glideinWMS

SCD Projects Meeting

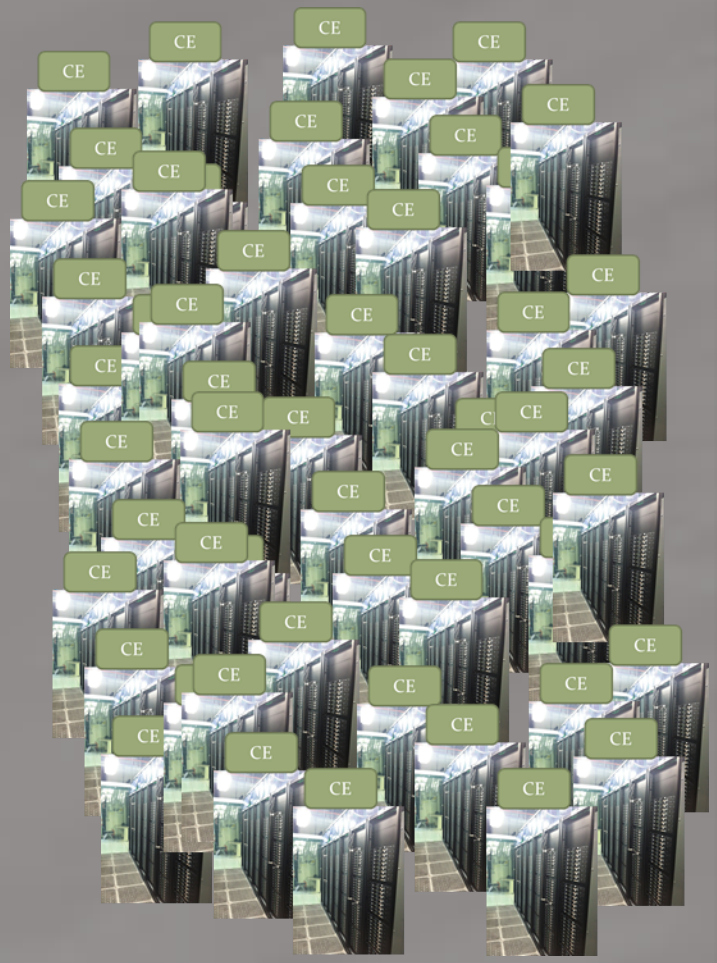
Burt Holzman

Challenges of Grid Computing: Distributed Compute Resources

3 Compute Elements are manageable By hand

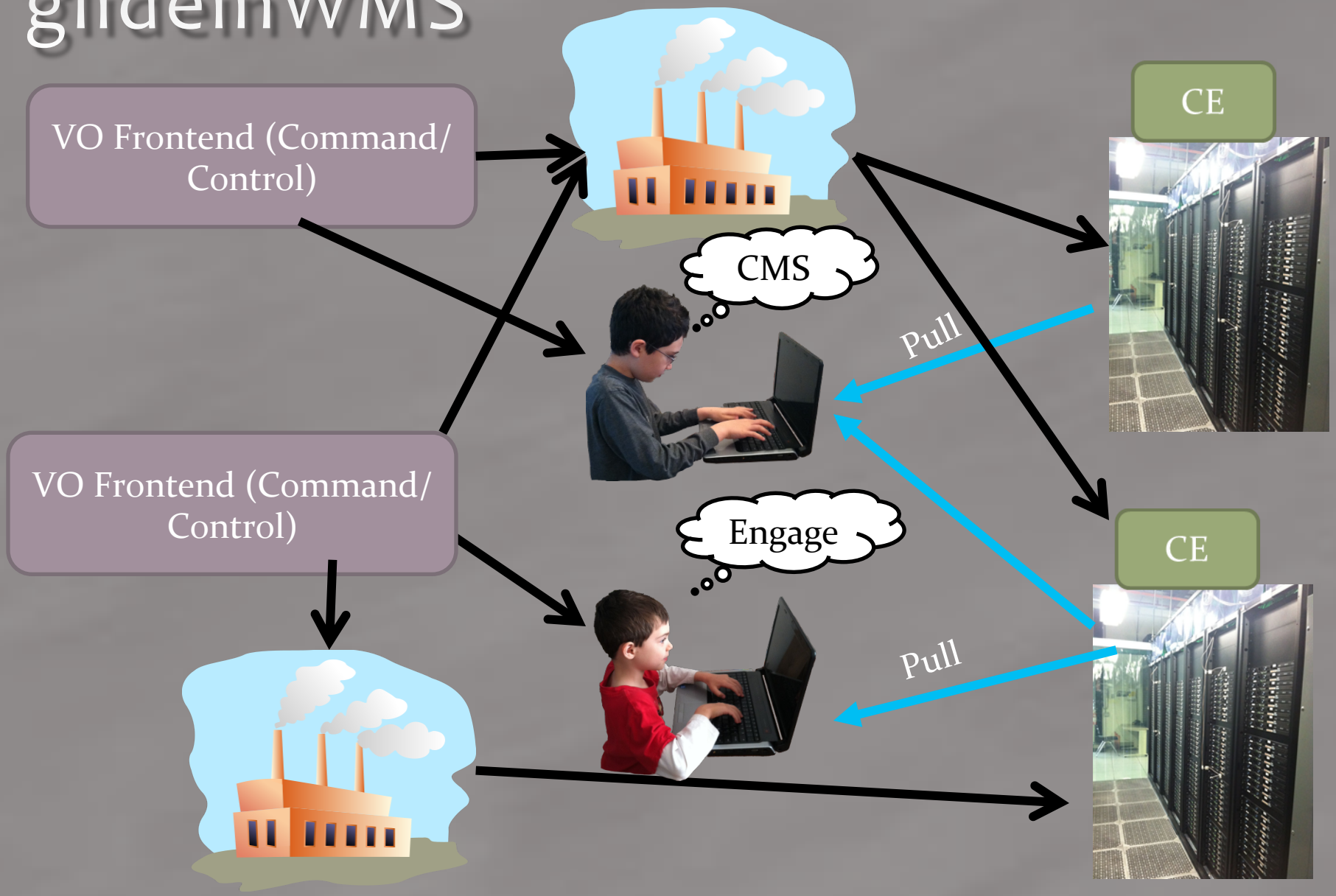
