GRID Resources - Computing

Corso di formazione per utenti DataCloud

WP2 DataCloud
Alessandro Pascolini
alessandro.pascolini@cnaf.infn.it

Outline

- GRID Resources
- High Throughput Computing (HTC)
- HTCondor
  - Cluster structure
  - Users and queues
  - Job flow
  - Commands
- HTCondor-CE
  - GRID AuthN/Z
  - GRID submission

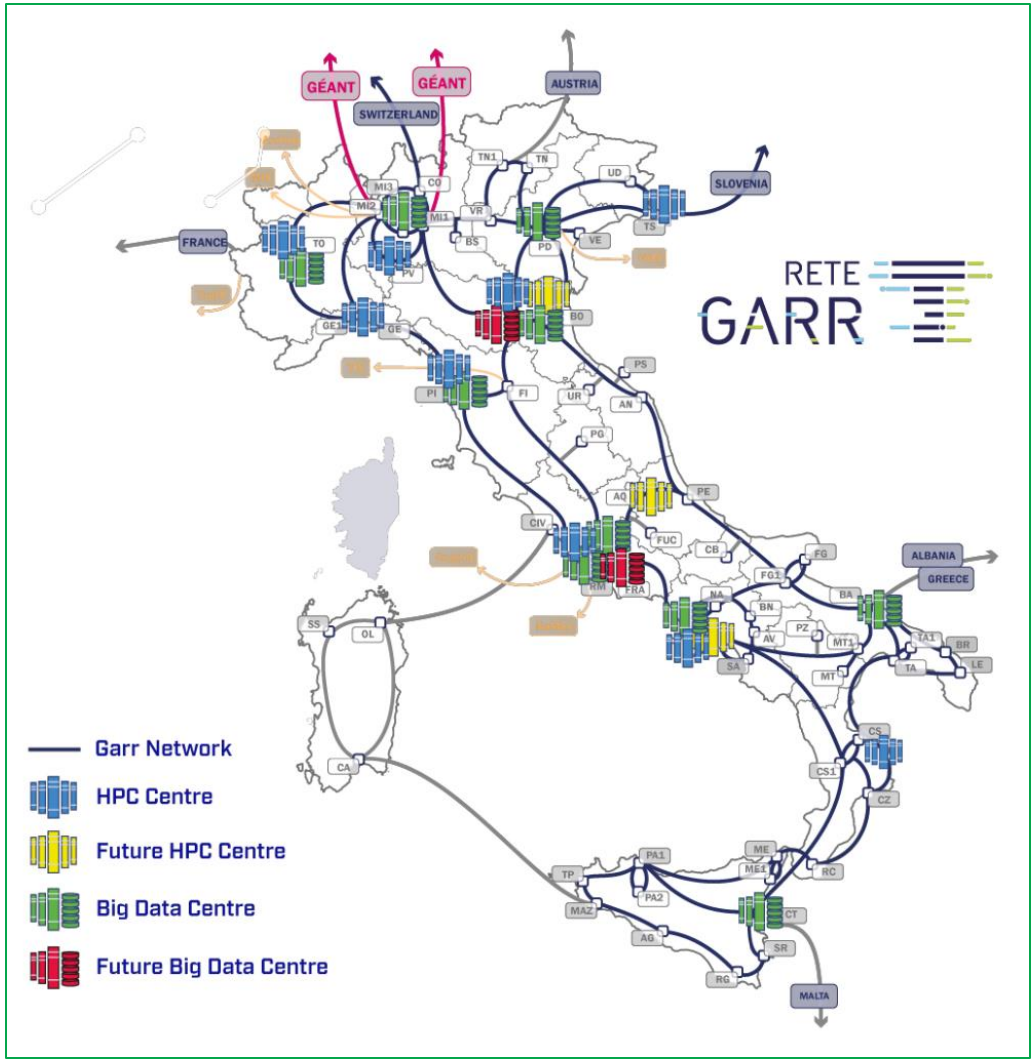
GRID Resources - Computing

Grid Computing @ INFN
- 1 Tier-1 → INFN – CNAF
- 9 Tier-2

~100k total cores

Provided mainly via HTC (High Throughput Computing) clusters
- **HTCondor**
- LSF
- SLURM
- …

GRID Access → AUTENTICATION
- **SCITOKEN**
- SSL - x509 (VOMS-Proxy)



GRID Computing

High Throughput Computing

and

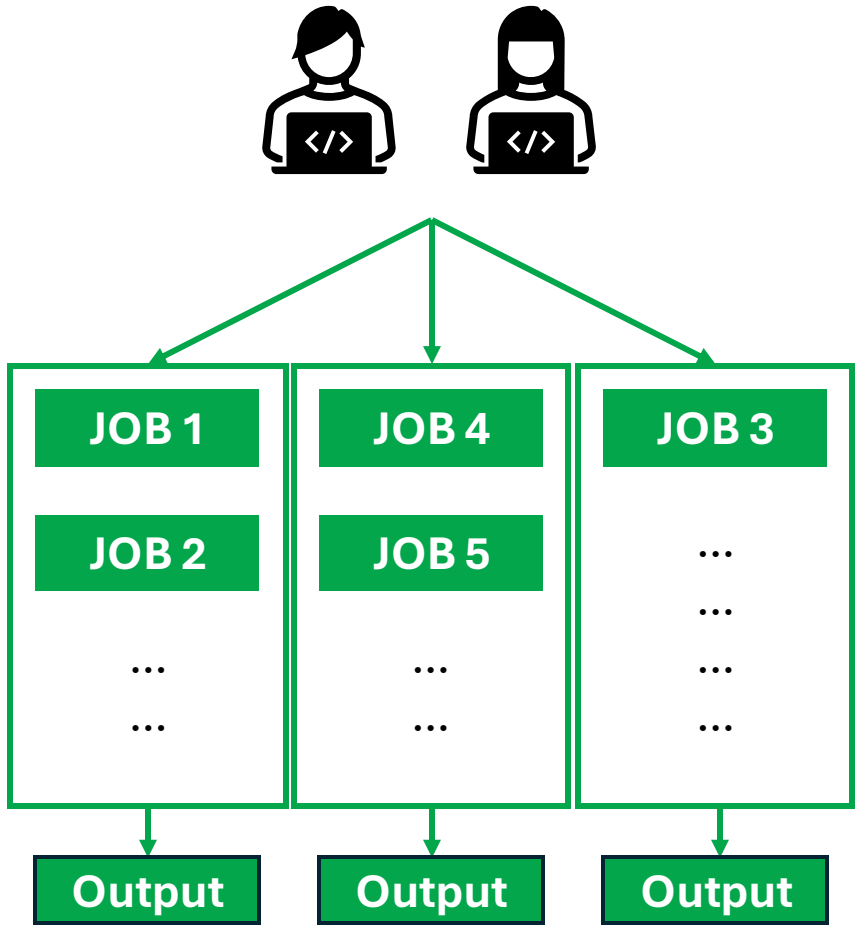
HTCondor

High Throughput Computing

- Optimized for **NON INTERACTIVE** (batch) jobs
- Jobs run on a single machine, using one or more CPUs
- High Throughput of jobs (i.e. many jobs running at the same time)

Why HTC?

