

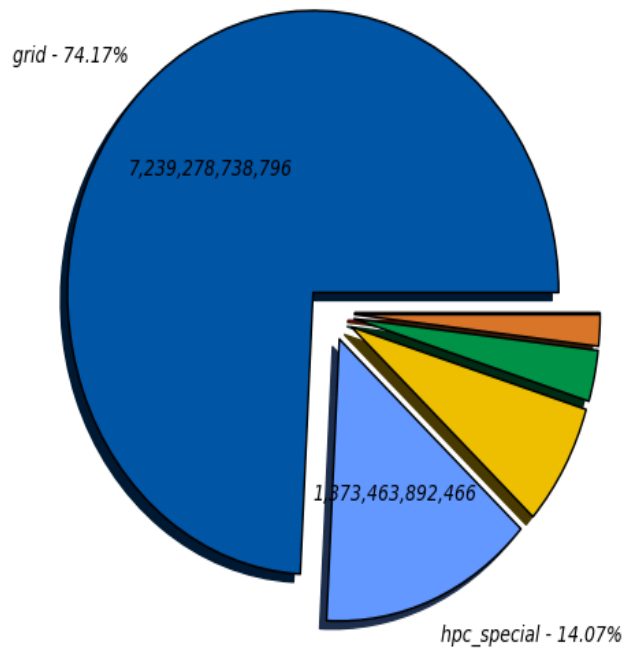
# **Experience running ATLAS applications in an HPC center: IDRIS of CNRS**

**14 Feb 2017 CC-IN2P3  
Vamvakopoulos E.**

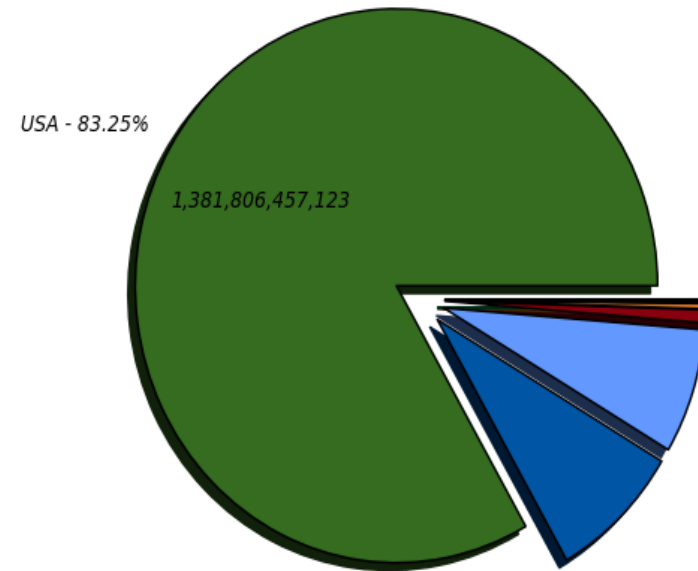
# ATLAS HPC STATUS



Wall Clock consumption All Jobs in seconds (Sum: 9,760,084,777,320)



Wall Clock consumption All Jobs in seconds (Sum: 1,659,887,471,512)



- grid - 74.17% (7,239,278,738,797)
- cloud - 7.02% (685,543,232,261)
- local - 1.80% (175,364,204,679)
- hpc\_special - 14.07% (1,373,463,892,467)
- hpc - 2.93% (286,423,579,045)
- None - 0.00% (11,130,071)
- USA - 83.25% (1,381,806,457,123)
- GERMANY - 7.29% (121,082,335,642)
- RUSSIA - 0.31% (5,161,738,393)
- CHINA - 0.06% (922,773,480)
- DENMARK, FINLAND, NORWAY, SWEDEN - 8.20% (136,180,792,538)
- SWITZERLAND - 0.75% (12,502,882,022)
- CZECH REPUBLIC - 0.11% (1,784,734,030)
- FRANCE - 0.03% (445,758,284)

Atlas have test, proposed and use different machinery on different HPC environment ~ 4 years

Significant contribution of HPC sites for 2017 → 17% (WallClock time)

