

# Lightweight on-demand computing with Elasticcluster 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)



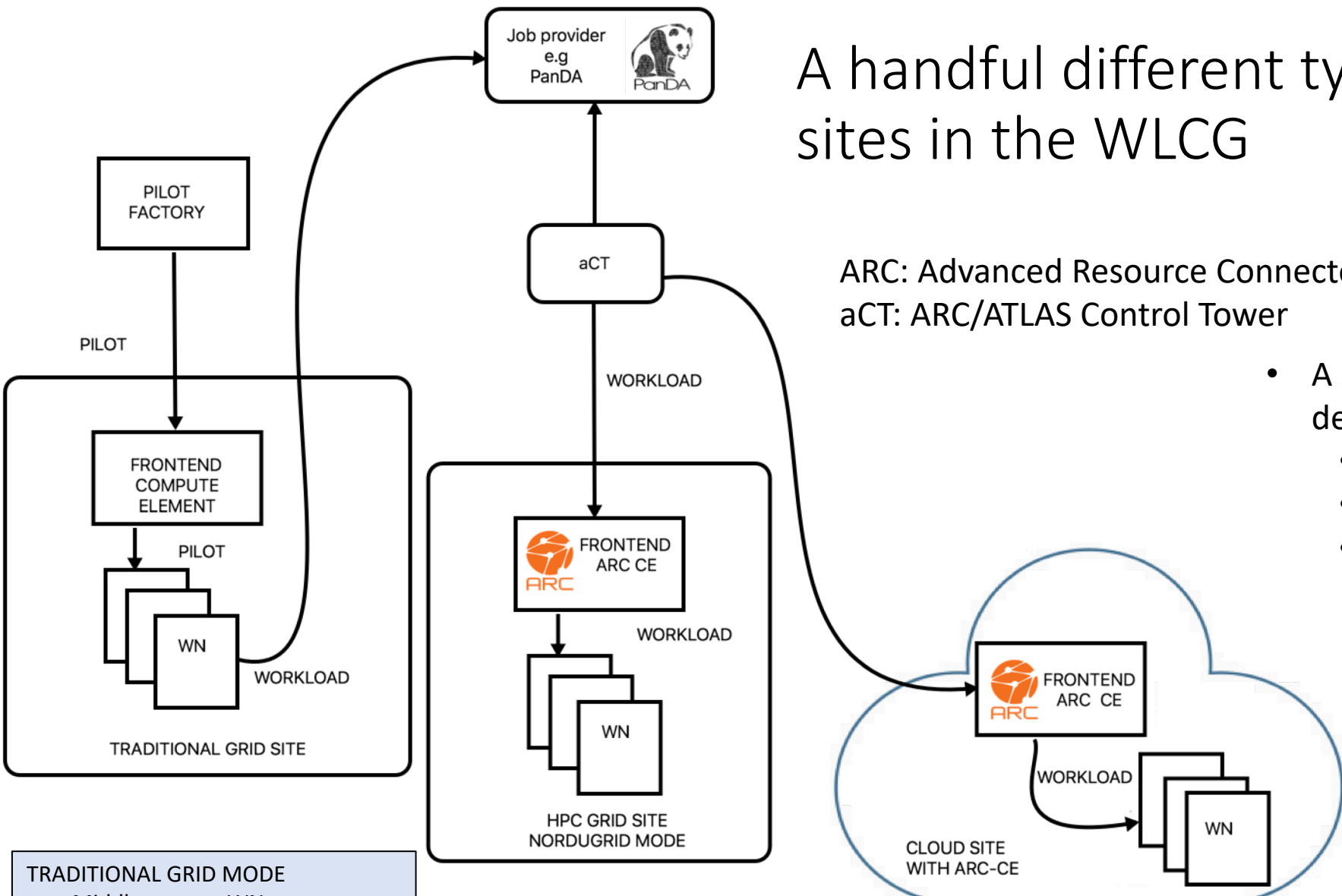
# Overview

- Types of ATLAS sites in WLCG
- Nordugrid ARC and aCT in INTERNAL mode on cloud resource
- Overview of the ARC-CE submission interfaces
- Setup and configuration of OpenStack grid site with Elasticcluster
- 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 depending on availability
  - Grid
  - HPC
  - Cloud



