HEPiX Fall/Autumn 2018 Summary

Andrei Dumitru ▪ Arkadiy Shevrikuko ▪ Julien Leduc

https://indico.cern.ch/e/hepix-autumn2018

Organized by Port d’Informació Científica (PIC) at Casa Convalescència, Barcelona

HEPiX

CERN IHEP DESY INFN PIC NIKHEP RAL FZU BNL …

Future plans

Working groups

Challenges

Recent work

Status reports

Experiences

2x / year

AGLT2 KIT Purdue TRIUMF FNAL KEK GridPP NERSC …

https://www.hepix.org

Andrei

Autumn 2018 Meeting and General HEPiX News
Computing and Batch Services
Basic IT Services
Site Reports

Security and Networking
Storage and Filesystems
Miscellaneous

Arkadiy

Julien

Grid, Cloud and Virtualisation
IT Facilities and Business Continuity
End-User IT Services and Operating Systems

HEPiX 2018 Autumn in numbers

137 registered participants (record!)

105 from Europe
14 from North America
9 from Asia
9 from companies

HEPiX 2018 Autumn in numbers

69 contributions

Site Reports (16)
Networking & Security (11)
End-User IT Services & OS (7)
Storage & Filesystems (11)

Computing & Batch Services (9)
IT Facilities & Business Continuity (4)
Basic IT Services (4) Misc. (4)
Grids, Clouds & Virtualisation (3)

Network & Security

“Network configuration management at CERN: status and outlook”

- Evolving current solution
- Moving from perl to python
- Adding support for new vendors
- Usage of open-source platforms for new cfmgr

“Update about Wi-Fi service enhancement at CERN”

- Nearing completion of planned indoor coverage
  - 180 building activated, 20 buildings to go
  - 11,000 unique devices, 7,000 users per day
  - Current APs vendor - Aruba
- Plan for outdoor coverage (in Meyrin and Prevessin sites)

Network & Security

WLCG/OSG

- PerfSONAR 4.1 released in August perfSONAR
  - Drops support of SLS6, introduce Docker support, new web interface for configuration mechanism, new plugins introduced
- SAND project
  - Extract useful insights and metrics from perfSONAR collected data
- IRIS-HEP project
  - Algorithms for data reconstruction and triggering
  - Data organization, management and access systems for upcoming Exabyte era.

IPv6 & WLCG - update from the HEPiX IPv6 working group

- Usage of IPv6-only CPUs resources by end of Run 2
- Tier-1s have production storage accessible over IPv6
- Tier-2s around 38% is done
- ~30% FTP transfers over IPv6
- ~50% perfSONAR hosts reporting IPv6 enabled

Network & Security

Network Functions Virtualization Working Group Update

- Evaluate SDN/NFV solutions
- Helping sites to set up test environments and share knowledge between the sites (phase 1)
- Data analysis of test environments and prepare implementation and derive recommendations (phase 2)

CERN campus network upgrade

- Current status of campus and technical networks at CERN
  - Campus -> old equipment, increased instability
  - TN -> architecture is not tolerant to router failures
- New vendor for network equipment -> Juniper
- New equipment deployment on campus network and TN (during LS2)

Network & Security

Challenges for connecting Internet of Thing devices at CERN

- Security is a priority for IoT network deployment
- Technologies: LoRa (outdoor), ZigB (indoor)
- Defined set of requirements for IoT network:
    - Scalability, wide coverage, traffic separation (VRF lite), automated detection of “bad behavior”

IHEP Campus Network design based on SDN Technology

- Evaluation of solutions for SDN network from HUAWEI and Ruijie
- Equipment of the campus network replacement until 2020
- Decision will be made in 6 month based on evaluation results.

