

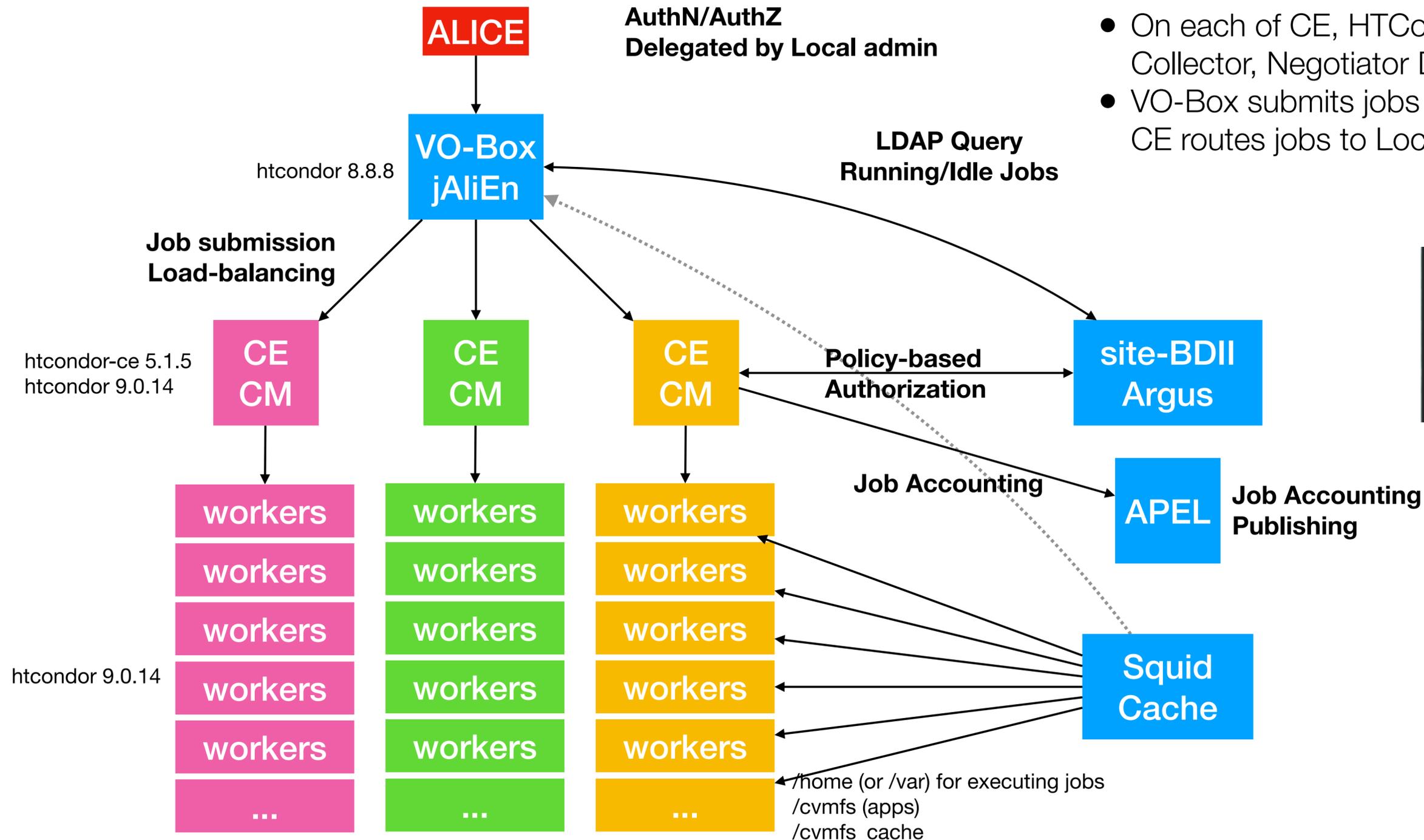
KISTI HTCCondor Experiences

Geonmo Ryu, Sangwook Bae, Sang-Un Ahn

Introduction

- HTCondor/HTCondor-CE configuration at KISTI-GSDC
 - ALICE Tier-1/Tier-3, CMS Tier-2/Tier-3, LDG Tier-2/Tier-3 and other local services share common HTCondor configuration
 - HTCondor-CE configurations should be experiment-specific
- Integrated farm configuration
 - A shared pool among ALICE Tier-3, CMS Tier-3 and ARGO (Genomic Research) at KISTI-GSDC with some specific configurations such as accounting group quota allowing surplus, preemption policy and singularity(apptainer)

