HTCondor Workshop Autumn 2025

Operating HTCondor for CPU and GPU Workloads at IFIC

Miguel G. Folgado
Javier Sánchez
Matias Salinero

Who We Are?

Instituto de Física Corpuscular (IFIC)

Joint Center University of Valencia & CSIC

- Research in High Energy Physics, Nuclear Physics, Medical Physics and technological applications.
- Strong participation in major international collaborations: LHC/CERN, neutrino experiments, astroparticle physics, among others.
- A multidisciplinary environment where fundamental science and technology meet.

Who We Are?

114
Senior Researchers

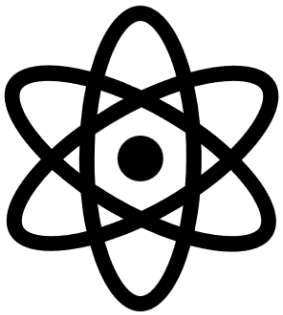
69
Postdoctoral Researchers

105
Students

59
Technical Personnel

14
Research groups

What Brings Us Together?



IT Infrastructure

Artemisa

GLUON
IFIC General Computing Infrastructure

ATLAS EXPERIMENT

Flavour and Origin of Matter
Sabor y Origen de la Materia

Storage	Nodes	CPU	GPU
361 TB	37	2,008	53
848 TB	26	2,592	N/A
6 PB	74	3,256	N/A
418 TB	23	1,368	N/A

Artemisa

GLUON
IFIC General Computing Infrastructure

ATLAS EXPERIMENT

Flavour and Origin of Matter
Sabor y Origen de la Materia

FS	OS	Scheduler	Monitoring
lustre		HTCondor Software Suite	influxdb + Grafana
lustre		HTCondor Software Suite	Prometheus + Grafana
lustre		HTCondor Software Suite	influxdb + Grafana
lustre		slurm	Prometheus + Grafana

Artemisa

GLUON
IFIC General Computing Infrastructure

ATLAS EXPERIMENT

Flavour and Origin of Matter
Sabor y Origen de la Materia

FS	OS	Scheduler	Monitoring
lustre		HTCondor Software Suite	influxdb Grafana
lustre		HTCondor Software Suite	Prometheus Grafana
lustre		HTCondor Software Suite	influxdb Grafana
lustre		slurm	Prometheus Grafana

Artemisa

GLUON
IFIC General Computing Infrastructure

ATLAS EXPERIMENT

Flavour and Origin of Matter
Sabor y Origen de la Materia

Metrics 

Metrics 

LDAP

User Acces
mluiXX.ific.uv.es
>_ SSH

User Acces
gluiXX.ific.uv.es
>_ SSH

Compute 

Compute 

Software Layer 

lustre

ML Proyects

Group Proyects

/lhome

Management Layer
ANSIBLE

GLUON

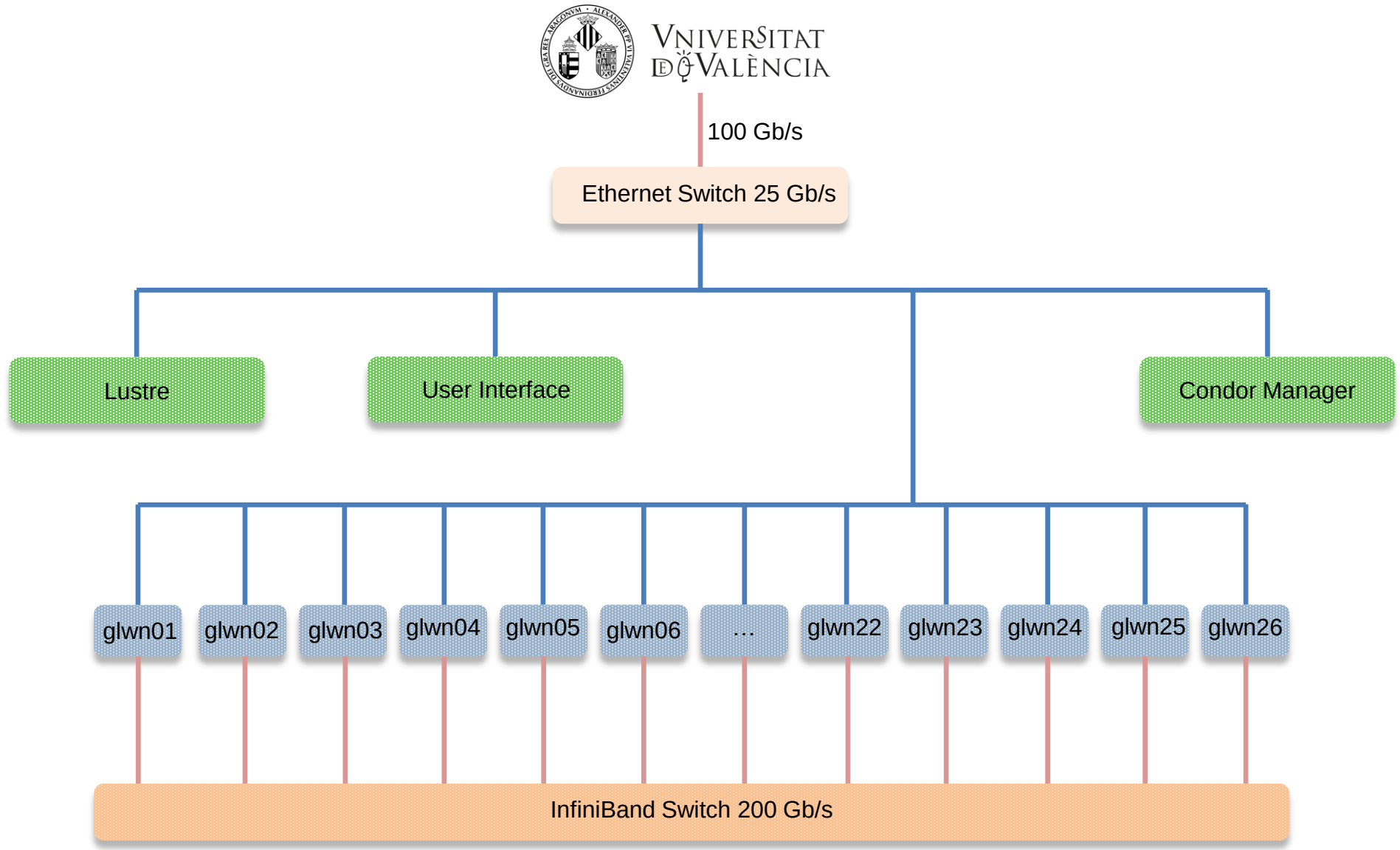
Storage	Nodes	CPU	GPU
361 TB	37	2,008	53
848 TB	26	2,592	N/A
6 PB	74	3,256	N/A
418 TB	23	1,368	N/A

