Summary from Track 7 - Computing Infrastructure

Henryk Giemza (NCBJ), Bruno Hoeft (KIT), Flavio Pisani (CERN), Christoph Wissing (DESY)

Track Statistics

- 42 oral presentations scheduled + 17 posters
  - 2 rather short term oral cancellations
  - typically between 40 and 60 people in the audience
  - Active discussions after almost all talks

Monitoring + DataCenter - 1

- Robin Hofsaess - A Lightweight Analysis & Grid Facility for the DARWIN Experiment
  - Joint project to design/create collaboration platform for new experiments.
  - Easy to deploy and use. Based on standard solutions.
  - Single entry point for job submission (same for storage). Token based authentication.
  - Integration with batch systems via TARDIS.

- Dr Marcus Ebert - The BaBar Long Term Data Preservation and Computing Infrastructure
  - Summary contains hints for retiring experiments.
  - one crucial issue --> keep the community as knowledgespace

Computing in a New(-ish) Collaboration

Situation

- Typically **no central IT services** available yet
- **No dedicated** infrastructure or IT personnel
- Often **no joint computing platform**, but self-made solutions for different sub-groups

Requirements

- **Main computing needs** in R&D phase: simulations and framework development
- Accessible for **all members**
- **Lightweight** setup with simple deployment
- Good **scalability** according to the computing need

Our Solution: A Lightweight Analysis and Grid Facility

A **unified**, **future-proof** concept for **users** and **central production** needs of an experiment

Redundancy/Reliability

login machine VM NIS Server VM interactive VM XRootD redirector VM --- hardware raid1 OS ZFS mirror data disks	XRootD proxy server --- hardware raid1 for OS ZFS raidz3 data disks	Web documentation VM Wiki VM Hypernews (HN) VM --- hardware raid1 for OS ZFS raidz3 for data disks	4 NFS server: NFS $HOME NFS job output NFS framework NFS documentation --- all use: ZFS raidz2/3 hardware raid1 for OS
• spare server setup the same way • ZFS send/receive	• multiple servers available • just cache, loose no data	• web content on NFS • HN content on NFS • images backed up • daily mysql dump to NFS	• spare server setup in the same way • ZFS send/receive • extra backup of framework and documentation

Monitoring + DataCenter - 2

➔ Xoan Carlos Cosmed Peralejo –

Monitoring particle accelerators with wireless IoT

- Allows to monitor wide space in a cost-effective way.
- Scales up to 1000 devices per gateway (tested).
- No changes to the lora standard

- **Industrial use cases:**
  - Temperature and humidity
  - Access control
  - Assets and vehicles tracki
  - Cranes usage monitoring
- **Underground uses cases:**
  - Measure radiation levels
    - Installation and cost reduced drastically
    - Number of monitored areas can grow
  - Tunnel displacement and cracks monitoring
    - Allows us to avoid manual campaigns
  - Vibration and electromagnetic field monitoring

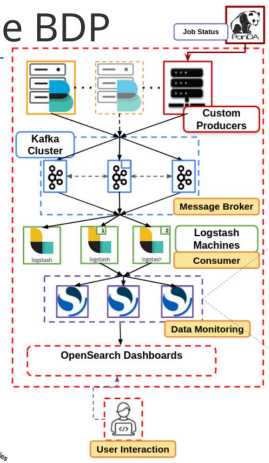
➔ Aksieniia Shtimmerman –

The Big Data Processing Infrastructure for monitoring and analysing the ATLAS experiment processing activities at INFN-CNAF Tier-1

- implementing BDP - joblogs analysis - for identifying issue and improving efficiency

INFN CNAF Structure of the BDP

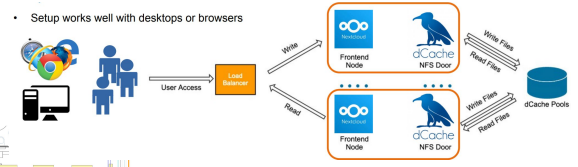
ATLAS Job addressed to Tier-1 machines in Bologna are selected with a PanDA API, a summarizing log report is ingested in the data indexing pipeline in a compact JSON format with a custom producer.