Network & Security

Computer Security Update

- Meltdown & Spectre
  - Updates are coming all the time
  - Microcode should be updated
- HP ILO authentication bypass
  - Fixed (affecting only HP iLO4(below 2.6) and iLO5(below 1.3)
- Quanta BMC SSH default credentials
  - SMASH have hardcoded default credentials
  - Isolate BMC, change default credentials, update firmware, disable unnecessary services
- Phishing, data leaks and everything you like is still there
- CERN developed a system for leaked credentials notification

Network & Security

Data Protection at CERN IT

- A GDPR group was created to determine what is private data in CERN and how to handle logs (we should have decent logs for troubleshooting)
- Privacy notice – document in understandable and transparent language, which explains data protection policy at CERN (currently in review)
- Established the max retention period for different categories of data.

A Framework for Open Science Cybersecurity Programs

- Talk about cybersecurity framework, which will aid to projects/facilities to set up protection against threads.
- Exists as a guide (document) with advices.
- Version 1.0 is planned for March 2019

Storage and Filesystems

Storage at CERN

- Types of storages at CERN: physics data, general and special purpose
- EOS – improvements on software stability and recovery time after failure (by splitting instances)
- CASTOR – still heavily used, but CTA development is ongoing
- CERNBox – backend improvements ongoing, turning into an application hub

CERN Tape Archive initial deployments

- New archive tool for EOS and tape storages, developed at CERN
- Aims to replace CASTOR
- Field test are ongoing right now, production foreseen for 2021

Storage and Filesystems

CASTOR

CERN Advanced STORage manager

Latest developments of the CERN Data Management Tools

- CTA integration for FTS
- DPM – storage system for Grid computing
  - Evolution to newer technologies (HTTP, REST, xrootd)
  - Improvements on QoS and performance
- XRootD – framework for remote access to data repositories
  - New release 4.9 with updates over client, server XrdHttp and Posix API
- EOS – disk storage system designed for physics analysis
  - Integration of XRootD as a native transport protocol
  - CTA integration and CI automation
- CERNBox – cloud synchronization and sharing service

DPM

Disk Pool Manager

Storage and Filesystems

The OSiRIS Project: A Multi-institutional Ceph Storage Infrastructure

- Pilot project for evaluating a software-defined storage infrastructure for Michigan research universities
- Integration of new science domains into OsiRIS
- dCache over Ceph experiment for ATLAS
- The main challenges are incorporating network orchestration into normal operations and improving users toolkit

FUJIFILM: Development of tape technology and challenges to overcome

- 3592 Tape Storage solution presentation
- Usage of StroniumFerrite (SrFe) – smaller particles with high magnetic output allows to create tapes with bigger capacity
- Robust tapes: thickness of shell is bigger than for LTO8
- Access to data is 50% faster than LTO8

Storage and Filesystems

LHC Long Shutdown 2 and database changes

- Migration of database hardware due to end-of-service
  - Replicate database (using Data Guard) and change DNS entries
  - Software patches (takes significantly less time due to the golden image concept)
- Enterprise solution for managing DB (from Oracle)
- Automation -> RUNDECK (job triggering using REST) idea is to get rid of routine tasks

Backup Infrastructure at CERN

- Overview of the backup infrastructures
- Recent update (increasing amount of scalable elements) results in significantly smaller amount of issues
- Rundeck for automation
- License limitations -> 15.9 PB (renewal in 2020)

Storage and Filesystems

RAID is dead

- Introduction of FlexiRemap algorithm (developed by Accelstor)
    - Splits data into 4kb write them sequentially over all SSD
    - Prevents from writing to damaged SSD, ensuring robust data protection
    - Optimized garbage collection (separates data into three tiers by write frequency)

LTO experiences at CERN

- LTO has a good value for price/TB, but reading speed almost 6x time slower than enterprise solution.
- One system already deployed in July 2018, some more to go during LS2
- Algorithm (inspired by 20 years old papers)
    - get physical location of the reading head
    - define costs for each hop between block I and j
    - Travelling salesman algorithm (minimal cost
- LTO good alternative, lots of room for improvement, but with this algorithm positioning time is improved by 3 times