Challenges of Grid Computing: Distributed Compute Resources

?



We need middleware – specifically
a Workload Management System
(and more specifically, “glideinWMS”)

glideinWMS



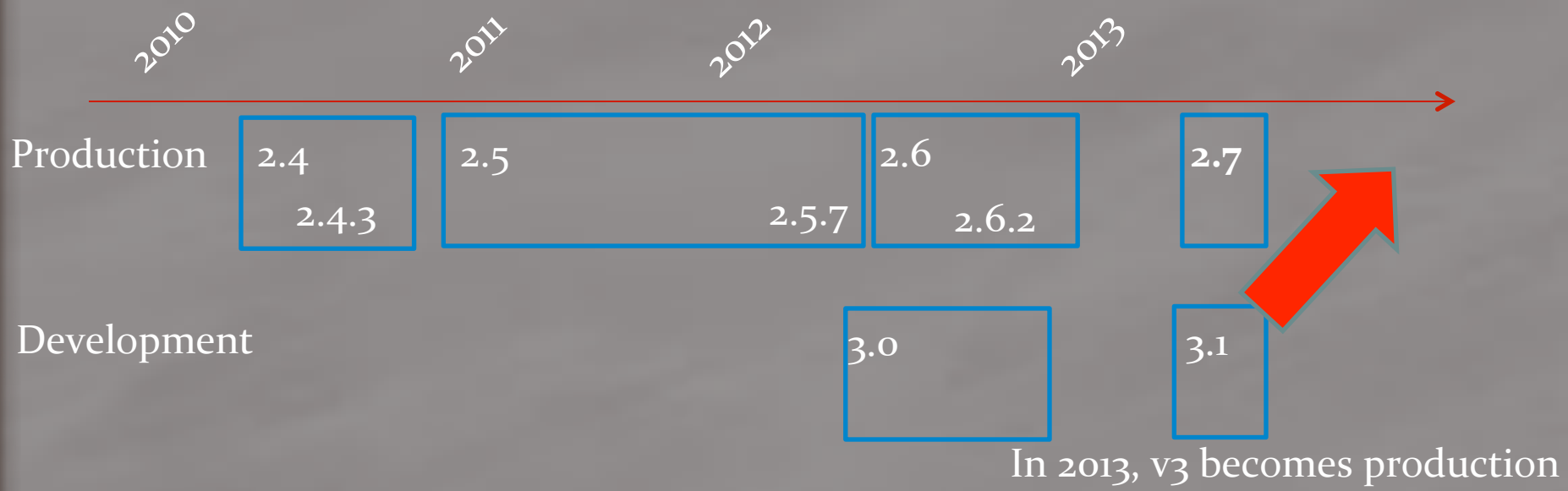
VO Frontend can talk to multiple factories

