- Motivation and challenges
- Non-invasive ARC approach
- The LHConCRAY project: ARC in the HPC centre

INTEGRATING PIZ DAINT @ CSCS IN WLCG WITH ARC

STATUS REPORT

Gianfranco Sciacca

AEC - Laboratory for High Energy Physics, University of Bern, Switzerland

Nordugrid Conference 2017 - 27-30 June 2017, Tromsø, Norway

UNIVERSITÄT BERN
AEC ALBERT EINSTEIN CENTER FOR FUNDAMENTAL PHYSICS

LABORATORIUM FÜR HOCHENERGIEPHYSIK LHEP UNIVERSITÄT BERN

MOTIVATION

More computing at flat budget

- **Our Swiss WLCG budget** (like everybody else's) **most likely to be flat** (more likely cut than increased)
- **The current computing models do not scale for High Luminosity LHC (beyond 2020)**
- The challenge is open, and use of HPC centres might play a major role
- In CH we operate a dedicated x86_64 WLCG cluster at CSCS (**Phoenix**) - ATLAS, CMS, LHCb
- **Opportunity to operate on Piz Daint, flagship HPC in Europe, 3rd ranking worldwide**

TOP500 The List. @ ISC High Performance The HPC Event.
JUNE 19 - 23, 2016
FRANKFURT, GERMANY

Rank	Site		Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)
1	National Supercomputing Center in Wuxi (/preview/site/50623) China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway (/preview/system/178764) NRCPC	10,649,600	93,014.6	125,435.9
2	National Super Computer Center in Guangzhou (/preview/site/50365) China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P (/preview/system/177999) NUDT	3,120,000	33,862.7	54,902.4
3	Swiss National Supercomputing Centre (CSCS) (/preview/site/50422) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 (/preview/system/177824) Cray Inc.	361,760	19,590.0	25,326.3
4	DOE/SC/Oak Ridge National Laboratory (/preview/site/48553) United States	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x (/preview/system/177975) Cray Inc.	560,640	17,590.0	27,112.5
5	DOE/NNSA/LLNL (/preview/site/49763) United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom (/preview/system/177556) IBM	1,572,864	17,173.2	20,132.7
6	DOE/SC/LBNL/NERSC (/preview/site/48429) United States	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect (/preview/system/178924) Cray Inc.	622,336	14,014.7	27,880.7