**TRADITIONAL GRID MODE**

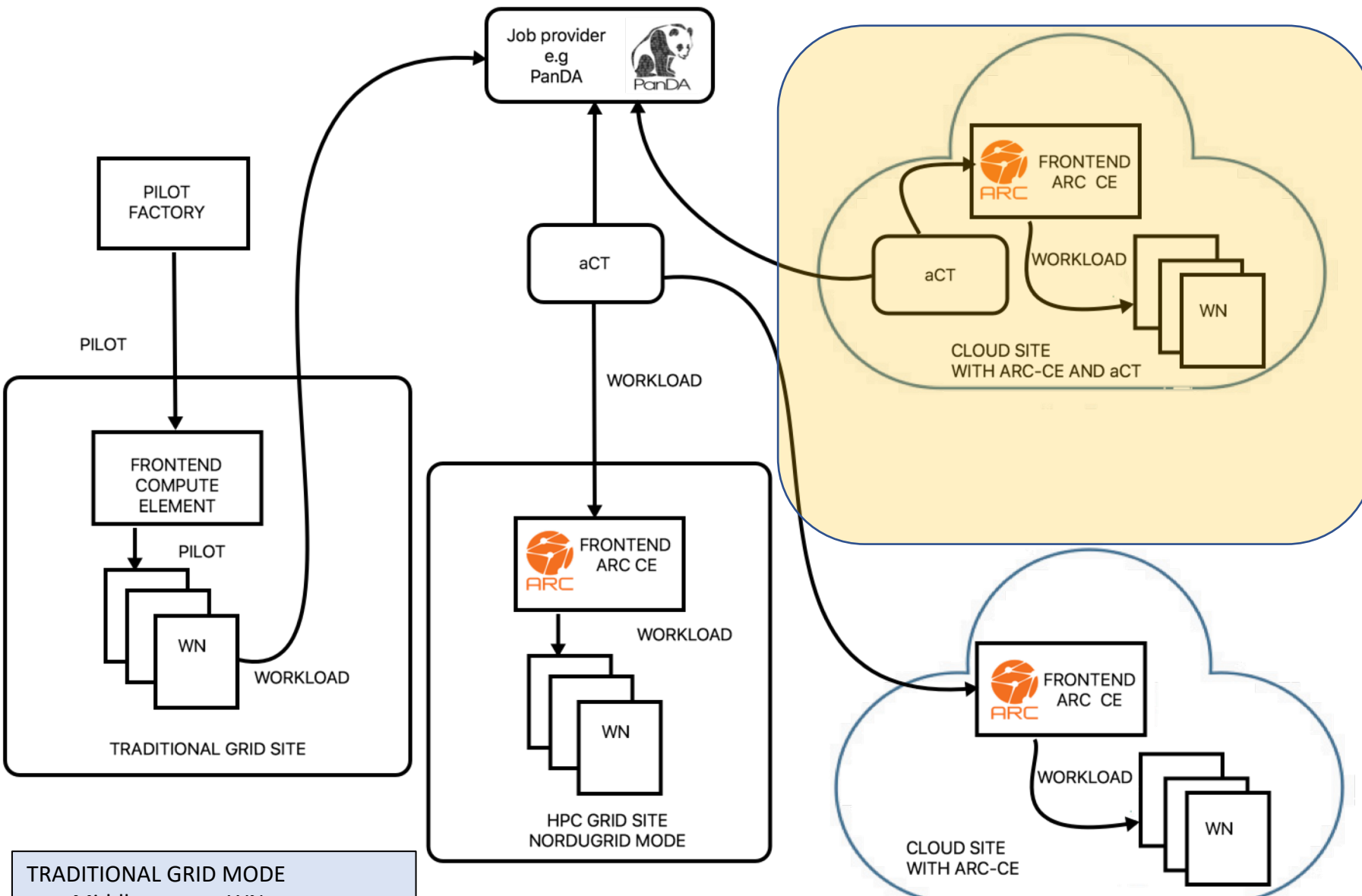
- Middleware on WN
- Inbound connectivity on WN and frontend
- Information publishing service for discovery

**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

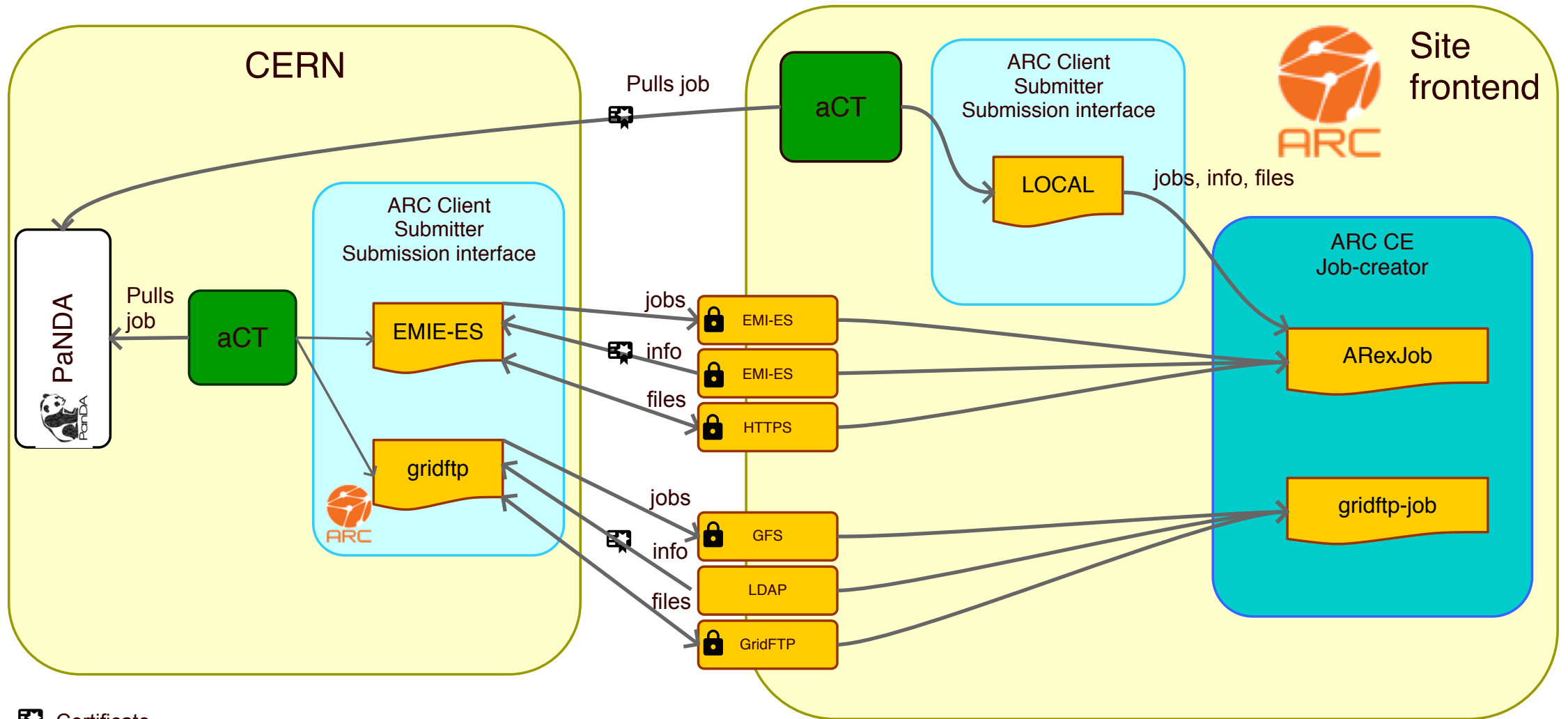
- NORDUGRID INTERNAL MODE**
- NO middleware on WN
  - NO inbound connectivity
  - NO information publishing