ALICE Tier-1 HTCondor Batch System



- On each of CE, HTCondor-CE & HTCondor Schedd, Collector, Negotiator Daemons (CM Role) are running
- VO-Box submits jobs to HTCondor-CE then HTCondor-CE routes jobs to Local (HTCondor) Queue

HTCondor-CE related packages

```
[root@alice-t1-ce06 ~]# rpm -qa htcondor\*
htcondor-ce-5.1.5-1.el7.noarch
htcondor-ce-condor-5.1.5-1.el7.noarch
htcondor-release-9.0-2.el7.noarch
htcondor-ce-client-5.1.5-1.el7.noarch
htcondor-ce-bdii-5.1.1-1.el7.noarch
htcondor-ce-apel-5.1.5-1.el7.noarch
```

HTCondor packages (on CE)

```
[root@alice-t1-ce06 ~]# rpm -qa condor\*
condor-procd-9.0.14-1.el7.x86_64
condor-9.0.14-1.el7.x86_64
condor-boinc-7.16.16-1.el7.x86_64
condor-externals-9.0.14-1.el7.x86_64
condor-classads-9.0.14-1.el7.x86_64
```

HTCondor packages (on WN)

```
[root@awn1052 ~]# rpm -qa condor\*
condor-procd-9.0.14-1.el7.x86_64
condor-boinc-7.16.16-1.el7.x86_64
condor-externals-9.0.14-1.el7.x86_64
condor-9.0.14-1.el7.x86_64
condor-classads-9.0.14-1.el7.x86_64
```

3 Different Clusters with different specification/vendors and years

Essential Configuration for HTCCondor(LRMS)

Common configuration for all (SN/CM/EN)

```
[root@alice-t1-ce06 ~]# cat /etc/condor/config.d/cluster.conf
#LOCAL_CONFIG_FILE = /nfs/condor/condor-etc/condor_config.$(HOSTNAME)
UID_DOMAIN = sdfarm.kr
COLLECTOR_NAME = "ALICE T1 HTCCondor"
FILESYSTEM_DOMAIN = sdfarm.kr
ALLOW_WRITE = *.sdfarm.kr
ALLOW_READ = *.sdfarm.kr
CONDOR_ADMIN = sahn@kisti.re.kr
CONDOR_HOST = alice-t1-ce06.sdfarm.kr
IN_HIGHPORT = 13999
IN_LOWPORT = 9000
SEC_DAEMON_AUTHENTICATION = required
SEC_DAEMON_AUTHENTICATION_METHODS = password
SEC_CLIENT_AUTHENTICATION_METHODS = password,fs,gsi
SEC_PASSWORD_FILE = /var/lib/condor/condor_credential
ALLOW_DAEMON = condor_pool@*
NEGOTIATOR_INTERVAL = 20
TRUST_UID_DOMAIN = TRUE
START = TRUE
SUSPEND = FALSE
PREEMPT = FALSE
KILL = FALSE
REQUIRE_LOCAL_CONFIG_FILE = False
STARTD_ATTRS = $(STARTD_ATTRS) GSDCScaling
```

Local configuration to define roles (e.g. SN+CM)

```
[root@alice-t1-ce06 ~]# cat /etc/condor/config.d/local.conf
DAEMON_LIST = MASTER, SCHEDD, COLLECTOR, NEGOTIATOR
```

Submission & Central Manager

Typical worker node (EN) configuration

```
[root@awn1052 ~]# cat /etc/condor/config.d/local.conf
DAEMON_LIST = MASTER, STARTD
```

Execution jobs

- Authentication among HTCCondor nodes via a common secret (host_based)
- From HTCCondor v9.0, host_based authentication is not recommended
- Keep this because we have upgraded from the legacy HTCCondor system (v8)

