Transitioning the CMS pools to ALMA9

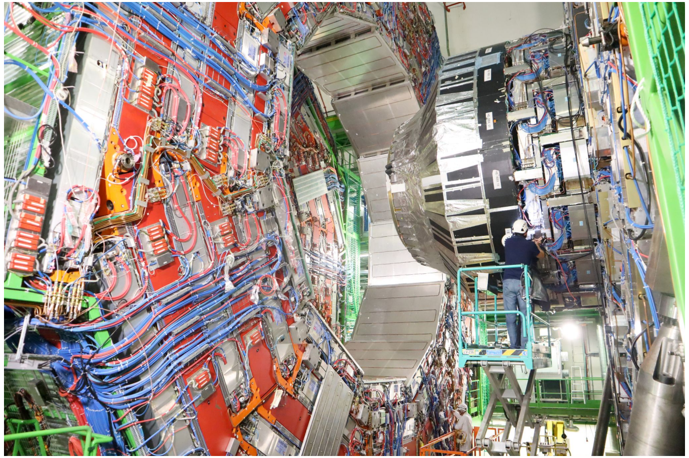
2024 European HTCondor Workshop

R. Florian von Cube

for the Submission Infrastructure Group in the CMS collaboration

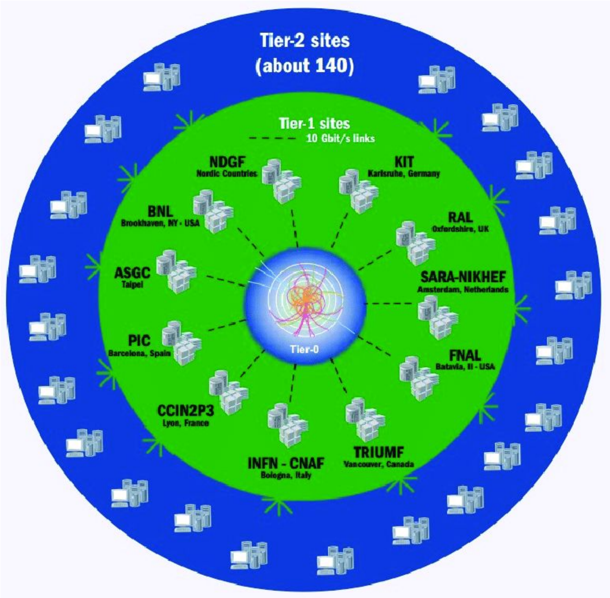
The **C**ompact **M**uon **S**olenoid

- One of the four major experiments at the LHC at CERN, where proton-proton interactions are produced at the world’s highest collision energy
- Studies elementary particles and fundamental forces in a broad range of physics processes
- Collects hundreds of petabytes of data for physicists to analyze and perform statistical analyses



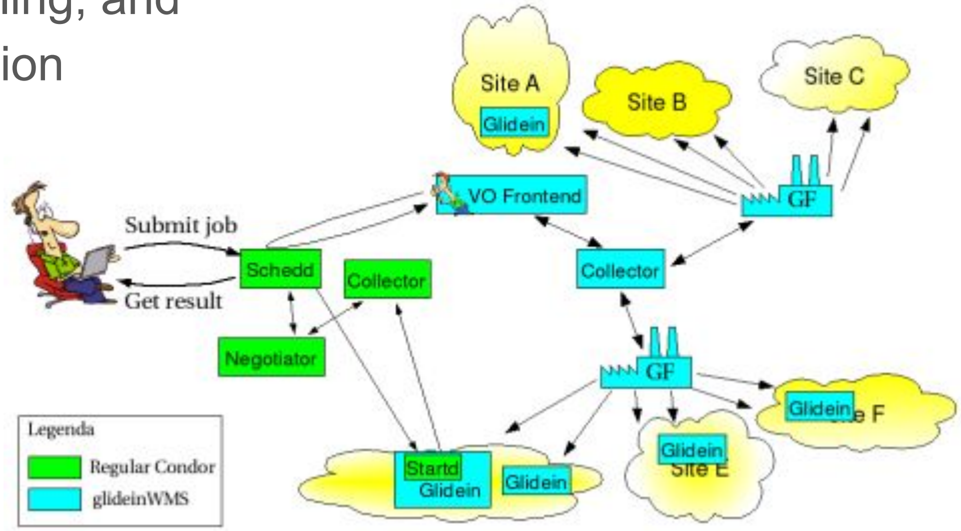
The WLCG

- About 170 computing centers in 42 countries organized in the “**W**orldwide **L**HC **C**omputing **G**rid”
- Aggregates 1.4M CPU cores, 1.5 exabytes of disk and tape storage to serve the scientific programmes of the LHC experiment collaborations