CHALLENGES

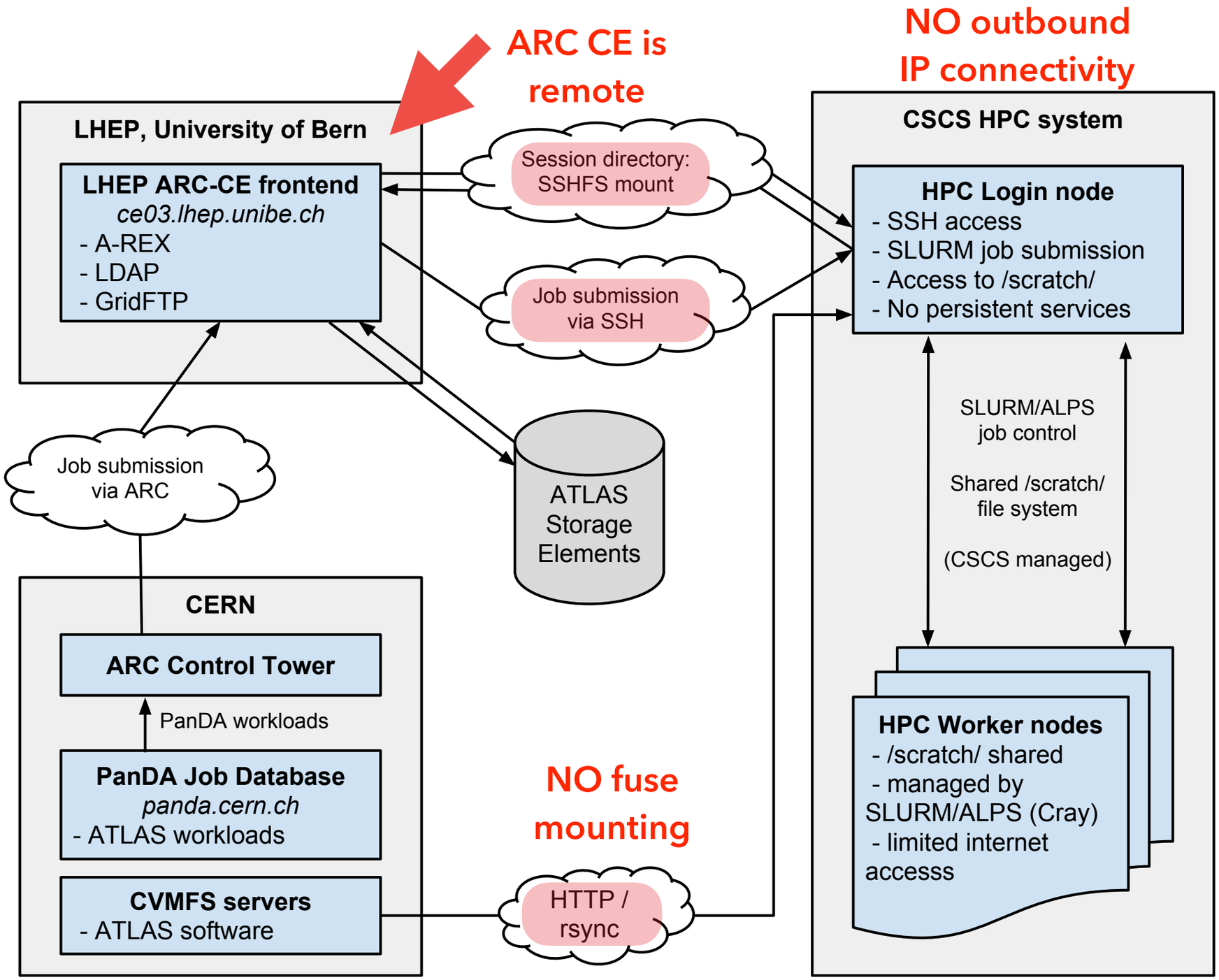
WLCG computing on HPC systems

- **Several challenges arise**
  - **Processor architecture and/or OS might not always be suitable**
    complex software re-builds, environment tweaking, etc..
  - **Compliance with tight access rules**
    single-user access, username/password
  - **Application provisioning**
    a single ATLAS release is ~20GB, release cycles are very short/unpredictable
  - **Workload management integration**
    requires in general outbound IP connectivity
  - **Data input and retrieval**
    for real data processing: ~0.2MB/s/core IN, ~0.1MB/s/core OUT

HPCS ARE VERY RESTRICTED AND SELF-CONTAINED ENVIRONMENTS!

NON INVASIVE ARC APPROACH

First approach to a Cray at CSCS, Tödi (2014/15)



- CRAY XK7 (Tödi)
  - Former CPU/GPU development system
  - 16-core AMD Opteron CPU, 32GB RAM
  - SLURM / ALPS
  - Same system as Titan

- Does NOT require:
  - middleware on the WN
  - outbound connectivity
  - synchronous data-staging

5 MONTHS IN PRODUCTION RUNNING ATLAS G4 UNMODIFIED BINARIES OUT OF CVMFS

- performed 30% better than binaries re-compiled with the Cray compiler

- Master thesis work by Michael Hostettler, Universität Bern , 2015

THE LHCONCRAY PROJECT

Consolidation project to run LHC jobs on Piz Daint

- Ran for over 1 year on a Cray XC40 Test Design System @ CSCS (Brisi)
  - Integrate all experiment workloads (ATLAS, CMS, LHCb)
  - Objective: from proof-of-concept to pre-production

Phase 1 Feasibility study

Objective: Evaluate whether it is actually possible to run WLCG jobs in a Cray system

1. **Enabling job submission from the grid**
   a. Adapt the infrastructure
   b. Adapt CE software
2. **Software validation**
   a. Base grid software
   b. Application software

SLURMmod.pm
submit-SLURM-job

(ensure the payload is called within Shifter)

glue-generator.pl

Phase 2 Integration

Objective: Attempt to integrate WLCG jobs on CSCS infrastructure.

3. **Scheduling & Accounting**
   a. Multiple jobs per node
   b. Fairshare complexity

CVMFS + RUNNING UNMODIFIED APPLICATIONS WITH SHIFTER

Phase 3 Evaluation at scale

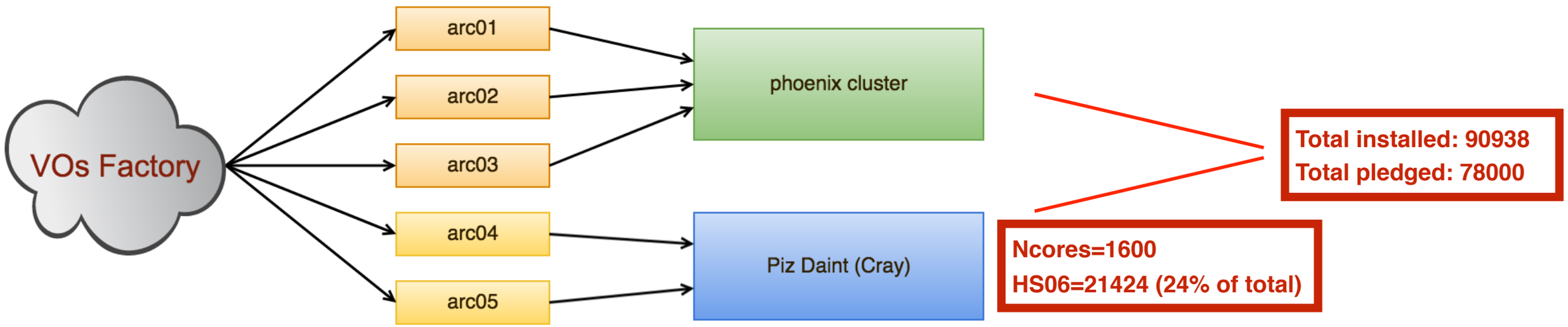
Objective: Evaluate and control impact of WLCG jobs on CSCS infrastructure.