Scaling factor defined in Workers

```
[root@awn1052 ~]# cat /etc/condor/config.d/00-node_parameters
GSDCScaling = 1.016
```

- Scaling factor for Job accounting (APEL)
- Multiplying to HepSpec06 benchmark assumed baseline score (probably 10HS06/core)

Configurations for HTCondor-CE (1/3)

Job routes for submission to LRMS

Job routes configuration

```
[root@alice-t1-ce06 ~]# cat /etc/condor-ce/config.d/61-job-routes.conf
#####
# Example Job Route
#
# This is an extraordinarily simple job route.
# All it does is route local condor and set a
# simple Accounting Group and default RequestMemory.
#####

# No custom functions for job router entries; these are causing crashes in 8.3.5.
# Can remove the eval_set_environment attribute below starting in 8.3.8.

# Modified 31 Mar 2020 thanks to Brian Lin
JOB_ROUTER_ENTRIES = \
[ \
#eval_set_environment = debug( mergeEnvironment(join(" ", "HOME=/tmp", \
eval_set_environment = debug( mergeEnvironment(join(" ", strcat("HOME=", userHome(Owner, "/")), \
    ifThenElse($(DISABLE_PILOT_ADS) =?= True, "", strcat("CONDORCE_COLLECTOR_HOST=", "$(COLLECTOR_HOST)")), osg_environment, \
    orig_environment, $(CONDORCE_PILOT_JOB_ENV), default_pilot_job_env)); \
TargetUniverse = 5; \
name = "Local_Condor"; \
eval_set_AccountingGroup = strcat("group_u_", x509userproxyvname, ".", Owner); \
delete_SUBMIT_lwd = true; \
set_WantIOProxy = true; \
set_default_maxMemory = 3000; \
]
```

- Mostly the default configuration shipped with HTCondor-CE package will work
- Job routes should be customized to fit with local batch system (default: HTCondor)
- In case that you need any help related to HTCondor-CE, please consult experts via <htcondor-users@cs.wisc.edu>

- Not to use '/tmp' for condor home (to execute job), instead to use owner's home
- After upgrading HTCondor-CE from v3 to v4, a deprecated syntax in v3 was removed (thanks to Brian Lin)



Configurations for HTCondor-CE (2/3)

AuthN/AuthZ based on GSI Mapping

GSI Mapping

```
[root@alice-t1-ce06 ~]# cat /etc/condor-ce/mapfiles.d/10-gsi.conf
#####
#
# HTCondor-CE manual GSI/VOMS authentication mappings
#
# This file will NOT be overwritten upon RPM upgrade.
#
#####
...
GSI "^VDC=chVDC=cernVOU=computersVCN=?([A-Za-z0-9.\-]*)$" \1@cern.ch
GSI "^VC\=KRVO\=KISTIVO\=KISTIVCN\=(hostV)?([A-Za-z0-9.\-]*)$" \2@sdfarm.kr
```

- If you use GSI authentication (for now), then HTCondor-CE should have a X509-based host certificate and a pair of certificate must be located in /etc/grid-security/
- To authenticate and authorize the submitter's identity and role, HTCondor-CE relies on GRIDMAP and the process is performed via ARGUS callouts

GSI callouts

```
[root@alice-t1-ce06 ~]# cat /etc/condor-ce/mapfiles.d/50-gsi-callout.conf
#####
#
# HTCondor-CE authentication mapping for GSI callouts
#
# This file will NOT be overwritten upon RPM upgrade.
#
#####
# The special token GSS_ASSIST_GRIDMAP indicates one should use the Globus Toolkit
# callout mechanism (which may involve plugins such as LCMAPS or Argus).
# Comment this out if you are not using a Globus Toolkit callout for mappings
GSI/(.*)/GSS_ASSIST_GRIDMAP
```

