Lightweight on-demand computing with Elasticluster and Nordugrid ARC

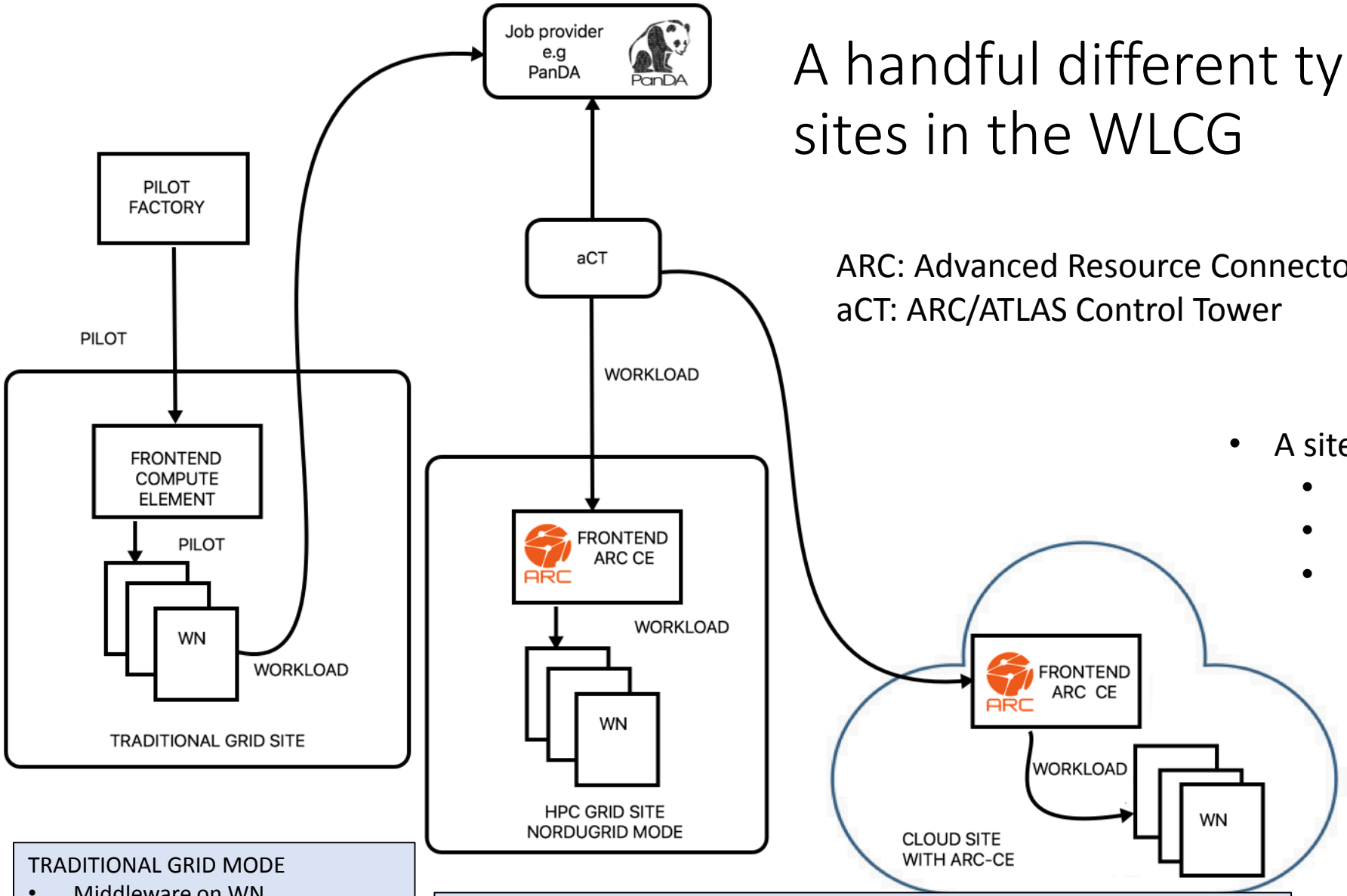
On behalf of the ATLAS Collaboration

Maiken Pedersen, University of Oslo (NO)
David Cameron, University of Oslo (NO)
Andrej Filipcic, Jozef Stefan Institute (SI)

Overview

- Types of ATLAS sites in WLCG including the Nordugrid ARC and aCT INTERNAL mode grid site
- Overview of the different ARC-CE submission interfaces
- Setup and configuration of OpenStack grid site with Elasticluster
- INTERNAL submission interface in use
- Conclusion