4. **Scalability**
   a. High rate of job submission/sec
   b. High impact on I/O systems

THE LHCONCRAY PROJECT

Consolidation project to run LHC jobs on Piz Daint

- **Recently started production on Piz Daint: 1600 cores (ATLAS:CMS:LHCb - 40:40:20)**
- **The goal is to run all T2 experiment workloads without changes to the workflows**



- **Normal workflow:**
  - Jobs submitted via ARC, SLURM LRMS
  - Running in containers (Shifter)
  - CVMFS Native on Cray nodes
- **Roadmap**
  - Measure performance in production environment, produce a cost study (until Dec. 2017)
  - Decision due: continue or revert to invest on Phoenix

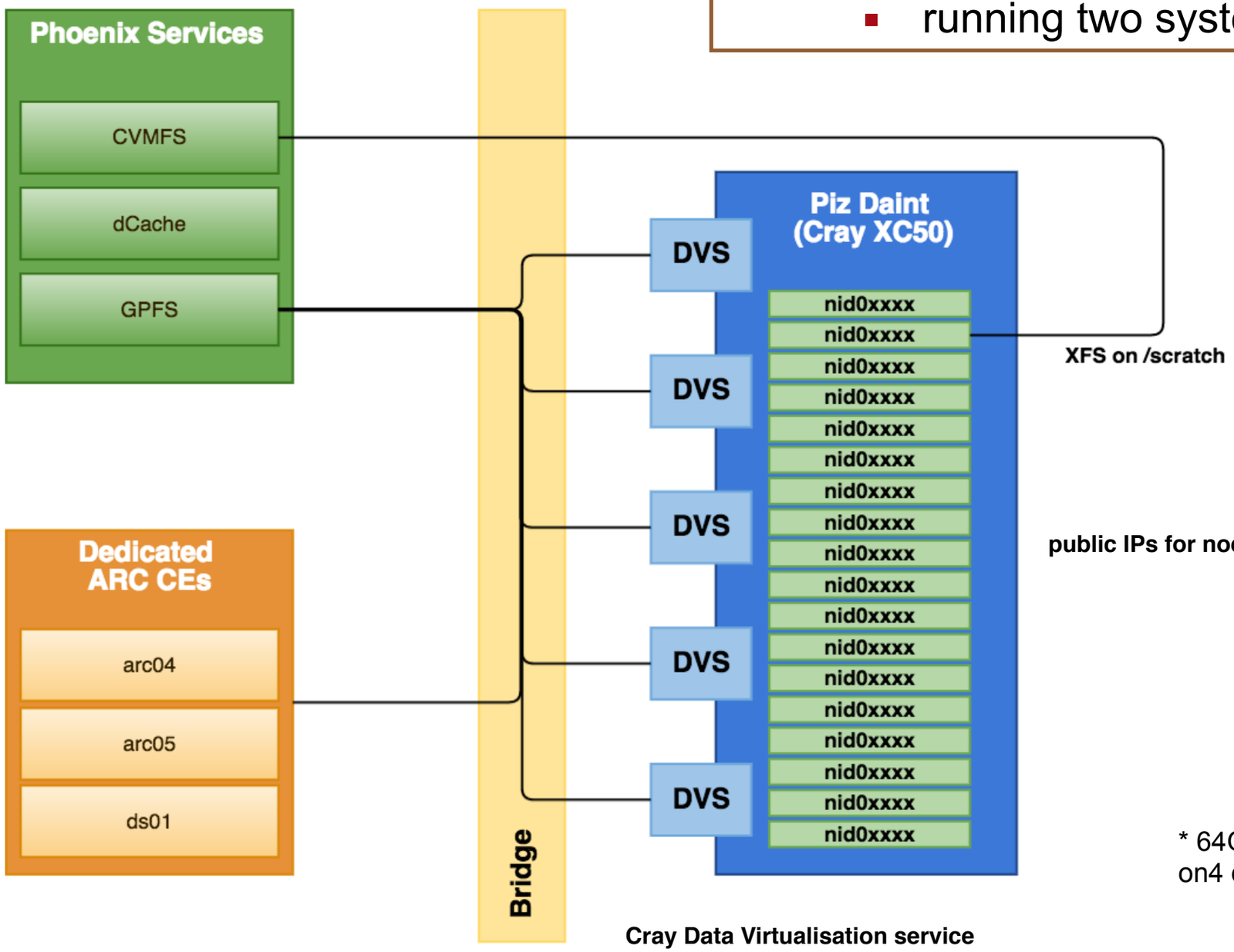
- **Broadwell (Intel Xeon E5-2695 v4 @2.10GHz)**
- **72 HT-cores (64 used), 128GB RAM**
- **CLE 6.0.UP02**
- **Cray Aries interconnect**
- **Native SLURM 17.02.3-1**
- **Nodes exclusive to WLCG jobs**

THE LHCONCRAY PROJECT

Consolidation project to run LHC jobs on Piz Daint

A SLIGHLTY MODIFIED ARC CE IS NOW FULLY INTEGRATED

- Key challenges are now:
  - memory management (just enabled swap*)
  - shared file system performance
  - running two systems in parallel



* 64GB SSD-based swap space per node on4 dedicated DataWarp nodes [1].

