



Archival, anonymization and presentation of HTCondor logs with GlideinMonitor

Thomas Hein, University of Illinois at Chicago

Mirica Yancey, Valparaiso University

Marco Mambelli, Fermilab

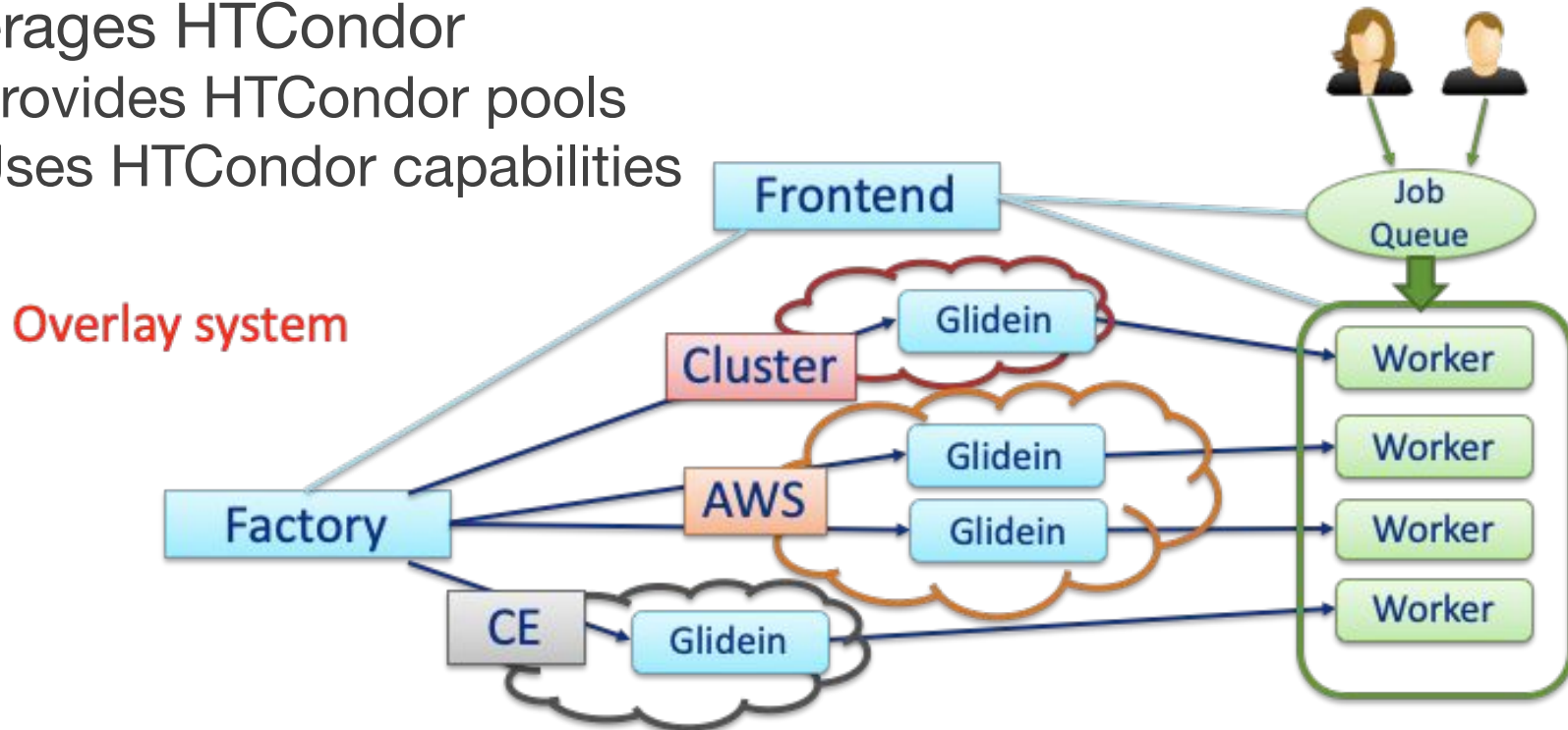
September 22, 2020

HTCondor Workshop Autumn 2020



GlideinWMS

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



Glidein: node testing and customization

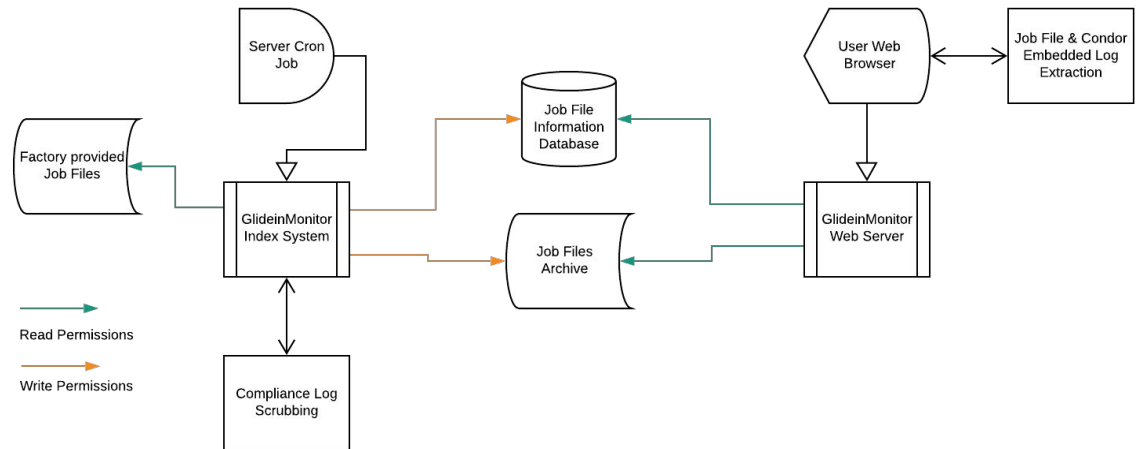
- Scouts for resources and validates the Worker node
 - Cores, memory, disk, GPU, ...
 - OS, software installed
 - CVMFS
 - VO specific tests
- Customizes the Worker node
 - Environment, GPU libraries, ...
 - Starting containers (Singularity, ...)
 - VO specific setup
- Provides a reliable and customized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

Log files

- Very useful to troubleshoot the whole system: Entries, Jobs, and GlideinWMS
- Sent back as Glideins' stdout and stderr
- Stored in a directory tree on the Factory
 - Available only to Factory operators
 - The only metadata are timestamp, Entry and HTCondor ID
- Custom format including compressed HTCondor logs
 - Semi-manual extraction with Factory tools
- Contain information about the jobs and the submitters that we want to restrict and may need to be protected by law
 - User IDs
 - Email addresses
 - IP addresses

GlideinMonitor Architecture

- Independent components
 - Indexer
 - Web server
 - Database



- File system spaces
 - Upload space
 - Archive (multiple versions)

GlideinMonitor Indexer

- Indexer parses & generates gzip archives of Factory Job Logs
 - Factory Job Logs are deleted periodically
 - Job Logs are in text format which is taxing on disk space over time