A handful different types of ATLAS sites in the WLCG



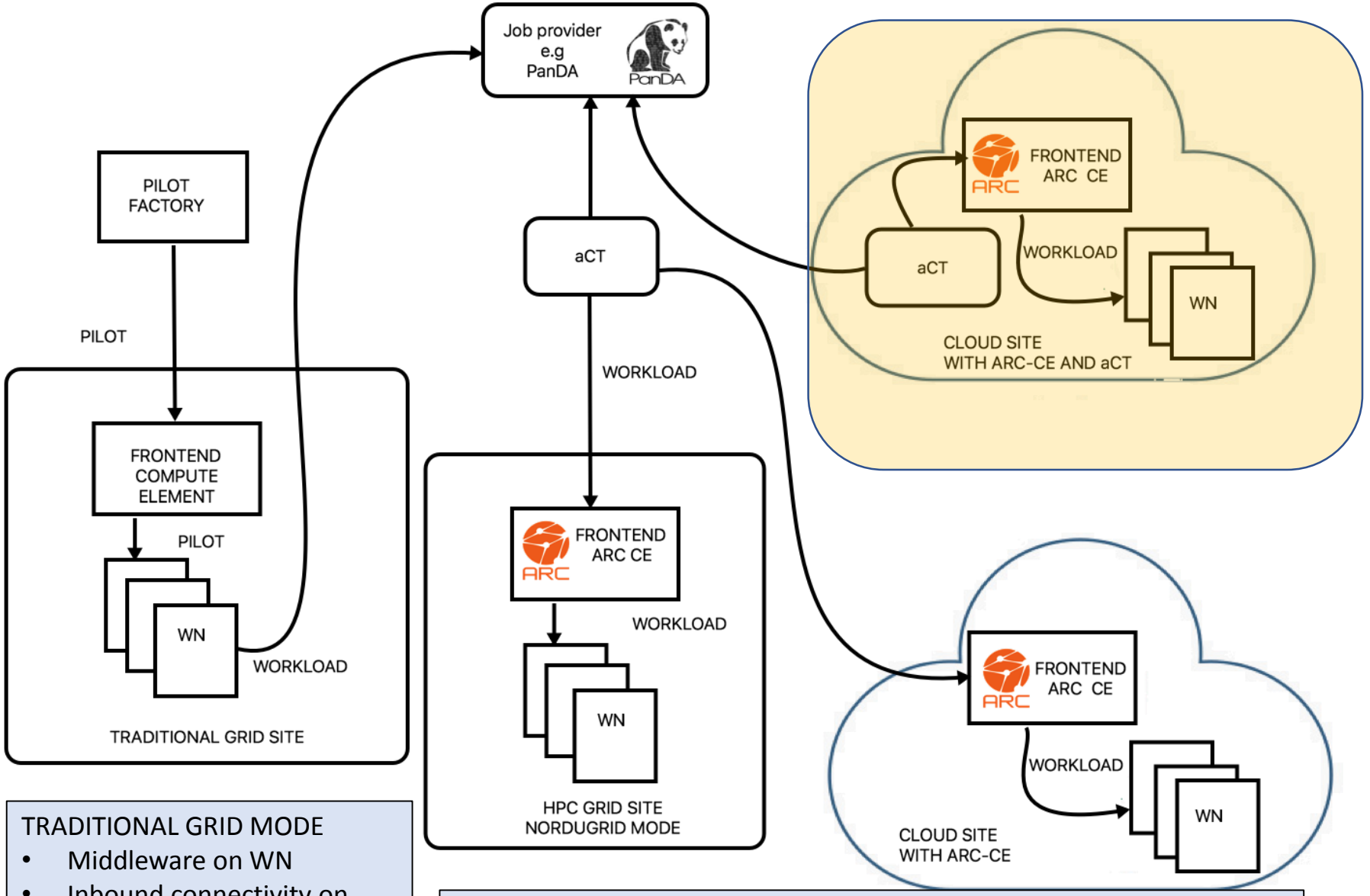
ARC: Advanced Resource Connector
aCT: ARC/ATLAS Control Tower

- A site might offer several grid flavours
  - Grid
  - HPC
  - Cloud

TRADITIONAL GRID MODE
- Middleware on WN
- Inbound connectivity on WN and frontend
- Information publishing service for discovery

NORDUGRID STANDARD MODE
- NO middleware on WN
- NO inbound connectivity on WN
- Inbound connectivity on frontend
- Information publishing service for discovery

Nordugrid ARC-CE and aCT INTERNAL MODE

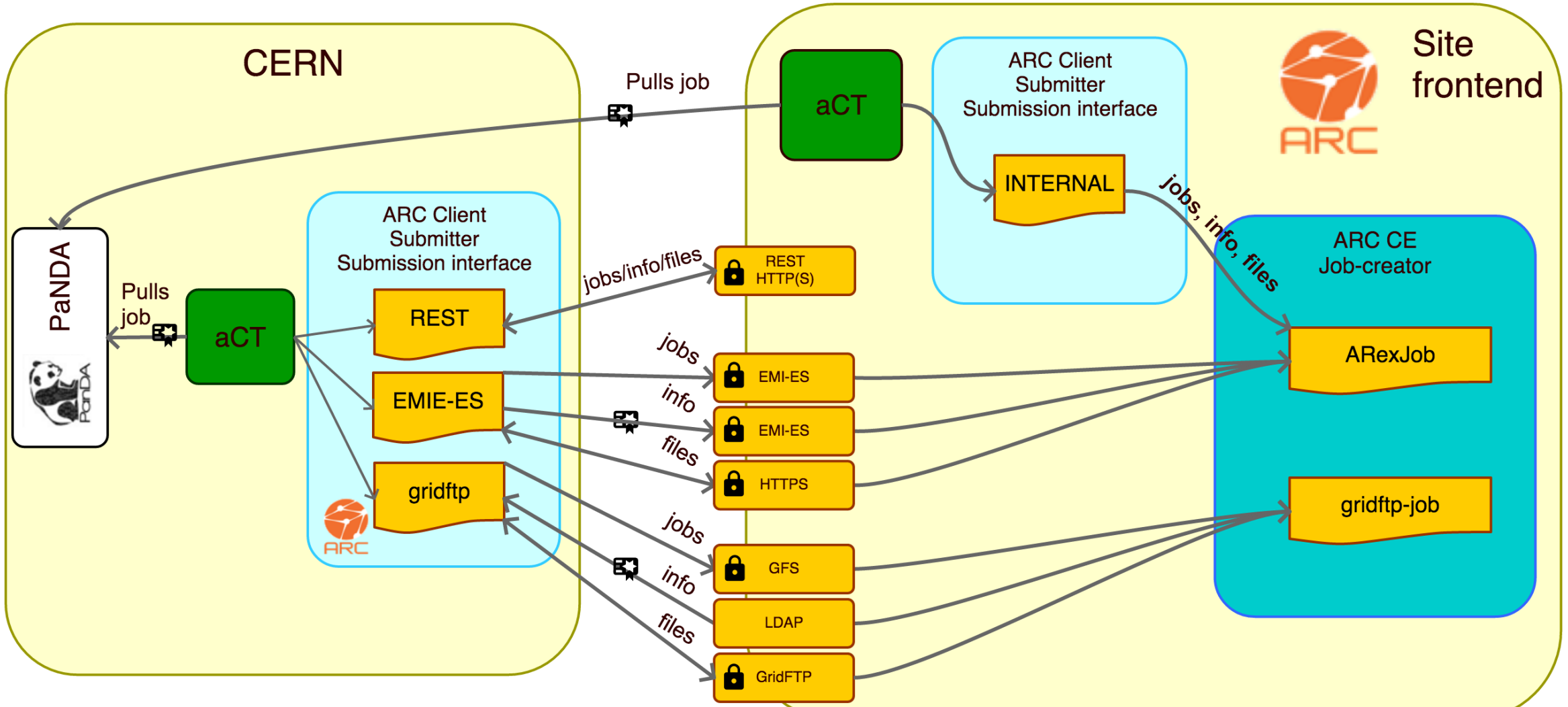


TRADITIONAL GRID MODE
- Middleware on WN
- Inbound connectivity on WN and frontend
- Information publishing service for discovery

