Using the glideinWMS System as a Common Resource Provisioning Layer in CMS

J. Balcas[1], S. Belforte[2], B. Bockelman[3], D. Colling[4],
O. Gutsche[5], D. Hufnagel[5], F. Khan[6], K. Larson[5], J. Letts[7],
M. Mascheroni[8], D. Mason[5], A. McCrea[7], S. Piperov[9],
M. Saiz-Santos[7], I. Sfiligoi[7], C. Wissing[10]

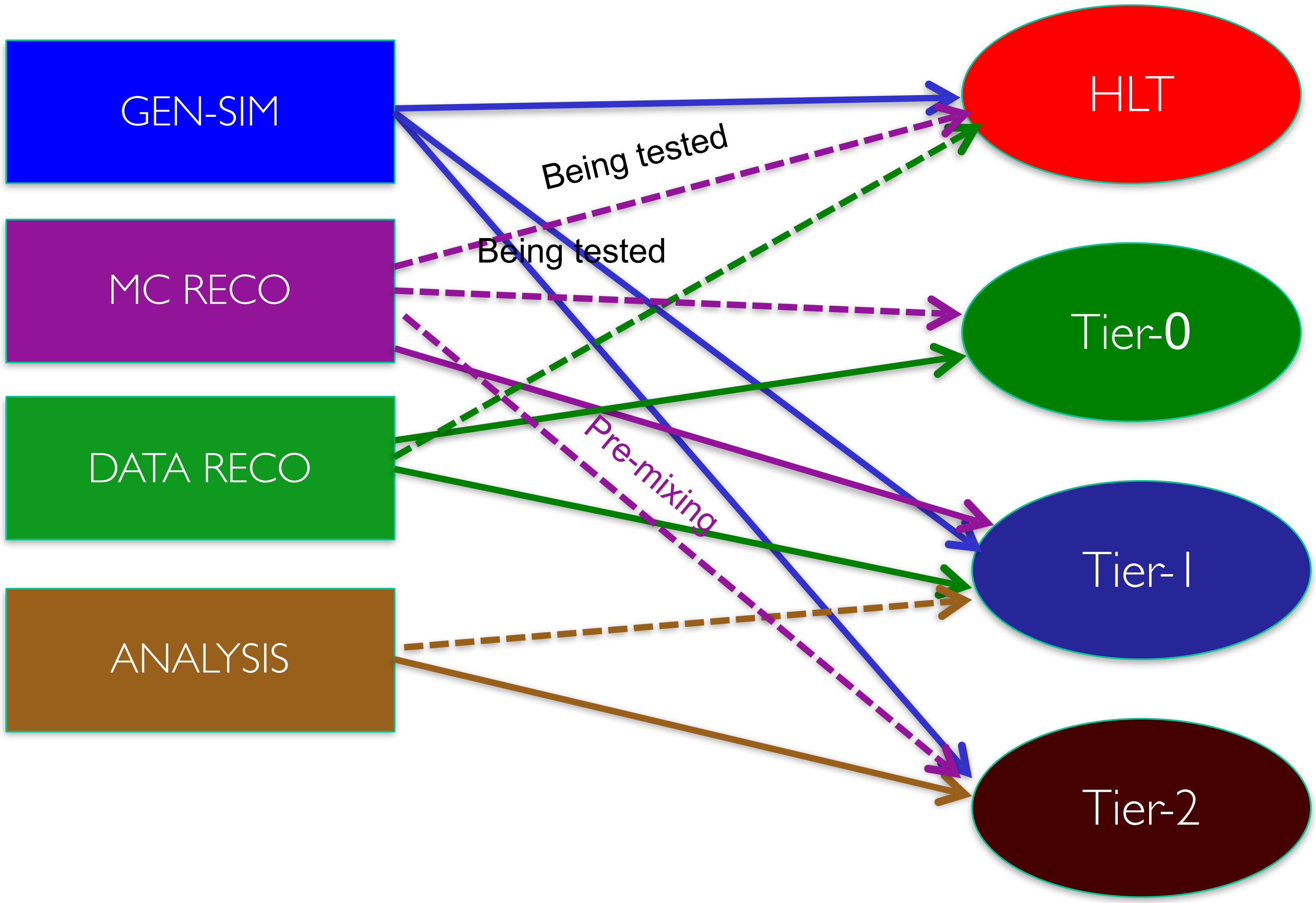
1. Vilnius Univ. (LT) 2. INFN-Trieste (IT) 3. Univ. Nebraska Lincoln (US) 4. Imperial College London (UK) 5. FNAL (US) 6. NCP (PK) 7. UCSD (US) 8. INFN-Milano (IT) 9. Brown (US) 10. DESY (DE)

