PIC
port d'informació
científica

HTCondor-CE at PIC

J. Flix, C. Acosta*

EGI Conference 2019, 6-8 May 2019, Amsterdam

IFAE 25 years 1991-2016 Institut de Física d'Altes Energies

EXCELENCIA SEVERO OCHOA

BIST Barcelona Institute of Science and Technology

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

EXCELENCIA MARÍA DE MAEZTU

cfp CIEMAT física de partículas

PIC
port d'informació
científica

*Disclaimer!

The talk and the work was done by my colleague C. Acosta, which is very helpful and welcoming feedback and technical questions that I won't be able (most likely) to answer

IFAE 25 years 1991-2016 Institut de Física d'Altes Energies

EXCELENCIA SEVERO OCHOA

BIST Barcelona Institute of Science and Technology

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

EXCELENCIA MARÍA DE MAEZTU

cfp CIEMAT física de partículas

PIC
port d'informació
científica

HTCondor-CE at PIC

J. Flix, C. Acosta [credits go to him!]

EGI Conference 2019, 6-8 May 2019, Amsterdam

IFAE 25 years 1991-2016 Institut de Física d'Altes Energies

EXCELENCIA SEVERO OCHOA

BIST Barcelona Institute of Science and Technology

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

EXCELENCIA MARÍA DE MAEZTU

cfp CIEMAT física de partículas

PIC
port d'informació
científica

Index

- Introduction
- HTCondor at PIC
- CEs for HTCondor
- HTCondor-CE
- Conclusions

PIC
port d'informació
científica

Introduction

Multi-VO site

- 82% of PIC computing resources for Tier-1 (ATLAS, CMS, LHCb)
- 7% Tier-2 ATLAS (IFAE site)
- ~10% for: Tier-0 Magic, Tier-3 ATLAS and other smaller VOs

PIC site is fully puppetized (3.8.7 → 4.10.10 soon)

PIC
port d'informació
científica

Introduction

- PIC and IFAE sites (IFAE co-located in PIC) were Grid sites based in CREAM-CE and Torque/Maui batch system
- Torque 4.2.10 and Maui 3.3.1 were a general solution for many Grid sites basically because CREAM-CE+Torque are well implemented
- Many issues for not trusting in Torque/Maui for the future:
  - Maui and Torque branches 2 and 4 were not supported for the long term
  - Scalability issues
  - Need of new features (cgroups for instance, docker and singularity integration)
  - IPv6 compatibility
  - RSS memory report
  - CREAM-CE end of life 2020!
- Other options (free, highly scalable, etc.)? Slurm or **HTCondor**
- HTCondor is the natural solution for Grid sites considering its versatility and that it is the default batch system for important Grid Sites (CERN, RAL, most of the OSG sites)

PIC
port d'informació
científica

HTCondor at PIC

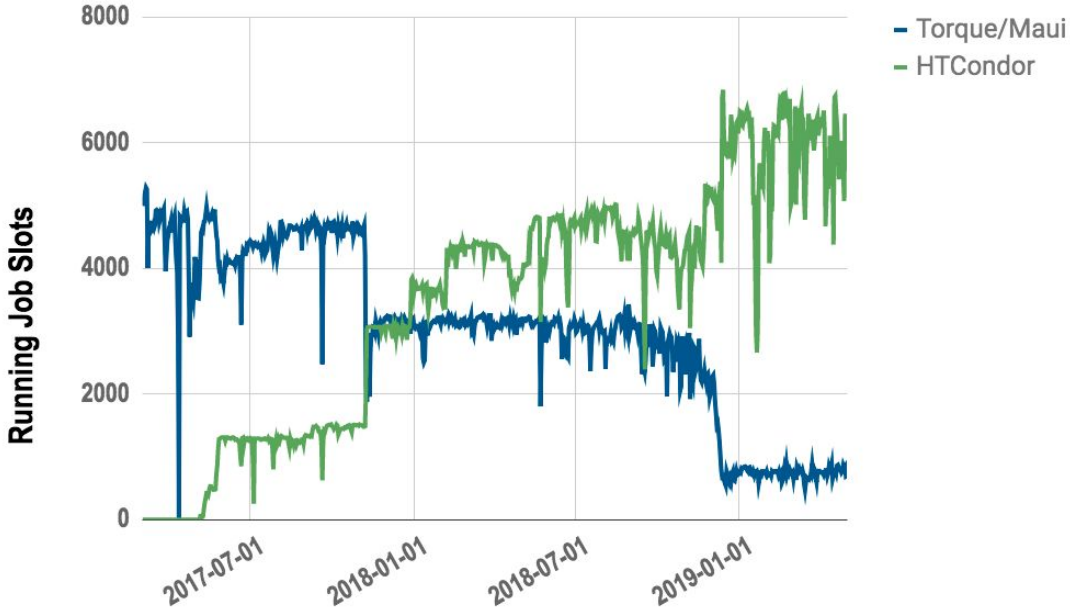
The machines in the HTCondor pool are defined by the daemons running in it