NORDUGRID STANDARD MODE
- NO middleware on WN
- NO inbound connectivity on WN
- Inbound connectivity on frontend
- Information publishing service for discovery

NORDUGRID INTERNAL MODE
- NO middleware on WN
- NO inbound connectivity neither on WN nor frontend
- NO information publishing

Overview of the ARC-CE submission interfaces



INTERNAL submission interface

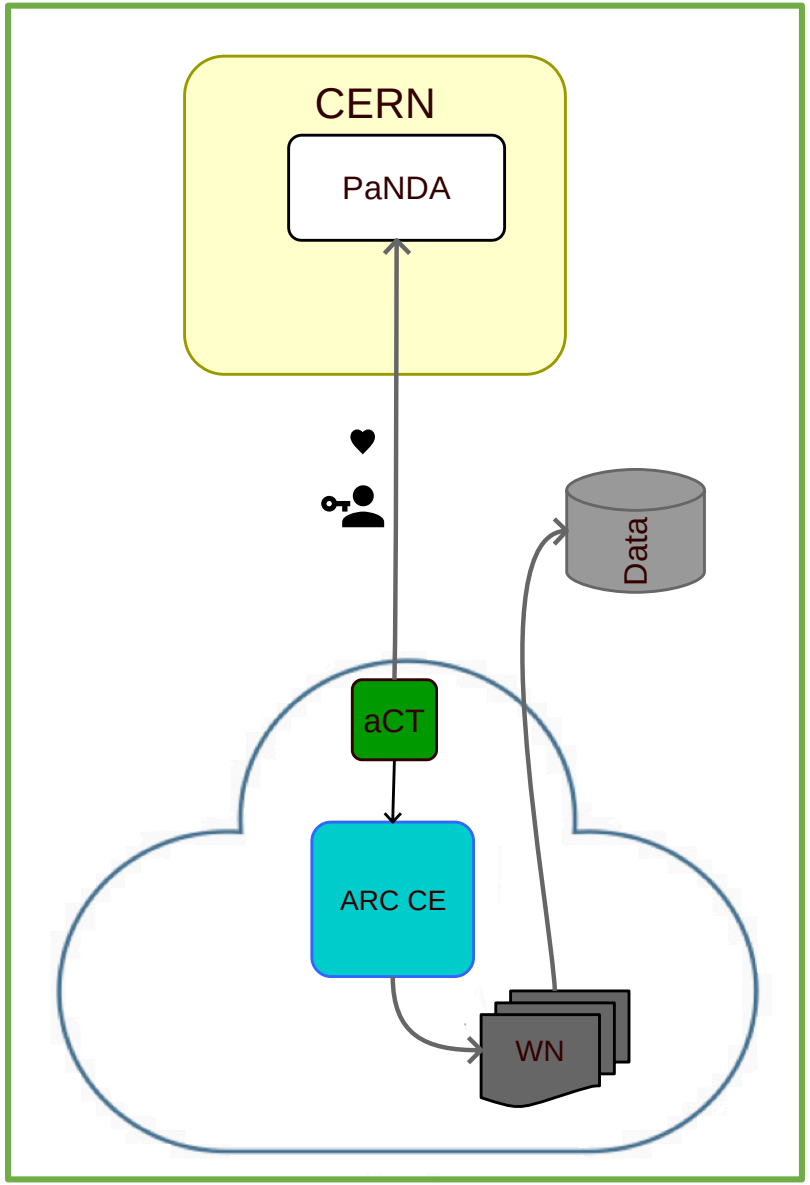
With aCT and ARC-CE installed at site running in “internal” mode: system administrator can run aCT and ARC-CE as non-root

- All files and jobs owned by this user

Since aCT and ARC are run on the same machine no host certificate is required

→ Minimal set of services, no gridftp server, no emi-es, no ldap, no host certificate

Lightweight ARC-CE beneficial for installation, configuration and maintenance



Setup and configuration of OpenStack grid site with Elasticluster

Elasticluster

http://elasticluster.readthedocs.io/en/latest/

Tool that uses ansible scripts to set up a cluster on a cloud service from inside or outside the cloud

- Elasticluster supported cloud providers
  - ec2_boto
  - Google
  - Openstack
  - Libcloud
- Batch system – slurm/gridengine/htcondor
- NFS setup
- HPC common software (… lmod, …), ganglia