Outline

- Motivation
- glideinWMS and HTCondor
- Clients
- Current Use Cases
- Future Use Cases
- Support Model
- A Note on Scalability

Motivation

- **Maximum flexibility in prioritizing different kinds of work on different types of resources, as needed.**



Motivation

- Run 1: Mix of grid sites and some local resources.
- During LS1, new types of resources became available (Cloud, Opportunistic, HPC Centers, etc.)
- Challenge: Simplify the submission infrastructure to cover flexibly all use cases, resource types.

glideinWMS and HTCondor

- glideinWMS and HTCondor were already being used in several separate pools in CMS.

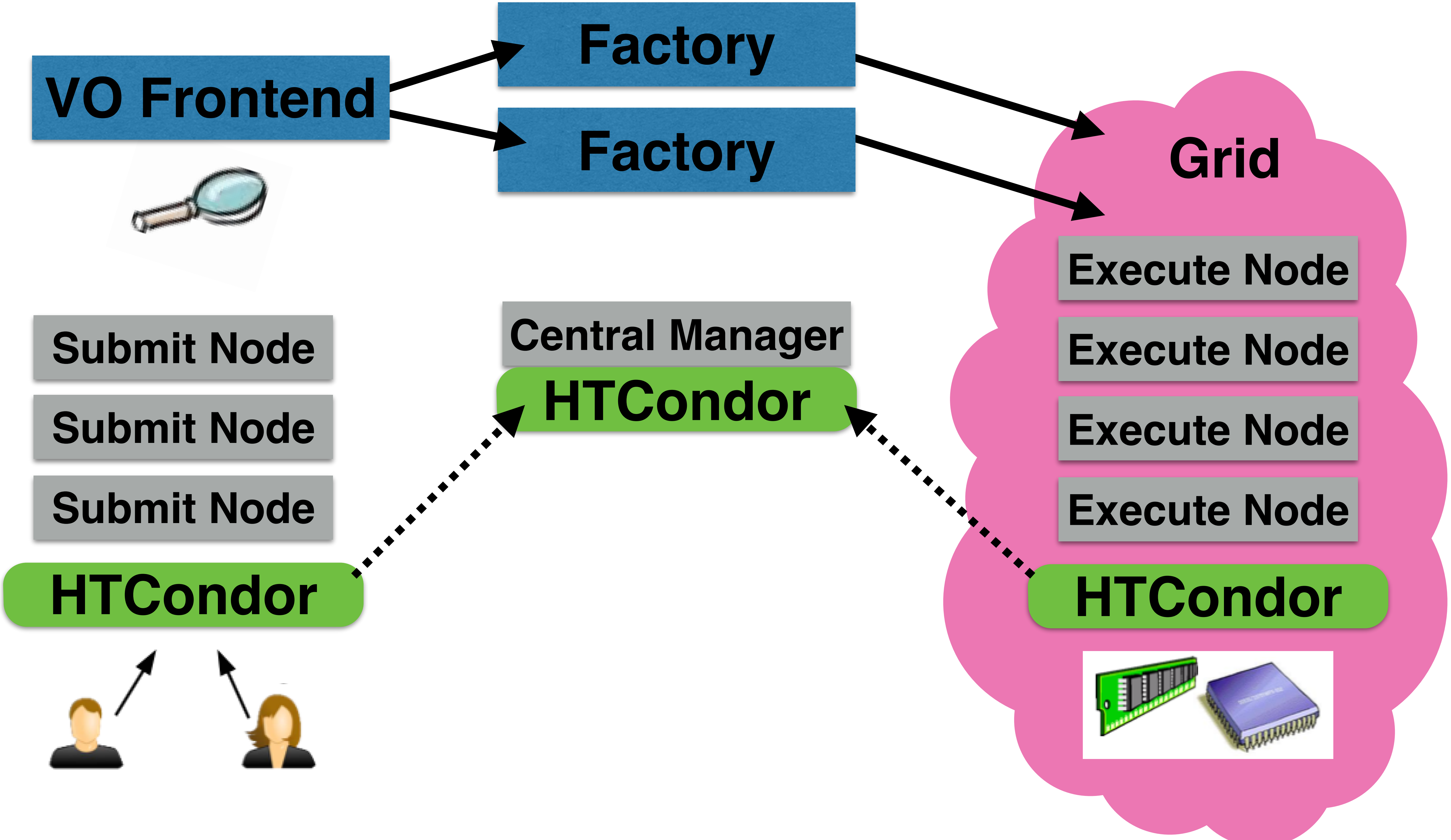
- Major use cases were for data analysis (CRAB) and central data production and processing (WMAgent).

