b
UNIVERSITÄT
BERN

AEC
ALBERT EINSTEIN CENTER
FOR FUNDAMENTAL PHYSICS

ARC as Cloud Front-End
Swiss Experience - Transient Compute

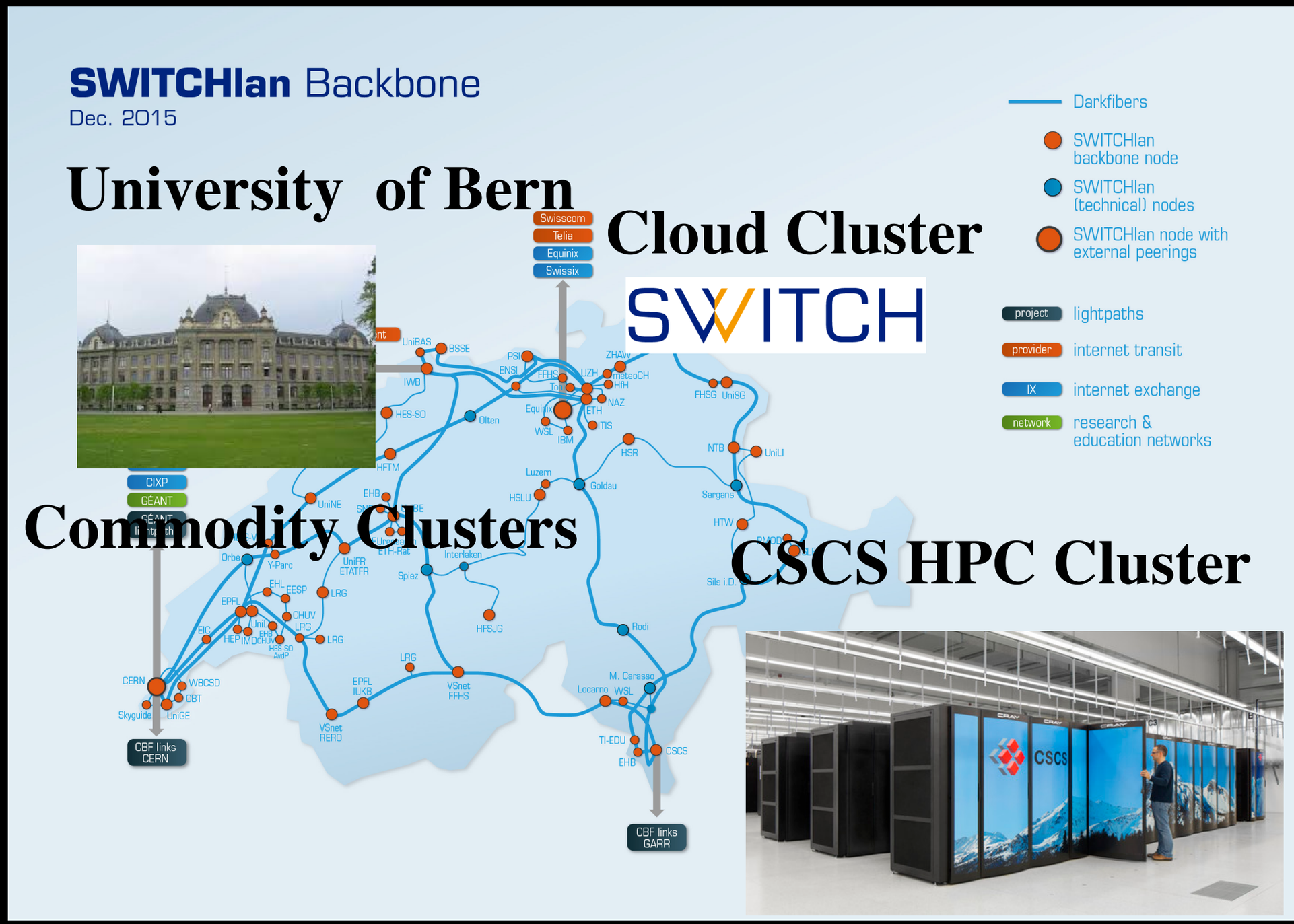
Sigve Haug, Gianfranco Sciacca, AEC-LHEP University of Bern

sigve.haug@lhep.unibe.ch, gianfranco.sciacca@lhep.unibe.ch

Answer from an CH ATLAS/Science perspective

Why ?

