HTCondor at Fermilab

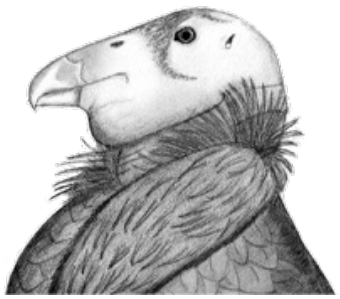
Farrukh Khan on behalf of
Compute Services Infrastructure group,
Computational Science & AI Directorate at Fermilab

July 13, 2023

Who are we?

- Fermi National Accelerator Laboratory, or simply Fermilab, is one of the leading centers for particle physics and accelerator research in the world
- Fermilab hosts many cutting-edge experiments and scientists doing research in quantum science, particle physics and dark matter/energy
- Research and technological advancements coming out of Fermilab has also supported physics research around the world in facilities like Sanford Underground Research Facility in South Dakota, the Large Hadron Collider in Switzerland and the South Pole Telescope.
- Fermilab is managed by Fermi Research Alliance LLC (FRA) for the U.S. Department of Energy Office of Science. FRA is a partnership of the University of Chicago and Universities Research Association Inc., a consortium of 89 research universities

HTCondor at Fermilab



HTCondor
Software Suite

HTCondor at Fermilab

(HT)Condor is KING!

HTCondor at Fermilab

- It wouldn't be wrong to say that HTCondor is THE most important component of our scientific computing infrastructure
- Other than scheduling workflows to enable science:
  - We have several tools built around it like GlideinWMS, JobSub, Decision Engine
  - One of Fermilab's flagship projects, HEPCloud, heavily utilizes HTCondor (candidate for understatement of the year..)
  - Our monitoring, fifemon, heavily relies on it and probes it for metrics – as Kevin Retzke likes to say, "there's a dashboard for that!"

HTCondor at Fermilab

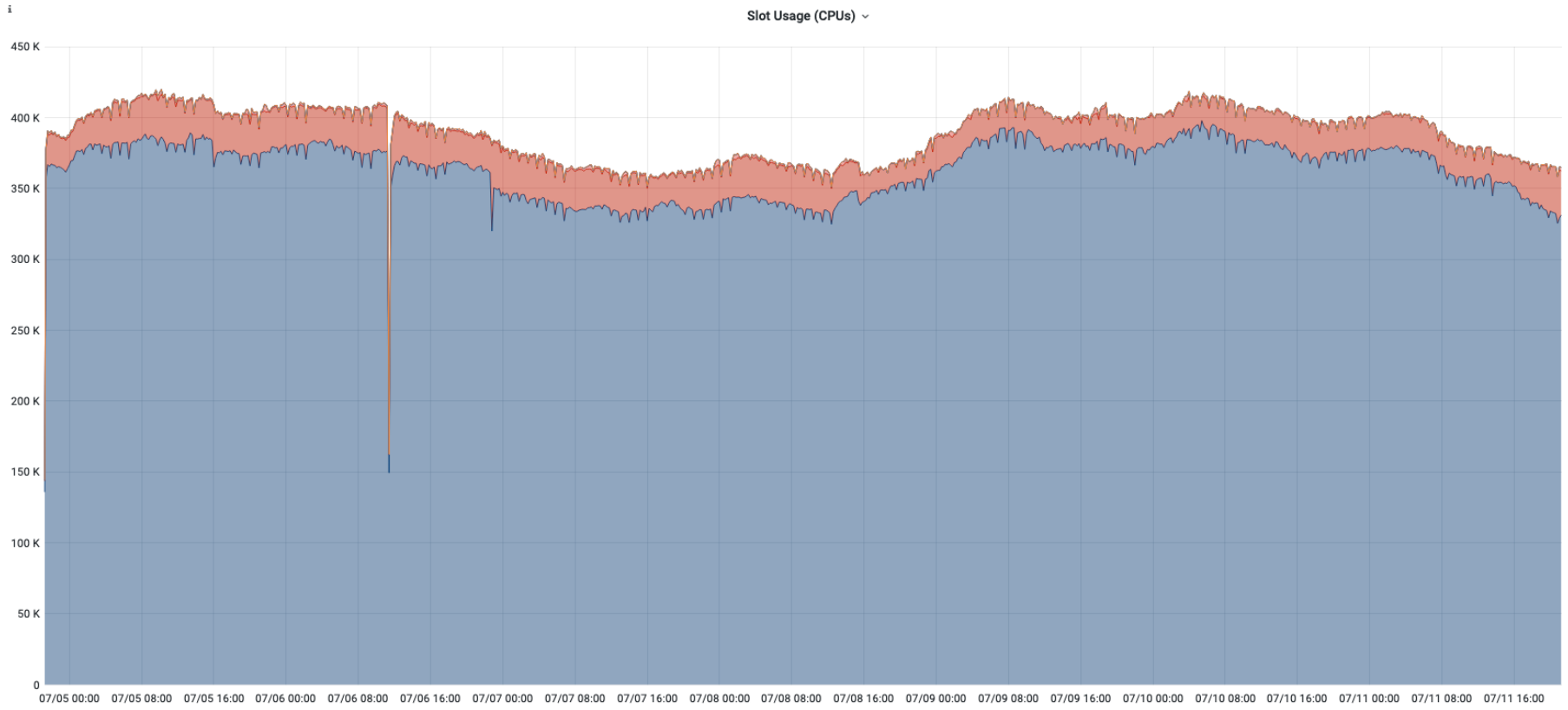
Excluding testbed and R&D pools, HTCondor is used in one form or another inside the following facilities/pools:

- CMS Tier1 and LHC Physics Center computing facility
- FermiGrid computing facility
- CMS Global pool
- DUNE Global pool
- Elastic Analysis Facility

CMS Global pool

- One pool to rule them all!
- CMS global pool is used to run CMS production and analysis workflows over the grid at collaborating institutes and sites
- Just like OS pool, it is a glideinWMS pilot based pool and utilizes HTCondor heavily
- We are part of the CMS submission infrastructure team and we operate half of the central services at Fermilab
- We also provide access points for CMS production team to run their workflows not only in the global pool but also at HPC centers in the U.S. through HEPCloud

CMS Global pool



DUNE Global pool

- Deep Underground Neutrino Experiment, or DUNE, is one of the flagship experiments at Fermilab
- This is a glideinWMS pilot based pool and was commissioned earlier this year so we are still dotting the Is and crossing the Ts
- All of the central submission infrastructure for the pool is operated at Fermilab by CSI group
- Access points for the pool are at Fermilab, Brookhaven National Lab and Rutherford Appleton Lab (so far)

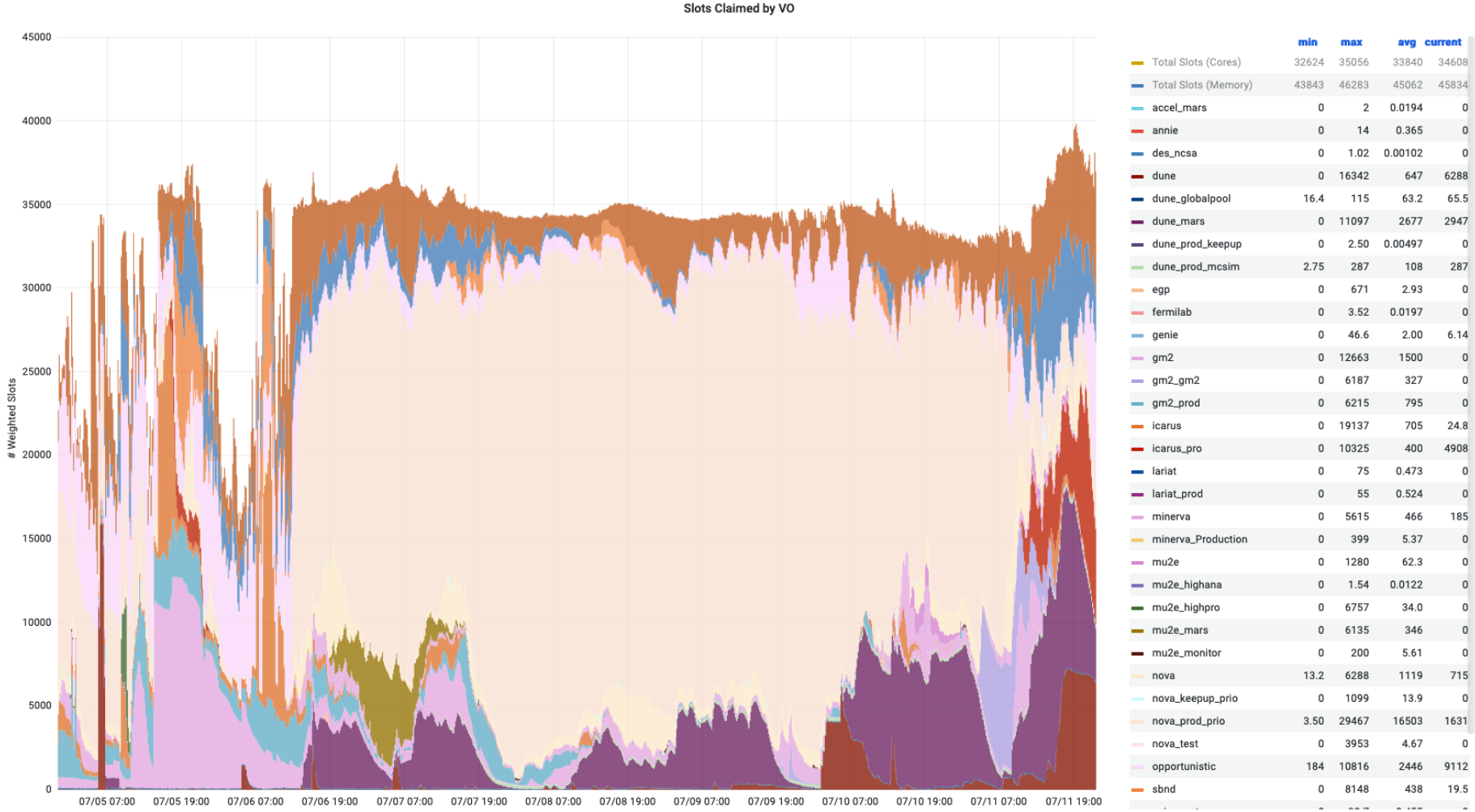
Elastic Analysis Facility

- A facility for the users!
- Led by Burt Holzman, Maria Acosta and Chris Bonnaud
- The facility is running several services to facilitate user analysis and is entirely containerized on the Fermilab OKD cluster
- It has interactive notebooks with HTCondor appropriately setup to allow users to submit their jobs to the on-premise facilities
- One of our goals is to also expand execute points into the OKD cluster on demand, we are not there yet.

FermiGrid compute facility

- FermiGrid compute facility is for all the experiments hosted at Fermilab. It is meant to provide a common interface to heterogenous computing resources at Fermilab and beyond.
- The facility is operated by collaborative effort of several teams in CS and AI directorate at Fermilab.
- This is an HTCondor facility with five access points, two compute entry points and around 500 execute points providing a total of 35000 cores.
- The pool has a glideinWMS frontend that allows it to tap into grid resources to expand the available resources
- The pool has a decision engine that allows experiments to run at HPC sites and cloud instances