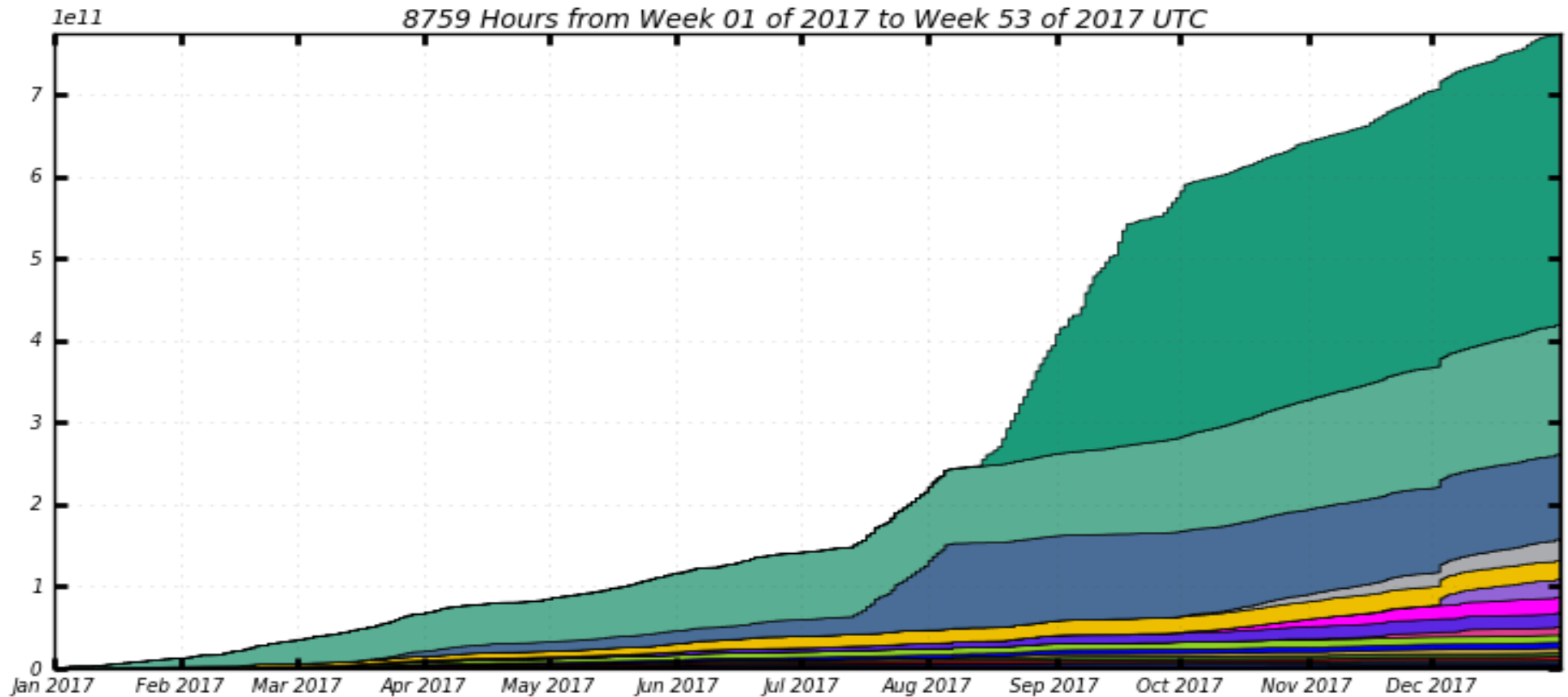
Contribution from US-HPC sites → 83.25%

# ATLAS HPC STATUS



CPU consumption Good Jobs in seconds

8759 Hours from Week 01 of 2017 to Week 53 of 2017 UTC










- NERSC\_Cori\_p2\_mcore (355,281,497,070)
- NERSC\_Edison\_2 (103,579,774,206)
- HPC2N\_MCORE (22,962,443,392)
- LRZ-LMU\_MUC\_MCORE1 (19,363,747,323)
- ALCF\_Theta (10,172,168,914)
- UIO\_MCORE\_LOPRI (6,998,438,257)
- UIO\_MCORE (4,599,243,235)
- NSC\_MCORE (2,858,094,718)
- CSCS-LCG2-HPC\_MCORE (1,492,874,247)
- Anselm\_MCORE (615,832,334)