 - The archived versions that the Indexer generates significantly reduce the size of these logs
- Indexer generates Global IDs for each Job Log
 - Allows for multiple separate factories
- Indexer saves the Original and Filtered archives in long term storage directory

GlideinMonitor Webserver

- GlideinMonitor Webserver allows users to search for and view job logs
 - Search based on timestamp and entry name
 - Listing based on JobID, Timestamp, Entry Name, Frontend Username, Instance, and Availability of Condor Logs
- Webserver offers HTTP authorization for a simple login based system
 - Setup through the configuration file
 - Allows for multiple users with different levels of access
- Job View pages contain client-side scripting to download the compressed archive
 - The webpage itself decompresses both the archive and the Condor Logs within the Log Files
 - Removes load from the GlideinMonitor Webserver

GlideinMonitor Webserver Job Search

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 

Search

Copy Excel Print

Show 10 entries

JobID ↑↓	FileSize ↑↓	Timestamp ↑↓	FrontendUsername ↑↓	InstanceName ↑↓	EntryName ↑↓	MasterLog ↑↓
job.7106.0	15904	2019-02-09T22:28:51-06:00	user_frontend	glidein_gfactory_instance	entry_ITB_FC_CE2_mc4	False
job.7108.0	15909	2019-02-09T22:29:51-06:00	user_frontend	glidein_qfactory_instance	entry_ITB_FC_CE2_mc4	False

GlideinMonitor Webserver Job View

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

Anonymization Plug-In

Goal: To only leave information for statistics and troubleshooting

Result: A plug in for GlideinMonitor developed in python capable of locating user data using regex and using irreversible anonymization to suppress that data.