Playbooks distributed with elasticluster

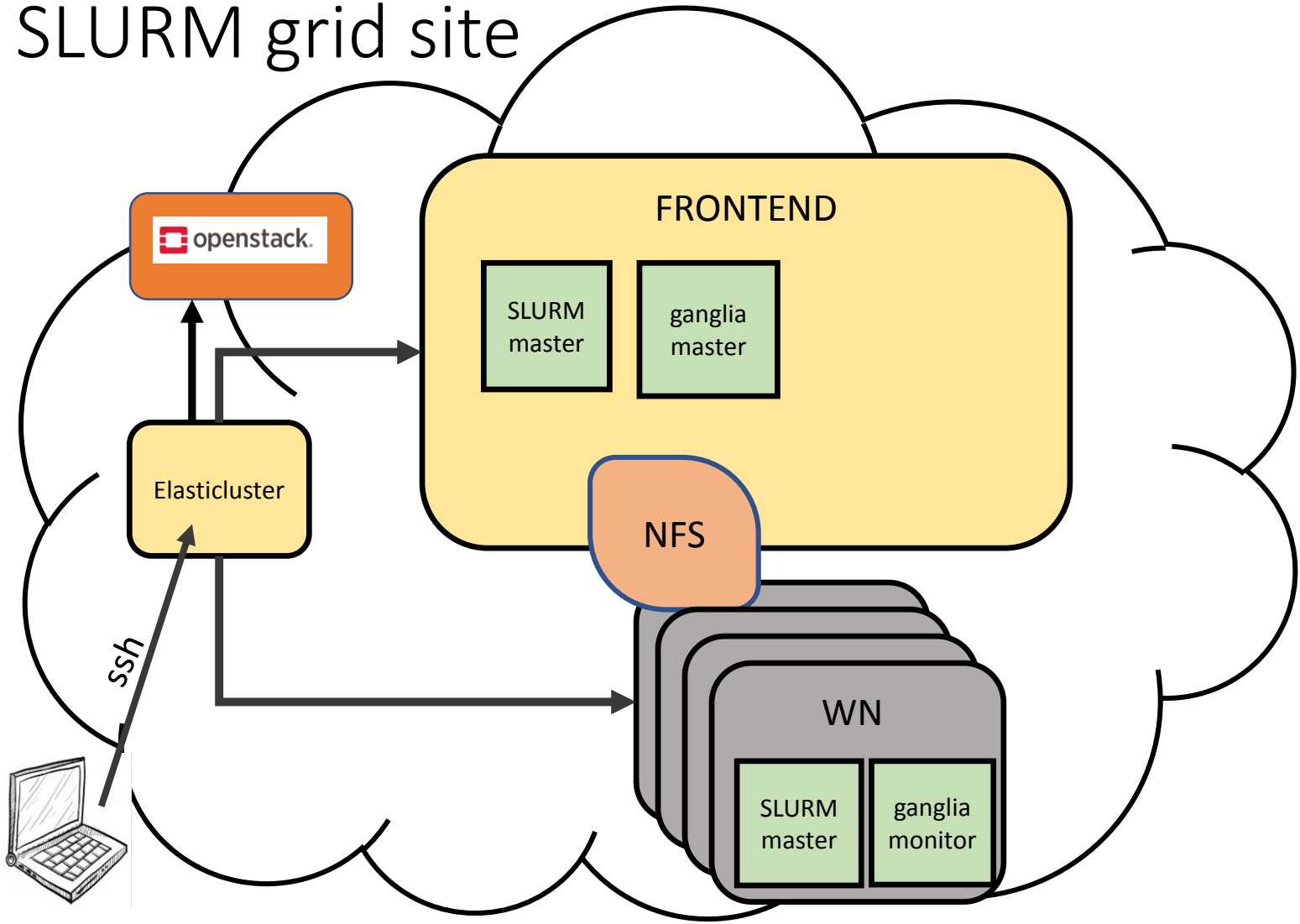
- Ansible
- SLURM
- GridEngine
- HTCondor
- Ganglia
- IPython cluster
- Hadoop + Spark
- CephFS
- GlusterFS
- OrangeFS/PVFS2
- Kubernetes

Available roles in Elasticluster:

anaconda	easybuild	glusterfs-server	hadoop.yml	htcondor.yml	jupyterhub	lua	pbs+maui	r.yml	spark-master
ansible	ganglia-gmetad	glusterfs.yml	hdfs-datanode	iptables	jupyterhub.yml	mcr	pbs+maui.yml	slurm-client	spark-worker
ansible.yml	ganglia-gmond	gridengine-common	hdfs-namenode	ipython	kubernetes-common	mcr.yml	pdsh	slurm-common	yarn-master
bigtop	ganglia-web	gridengine-exec	hive	ipython.yml	kubernetes-master	nfs-client	postgresql	slurm-master	yarn-worker
ceph	ganglia.yml	gridengine-master	hive-server	jenkins	kubernetes-worker	nfs-server	pvfs2	slurm-worker	
ceph.yml	glusterfs-client	gridengine.yml	hpc-common	jenkins.yml	kubernetes.yml	nis	pvfs2.yml	slurm.yml	
common	glusterfs-common	hadoop-common	htcondor	jupyter	lmod	ntpd	r	spark-common	

Elasticluster in work for SLURM grid site

- Elasticluster contacts the cloudprovider through the API
- Fires up specified number of frontends and compute nodes with specified OS, size, memory, and what ports to open (through predefined security group)
- Installs slurm server for frontend and client on compute nodes, NFS, ganglia (+ whatever else specified)
- Elasticluster ”after” play used to customize the frontend and compute elements