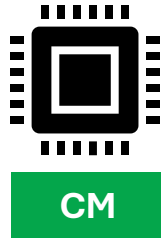
- Processes may take a lot of time to complete
- Resource heavy workflows that can't be run on user's PC
- Offers a way to delegate the execution on a remote machine or cluster

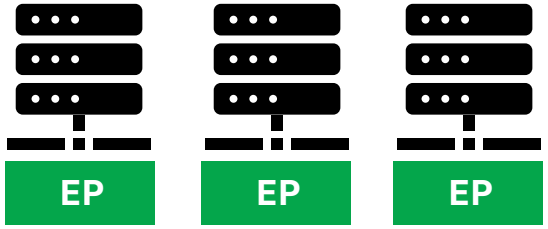


HTCondor

Software optimized to manage:
- High number of resources
- Many users assigned to different queues

HTCondor roles:
- **Central Manager**
  controls the whole cluster
- **Access Point**
  machine where the jobs are submitted to
- **Execution Point**
  executes the jobs





GRID Computing

HTCondor

Users and queues

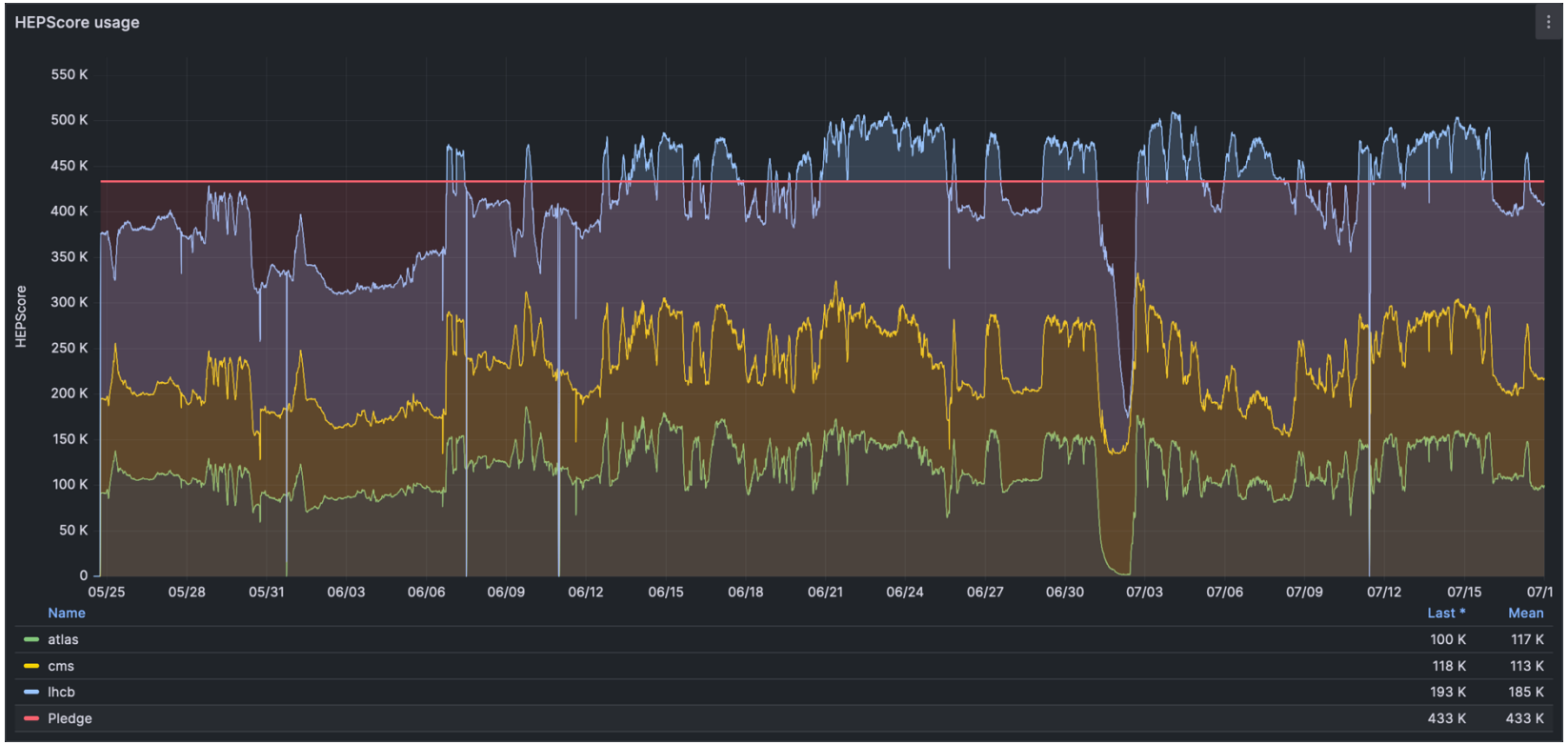
HTCondor – Users & Queues

HTCondor users:

- Each job is assigned to the unix user that submitted it
- On the **EP** the job is executed with the unix user corrisponding to the owner
- A user can **belong to more AcctGroups**

HTCondor queues (AcctGroups):

- To each AcctGroup can be assigned some resources:
  - % of all cluster resources
  - Absolute unit of cores
- An AcctGroup **may use more than what is assigned**