- Plan was to unify these instances in a "Global Pool"

Global Pool

- Flexibly re-prioritize work:
    - among major tasks
    - among tiers (resource types)
- And then add the other types of use cases (i.e. Tier-0) and resources (HLT, Cloud, Opportunistic etc.)

glideinWMS and HTCondor



Inspired by a diagram from I. Sfiligoi, “Introduction to glideinWMS”, Dec. 2012.

Clients

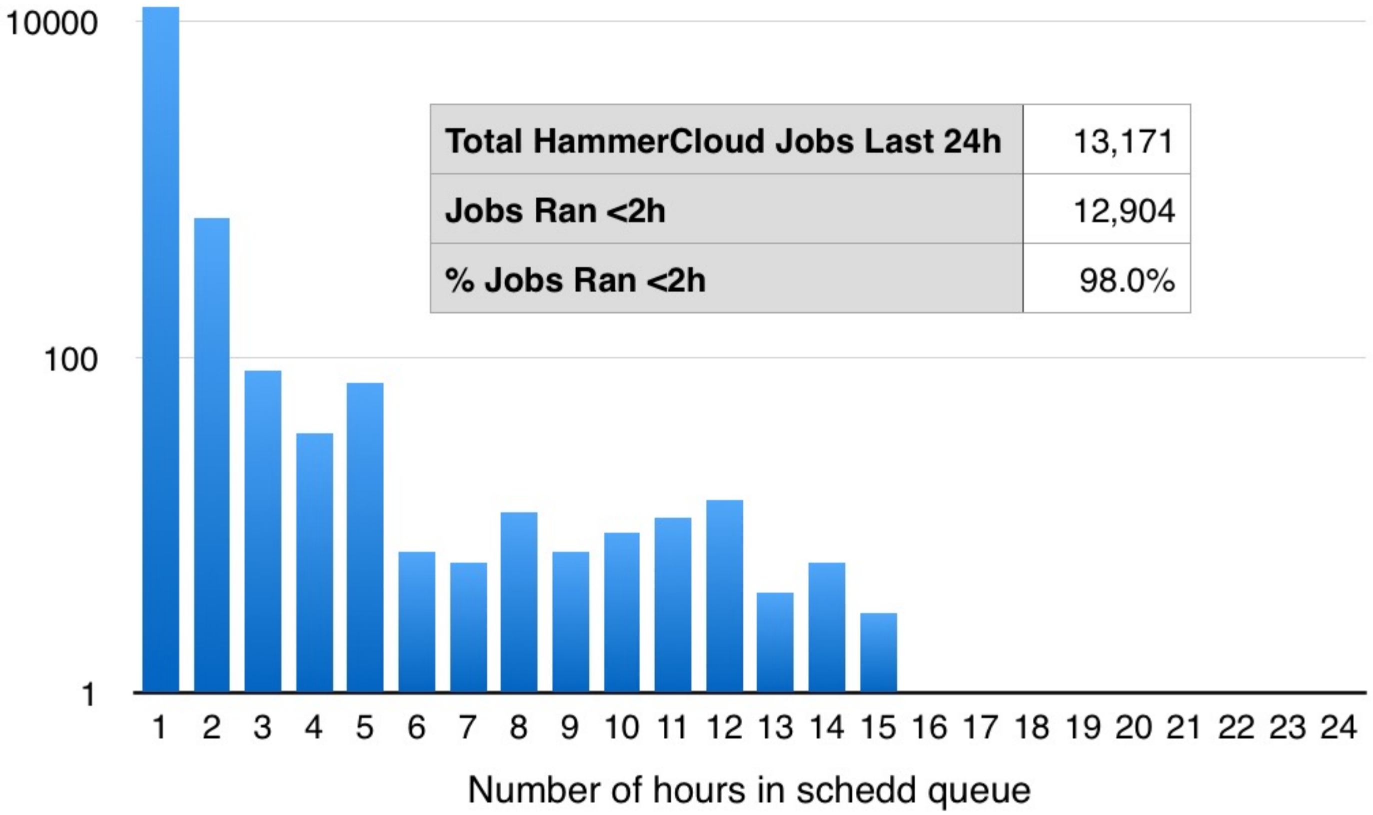
- Major clients:
    - CMS data analysis (CRAB)
    - Central Data Production and Processing (WMAgent)
    - Tier-0
- Each has servers which submit tasks to HTCondor schedulers at CERN and 4 other availability zones.
- However, there are even more use cases, some uncovered as yet.