➔ Christian Voss –

Monitoring large-scale dCache installations with storage events using Kafka streams

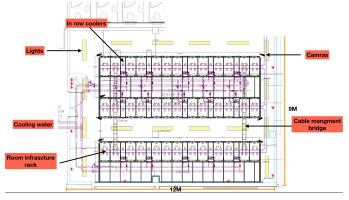
- All RAW data produced onsite go to dCache.
- 3 step monitoring: hardware, linux services, dCache services
- Trigger system: e.g. automatic restart of pools when something goes wrong



➔ Daniel Peter Traynor –

A Successful Data Centre Refurbishment Project

- additional rack space + deeper racks
- more power (20kW)
- more efficiency - heat recycling → use for heating building



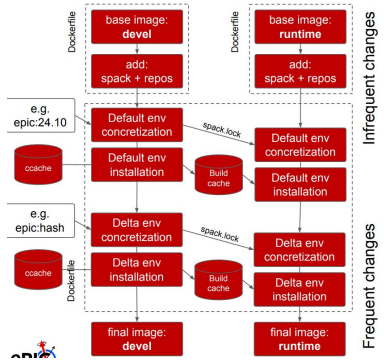
Kubernetes/Cloud - 1

➔ Wouter Deconinck

Cache Rules – Everything Around Me

- Building ePIC Containers With Spack
- speck for container build - speedup with ccache
- reuse build caches
- use github as code repository

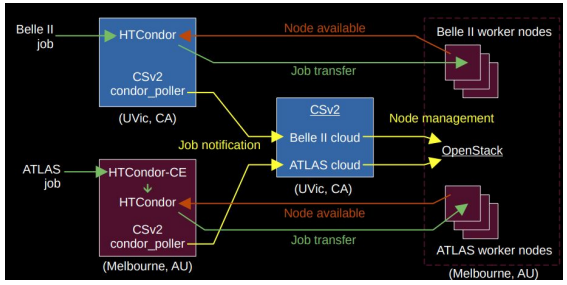
it was shown how to build a full-stack container and that it is possible for integrating full-system testing, validation, and benchmarking on every commit, but it requires judicious use of caching strategies



➔ Jonathan Mark Woithe –

An implementation of cloud-based grid CE and SE for ATLAS and Belle II

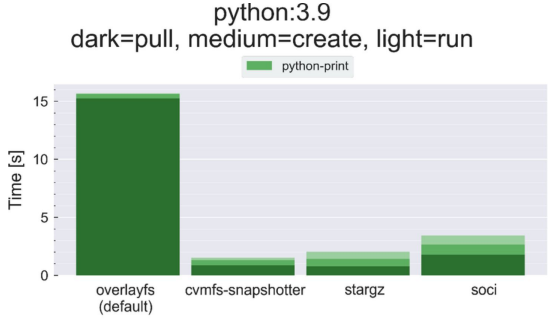
- melbourne research cloude offers (MRC) VMs, orchestrated by OpenStack, with 750TB S3 object store
- use resources for both Atlas and Belle II



➔ Clemens Lange –

Efficient and fast container execution using image snapshotters

- read only as much as you need (lazy-pulling of container images)
- up to 10 times faster (as reading and decompressing the full image)
- cvmfs snapshotter is fast