GRID Computing

HTCondor

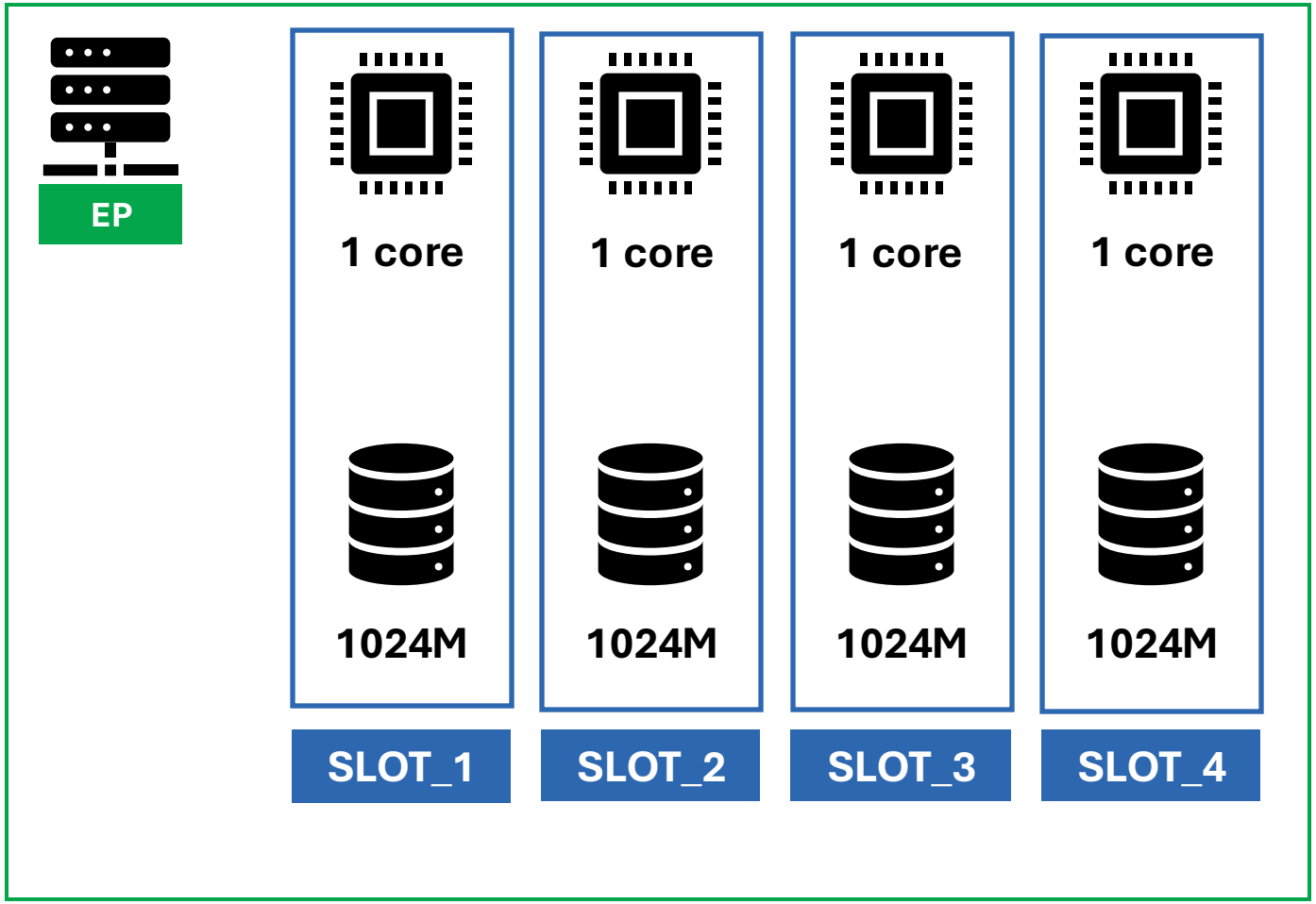
Execution Point

HTCondor – EP Static Slots

HTCondor detects all **hardware resources on the EP**
These resources will be assigned to **one or more slots**

STATIC SLOTS
- Default in HTCondor < 10
- By default assigns 1 core/slot
- Memory equally distributed to each slot

- **<u>Issues with job requests not matching slots flavour</u>**

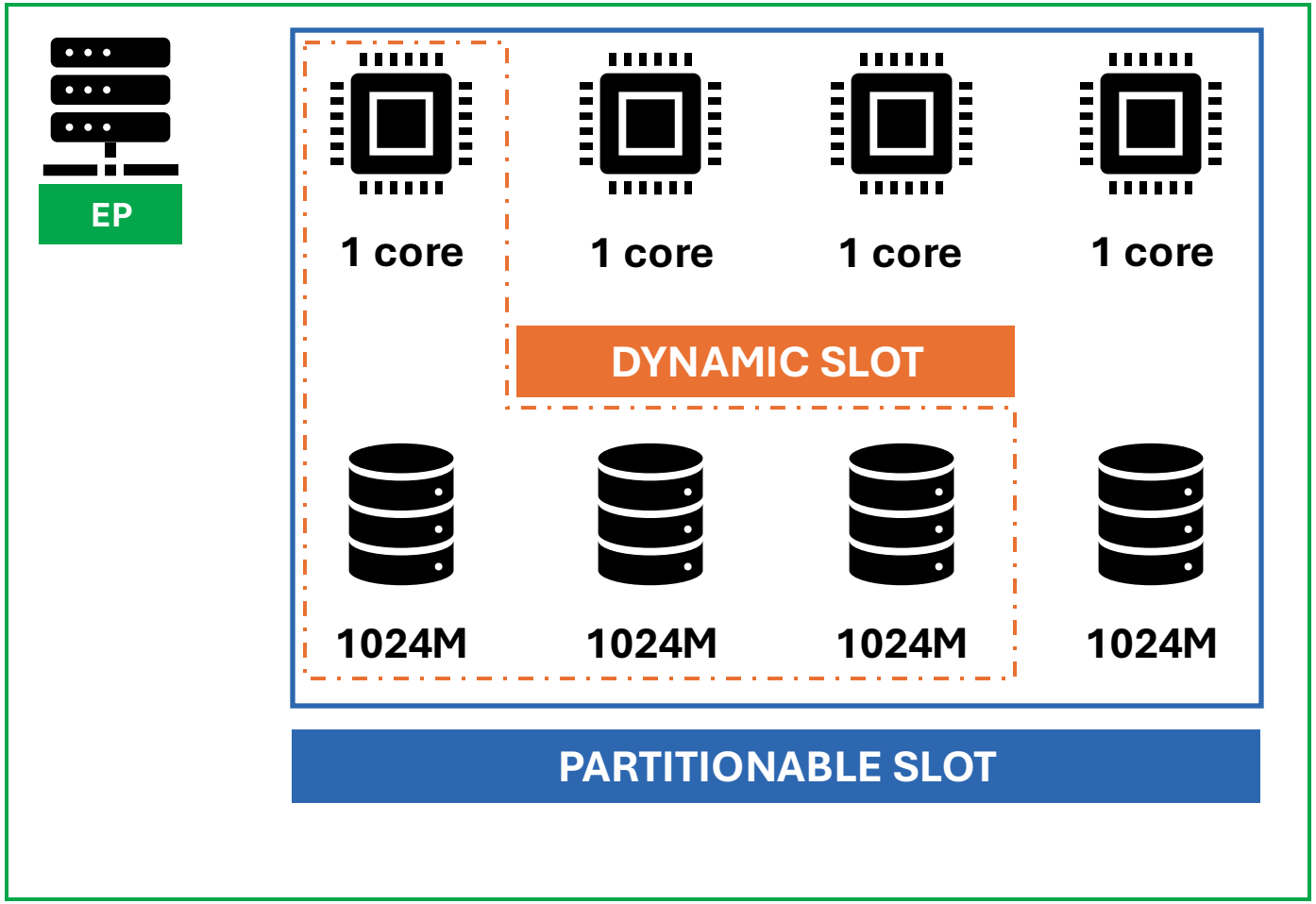


HTCondor – EP Dynamic Slots

HTCondor detects all **hardware resources on the EP**
These resources will be assigned to **one or more slots**