Storage and Filesystems

Future-Looking Data Storage for Peak Performance in HPC Environments

- Tfinity ExaScale storage library presentation
  - 641 PB (with LTO-8), dual robotics for availability and performance
  - Zoning for optimized work of both robots
  - Ordered recalls based of LPOS instead of linear

Updates on ATLAS Data Carousel R&D

- Testing is ongoing:
  - To discover systems settings optimization problematic parts of the set-up
  - Check writing capabilities(for files up to 10GB)
  - Find a way to control bulk requests limits (Rucio)
- Almost all sites finished initial testing and "near production environment" tests are coming

End-User IT Services and Operating Systems

Service management at CERN: lessons learnt

- Bringing services on board
- User experience
- Tool configuration

CERN Linux service

- SLC6 (→11/2020), CC7(→04/2024), CC8?
- Moving Linux service to the cloud
- Koji/gitlab integration

Indico 2.x

- Major code rewrite: upgrade to 2.1!
- Future: new room booking, internationalization, paper reviewing

End-User IT Services and Operating Systems

Exploring the Alternatives...

- Deliver the same service to every CERN user
- Avoid vendor lock-in
- Keep hands on the data
- Address majority use cases

Evolution of CERN Web Services

- PaaS infractructure based on openshift (OKD)
- Simplifies, streamlines and consolidate web application deployment more efficiently

BNL Jupyter Based analysis portal

- DaaaS infrastructure well suited for interactive analysis
- HTCondor integration: from batch job submission to resulting analysis

RUST programming language

Tape BoF

Several site specific presentation

- focused on ATLAS Data Carousel results

No time left for discussion

Tape infrastructure evolution?

IT Facilities and Business Continuity

Technology watch WG

- Kick-off meeting report

HSF/WLCG cost and performance modeling WG

- Better understanding the current workloads, resource utilization and site costs
- Define a common framework for estimating resources
- Identify representative experiment workloads and evaluate impact of various parameters on performance
- HL-LHC workloads?
- Exploratory work following various technology trends (colder data, vectorization,...)

IT Facilities and Business Continuity

CERN procurement update

- Changes in the team
- Technology changes: more SSDs, accelerators coming
- Update on hardware repair workflow

NERSC Superfacility

- “A superfacility is two or more interconnected facilities using workflow and data management software such that the scientific output of the connected facilities is greater than it otherwise could be”
- Early stage of defining such a facility and API for various aspects

Grid, Cloud and Virtualisation

Extending local computing facilities using Helix Nebula Science Cloud

- Two commercial cloud providers remaining in the current Pilot phase: T-system (openstack) and RHEA (cloudstack)
- Three evaluated use cases: MAGIC, CTA and HTCondor
- **Conclusions available @CERN on 29 November 2018**

Good time with data using FaaS

- FaaS implementation based on the fn project https://github.com/fnproject/fn
- Nanoservices providing data API: data access backend can change client code is identical
- Example of PUE monitoring @NERSC (Elasticsearch backend)

Grid, Cloud and Virtualisation

A Data Lake Prototype for HL-LHC

- HL-LHC will be exceeding what funding agencies can provide by an order of magnitude.
- All the current data management concepts must be revisited to optimize cost
- Major technical and cultural changes to focus on QoS associated to various datasets
- Eulake federated storage PoC has been integrated and tested
- Several DOMA WG have been started on data ACCESS, DISTRIBUTION and STORAGE CLASS QoS

Site Reports

CSCS

- Decommissioning of Phoenix compute (LCG cluster, Swiss Tier2) by April 2019 in favour of HPC resources
- Tier-0 spillover tests on HPC
- Plan to activate IPv6

AGLT2 (ATLAS Great Lake Tier-2)