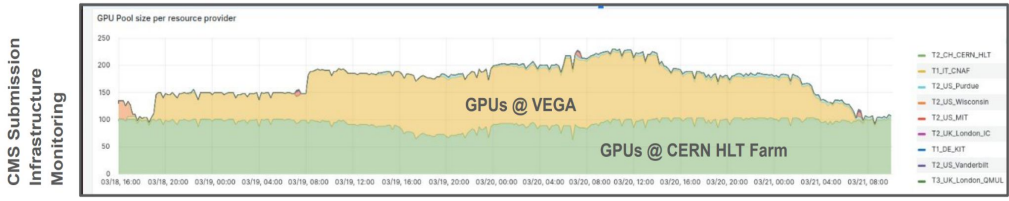


Kubernetes/Cloud - 2

➔ Daniele Spiga

Exploiting GPU Resources at VEGA for CMS Software Validation

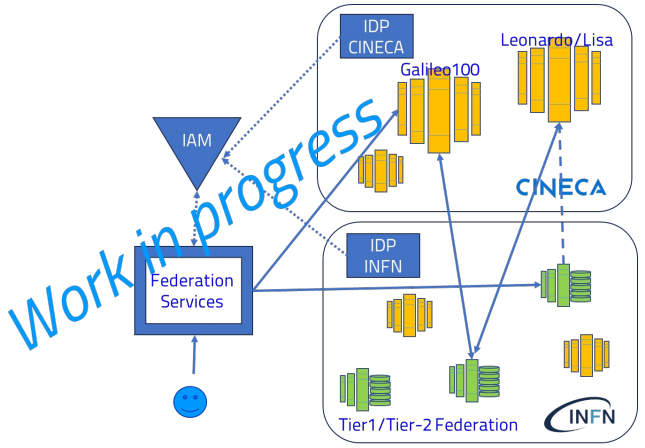
- use HPC for WLCG --> CMS Grant at VEGA in Slovenia
- transnational Site Extension (CNAF - Bologna to VEGA)
- use GPU at VEGA transparent for the user as if the submitted job would complete run at CNAF Bologna



➔ Claudio Grandic

The evolution of INFN's Cloud Platform: improvements in Orchestration and User Experience

- data cloud evolving into cloude federation (ease of use through PaaS orchestrator and dashboard)
- offer through INFN, CINECA, GARR the cloud federration
- Grid user will get access to HPC Leonardo and Galileo resources



➔ Matthias Richter

Provisioning of Grid computing resources in the Norwegian Research and Education Cloud

- nordic e-infrastructur collaboration (neic) tier-1
- wn harware - middelware at VM - storage (distributed dCache cluster) network connected

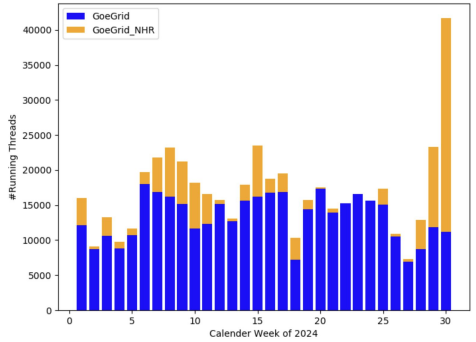
- Compute workers and dCache pool nodes are running on dedicated hardware
- All other service instances, like VO Box, monitoring instances running on common hardware
- Connection to Nordic LHC network, all instances have IPv6

Leverage HPC resources - 1

➔ Uday Saidev Polisetty

Integration of the Goettingen HPC cluster Emmy to the WLCG Tier-2 centre GoeGrid and performance tests

- Consolidation of WLCG infrastructure in Germany. Focus on the ATLAS site in Goettingen.
- CPU resources are to provided from the HPC site Emmy (hosted in Goettingen).
- CPU_eff on Emmy slightly worse compared to GoeGrid dedicated grid site.

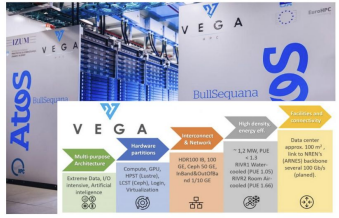


➔ Diego Ciangottini

Unlocking the compute continuum: scaling out from cloud to HPC and HTC resources

- Matching heterogeneous payloads to heterogenous resources.
- Heavily builds on Kubernetes.
- Implementation on two HPC sites, VEGA and systems in Jülich.

Our first case studies



HPC Vega is the first EuroHPC JU supercomputer hosted at the Institute of Information Science in Maribor, in Slovenia.

First volunteer HPC provider, enabling super early prototyping

The Jülich Supercomputing Centre operates one of the most powerful supercomputers in Europe, JUWELS, and JUNIQ the first European infrastructure for quantum computing. UNICORE offers seamless access to the Supercomputers.

First volunteer for an external plugin based on UNICORE

➔ Muhammad Imran

Development of machine-learning based app for anomaly detection in CMSWEB

- formular for Live Anomaly Detection

$$T_{\text{adaptive}} = \left(1 - e^{-\lambda \cdot t_{\text{diff}}}\right) \cdot \left(w \cdot T_{\text{prev}} + (1 - w) \cdot T_{\text{recent}}\right) + e^{-\lambda \cdot t_{\text{diff}}} \cdot T_{\text{recent}}$$