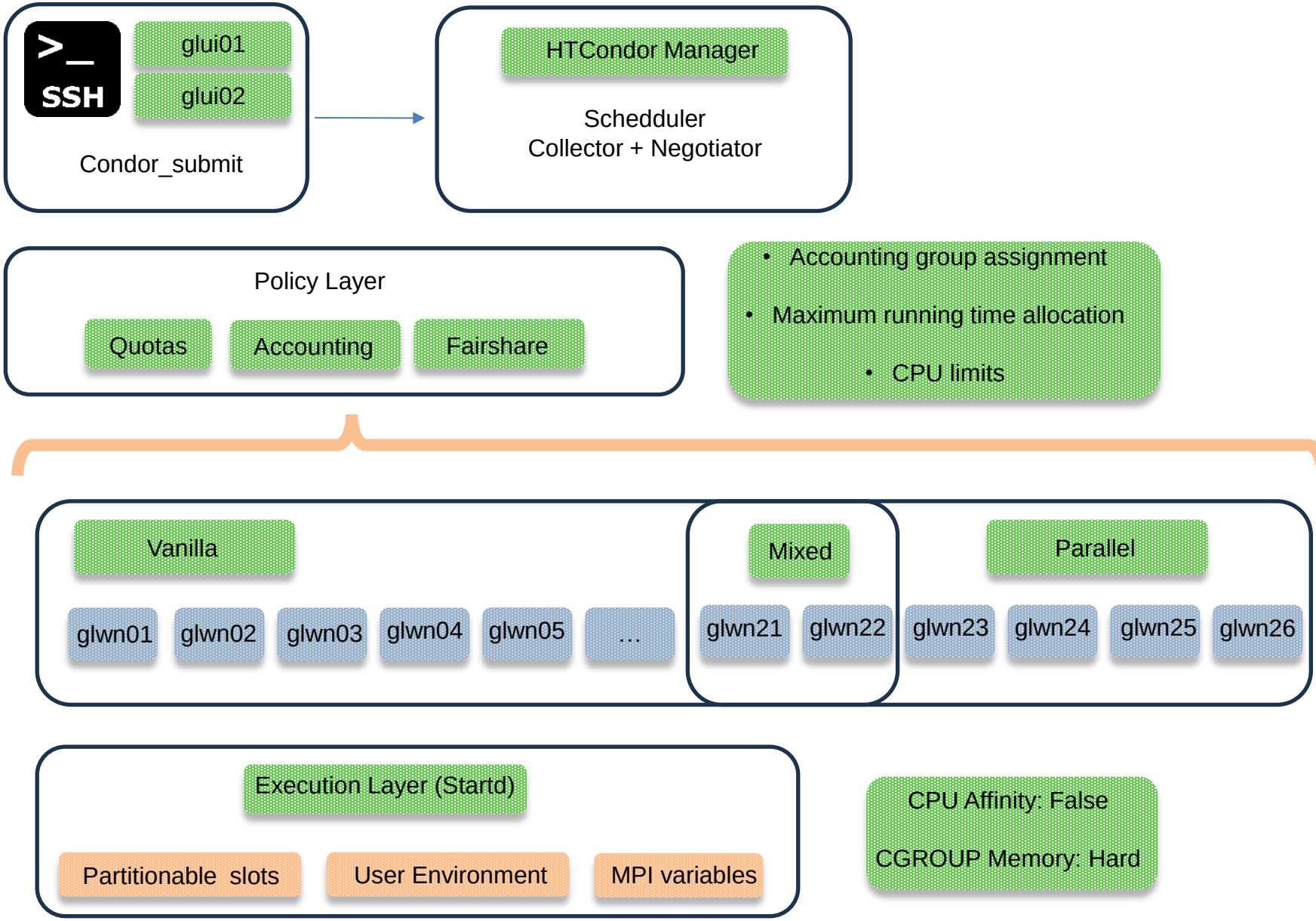
Artemisa

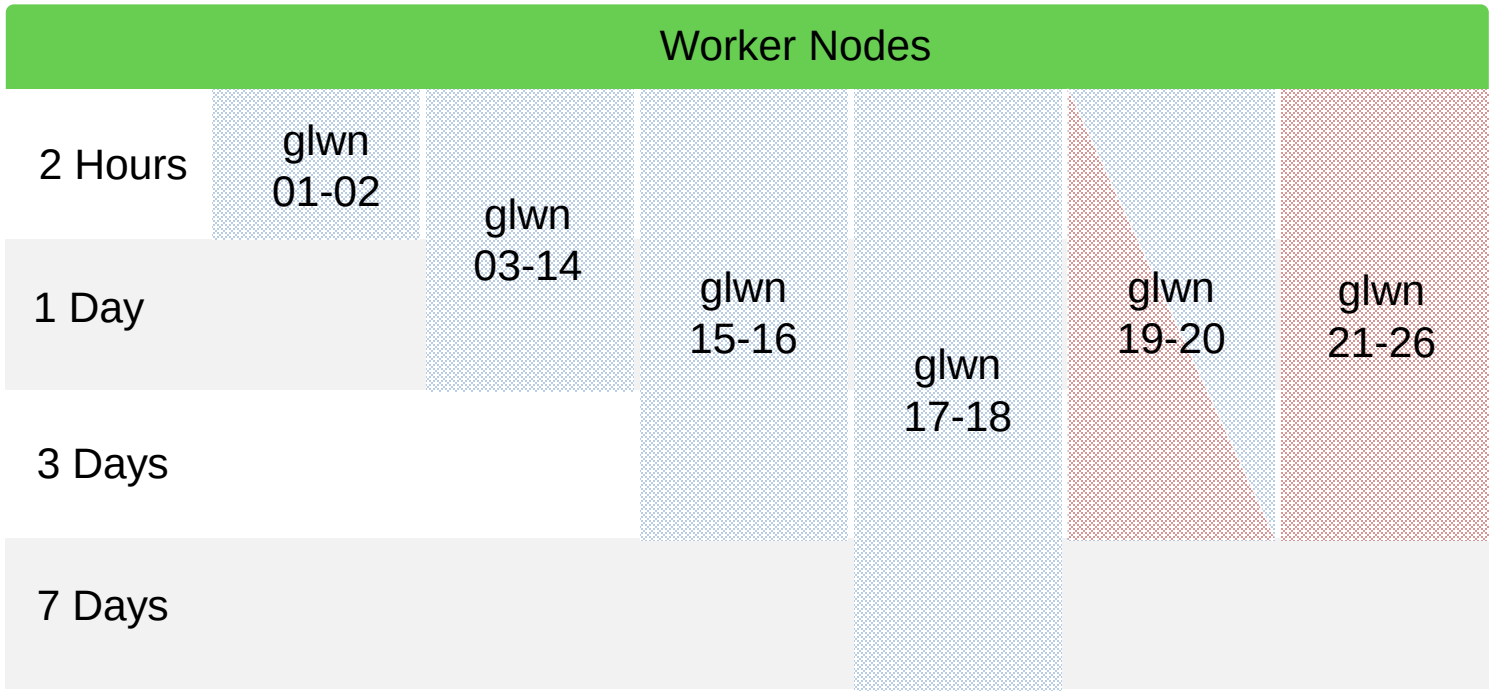
GLUON
IFIC General Computing Infrastructure

ATLAS EXPERIMENT

Flavour and Origin of Matter
Sabor y Origen de la Materia


VNIVERSITAT DE VALÈNCIA
100 Gb/s
Ethernet Switch 25 Gb/s
Lustre
User Interface
Condor Manager
glwn01
glwn02
glwn03
glwn04
glwn05
glwn06
…
glwn22
glwn23
glwn24
glwn25
glwn26
InfiniBand Switch 200 Gb/s