- NetFlow/sFlow monitoring via ELK stack and ElastiFlow
- Experimenting with SDN/NFV/OVS in their Tier-2 and as part of LHCONE point-to-point testbed

KEK

- KEK Central Computer System (KEKCC) providing stable service and computing resource for the Belle II and other experiments
- Renewal of KEK campus network completed in Sep. 2018

Site Reports

PIC

- Lots of microcode updates (done Variant 1, 2, 3a. Working on L1TF)
- Problem with a TK10C tape drive (roller damaged)
- IPv6: WNs (80%) and Storage in dual-stack (IPv6 only data transfer expected but still some issues to investigate)

INFN

- Anti-flooding system fully operational with telephone notification
- Structural works in order to reduce risk of flooding almost complete
- All R&D activities slowly recovering after flooding
- 75 “wet” tapes containing unique data sent to Oracle Lab for recovery
- 6 tapes partially unrecoverable (~20TB over a total of 630TB)

Site Reports

BNL/RACF/SDCC

- All compute nodes are SL7
- Tape storage ready to accept Belle-II data

SURFsara

- Very active on GRID activities for non WLCG communities
- WebDav security: upload & download through dCache WebDAV, to use TLS protocol (user/pass authentication)
- RcAuth - proxies without certificates: gives users a grid proxy after authentication in a portal with username/password (service created by NIKHEF)
- SURFdrive - data 100% in the Netherlands(ownCloud & Galera & Scality)

Site Reports

NDGF - Nordic Data Grid Facility

- Central dCache: proper pre-production setup; test before deployment with real user load
- Some “hardware” issues at NSC…

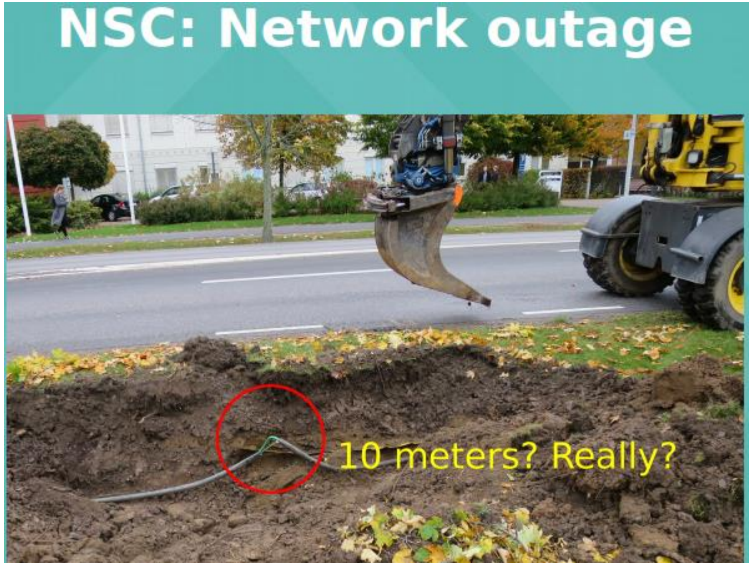
NSC: Network outage

NSC: UPS battery meltdown

Site Reports

NDGF - Nordic Data Grid Facility

- Central dCache: proper pre-production setup; test before deployment with real user load
- Some “hardware” issues at NSC…