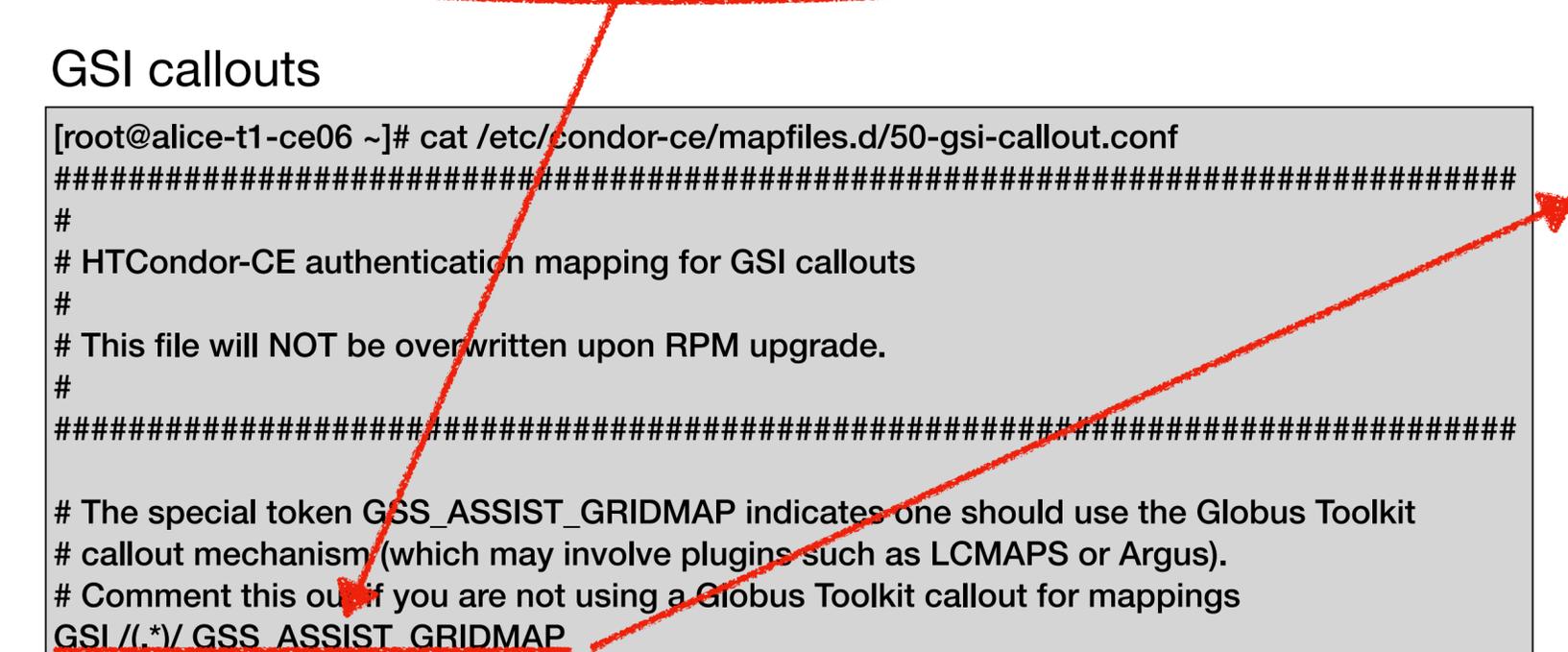
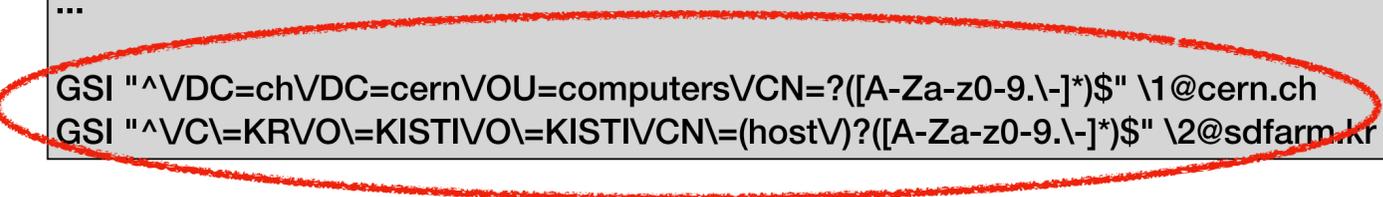
GSI mapping done through ARGUS callout

```
[root@alice-t1-ce06 ~]# cat /etc/grid-security/gsi-authz.conf
globus_mapping /usr/lib64/libgsi_peg_callout.so argus_peg_callout
```

