Nordugrid Conference 2017

2017-06-29, University of Tromsoe

About 23 min slot

b
UNIVERSITÄT
BERN

Science IT Support

ARC Front-End to Clouds

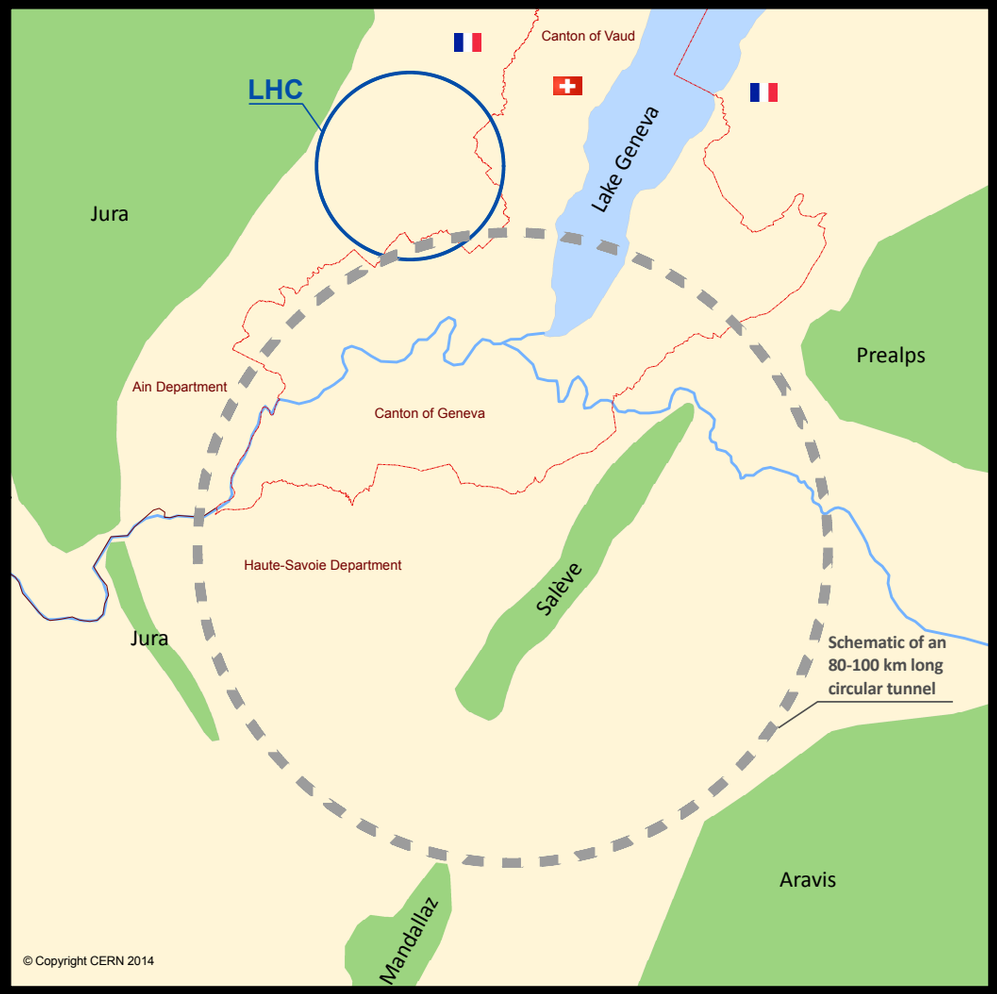
Demonstration

Sigve Haug, University of Bern