- 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

# Overview of the ARC-CE submission interfaces



-  Certificate
-  Credentials

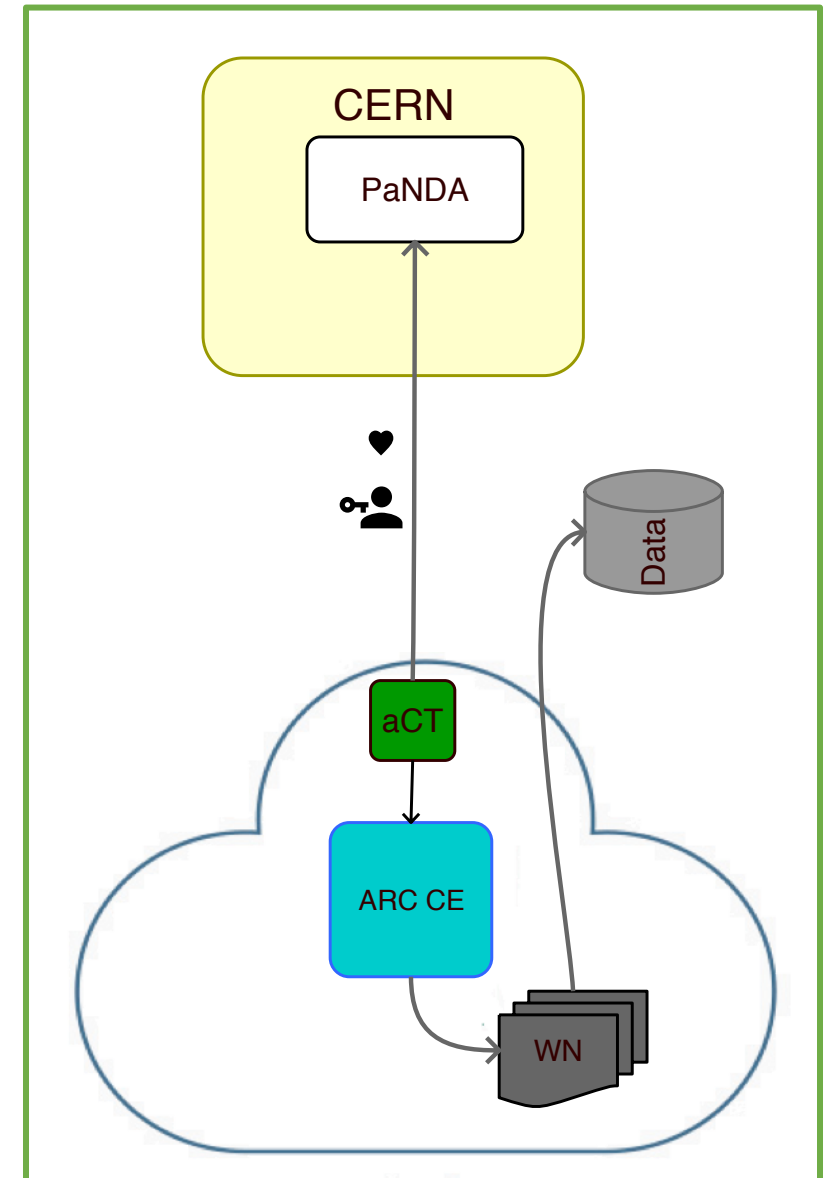
# INTERNAL submission interface

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

→ 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 Elasticcluster

# Elasticcluster

<http://elasticcluster.readthedocs.io/en/latest/>

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

- Elasticcluster supported cloud providers
  - ec2\_boto
  - Google
  - Openstack
  - Libcloud
- Batch system – slurm/gridengine/htcondor
- NFS setup
- HPC common software (... lmod, ...), ganglia

