Workload management landscape

GDB meeting at 

8th March 2017

Maarten Litmaath, CERN

v1.0

Table of contents

Plus anything that I managed to miss …

Current stats in the BDII

CE type	Instances
ARC	88
CREAM	352
GLOBUS	15
HTCONDOR	40

Batch system	Instances
HTCONDOR	86
LOADLEVELER	2
LSF	52
OGS/GE	2
PBS	88
PBSPRO	1
SGE	40
SLURM	57
TORQUE	169

CE instances per batch system

	ARC	CREAM	GLOBUS	HTCONDOR
HTCONDOR	39	10	8	29
LOADLEVELER	2			
LSF		51	1	
OGS/GE	2			
PBS		80	4	4
PBSPRO	1			
SGE	2	36	1	1
SLURM	35	13	1	6
TORQUE	7	162		

CE and batch system trends

- CE types
  - CREAM by far the most numerous
  - ARC and HTCondor on the rise
- Batch systems
  - PBS/Torque by far the most numerous
  - HTCondor and SLURM on the rise
- PIC working on APEL parser for HTCondor CE + batch system

Configuration methods

- YAIM – still there, but minimally maintained
- Puppet – on the rise
- Ansible – ditto?
- Quattor / Aquilon
- A bunch of others used at some sites
  - Chef
  - SaltStack
  - CFEngine
  - …

Lightweight sites (1)

- How might sites provide resources with less effort?
  - Storage → followup in Data Management steering group
  - Computing → followup in Ops Coordination

- Here we are mostly concerned with EGI sites
  - US-ATLAS and US-CMS projects: see CHEP talk

- Site responses to a questionnaire show the potential benefits of shared repositories
  - OpenStack images
    - Pre-built services, pre-configured where possible
  - Docker containers
    - Ditto
  - Puppet modules
    - For site-specific configuration

Lightweight sites (2)

- CE + batch system not strictly needed
- Cloud VMs or containers could be sufficient
- They can be managed e.g. with Vac or Vcycle
  - Several GridPP sites are doing that already
    - All 4 experiments are covered
  - The resources are properly accounted
- They can directly receive work from an experiment's central task queue
- Or they can join an HTCondor pool to which an experiment submits work
  - Proof of concept used by GridPP sites for ALICE

Containers

- An isolation paradigm much lighter than VMs
- See this recent talk by Andrew Lahiff on usage at RAL
- A new tool to launch containers is gaining momentum in our community and beyond: Singularity
  - US-CMS already use it at their T2 sites
  - Provide SL6 environment on CentOS/EL7 WN
  - And isolate the user payload from the pilot
    → replacement for gLExec!
    - See today's Traceability & Isolation WG talk by Vincent Brillault
  - It currently needs to be *setuid root* because the kernel requires root for mount namespace operations
    - A future EL update (7.4?) is expected to lift that restriction
- See the workshop on containers co-located at ISGC!
- See the ATLAS mini workshop on containers later today!

CentOS/EL7

- ALICE, ATLAS and LHCb can run on it today
- For ATLAS please see this recent talk by Alessandra Forti
- CMS are making good progress
- For now CMS can run on CentOS/EL7 if the site provides *Singularity* at the same time
  - To allow the jobs to have an SL6 environment
- EMI WN and UI meta packages foreseen to be released in the May update of UMD-4
  - Preliminary versions available in the community preview repository

IPv6-only computing resources

- Please see Dave Kelsey's Jan GDB talk and his talk earlier today
- An experiment may be able to use IPv6-only computing resources efficiently only when a large part of its SEs are dual-stack!
  - Depends on the experiment
- April this year: CERN + T1 sites should have (some) dual-stack storage
- April next year: many/most sites *ought* to have dual-stack storage…

Benchmarking and accounting

- Please see the Feb GDB report of the HEPiX Benchmarking WG
  - Studying fast benchmarks
  - Looking into replacement of the archaic HS06
- Please see the Jan GDB report of the Accounting TF
  - The new WLCG accounting portal and reports have been validated
  - Further work on open issues, e.g. which changes will be needed to replace HS06
- Please see this recent talk on both subjects by Alessandra Forti

Experiment trends

- ATLAS, CMS: multi-core vs. single-core jobs
  - "*Tetris*" problem: up to 7 unused single-core slots to create an 8-core slot
- LHCb: job "masonry"
  - Use a job slot's remaining time as efficiently as possible
  - Would benefit a lot from MJF deployment
    - See next page
- ATLAS Event Service jobs
  - Save often, lose little work when killed by the batch system

Machine/Job Features

- Please see Andrew McNab's recent talk
- A common API that jobs can use to discover the parameters of their environment
- Also provides graceful shutdown mechanism
  - Allow jobs to exit cleanly
- Can thus enhance QoS on pre-emptible resources
- And help avoid the "*Tetris*" problem in single/multi-core mixes
  - Make 8 single-core jobs finish at the same time

Computing resource SLAs

- Please see Gavin McCance's recent talk
- Extra computing resources could be made available at a lower QoS than usual
  - Jobs might e.g. get lower IOPS and would typically be pre-emptible
  - MJF functionality can help smooth the use
- They would have an SLA between those of standard and volunteer resources

Volunteer computing

- The LHC@home project coordinates volunteer computing activities across the experiments
- ATLAS have benefited from 1-2% extra resources for simulation workloads
- See this recent talk by David Cameron
- It might even become a way for a computing-only lightweight site to provide its resources
  - The central infrastructure can scale at least for simulation jobs
  - The resources can be properly accounted in APEL

HPC resources

- HPC is not a natural match for experiment jobs
  - Fast interconnects not needed
  - Often no external network, no local disk
    - Getting better on new machines
- NorduGrid HPC sites use ARC functionality to have data staged in and out for jobs
- US HPC facilities each have different edge services and operational policies
  - Has led to ad-hoc complexities in the job frameworks of ATLAS and ALICE
- ATLAS working on new *Harvester* service as intermediary between PanDA and pilots
  - See this recent talk by Tadashi Maeno
  - Common architecture for HPC, grid, cloud, …

Cloud initiatives

- Commercial cloud usage is expected to become important in the next years
- Experiments and several sites have been gaining experience with it in the last years
- US-ATLAS/BNL and US-CMS/FNAL have been doing big exercises e.g. with Amazon
- CERN, European T1 sites and other big-data institutes are working together in Helix Nebula – The Science Cloud
- CERN will make such computing resources available to the experiments via HTCondor
  - As already done successfully in previous exercises

Conclusions

- Classic grid services still constitute the backbone of the computing resources for now
- Lightweight site initiatives are expected to help reduce their complexity
- New technologies and paradigms are gaining popularity
  - In particular containers and Singularity