- ORNL\_Titan\_MCORE (158,238,270,129)
- Titan\_long\_MCORE (25,889,608,518)
- NERSC\_Cori\_p2\_ES (21,463,618,412)
- MPPMU-DRACO\_MCORE (16,759,044,668)
- LRZ-LMU\_MUC1\_MCORE (8,877,872,638)
- NERSC\_Edison\_mcore (5,825,190,964)
- MPPMU-HYDRA\_MCORE (4,532,458,599)
- LRZ-LMU\_C2PAP\_MCORE (2,503,763,289)
- BNL\_KNL\_MCORE (819,618,850)
- ... plus 12 more

Total: 774,331,000,062 , Average Rate: 24,553 /s

# ***HPC environments***

## **General Challenges on HPC environments :**

-  **Total intergraded solution**
-  **Different CPU architectures (not always x86\_64 INTEL)**
-  **Execution of long-lived applications (i.e. services) is not allowed**
-  **Tight access rules (i.e. only via ssh)**
-  **Shared network filesystem (i.e. GPFS)**
-  **Lack of local disk device on the worker nodes**
-  **Outbound internet connections from the compute nodes are not permitted**

# HPC IDRIS of CNRS



At IDRIS, there are two machines currently in operation for intensive numerical computing:

✚ IBM BlueGene-Q with PowerPC-A2 CPU, ~90000 slots and 2GB per CPU (named Turing)

✚ IBM x86 machine based on Intel E5-4650@2.7GHz with ~10000 CPU slots and InfiniBand interconnection (named Ada) ~4GB per core, 4 way x 8 cores .

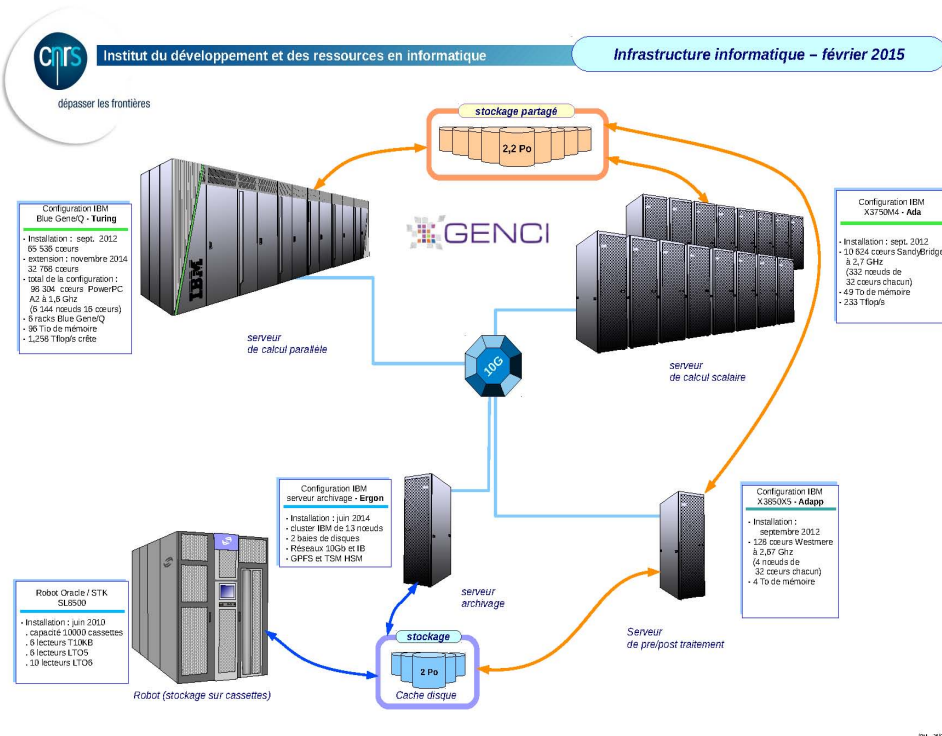
✚ IBM Load-Leveler 5.1 as batch system (two instances)

✚ Common local GPFS file-system

 User homes

 Working Directories

 Temporary space (only on ADA)