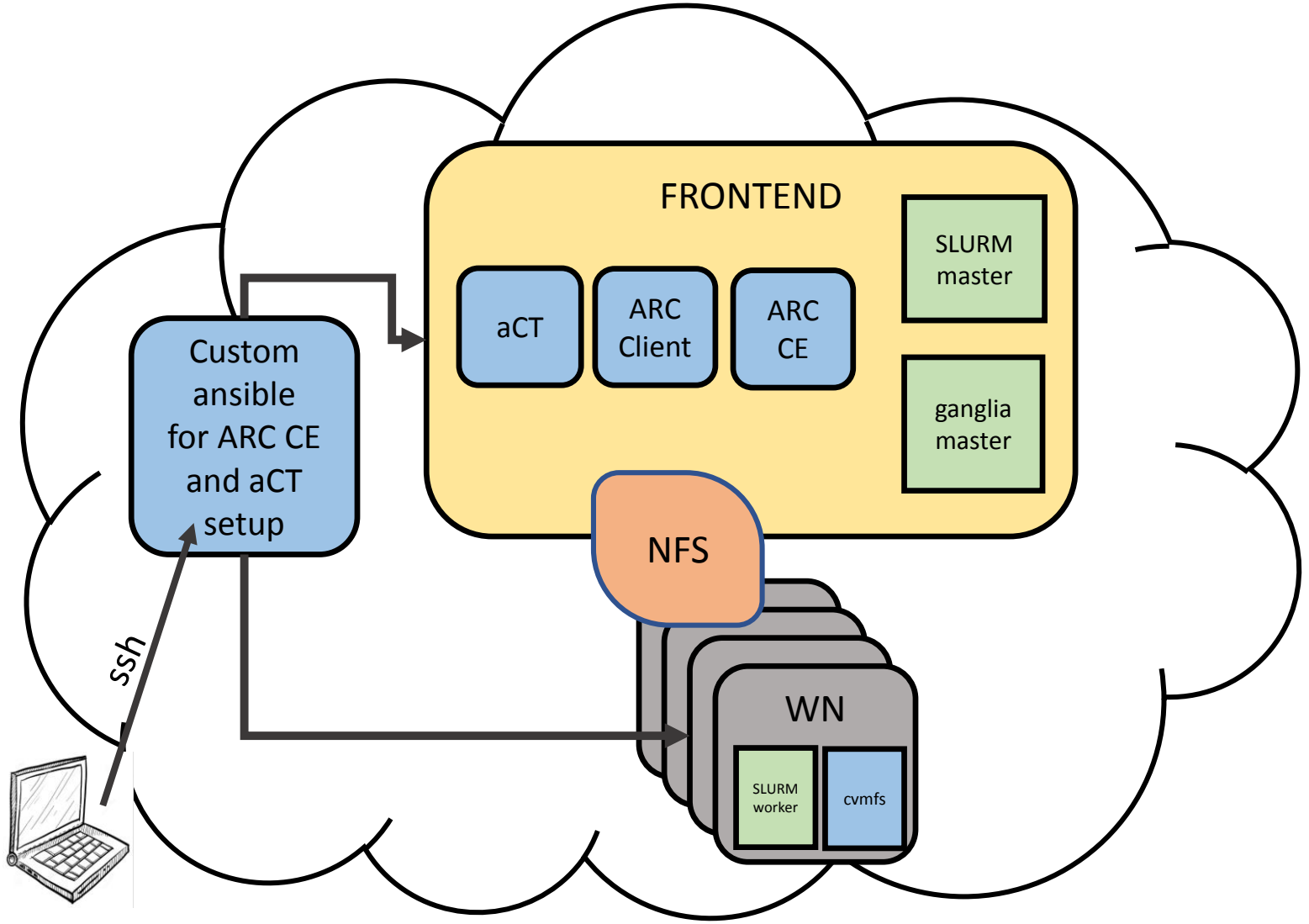
Steps to create an ARC-CE INTERNAL site
Ansible script tasks

On frontend

- Install, configure ARC, aCT
- Mounting of extra block storage for shared session directory, cache and runtime directory
- Install CA's for verification of incoming jobs
- Modify $PATH and $PYTHONPATH for non-default installation and as non-root
- Create griduser and add user to SLURM

On compute node

- Cvmfs setup plus extra block storage to contain it
- Create griduser and add user to SLURM



Elasticluster and ansible sequence

```
step1)
elasticluster -v start slurm -n $clustername

step2)
elasticluster -v setup $clustername -- elasticluster/src/elasticluster/share/playbooks/after_custom.yml \
--tags "after" \
--extra-vars="localuser=centos lrms_type=slurm cluster_name=$clustername" \
--extra-vars="@$play_vars/blockstorage.yml" \
--extra-vars="@$play_vars/griduser_local.yml" \
--extra-vars="@$play_vars/os_env.yml" \
--extra-vars="@$play_vars/nfs_export_mounts_local.yml"

step3)
ansible-playbook grid-uh-cloud/ansible/site_arc-ce_act.yml \
-i ~/.elasticluster/storage/$clustername.inventory \
--skip-tags="installarc,private-act,cvmfs,apache" \
--extra-vars="localuser=centos installationtype=local arc_major=6 lrms_type=slurm"  \
--extra-vars="@$play_vars/griduser_local.yml" \
--extra-vars="@$play_vars/os_env.yml"  \
--extra-vars="@$play_vars/host_env.yml"  \
--extra-vars="@$play_vars/slurm_pwd.yml"
```

Link to playbook to install ARC and aCT (step 3)
https://source.coderefinery.org/nordugrid/contrib/tree/master/ansible/arc-ce

Testing submission with the INTERNAL submission mode

Specify local interface `–S org.nordugrid.internal` or leave blank as it is the default

```
[centos@frontend001 testing]$ arcsub -c localhost -S org.nordugrid.internal hello.xrls
[centos@frontend001 testing]$ arcsub -c localhost hello.xrls
Job submitted with jobid: file:///wlcg/session/5alNDm1r9vsnrpO2tmaBI5UnABFKDmABFKDmB2KKDmABFKDmXr0rKm

[centos@frontend001 testing]$ arcstat --long --all
Job: file:///wlcg/session/5alNDm1r9vsnrpO2tmaBI5UnABFKDmABFKDmB2KKDmABFKDmXr0rKm
 Name: hello_ARCTEST1
 State: Queuing
 Specific state: INLRMS
 ID on service: 5alNDm1r9vsnrpO2tmaBI5UnABFKDmABFKDmB2KKDmABFKDmXr0rKm
 Service information URL: file://localhost (org.nordugrid.internal)
 Job status URL: file://localhost (org.nordugrid.internal)
```

UIO_CLOUD queue Hammercloud jobs with local submission in PanDA monitor

- An ARC-CE and aCT INTERNAL test cluster has successfully been installed in the University of Oslo's Openstack cloud service
- Collects jobs from PanDA as the UIO_CLOUD queue
- The jobs are so-called Hammercloud jobs
  - Testing framework using realistic ATLAS jobs
  - Jobs require cvmfs, download of input files etc.

