Stability of the CMS Submission Infrastructure for the LHC Run 3

A. Pérez-Calero Yzquierdo[1], Edita Kizinevic[2], Farrukh Aftab Khan[3], Hyunwoo Kim[3], Marco Mascheroni[4], Maria Acosta Flechas[3], Nikos Tsipinakis[2] and Saqib Haleem[5], on behalf of the CMS collaboration

1. CIEMAT and PIC (ES), 2. CERN, 3. Fermi National Accelerator Lab. (US), 4. Univ. of California San Diego (US), 5. National Center for Physics (PK)

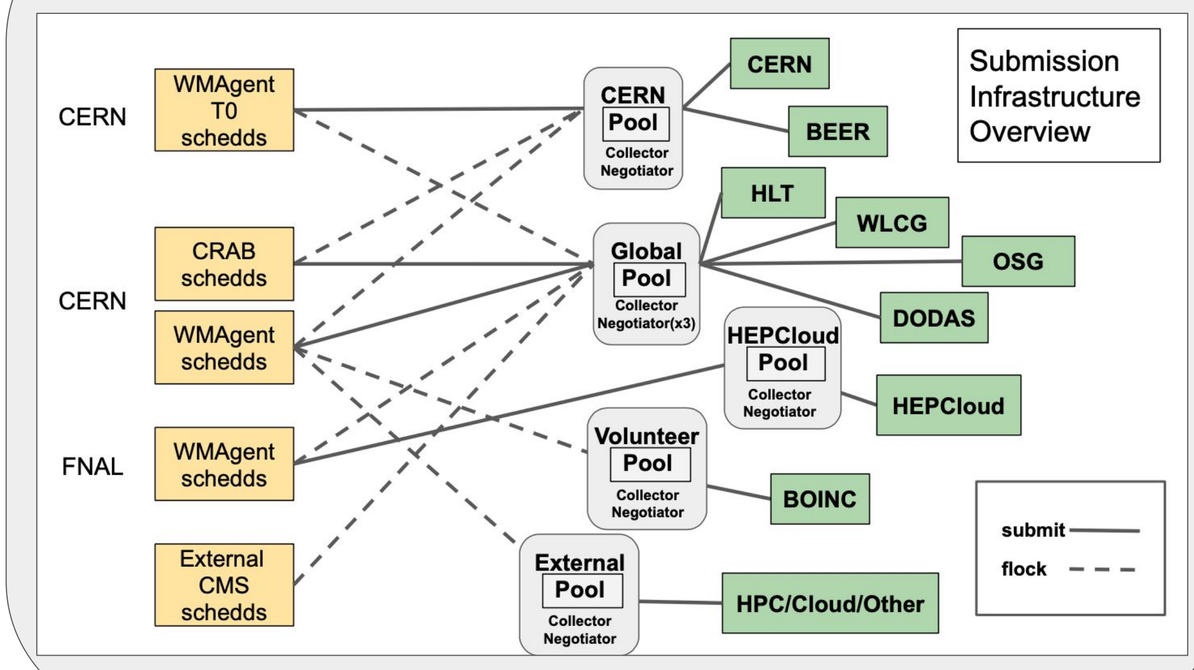
The CMS Submission Infrastructure

- **The CMS Submission Infrastructure:** team in CMS Offline and Computing in **charge** of:
  - Organizing **HTCondor** and **GlideinWMS** operations in CMS
  - Maintaining a **Global Pool, an infrastructure of distributed compute resources where reconstruction, simulation, and analysis of physics data takes place**
  - Communicate CMS **priorities to the development teams** of glideinWMS and HTCondor

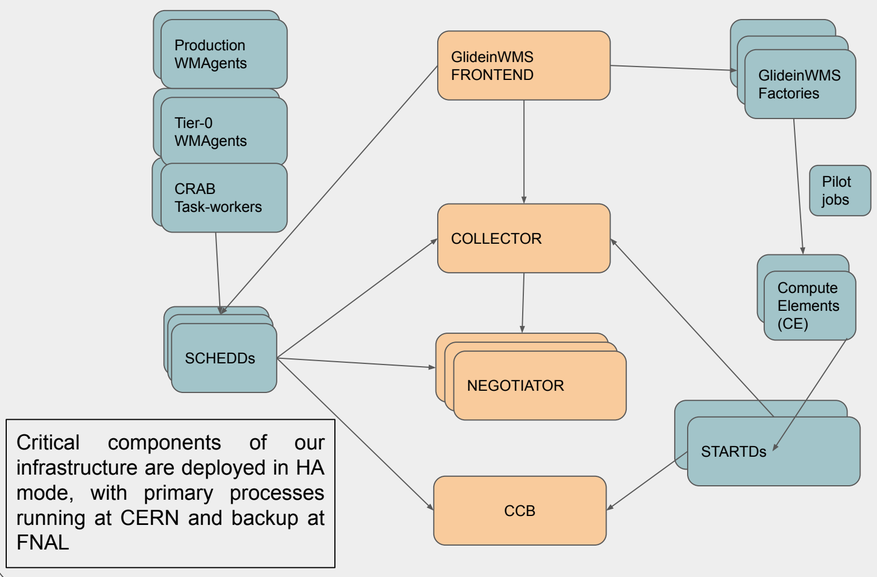
- **In practice:**
  - We operate a set of federated HTCondor pools which aggregate **resources from 70 Grid sites, plus non-Grid resources**
  - We regularly hold **meetings with** HTCondor and glideinWMS **developers** where we discuss current **operational limitations, new feature requests and future scale requirements**