glideinWMS: quick facts

- glideinWMS is an open-source Fermilab Computing Sector product driven by CMS
- Heavy reliance on HTCondor from UW Madison and we work closely with them
- http://tinyurl.com/glideinWMS

- Contributors include:
  - Krista Larson (FNAL/Corral)
  - **Parag Mhashilkar (FNAL/Corral)**
  - Mats Rynge (ISI/USC/Corral)
  - Igor Sfiligoi (UCSD)
  - ~~Doug Strain (FNAL)~~
  - **Anthony Tiradani (FNAL/CMS)**
  - John Weigand (CMS)
  - Derek Weitzel (UNL)

glideinWMS: version timeline



- 2.4.x: privilege separation, aggregate monitoring, glexec control, glidein lifetime control
- 2.5.x: HTCondor TCP bulk updates, efficiency improvements, factory limits per frontend, excess glidein removal, shared ports, better user pool matchmaking
- 2.6.x: Better multislot support, ARC CE, more glidein lifetime controls, factory limits per frontend security class
- 2.7.x: Refactor for factory scaling, performance fixes, partitionable slot support

- 3.x: Cloud support, CorralWMS frontend support

glideinWMS: OSG tasks

- 2811: Plug-in architecture for frontends
  - glideinWMS configuration language is XML: add the ability to use XSLT
- 3203: condor_q –analyze analogue for glideins
  - VOs would like to know why their job didn't trigger a glidein request?
- 3204: http file transfer plugins in pilot
  - Input file transfer by default can be inefficient. It would be great to configure a local squid when possible

Roadmap for glideinWMS: 1/3

- Scalability
  - We can handle hundreds of destination CEs, but treat each permutation as a new CE
    - Refactored code from "one process per CE" **v2.7, v3.1**
  - We can better scale HTCondor components (separate factory collector, for example)
  - Work with HTCondor team on increasing HTCondor scalability (number of submitted jobs)

Roadmap for glideinWMS: 2/3

- Usability
  - Easier installation and configuration
    - We provide RPMs and tarballs, but it's far from push-button
  - More consistent monitoring and logging (**v3.1**)
  - Give more useful information in the hands of the VOs and users
    - APFMon interface, for example (**v3.1**)

Roadmap for glideinWMS: 2/3

Home Clouds Queues Faults Help

Panda Queue view - entry_CMSHTPC_T1_US_FNAL_ce - unknown (AGIS) (schedconfig)

Label factory created running exiting done miss fault when Message

Factory	job	state	payload?	created	last modified
burt-TEST-cms-xen43-FNAL	job.2533.0	CREATED	-	1 week ago	1 week ago
burt-TEST-cms-xen43-FNAL	job.2532.0	CREATED	-	1 week ago	1 week ago
burt-TEST-cms-xen43-FNAL	job.2531.0	CREATED	-	1 week ago	1 week ago

1 - 3 of 3

Most Visited Getting Started Latest Headlines Google InstanceInitiatedShutd... Google Calendar US CMS

