GRIF: site report

CREAM-CE/HTCondor site

What's GRIF ?

- A distributed grid site mainly for LHC experiences (started 2005)
  - 6 labs around Paris
  - ~ 7PB
  - ~ 15k logical core
- Goal
  - Be seen and used as a unique site
  - Create 1 tech. Team distributed on labs

HTCondor @GRIF

- 3 sub-sites leads the migration to HTCondor
  - Due to scalability issue on some site
  - Better multicore support
  - Leak of support on some historical component (maui)
- Migration started on July 2014
  - One ARC-CE/HTCondor for multicore support
  - 2 other sub-site chose to keep CREAM-CE and HTCondor as batch system (January 2015)

Why CREAM-CE ?

- We support a lot of differents VO
  - Not want to do some validation stuff for VO frontend
- We know the CREAM-CE so easiest for us to understand where the problem is
- We were happy with CREAM-CE

CREAM-CE/HTCondor

- Stuff are implemented on CREAM-CE since a long time
  - 95% of the job was done
  - But since now no sites were use it
    - Some stuff were broken
    - Some stuff missed
  - But a shell script can be put by site between CREAM-CE and HTCondor schedd
    - /usr/libexec/condor_local_submit_attributes.sh

Queue vs ClassAd

- CREAM-CE is based on Queue
  - But it use queue only as a endpoint
    - No need to linked it with anything behind
- What we want to do with queue
  - Mainly apply some policy to a queue (multicore / long-time job/ short-time job / …)
    - Put a ClassAd to specify which queue was used

Scheduling policy

- Main idea of our grid site is
  - You have a priority to access to X% of our cluster
  - But you can have more if available
  - And we ensure your job will run up to 72H
- This is what you can do with
  - Dynamic quota (based on accounting_group)
  - Accept_surplus & autoregroup
  - Preempting should not be enabled

Hierarchic Fair Share

- A VO request since a long time
  - Could you have 50% of my fair share dedicated to analysis
  - And 50% to production
- On CREAM-CE side we create one queue per activities
  - And put the right accounting_group information on submit config

Group assignment

- Key point of scheduling policy
  - No way provide by CREAM-CE out-of-the-box
- But shell script have access to user proxy
  - So we can compute vo name and role
  - Associate it to a group and or subgroup
- Shell script also have access to CREAM queue name
  - So we can use it for subgroup instead of role

Multicore support

- Some VO want to use multicore job
  - Based on the agreement that 1 multicore job will take 8 cores
- We just put all our slot as partitionable
  - But hardest to monitor the effective usage of our cluster as HTCondor provide a job based view

Single / Multicore

- For fair share we create a new subgroup ‘multicore’
  - With it s own quota
- To avoid infinite idle job we active DEFRAG
  - Based on the idea that drain stop when 8 core is available (thx Andrew from RAL)
  - But no dynamic configuration modification (DrainBoss)

Partitionable slot and BDII

- A BDII is a information system database
  - Regroup all information about your site
  - Based on a script that run on your CREAM-CE
- But as CREAM-CE/HTCondor BDII plugin not supported partitionable slot
  - We re-write it in python with HTCondor python bindings

Accounting in HTCondor

- Based on work done @RAL
- We convert HTCondor history to PBS accounting log
  - And push it on national accounting as a classic CREAM-CE / Torque cluster

1 year production

- No big issue
- No regrets
- But…

Job HELD

- Some job go HELD
  - CREAM-CE submit the job but user proxy is not available on CREAM sandbox
  - It was a pure CREAM-CE issue
  - Was very easy to debug
    - HELD message was clear (file missing)

Accountantnew.log corruption

- Just trigged last week
- When CREAM-CE is overloaded, shell script raise « Java out of memory .. » exception
  - Lot of space on error message
  - But not catch by our script that pass the message as a group name
  - On restart Negiciator complains that accountantnew.log is corrupt

Future plans

- HTCondor CE
  - CREAM-CE is now our bottleneck
  - 1 middleware
- Create 1 GRIF HTCondor pool
  - 1 CE (submitter) per site
  - H.A Central Manager
  - Each worker node available for every CE
    - Probably add some ClassAd (WorkerLocation) to allow Ranking