# ***CC-IN2P3 and IDRIS activity***



## **Objective**

 **Run Atlas EventGen and MonteCarlo (MC) production on IDRIS HPC environment**



**CPU jobs with low I/O**



## **Focus on ADA machine**

 **X86\_64 architecture**

 **RHEL 6.x OS**

 **IBM BlueGene-Q will be replaced soon ...**



## **Deploy and test ARC-CE integration solution**

 **University of Bern**

 **Nordu-Grid**

 **SuperMuc at LPZ**





 **Hydra at Max Planck Computing and Data Facility**

 **C2PAP at LRZ**

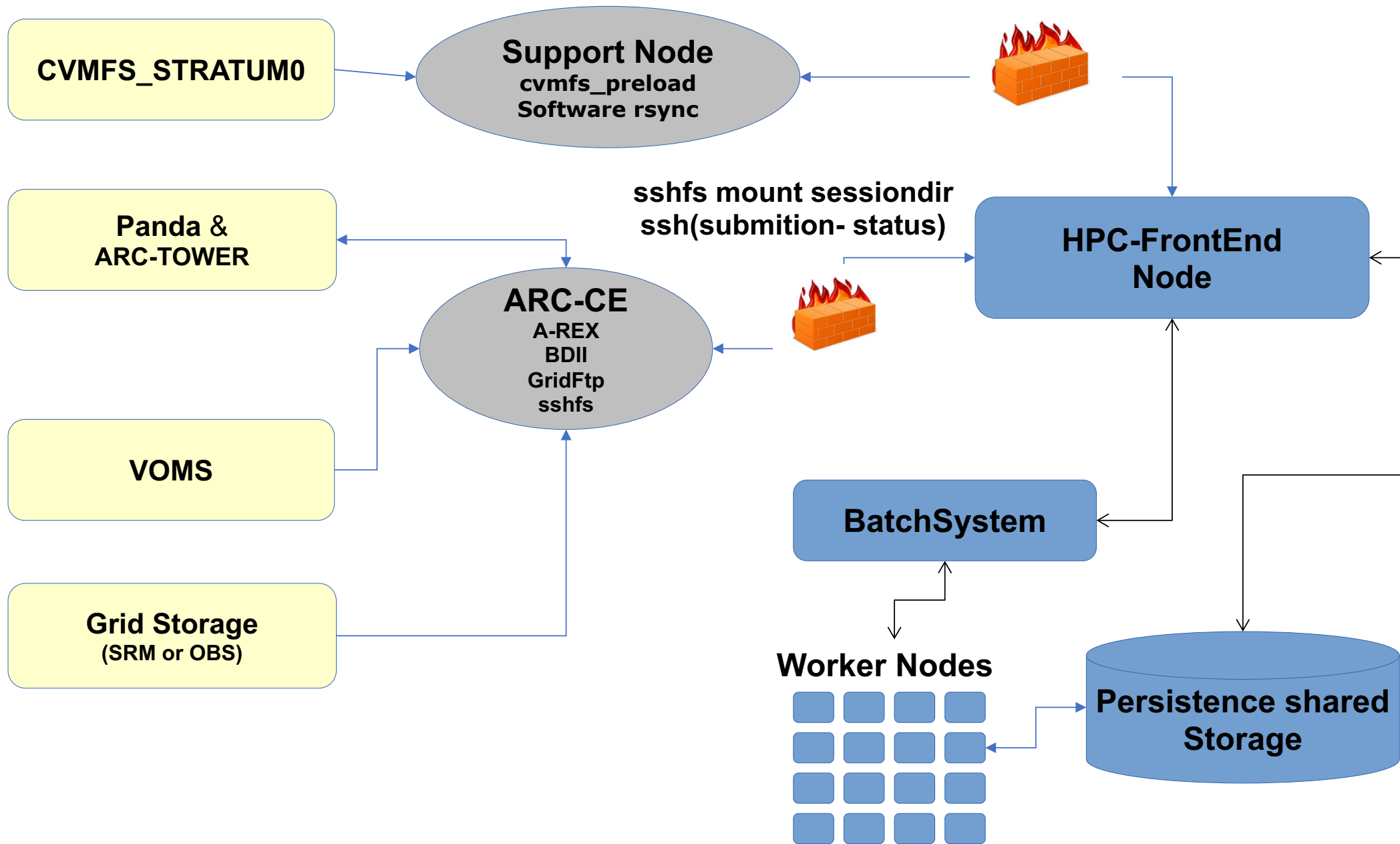
# ***ATLAS to HPC***

 **Current proposed solution based on ARC-CE computer element**

 **The solution is trying to address the following challenges:**

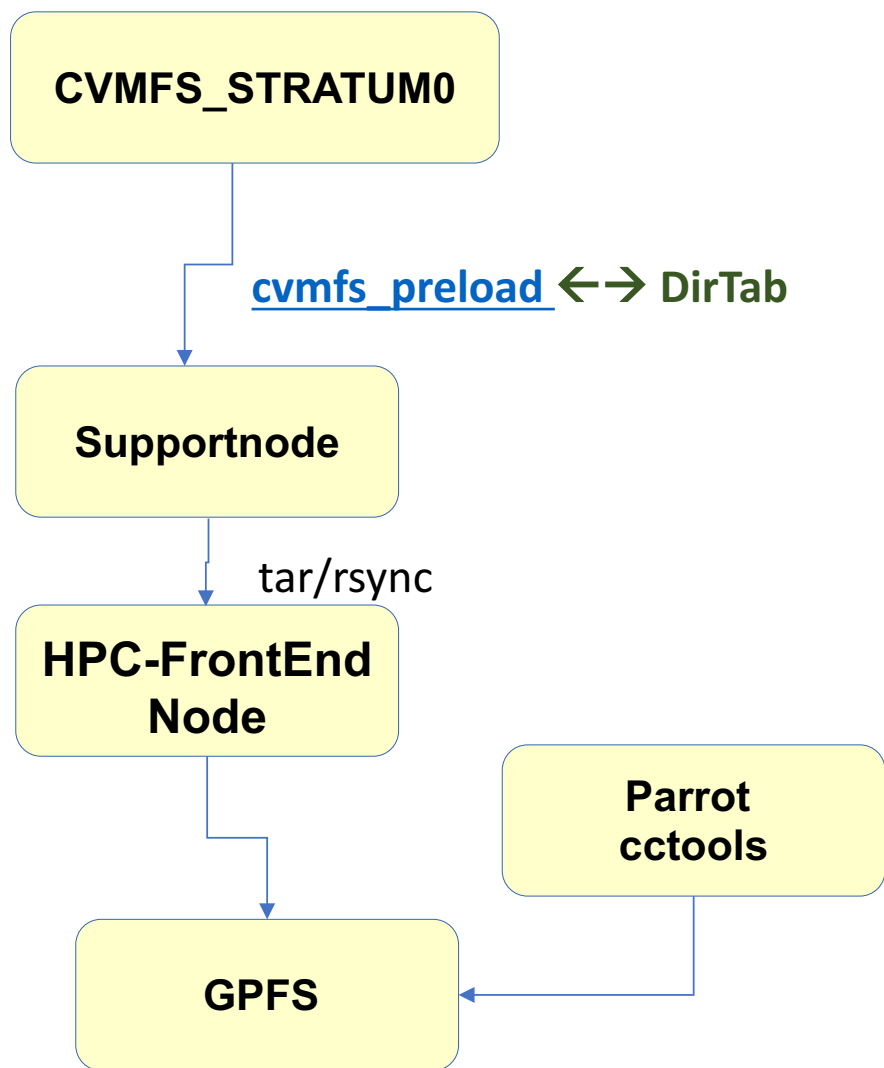
-  **Restrict access pattern (e.g. ssh/IP filtering, authentication)**
-  **Atlas Software delivery**
-  **Interface the ATLAS Work Load Management with HPC machines**
-  **Stage in and Stage out of the data**

# Context Diagram





# Software CVMFS preload



- 📎 The [cvmfs\\_preload utility](#) can be used to preload a CernVM-FS cache into the shared cluster file system
- 📎 Run cvmfs preload in order to populate the cache for first time 2.2 Tbytes, 1.8M files (**~17h**)
- 📎 Initial copy of cache with tar/ssh (**~48h**) → **too low ~12MB/sec !**
- 📎 Periodically run preload/rsync once per day (**~1h/2h**)
  - 📌 Agis Configuration
  - 📌 Software
  - 📌 Conditions data
- 📎 Use Parrot to read the cache
  - 📌 Not need to reallocate the software
  - 📌 Parrot is only for linux OSes

<http://cvmfs.readthedocs.io/en/stable/cpt-hpc.html>

# Software delivery alternatives on HPCs



## Containers: Shifter (NERSC's Cori) and Singularity (Titan, Theta)

- ✦ Fat containers all ATLAS offline release (~600GB)
- ✦ Thin containers single ATLAS offline release (~ 50 GB uncvmfs )
- ✦ Single release container rpm - size (rpm) ~ 12 GB