## Playbooks distributed with elasticcluster

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

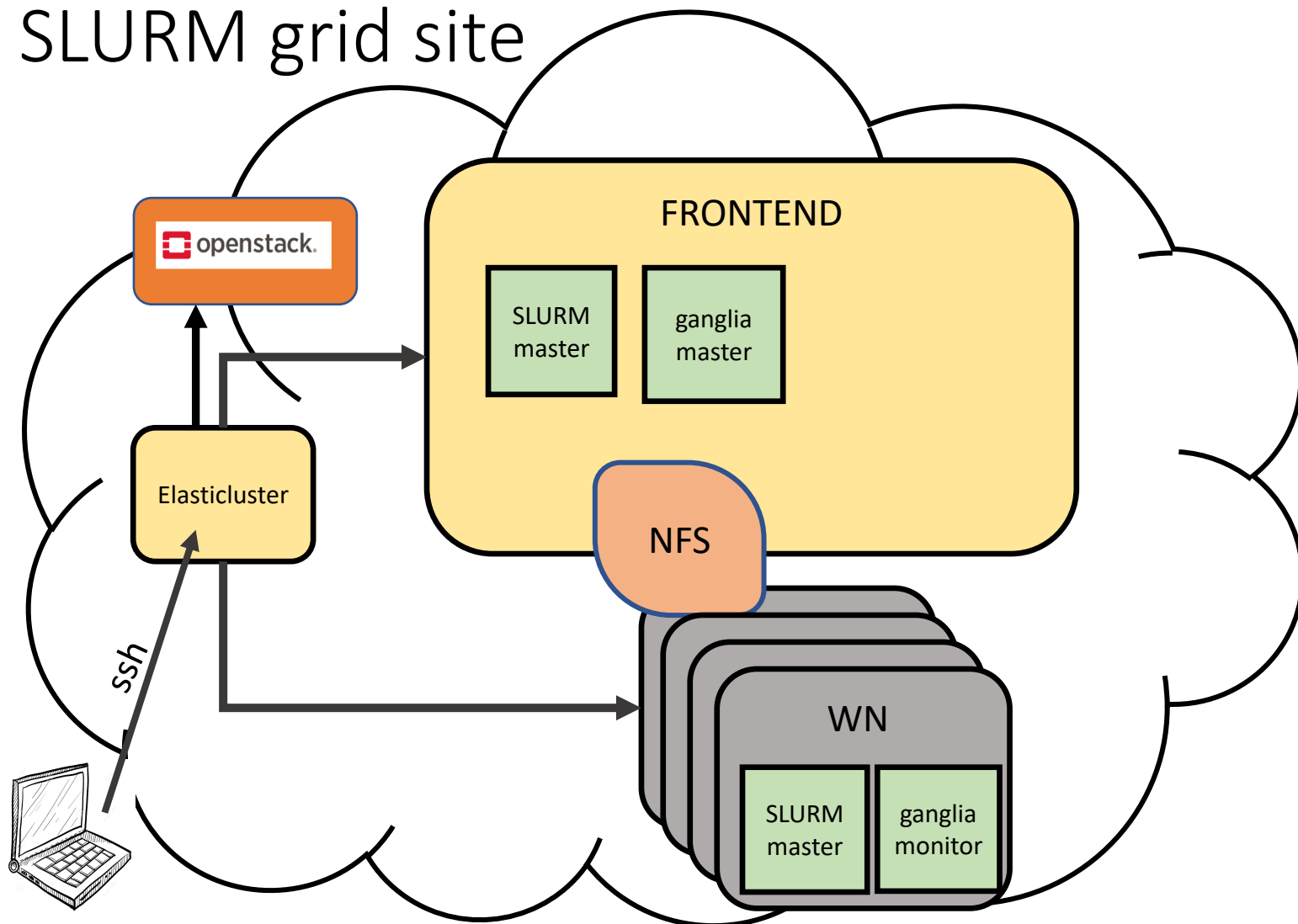
## Available roles in Elasticcluster:

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

- Elasticcluster 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
- Custom Elasticcluster "after" play can be configured to attach extra storage volumes needed for the ARC-CE, and distribute shared folders



