- Sites
- Storage
- Networking
- Containerisation and EL7
- Miscellaneous

ND Cloud Report

Gianfranco Sciacca

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

ATLAS Sites Jamboree - 5 March 2018

ND CLOUD SITES

10 computing sites in 5 countries

Common denominator: ARC CE in push mode (sitemover mv)

- **NDGF-T1**
  - **DCSC** (Steno - Denmark)
  - **HPC2N** (Abisko - Sweden)
  - **NSC** (Triolith - Sweden)
  - **UIO** (Abel - Norway)

  - 69k cores, 3k for ATLAS (pledged)
  - *average running: 5.6k, peaks of 15k*
  - 9.3PB distributed dCache (incl. Slovenian pools)
  - ARC-CE cache at every site

- **Slovenian T2 + T3**
  - **ARNES**
  - **SiGNET** (2 clusters: NSC IJS, SiGNET)

  - 12.6k cores, 8k for ATLAS
  - 3.3PB in NDGF-T1 dCache pools
  - 1PB in ARC-CE caches

- **Swedish T2**
  - **SE-SNIC-T2 / LUNARC** (Aurora)

  - 0.4k cores for ATLAS with ARC cache

- **Swiss T2**
  - **UNIBE-LHEP** (3 clusters: 2*LHEP, UBELIX)

  - 8.8k cores, 2.5k for ATLAS
  - 0.6PB DPM
  - 0.3PB in ARC-CE caches

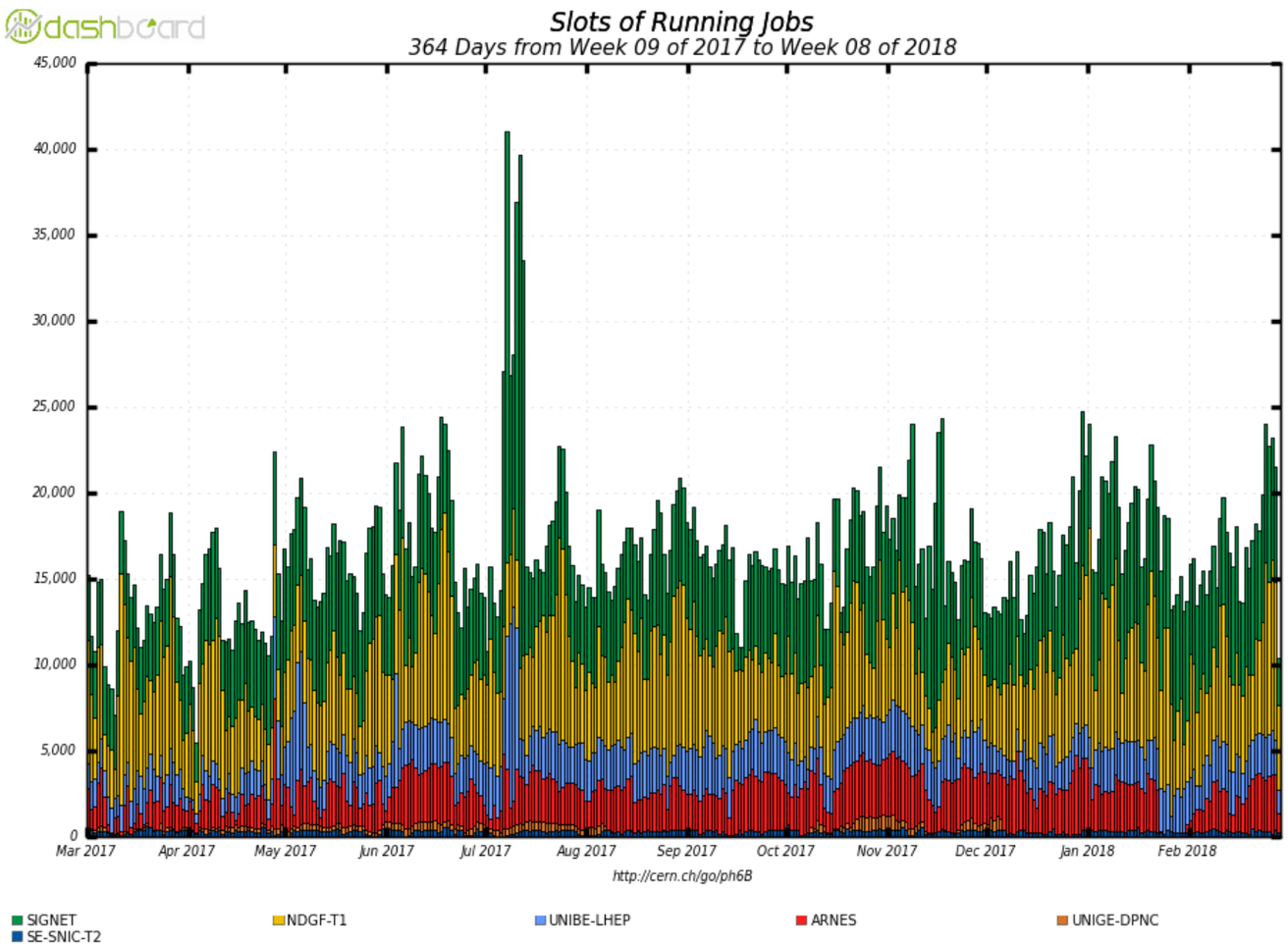
- **Swiss T3**
  - **UNIGE-DPNC** (Baobab, being re-commissioned)