3977944112 Attempt 0	gangarbt		Sim_tf.py	finished	2018-06-28 15:35:55	0:0:05:16	0:0:17:02	2018-06-28 16:05:06	ND UIO_CLOUD test testing	10000	2 (1)
	Job name: 9fe61744-6963-48e0-93a2-1e67fbc53743_779 #0										
	Datasets: **In:** mc15_13TeV.361106.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zee.evgen.EVNT.e3601_tid04972714_00 || Rucio link **Out:** hc_test.gangarbt.hc20116895.tid957.UIO_CLOUD.110										
3977936140 Attempt 0	gangarbt		Sim_tf.py	finished	2018-06-28 15:20:59	0:0:10:12	0:0:17:02	2018-06-28 15:52:44	ND UIO_CLOUD test testing	10000	2 (1)
	Job name: fc1d1bd4-db17-480a-96eb-8254c5ee94ae_41732 #0										
	Datasets: **In:** mc15_13TeV.361106.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zee.evgen.EVNT.e3601_tid04972714_00 || Rucio link **Out:** hc_test.gangarbt.hc20116895.tid957.UIO_CLOUD.110										
3977929803 Attempt 0	gangarbt		Sim_tf.py	finished	2018-06-28 15:06:04	0:0:05:02	0:0:17:04	2018-06-28 15:41:53	ND UIO_CLOUD test testing	10000	2 (1)
	Job name: 9fbe2cfb-3b01-4078-84b5-e591414c728b_5435 #0										
	Datasets: **In:** mc15_13TeV.361106.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zee.evgen.EVNT.e3601_tid04972714_00 || Rucio link **Out:** hc_test.gangarbt.hc20116895.tid957.UIO_CLOUD.110										
	gangarbt		Sim_tf.py	finished	2018-06-28 14:53:28	0:0:02:14	0:0:17:05	2018-06-28 15:17:04	ND UIO_CLOUD test testing	10000	2 (1)

Conclusion

- ARC and aCT gives a new site configuration option for ATLAS sites
  - Lightweight
  - Good option for restrictive sites
  - Suitable for cloud and HPC
- Will be available in upcoming release of ARC 6
  - Pre-release version already available
  - https://source.coderefinery.org/nordugrid/arc

Extra material

Minimalistic configuration of ARC for INTERNAL submission only running ARC as normal user

```
[lrms]
lrms=slurm

[arex]
logfile=/grid/arex.log
joblog=/grid/gm-jobs.log
controldir=/grid/control
sessiondir=/wlcg/session
runtimedir=/wlcg/runtime
shared_scratch=/wlcg

[arex/cache]
logfile=/grid/cache-clean.log
cachedir=/wlcg/cache
cachesize=80 70
cachelifetime=1d

[infosys]
logfile=/grid/infoprovider.log

[queue:main]
```

For production site you would add VO configuration

Example configuration of elasticluster

Openstack auth

```
[cloud/iaas]
provider=openstack
auth_url=https://api.uh-iaas.no:5000/v3
username=maiken.pedersen@usit.uio.no
password=xxxxxx
project_name=uio-test-hpc-grid
user_domain_name=dataporten
project_domain_name=dataporten
region_name=osl
identity_api_version=3
```

Cluster login

```
[login/centos]
image_user=centos
image_user_sudo=root
image_sudo=True
user_key_name=cloud
user_key_private=~/.ssh/cloud.key
user_key_public=~/.ssh/cloud.key.pub
```

Ansible groups

```
[setup/ansible-slurm]
provider=ansible
frontend_groups=slurm_master,ganglia_master,ganglia_monitor,frontend,cluster
compute_groups=slurm_worker,ganglia_monitor,compute,cluster
global_var_multiuser_cluster=no
```

Cluster setup

```
[cluster/slurm]
cloud=iaas
login=centos
setup=ansible-slurm
security_group=default
image_id=df3dedc6-f98c-4eb0-b77e-7f8f24f857e4
frontend_nodes=1
compute_nodes=1
ssh_to=frontend
network_ids=c97fa886-592e-4ad1-a995-6d55651bed78
```

Instance flavours

```
[cluster/slurm/frontend]
flavor=m1.medium

[cluster/slurm/compute]
flavor=m2.4xlarge
```

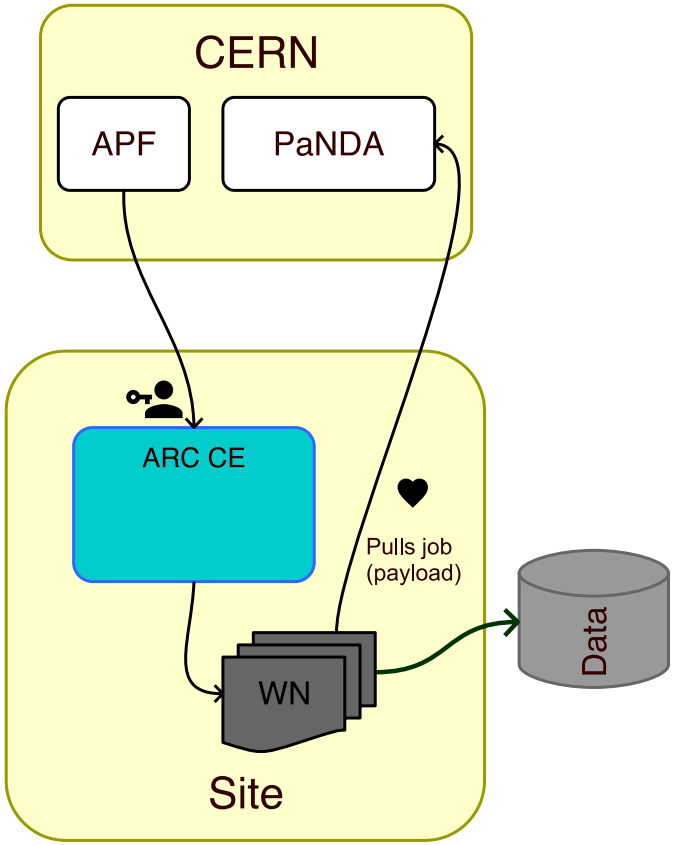
Configuration of aCT for INTERNAL mode

```
<config>

<db>
  <type>mysql</type>
   <name>act</name>
  <user>centos</user>
  <password>secret</password>
  <host>localhost</host>
  <port>3306</port>
</db>

<loop>
  <periodicrestart>
    <actsubmitter>120</actsubmitter>
    <actstatus>600</actstatus>
    <actfetcher>600</actfetcher>
    <actcleaner>600</actcleaner>
  </periodicrestart>
</loop>

<tmp>
  <dir>/tmp</dir>
</tmp>

<actlocation>
  <dir>/grid/software/aCT/src/</dir>
  <pidfile>/grid/act.pid</pidfile>
</actlocation>
```