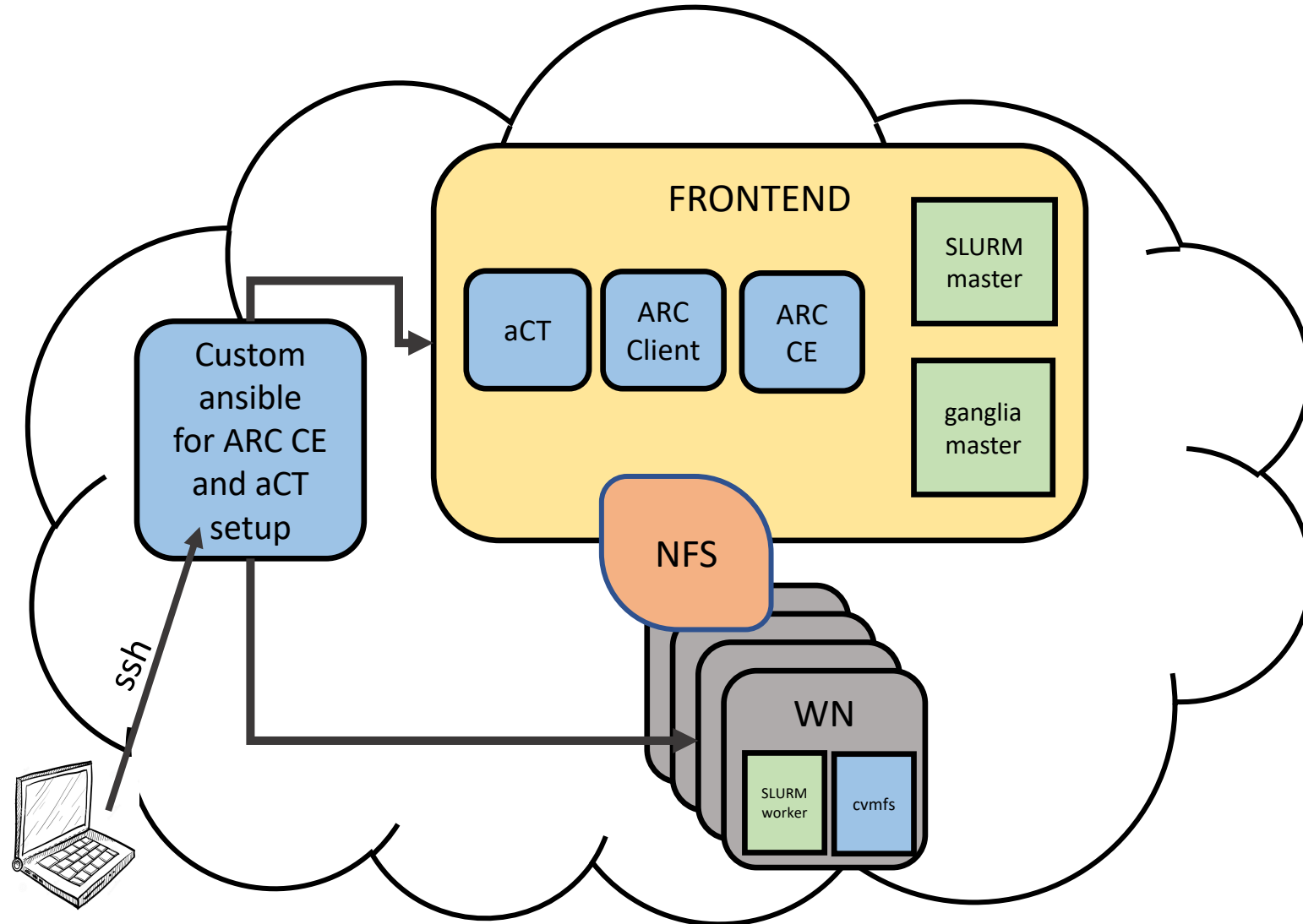
# Creating an ARC-CE with aCT and preparing compute nodes

## 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



# Elasticcluster and ansible sequence

step1)

```
elasticcluster -v start slurm -n $clustername
```

step2)

```
elasticcluster -v setup $clustername -- elasticcluster/src/elasticcluster/share/playbooks/after_custom.yml \  
--tags "after" \  
--extra-vars="localuser=centos lrms_type=slurm cluster_name=$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 ~/.elasticcluster/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"
```

# Testing submission with the INTERNAL submission interface

On compute-element use INTERNAL job submission

```
arcsub --direct -c localhost -S org.nordugrid.internal hello.xr1s
```

```
[[centos@frontend001 arctestng]$ arcstat -c localhost --long
```

```
Job: local://localhost/q5ZNDmJ4CdrnzfEJwm4kCpGoABFKDmABFKDm6SIKDmABFKDmmXOthn  
Name: hello_LOCAL-CLOUD-ARC  
State: Finishing  
Specific state: FINISHING  
ID on service: q5ZNDmJ4CdrnzfEJwm4kCpGoABFKDmABFKDm6SIKDmABFKDmmXOthn  
Service information URL: local://localhost (org.nordugrid.local)  
Job status URL: local://localhost (org.nordugrid.local)  
Job management URL: local://localhost (org.nordugrid.local)
```

```
Status of 1 jobs was queried, 1 jobs returned information
```

# UIO\_CLOUD

## Hammercloud jobs with local submission in PanDA monitor

- An ARC-CE and aCT INTERNAL test cluster has successfully been installed
- 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.