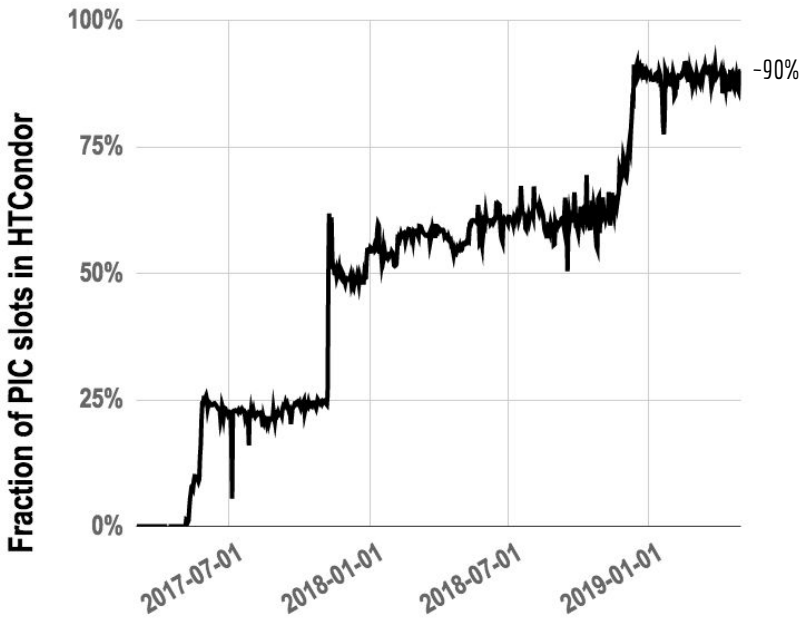
Relevant HTCondor daemons:

- **condor_master**: responsible for controlling the rest of daemons
- **condor_collector**: takes the information from an HTCondor pool
- **condor_negotiator**: makes the matching between ClassAds jobs and ClassAds machines
- **condor_schedd**: controls the queue and one can queries/modify it with several commands. The condor_schedd generates a **condor_shadow** process for each matching
- **condor_startd**: defines an execute machine (WNs). The **condor_starter** daemon is the one that starts the job when condor_startd is ready

HTCondor at PIC

- At April 2019, the HTCondor pool at PIC consists in a pool of around 7300 CPU-cores and ~100 kHS06; Shrinking partition of ~11 kHS06 in Torque/Maui for small VOs… soon to be integrated into HTCondor