A new player in the CH compute provider business, can we use her ?



WHY cont.

- Suddenly the Swiss NREN (SWITCH) decided to become a compute and storage (IaaS) provider with OpenStack
- Several CH universities also have OpenStack infrastructures by now, so better check it out for our science at AEC
- Federal infrastructure funding for free trial on SWITCHengines, ok … if it is for free … let's release the ATLAS beast onto it
- Some nice tools make it easy : ARC and elasticluster
- Alternative to buying own hardware, fight for HPC allocations or deal with rigid central batch clusters and policies ?
- Now a small academic compute market in CH, can science benefit from this ?

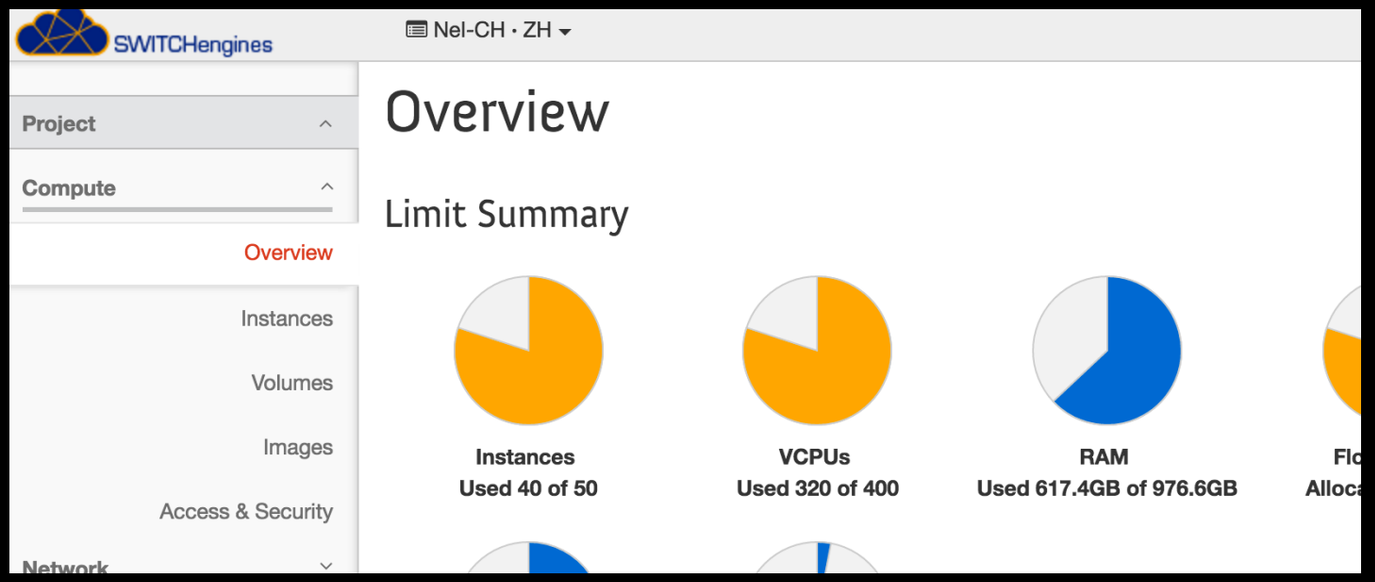
www.switch.ch/engines/

HOW ?

- Got an account on SWITCHengines (an email) with some quota
- Made an instance for elasticluster (ubuntu)
- Made an instance for ATLAS with CentOS, mounted /cvmfs, installed some stuff to make ATLAS run. Made a snapshot (image).
- Fired up a SLURM cluster with 304 cores with that image (30 min)

- SWITCHengines

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



www.s3it.uzh.ch/software/elasticluster/

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

- Riccardo Murri
- Sergio Maffioletti

Elasticluster in action

- Get a cluster in 30 minutes

```
(elasticluster)ubuntu@elasticluster-nei:~$ tail .elasticluster/config
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 start slurm -n ATLAS
(elasticluster)ubuntu@elasticluster-nei:~$ elasticluster list

The following clusters have been started.
Please note that there's no guarantee that they are fully configured:

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

(elasticluster)ubuntu@elasticluster-nei:~$
(elasticluster)ubuntu@elasticluster-nei:~$ elasticluster resize -t slurm -a
5:compute ATLAS
(elasticluster)ubuntu@elasticluster-nei:~$ elasticluster stop ATLAS
```

HOW cont.