3754909557 Attempt 0	gangarbt	Sim_tf.py	finished	2017-12-17 13:17:08	0:0:02:55	0:0:14:07	2017-12-17 13:43:10	ND UIO_CLOUD
	Job name: e6b3d63a-7719-4f8b-a525-0f0ed6c1bda6_28051 #0							
	Datasets: In: mc15_13TeV.361106.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zee.evgen.EVNT.e3601_tid04972714_00 Out: hc_test.gangarbt.hc20112674.tid916.UIO_CLOUD.141							
3754909568 Attempt 0	gangarbt	Sim_tf.py	finished	2017-12-17 13:17:09	0:0:02:58	0:0:14:03	2017-12-17 13:41:10	ND UIO_CLOUD
	Job name: 4dde54bd-362a-4626-af48-8f93e26a271d_36051 #0							
	Datasets: In: mc15_13TeV.361106.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zee.evgen.EVNT.e3601_tid04972714_00 Out: hc_test.gangarbt.hc20112674.tid916.UIO_CLOUD.141							
3754909822 Attempt 0	gangarbt	Sim_tf.py	finished	2017-12-17 13:17:31	0:0:02:35	0:0:10:55	2017-12-17 13:39:41	ND UIO_CLOUD
	Job name: e49a5426-e8c1-4b4d-abad-e1e96510ca32_38501 #0							
	Datasets: In: mc15_13TeV.361106.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zee.evgen.EVNT.e3601_tid04972714_00 Out: hc_test.gangarbt.hc20112683.tid957.UIO_CLOUD.239							
3754909500 Attempt 0	gangarbt	Sim_tf.py	finished	2017-12-17 13:17:07	0:0:02:47	0:0:12:56	2017-12-17 13:37:58	ND UIO_CLOUD
	Job name: 5f59d569-04c4-43b5-8e65-7147de5fa9cf_32178 #0							
	Datasets: In: mc15_13TeV.361106.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zee.evgen.EVNT.e3601_tid04972714_00 Out: hc_test.gangarbt.hc20112675.tid839.UIO_CLOUD.172							
3754906720 Attempt 0	gangarbt	Sim_tf.py	finished	2017-12-17 13:12:30	0:0:02:10	0:0:11:59	2017-12-17 13:34:21	ND UIO_CLOUD
	Job name: 807454fd-bc31-4ce4-b3d8-b9615e7e606c_32299 #0							
	Datasets: In: mc15_13TeV.361106.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zee.evgen.EVNT.e3601_tid04972714_00 Out: hc_test.gangarbt.hc20112683.tid957.UIO_CLOUD.239							
3754894693 Attempt 0	gangarbt	Sim_tf.py	starting	2017-12-17 12:51:43	0:0:02:48	0:0:55:53	2017-12-17 13:24:34	ND UIO_CLOUD
	Job name: 12ad44bb-681f-4a27-9370-471e8530026c_32730 #0							
	Datasets: In: mc15_13TeV.361106.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zee.evgen.EVNT.e3601_tid04972714_00 Out: hc_test.gangarbt.hc20112671.tid841.UIO_CLOUD.89							
3754894695 Attempt 0	gangarbt	Sim_tf.py	finished	2017-12-17 12:51:43	0:0:07:45	0:0:14:22	2017-12-17 13:24:31	ND UIO_CLOUD
	Job name: a4436ab7-4ea2-4e08-aba3-a929a9f11cb2_9002 #0							
	Datasets: In: mc15_13TeV.361106.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zee.evgen.EVNT.e3601_tid04972714_00 Out: hc_test.gangarbt.hc20112671.tid841.UIO_CLOUD.89							
	gangarbt	Sim_tf.py	finished	2017-12-17 12:59:33	0:0:04:54	0:0:11:12	2017-12-17 13:21:09	ND UIO_CLOUD

# 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 at <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 elasticcluster

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=ui0-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>

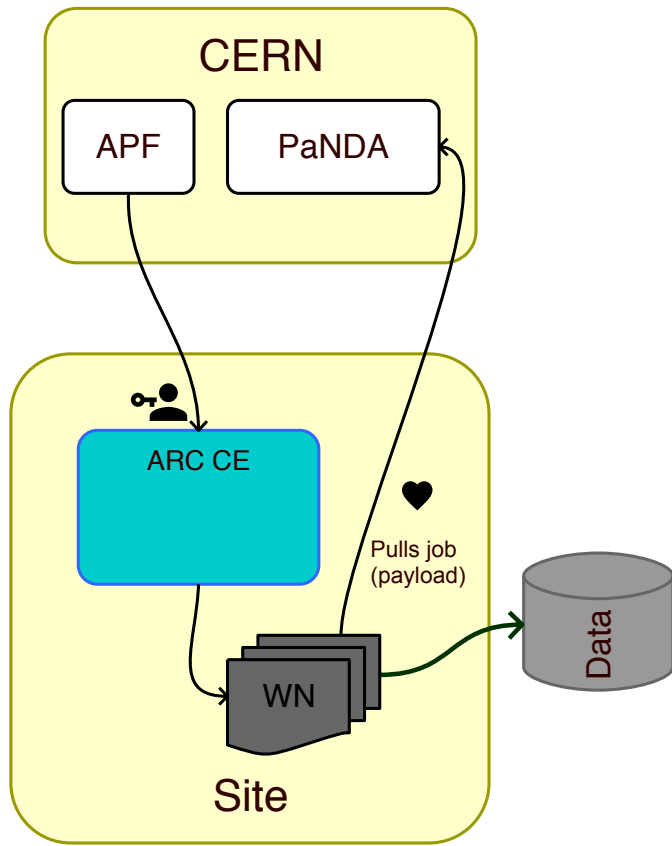
<atlasgiis>
  <timeout>20</timeout>
</atlasgiis>