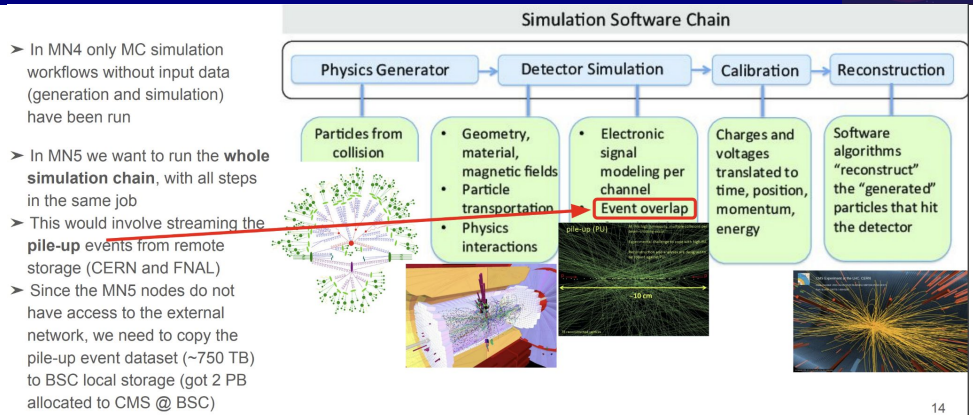
$$w = \frac{\sigma_{\text{prev}}}{\sigma_{\text{prev}} + \sigma_{\text{recent}}}$$

leverage HPC resources - 2

➔ Jose Hernandez

Commissioning and exploitation of the MareNostrum5 cluster at the Barcelona Supercomputing Center for CMS computing

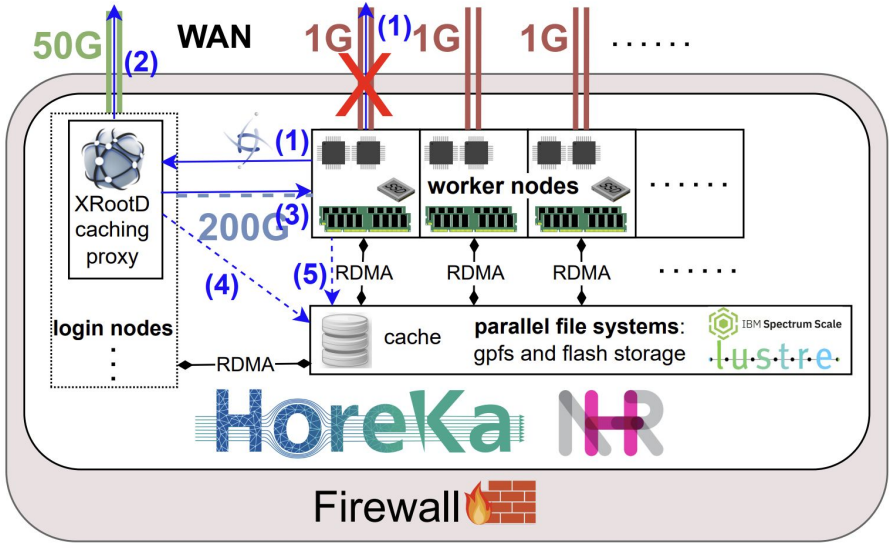
- BSC HPC very restrictive regarding network connecetions. A number of workarounds needed to be implemented by Spanish CMS community.
- Pileup needs to be copied to BSC, it is quite large, about 650TB. Import quite time consuming,.



➔ Robin Hofsaess

First Deployment of XCache for Workflow and Efficiency Optimizations on Opportunistic HPC Resources in Germany

- HEP jobs on HoreKa HPC show higher failure rate and worse CPU efficiency.
- Working on a caching/buffering strategy employing an Xcache/proxy on one of the login nodes.