THE LHCONCRAY PROJECT

Problem solved?

✓ **Processor architecture and/or OS might not always be suitable**

jobs run within **Shifter containers** [1]. The container itself is a CentOS 6.8 full image with the same packages as in the dedicated WLCG T2 cluster (Phoenix) and configured accordingly [2-3]

✓ **Compliance with tight access rules**

access policies relaxed (project endorsed by the RC). Middleware **INSIDE** the centre

✓ **Application provisioning**

CVMFS, with one loopback XFS FS (one single sparse XFS file) **for the cache per node** [4]

✓ **Workload management integration**

leverage the ARC CE technology. Integrated seamlessly with the ATLAS, CMS and LHCb factories. Nodes inside the Cray High Speed Network now use **public IPs** with standard Linux IP packet forwarding

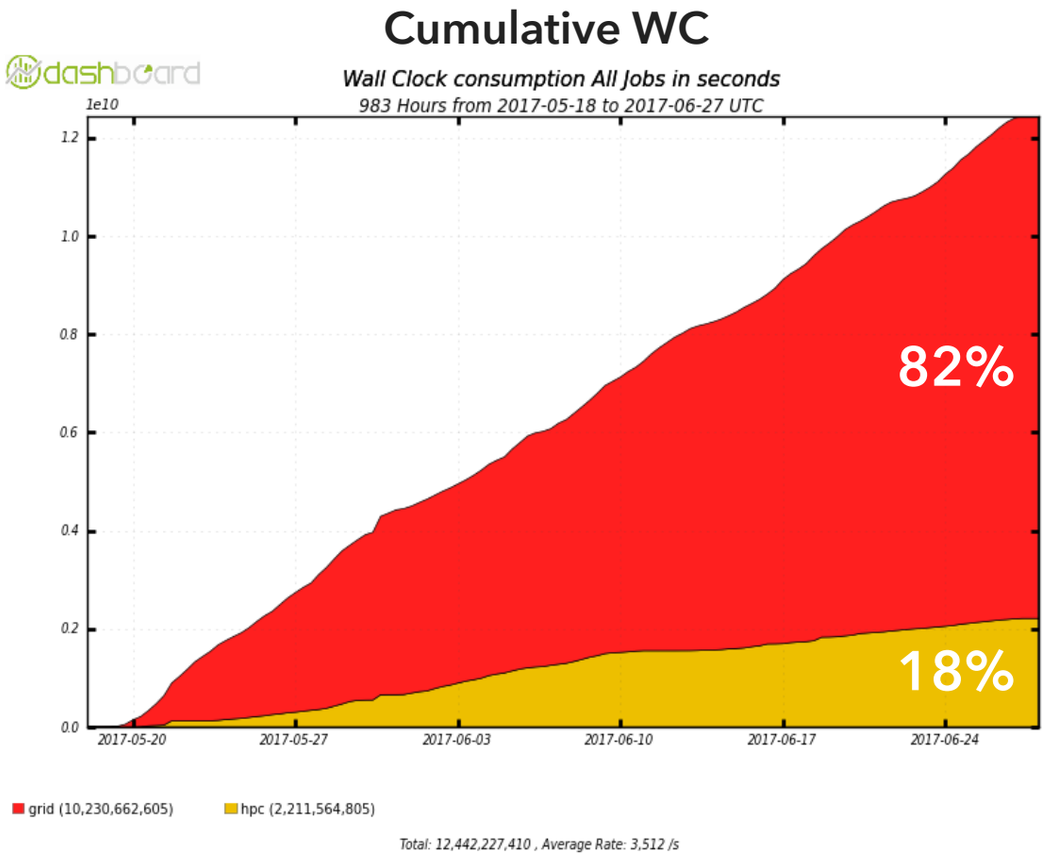
✓ **Data input and retrieval**

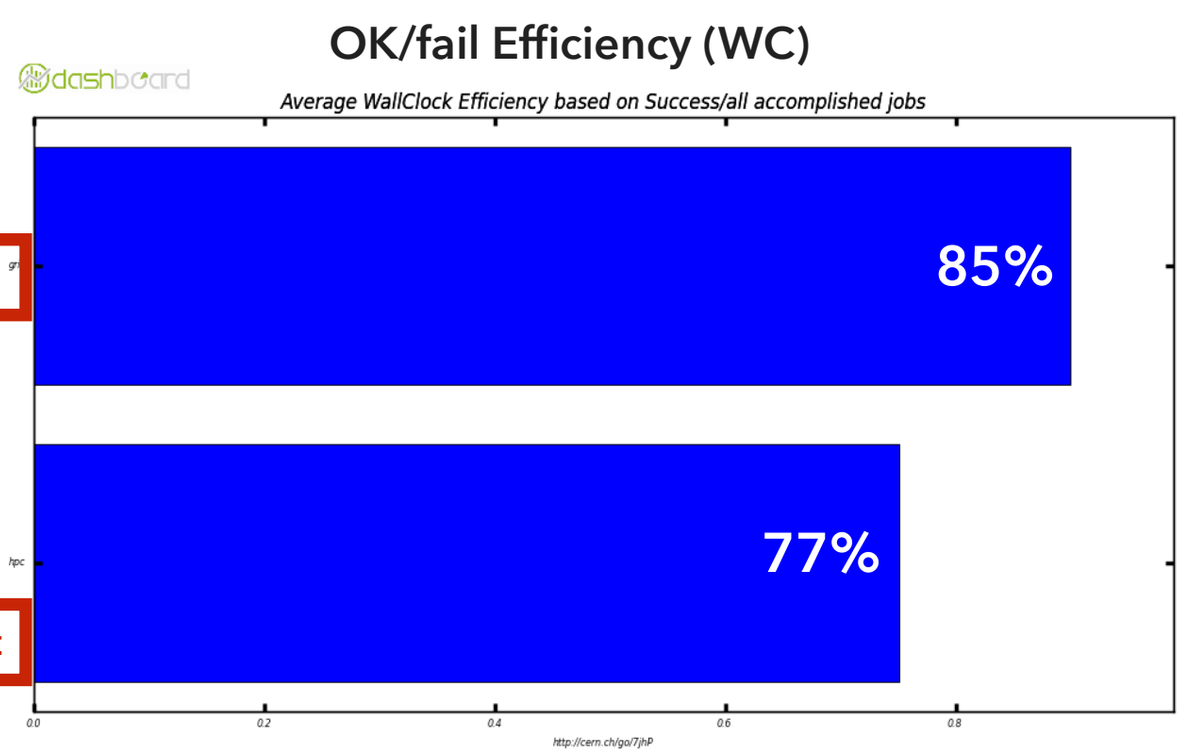
leverage the ARC CE technology. Will need testing at scale

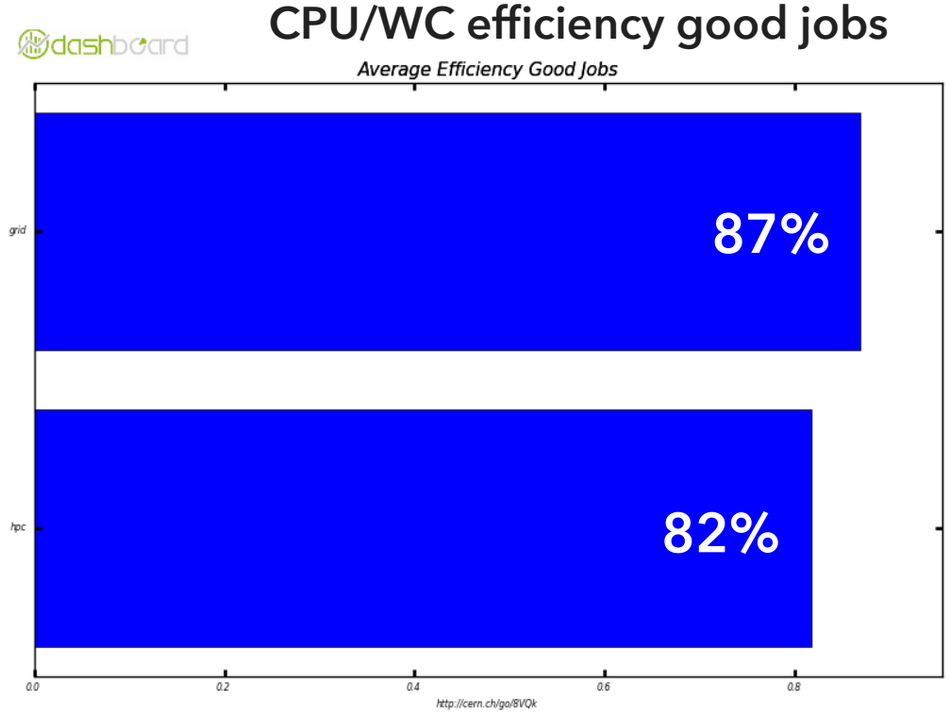
[1] http://www.nersc.gov/research-and-development/user-defined-images
[2] https://hub.docker.com/r/cscs/wlcg_wn/
[3] https://github.com/miguelgila/docker-wlcg_wn/blob/master/Dockerfile
[4] http://cvmfs.readthedocs.io/en/stable/cpt-hpc.html#loopback-file-systems-for-nodes-caches