ARGUS instance information

```
[root@alice-t1-ce06 ~]# cat /etc/grid-security/gsi-peg-callout-condor.conf
pep_ssl_server_cpath /etc/grid-security/certificates/
pep_ssl_client_cert /etc/grid-security/condorcet.pem
pep_ssl_client_key /etc/grid-security/condorkey.pem
pep_url https://alice-t1-sbdii02.sdfarm.kr:8154/authz
pep_timeout 30 # seconds
xacml_resourceid http://alice-t1-ce06.sdfarm.kr/htcondor-ce
```

Certs for SSL communication
Argus PEP endpoint
Argus Resource ID for CE



Configurations for HTCondor-CE (3/3)

Information Service & Accounting

site-BDII

```
[root@alice-t1-ce06 ~]# cat /etc/condor-ce/config.d/06-ce-bdii.conf
#####
#
# HTCondor-CE BDII/GLUE Publication configuration file.
#
#####

# For multi-CE sites, only one CE publishes certain values.
HTCONDORCE_BDII_ELECTION = LEADER
HTCONDORCE_BDII_LEADER = alice-t1-ce04.sdfarm.kr

# BDII Static Info and VOs
HTCONDORCE_VONames = alice, dteam, ops
HTCONDORCE_SiteName = KR-KISTI-GSDC-01
HTCONDORCE_HEPSPEC_INFO = 10.13-HEP-SPEC06
HTCONDORCE_CORES = 40 # cores per node
```

- Note that not all configurations are presented here, some essentials might be missing
- For HTCondor/HTCondor-CE, referring to HTCondor Manual or HTCondor-CE documentation will further help

- Additional configurations to work with site-BDII (Information System) & APEL (Job Accounting)
- CE should run BDII and APEL parser services
- Note that site-BDII and APEL are independent services (They can run on CE hosts for sure)

APEL

```
[root@alice-t1-ce06 ~]# cat /etc/condor-ce/config.d/50-ce-apel.conf
#####
#
# HTCondor-CE APEL configuration file.
#
#####

APEL_CE_HOST = $(CONDOR_HOST)
APEL_BATCH_HOST = $(CONDOR_HOST)

# Directory to write batch and blah records
APEL_OUTPUT_DIR = /var/lib/condor-ce/apel/

APEL_CE_ID = $(APEL_CE_HOST):$(PORT)/$(APEL_BATCH_HOST)-condor

# If False, skip 'apelclient' and 'ssmsend' in the cron script
# because they are executed on the site's central publisher
# (True by default)
#APEL_SEND_RECORDS = True

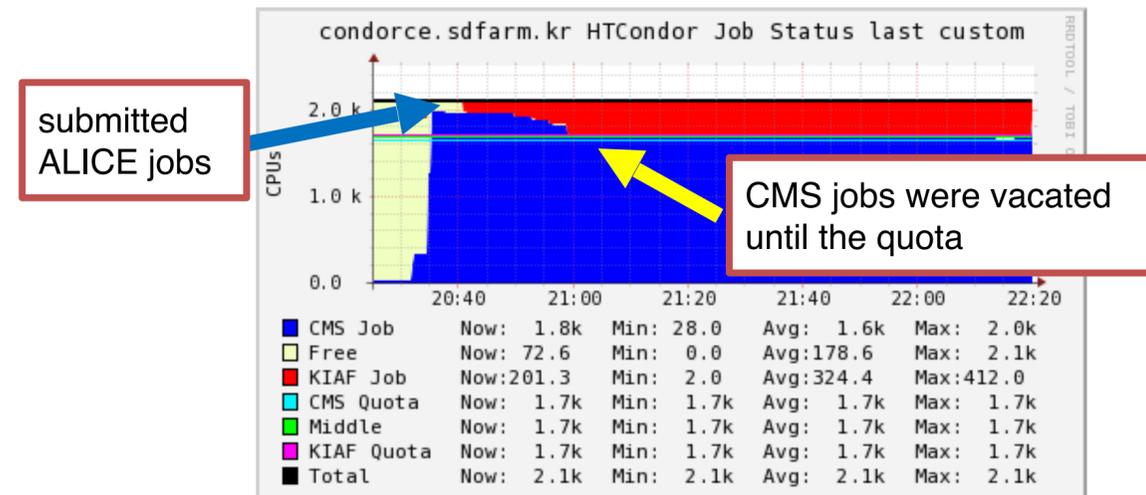
APEL_SCALING_ATTR = MachineAttrGSDCScaling0
```