  - 0.5k cores for ATLAS
  - 0.4PB DPM (to be merged to the UNIBE-LHEP SE)

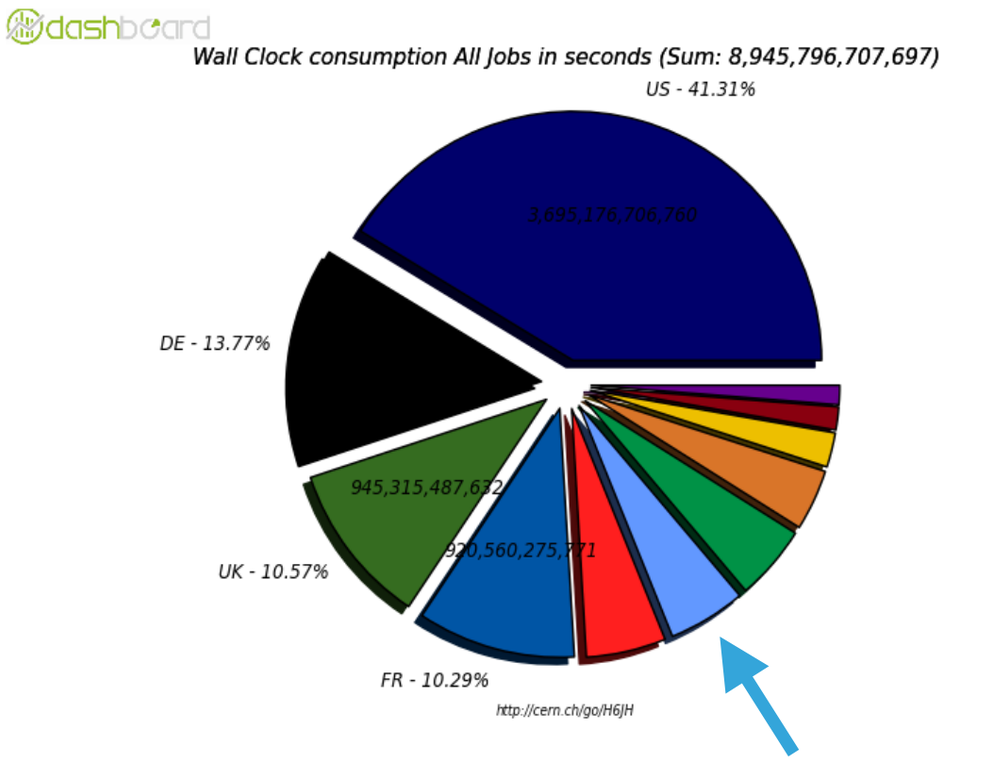
ND CLOUD SITES

10 computing sites in 5 countries

Average 16-17k running slots



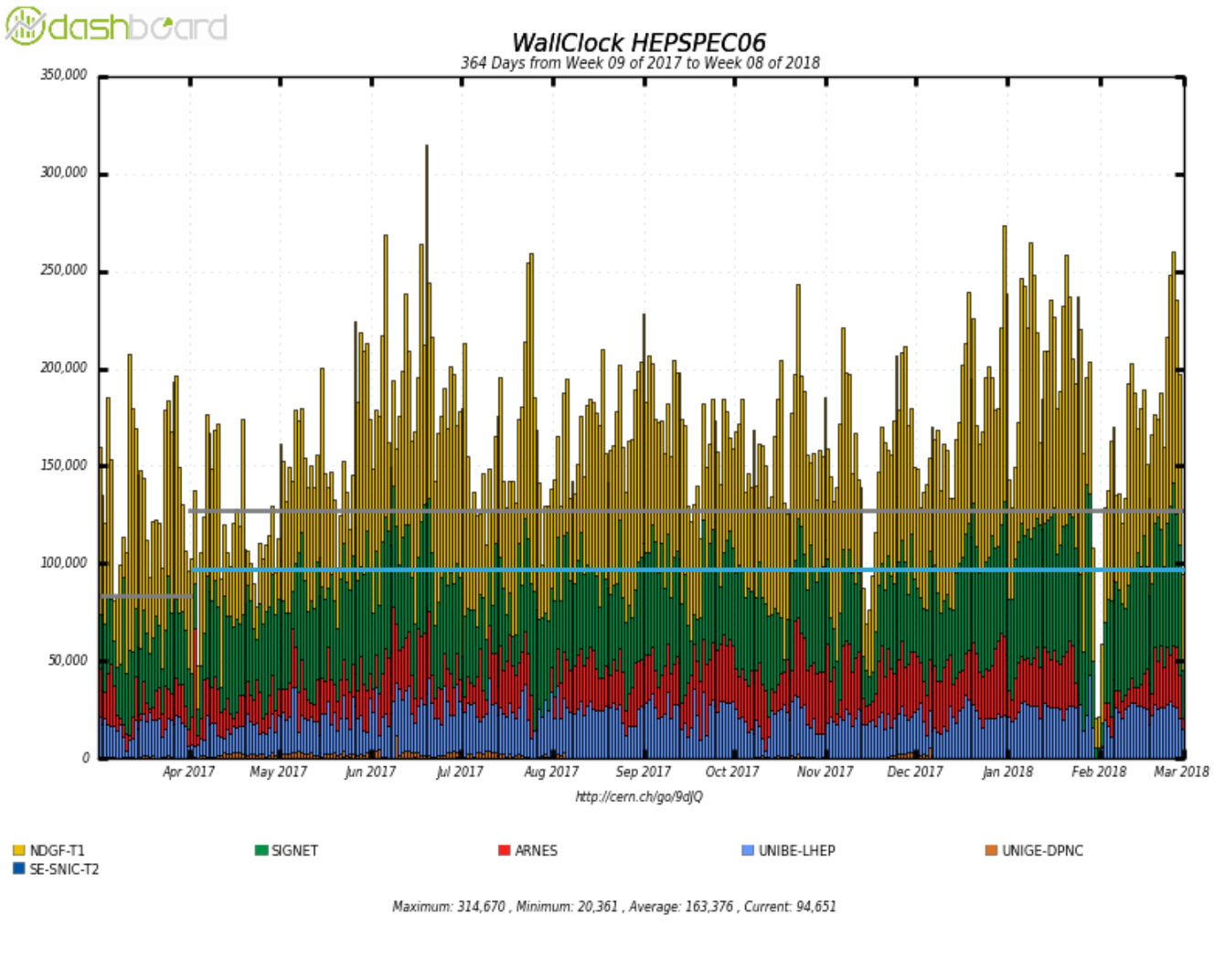
5% of ATLAS T1-T2s



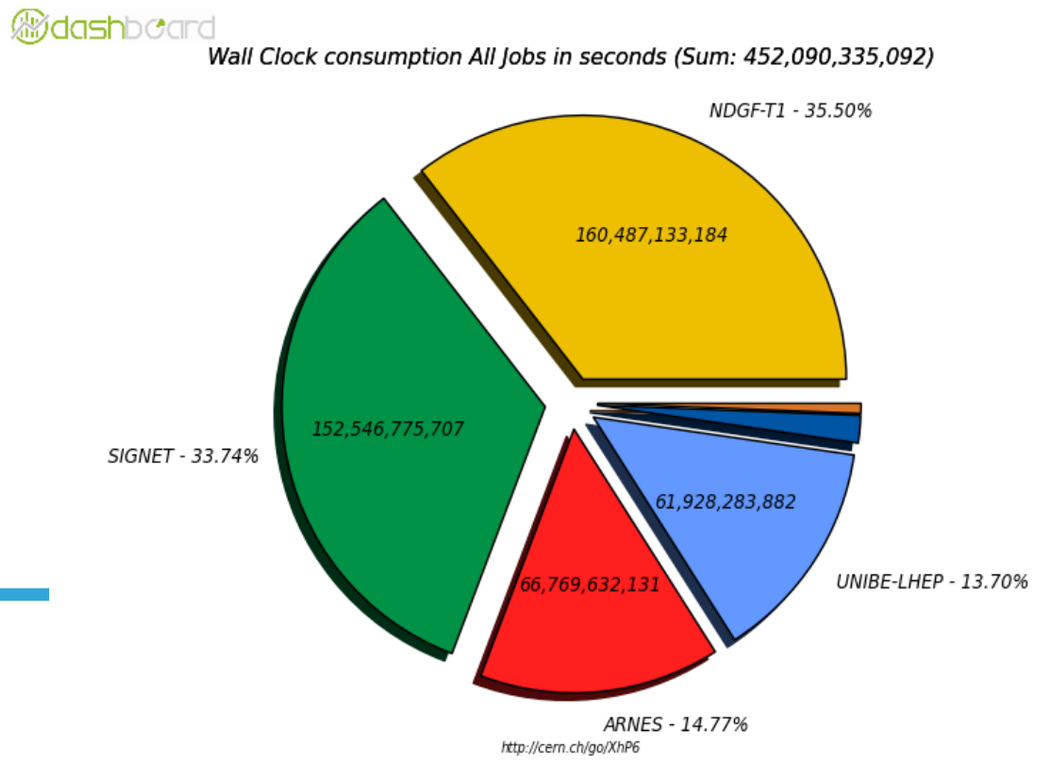
ND CLOUD SITES

10 computing sites in 5 countries

Combined pledge 96.3 kHS06



WC delivery per site



NDGF-T1 STORAGE

How do you provide reliable storage services

- **High Availability distributed dCache**
  - **Disk and tape pools are distributed among 7 sub-sites (CSC,HPC2N,IJS,NSC,UCPH,UiB,UiO) in the Nordic countries + Slovenia. 2nd DATADISK for size in ATLAS.**
  - **The central site (Ørestaden, UCPH) uses dCache load-balancing, HAProxy, UCARP, a replicated postgresql cluster with repmgr management of database failover, Ganeti to supply doors**
    - **This setup allows to carry out rolling upgrades to both pools and central services without impacting the storage service**
    - **Storage arrays protected at RAID6 level**
  - **Network-to-storage is generally not saturated. Some sub-sites use shared links, upgrades in the works (e.g. 100G for UCPH)**