Anonymization Plug In: Research

- Reversible vs. Non Reversible Anonymization
 - Reversible
 - Allows for recovery of data
 - Protects data while maintaining data use
 - Non Reversible
 - Permanently changes data so it is unavailable for recovery
 - Insures the permanent loss of personal identifiers
- **Non Reversible**
 - Original log preserved
 - Removed information not needed for troubleshooting
- Named Entity Recognition vs. Regular Expressions
 - Named Entity Recognition
 - Recognizes data by predefined categories
 - Able to learn and be trained
 - Regular Expressions
 - Recognizes data by patterns
 - Separate language all together
 - Supported by most coding languages
- **Regular Expressions**
 - Easier and quicker to implement
 - No need to categorize data

Anonymization Plug In: Implementation (Take 1)

- Plug In opens the file and reads it line by line
- Uses a regex script to locate user data in that line
- Does an inline replacement of that data
- Writes the new line to the file.
- At end of all lines, closes the file

Cons:

- Long run time
- Can leave incomplete files
- Location regex too specific, making it inaccurate

```
192408 06/03/20 22:18:23 (pid:23220
192409 06/03/20 22:18:23 (pid:23220
192410 06/03/20 22:18:23 (pid:23220
192411 06/03/20 22:18:23 (pid:23220
192412 06/03/20 22:18:23 (pid:23220
192413
```

```
FAILED (failures=1)
PS C:\Users\miric\Documents\Important DOCS\Fermilab_PDFs\code> & C:/Users/
jost s-w ['AddTrust External CA Root has expired.\r\n\r\n06'
.
-----
Ran 1 test in 3.790s
```

Anonymization Plug In: Implementation (Take 2)

- Opens file
- Runs through file until it finds the replacement indicator
- Picks up the information from that line
- Strips it to the information we want
- Closes file
- Opens file again for actual anonymization
- Does general replacement

```
def findEmail(filename): #CONDOR SPECIFIC finds and returns user email
    lis = ''
    with open(filename, 'rb', 0) as file, \
        mmap.mmap(file.fileno(), 0, access=mmap.ACCESS_READ) as s:
        if s.find(b'x509UserProxyEmail') != -1:
            x = s.find(b'x509UserProxyEmail')
            end = s.find(b'@',x)
```

- Pros
 - Takes less time
 - Shorter, more contained scripts
 - Works in memory, writes the file at once
- Cons
 - Uses more memory

Anonymization Plug In: Testing + Integration

Factory Monitoring Job View

Filters below alter the data in the table
Click search once you have narrowed the query

Timestamp From: Timestamp To: Entry Name:

Show entries

JobID	FileSize	Timestamp	FrontendUsername	InstanceName	EntryName	MasterLog	StartFile
job.2682077.1	2033713	2020-05-31T04:16:10-05:00				True	True
job.2717803.12	220085	2020-06-03T05:08:55-05:00				True	True
job.2224810.3	1334646	2020-06-03T15:20:30-05:00				True	True
job.2224810.0	1902945	2020-06-03T15:21:48-05:00				True	True
job.2918326.1	339470	2020-06-06T10:54:43-05:00				True	True

Showing 1 to 5 of 5 entries

GWMS Factory Job Logs

Job 900

Time: 2019-07-20T23:40:15-05:00

General Information

Timestamp: 1902945

FileSize: 1902945

Entry Name: entry_TB_FC_C02

Instance Name: glomn_gfactory_inst000

Frontend Username: user_frontend

GUID: user_frontend@deisfactory_inst000@entry_TB_FC_C02@1902945-1902945

Data Files

Full Logs

900.out

900.err

900.start

900.join

Condor Logs

Master Log

Startd Log

Starters Log

Stand-Exec Log

XML Descriptions

Log Search

BRAN: 009: 7679 /bc-stdb/ELLS/ELLS.E.Leng/hom: ch04: 64730034/030963730034/0309: 62386/ra6/vorror: job: 04: entry_TB_FC_C02@1902945-1902945

```

CRAB_SubmitterIpAddress = "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"
    
```

Summary

- GlideinMonitor
 - 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>

<https://github.com/miricayancey2025/AnonymizeScript>