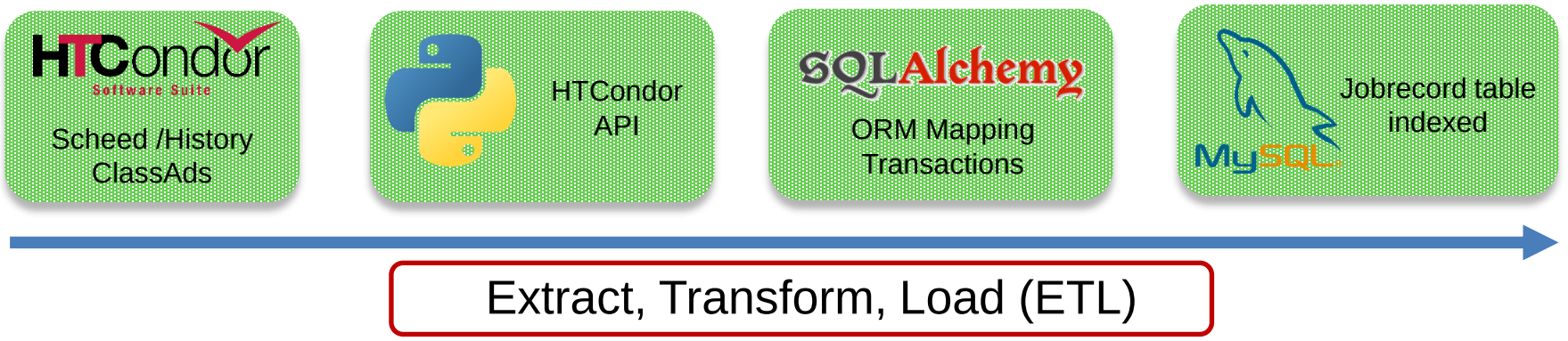

SSH
glui01
glui02
Condor_submit
HTCondor Manager
Schedduler
Collector + Negotiator
Policy Layer
Quotas
Accounting
Fairshare
• Accounting group assignment
• Maximum running time allocation
• CPU limits
Vanilla
glwn01
glwn02
glwn03
glwn04
glwn05
…
Mixed
glwn21
glwn22
Parallel
glwn23
glwn24
glwn25
glwn26
Execution Layer (Startd)
Partitionable slots
User Environment
MPI variables
CPU Affinity: False
CGROUP Memory: Hard



Key points on scheduling policy

- All nodes are equipped with InfiniBand → no need for complex labels or constraints.
- The split between *Vanilla* and *Parallel* is done purely via **job universe** and **runtime limits**.
- This makes the policy simple, flexible, and easy to adjust quickly if user needs change.

Accounting pipeline



What do we do with this data?

- Currently: periodic reports and usage statistics (per user, group, walltime, efficiency).
- Potential: real-time dashboards, advanced analytics, integration with monitoring stack.