THE LHCONCRAY PROJECT

Piz Daint in production with ARC CE

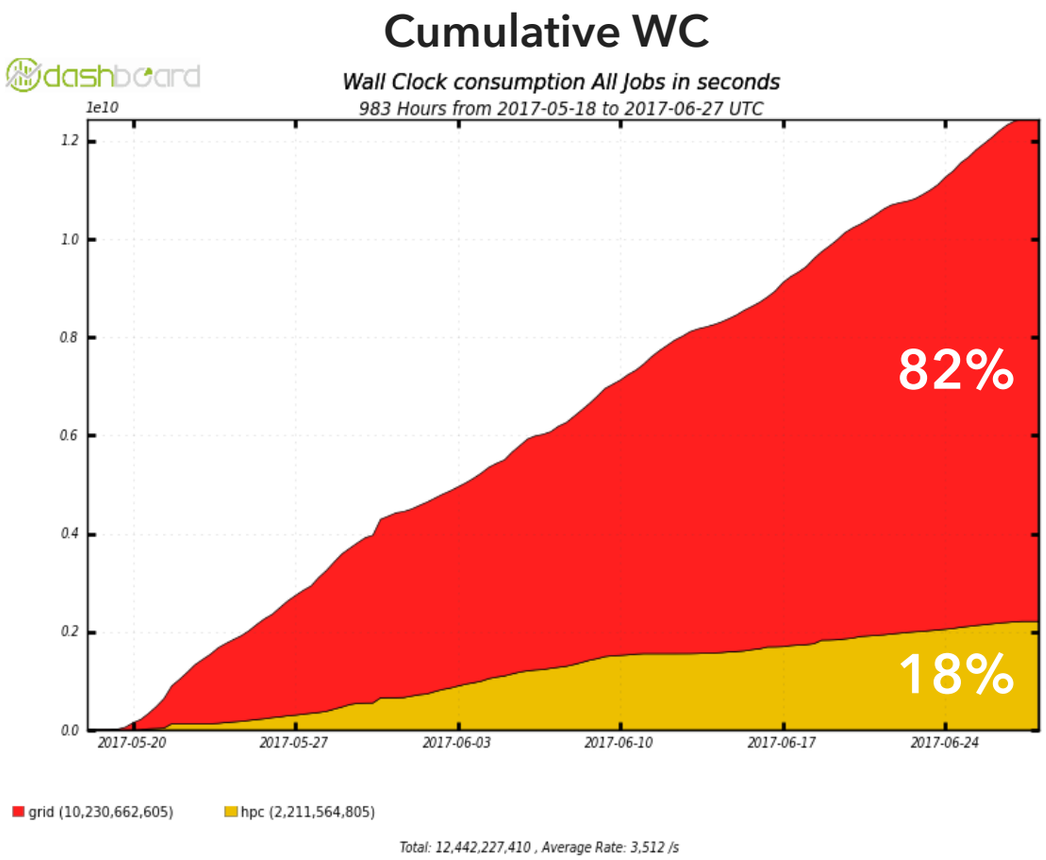


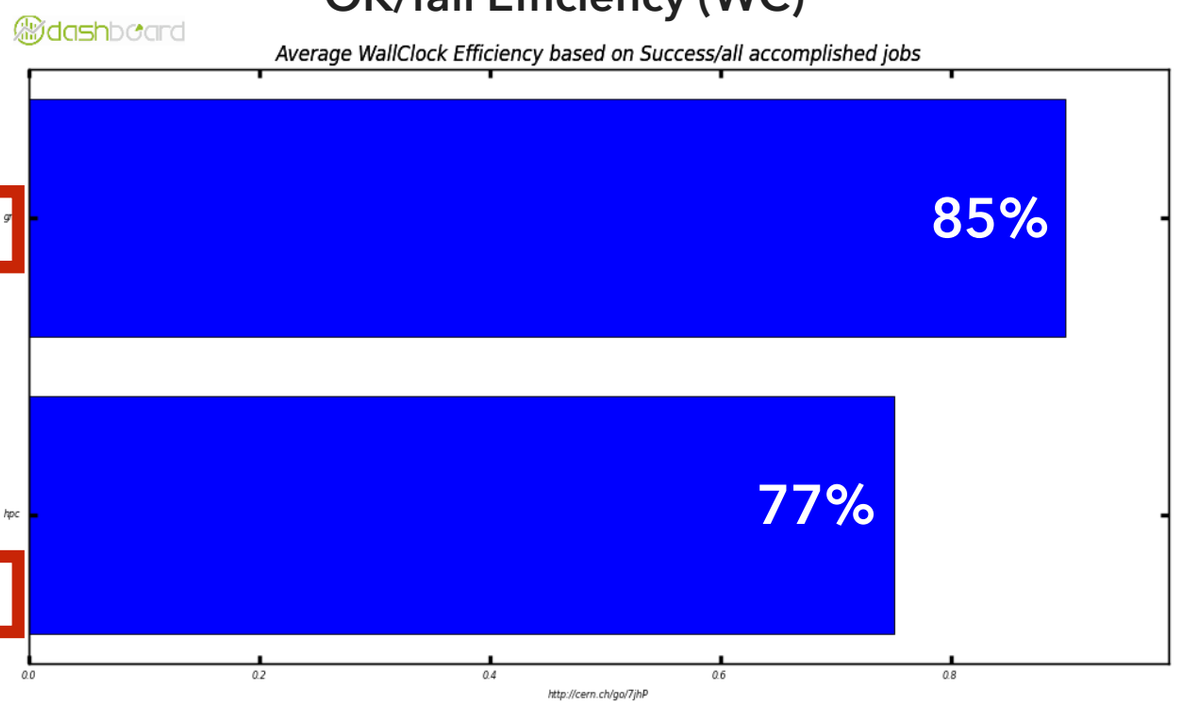




THE LHCONCRAY PROJECT

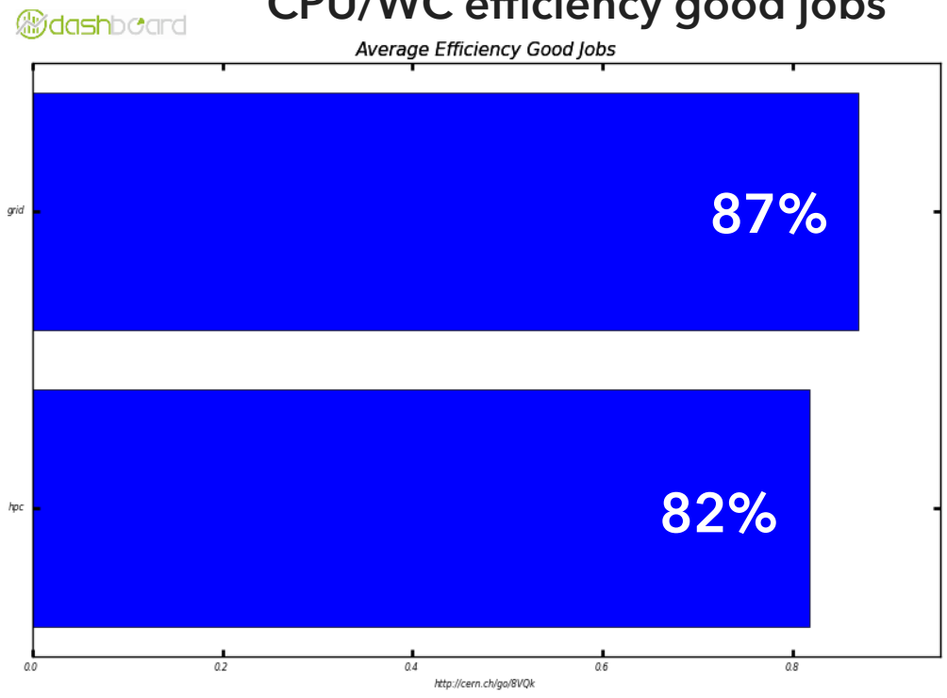
Piz Daint in production with ARC CE





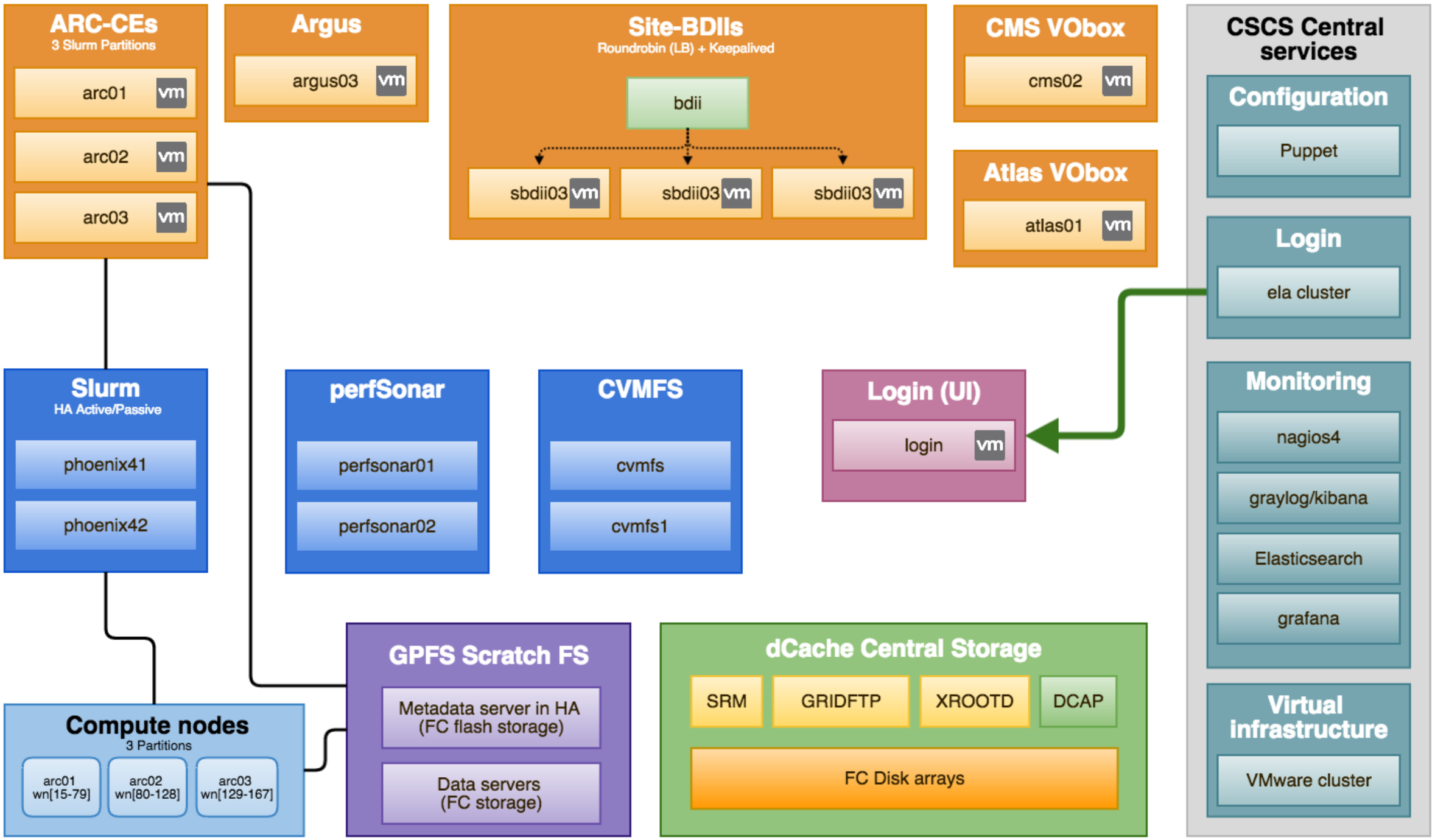
DEMO

http://ganglia.lcg.cscs.ch/ganglia/sltop_lhconcray.html



BACKUP

PHOENIX @ CSCS



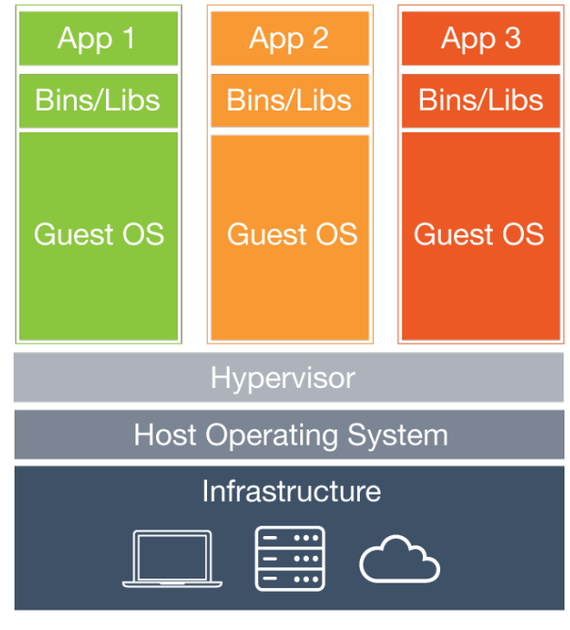
SHIFTER @ CSCS

Phase 2: Running unmodified applications with Shifter

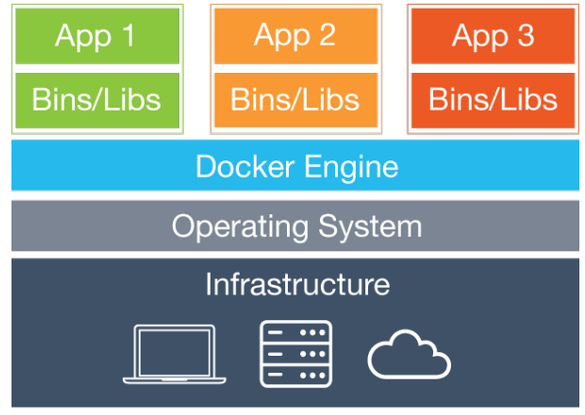
- Shifter basically
  1. Pulls an image to a shared location (/scratch)