- **Tape**
  - **Sub-sites are running IBM TSM robots. Using their own ENDIT interface: https://github.com/neicnordic/endit**
  - **Don't like to sometimes see recall requests trickling in rather than coming in big blocks**
- **No plans to move to different storage technologies (for the persistent storage)**
- **ARC caches**
  - **generally multiple NFS servers (~100+ spindles per site), lustre at HPC2N**
  - **UiO plans to move to CEPH, UCPH phasing out ZFS**

NDGF-T1 NETWORKING

How do you provide reliable networking infrastructure

- **Almost all sub-sites (DK is on the way) have dual networks with automatic failover**
  - **Internally at sub-sites network is handled by the local team**
  - **Between countries the connection is handled by Nordunet and their engineers**
  - **Between the sub-sites and the Nordunet infrastructure the connection is handled by the national provider**
  - **Nordunet is connected to LHCONE, but the traffic is not routed that way**
  - **No experience or plans for SDN**

- **Network capacity**
  - **Virtual 100 Gbps, 40 to sub-sites (still 10 in NO, undergoing upgrade) , overall utilisation about 50%**

- **perfSONAR**
  - **Is used to monitor the connection of the central services. Data used to debug networking issues**
  - **Work ongoing on installing perfSONAR at sub-sites to debug internal networking issues**

- **IPv6 transition**
  - **Storage is fully dual stacked. There are not firm plans to move compute nodes to IPv6**

TIER-2 STORAGE

How do you provide reliable storage services

- **Slovenian storage as dCache pools for NDGF-T1**
  - **3.3PB, RAID6 very reliable**
  - **few disk failures per year (~5 out of ~500 disks), standard raid recovery**
  - **occasional NIC issue, no data loss or corruption in ~10 years**
  - **will stay with dCache for now, to be seen in the future if CEPH can be used**

- **Swiss storage is DPM**
  - **0.6PB, RAID6 very reliable, strict local monitoring,**
  - **25 controllers, ~300 disks, negligible failure rate, recovered without data loss**
  - **2 controllers replaced in ~8 years**
  - ***Recent downtime due to outdated OS on the head node***
  - **The Tier-3 storage is in the process of being merged to the Tier-2 (localgroupdisk)**
  - **No immediate plans to move away from DPM (lightweight and reliable)**