Next steps in Job Accounting

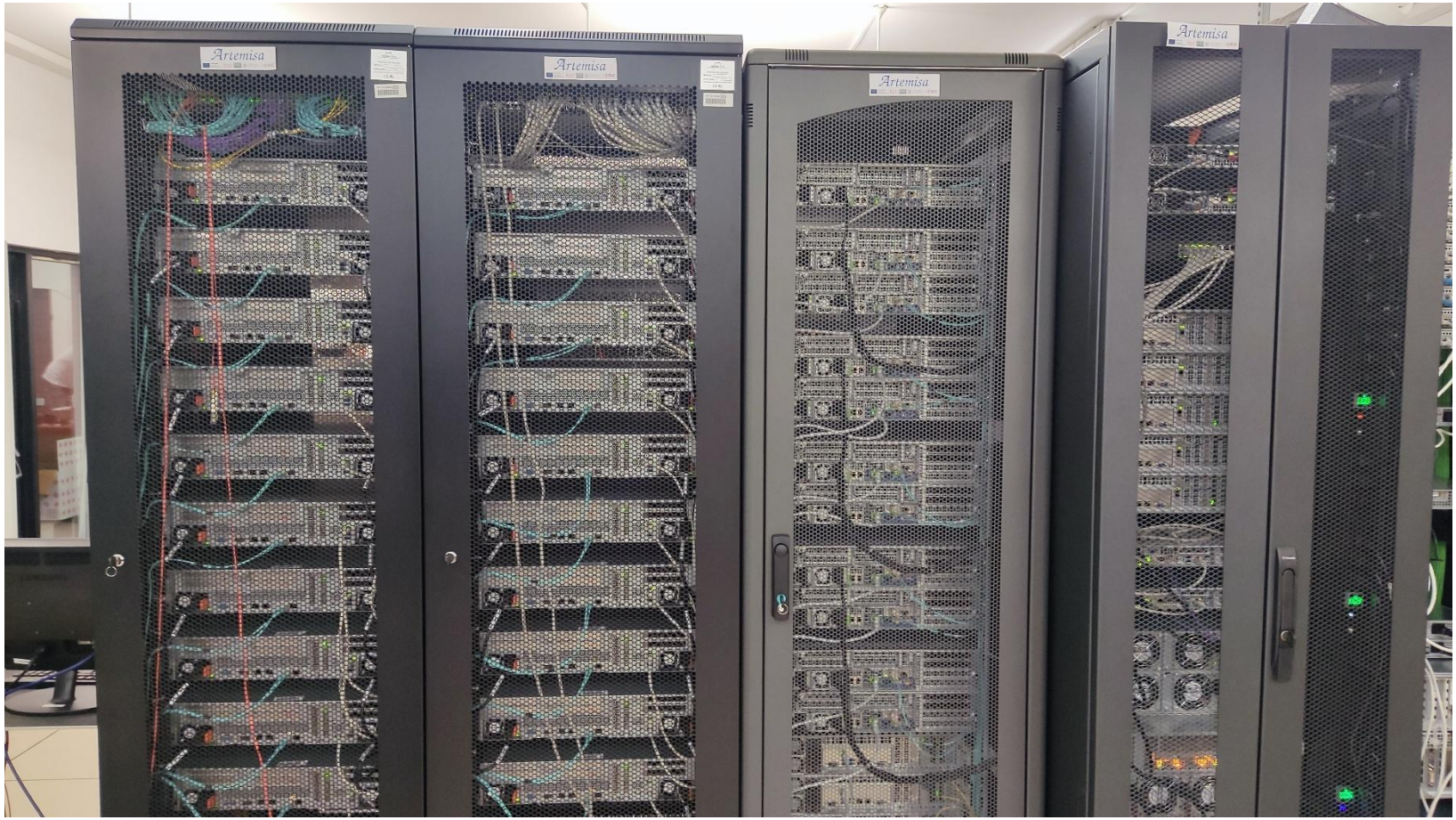
Real-time Monitoring

Prometheus + Grafana

- **Real-time dashboards** for jobs and usage
- **Seamless integration** with our existing monitoring stack
- **Custom Prometheus exporter** under development for HTCondor

ELK (Elastic + Logstash + Kibana)

- **Advanced historical search** on job records
- **Powerful analytics** (filters, aggregations, heatmaps)
- **Integration with system & audit logs** (compliance, debugging)
- **Heavier deployment/maintenance cost**



Artemisa

Storage	Nodes	CPU	GPU	
361 TB	37	2,008	53	Artemisa
848 TB	26	2,592	N/A	GLUON IFIC General Computing Infrastructure
6 PB	74	3,256	N/A	ATLAS EXPERIMENT
418 TB	23	1,368	N/A	Flavour and Origin of Matter Sabor y Origen de la Materia