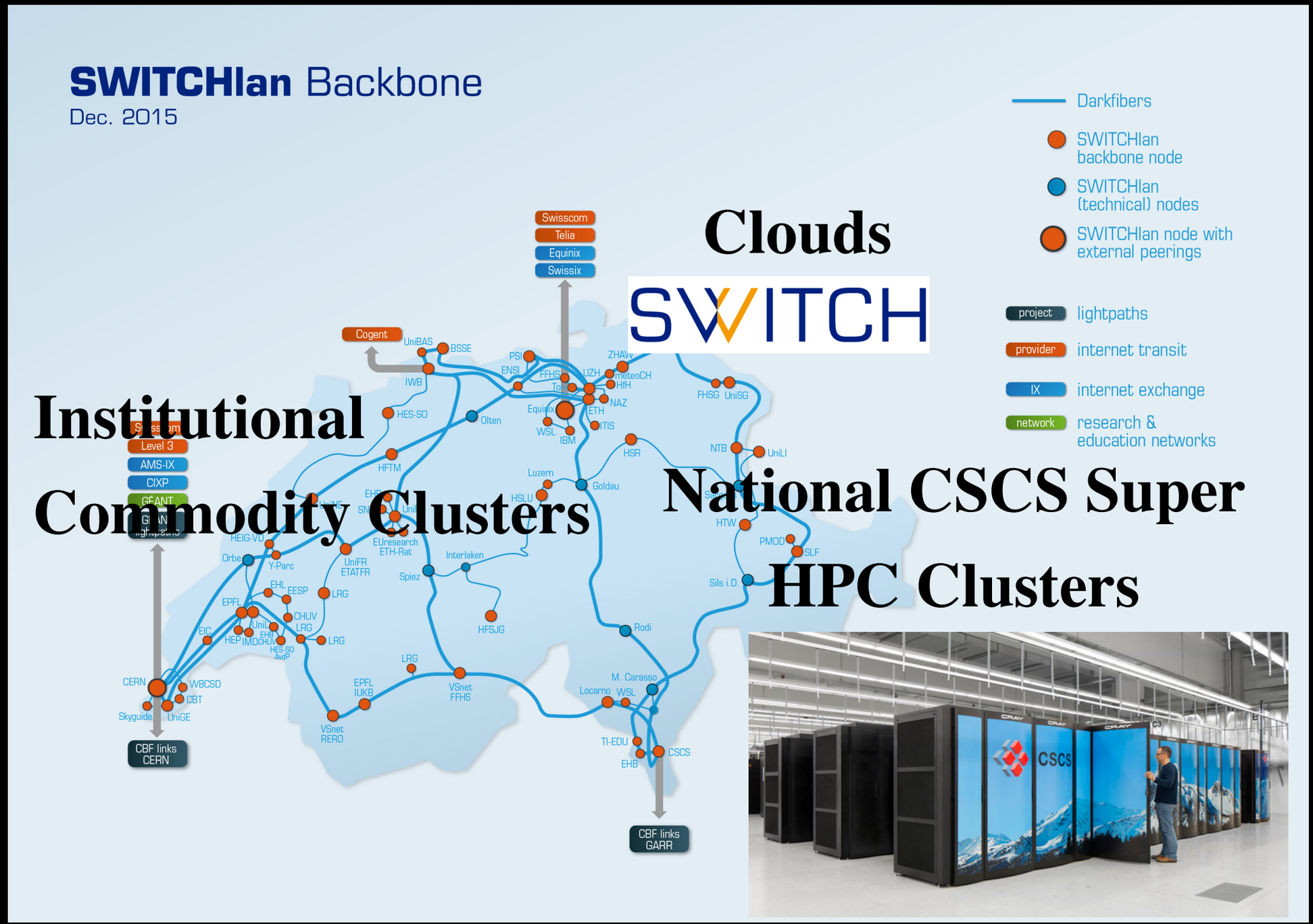
sigve.haug@math.unibe.ch

Motivations

- Get a O(1000) cores linux cluster within hours
- Not run hardware/infrastructure yourself (like for small LHC)
- Transparent costs
-
- Run every where … also on clouds



