes much sooner than expected either
  - mix of job lengths, including short
- Panda pull model not well suited
  - maybe can add finer-grained queues – with JEDI
    - for now, can add short analysis queue
  - ND+few sites receive pre-loaded pilots with per job granularity requirements
    - needs ARC CE, shared-FS, aCT and some ambition

Accounting

- 100% more at T2 – needs to be recorded and displayed in portal correctly
- Who pays for empty cores?
  - Accounting is hard
    - sites running hi mem jobs, not filling cores
      - but vital service
    - probably only possible with whole node accounting(a la cloud)
  - I think site pays, and works to minimize
    - good scheduling, mix of accurate walltime jobs, always 8 core queued (from any VO)
      - needs help from users
  - if VO pays, e.g. CMS backfill, then want very long jobs to minimize loss
- Either:-
  - site does it all = HPC
  - user does it all = cloud
- WLCG stuck in the middle - seems need to move one way or the other

Conclusion

- No big pressure on new sites to deliver before April
  - but please really plan now
- Move working sites to mostly mcore