DYNAMIC SLOTS
- Default in newer versions of HTCondor
- By default assigns ALL resources to a Partitionable slot
- Dynamic slots created according to job requests

- **<u>More flexibility</u>**



GRID Computing

HTCondor

Job life and management

HTCondor – What is a Job?

- Bunch of instruction to launch an executable on a remote machine

- In HTCondor every entity is described by some Attributes, for example in Jobs we have:
  - **Requirements**
    expression that defines the conditions that an **EP** need so satisfy to execute the job
  - **Request<metric>**
    CPU,memory,disk that the Job request to the **EP**
  - **many more…** [1]

- Many of these features are declared in the submit file

local-submit.sub

```
1  # Unix submit description file
2  # sleep.sub -- simple sleep job
3
4  batch_name              = Sleep
5  executable              = /usr/bin/sleep
6  arguments               = 300
7  log                     = Sleep.log
8  output                  = Sleep.out
9  error                   = Sleep.err
10
11
12 # require to run on LINUX machines
13 requirements = OpSys == "LINUX"
14
15 # ask for 1 core, 1024MB memory and 1024kB disk
16 request_cpus   = 1
17 request_memory = 1024M
18 request_disk   = 10240K
19
20 # submitting 3 jobs (default 1, if no number specified after queue)
21 queue 3
```

Output Files

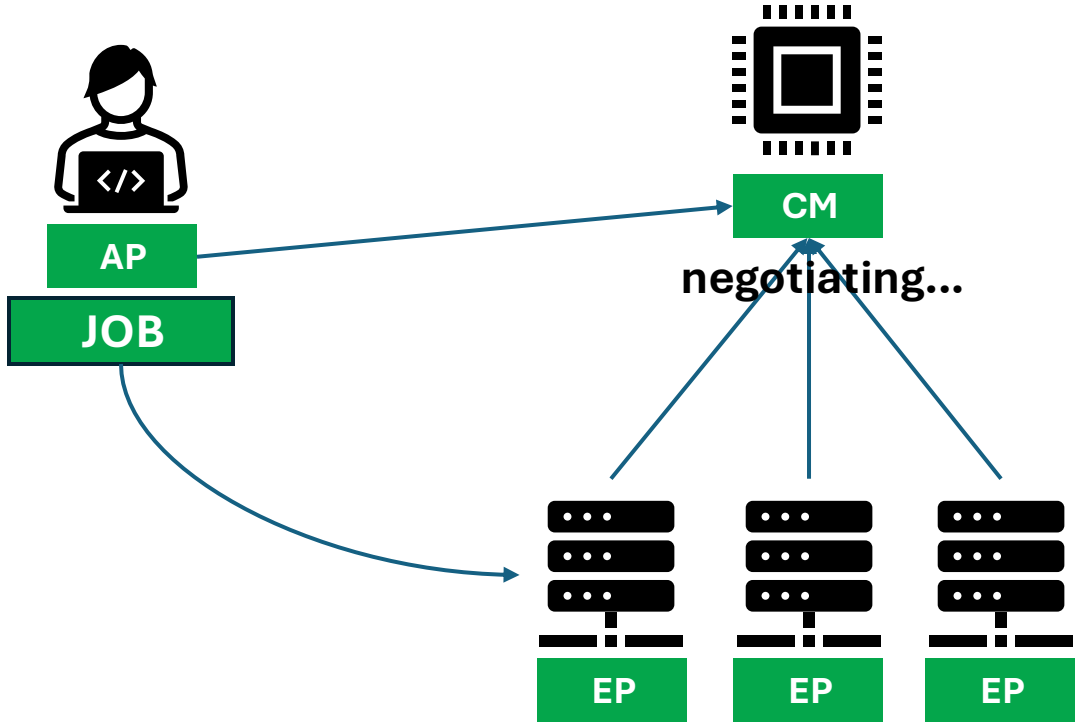
Requests to EP

Simple submit file that sleeps for 300 seconds

[1] **Job ClassAd Attributes**
https://htcondor.readthedocs.io/en/latest/classad-attributes/job-classad-attributes.html#job-classad-attributes

HTCondor – Job Flow

1. A Job is submitted to the **AP**
2. The **CM** periodically retrieves all info from other machines
   - Job status
   - **EP** status
3. Negotiation stage (idle Jobs are assigned to an **EP**)
4. The Job starts running on an **EP**



HTCondor – Job Life

IDLE
- Job submitted, waiting to be assigned to EP
- Attributes:
  - Qdate → submission timestamp
  - JobID/ClusterID → unique identifiers of Job/Cluster

RUNNING
- Job assigned to EP and statred execution
- Attributes:
  + JobStartDate → timestamp when job started running
  + RemoteHost → hostname of EP

COMPLETED

If everything goes well…
- Job has ended
- Attributes:
  + CompletionDate → timestamp when job ended
  + LastRemoteHost → hostname of EP

REMOVED
- Job removed from queue by User or Admin
- Attributes:
  + RemoveReason → String with the reason of job removal

If something doesn't…

HOLD
- Job put in hold due to several reasons
- HTCondor may put a job in hold to prevent it from using more resources than requested
- Attributes:
  - HoldReason → String with the reason