ALICE Specific Configuration

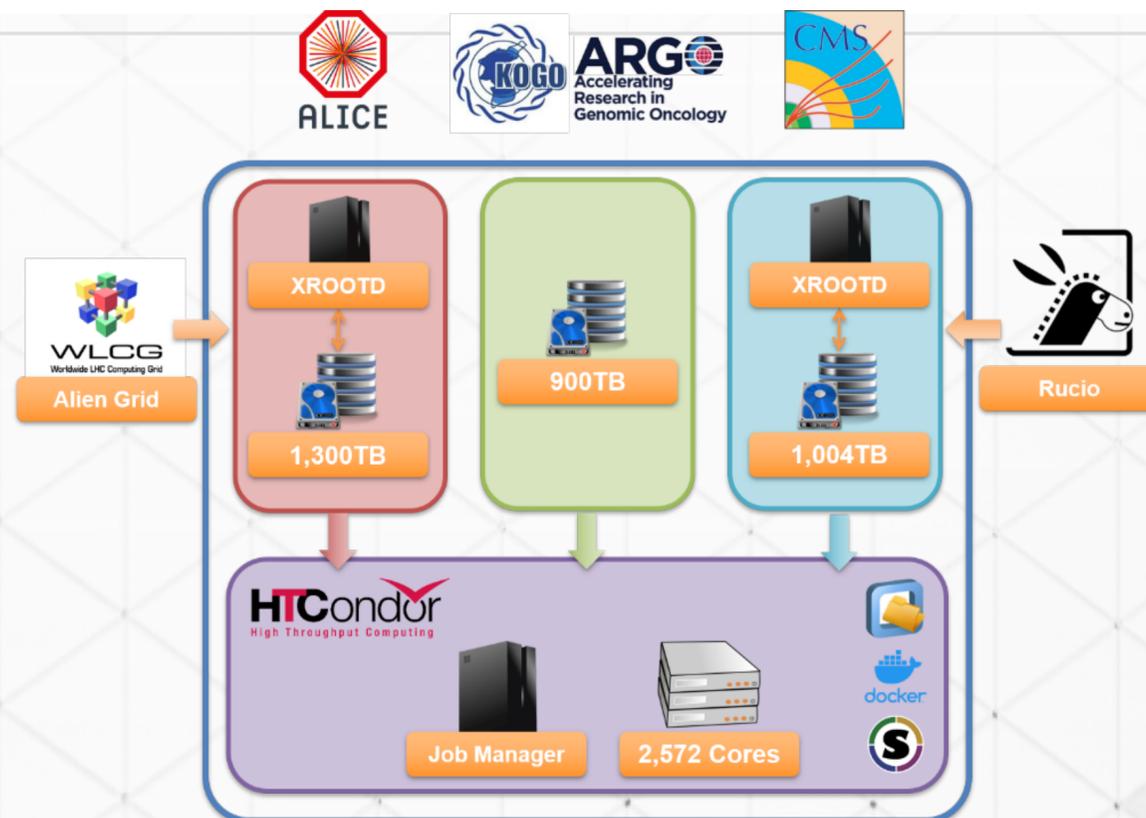
- Once HTCondor-CE is configured successfully,
 - How to validate a HTCondor-CE:
<https://htcondor.com/htcondor-ce/v5/operation/>
- Since ALICE uses HTCondor to submit jobs into HTCondor-CE enabled LRMS, one should configure HTCondor-based AliEn on VO-Box
 - HTCondor-based AliEn Installation on VO-Box:
<https://alien.web.cern.ch/content/documentation/howto/site/htcondor-based-alien-site-installation>
 - If one might need to contact ALICE experts, Maarten Litmaath

- Introduction
 - An analysis farm that can share and use resources among different research groups
 - Supporting quota and withdraw(preemption) of resource by using HTCondor