Storage	Nodes	CPU	GPU
361 TB	37	2,008	53

4 x P100

19 x V100

26 x A100

4 x H100

Storage	Nodes	CPU	GPU
361 TB	37	2,008	53

NVIDIA

Artemisa

4 x P100

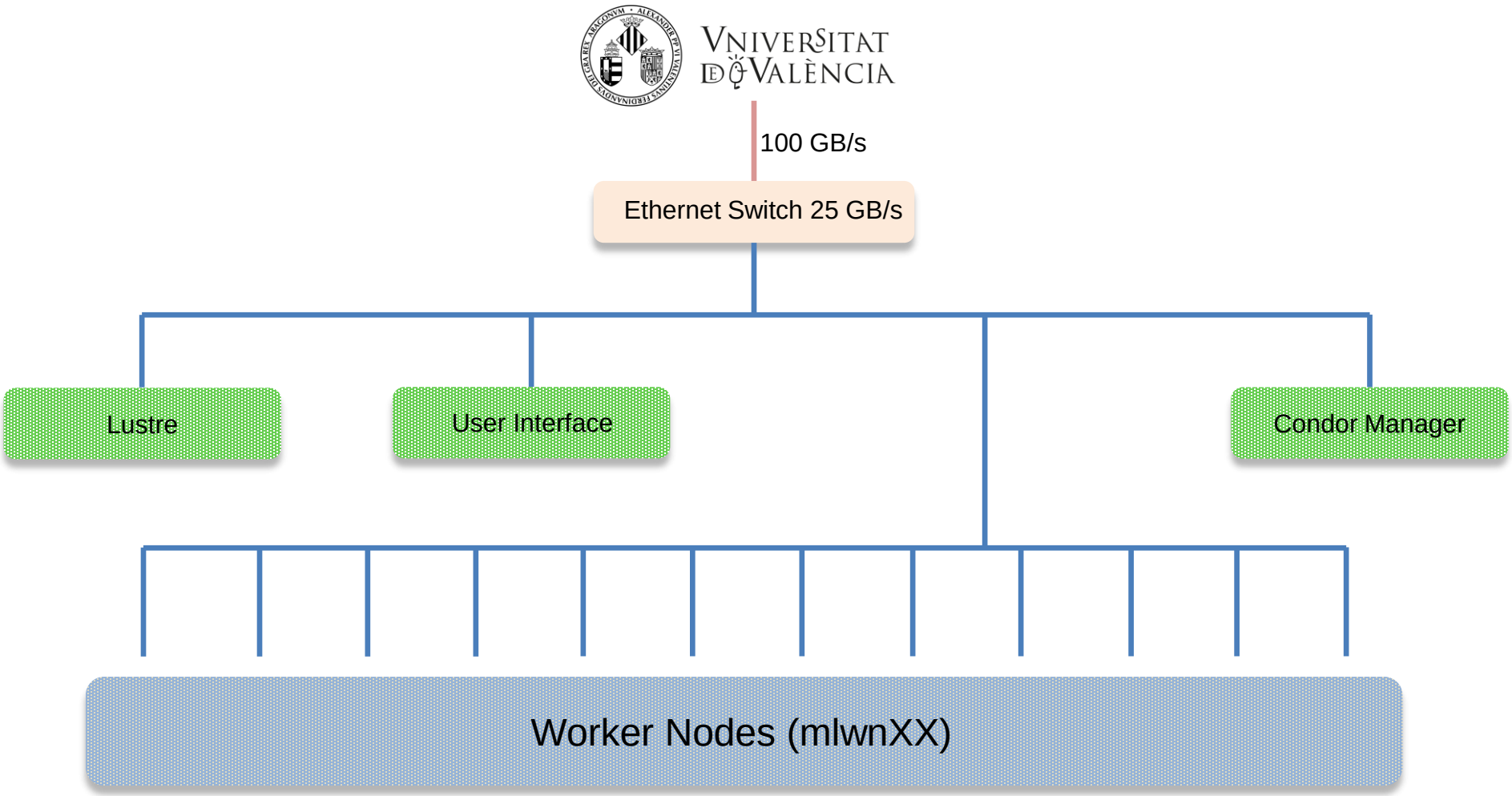
19 x V100

26 x A100

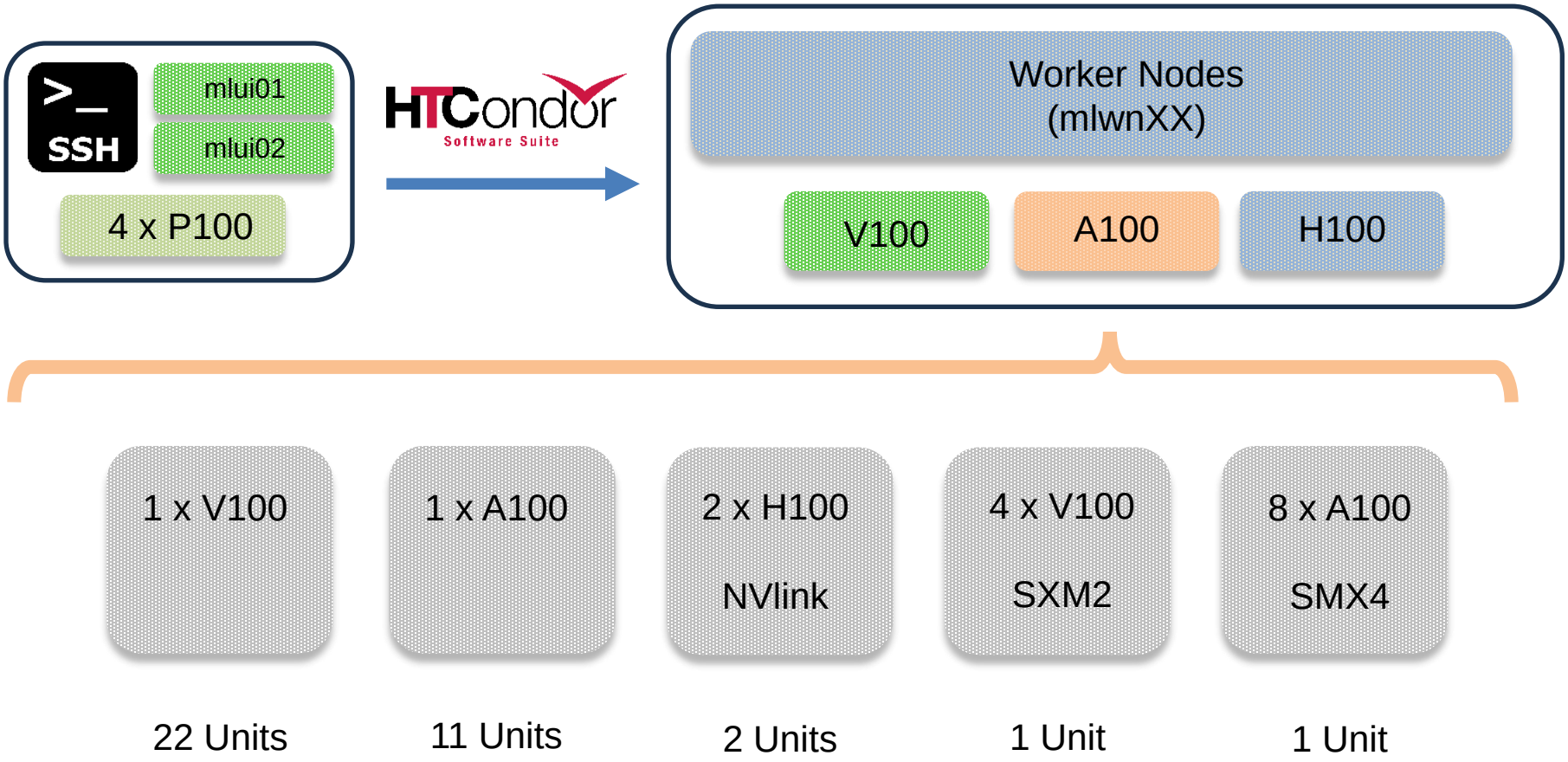
4 x H100

40 x H200

2026


VNIVERSITAT DE VALÈNCIA
100 GB/s
Ethernet Switch 25 GB/s
Lustre
User Interface
Condor Manager
Worker Nodes (mlwnXX)

Distribution of GPUs



Example partition 1 GPU

Slot Type 1 (CPU)

CPU: 72 MEM: 40%

Partitionable

Slot Type 2 (GPU)

CPU: 56 MEM: 60%

GPU: 1

Non Partitionable!

1 x A100 / 128 CPUs

```
SLOT_TYPE_1 = cpus=72, memory=40%
SLOT_TYPE_1_PARTITIONABLE = True
NUM_SLOTS_TYPE_1 = 1
SLOT_TYPE_1_CONSUMPTION_POLICY = True
SLOT_TYPE_1_CONSUMPTION_CPUS = quantize(TARGET.RequestCpus,{8})
SLOT_TYPE_1_CONSUMPTION_MEMORY =
quantize(TARGET.RequestMemory,{16384})