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

Federated HTCondor pools



CMS Global Pool and stability: High-Availability deployment and redundancy and horizontal scaling of its components



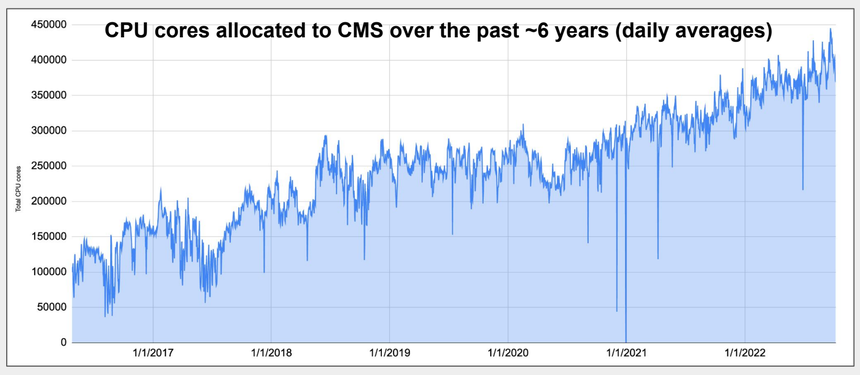
Critical components of our infrastructure are deployed in HA mode, with primary processes running at CERN and backup at FNAL

Pilot factories are deployed at CERN, FNAL and UCSD, overlapping with OSG's Factory Ops team

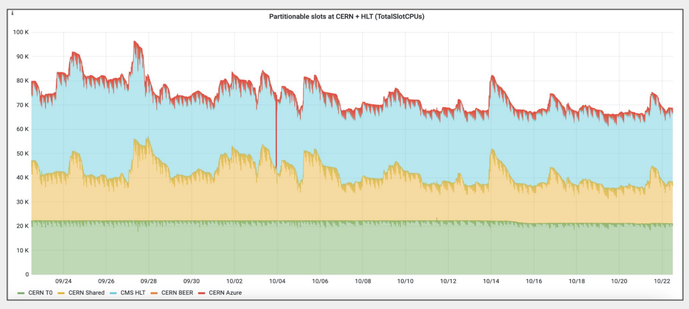
Schedulers are also deployed at CERN and FNAL, in numbers sufficient to manage CMS workloads for each major task category: centralized production (MC production and data processing campaigns), analysis tasks and Tier-0 jobs (including prompt data reconstruction at CERN).

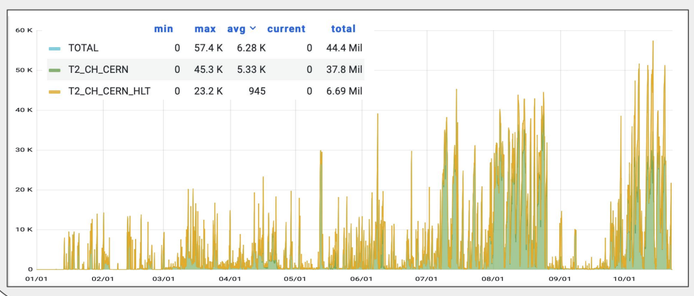
Scalability Frontiers

- **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
    - Front-end and factories capacity to react to oscillating resource demands and provision them
    - Collector capacity to process the stream of slot updates and keep resource status fresh
    - Negotiator matchmaking cycle time under control
  - Total number of workflows we can manage and jobs we can run simultaneously with our pool of schedds



The LHC Run3 and CMS Tier-0 operations





As a critical part of its mission in support of CMS, the Submission Infrastructure aggregates and manages all CPU resources available for CMS at CERN:

- CERN "Tier0" and "Shared" clusters
- The former Run2 HLT CPUs ("P5 permanent cloud")
- Opportunistic (BEER) and cloud (Azure) CPUs provided via CERN