Use Cases

- Analysis on the Grid with CRAB.

- In the Global Pool, we can boost the priority and fair-share by user as needed.

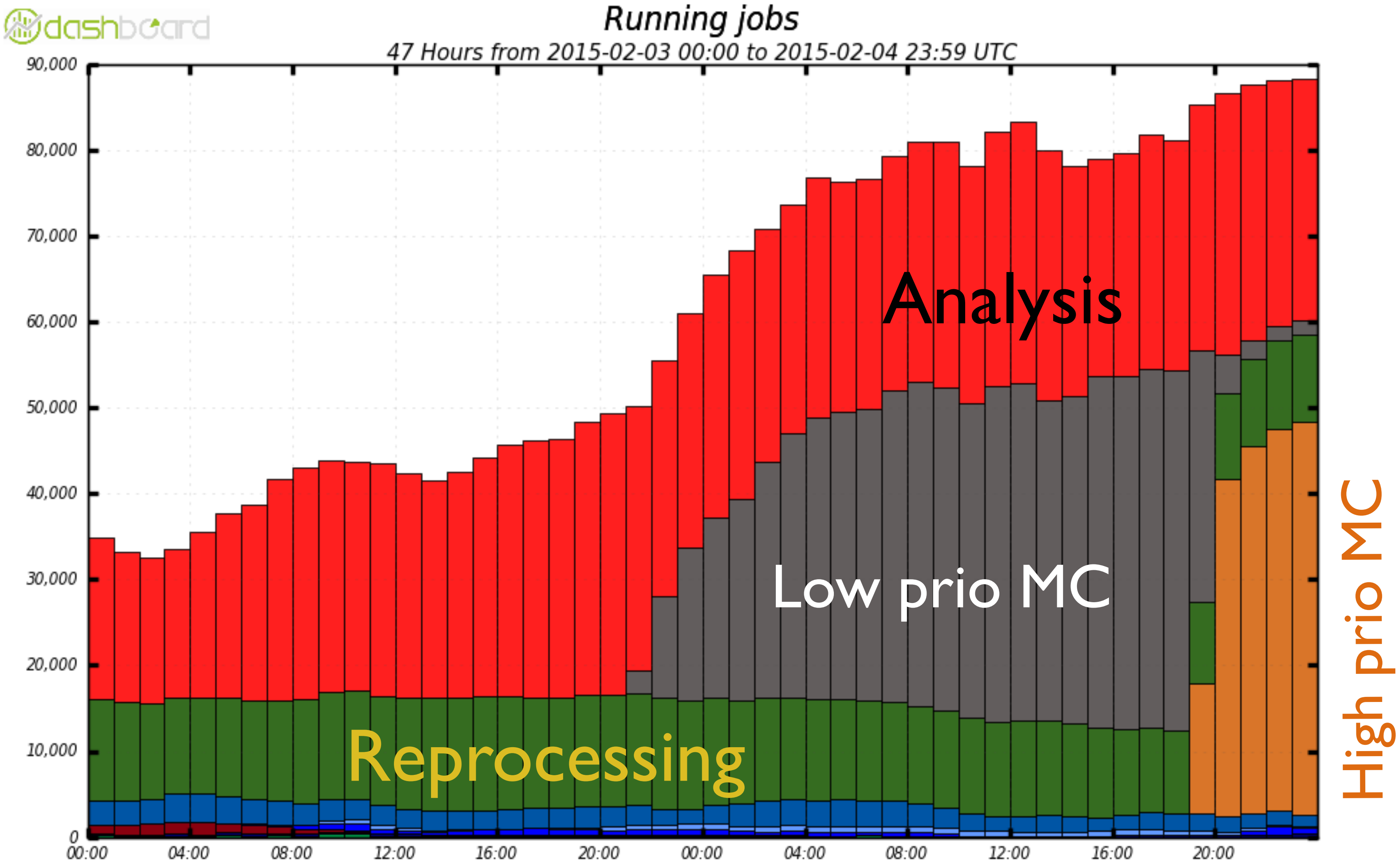
- We are ready to fast track as needed any hot analyses during Run II.