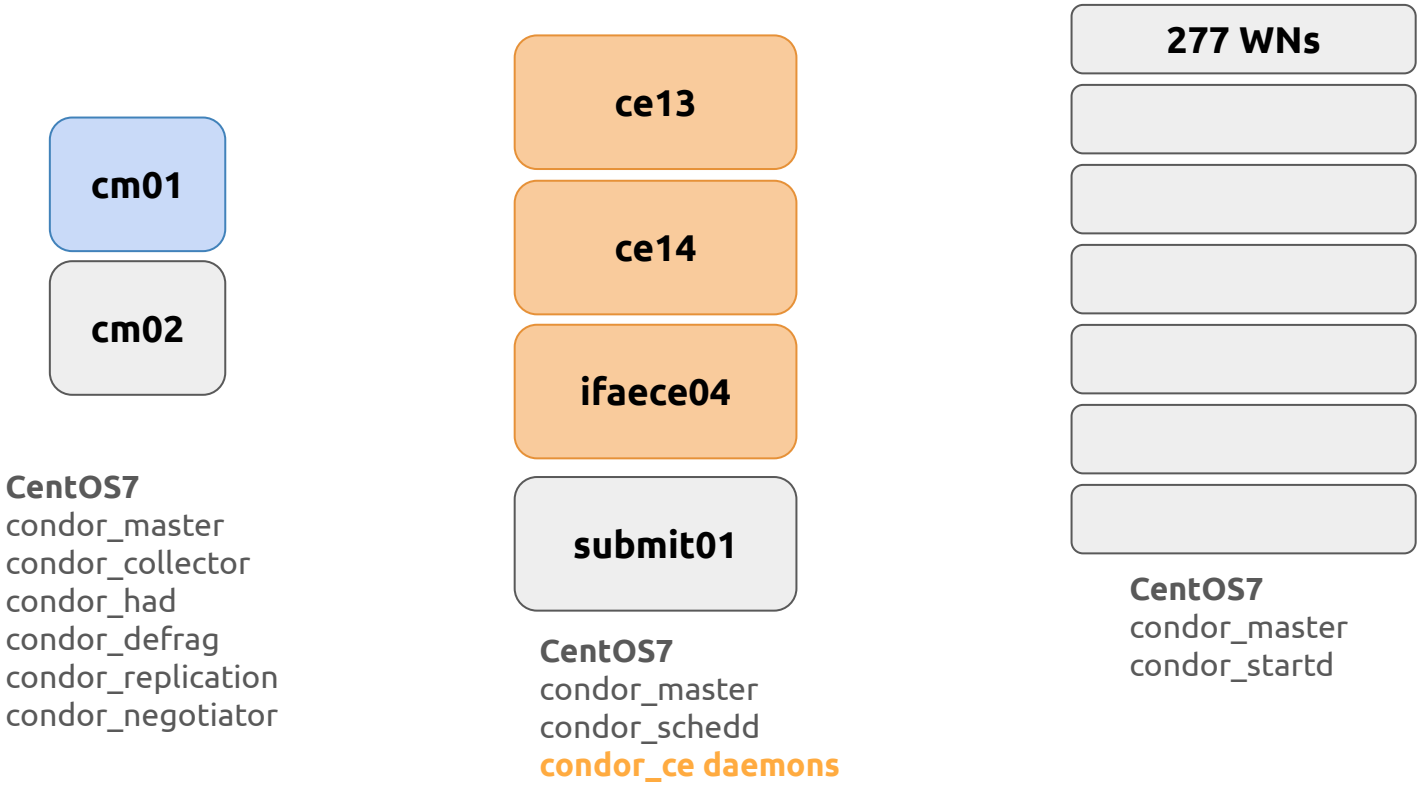




PIC
port d'informació
científica

HTCondor at PIC

- At April 2019, the HTCondor pool at PIC consists in a pool of around 7300 CPU-cores and ~100kHS06
- Current version: Stable 8.8.1 (migrating to 8.8.2 soon)
- We have 2 Central Managers (CMs - HA), 277 Worker Nodes, 1 local schedd and 3 HTCondor-CEs (let's talk about them later)



PIC
port d'informació
científica

HTCondor at PIC

Central Managers

- 2 CM in high availability at PIC
- Priority and Fair-Share
  - Establish % of use for each VOs using GROUP_QUOTA_DYNAMIC
  - GROUP_ACCEPT_SURPLUS allows groups to use more of their quota and accept new matching even if the group is over quota (GROUP_AUTOREGROUP)
  - Define GROUP_NAMES and priority factors
  - GROUP_SORT_EXPR to consider first highprio, then mcore, etc… for the negotiator
- Defrag daemon
  - condor_defrag to drain machines and to obtain a hole for next mcore job (similar concept as mcfloat used in previous Torque/Maui configuration)
- Condor_gangliad
  - Several metrics are directly submitted form HTCondor to Ganglia

PIC
port d'informació
científica

HTCondor at PIC

Execute Machines

- Dynamic slots configuration
  - Slots created due to jobs requirements (cpu)
- MOUNT_UNDER_SCRATCH
  - Very helpful: the job believes that /tmp and /var/tmp are in execute directory
  - Issue with glexec solved
- PID_NAMESPACES and cgroup
  - PID_NAMESPACES: same process ID number space for all job processes
  - cgroup to limit memory (soft) and use of procs
- Tag inclusion in startd configuration
  - Adding tags to WNs (WN="IPv6" e.g. in STARTD_ATTRS)
- Node Health Check to check if a WN is healthy (HW issues, CVMFS working, etc)
- WN grid configuration: CVMFS, Machine Job Features, singularity, etc.