HTCondor – Job Cheat Sheet

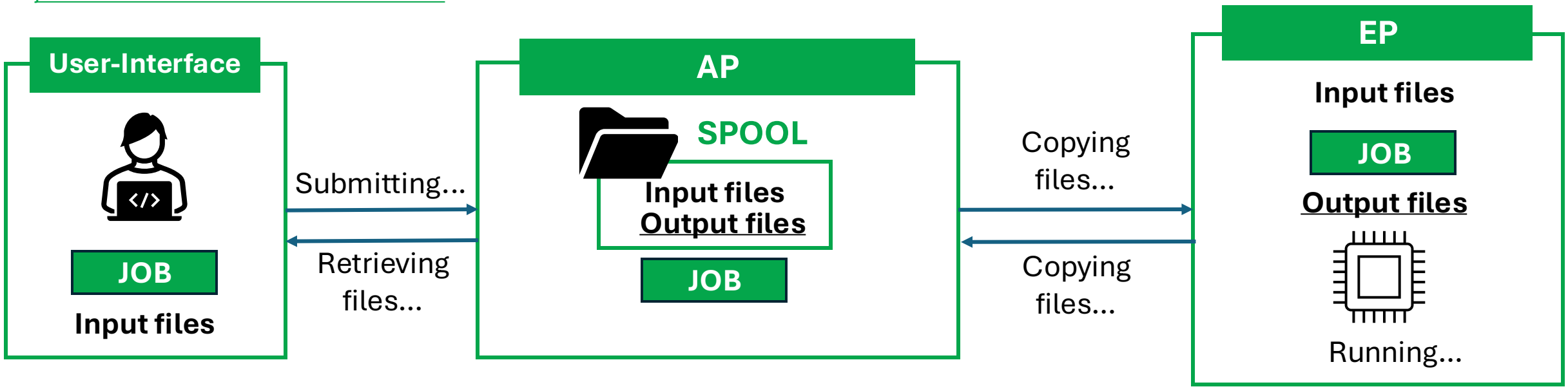
Useful Job Attributes	
ClusterID	ID of a Cluster of jobs
JobID	unique ID of a job [**ClusterID.Job_number**]
Owner	Local user executing the job
AcctGroup	Queue the owner/job belongs to
JobStatus	Number associated to the status of the job
Qdate	Unix timestamp of submission
JobstartDate	Unix timestamp of job start
CompletionDate	Unix timestamp of job completion
(Last)RemoteHost	EP slot where the job was/is running
HoldReason	String indicating the reason the job was put in HOLD state
RemoveReason	String indicating the reason the job was REMOVED

JOB STATUS	
1	IDLE
2	RUNNING
3	REMOVED
4	COMPLETED
5	HELD

These Job Attributes may note be all defined at the same time!!

HTCondor – Spooling

Spooling Mechanism in HTCondor
https://htcondor.readthedocs.io/en/latest/users-manual/submitting-a-remote-job.html#file-transfer-with-remote-submission



1. User submits job from machine without shared FS with AP
2. Sends files to the spool directory on the AP
3. The input files are copied to the EP
4. After completion the job output files are copied back in the spool directory
5. User can retrieve the output files from the AP spool

GRID Computing

HTCondor

Commands

HTCondor – Commands

condor_q [2]

- Shows info about the user's submitted jobs
- Allows to examine attributes of the jobs
- Can query using:
  - Constraints
  - JobID
  - Usernames
  - Job status

```
apascolinit1@ui-tier1 ~
$ condor_q


-- Schedd: sn01-htc.cr.cnaf.infn.it : <131.154.192.242:9618?... @ 07/10/24 15:10:03
OWNER        BATCH_NAME    SUBMITTED   DONE   RUN    IDLE  TOTAL JOB_IDS
apascolinit1 Sleep        7/10 12:58      _      _      3      3 408036.0-2

Total for query: 3 jobs; 0 completed, 0 removed, 3 idle, 0 running, 0 held, 0 suspended
Total for apascolinit1: 3 jobs; 0 completed, 0 removed, 3 idle, 0 running, 0 held, 0 suspended
Total for all users: 39917 jobs; 14676 completed, 0 removed, 15238 idle, 9994 running, 9 held, 0 suspended

apascolinit1@ui-tier1 ~
$ condor_q -af:jh owner jobstatus 'formattime(qdate)'*
 ID        owner        jobstatus formattime(qdate)
408036.0   apascolinit1 1         Wed Jul 10 12:58:52 2024
408036.1   apascolinit1 1         Wed Jul 10 12:58:52 2024
408036.2   apascolinit1 1         Wed Jul 10 12:58:52 2024
```

*** qdate → submission timestamp**

[2] **HTCondor doc on condor_q**
https://htcondor.readthedocs.io/en/latest/man-pages/condor_q.html

HTCondor – Commands

condor_submit [3]

- Command to submit new jobs
- -spool option to submit on remote **AP** with no shared FS [4]