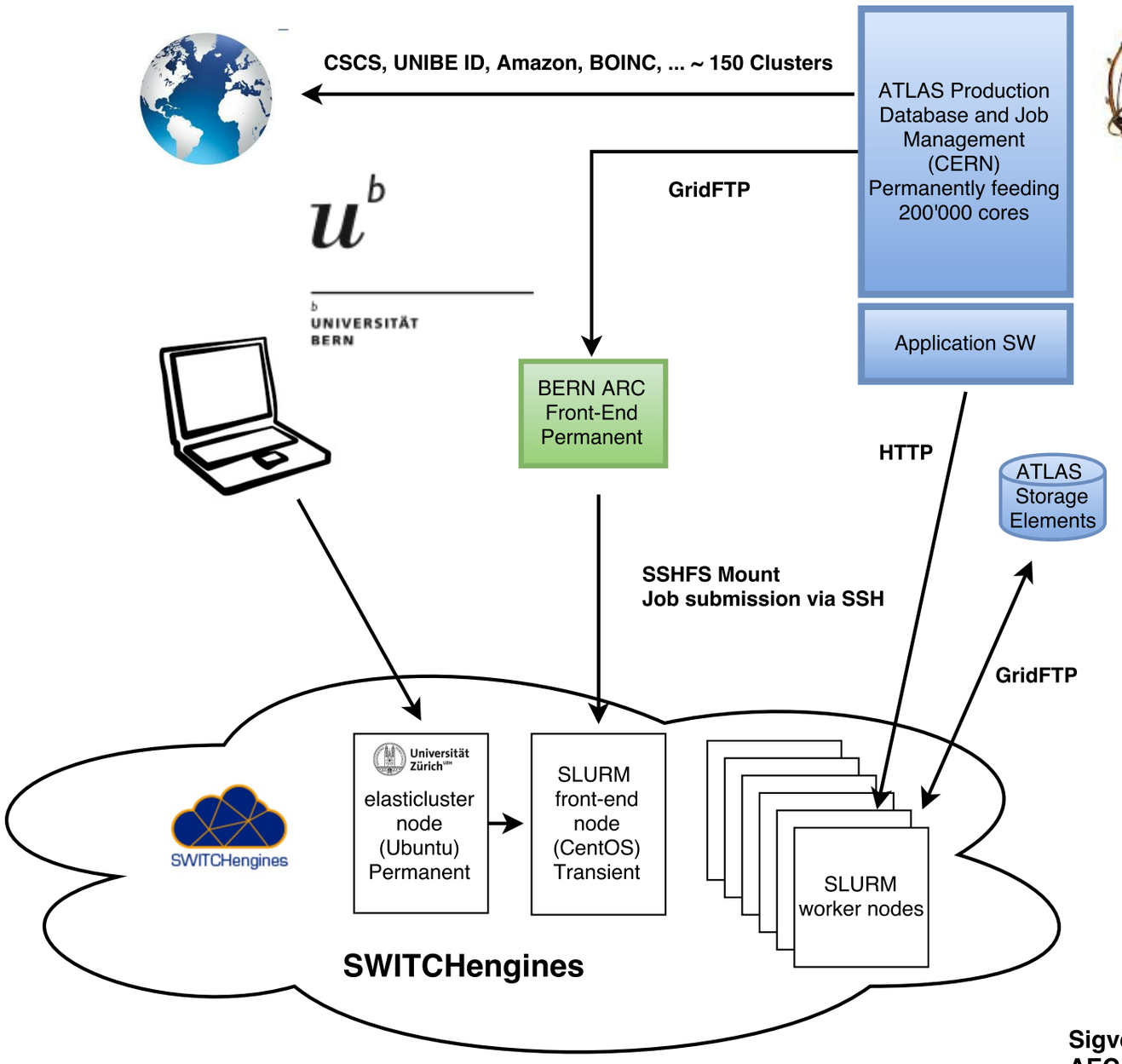
Swiss HPC landscape



Demonstration outline

- Sketch of the solution
- Access the cloud
- Create the cluster (Slurm)
- Connect to a grid with ARC (ATLAS production system)
- Wrap up

Sketch of the solution



ATLAS Experiment at CERN is using SWITCHengines 24/7 (we can set up a 1000 cores cluster for ATLAS in 1h)

ATLAS Instances on SWITCH

- 8 Cores
- 16 GB RAM
- 20 GB Disk

Currently running 300 cores 24/7
(CSCS running ~1500 cores 24/7)