```

```
MACHINE_RESOURCE_NAMES = Gpus
MACHINE_RESOURCE_Gpus = 1
SLOT_TYPE_2 = cpus=56, memory=60%, Gpus=1
SLOT_TYPE_2_PARTITIONABLE = False
NUM_SLOTS_TYPE_2 = 1
SLOT_TYPE_2_CONSUMPTION_POLICY = True
SLOT_TYPE_2_CONSUMPTION_CPUS = quantize(TARGET.RequestCpus,{8})
SLOT_TYPE_2_CONSUMPTION_MEMORY =
quantize(TARGET.RequestMemory,{32768})
SLOT_TYPE_2_CONSUMPTION_GPUS =
ifthenelse(TARGET.RequestGpus =?= undefined,0,TARGET.RequestGpus)
SLOT2_STARTD_ATTRS = HasNvidiaA100
SLOT2_HasNvidiaA100 = True
```

```
START = (NODE_IS_HEALTHY is true) && (StartJobs is true) && \
        ifThenElse(TARGET.RequestGpus isnt UNDEFINED, \
        (TARGET.RequestGpus == 1) && MY.SlotID == 2 , MY.SlotID == 1))
```

Example partition 4 GPU

Slot Type 1 (CPU)

CPU: 56 MEM: 40%

Partitionable

Slot Type 2 (GPU)

CPU: 56 MEM: 60%
GPU: 4

Partitionable!