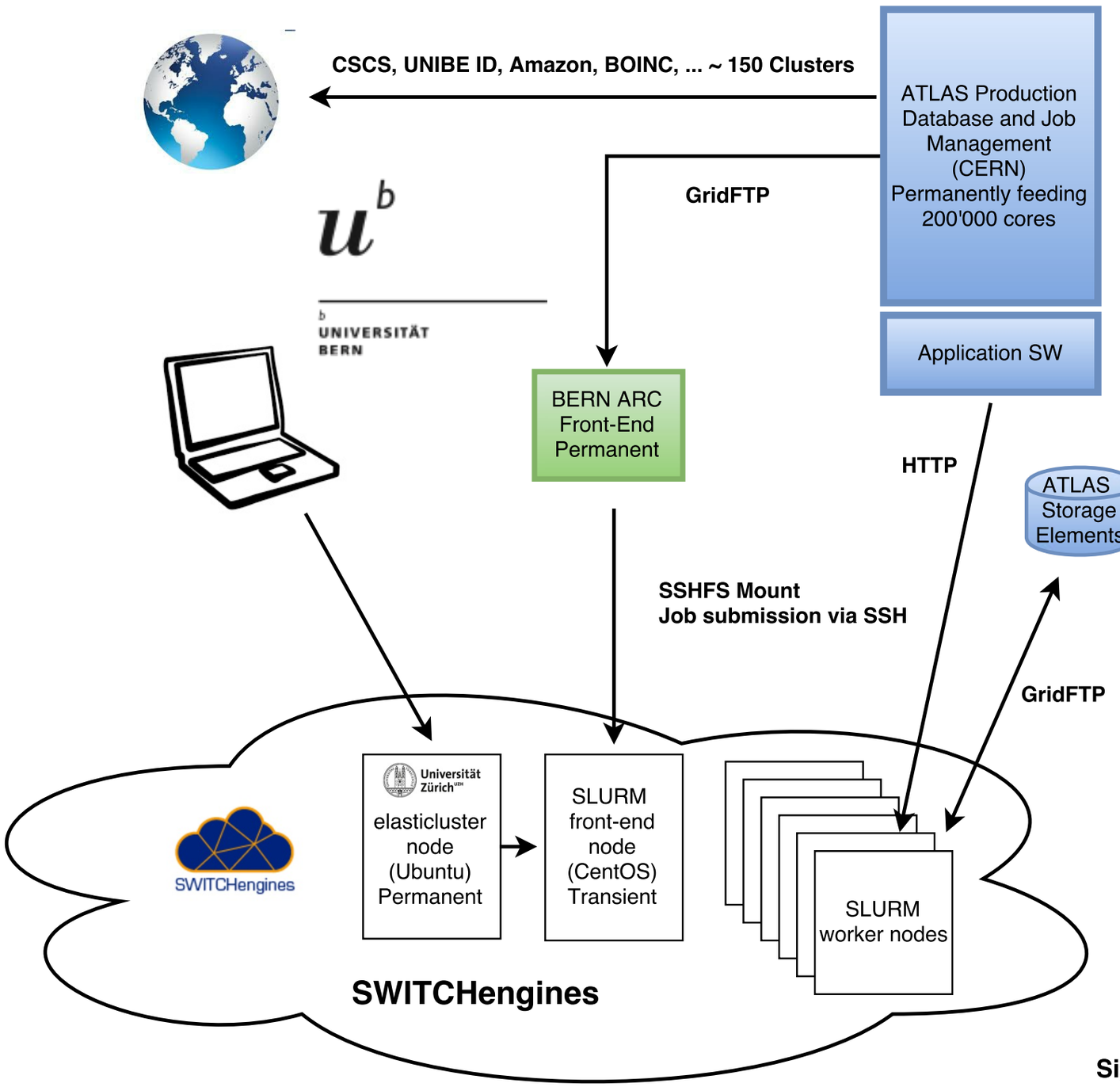
www.nordugrid.org/atlas

- Cloned our ARC HPC VM front-end in Bern
- ssh mounted /home/atlas from SWITCHengines and activated our ssh back-end (small wrapper around standard ARC slurm back-end)
- Registered front-end in ATLAS production system
- Started running