  2. Creates a loop device with the image (=container)
  3. Creates a chrooted environment on the loop device
  4. Runs our application in chrooted environment

- A container in our context is basically an image with a full CentOS distribution and a chroot

```
[miguelgi@brisi01]-[02:46:29]-[~]:-)$ salloc -t 01:00:00 -n1 --image=docker:centos:6.7 -N1
salloc: Granted job allocation 82463

[miguelgi@brisi01]-[02:57:20]-[~]:-)$ srun --pty shifter /bin/bash

[miguelgi@nid00035]-[01:57:30]-[~]:-)$ uname -r
3.0.101-0.46.1_1.0502.8871-cray_ari_c

[miguelgi@nid00035]-[01:57:31]-[~]:-)$ cat /etc/redhat-release
CentOS release 6.7 (Final)
```



Virtual Machines



Containers

BENCHMARKING WITH HAMMERCLOUD

LHConCRAY - Performance with ATLAS HammerCloud stress tests

- **Reproducible tests** (24h runs)
  - single core and multicore
  - same input file for every job
  - same number of events for every job
  - running simultaneously on both clusters
  - mean WC times measured for each site are directly comparable

- **CPU-bound workload:**
  ATLAS detector simulation

MEASURE THE WC PERFORMANCE INCREASE OF THE CRAY RELATIVE TO PHOENIX:

$$\frac{(\text{WC PHOENIX}) - (\text{WC CRAY})}{(\text{WC PHOENIX})}$$

CRAY HEPSPEC RATING IS 16.8% BETTER THAN THE PHOENIX TIER-2 NODES