Preemption

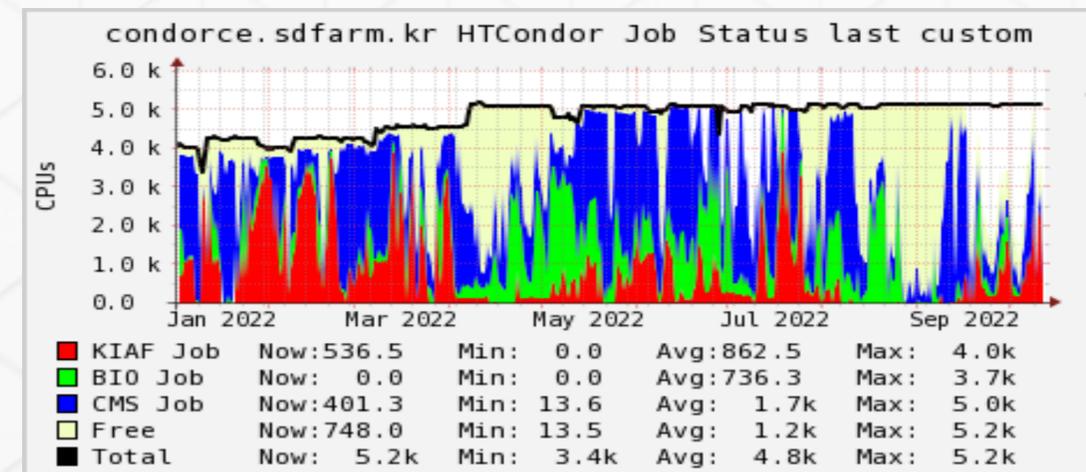


GSDC Integration Farm



Job Processed

	2018	2019	2020	2021
KoALICE	450,837	776,554	1,977,302	6,284,506
BIO	20,683	23,774	98,668	2,766,617
KCMS	5,966,699	3,429,751	3,873,006	9,163,872



Negotiator

```

NEGOTIATOR_CONSIDER_PREEMPTION = True
PREEMPTION_REQUIREMENTS = ((SubmitterGroupResourcesInUse < SubmitterGroupQuota) && (RemoteGroupResourcesInUse > RemoteGroupQuota))
PREEMPTION_RANK = 2592000 - ifThenElse(isUndefined(TotalJobRuntime),0,TotalJobRuntime)
PREEMPTION_RANK_STABLE = False
ALLOW_PSLOT_PREEMPTION = True
    
```

- To enable the preemption feature,
 - The settings must be set in the Condor Negotiator because we need to use a policy to preempt
 - The above PREEMPTION_REQUIREMENTS formula is an official guideline.
 - Available ClassAds variables are described on the "Priority in Negotiation and Preemption" page in the HTCondor Manual
 - You can only access new job or machine classads
 - The classads of the current running job must be approached in a different way
 - TotalJobRuntime is defined on condor_startd (==machine classads)
 - PREEMPTION_RANK is used to select the machine slot to be preempted.
 - We set the above setting because we want to cancel the job the most recent submission
 - You can set a preemption policy by considering the priority of the job.

Negotiator

```
GROUP_NAMES = group_alice, group_alice.yonsei, group_cms, group_genome, group_genome.argo, group_genome.bio, group_etc
GROUP_QUOTA_group_etc = 0
GROUP_QUOTA_group_alice = 1160
GROUP_QUOTA_group_alice.yonsei = 200
GROUP_QUOTA_group_cms = 2288
GROUP_QUOTA_group_genome = 1696
GROUP_QUOTA_group_genome.argo = 1696
GROUP_QUOTA_group_genome.bio = 1696
GROUP_ACCEPT_SURPLUS = true
GROUP_ACCEPT_SURPLUS_alice = true
GROUP_ACCEPT_SURPLUS_alice.yonsei = false
GROUP_ACCEPT_SURPLUS_cms = true
GROUP_ACCEPT_SURPLUS_genome = true
GROUP_ACCEPT_SURPLUS_genome.argo = true
GROUP_ACCEPT_SURPLUS_genome.bio = true
```

```
[geonmo@ui20 ~]$ condor_userprio -quota
Last Priority Update: 10/13 17:54
Group      Effective  Config   Use      Subtree  Requested
Name       Quota     Quota   Surplus  Quota    Resources
-----
group_alice      1160.00  1160.00  ByQuota  1160.00  6643
group_alice.yonsei  200.00  200.00  ByQuota  200.00  146
group_cms        2288.00  2288.00  ByQuota  2288.00  9705
group_genome.bio  1696.00  1696.00  ByQuota  1696.00  10
-----
Number of users: 8                               ByQuota
```