PIC
port d'informació
científica

HTCondor at PIC

Local Schedd

- We are in the process to integrate local users, since grid jobs have been running in HTCondor through HTCondor-CEs for a while
- 1 local schedd available for local users
  - Automatic mapping to an Accounting Group according to job Owner using JOB TRANSFORM
  - As CERN does, we create “flavours” to select the maximum walltime and other limits of the jobs (like memory, for example)
  - Established default requirements (1 cpu, 2 GB/cpu of RAM, 15 GB/cpu of disk)
  - Allowing remote submission from dedicated UIs
  - Creating documentation and training our users

PIC
port d'informació
científica

CEs for HTCondor

- 3 CEs were evaluated

CREAM-CE

- Well know and integrated with all our resources
- Doubt about its support (finally, the end-of-life was published)
- Not really ready for HTCondor (old scripts) and a complex configuration based in YAIM

ARC-CE

- Easy to configure (just one configuration file /etc/arc.conf) and widely use by Nordugrid and RAL
- Delay in query responses in our test environment
- We installed it and used in production with HTCondor for a while

PIC
port d'informació
científica

CEs for HTCondor

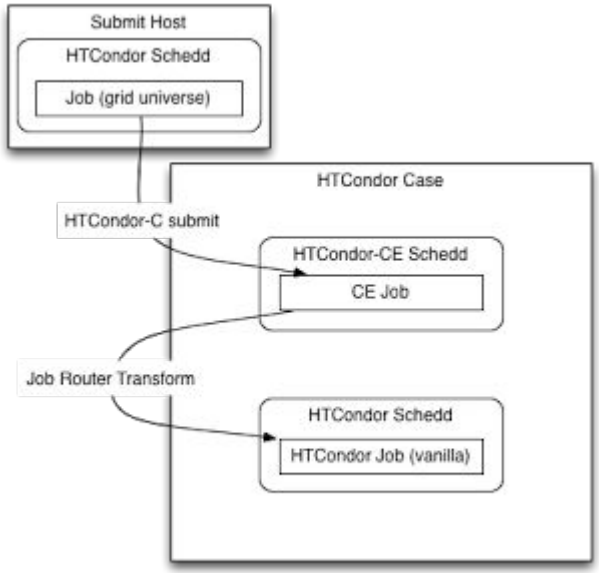
- 3 CEs were evaluated

HTCondor-CE

- It is just using HTCondor, in fact, uses HTCondor binaries and seems to be the most logical option if your batch system is HTCondor
- HTCondor-CE widely used in OSG sites but missing EGI features
- CERN interest turned to change things for us…
- New elements to interconnect PIC to HPCs rely on HTCondor

HTCondor-CE at PIC

- Current CEs setup: 2 HTCondor-CEs for Tier-1 (ATLAS, CMS, LHCb) and 1 HTCondor-CE for ATLAS Tier-2
- Current version: 3.2.1 (first time including bdii extension from htcondor repos)



- 2 set of condor daemons in the CE, the ones related to your condor pool and the ones from the HTCondor-CE
- JobRouter is the one which transform the job from the HTCondor-CE queue to the local condor queue

PIC
port d'informació
científica

HTCondor-CE at PIC

JobRouter

- Very powerful tool, it transforms the job according to request requirements (adding one requirement to the job, the job can be routed by this new tag that points to a specific WN, e.g.)
- According to the x509UserProxyVOName, the job is routed and the GROUP_NAME is established (mandatory for FairShare)
- Easy to make syntax mistakes, treat the configuration file carefully!
- JOB_TRANSFORM allows us to map users to Accountings Groups and create easier routes