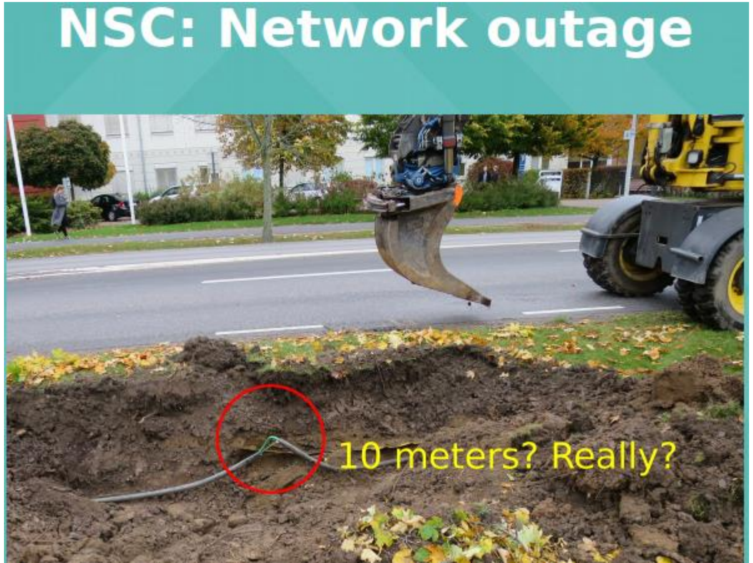
NSC: Network outage

NSC: UPS battery meltdown

Site Reports

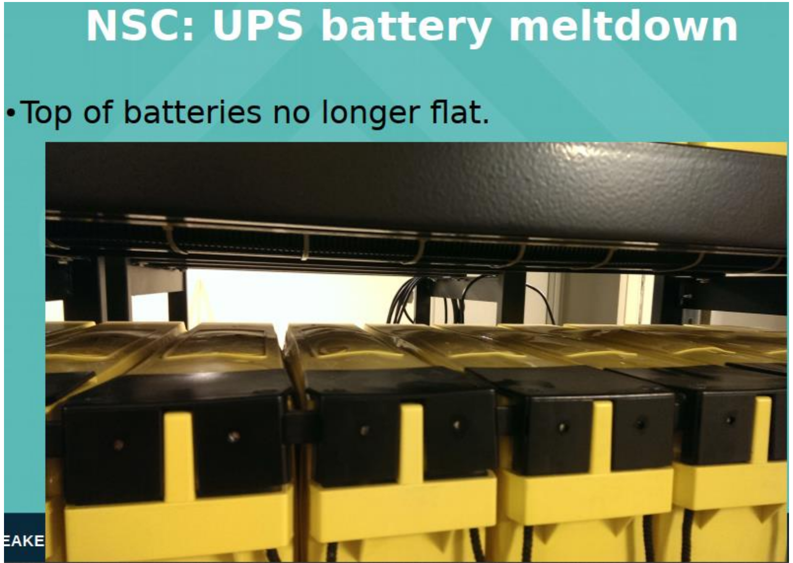
NDGF - Nordic Data Grid Facility

- Central dCache: proper pre-production setup; test before deployment with real user load
- Some “hardware” issues at NSC…



NSC: Network outage

10 meters? Really?



NSC: UPS battery meltdown

- Top of batteries no longer flat.

Site Reports

JLAB - Thomas Jefferson National Accelerator Facility

- Switching to Slurm from PBS/Torque/Maui
- Cloud Services as offsite computing resources (to cover bursts)

NERSC / PDSF (Parallel Distributed Systems Facility)

- Move from PDSF to SLURM completed
- Move to CORI (Cray XC40): CVMFS on the Cray

LAL + GRIF

- Currently using the datacentre phase 1 built in 2012/2013
- Expansion to new data centre slower then expected, mainly due to administrative problems
- SL5 finally over! CentOS 7 on most of the service machines (DB, web servers, etc.)
- SL6 is the dominating version currently for grid services and resources (upgrade planned)

Site Reports

Prague Site Report

- Computing Center of Institute of Physics of the Czech Academy of Sciences
- IPv6: most services dual stack
- HT-Condor configuration issues (fairshare, limits, draining)

Tokyo Tier-2

- Migration from SL6 to CentOS 7 ongoing
- Migrating to a new hardware at the end of this year

KISTI

- Moving towards virtual infrastructure based on containers
- Data Center Relocation (aging power supply and cooling system)
- KISTI Grid CA System redesigned: based on openXPKI project with HSM supported