## Stratum-R (Stampede machines, Blue Waters at NCSA)

- ✦ Local replication of CVMFS repos from a local Stratum snapshot
- ✦ Remote synchronization (rsync) to the target Share FS
- ✦ [HPC in ATLAS, Doug Benjamin, CernVM Users Workshop Feb 2018](#)

## Native CVMFS + Workspace + Tiered Cache (CSCS/CRAY)

- ✦ CVMFS\_CACHE\_PLUGIN employment
  -  High level cache → RW cache in memory
  -  Low level cache → RO Preload-cache on share FS (xfs image)
- ✦ [Running native CVMFS on a Cray supercomputer, Miguel Gila et al., CernVM Users Workshop Feb 2018](#)

# sshFS



## ARC-CE Session Directory

📌 Job input and output data, job scripts, stdin/stderr files and status

📌 This is a filesystem client based on the SSH File Transfer Protocol.

📌 On the server side there's nothing to do.

📌 On the client side mounting the filesystem is as easy as logging into the server with ssh.

📌 Based on FUSE (the best userspace filesystem framework for Linux)

📌 Reconnect on failure

📌 Translate of UID/GID between client and server node

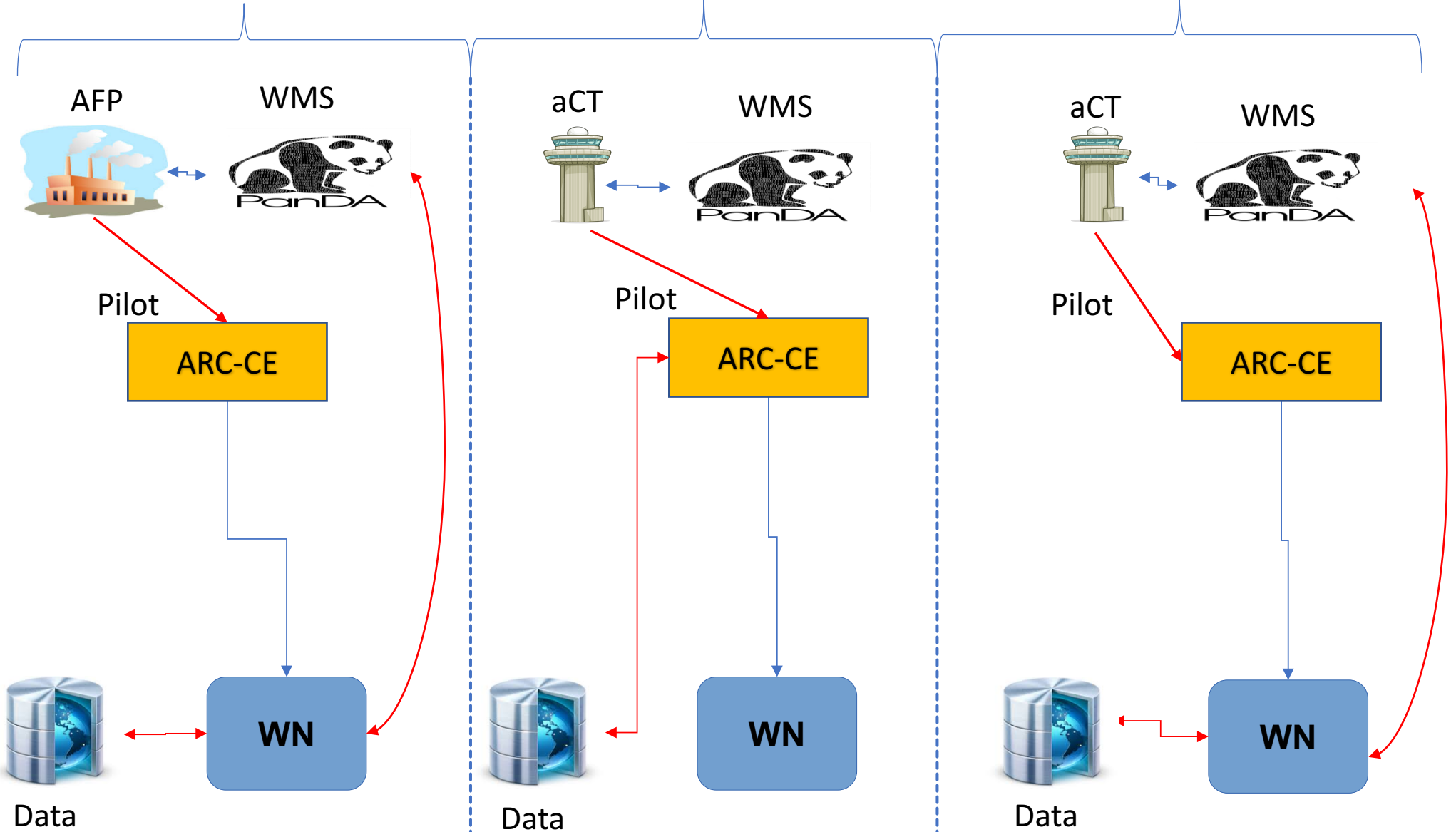
📌 With this component, We can share the ARC-CE session directory between ARC-CE node and shared file system on HPC machine