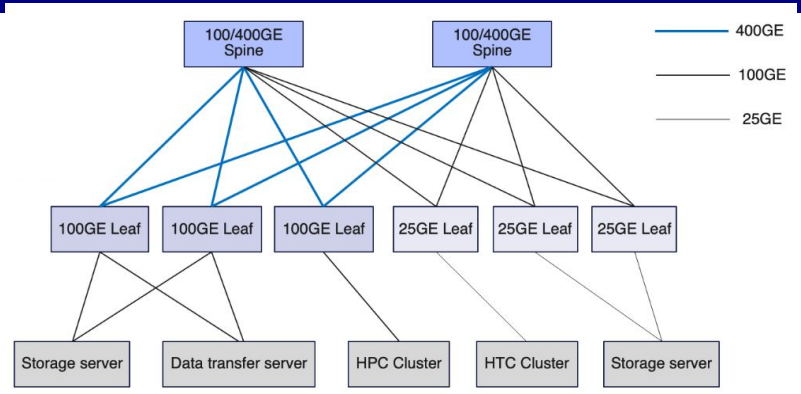


CPU Architecture + RDMA/ROcE - 1

➔ Shan Zeng

A RoCE-based network framework for science workloads in HEPS data center

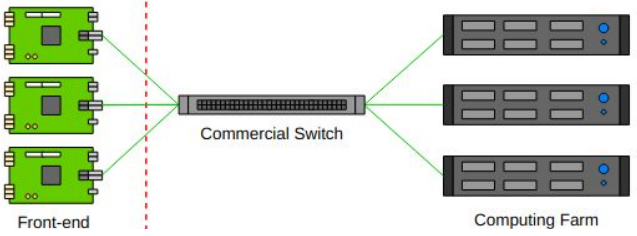
- New deployment of an RDMA capable Ethernet network
- Spine-Leaf topology for high throughput low latency interconnection



➔ Gabriele Bortolato

Front-End RDMA Over Converged Ethernet, lightweight RoCE endpoint

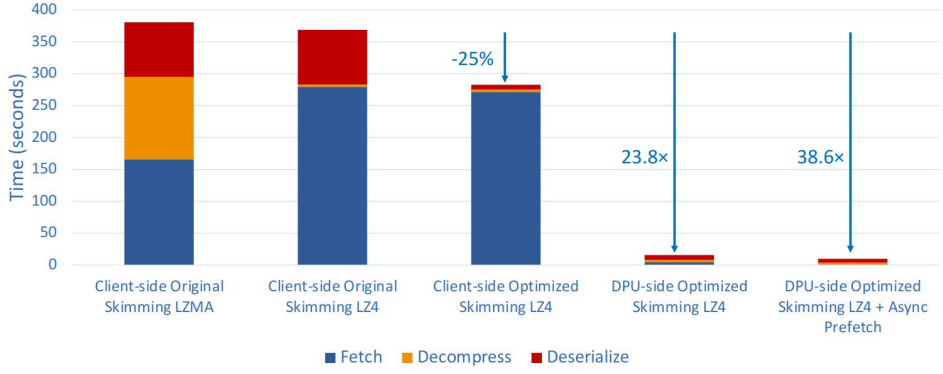
- RnD on RDMA capable Front End electronics
- Reduce the cost of DAQ by using COTS components



➔ Philip Chang

Near-Data Compu.ng Model for Accelera.ng LHC Data Filtering

- Using DPUs to accelerate the data skimming process
- Performing data skimming of the server side has a significant impact on the file transfer time



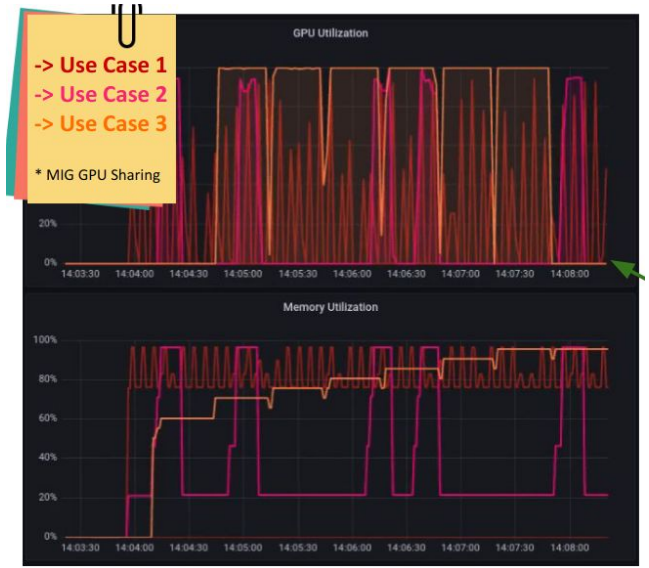
CPU Architecture + RDMA/ROcE - 2

➔ Johannes Elmsheuser

Using the ATLAS experiment software on heterogeneous resources

- The use of heterogeneous resources will help to address the HL-LHC computing resources challenge
- ARM is fully supported for MC production
- Extensive RnD to port workflows on GPU
- ATLAS job distribution system PANDA ready to target GPU resources
- variety of CUDA versions and kernel modules lead to challenging environment
- further coordination with other experiments & WLCG needed