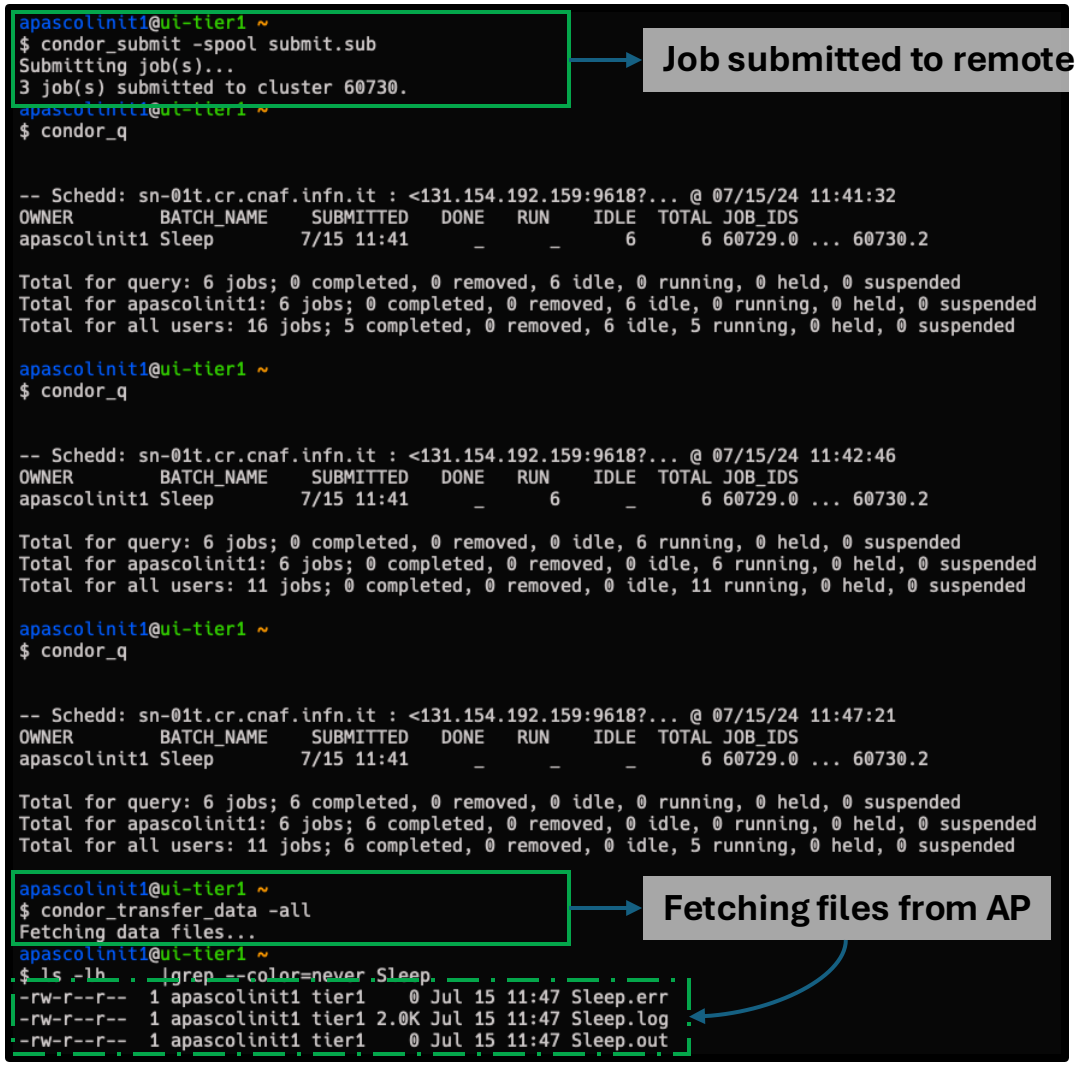
```
apascolinit1@ui-tier1 ~
$ condor_submit submit.sub
Submitting job(s)...
3 job(s) submitted to cluster 408339.
apascolinit1@ui-tier1 ~
$ condor_q


-- Schedd: sn01-htc.cr.cnaf.infn.it : <131.154.192.242:9618?... @ 07/10/24 15:16:22
OWNER        BATCH_NAME    SUBMITTED   DONE   RUN    IDLE  TOTAL JOB_IDS
apascolinit1 Sleep        7/10 12:58      _      _      6      6 408036.0 ... 408339.2

Total for query: 6 jobs; 0 completed, 0 removed, 6 idle, 0 running, 0 held, 0 suspended
Total for apascolinit1: 6 jobs; 0 completed, 0 removed, 6 idle, 0 running, 0 held, 0 suspended
Total for all users: 39840 jobs; 14651 completed, 0 removed, 15181 idle, 9999 running, 9 held, 0 suspended

apascolinit1@ui-tier1 ~
$ condor_q --nobatch


-- Schedd: sn01-htc.cr.cnaf.infn.it : <131.154.192.242:9618?... @ 07/10/24 15:16:31
 ID        OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
408036.0   apascolinit1    7/10 12:58   0+00:00:00 I  0    0.0 sleep 300
408036.1   apascolinit1    7/10 12:58   0+00:00:00 I  0    0.0 sleep 300
408036.2   apascolinit1    7/10 12:58   0+00:00:00 I  0    0.0 sleep 300
408339.0   apascolinit1    7/10 15:16   0+00:00:00 I  0    0.0 sleep 300
408339.1   apascolinit1    7/10 15:16   0+00:00:00 I  0    0.0 sleep 300
408339.2   apascolinit1    7/10 15:16   0+00:00:00 I  0    0.0 sleep 300

Total for query: 6 jobs; 0 completed, 0 removed, 6 idle, 0 running, 0 held, 0 suspended
Total for apascolinit1: 6 jobs; 0 completed, 0 removed, 6 idle, 0 running, 0 held, 0 suspended
Total for all users: 39840 jobs; 14654 completed, 0 removed, 15181 idle, 9996 running, 9 held, 0 suspended
```


[3] **HTCondor doc on condor_submit**
https://htcondor.readthedocs.io/en/latest/man-pages/condor_status.html
[4] **Spooling Mechanism in HTCondor**
https://htcondor.readthedocs.io/en/latest/users-manual/submitting-a-remote-job.html#file-transfer-with-remote-submission

HTCondor – Commands

condor_transfer_data [5]

- Command to retreive output files
- To be used when submitting to remote **AP** with no shared FS [4]



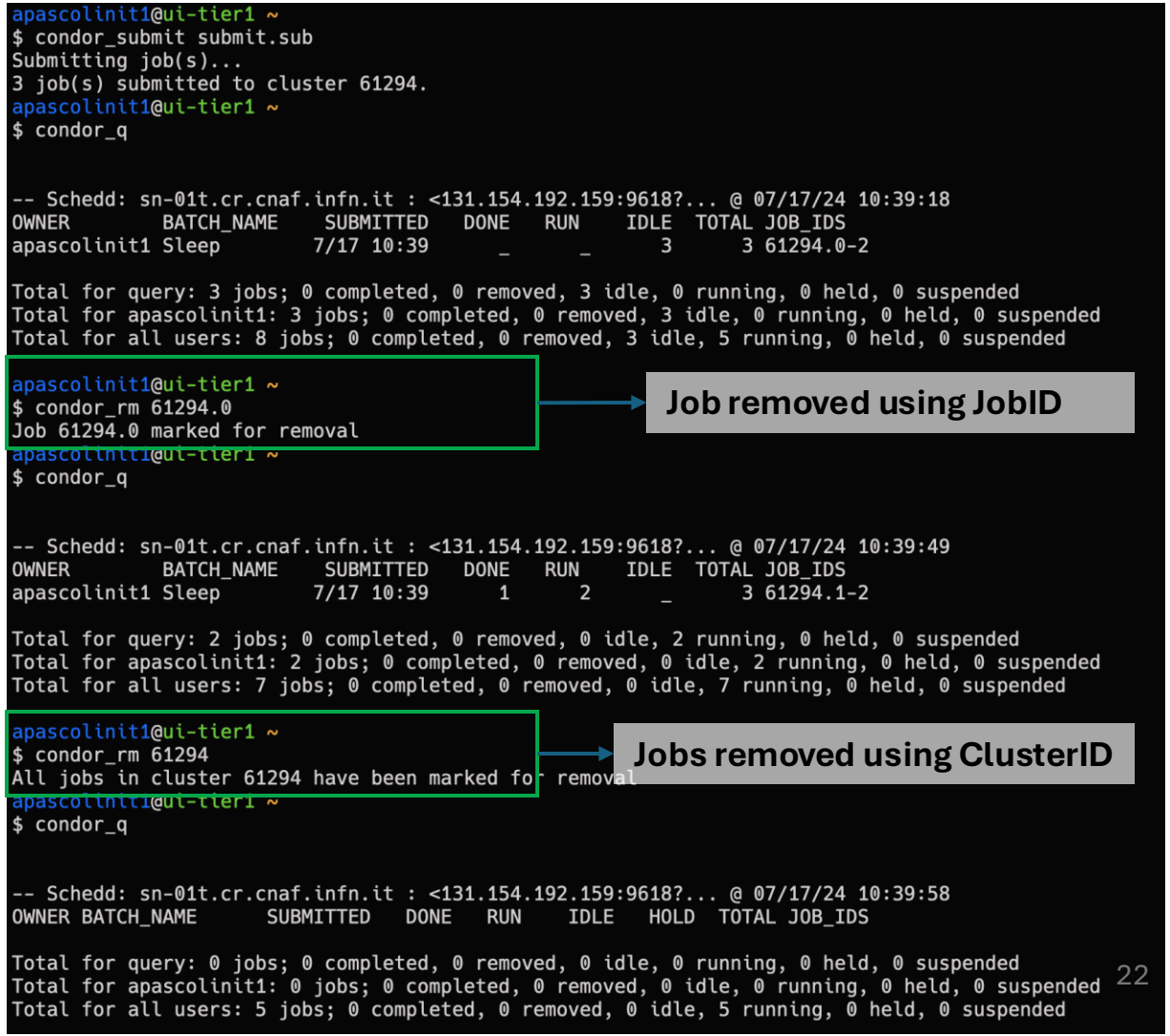
[4] **Spooling Mechanism in HTCondor**
https://htcondor.readthedocs.io/en/latest/users-manual/submitting-a-remote-job.html#file-transfer-with-remote-submission