Total HammerCloud Jobs Last 24h	13,171
Jobs Ran <2h	12,904
% Jobs Ran <2h	98.0%

Use Cases

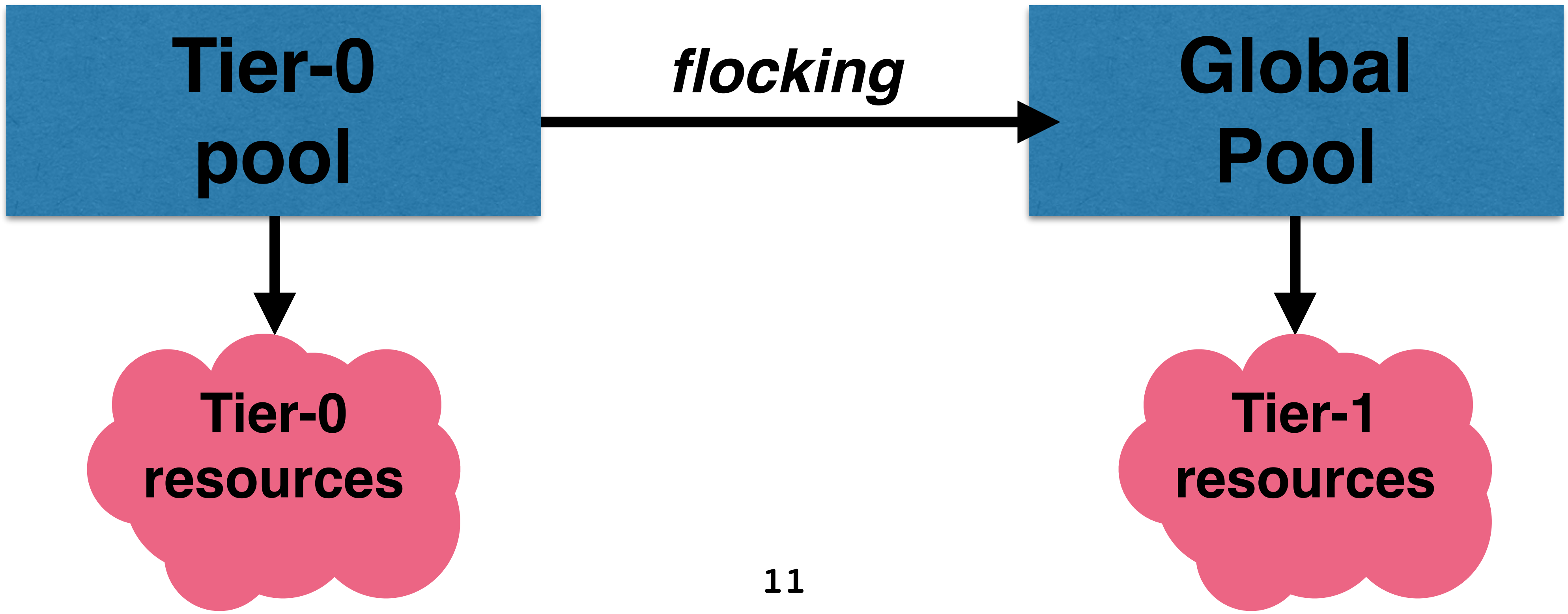
- **High priority Monte Carlo production can quickly take over the resources.**



Use Cases

- Resource provisioning on the Tier-0 is now also done with glideinWMS and HTCondor.
- Independent HTCondor pool for enhanced reliability, risk management.
- Can "flock" work to the Global Pool.

See related work: D. Hufnagel et al., "The CMS Tier-0 Goes Cloud and Grid for LHC Run 2", CHEP15 Oral Presentation #119



Use Cases

- In addition, we can also send work to the HLT farm from the Global Pool during LHC inter-fill periods, which are expected to last several hours.
    - No Grid submission: special VM with startd.
- CMS has first CPU allocations at NERSC and SDSC.
    - Each site presented its own challenges.
    - No Grid submission, only ssh
    - At NERSC investigated BOSCO & Parrot: no CVMFS
- Establishing procedures and testbeds to bring new types of resources like these into the Global Pool so knowledge is propagated.

