



# Archival, anonymization and presentation of HTCondor logs with GlideinMonitor

Marco Mambelli, Fermilab

Thomas Hein, University of Illinois at Chicago

Mirica Yancey, Valparaiso University

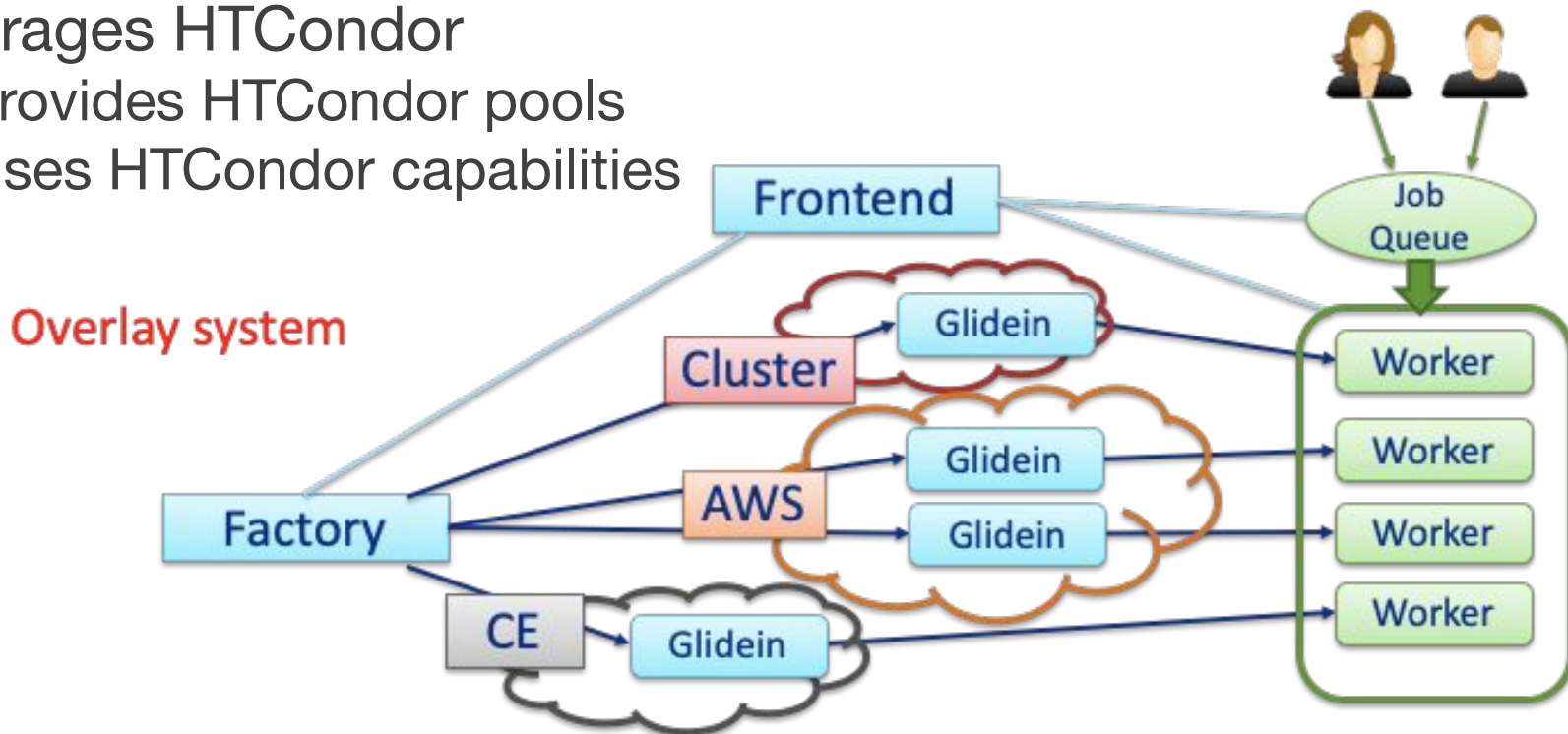
May 20, 2021

vCHEP 2021



# GlideinWMS

- GlideinWMS is a pilot based resource provisioning tool for distributed High Throughput Computing
- Provides reliable and uniform virtual clusters
- Submits Glideins to heterogeneous resources
- Leverages HTCondor
  - Provides HTCondor pools
  - Uses HTCondor capabilities