HTCondor-CE at PIC

JobRouter

- Example of a route:

```
   [ \
    name = "Cms_Condor"; \
    MaxIdleJobs = 2000; \
    TargetUniverse = 5; \
    set_LivAcctSubGroup = ifThenElse(NumberCpus > 1 ,"_mcore","_score");\
    eval_set_AccountingGroup = strcat(LivAcctGroup, LivAcctSubGroup, ".", Owner); \
    Requirements = x509UserProxyVOName == "cms"; \
    set_Requirements = WN_property == "default"; \
    eval_set_NumberCpus = ifThenElse(xcount is undefined, orig_RequestCpus, ifThenElse(xcount
isnt undefined, xcount, 1));\
    set_PICScaling = "$$(PICScaling:1)"; \
    set_JobMemoryLimit = (6144*NumberCpus); \
    set_PeriodicRemove = (JobStatus == 2 && (CurrentTime - EnteredCurrentStatus) > 360000) ||
(JobStatus == 4 && (CurrentTime - EnteredCurrentStatus ) > 3600*24*2) ||
ifThenElse(ResidentSetSize isnt undefined, ResidentSetSize > JobMemoryLimit*1024, false); \
   ] \
```

PIC
port d'informació
científica

HTCondor-CE at PIC

Publication

- One old issue related to EGI integration but CERN worked to create htcondor-ce-bdii package
- In the past, we built all our htcondor bdii package but the htcondor-ce-bdii package was released for first time in HTCondor repo from version 8.8.1

PIC
port d'informació
científica

HTCondor-CE at PIC

Authorization/Authentication

- HTCondor-CE uses GSI-based authentication and authorization and you can easily use an Argus server as we did with CREAM-CE

```
# cat /etc/grid-security/gsi-authz.conf
globus_mapping /usr/lib64/libgsi_pep_callout.so argus_pep_callout

# cat /etc/grid-security/gsi-pep-callout-condor.conf
pep_ssl_server_capath /etc/grid-security/certificates/
pep_ssl_client_cert /etc/grid-security/condorcert.pem
pep_ssl_client_key /etc/grid-security/condorkey.pem
pep_url https://argus.pic.es:8154/authz
pep_timeout 30 # seconds
xacml_resourceid http://authz-interop.org/xacml/resource/resource-type/htcondor-ce
```

PIC
port d'informació
científica

HTCondor-CE at PIC

Authorization/Authentication

- HTCondor-CE uses GSI-based authentication and authorization and you can easily use an Argus server as we did with CREAM-CE

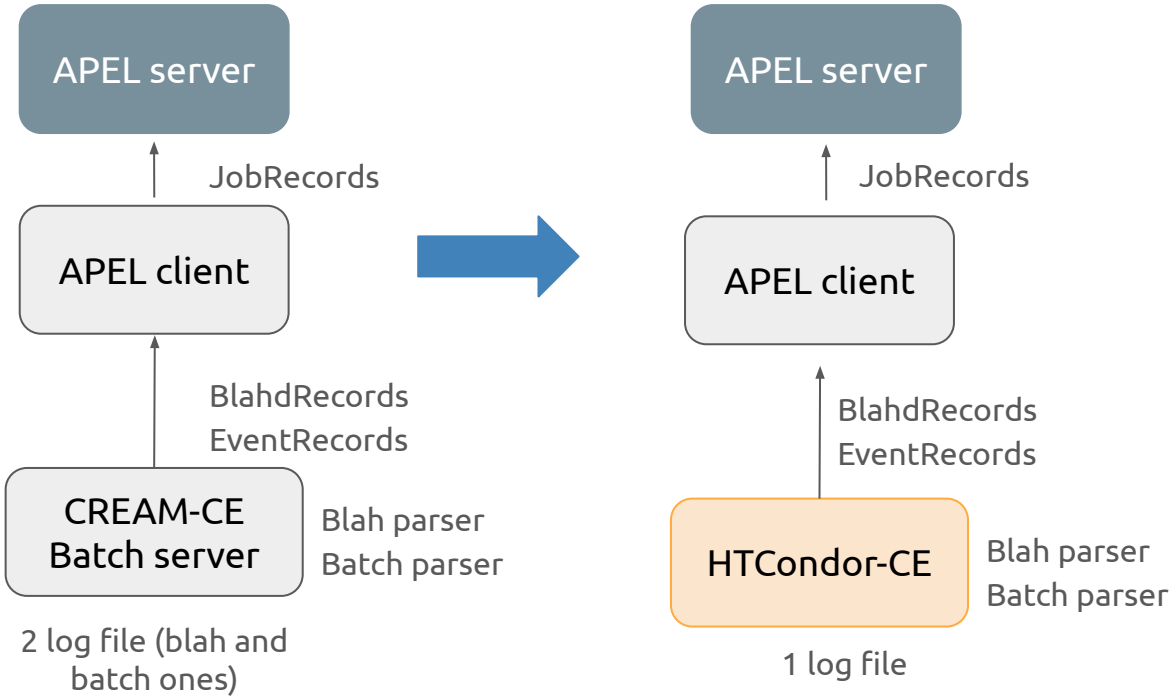
```
resource "http://authz-interop.org/xacml/resource/resource-type/htcondor-ce" {
    obligation "http://glite.org/xacml/obligation/local-environment-map" {
    }

    action ".*" {
        rule permit { pfqan="/atlas/Role=lcgadmin/Capability=NULL" }
        rule permit { pfqan="/atlas/Role=lcgadmin" }
        rule permit { pfqan="/atlas/Role=pilot/Capability=NULL" }
        rule permit { pfqan="/atlas/Role=pilot" }
        rule permit { pfqan="/atlas/Role=production/Capability=NULL" }
        rule permit { pfqan="/atlas/Role=production" }
        rule permit { pfqan="/atlas/Role=software/Capability=NULL" }
        rule permit { pfqan="/atlas/Role=software" }
        rule permit { pfqan="/atlas/Role=NULL/Capability=NULL" }
        rule permit { pfqan="/atlas" }
[...]
```

