HPC resources for CMS offline computing: an integration and scalability challenge for the Submission Infrastructure

The CMS Submission Infrastructure

The CMS Submission Infrastructure team in CMS Offline and Computing is in charge of operating a set of federated HTC/Condor pools which aggregates resources from 70 Grid sites, plus non-Grid resources, where reconstruction, simulation, and analysis of physics data takes place.

The challenges:
- Operate our infrastructure managing an ever-growing collection of computing resources
- Connecting new and more diverse resource types (including non-x86 CPU architectures and GPUs) and resource providers (WLCG and OSG, HPC, Cloud, volunteers)
- Use all of our resources efficiently, maximizing data processing throughput
- Enforce task priorities according to CMS research programs

HPC growing contribution to CMS computing

HPCs can help supporting the computing needs of CMS:
- Substantial national and supranational investments made and planned
- HPCs are part of the scientific computing infrastructure, and there to stay
- Several HPC integration efforts in CMS

Two main models of HPC integration to CMS computing:
- HEPCloud [1], based on dedicated sites, for HPC resources in the U.S.
- Transparent WLCG site-extension, employed for the EU HPC facilities.

It appears that the compute capacity used by CMS at HPC facilities could be maintained at the current level, if not higher, in the future years.

(1) HEPCloud: eSci/1715.01300

Challenges towards use of HPC resources by CMS

- Each HPC is different:
  - Diversity in technical and policy constraints (networking, security, edge services, etc)
- Negotiation and deployment of resources vary by HPC:
  - Simplicity, CVMFS
  - Data cache services (for application software, conditions data, event data)
  - Job gateway (CE)

- Diversity in technical and policy constraints (networking, security, edge services, etc)

Data cache services (for application software, conditions data, event data)
- Singularity, CVMFS

Scalability in the CMS SI

- Operate away from any scalability limiting factor: critical aspect for a system that is designed to perform in a dynamic environment, adapting itself to growing resource demands by CMS, resource availability in the WLCG, and the mix of workloads it has to manage:
  - Proactively find those limits, in every direction, and evolve the infrastructure to push them further away:
    - Total computing power our HTCondor pools can harness and use efficiently
    - Collector capacity to process the stream of slot updates and keep resource status fresh
    - Negotiator matchmaking cycle time under control
    - Total number of workfows we can manage and jobs we can run simultaneously with our pool of schedulers

The SI 2023 scale tests (I)

Our main goal for the 2023 tests is to assess the scalability of our Global Pool, considering the following updates since our latest tests (2021):
- Evolution in HTCondor software (tested version 10.0.1)
- New physical host for the central manager (AMD EPYC 7302 at 3 GHz)
- Adoption of token-based authentication for HTCondor services in our SI
- Incrementally improved configuration of our infrastructure

As in previous tests:
- Use Glokins and HTCondor to generate enlarged copies of our Global Pool
- Secondary collector service in the pool as a source of monitoring data, to reduce stress from non-essential queries on the main collector

Initial results and mitigations:
- Found first bottleneck in the total capacity of our pool of schedulers to supply jobs to the resource pool:
  - At 1 MB of RAM per running job, our 10 schedulers with 50 GB of memory saturated at 50k running jobs each
  - Saturation at 0.5M running jobs, comparable to our 2021 result
- Sched pool capacity subsequently enlarged to, in principle, be able to support nearby 1M running jobs

The SI 2023 scale tests (II)

Our latest results:
- The main Global Pool collector, which processes slot status updates, becomes increasingly stressed as the size of the pool and the rate of updates keeps growing.
- When the collector becomes fully saturated (collector duty cycle < 1), it can’t provide further updated slot status information to the negotiator, thus job slot matchmaking becomes slow and inefficient

This effectively limits the total number of running jobs that a pool can manage:
- Our latest tests pushed the scalability of our Global Pool to about 800k simultaneously running jobs

Conclusions and Future work

- HPC resources contribution to CMS Computing has increased in the recent past and is expected to continue growing in the near future
- The integration of these resources presents a number of challenges to our infrastructure, including that of scalability
- Understanding and managing all aspects of our CMS baseline compute resources, and especially due to CPU saturation limits at HPCs
- The CMS SI team, in close collaboration with the HTCondor and GlideinWMS developers, periodically evaluates the scalability of our infrastructure, in order to anticipate and remedy future scaling and stability problems and stay off the bleeding edge
- Our 2023 scalability tests showed the capacity of our SI to support up to 0.8M simultaneous running jobs
- The LHC program extends well into the future, so we need to continue pushing the SI for higher scales, as required by CMS needs, while maintaining stability and efficiency