FermiGrid compute facility

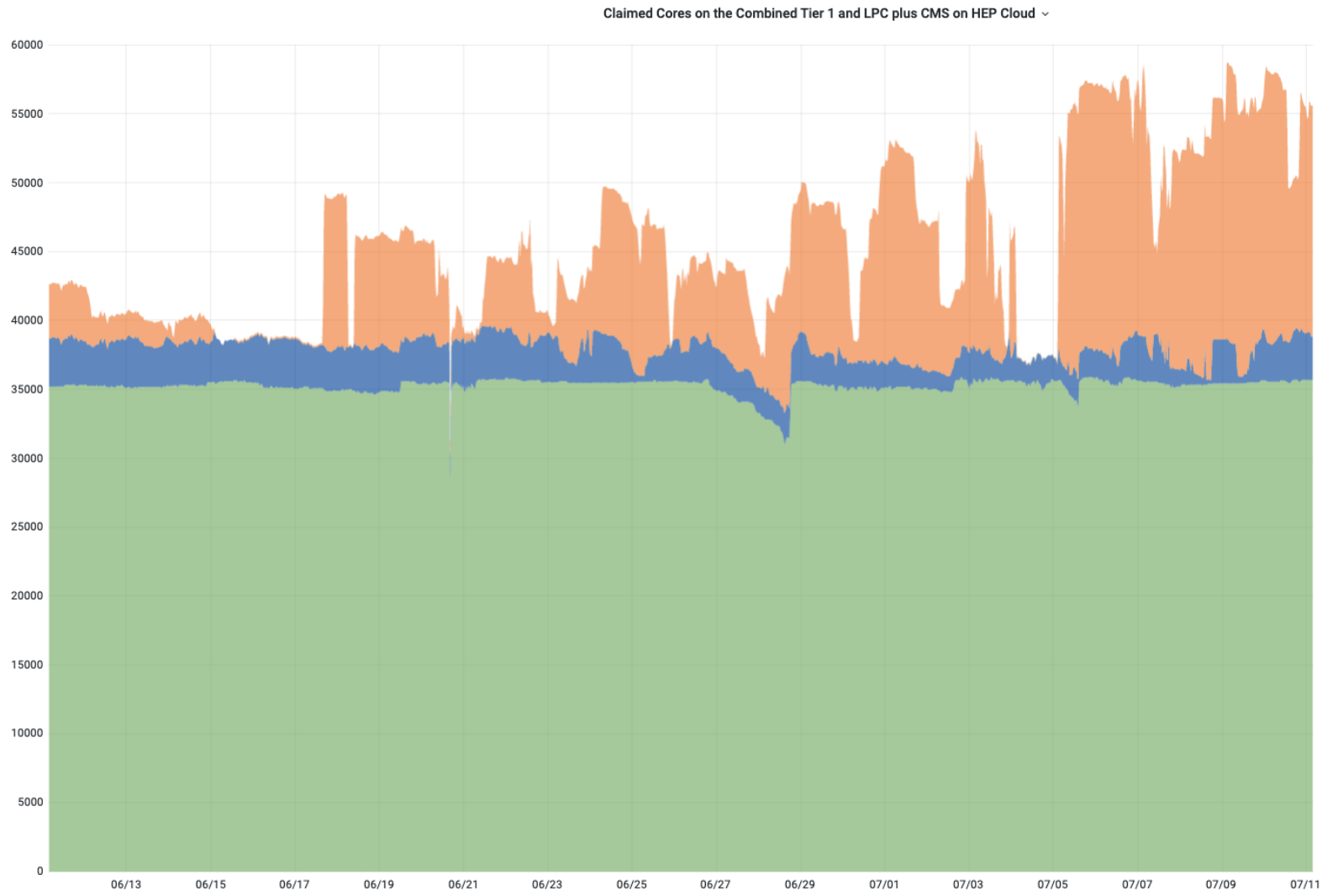


	min	max	avg	current
Total Slots (Cores)	32624	35056	33840	34608
Total Slots (Memory)	43843	46283	45062	45834
accel_mars	0	2	0.0194	0
annie	0	14	0.365	0
des_ncsa	0	1.02	0.00102	0
dune	0	16342	647	6288
dune_globalpool	16.4	115	63.2	65.5
dune_mars	0	11097	2677	2947
dune_prod_keepup	0	2.50	0.00497	0
dune_prod_mcsim	2.75	287	108	287
egp	0	671	2.93	0
fermilab	0	3.52	0.0197	0
genie	0	46.6	2.00	6.14
gm2	0	12663	1500	0
gm2_gm2	0	6187	327	0
gm2_prod	0	6215	795	0
icarus	0	19137	705	24.8
icarus_pro	0	10325	400	4908
lariat	0	75	0.473	0
lariat_prod	0	55	0.524	0
minerva	0	5615	466	185
minerva_Production	0	399	5.37	0
mu2e	0	1280	62.3	0
mu2e_highana	0	1.54	0.0122	0
mu2e_highpro	0	6757	34.0	0
mu2e_mars	0	6135	346	0
mu2e_monitor	0	200	5.61	0
nova	13.2	6288	1119	715
nova_keepup_prio	0	1099	13.9	0
nova_prod_prio	3.50	29467	16503	1631
nova_test	0	3953	4.67	0
opportunistic	184	10816	2446	9112
sbnd	0	8148	438	19.5

CMS Tier1 & LHC Physics Center facility

- CMS Tier1 & LHC Physics Center facility is primarily reserved for CMS usage
- The facility is virtually divided into three partitions; the CMS tier1 for CMS global pool, LPC for US CMS users at Fermilab and HPC/Cloud for CMS production
- We operate six compute entry points, nine access points and have around 500 execute points that provide a total of 40000 cores
- The tier1 facility at Fermilab is the largest compute resource provider to CMS experiment outside of CERN
- The facility also has a decision engine to tap into CMS allocations at HPC centers and into the cloud

CMS Tier1 & LHC Physics Center facility



Uses and abuses of HTCondor

- Docker
- Job Transforms
- Job Machine Attributes
- Worker node health checks
- Multiple negotiators
- HA defragmentation
- FS_REMOTE authentication
- JDL magic
- Job Router Hooks
- Virtual Slots
- Whole node scheduling
- Daemon health monitoring

Uses and abuses of HTCondor

- Docker
- Job Transforms
- Job Machine Attributes
- Worker node health checks
- Multiple negotiators
- HA defragmentation
- FS_REMOTE authentication
- JDL magic
- Job Router Hooks
- Virtual Slots
- Whole node scheduling
- Daemon health monitoring