HTCondor-CE at PIC

APEL integration

- It was the first stopper to put HTCondor-CEs in production
- We create our own APEL integration based in CREAM-CE/Torque scheme



HTCondor-CE at PIC

APEL integration

- It was a difficulty to solve in order to put HTCondor-CEs in production
- We created our own APEL integration based in CREAM-CE/Torque schema
- We use a condor_history query to create our accounting records in text files
- PICScaling is the cputmult used to normalize the CPU
- Adding and changing scripts
  - Apel-parsers and apel-lib pacakges installed in the CE
  - Changes in /etc/apel/parser.cfg file
  - Modify the apelparser script
  - Add htcondor.py and htcondorce.py parsers
- This solution is thought to work using our environment -> Stephen Jones is working in a general solution for any site

https://twiki.cern.ch/twiki/bin/viewauth/LCG/HtCondorCeAccounting
https://github.com/apel/apel/releases
https://wiki.egi.eu/wiki/Preview_Repository#Updates_Released

HTCondor-CE at PIC

CE management

- You can use the same commands as condor (condor_ce_q, condor_ce_history for instance)

```
# condor_ce_q


-- Schedd: ce14.pic.es : <193.109.175.230:13903> @ 04/24/19 17:12:19
OWNER      BATCH_NAME     SUBMITTED   DONE   RUN    IDLE   HOLD  TOTAL JOB_IDS
lhpilot007 ID: 4980889   4/20 03:58      _      _      _      _     25 4980889.0-24
lhpilot007 ID: 4980895   4/20 03:58      _      _      _      _     25 4980895.0-24
lhpilot007 ID: 4980896   4/20 03:58      _      _      _      _     25 4980896.0-24
[...]
atpilot012 ID: 5029317   4/24 17:09      _      1      _      _      1 5029317.0
atpilot012 ID: 5029318   4/24 17:09      _      1      _      _      1 5029318.0
atpilot012 ID: 5029319   4/24 17:09      _      1      _      _      1 5029319.0
atpilot012 ID: 5029320   4/24 17:09      _      1      _      _      1 5029320.0
lhpilot007 ID: 5029321   4/24 17:11      _     24      _      _     24 5029321.0-23
atpilot012 ID: 5029322   4/24 17:12      _      _      _      1      1 5029322.0

Total for query: 5135 jobs; 3885 completed, 0 removed, 233 idle, 968 running, 49 held, 0
suspended
Total for all users: 5135 jobs; 3885 completed, 0 removed, 233 idle, 968 running, 49 held, 0
suspended
```

HTCondor-CE at PIC

CE management

- You can use the same commands as condor (condor_ce_q, condor_ce_history for instance)
- There are different log files stored for all the daemons in /var/log/condor-ce (JobRouterLog, SchedLog and AuditLog are important to check)
- AuditLog kepts for 90 days information about what is happening in the queue, new jobs, transfer files, proxies, etc.