See related work: D. Hufnagel et al., "Enabling Opportunistic Resources for CMS Computing Operations", CHEP15 Oral Presentation #123

Future Work

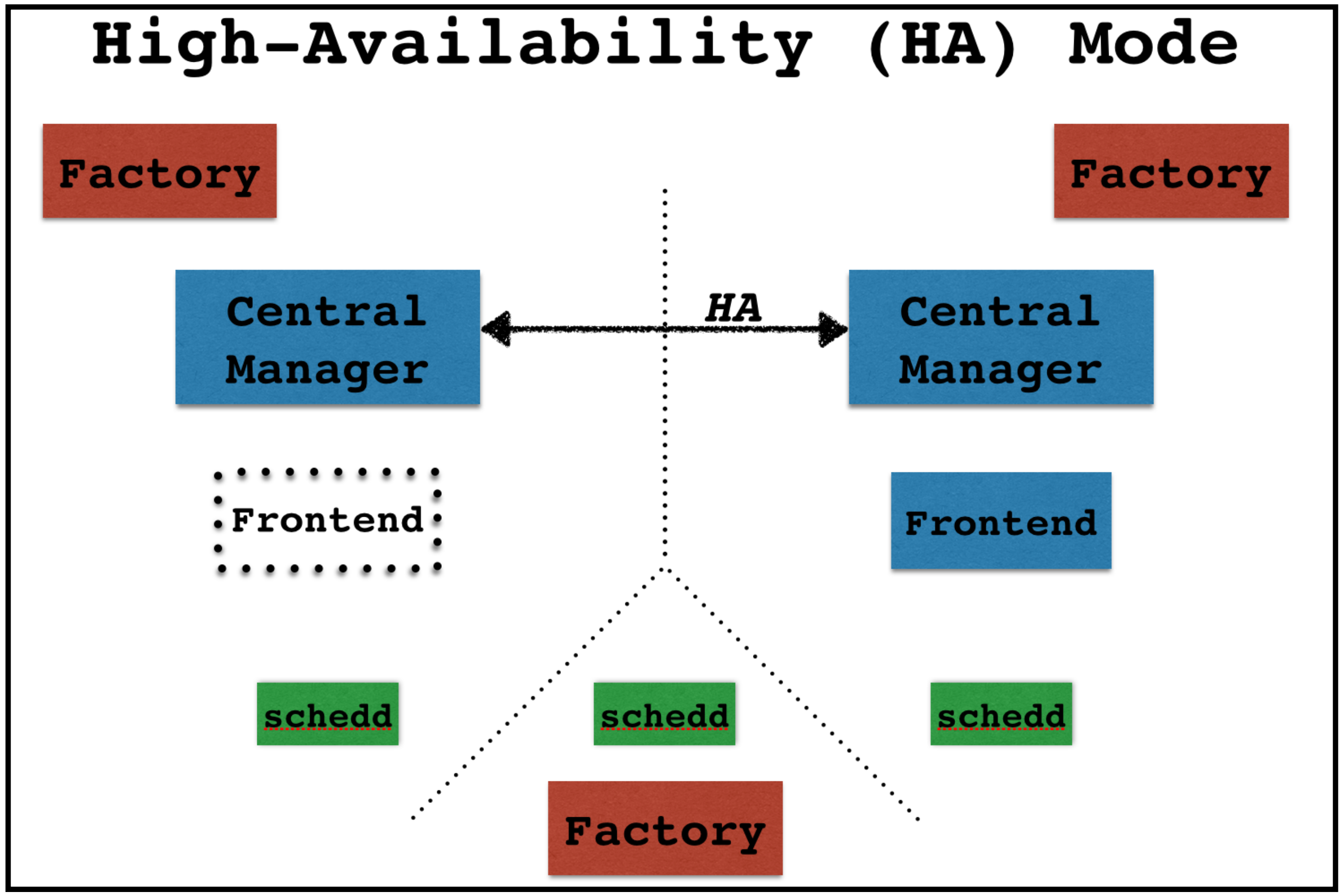
- Access to local resources:
    - With or without a Grid CE, i.e. Tier-3.
    - Local control of queue, user and task prioritization.
- Integrating non-CMS sites seamlessly, i.e. OSG.