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

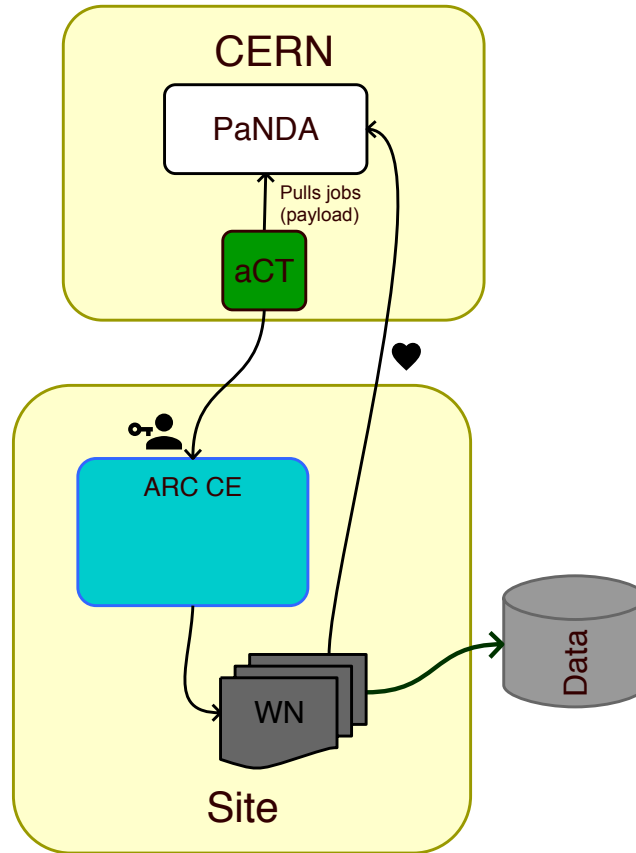
<jobs>
  <checkinterval>30</checkinterval>
  <checkmintime>20</checkmintime>
  <maxtimerunning>259200</maxtimerunning>
  <maxtimehold>172800</maxtimehold>
  <maxtimeundefined>3600</maxtimeundefined>
</jobs>