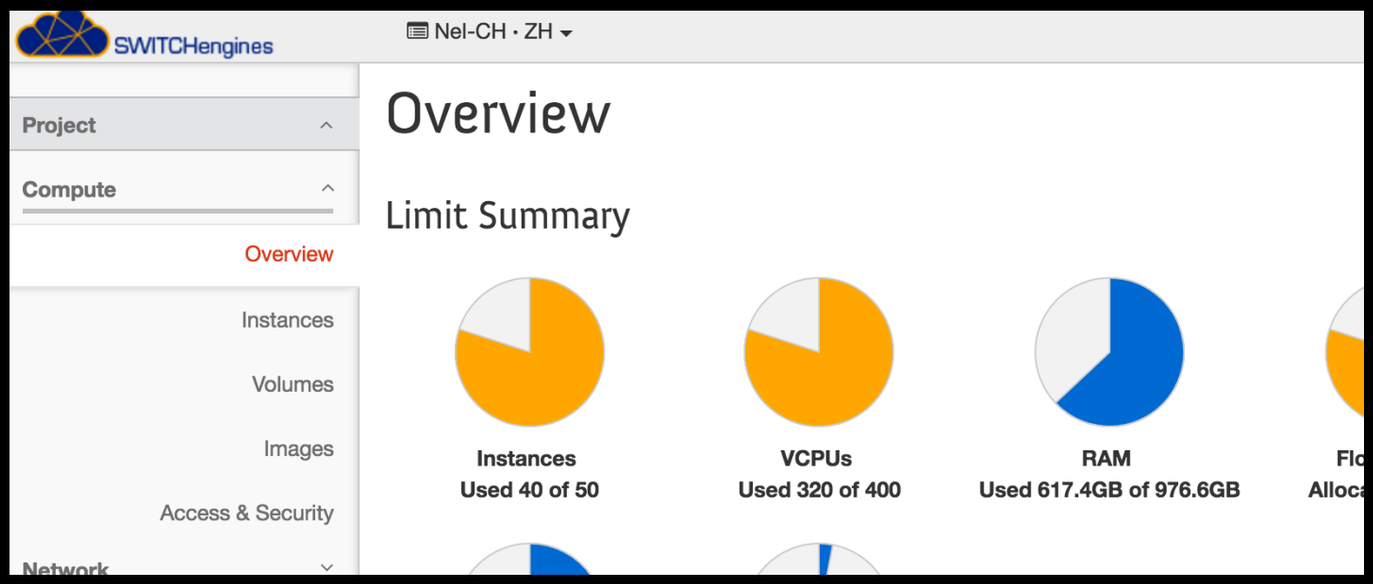
Sigve Haug, Hang Liu, Michi Hostettler, Gianfranco Sciacca, AEC University of Bern 2016

Compute cluster becomes transient (30 min setup), ARC front-end is the persistent part

Access SWITCHengines

- Get account on IaaS (an email) with some quota
- Make an instance for elasticluster (ubuntu)
- Make the worker node instance
- (here CentOS, mounted cvmfs, installed some stuff to make ATLAS applications run)

Jens-Christian Fischer
Product Owner
+41 44 268 15 71



ATLAS-compute038	8	20GB	15.6GB
ATLAS-frontend001	8	20GB	15.6GB
elasticluster-Nel	8	20GB	8GB

- https://www.switch.ch/engines/

Lets look at the OpenStack

- https://www.switch.ch/engines/

Setup HPC on SWITCHengines

- Fire up a SLURM cluster in 30 min (here Slurm) with one command

http://gc3-uzh-ch.github.io/elasticluster/

- Riccardo Murri, Sergio Maffioletti et al.

```
(elasticluster)ubuntu@elasticluster-nei:~$ tail .elasticluste
security_group=mpi_test
image_id=92cf2dc2-547c-4ab6-8d4f-9a383a4cf6e6
flavor=NeI-CH-8CPU-16GB_Ram
frontend_nodes=1
compute_nodes=38
image_userdata=
ssh_to=frontend
network_ids=c9e33fb0-5adf-4c81-97a6-a6eba639d0b1
(elasticluster)ubuntu@elasticluster-nei:~$ elasticluster star
(elasticluster)ubuntu@elasticluster-nei:~$ elasticluster list
(elasticluster)ubuntu@elasticluster-nei:~$ elasticluster list
The following clusters have been started.
Please note that there's no guarantee that they are fully con

ATLAS
-----
  name:           ATLAS
  template:       slurm
  - frontend nodes: 1
  - compute nodes: 38

(elasticluster)ubuntu@elasticluster-nei:~$
(elasticluster)ubuntu@elasticluster-nei:~$ elasticluster resi
5:compute ATLAS
(elasticluster)ubuntu@elasticluster-nei:~$ elasticluster stop
```

Let's look at the elasticluster

- Login, setup environment
- cat .elasticluster/config
- elasticluster start slurm -n NGC17
- Login to NGC17
- elasticluster resize -t slurm -a 5:compute NGC17

- elasticluster list
- elasticluster list-nodes ATLAS-EL
- ls -l /home/atlas on slurm front-end