Support Model

- Run critical services in several availability zones in HA-mode.

- Global Pool testbed to roll out changes to production.

- Development pool to try out new resource types.



Support Model

- Team of CMS operators at CERN and FNAL.
- OSG runs the glideinWMS factories.
- Bi-weekly dedicated meetings with HTCondor developers, OSG, glideinWMS developers.
    - Communicate priorities through these meetings and ticketing systems.
- Weekly meeting between our major clients, the operations team, and management.

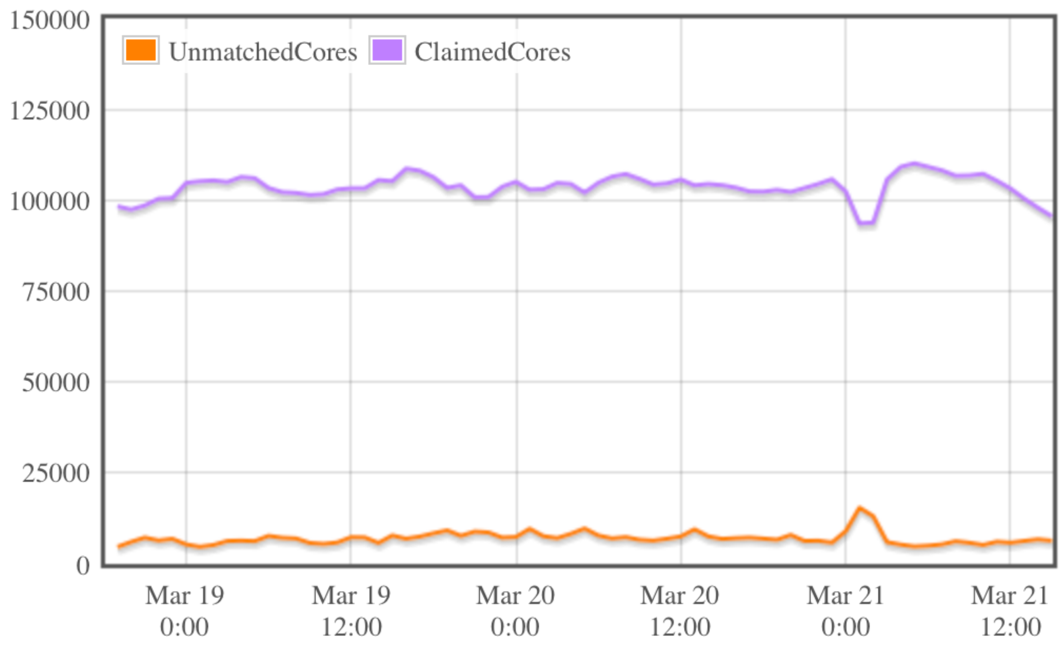
Scalability

- **Over the past year, we have sought to push the limits of scalability of the system in a stable way.**

- **Worked directly with the HTCondor and glideinWMS developers to make improvements in the inter-component communication, negotiator stability, etc.**

- **Also worked closely with the OSG, who proved scalability up to 200K parallel running jobs.**

- **Running the Global Pool at the scale of the pledged resources available (~110K) is routine now.**



See related work:
E. Fajadro et al., “How Much Higher Can HTCondor Fly?”, CHEP15 Poster #6.
J. Letts et al., “Pushing HTCondor and glideinWMS to 200K+ Jobs in a Global Pool for CMS before LHC Run 2”, CHEP15 Poster #371.