HTCondor at Fermilab

- Did you know that the DOCKER knob in HTCondor doesn't need to point to the docker binary?
- Did you know that HTCondor allows you add as many auxiliary knobs as you want?
- We use/abuse this at Fermilab to get greater control over our containers!

```
# Docker related configurations start from here onwards
# Setting docker path
# DOCKER = /usr/bin/docker
#
# The default path is commented out to run docker from
# a wrapper script. This allows us to set certain
# capabilities to the container if/when needed
# (Farrukh 2017-08-10))
DOCKER = /usr/local/libexec/condor-docker.py
```

HTCondor at Fermilab

```
FERMIHTC_APPTAINER_BIN =  /cvmfs/oasis.opensciencegrid.org/mis/apptainer/current/bin/apptainer

FERMIHTC_APPTAINER_ARGS = --pid --ipc --contain

FERMIHTC_APPTAINER_PWD = /srv

FERMIHTC_DOCKER_CAPABLE = True
STARTD_ATTRS = $(STARTD_ATTRS) FERMIHTC_DOCKER_CAPABLE

FERMIHTC_DOCKER_ENV = [ \
   GLIDEIN_Site=FermiGrid; \
   GLIDEIN_DUNESite=US_FNAL-FermiGrid; \
]

FERMIHTC_DOCKER_PULL = True
```

HTCondor at Fermilab

```
#
# Comma or space separated list of capabilities to be enabled
# inside the containers listed under FERMIHTC_DOCKER_TRUSTED_IMAGES.
# These capabilities are NOT enabled by default for all images.
# List of all kernel capabilities can be seen here:
# http://man7.org/linux/man-pages/man7/capabilities.7.html
#
# These values are taken as is from the script forwarded by Nebraska
# admins for signularity. If undefined, no capabilities are added to the
# containers
FERMIHTC_DOCKER_CAPABILITIES = SYS_ADMIN
#
```

```
FERMIHTC_DOCKER_TRUSTED_IMAGES = 'ssiregistry.fnal.gov/ecf-gco/cmst1-slf7:production', 'ssiregistry.fnal.gov/ecf-gco/cmslpc-sl7:production','ssiregistry.fnal.gov/ecf-gco/cmst1-slf7:test', 'ssiregistry.fnal.gov/ecf-gco/cmslpc-sl7:test', 'ssiregistry.fnal.gov/ecf-gco/gpgrid-sl7:production','ssiregistry.fnal.gov/ecf-gco/gpgrid-sl7:test'
```

HTCondor at Fermilab

- Did you know that HTCondor has job transforms that allow administrators to transform jobs at submit time?
- Did you know job machine attribute feature lets you pull in any classAd from the worker node into your job?

```
# To make sure a job runs inside Docker
#
# To fetch FERMIHTC_DOCKER_CAPABLE attribute from the STARTD into the job
# This is used to determine whether the startd supports Docker or not
#
SYSTEM_JOB_MACHINE_ATTRS = $(SYSTEM_JOB_MACHINE_ATTRS) FERMIHTC_DOCKER_CAPABLE
#
# Job transform to set relevant Docker attributes
#
JOB_TRANSFORM_NAMES = FERMIHTC_Docker
JOB_TRANSFORM_FERMIHTC_Docker = \
  [ \
   set_WantDocker = isUndefined(MachineAttrFERMIHTC_DOCKER_CAPABLE0) ? FALSE : \
                                    MachineAttrFERMIHTC_DOCKER_CAPABLE0; \
   set_DockerImage = "ssiregistry.fnal.gov/ecf-gco/cmslpc-sl7:production"; \
  ]
```

Uses and abuses of HTCondor

- Docker
- Job Transforms
- Job Machine Attributes
- Worker node health checks
- Multiple negotiators
- HA defragmentation
- FS_REMOTE authentication
- JDL magic
- Job Router Hooks
- Virtual Slots
- Whole node scheduling
- Daemon health monitoring

Summary

- Fermilab relies heavily on HTCondor. We not only use it as a workload scheduler but also have several tools built around it
- We absolutely love the flexibility HTCondor gives us as pool administrators
- We also wanted to give a shout out to the HTCondor development team. You guys are a huge part of making our computing systems and services a success!

HTCondor at Fermilab

If anything I talked about felt interesting and if you are in the market for a new career opportunity, don't hesitate to come talk to me!

We are hiring for multiple positions at Fermilab:
https://fermilab.wd5.myworkdayjobs.com/en-US/FermilabCareers

Questions? Comments?