Integrate cloud in a grid

www.nordugrid.org/atlas

- Cloned ARC HPC VM front-end for Cray (at lab, not in IaaS)
- ssh mounted /home/atlas from SWITCHengines activated our slum ssh back-end for ARC)

SWITCH OpenStack

Switzerland	ATLAS BOINC	85368	[illegible]+5406	1037+981
	ATLAS BOINC 3	85368	5336+5720	1005+1022
	ATLAS BOINC TEST	346	[illegible]	62+4499
	Bern ce01 (UNIBE-LHEP)	1497	1072+0	150+0
	Bern ce02 (UNIBE-LHEP)	770	576+0	145+0
	Bern ce04 (UNIBE-LHEP>	304	304+0	69+1
	Bern UBELIX T3	2968	141+2107	107+4203
	CSCS BRISI Cray XC40	240	261+5	0+0
	Geneva (UNIGE-DPNC)	568	8+205	[illegible]
	Lugano PHOENIX T2 arc>	2048	1818+3662	228+6
	Lugano PHOENIX T2 arc>	1920	1657+3823	242+0
	Lugano PHOENIX T2 arc>	1920	1709+3767	232+4

Volunteer Computing

Cray HPC

Let’s look at the ARC front-end

- Login to ce04
- ls -l /home/atlas (on the cloud frontend)
- how to mount it:

```
sshfs atlas@86.119.38.218:/home/atlas/ /home/atlas/ -o reconnect -o
allow_other -o workaround=rename -o idmap=file -o uidfile=/opt/sshslurm/
config/sshfs-cloud.uidmap -o gidfile=/opt/sshslurm/config/sshfs-cloud.gidmap
-o nomap=ignore -o ServerAliveInterval=30 -o ServerAliveCountMax=2 -o
IdentityFile=/opt/sshslurm/config/id_rsa.root -s -o nonempty
```

- ARC backend hacks: ls -l /opt/sshslurm
- cat /opt/sshslurm/sshslurm

- cat /opt/sshslurm/sshslurm

```
[root@ce04 ~]# cat /opt/sshslurm/sshslurm
#!/bin/bash

# config
source /opt/sshslurm/config/sshslurm-config

SBINARY=$(basename "$0")
SARGS=""
for token in "$@"; do
   SARGS="$SARGS '$token'"
#  echo $SARGS
done

echo $(date) - $SBINARY "$SARGS" >> /tmp/sshslurm.log

if [[ "$SBINARY" == "sbatch" && "$1" != "" ]]; then
   SARGS=$REMOTE_TEMP_PATH/$(basename "$1")
   $SCP_CMDLINE -q "$1" "$SSHSLURM_HOST:$SARGS"
   $SSH_CMDLINE $SSHSLURM_HOST -- [ -d "$PWD" ] \&\& cd "$PWD"\;
$REMOTE_SLURM_PATH/$SBINARY "$SARGS" \&\& rm -f "$SARGS"
   exit $?
fi

$SSH_CMDLINE $SSHSLURM_HOST -- [ -d "$PWD" ] \&\& cd "$PWD"\;
$REMOTE_SLURM_PATH/$SBINARY "$SARGS"

exit $?
[root@ce04 ~]#
```

```
[root@ce04 ~]# cat /opt/sshslurm/config/
id_rsa.griduser-michi      sshfs-cloud.gidmap        sshslurm-config
id_rsa.griduser-michi.old  sshfs-cloud.uidmap        sshslurm-config.test
id_rsa.root                sshfs-todi.gidmap         sshslurm-config.todi
id_rsa.root.old            sshfs-todi.uidmap


[root@ce04 ~]# cat /opt/sshslurm/config/sshslurm-config
SSHSLURM_HOST="atlas@86.119.38.218"
SSH_CMDLINE="/opt/openssh-6.6/bin/ssh -o "ControlPath=~/.ssh/controlmaster-%r@%h:%p" -o
"ControlMaster=auto" -o "ControlPersist=2h" -o "ServerAliveInterval=120" -i /opt/sshslurm
config/id_rsa.$(whoami)"
SCP_CMDLINE="/opt/openssh-6.6/bin/scp -o "ControlPath=~/.ssh/controlmaster-%r@%h:%p" -o
"ControlMaster=auto" -o "ControlPersist=2h" -o "ServerAliveInterval=120" -i /opt/sshslurm
config/id_rsa.$(whoami)"
REMOTE_SLURM_PATH="/usr/bin"
REMOTE_TEMP_PATH="/tmp"
[root@ce04 ~]#
```