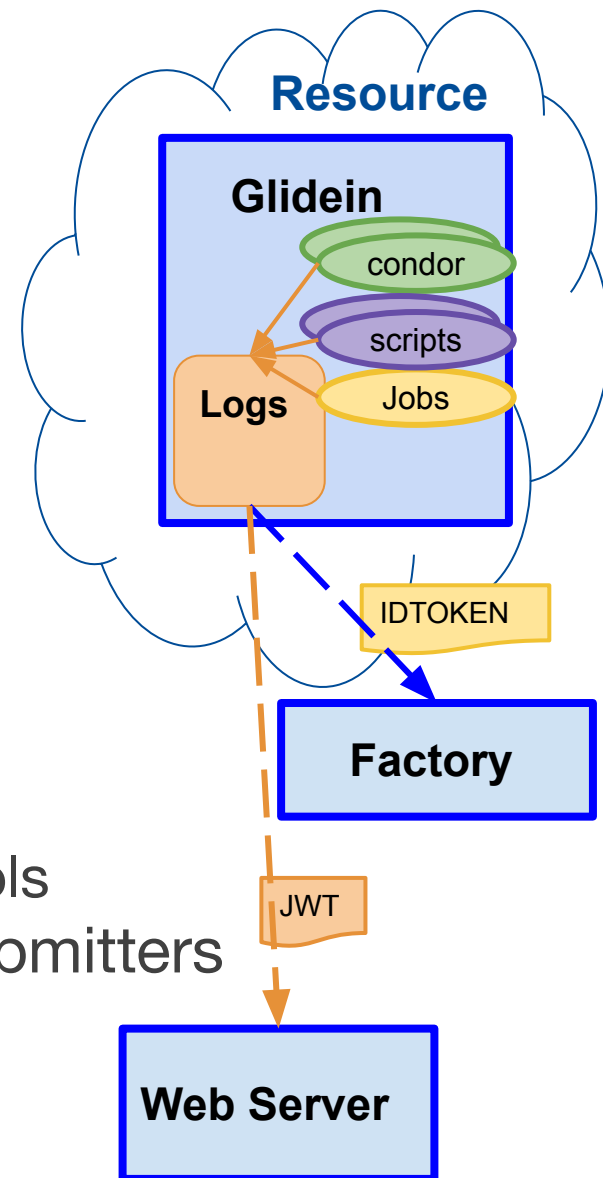


# Glidein: node testing and customization

- Scouts for resources and validates the Worker node
  - Cores, memory, disk, GPU, ...
  - OS, software installed
  - CVMFS
  - VO specific tests
- Customizes the Worker node
  - Environment, GPU libraries, ...
  - Starting containers (Singularity, ...)
  - VO specific setup
- Provides a reliable and customized execute node to HTCondor
- Reports back to the Factory

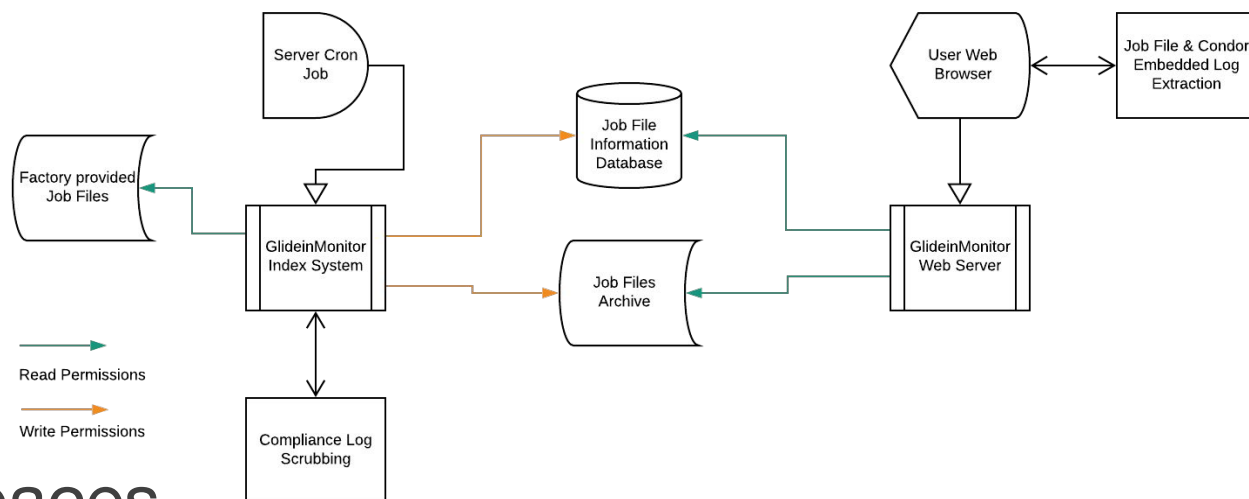
# Glidein Logging

- Very useful to troubleshoot
  - Resource
  - Jobs
  - GlideinWMS
- Authenticated channels
  - HTCondor stdout/err, WebDAV
- Factory and Monitoring servers
  - Available only to operators
  - Limited metadata
- Custom format
  - Semi-manual extraction with GWMS tools
- Sensitive information about jobs and submitters
  - User IDs
  - Email addresses
  - IP addresses