# aCT/ARC modes

APF

NorduGrid & HPC

aCT TruePilot

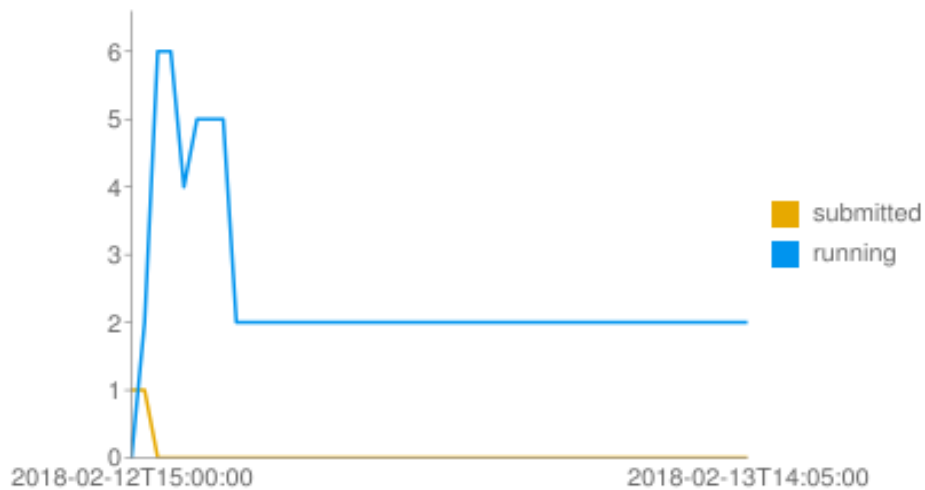


[ARC Control Tower and ARC CE, David Cameron Andrej Filipic, ADC TCB 25.4.16](#)

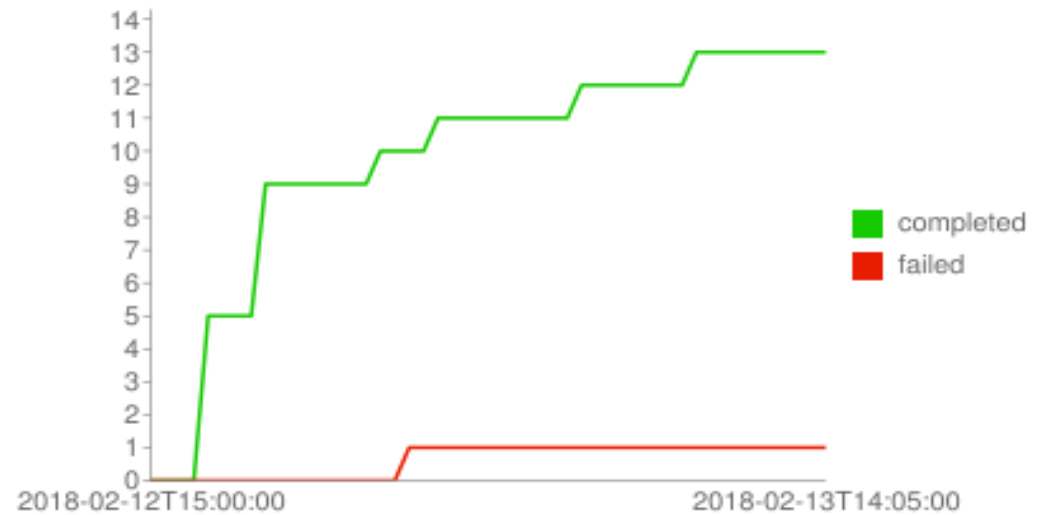
# Testing ....