Wrap up

- Setup of application dedicated O(1000) core clusters with elasticluster on an OpenStack IaaS within an hour is possible
- Hook this cluster to a remote ARC front-end works well for tested LHC tasks
- Performance is sufficient, very stable
- The cluster back-end becomes transient, i.e. can be reinstalled on the time-scale of changing a disk drive

Cluster compute becomes transient

Some previous presentations

- Master Thesis Hosttler https://cds.cern.ch/record/2253719

- At Digital Infrastructures for Research 2016, Cracow
  - https://www.digitalinfrastructures.eu/content/transient-compute-arc-cloud-front-end

- At Computing in High Energy Physics CHEP 2016, San Francisco
  - https://indico.cern.ch/event/505613/contributions/2230742/

Additional Material

ARC Bern ssh back-end

```
[root@ce04 ~]# ll /opt/sshslurm/
total 8
drwxr-xr-x. 2 root root 4096 Dec 18 17:50 config
lrwxrwxrwx. 1 root root    8 Apr 15  2014 sacct -> sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 sacctmgr -> sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 salloc -> sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 sattach -> sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 sbatch -> sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 sbcast -> sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 scancel -> sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 scontrol -> sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 sdiag -> sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 sinfo -> sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 sprio -> sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 squeue -> sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 sreport -> sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 srun -> sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 sshare -> sshslurm
-rwxr-xr-x. 1 root root  604 Nov 13  2014 sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 sstat -> sshslurm
lrwxrwxrwx. 1 root root    8 Apr 15  2014 strigger -> sshslurm
[root@ce04 ~]#


[root@ce04 ~]# cat /opt/sshslurm/config/sshslurm-config
SSHSLURM_HOST="atlas@86.119.38.88"
SSH_CMDLINE="/opt/openssh-6.6/bin/ssh -o "ControlPath=~/.ssh/controlmaster-%r@%h:%p" -o
"ControlMaster=auto" -o "ControlPersist=2h" -o "ServerAliveInterval=120" -i /opt/sshslurm/
config/id_rsa.$(whoami)"
SCP_CMDLINE="/opt/openssh-6.6/bin/scp -o "ControlPath=~/.ssh/controlmaster-%r@%h:%p" -o
"ControlMaster=auto" -o "ControlPersist=2h" -o "ServerAliveInterval=120" -i /opt/sshslurm/
config/id_rsa.$(whoami)"
REMOTE_SLURM_PATH="/usr/bin"
REMOTE_TEMP_PATH="/tmp"
[root@ce04 ~]#
```

ARC Bern ssh back-end

```
[root@ce04 ~]# cat /opt/sshslurm/sshslurm
#!/bin/bash

# config
source /opt/sshslurm/config/sshslurm-config

SBINARY=$(basename "$0")
SARGS=""
for token in "$@"; do
    SARGS="$SARGS '$token'"
#   echo $SARGS
done

echo $(date) - $SBINARY "$SARGS" >> /tmp/sshslurm.log

if [[ "$SBINARY" == "sbatch" && "$1" != "" ]]; then
    SARGS=$REMOTE_TEMP_PATH/$(basename "$1")
    $SCP_CMDLINE -q "$1" "$SSHSLURM_HOST:$SARGS"
    $SSH_CMDLINE $SSHSLURM_HOST -- [ -d "$PWD" ] \&\& cd "$PWD"\; $REMOTE_SLURM_PATH/
$SBINARY "$SARGS" \&\& rm -f "$SARGS"
    exit $?
fi

$SSH_CMDLINE $SSHSLURM_HOST -- [ -d "$PWD" ] \&\& cd "$PWD"\; $REMOTE_SLURM_PATH/$SBINARY
"$SARGS"

exit $?
[root@ce04 ~]#


sshfs atlas@86.119.38.88:/home/atlas/ /home/atlas/ -o reconnect -o allow_other -o
workaround=rename -o idmap=file -o uidfile=/opt/sshslurm/config/sshfs-cloud.uidmap -o
gidfile=/opt/sshslurm/config/sshfs-cloud.gidmap -o nomap=ignore -o ServerAliveInterval=30
-o ServerAliveCountMax=2 -o IdentityFile=/opt/sshslurm/config/id_rsa.root -s -o nonempty
```