OpenStack

Switzerland	ATLAS BOINC	85368	5356+5406	1037+981
	ATLAS BOINC 3	85368	5336+5720	1005+1022
	ATLAS BOINC TEST	346	0+4	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	0+0
	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

The poster sketch



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

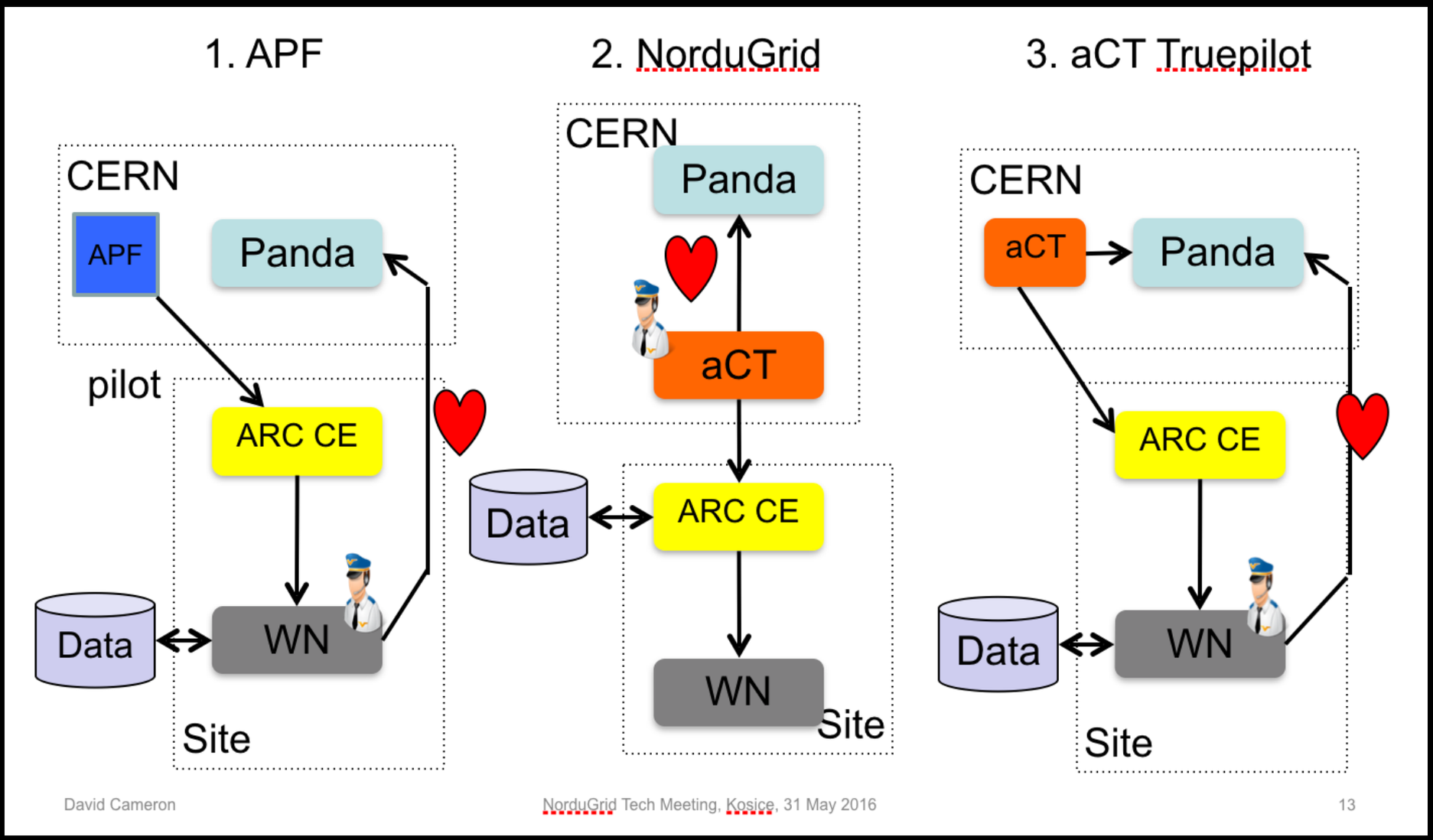
- 8 Cores
- 16 GB RAM
- 20 GB Disk

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

- **The compute cluster has become transient**

Running in “true pilot” mode



APF : ATLAS Pilot Fabric

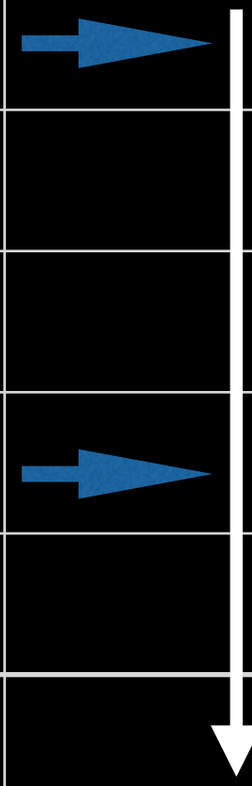
aCT : ATLAS Control Tower

Panda: Workload Manager

No I/O restrictions on SWITCH

So it makes sense to let WN do I/O

Workflow per ATLAS event

Input	Flow	Processing Step	Format	V / MB	t / s	x
Theory		Generation	EVNT	0.04	980	T2
		Detector Simulation	HITS	1.0	3600/300	T1/T2
		Digitization	RDO	3.6		T1/T2
Data			Raw	1.0		T0
		Reconstruction	ESD/AOD	3.6/0.5 2.6/0.3	570 230	T0/T1
		Analysis	NTUP		0.4	T2

So far running only Generation and Detector Simulation step on SWITCHengines - moderate I/O

Performance

You don’t check the mouth of a horse you get for free …

However, impression is that performance is good. So far jobs are CPU bound when using local disk

Will quantify the performance

Pleasant stability, basically 100% up, no downtimes

and prices for “ATLAS/LHC” cores

Table 1 : End 2015 prices in CHF per logical core with 2 GB RAM and about 2.5 GB (CSCS) and 7.5 GB (LHEP) and 2.5 (SWITCHengines) disk. HW prices from Dalco offers. Lifetime is 5 years.

2016	LHEP (own cluster)	CSCS (HPC)	SWITCHengines (SE)	SWITCHengines Preemptive
HW	116	114		
FTE		233		
Scratch license		33		
Installation		2		
