# Job Accounting with HTCondor in Heterogeneous Systems

HEPiX Spring 2023 at ASGC, Taipei, Taiwan

Marcus Ebert
on behalf of the
HEP-RC group at the University of Victoria, Canada

# Why looking into it?

- (WLCG) accounting needs CPU time and number of cores used by user jobs
- CPU time used for the same job varies depending on used CPU
- need a scaling factor depending on CPU type used by a job
- WLCG/EGI choose HEPSpec06 (HS06)
  - HEPScore will replace HS06
  - different value for different CPUs/configurations
- sites usually have different machines
  - new purchases do not replace everything that existed before
  - with time a large heterogeneous system can develop

# Why looking into it?

- ARC-CE and probably others allow for one benchmark value in the configuration
  - at least a few years back
- average over whole site needs to be made
  - taking HS06 for different CPU types and number of cores for each type into account
- works well when all CPUs are used and all jobs are of the same kind/from same VO or user group
- becomes incorrect when
  - not all machines are up or used
  - different VOs run jobs, likely not evenly distributed over all machine types
  - throttling CPUs due to power saving needs
- using a HS06 value per job instead per site would be better

# Why we look into it?

- we do not use static, bare-metal worker nodes

- running on VMs in different clouds
  - Canada, USA, Australia, Europe
  - large variation of CPUs and machine configurations

- running Belle-II and ATLAS jobs as service for other sites
- running single HTCondor for each experiment
  - no HTCondor-CE or ARC-CE used

- using a HS06 value per job instead per site is needed due to reporting for different sites

# How to use HS06 per job

- HTCondor worker node needs to know its HS06 value
- HS06 value gets added to job attributes when entering a machine
- condor_history can be parsed to get all information needed for accounting
- for EGI/WLCG accounting, using apel's ssmsend container to report

# How we do it with HTCondor

## machine needs to know its HS06 value

- make values known to worker node
  - specific for the CPU/type on the worker node
    - run HS06 previously and keep table, use value depending on machine
    - run parts during machine startup or before job starts
      - DB12, parts of HEPScore

- add those to the HTCondor config on the worker node, e.g. */etc/condor/config.d/benchmark*:
  *HEPSPEC = "30.786"*
  *B2BMK = "2.5014"*
  *HS06EQ = "25.410"*
  *STARTD_ATTRS = $(STARTD_ATTRS) HEPSPEC B2BMK HS06EQ*

- (re)start HTCondor (all done during worker node boot)

# How we do it with HTCondor

## HS06 value gets added to job attributes when entering a machine

- use job wrapper file to prepare job environment
  - executed as part of the job from a HTCondor point of view

```
hs06bmk=$(condor_config_val HEPSPEC)
b2bmk=$(condor_config_val B2BMK)
hs06eq=$(condor_config_val HS06EQ)

condor_chirp set_job_attr HEPSPEC $hs06bmk
condor_chirp set_job_attr B2BMK $b2bmk
condor_chirp set_job_attr HS06EQ $hs06eq
```

- *WantIOProxy* needs to be set to have *condor_chirp* working
  - done in condor server config

```
WantIOProxy = True
SUBMIT_ATTRS = $(SUBMIT_ATTRS) WantIOProxy
```

# How we do it with HTCondor

## condor_history can be parsed

- Any value added on the worker nodes via *condor_chirp* becomes part of the job attributes
- Jobs have all needed information
- have script that parses “*condor_history -long*”
  - e.g. daily cronjob looking for jobs finished one day ago
  - collect all information needed for accounting
    - host, CPU time, wall time, VO, HS06,....
  - output information in a format needed for the reporting
    - WLCG: APEL conform text file

# How we do it with HTCondor

## condor_history can be parsed

- WLCG: APEL conform text file:

```
APEL-individual-job-message: v0.3
Site: CA-UVic-Cloud
SubmitHost: bellecs.heprc.uvic.ca
MachineName: belle--arbutus--3456932580--81292352679336-5.novalocal
LocalJobId: 1585696.77
LocalUserId: "dirac"
VO: belle
WallDuration: 8809
CpuDuration: 5696
Processors: 1
StartTime: 1675287181
EndTime: 1675295990
ServiceLevelType: HEPSPEC
ServiceLevel: 19.00
%%
Site: CA-UVic-Cloud
SubmitHost: bellecs.heprc.uvic.ca
...
```

# How we do it with HTCondor

## EGI/WLCG accounting

```
sudo podman run --rm -d --entrypoint ssmsend \
    -v apel_container/config/sender.cfg:/etc/apel/sender.cfg \
    -v test_out:/var/spool/apel/outgoing \
    -v /etc/grid-security:/etc/grid-security \
    -v log/:/var/log/apel \
    stfc/ssm
```

<----------- contains site/host specific configurations
<----------- maps local dir with accounting files to container dir
<----------- makes hostcer/key and others available in container
<----------- keeps log files

```
[root@accounting apel_container]# podman image ls
REPOSITORY          TAG      IMAGE ID       CREATED        SIZE
docker.io/stfc/ssm  latest   f802e4fb3088   23 months ago  908 MB
```

# Summary

- HTCondor job attributes have nearly all information needed for WLCG accounting
- missing information, like benchmark values, can be added via condor_chirp
  - approach usable for anything that would be useful to be associated with specific jobs
- condor_history has information about finished jobs
  - each job contains all needed information
    ------> parse for information needed for a specific accounting