Autopyfactory Monitoring

Home Clouds Queues Faults Help

FACTORY	burt-TEST-cms-xen43-FNAL
JOB_ID	job.2533.0 (stdout, stderr, stdlog, only recent logs)
PANDAQ	entry_CMSHTPC_T1_US_FNAL_ce
CREATED	2013-02-22 10:02:33 (1 week, 4 days ago)
LAST MODIFIED	2013-02-22 10:02:33 (1 week, 4 days ago)
STATE	CREATED
PAYLOAD?	-
PILOTCODE	-
FLAG	False

2013-02-22 10:02:33 131.225.204.217 CREATED

Roadmap for glideinWMS: 3/3

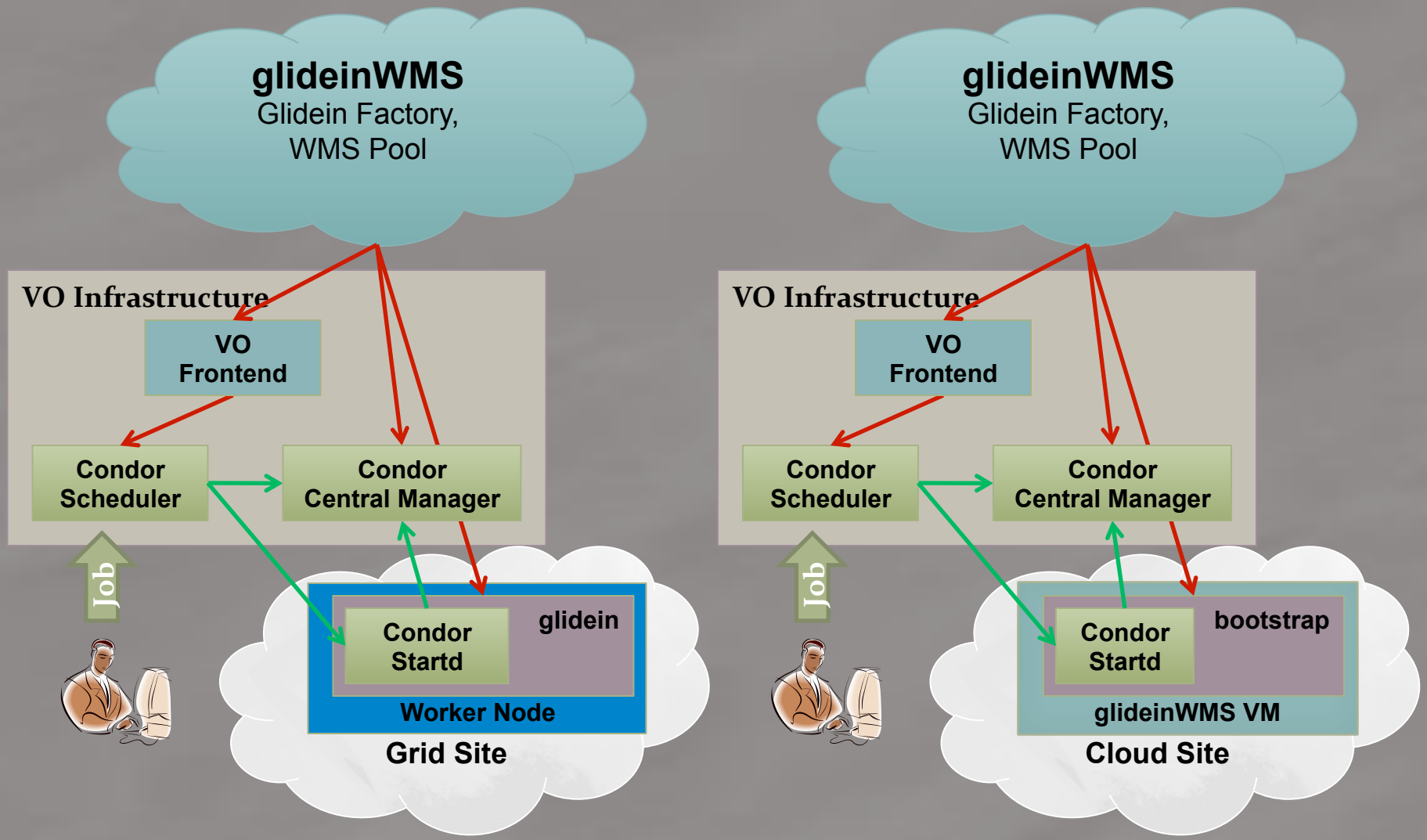
- Extendability: beyond the grid

- This works **now** in v3.0, but is not our production release yet
  - Different cost models lead to new needed features: in progress