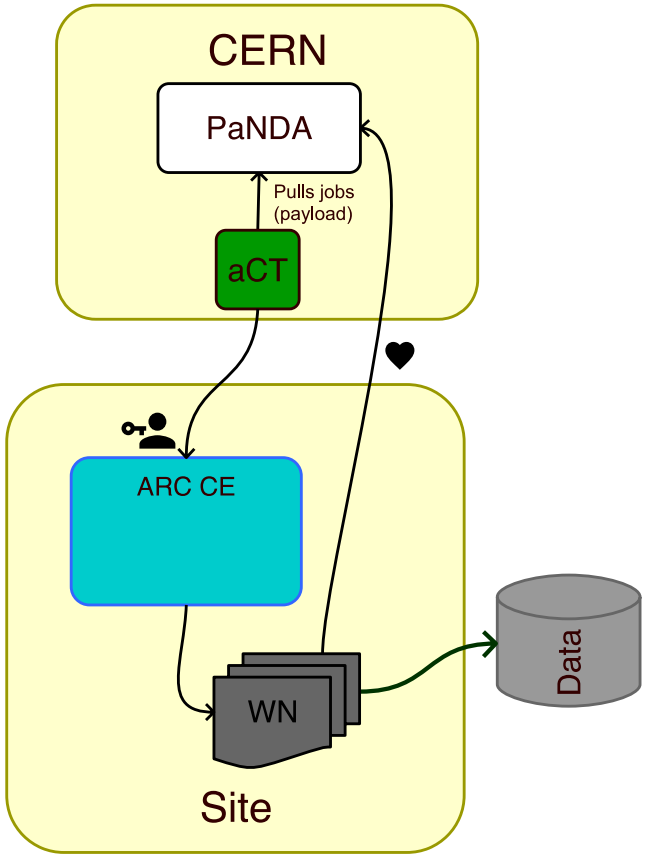
```
<logger>
  <level>debug</level>
  <arclevel>debug</arclevel>
  <logdir>/grid</logdir>
  <rotate>25</rotate>
</logger>

<atlasgiis>
<timeout>20</timeout>
</atlasgiis>

<queuesreject>
  <item>bigmem</item>
  <item>tier3</item>
  <item>infiniband</item>
  <item>gridsim</item>
</queuesreject>
```

```
<jobs>
  <checkinterval>30</checkinterval>
  <checkmintime>20</checkmintime>
  <maxtimerunning>259200</maxtimerunning>
  <maxtimehold>172800</maxtimehold>
  <maxtimeundefined>3600</maxtimeundefined>
</jobs>

<voms>
  <vo>atlas</vo>
  <roles>
    <item>production</item>
  </roles>
  <bindir>/grid/software/bin</bindir>
  <proxylifetime>345600</proxylifetime>
  <minlifetime>259200</minlifetime>
  <proxypath>/grid/atlact1.rfc.long.proxy</proxypath>
  <cacertdir>/etc/grid-security/certificates</cacertdir>
  <proxystoredir>/grid/proxies</proxystoredir>
</voms>
```