TIER-2 STORAGE

How do you provide reliable storage services

- **ARC caches and scratch file systems**
  - **NFS4.1 at ARNES, NSC IJS, and LUNARC , typically multiple ARC data delivery servers (5 at ARNES)**
    - **One outage at NSC IJS last year, hardware related**
    - **planning to move to CEPH with CephFS**
  - **CHEP with CephFS at SiGNET**
    - **750 TB on 9 CEPH servers with 300 disks, 4 ARC data delivery servers**
    - **very resilient to disk failures (old hardware)**
    - **better performance vs e.g. lustre or parallel NFS**
  - **Lustre (x2) at UNIBE-LHEP**
    - **216 disks JBODS (6 servers) + 192 disks RAID10 arrays (4 servers) negligible disk failure rate**
    - **ARC-CE acts as data delivery server for each cluster**
    - **planning to evaluate CEPH with CephFS**
  - **GPFS at UNIBE-LHEP-UBELIX**
    - **large shared installation, very performant**

TIER-2 NETWORKING

How do you provide reliable networking infrastructure

- **The Slovenian network is handled by the NREN ARNES**
  - **30Gbs LHCONE link to Slovenia, 20Gbs to ARNES and 20Gbs to IJS (upgrades to 100Gbps in the planning)**
  - **Peak usage is ~15Gbps, LHCONE link to NDGF pools is often exhausted (there are some bottlenecks)**
  - **LAN and the connection to ARNES is managed by the institute networking team**

- **The Swiss network is managed by the NREN SWITCH**
  - **Redundant 100Gbps links to GEANT and to CERN (direct)**
  - **Not connected to LHCONE. The SWITCH team does not see a reason to run a separate (virtual) infrastructure for this.**
  - **Uni Bern is directly connected to the SWITCH 100G backbone and currently has an access of 40G**
  - **Uni Geneva is also directly connected to the SWITCH 100G backbone**
  - **Connection to SWITCH handled by the institute network teams, LAN by the local admins (2x10Gbps in Bern)**
  - **We are far from saturation (peaks at ~60%)**

- **The Swedish Tier-2 is connected to Nordunet**
  - **Network handled by the institute network team**

TIER-2 NETWORKING

How do you provide reliable networking infrastructure

- **perfSONAR**
  - **At IJS, published at the standard location**
  - **No perfSONAR in Bern. The SWITCH team is very responsive in case help is needed**

- **IPv6 transition**
  - **Done in Slovenia. Dual-stack, also IPv6 only wn's at NSC IJS**
  - **Planning dual-stack storage and CEs in Switzerland. No plans for the wn's yet**

- **No experience or plans for SDN**

CONTAINERISATION AND WN ENVIRONMENT

Docker and/or singularity, migration to CentOS 7

- **Singularity**
  - **Used at HPC2N to provide EL6 environment un Ubuntu nodes**
  - **UiO looking into it**
  - **Used at NSC IJS (Fedora), SiGNET (gentoo) and ARNES (CentOS 7)**
  - **Available for testing at UBELIX on CentOS 7 nodes**

- **Not using container orchestration**

- **Cloud provisioning**
  - **Used in the past virtual clusters on top of openstack (elasticluster) at UNIBE-LHEP**
  - **Currently being evaluated at UiO**

- **CentOS 7**
  - **Planning a migration at some of the NDGF-T1 sub sites**
  - **ARNES on CentOS 7**
  - **UBELIX is transitioning**
  - **No plans for UNIBE-LHEP and LUNARC at least until end 2018**

MISCELLANEOUS

Unified PanDA queues, monitoring

- **Unified PanDA queues**
  - **Done for NSC IJS**
  - **No obstacles, not too urgent for ARC CE in push mode**

- **Operations, Monitoring**
  - **Monitoring is quite advanced at all sites**
  - **Standard ingredients: nagios, ganglia, graphana**
  - **Operator on Duty shifts at the NDGF-T1**
    - **With special additional night, weekend and public holiday shifts**
  - **Weekly operation Slack meeting at the NDGF-T1**
  - **Live dashboards available at all sites, but private to operators and cloud support people**
  - **Critical alarms are broadcasted via SMS, Slack**
  - **Very active admin and support rooms in Slack, Skype**

QUESTIONS?

?