Name	nCPU	HepScore23	HepScore23 per nCPU
HepScore23 reference	64	1018	15.9
Grace Hopper	72	2319	32.2
Apple M2 Air	8	141.4	17.7
Ampere Neoverse-N1	20	349.4	17.5
Intel Xeon E5-2683 v4	16	258.5	16.2



➔ Diana Gaponcic

Improving overall GPU sharing and usage efficiency with Kubernetes

- Exclusive GPU allocation can be inefficient for certain tasks
- Time-slicing and GPU partitioning provides better resource utilization
- reduce idle time by GPU sharing via kubernetes
- Various ways to configure the sharing via helm charts

➔ Luis Guilherme Neri Ferreira

The Glance project common infrastructure dependencies upgrade from the ATLAS Glance perspective

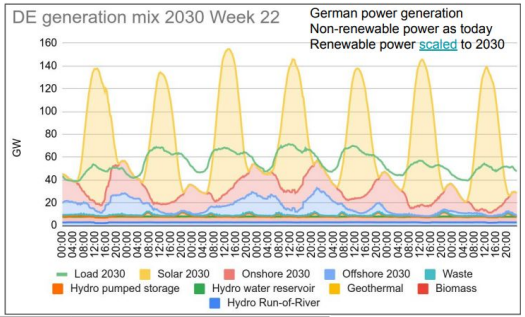
- Successfully update GLANCE to RHEL 9

carbon footprint (or efficiency) - 1

➔ Zach Marshall –

Carbon, Power, and Sustainability in ATLAS Computing

- Improving software will reduce the resources at the next pledge request.
- Reduce waste (unsused simulations, ecc) job inefficiency, job failure.
- Different sites useing different strategies, we need a full study of the problem.
- Job scheduling and CPU frequency scaling based on power availabily



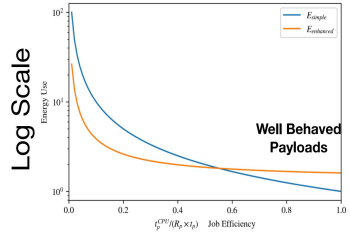
➔ Alex Owen

Allocating Carbon Costs to Computing Payloads across Heterogeneous Infrastructures.

- Mathematical model for carbon mapping for CPU power based on idle consumption and CPU job time (see slides)
- Similar model for storage based on data access time

	Scope 2 - Energy	Scope 3 - Carbon
Payload	$E_p = P_f^{idle} \cdot \frac{R_p}{R_f} \cdot t_p + P_{slot}^{CPU} \cdot t_p^{CPU}$ Where: $P_{slot}^{CPU} = \frac{E_f^t - P_f^{idle} \cdot t}{t_f^{CPU}}$	$C_{ep} = \frac{R_p}{R_f} \cdot t_p \cdot Q_{ef}$ Where: $Q_{ef} = \sum_{x=1}^{items} \frac{C_{ex}}{T_x}$
Idle	$E_{idle}^t = E_f^t - \sum_{p=1}^{payloads} Ep$	$C_{e\,idle}^t = t \cdot Q_{ef} - \left(\sum_{p=1}^{payloads} C_{ep} \right)$

Table 3: Summary of the Enhanced Payload Model showing allocations of Scope 2 energy and Scope 3 carbon to user payloads and the remaining idle allocation to the provider