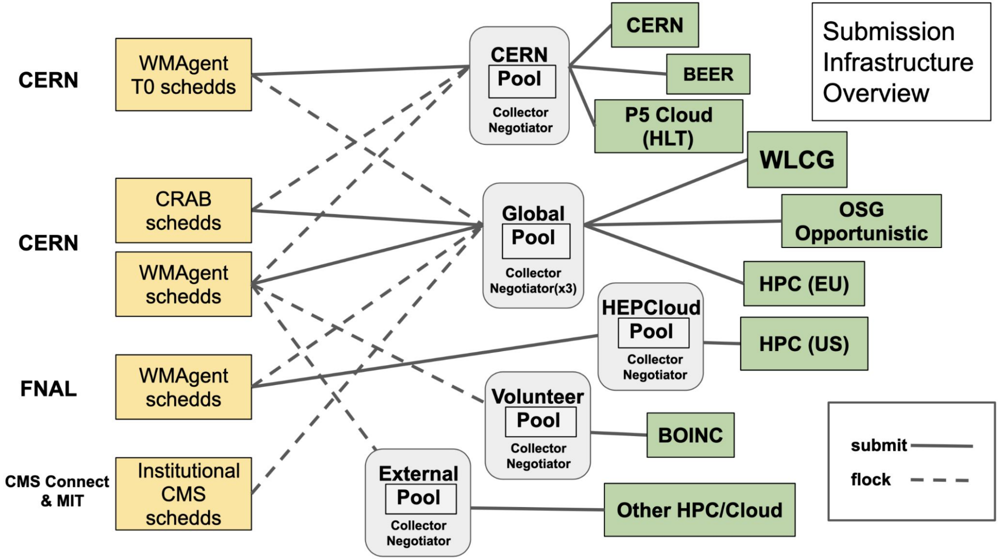
Submission Infrastructure

- Group within CMS Offline & Computing coordination area, responsible for providing compute resources to experiment and user groups
- We use HTCondor for job scheduling, and GlideinWMS for resource acquisition
- Regular handling of 500k cores in several federated pools
- Close and regular contact to HTCondor developers for guidance and tests
    - Big thanks to Jaime et al.!



The CMS Submission Infrastructure setup

- Different HTCondor pools for different “clients”
- ~70 schedds (APs) at CERN
- One primary central manager and CCB per pool at CERN and a secondary at Fermilab for high availability
- Heavy use of flocking for optimal use of the resources
- Pools comprising a total of ~500k CPU cores



The *Global Pool* setup

- Main pool for running the majority of CMS jobs
- Negotiator and collector run on dedicated (hardware) machines
  - With backups at Fermilab
  - Shared configuration repository, deployed via puppet
- GlideinWMS factory and frontend each on a separate machine
- Several schedds in virtual machines, dedicated to specific groups
- Resources mainly from WLCG computing centers, but also growing share from HPC centers around Europe and the US

The *CERN Pool* setup

- Pool mainly for immediate data reprocessing
  - Include highly critical prompt processing of newly recorded experimental data
  - Of highest priority as unavailability can lead to loss of experiment data
- General setup same as the Global pool:
  - Negotiator, and collector run on dedicated machines with backups at Fermilab
  - GlideinWMS factory and frontend each on a separate machine
  - Dedicated schedulers
- Resources mainly at CERN for data locality and quick interventions
  - Includes several resources from main CERN data center
  - Run 2 “trigger farm” (CPU, superseded)
  - “Opportunistic use” of current (Run 3) trigger farm (CPU+GPU), when not needed for data taking

The ITB(DEV) Pool

- Two internal test bed (development) pools
- Used for testing “new” setups, integrating development resources
- ITBDEV (unstable) set of machines, used for major debugging
  - Used for adapting puppet manifests of different machine types to ALMA9
- ITB mirrors global pool setup, allows for rigorous testing of new resources, plus periodic scale testing
  - On newer HTCondor version, testbed for new features
  - Allows us to provide test machines
  - Used by other groups to validate implemented changes

The Transitioning Strategy

- The CMS Submission Infrastructure (SI) must remain operational at all times, as it is critical for CMS data taking
  - Transition the setup to new ALMA9 hosts while fully in production
  - Relying on high availability to ensure no interruption of production

- Migration stages:
  - Start by setting up test machines on ALMA9 in ITBDEV, deploy HTCondor, weed out any issues
  - Think of strategy of transitioning whole pool, transition ITB.
  - Test new setup in ITB with several groups
  - Transition Global pool
  - Transition CERN pool