<voms>
  <vo>atlas</vo>
  <roles>
    <item>production</item>
  </roles>
  <bindir>/grid/software/bin</bindir>
  <proxylifetime>345600</proxylifetime>
  <minlifetime>259200</minlifetime>
  <proxypath>/grid/atlas1.rfc.long.proxy</proxypath>
  <cacertdir>/etc/grid-security/certificates</cacertdir>
  <proxystoredir>/grid/proxies</proxystoredir>
</voms>
```

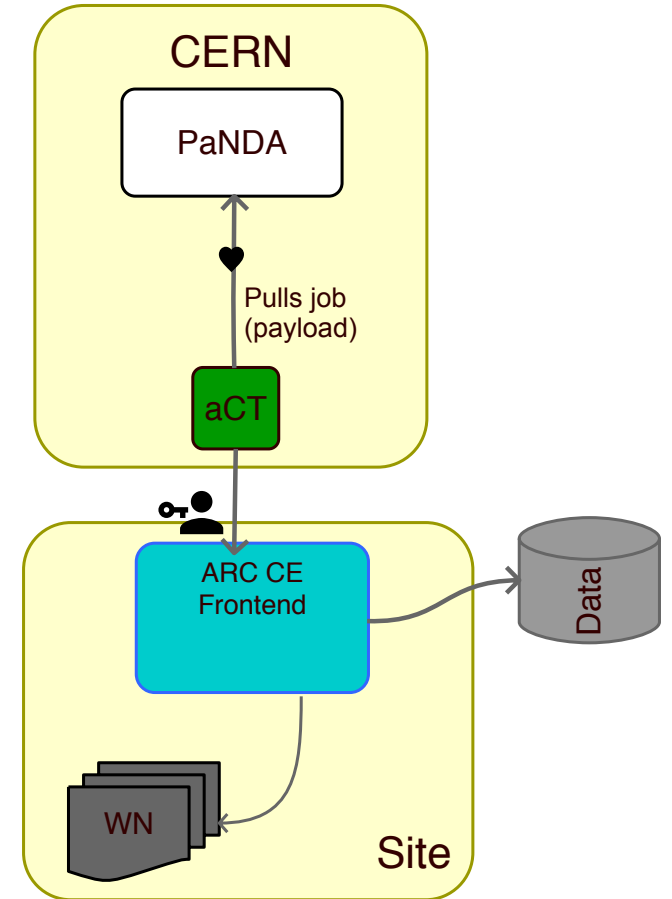
# Nordugrid ARC CE modes



Pilot factory

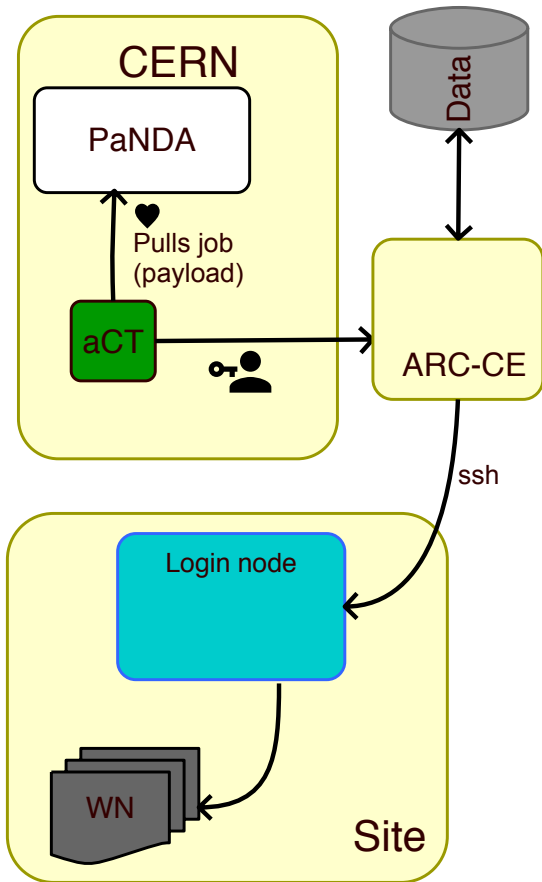


True pilot

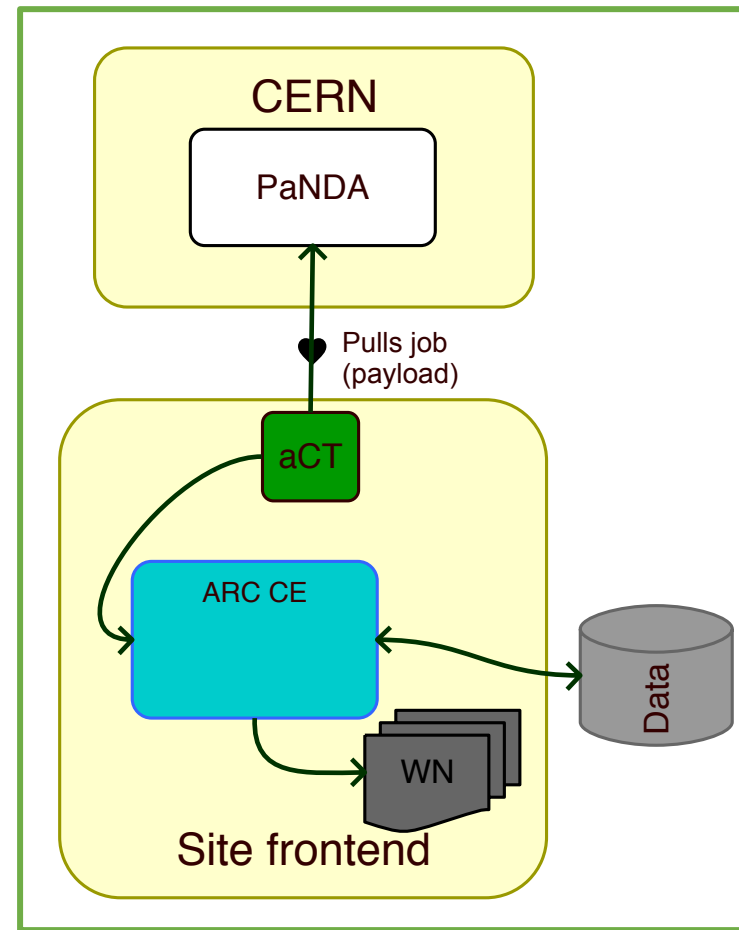


NDGF mode

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

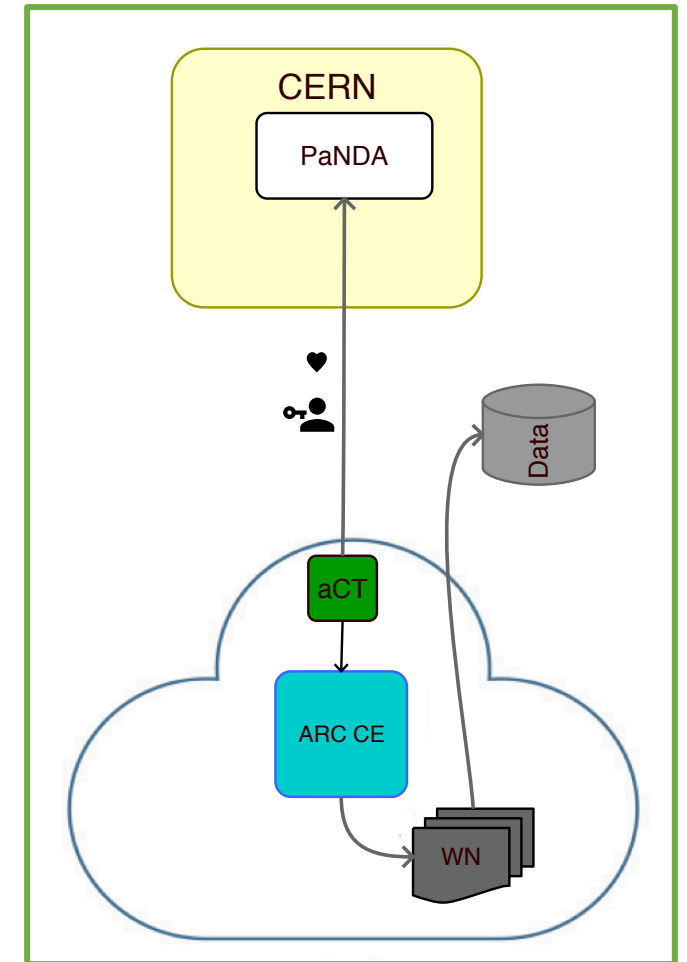


ssh-mode



INTERNAL mode HPC

Maiken Pedersen - UiO - CHEP 2018



INTERNAL mode cloud