Since the start of the LHC Run3, the Submission Infrastructure has been providing resources for Tier-0 tasks (data repacking and prompt reconstruction)

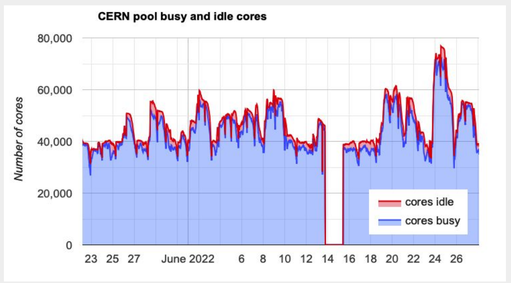
Given the mission-critical role of Tier-0 tasks for the CMS collaboration, several scale tests were performed in the first months of 2022 to ensure a smooth restart of data taking operations. Tier-0 test workloads were injected in bursts to ensure CPU allocation in sufficient quantity and with required responsiveness

The Submission Infrastructure "fire drills" (I)

Before the start of the LHC Run3 phase with major data taking operations ("stable beams" available since July 2022), the Submission Infrastructure team performed a series of exercise to validate the effectiveness of safety mechanisms embedded in its design.

In these exercises ("fire drills") carried out in May and June 2022, critical services associated with both the Global and CERN pool infrastructures were intentionally disabled, in order to force secondary services reaction. These actions, in most cases, had no impact on the overall performance of the infrastructure (as expected!). In others, some second-order effects were detected, followed by corrective actions. Some of the actions that contributed to improvements were:

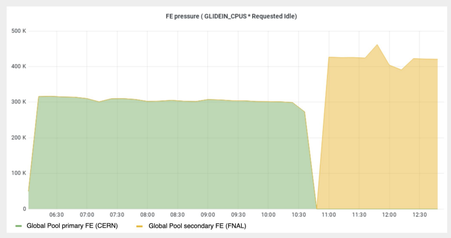
- June 13th: Condor service stopped on primary central manager (host for collector and negotiator processes) for for the CERN pool. Observations:
  - **Negotiator:** Automatically started on backup central manager node at FNAL
  - **FE:** primary collector loss did not affect performance of the primary CERN FE, as the CERN FE is configured query both collectors (primary and secondary) at once
  - **Schedds and startds connectivity:** all remained connected to the pool via the backup collector at FNAL
  - **Monitoring:** Service interruption due to exception in the monitoring script (triggered when trying to connect to the primary collector). The script was corrected and the infrastructure monitoring service recovered.



- June 16th: Backup central manager for the CERN pool disabled. Observation:
  - Pool performance, driven by primary central manager, was not affected.
  - The only (minor) observed effect was the interruption of the monitoring service, which is configured to mainly query the secondary collector, in order to minimize load in the primary one. An alarm was introduced to alert from the loss of secondary collector.

The Submission Infrastructure "fire drills" (II)

- June 30th: Global Pool primary FE service (at CERN) stopped. Observations:
  - Backup FE (at FNAL) started requesting new pilots from the factories
  - However, the pilot pressure generated by the secondary FE was being overestimated. Root cause found as the secondary FE was incorrectly querying both pool collectors in parallel, leading to some workload double-counting. Backup FE configuration was corrected.



Conclusions of the fire drill exercises:

- The Submission Infrastructure system is **stable and fault tolerant** in the context of any one of the central components (GlideinWMS Front-End and HTcondor collector, negotiator or CCB) accidentally becoming unavailable.
  - Some corrections were introduced as a consequence of the tests, mainly in relation to the service monitoring the submission infrastructure.

- However, a number of secondary sources of potential instability remain and should also eventually be tested.
  - For example, an outage of the GIT repositories storing the configuration information for each of our services cause additional delays in the deployment of new hosts as recovery measure to counter the loss of any main pool element (for example, additional pilot factories or schedds, in case a substantial fraction of them would be unavailable).
  - Several services as hosted by VMs whose configuration is located in separated CEPH disk volumes. In case of outages this would affect for example, the FE and some HTCondor services running on VM (e.g. schedds). On the other hand, main services such as collector, negotiator and CCBs are running in physical nodes, with a separate disk but no CEPH mounted.

Conclusions and Future work

- Key to the success of the Submission Infrastructure stably operating, while reaching ever higher scales, has been CMS's close coordination with the HTCondor developers, the glideinWMS developers and the OSG factory operations team
- Thanks to these collaborations, we have been able to anticipate and remedy future scaling and stability problems and stay off the bleeding edge of limitations
- The LHC program extends well into the future, so we need to continue pushing for **higher scales**, as required by CMS needs, while maintaining **stability** and **efficiency** and remain relevant by **adapting** to tools/technology changes