glideinWMS and Cloud Cost model

- Running on Open Science Grid is "free"
- Running on Commercial clouds (e.g. Amazon EC2) requires a credit card
  - Now, how do we make sure that we get what we are paying for?
    - New credentials, new credential handling
    - Traditional match expressions to ensure jobs are matched correctly
  - How do we optimize usage of Cloud resources?
    - Multi-slot pilots
    - Options for Frontend or Factory to select Cloud resource "profiles" (e.g. AMI ID + AMI Type)

glideinWMS: Grid vs. Cloud



VM Images for glideinWMS

- Provide recipes for creating VM images for use with glideinWMS on the cloud
  - Using BoxGrinder
  - Using OZ (New RedHat sponsored tool, part of Aeolus)
- glideinWMS pilot bootstrap rpms destined for OSG-Contrib repository

OpenStack Compatibility 1/2

- OpenStack appears to be the leading candidate for private and Science clouds.
  - For HEP, CERN is moving all of their infrastructure to a private OpenStack cloud called AgileInfrastructure
- However, there are plenty of … "quirks"
  - OpenStack introduces a new terminal state: "SHUTOFF"
  - OpenStack now ignores the "InstanceInstantiatedShutdownBehavior" flag
  - Following comments appear in OpenStack code:

```
# Note(maoy): We do not provide EC2 compatibility
# in shutdown_terminate flag behavior. So we ignore
# it here.
```

OpenStack Compatibility 2/2

- glideinWMS team has contributed code to OpenStack to bring support for the EC2 Query API into better alignment with Amazon EC2 and other cloud projects
  - Patches accepted for next release
- The HTCondor team has been very responsive in addressing the issues as they come up
  - SHUTOFF and STOPPED VMs are now explicitly terminated by HTCondor
  - Status queries are now batched to reduce stress on the OpenStack controller
  - Cloud status now reported in the classad

CMS: glideinWMS on HLT

- Environment
  - CMS High Level Trigger (HLT) Cluster ~13K cores
  - GlideinWMS v3 development release
  - OpenStack Folsom Release
  - SLC5 VMs
    - Internal Frontier/Squid server
    - One VM image for all glideinWMS workers

CMS: glideinWMS on HLT (VM)

- The appeal of the Cloud is that you have complete control over your VM image.
  - You can customize it to your specific needs
- This is also the downside of the Cloud
  - You must maintain your image
- For CMS on HLT we:
  - Use CVMFS for the CMS software
    - We don't have to install CMSSW on the image
    - We don't have to update the image for every (frequent) release of CMSSW
  - Use Xrootd for both stage in and stage out of data
  - New images for security updates (e.g. kernel, etc.)

CMS: glideinWMS on HLT (Issues)

- Easy to overload OpenStack controller
  - Effectively, one can DOS the controller from one node
  - Feature request (Done): HTCondor should batch status queries into one call, rather than make a status call per VM
  - Use existing (in 7.9.x) rate limiting knobs
- XML generated using the –xml argument on condor_status changed the escaping "\" rules.
  - Extra slashes were being generated causing the embedded new line characters to be misinterpreted
  - Difficult to debug since the shell "fixes" things for you
  - Patched glideinWMS to handle any number of escape characters

CMS: glideinWMS on HLT

- Performance
  - Once a job lands on a pilot VM the performance of the VM is acceptable.
  - Only lose about 7% due to virtualization overhead
  - Ramping up short or small workflows can be inefficient
    - Initially, more VMs will be requested than necessary
  - Large or steady-state workflows will not see this issue
- Overall successful, nearing production capability

Acknowledgments

- HTCondor Team
  - Jaime Frey
  - Todd Miller
  - Todd Tannebaum
  - Timothy St. Clair

Questions?