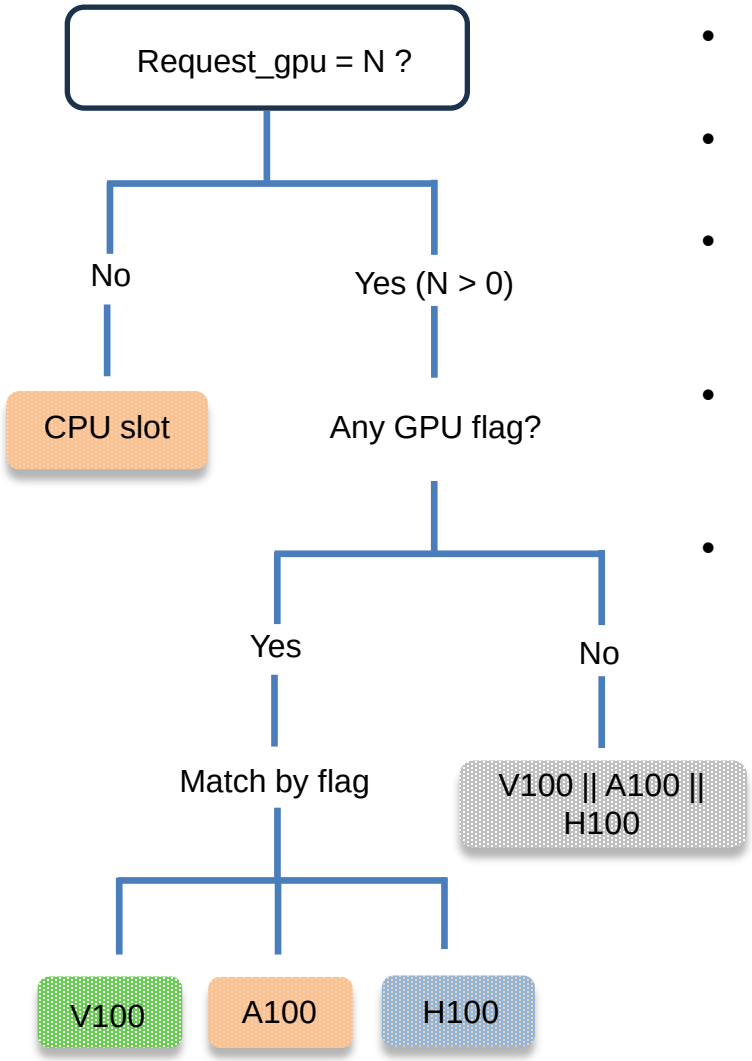
4 x V100 / 112 CPUs

```
SLOT_TYPE_1 = cpus=56, memory=40%
SLOT_TYPE_1_PARTITIONABLE = True
NUM_SLOTS_TYPE_1 = 1
SLOT_TYPE_1_CONSUMPTION_POLICY = True
SLOT_TYPE_1_CONSUMPTION_CPUS = quantize(TARGET.RequestCpus,{8})
SLOT_TYPE_1_CONSUMPTION_MEMORY =
quantize(TARGET.RequestMemory,{16384})
```

```
MACHINE_RESOURCE_NAMES = Gpus
MACHINE_RESOURCE_Gpus = 4
SLOT_TYPE_2 = cpus=56, memory=60%, Gpus=4
SLOT_TYPE_2_PARTITIONABLE = True
NUM_SLOTS_TYPE_2 = 1
SLOT_TYPE_2_CONSUMPTION_POLICY = True
SLOT_TYPE_2_CONSUMPTION_CPUS = quantize(TARGET.RequestCpus,{8})
SLOT_TYPE_2_CONSUMPTION_MEMORY =
quantize(TARGET.RequestMemory,{32768})
SLOT_TYPE_2_CONSUMPTION_GPUS =
ifthenelse(TARGET.RequestGpus =?= undefined,0,TARGET.RequestGpus)
SLOT2_STARTD_ATTRS = HasNvidiaA100
SLOT2_HasNvidiaA100 = True
```

```
START = (NODE_IS_HEALTHY is true) && (StartJobs is true) && \
        ifThenElse(TARGET.RequestGpus isnt UNDEFINED, \
        (TARGET.RequestGpus == 2 || TARGET.RequestGpus == 4) && MY.SlotID == 2 , MY.SlotID == 1))
```

GPU Matching Policy



- Users request GPUs with request_gpus = N
- Optional flags restrict jobs to specific GPU models.
- These flags map to ClassAd attributes (HasNvidiaV100/A100/H100).
- If no flag is set HTCondor assigns **any available GPU slot**
- If request_gpus is not defined → job runs in the CPU pool only

```
+UseV100 = True
requirements = (TARGET.HasNvidiaV100 == True) && \
(TARGET. RequestGpus >= N)

+UseA100 = True
requirements = (TARGET.HasNvidiaA100 == True) && \
(TARGET. RequestGpus >= N)

+UseH100 = True
requirements = (TARGET.HasNvidiaH100 == True) && \
(TARGET. RequestGpus >= N)
```

IFIC Computing Clusters: Overview & Future work

Common Layer

- Shared /home on Lustre, CVMFS software stack
- Central auth (LDAP) & config (Ansible)

GLUON
IFIC General Computing Infrastructure

- CPU cluster with 25 Gb Ethernet (Lustre), 200 Gb IB (compute)
- Vanilla / Parallel split, fairshare & runtime policy
- Accounting: ClassAds Python ETL → MySQL → usage reports

Artemisa

- CPU partitionable slots + dedicated GPU slots (V100/A100/H100)
- Consumption policy (CPU=8, Mem=32 GB, GPU=exact)
- GPU flags (+UseV100/+UseA100/+UseH100) map to ClassAd attributes

IFIC Computing Clusters: Overview & Future work

Monitoring & Next Steps

- Grafana central server; GLUON via Prometheus

- Work in progress: Condor exporter for real-time dashboards

- ELK under evaluation for advanced analytics