Conclusions

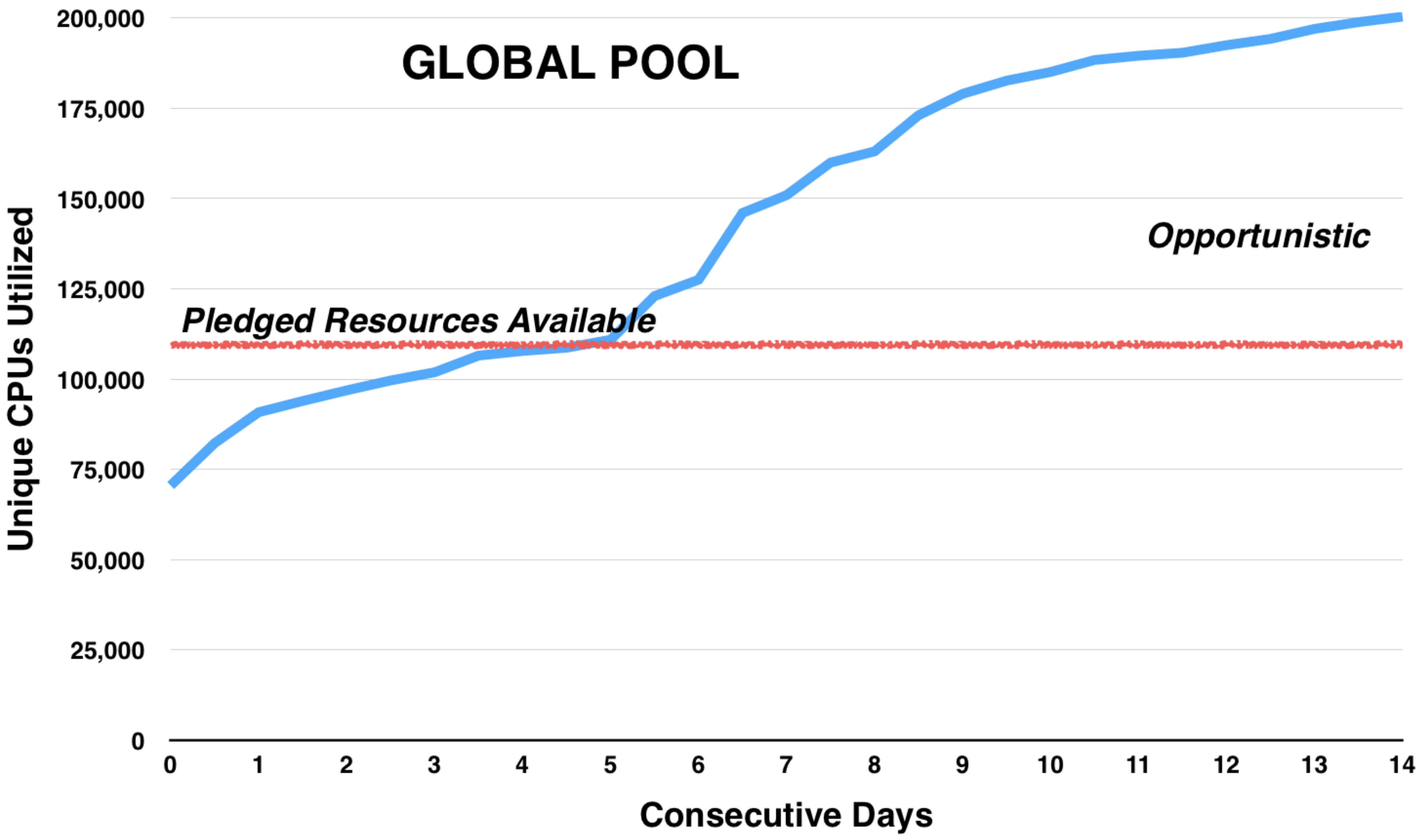
- We have deployed a single Global Pool based on HTCondor and glideinWMS which provides stable, flexible, scalable, and diverse resource provisioning to CMS for Run 2 of the LHC, backed by a strong operations team working under a written support model document, with close cooperation with the various software development teams.

- We plan to expand this model to cover new and different types of resources in the future, reaching ever higher scales.

Abstract

CMS will require access to more than 125k processor cores for the beginning of Run2 in 2015 to carry out its ambitious physics program with more and higher complexity events. During Run1 these resources were predominantly provided by a mix of grid sites and local batch resources. During the long shut down cloud infrastructures, diverse opportunistic resources and HPC supercomputing centers were made available to CMS, which further complicated the operations of the submission infrastructure. In this presentation we will discuss the CMS effort to adopt and deploy the glideinWMS system as a common resource provisioning layer to grid, cloud, local batch, and opportunistic resources and sites. We will address the challenges associated with integrating the various types of resources, the efficiency gains and simplifications associated with using a common resource provisioning layer, and discuss the solutions found. We will finish with an outlook of future plans for how CMS is moving forward on resource provisioning for more heterogenous architectures and services.

Backup Slide


GLOBAL POOL
Opportunistic
Pledged Resources Available
Unique CPUs Utilized
Consecutive Days