Computing & Batch Services

Evaluation of AMD EPYC @BNL

- New line of x86_64 server CPUs from AMD (June 2017)
- New high performance series of server CPUs since 2012
- AMD EPYC vs Intel Skylake-SP: similar HEP/NP benchmark performance and pricing

Data analysis and reduction at ALBA synchrotron

- Remote data analysis using Virtual Desktop Infrastructure (Citrix)
- Windows based software: OPUS and Unscrambler
- Centralized logging with ELK stack

Report on the Workshop on Central Computing Support for Photon Sciences @BNL

- Discussed the specific issues, in particular those caused by the loose links between users and the photon facility
- Lack of sufficient forums to share experiences and collaborate
- Common themes at Light Source Facilities: real time data analysis, re-processing campaigns, data retention, etc.

Computing & Batch Services

Improving OpenMP scaling using openssl @NIKHEF

- Test OpenMP scalability without disabling Turbo Boost or Hyperthreading
- `turbostat openssl speed -evp bf-cbc`
- Same set of 'openssl speed' commands can be used to quickly determine the configuration and performance of an (unknown) CPU in userspace

Commissioning CERN Tier-0 reconstruction workloads on Piz Daint at CSCS

- Goal: implementation of an environment supporting ATLAS and CMS Tier-0 spill-over to Piz Daint (Cray XC40/XC50 supercomputer)
- Conclusion: LHC experiments can use a general purpose HPC system transparently for all their workflows (plenty of complexities to overcome)
- Integration efforts were costly, first time Tier-0 workloads go to an HPC system

Computing & Batch Services

Batch On EOS Extra Resources (BEER) and containers @ CERN

- BEER goal: use spare compute capacity on storage servers for the batch service
- Condor + Containers on EOS servers: run user jobs in containers
- Limits to protect EOS (memory, reserved CPU cores, controlled I/O scheduling, etc.)

Basic IT Services

Deployment of WLCG Compute Nodes @UCLouvain

- Cobbler – install OS, hardware config (disk partitions, IPMI, etc) and minimal config (SSH Keys and Salt Minion)
- Ansible - one-off operations (build config files, etc.)
- Salt stack - central configuration management server

Workflows automation with CERNMegabus

- CERNMegabus - service that provides for instant communication between services
- Improves scalability and reliability of systems management
- CERN Computer centre (CC) power cut management

Basic IT Services

Network Configuration Management Tool Evolution @ CERN

- Overview of “LanDB” changes in architecture and development procedures
- Homogenise technologies, Microservices, Business Logic library

The new CERN Authentication and Authorization

- Present: Kerberos, Single Sign-On and WLCG authentication and the authorization process described
- New authentication
  - provide uniform access schemes and user experience
  - token based, token conversion service, Sigle Sign-on everywhere
- New authorization
  - full federation support, map accounts to an identity, application specific roles

AAI - BoF Session

“Scheduled spontaneously” after the CERN AA talk

- More than 30 participants
- Multiple sites revamping their authentication system
- Keycloak (identity and access mgmt.) being investigated in several institutes
- Legacy applications require "token translation“
- Lots of needs - support both HEP and Photoscience communities
- Granularity required but not always easy to define given the divers users

HEPiX AAI Working group created

- Kick-off meeting on 9th of November https://indico.cern.ch/event/769924/

*Birds of a Feather (computing) = an informal discussion group | AAI = Authentication and Authorization Infrastructure

HEPiX

Spring 2019

25 - 29 March

SDSC SAN DIEGO SUPERCOMPUTER CENTER

Next meetings

SDSC
San Diego, USA
25 - 29 March 2019

SDSC SAN DIEGO SUPERCOMPUTER CENTER

NIKHEF
Amsterdam, The Netherlands
14 - 18 October 2019

Nikhef

HEPiX

Spring 2019

25 - 29 March

SDSC SAN DIEGO SUPERCOMPUTER CENTER