SUM	116	382	919	?

- HW is not cheaper at (this) scale
- “Market” is not fair due to subventions
- The cloud solution is most convenient, but still too expensive

1 CPU - 45 Cloud Credits
1 GB RAM - 22.5 Cloud Credits
1 GB Disk Storage - 0.1375 Cloud Credits
1 GB Object Storage - 0.175 Cloud Credits
1 IP address - 2.5 Cloud Credits

Wrap up - connecting a cloud to grid

- In CH academic resource providers now expose OpenStack

- Fire up application dedicated O(1000) core clusters with elasticluster within an hour is possible

- Hook this cluster to a remote ARC front-end works well for some LHC tasks

- A convenient alternative to dedicated or shared HPC resources. Funding scheme/unfair competition is the current barrier in CH.

- compute becomes transient

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
```

LHC and data taking parameters		2016 pp μ=30 @ 25 ns	2017 pp μ=35 @ 25 ns	2018 pp μ=35 @ 25 ns
Rate [Hz]	Hz	1000	1000	1000
Time [sec]	MSeconds	5.0	**5.5 (was 5.1)**	5.5
Real data	B Events	5.0	**5.5 (was 5.1)**	5.5
Simulated Data				
Full Simulation	B Events	5.0	**5.15 (was 3.8)**	**5.15**
Fast Simulation	B Events	2.5	**3.1 (was 7.5)**	**3.6**
Generator	B Events	**15.0**	**14.8**	**14.0**
Event sizes				
Real RAW	MB	1	1	1
Real ESD	MB	2.6	2.6	2.6
Real AOD	MB	0.28	0.32	0.32
Sim HITS	MB	1	1	1
Sim ESD	MB	3.6	3.6	3.6
Sim AOD	MB	0.45	0.50	0.50
Sim RDO	MB	3.6	3.6	3.6
EVNT	MB	0.04	0.04	0.04
CPU times per				
Full sim	HS06 sec	3500	3500	3500
Fast sim	HS06 sec	300	300	300
Generators	HS06 sec	980	980	980
Real recon	HS06 sec	210	230	230
Sim recon	HS06 sec	533	567	567
AOD2AOD data	HS06 sec	25	25	25
AOD2AOD sim	HS06 sec	60	60	60
Group analysis	HS06 sec	3	3	3
User analysis	HS06 sec	0.4	0.4	0.4