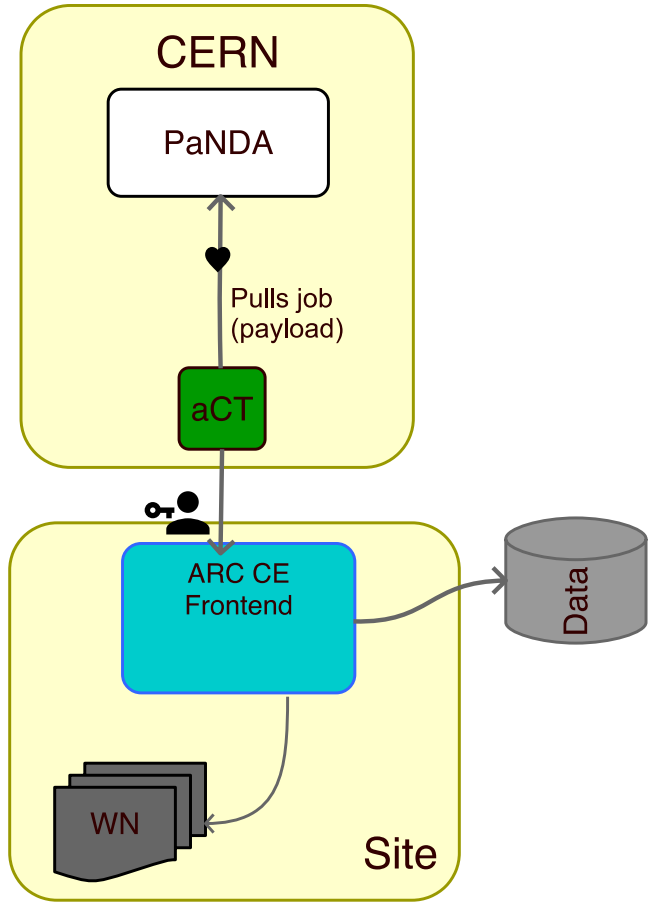
Nordugrid ARC CE modes



Pilot factory

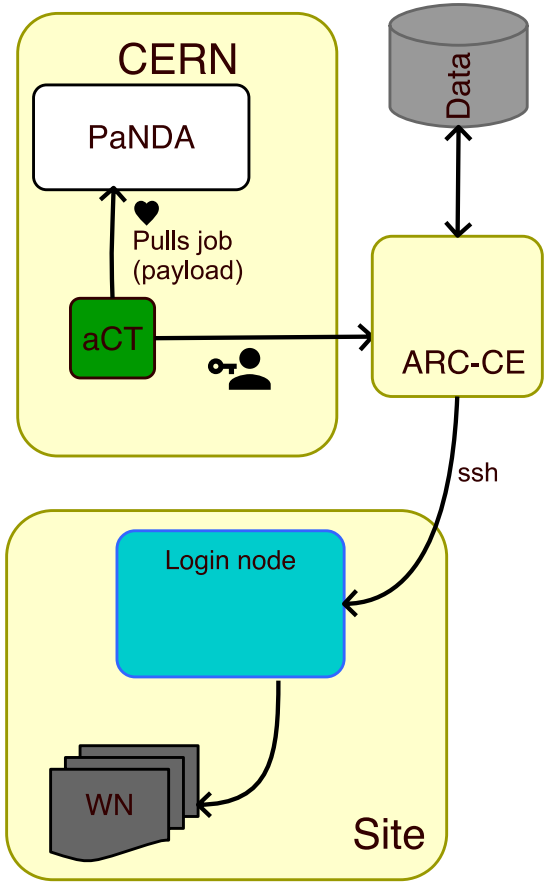


True pilot

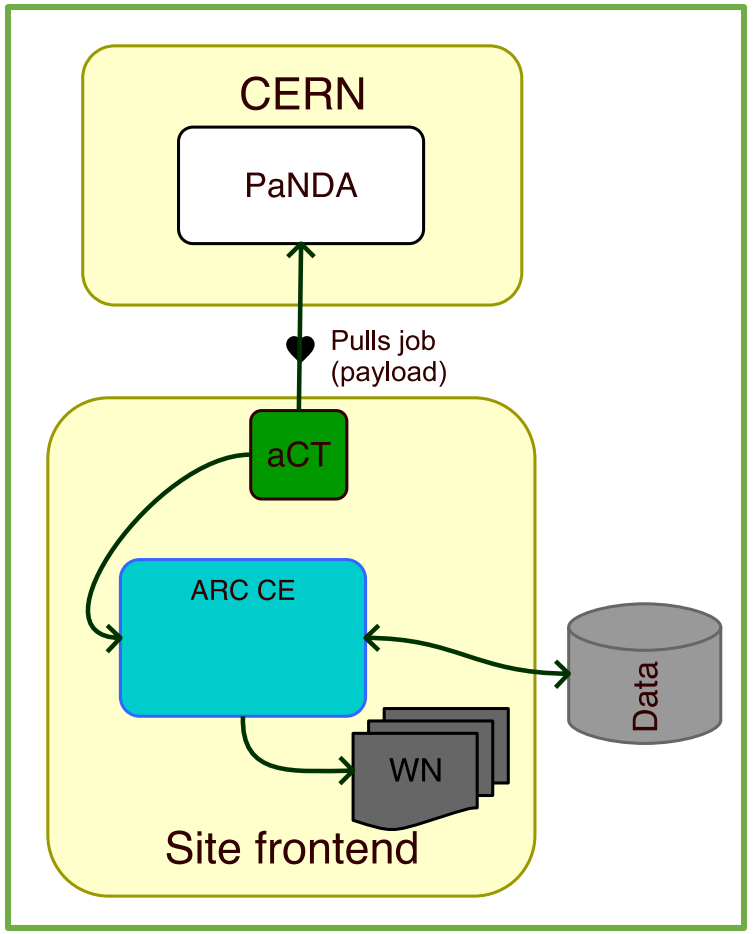


NDGF mode

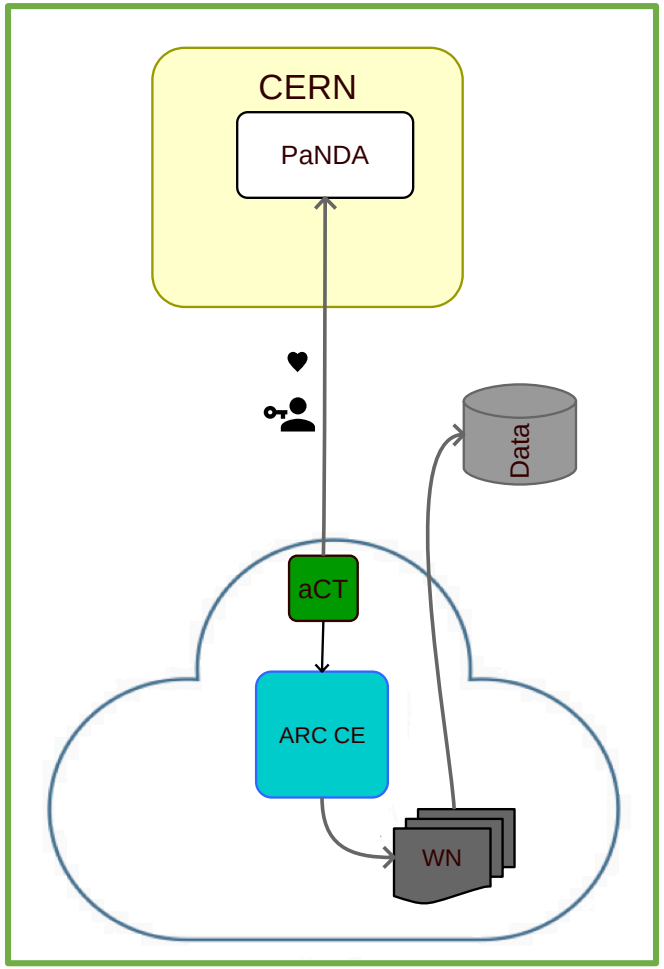
Nordugrid ARC CE modes
for restrictive (HPC) sites and lightweight sites, including clouds



ssh-mode



INTERNAL mode HPC



INTERNAL mode cloud