## HammerCloud tests based on 914 template and preliminary MC jobs

Submitted / Running IN2P3-CC\_HPC\_DEBUG

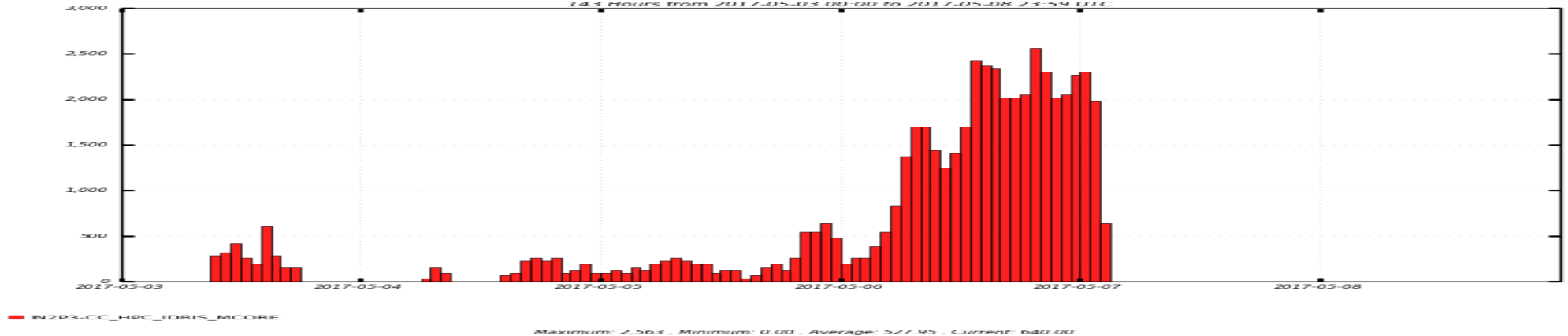


Completed / Failed IN2P3-CC\_HPC\_DEBUG



Slots of Running Jobs

143 Hours from 2017-05-03 00:00 to 2017-05-08 23:59 UTC



# Comments and Next Steps

## HPC integration into Grid is a custom-make Process

- 📌 Initial object (type of Jobs, pledge or not)
- 📌 Architecture of HPC environment
- 📌 floor space

## Software delivery mechanism is the most important aspect of our project

- 📌 Cvmfs\_preload cache and parrot looks straight forward solution
- 📌 Initial Population takes time / Need to re-check synchronization issues (?)
- 📌 We need a notify process for the new files on CVMFS stratum-0
- 📌 A native/adapted CVMFS solution on HPCs is welcome

## New queue for ATLAS

- 📌 Low priority with 6h wallclock limit (preempt queue)
  - 📌 **Run in whole node mode - 32 cores (non shared)**
- 📌 We are starting to setup EventService mode
  - 📌 **Fault tolerant mode**

## Tune ARC-CE timeouts

- 📌 Tune BDII publication and job status
- 📌 Should avoid frequent call of llq, llstatus and llclass command

## sshFS looks stable

- 📌 Performance should be verified under stress

# Acknowledgements

## **Atlas HPC team: solution based on ARCs**

 Rod Walker, Andrej Filipcic and David Cameron

## **Local teams of CC-IN2P3**

 Openstack and Network Teams

 Fabio Hernandez

## **Personnel of IDRIS**

 Agnes Ansari

 BatchMasters

**Thank you for your  
attention**

# ***Backup slides***



# Grid (WLCG) Site's Basic Services

## Virtual Organization

## Information System (BDII)

Membership Service (VOMS)    Computer Element (CE)



x.509

Job submission



Storage Element (SE)

User Interface (UI)

x.509

Batch System



x.509

Connectivity to Internet



CA authorities

Worker nodes farm  
Based on X86 64 arch

x.509



x.509



# Atlas Panda Framework

AGIS



Pilot Factories



WLCG/Grid Sites



Users



Panda Workload System



Rucio : Data Management system

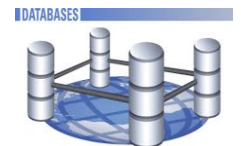


FTS3



CVMFS

Frontier



# Atlas Pilot