HTCondor to exploit HPC resources

HTCondor for BSC

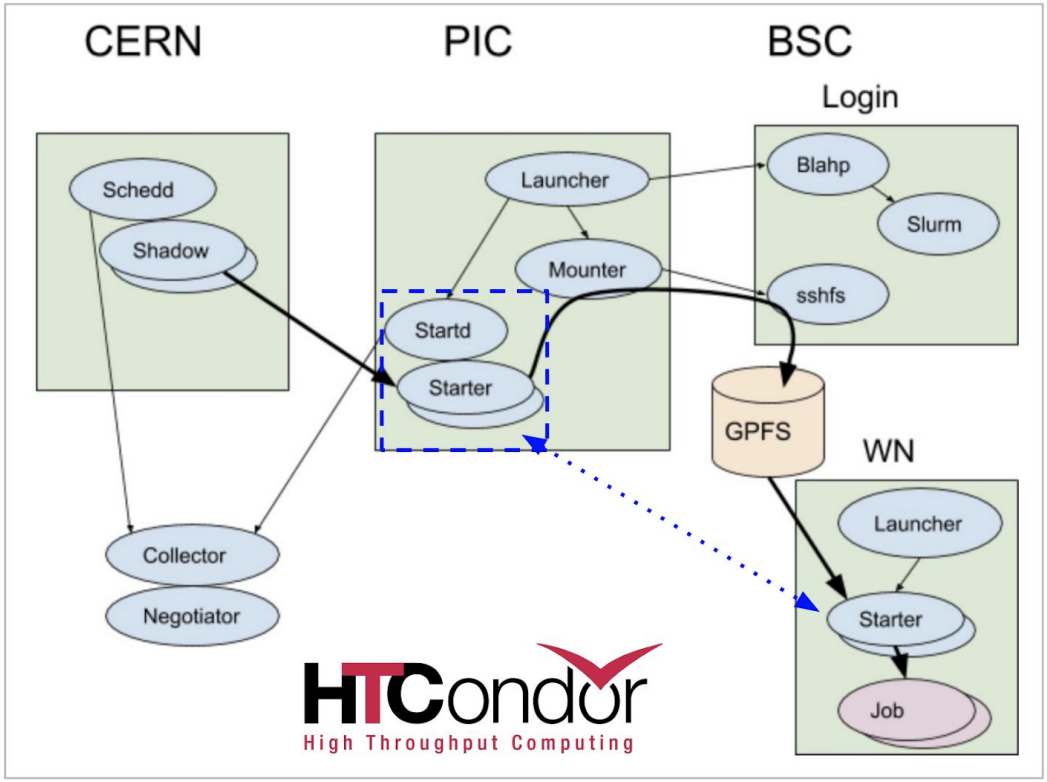
Main idea: a **bridge node** at PIC which allows access to **starter** processes running in the BSC nodes, mirroring **starter** at PIC

Input sandbox, status, etc, **passed as .tar files** via gpfs from bridge to WN

The **startd** process remains at PIC, where it can be accessed to **negotiate** by CMS/PIC Central Managers

From a functional perspective, **the node at BSC has joined the HTCondor pool at PIC**

See more details in the backup slides



PIC
port d'informació
científica

PIC experience so far...

- HTCondor (and HTCondor-CE in extension) is a free software with a dynamic team working continuously to always improve it
- User support and extensive documentation available
- Once you get a stable configuration, HTCondor-CEs works fine without any remarkable issues
- Easy to play with HTCondor-CE as works with the same concepts as HTCondor
- HTCondor and HTCondor-CE flexibility allow us to test and implement new features: PIC-CIEMAT federation, connect to cloud resources (HelixNebula, AWS), interfacing to HPCs (collaboration PIC with HTCondor developers)
- Some issues and bugs, but HTCondor developers very proactive and helpful
    - Dual-stack issue with High Availability
    - condor_annex bug reported (to connect resources with AWS)
    -
- Lack of documentation supporting HTCondor-CE+EGI resources

PIC
port d'informació
científica

Questions?

IFAE 25 years 1991-2016 Institut de Física d'Altes Energies

EXCELENCIA SEVERO OCHOA

BIST Barcelona Institute of Science and Technology

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

EXCELENCIA MARÍA DE MAEZTU

cfp CIEMAT física de partículas