[5] **HTCondor doc on condor_transfer_data**
https://htcondor.readthedocs.io/en/latest/man-pages/condor_transfer_data.html

HTCondor – Commands

condor_rm [6]

- Removes user's jobs
- A job to be removed can be specified by:
  - JobID
  - ClusterID
  - Owner
  - BatchName
  - Constraint



[6] **HTCondor doc on condor_transfer_data**
https://htcondor.readthedocs.io/en/latest/man-pages/condor_rm.html

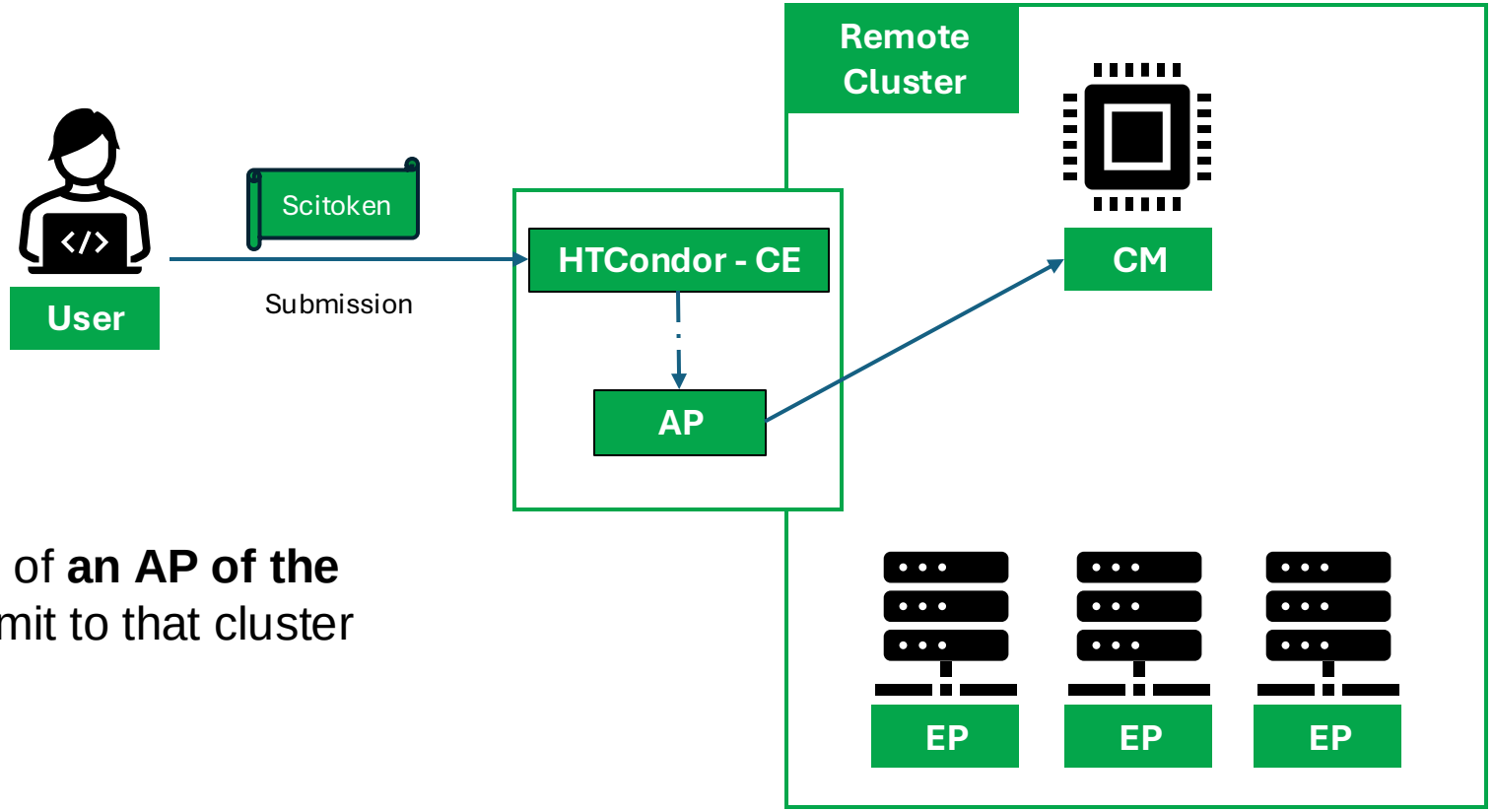
GRID Computing

HTCondor-CE

HTCondor-CE

HTCondor-CE (**Compute Entrypoint**) is an additional software to manage **GRID Jobs**

- AuthN/Z (**SCITOKEN** o SSL)
- Submission to a **remote cluster**



HTCondor-CE runs on the same machine of **an AP of the remote cluster** in order to be able to submit to that cluster

HTCondor-CE – AuthN/Z

To submit a job to an HTCondor-CE it is necessary to use a supported AuthN method:

- **SSL (VOMS-proxy)**
- **Scitokens (JWT)**

Token issued by IAM [7]