Transitioning Central Managers to ALMA9 in Production

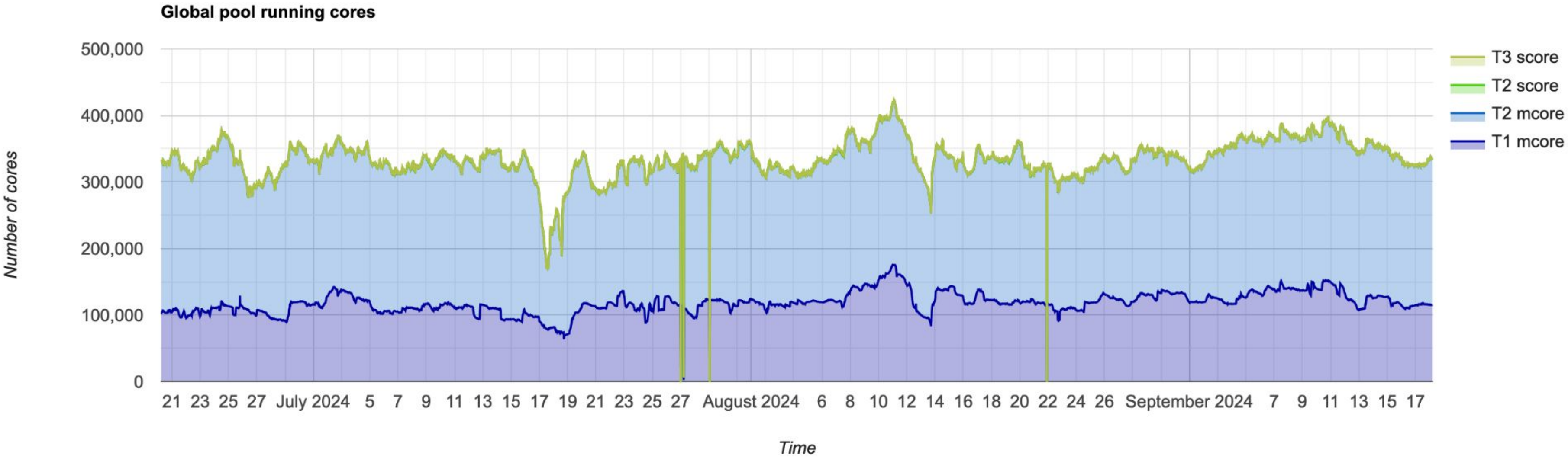
- Collector transition was rather smooth:
  1. Switch off main collectors at CERN, wait for Fermilab to take over
  2. Change collector hostname in all configurations and swap order (now Fermilab first, CERN second), wait for CERN collector to be populated
  3. Change order in configuration back
- Same for negotiator:
  1. Switch off negotiator, wait for Fermilab to take over
  2. Change order to Fermilab first, (new) CERN host second, wait
  3. Change order in configuration back

Transitioning APs to ALMA9

- Pools are accessed through ~70 schedds (APs) provided for different groups, namely production, analysis and Tier-0
- For the ALMA9 transition, groups were provided with…
  - … ALMA9 schedds in Global/CERN pool to test with existing (non-ALMA9) setup
  - … schedds in ITB to test with (new) ALMA9 setup
- Transitions were coordinated with groups as they need to adapt their setups
  - Services using schedds need to be tuned for efficient use of schedd in terms of memory, no. of concurrent jobs running…

- Transitioned all of the schedds to ALMA9 machines without major interferences

Transition without major Interruption



Other transitions: HTC v2 python bindings

- CMS analysis tasks submission tool (CRAB) as alpha tester
  - We got access to pre-release code starting in March
- Very positive experience
- Made us fix code untouched since 6+ years and long deprecated in HTC
  - Our code is cleaner now. Python API had improved over the years.
- Once that was done, updating to v2 bindings was 1-day work
  - Only change is semantic of schedd.spool(), more clear now, thanks !
- Pleased to have helped developers fix small things before release
- Had excellent support along the way. Thank you ToddM !
- Eager to deploy coming v24 to be able to have htcondor2/classad2 available on our production schedulers for large scale testing and hopeful migrate by end of the year

Summary

- Essential resources to CMS data taking processing and analysis (about 500k CPU cores) unaffected by transition to ALMA9 of
    - Central machines and their HA counter parts
    - 70 APs

- Modularity and high-availability of HTCondor allowed us to replace one service after another while keeping the pools running

- Transition was mostly transparent to users and groups