- To set account group and quota,
 - Restarting the Negotiator for a short period of time does not cause problems
 - To apply the settings by restarting the condor service
 - According to the above settings, the alice.yonsei group **cannot** use more than **200 cores**.
 - However, when the entire alice group is using more than their quota and needs to return resources, 200 cores is preserved for the alice.yonsei group

```
SUBMIT_REQUIREMENT_NAMES = GROUP
SUBMIT_REQUIREMENT_GROUP = (AcctGroup =?= "group_cms")
SUBMIT_REQUIREMENT_GROUP_REASON = "Bad accounting group. Your group is group_cms"
```

SCHEDD

Job Description File(.sub)

```
accounting_group=group_cms
```

- To prevent submission of jobs to the wrong group,
 - Prepare job submission servers for each group
 - SUBMIT_REQUIREMENT keyword can be used to restrict the submission of jobs to an invalid accounting group name
 - If it is difficult to separate, it is necessary to train users to submit a job as their accounting group name
 - Alternatively, if you use HTCondor 9.0 or later, you can automatically separate it using a accounting group map file.
 - For user convenience, we are using the alias below.
 - alias condor_submit='condor_submit -append accounting_group="group_cms" '

STARTD

```
### Singularity(Apptainer) Part
SINGULARITY_JOB = !isUndefined(TARGET.SingularityImage)
SINGULARITY_IMAGE_EXPR = TARGET.SingularityImage
SINGULARITY_TARGET_DIR = /srv
MOUNT_UNDER_SCRATCH = /tmp, /var/tmp
SINGULARITY_BIND_EXPR=ifThenElse( isUndefined(TARGET.SingularityBind),"/cvmfs, /cvmfs, /cms, /share, /tmp",TARGET.SingularityBind)
SINGULARITY_EXTRA_ARGUMENTS=ifThenElse( isUndefined(TARGET.SingularityExtraArgs),"",TARGET.SingularityExtraArgs)
```

Job Description File(.sub)

```
+SingularityImage = "/cvmfs/singularity.opensciencegrid.org/opensciencegrid/osgvo-el8:latest"
+SingularityBindCVMFS = True
+SingularityBind = "/cvmfs, /cms, /cms_scratch"
## If you want to use apptainer's extra arguments like as "--nv",
#+SingularityExtraArgs= "--nv"
```

- No preemption configuration for WNs
 - To increase the Preemption speed, a setting that divides the dynamic slots on the WN into multiple is required
 - The vacating is 1 per primary slot in a negotiation cycle.
- The integrated farm mainly uses Apptainer runtime rather than Docker
 - This is because there is no need to enable a docker daemon or manage the docker user list
 - Also, Docker container images are available at apptainer runtime
- To enable Singularity(apptainer) container environment,
 - Setup above configuration
 - Add related classads to a job description file
- If fusermount3 error occurs after updating the apptainer instead of the singularity,
 - If you are not familiar with rootless (unprivileged) container settings, install the apptainer-suid package.
- If you download an image from each WN, there may be a limit on the number of downloads of dockerhub homepage
 - It is also used to download the container image to a file and put it in a shared directory

References

- HTCondor-CE Documentation
 - <https://htcondor.com/htcondor-ce/>
 - It is well documented to setup a HTCondor-CE and configure with local batch cluster (HTCondor and others) in step-by-step
- HTCondor Documentation
 - v9.0: https://htcondor.readthedocs.io/en/v9_0/
 - v10.0 (LTS): <https://htcondor.readthedocs.io/en/lts/>
- HTCondor-based AliEn Site Installation
 - <https://alien.web.cern.ch/content/documentation/howto/site/htcondor-based-alien-site-installation>