# GlideinMonitor

- Web application to examine, filter, archive and serve Glidein Logs
- Independent components
  - Indexer
  - Web server
  - Database
- File system spaces
  - Log ingestion
  - Archives (multiple log versions)



# GlideinMonitor Indexer

- Extracts metadata for processing and DB
- Has a Global IDs for each Glidein logs set
  - Can collect logs of Glideins from multiple Factories
  - Allows logs updates
  - Allows logs from via multiple channels
- Generates archives of compressed log bundles
  - Less space on disk
  - Faster to serve
- Supports log-filtering plugins
  - on raw logs
  - on expanded components
  - parallel processing
- Applies filters and saves logs to multiple archives
- Manages Logs lifetime

# Anonymization Plug-In

- GlideinMonitor filtering plug-in (python script)
- Acts on expanded Glidein logs
- Non-reversible (vs reversible) anonymization
  - Troubleshooting information is kept
  - Original log can be preserved
- Uses regular expressions (vs Named Entity Recognition)
  - Easier and quicker to implement
  - No need to categorize data
- In-memory stream processing

```
CRAB_SubmitterIpAddr = "XXXX:XXXX:XXXX:XXXX::127"  
CRAB_TaskEndTime = 1593944968  
CRAB_TaskLifetimeDays = 30  
CRAB_TaskWorker = "crab-prod-tw02"  
CRAB_TFileOutputFiles = { }  
CRAB_TransferOutputs = 1  
CRAB_UserDN = "/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=USER/CN=USER/CN=USER USER"  
CRAB_UserGroup = undefined  
CRAB_UserHN = "USER"
```



# GlideinMonitor Webserver

- GUI
  - search or browsing of Glideins
  - inspection of Glidein logs
- Configurable multi-user authorization for log archives access
  - Based on archive, Factory, VO
- Client-side scripts in Job View pages
  - log bundle download
  - logs expansion
  - lighter on the server

Job 559
Time: 2018-09-21T18:10:54-05:00

**General Information**

Timestamp	1537571454
FileSize (.err + .out)	68002
Entry Name	entry_ITB_FC_HTCE1
Instance Name	glidein_gfactory_instance
Frontend Username	user_frontend
GUID	user_frontend@glidein_gfactory_instance@entry_ITB_FC_HTCE1@job.1692.0

**Data Files**

Open

**Full Logs**

[559.out →](#)

[559.err →](#)

[559.tar.gz →](#)

[559.json →](#)

**Condor Logs**

[Master Log →](#)

[Startd Log →](#)

[Starter Log →](#)

[StardHist Log →](#)

[XML Description →](#)

**Log Search**

Output & Error Log Combined

condorg\_cluster = '1692'

1692

## Factory Monitoring Job View

Filters below alter the data in the table

Click search once you have narrowed the query

Timestamp From

05/14/2015 8:56 AM

Timestamp To

Entry Name

entry\_ITB\_FC\_CE2\_mc4, er ▼

entry\_HCC\_US\_Omaha\_crane\_gpu

entry\_ITB\_FC\_CE2

entry\_ITB\_FC\_CE2\_mc4 ✓

entry\_ITB\_FC\_HTCE1 ✓

entry\_Lucille\_CE ✓

Show  entries

JobID	FileSize	Timestamp	FrontendUsername	InstanceName	EntryName	MasterLog
<a href="#">job.7106.0</a>	15904	2019-02-09T22:28:51-06:00	user_frontend	glidein_gfactory_instance	entry_ITB_FC_CE2_mc4	False
<a href="#">job.7108.0</a>	15909	2019-02-09T22:29:51-06:00	user_frontend	glidein_gfactory_instance	entry_ITB_FC_CE2_mc4	False



# Summary

- GlideinMonitor, modular Web application to manage glidein logs
  - Indexer, Web server, Database
- Organizes the logs in an efficient compressed archive
- Allows to search, unpack, and inspect them
- Convenient and secure Web UI
- Custom filters via plug-ins
- The anonymization plug-in is an automated filter that locates and suppresses personal information
  - Makes the Log files easier to share

# Acknowledgements

This work was done under the GlideinWMS project and the TARGET and SIST internship programs at Fermilab

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

## References

<https://github.com/glideinWMS/glideinmonitor>