Needed scopes:
compute.create
compute.modify
compute.read
compute.cancel

```
apascolinit1@ui-tier1 ~
$ payload $(oidc-token htc-grid)
{
  "sub": "2c5255ba-9480-4815-aea0-88159f6602b7",
  "iss": "https://iam.cloud.infn.it/",
  "preferred_username": "apascolini",
  "client_id": "9b01323c-c74b-44f4-bffb-d439300f453e",
  "wlcg.ver": "1.0",
  "aud": "https://wlcg.cern.ch/jwt/v1/any",
  "nbf": 1720621050,
  "scope": "openid compute.create offline_access profile compute.read compute.cancel compute.modify wlcg wlcg.groups",
  "name": "Alessandro Pascolini",
  "exp": 1720624650,
  "iat": 1720621050,
  "jti": "35503676-f800-43dc-9e51-3fc605b26932",
  "wlcg.groups": [
    "/user-support"
  ]
}
```

[7] Tier-1 guide on token submission
https://confluence.infn.it/display/TD/HTCondor+jobs#HTCondorjobs-SubmitgridjobsSubmitgridjobswithoutenvironmentmodules

HTCondor-CE – User Mapping

HTCondor-CE relies on **static or dynamic (PLUGINS) mapping** to **associate a SCITOKEN to a local user** on the remote cluster

- Users that want to submit to a CE should confirm that it is allowed (i.e. **they are mapped**)

```
apascolinit1@ui-tier1 ~
$ payload $(oidc-token htc-grid)
{
  "sub": "2c5255ba-9480-4815-aea0-88159f6602b7",
  "iss": "https://iam.cloud.infn.it/",
  "preferred_username": "apascolini",
  "client_id": "9b01323c-c74b-44f4-bffb-d439300f453e",
  "wlcg.ver": "1.0",
  "aud": "https://wlcg.cern.ch/jwt/v1/any",
  "nbf": 1720621050,
  "scope": "openid compute.create offline_access profile compute.read compute.cancel compute.modify wlcg wlcg.groups",
  "name": "Alessandro Pascolini",
  "exp": 1720624650,
  "iat": 1720621050,
  "jti": "35503676-f800-43dc-9e51-3fc605b26932",
  "wlcg.groups": [
    "/user-support"
  ]
}
```

```
[root@ce01t-htc ~]# cat /etc/condor-ce/mapfiles.d/00-training.conf
# users from IAM-Cloud
SCITOKENS /^https:\/\/iam\.cloud\.infn.it\/,/ PLUGIN:A

# Daniele L. & Alessandro P.
SCITOKENS "https://iam.cloud.infn.it/,039e2956-2e56-44c6-987c-25b240430d89" datacloud-proto
SCITOKENS "https://iam.cloud.infn.it/,68bfc01b-f8ad-440d-98f3-83d2d9c00620" datacloud-proto
```

→ PLUGIN-based mapping

→ Static mapping to the same user

HTCondor-CE – Commands

With HTCondor-CE all the previous commands can be used

Some small changes…

condor_q → condor_q **–pool <ce-fqdn>:9619 –name <ce-fqdn>**

condor_submit → condor_submit **–pool <ce-fqdn>:9619 –remote <ce-fqdn>**

Useful commands:

export _condor_CONDOR_HOST=<ce-fqdn>:9619

export _condor_SCHEDD_HOST=<ce-fqdn>

alias condor_submit='condor_submit –spool'

HTCondor-CE – Commands

With HTCondor-CE all the previous commands can be used

```
apascolinit1@ui-tier1 ~
$ export BEARER_TOKEN=$(oidc-token htc-grid)
apascolinit1@ui-tier1 ~
$ export _condor_SEC_CLIENT_AUTHENTICATION_METHODS=SCITOKENS
apascolinit1@ui-tier1 ~
$ export _condor_CONDOR_HOST=ce01t-htc.cr.cnaf.infn.it:9619
apascolinit1@ui-tier1 ~
$ export _condor_SCHEDD_HOST=ce01t-htc.cr.cnaf.infn.it
apascolinit1@ui-tier1 ~
$ condor_q


-- Schedd: ce01t-htc.cr.cnaf.infn.it : <131.154.192.69:9619?... @ 07/10/24 17:54:36
OWNER BATCH_NAME      SUBMITTED   DONE   RUN    IDLE   HOLD  TOTAL JOB_IDS

Total for query: 0 jobs; 0 completed, 0 removed, 0 idle, 0 running, 0 held, 0 suspended
Total for apascolinius: 0 jobs; 0 completed, 0 removed, 0 idle, 0 running, 0 held, 0 suspended
Total for all users: 365 jobs; 364 completed, 0 removed, 0 idle, 0 running, 1 held, 0 suspended
```

Export to authenticate to the CE

Successfully contacted the CE

HTCondor-CE – Commands

With HTCondor-CE all the previous commands can be used

token-submit.sub

```
# Unix submit description file
# sleep.sub -- simple sleep job

# Grid-specific options
+owner                  = undefined
scitokens_file          = $ENV(HOME)/token

batch_name              = Token-Sleep
executable              = /usr/bin/sleep
arguments               = 300
log                     = Sleep.log
output                  = Sleep.out
error                   = Sleep.err

# require to run on LINUX machines
requirements = OpSys == "LINUX"

# ask for 1 core, 1024MB memory and 1024kB disk
request_cpus    = 1
request_memory = 1024M
request_disk    = 10240K

# submitting 3 jobs (default 1, if no number specified after queue)
queue 3
```

GRID stuff

NOTA
some CEs may override the requirements specified in the submit file

```
apascolinit1@ui-tier1 ~
$ alias condor_submit='condor_submit -spool'
apascolinit1@ui-tier1 ~
$ MASK=$(umask); umask 0077 ; echo $BEARER_TOKEN > $HOME/token; umask $MASK
apascolinit1@ui-tier1 ~
$ condor_submit token-submit.sub
Submitting job(s)...
3 job(s) submitted to cluster 9424.
apascolinit1@ui-tier1 ~
$ condor_q


-- Schedd: ce01t-htc.cr.cnaf.infn.it : <131.154.192.69:9619?... @ 07/10/24 18:10:58
OWNER        BATCH_NAME     SUBMITTED   DONE   RUN    IDLE  TOTAL JOB_IDS
apascolinius Token-Sleep   7/10 18:10      _      _       3      3 9424.0-2

Total for query: 3 jobs; 0 completed, 0 removed, 3 idle, 0 running, 0 held, 0 suspended
Total for apascolinius: 3 jobs; 0 completed, 0 removed, 3 idle, 0 running, 0 held, 0 suspended
Total for all users: 368 jobs; 364 completed, 0 removed, 3 idle, 0 running, 1 held, 0 suspended
```

Thanks for your attention!

…any questions?