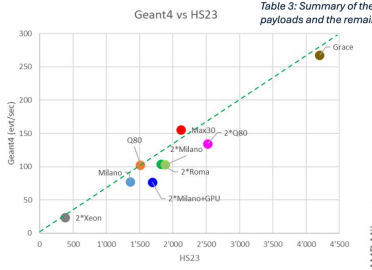
➔ Emanuele Simili –

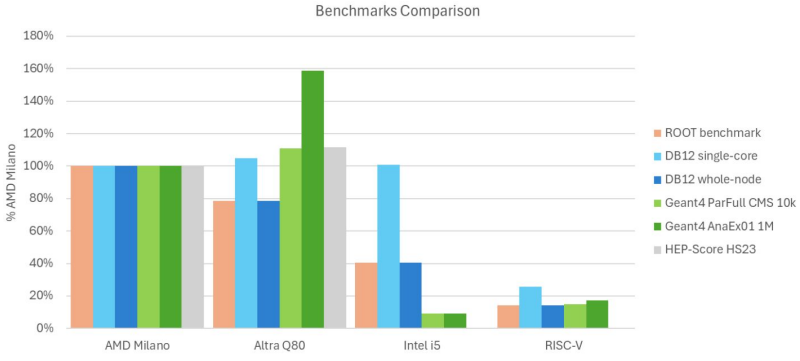
Heterogeneous Computing and Power Efficiency in HEP

- tests of several x86 and ARM systems
- Metrics: events/sec/core and events/watt
- ARM comparable to x86

Taking on RISC-V for Energy-Efficient Computing in HEP

- first look at RISK-V architecture
- slow performance per core, but # of cores increasing
- Packages like ROOT, GEANT4 and XrootD compiled from source





carbon footprint (or efficiency) - 2

➔ Daniele Lattanzio –

Heterogeneous computing at INFN-T1

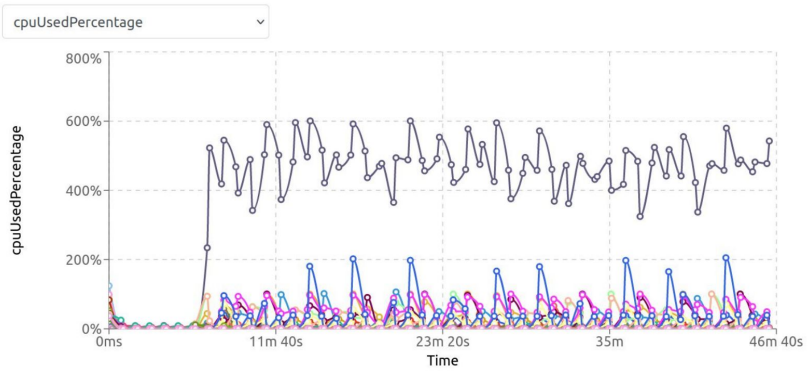
- ARM in production for ATLAS
- ARM validation ongoing for ALICE, LHCb and CMS
- First experiences with RISC-V with compatible finding (s. above)

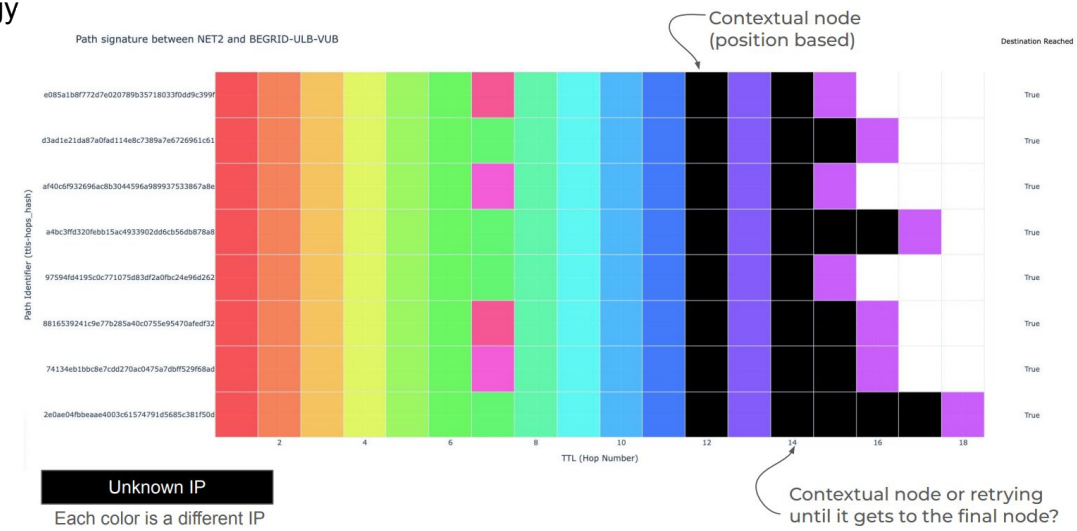
➔ Petya Vasileva –

Enhancing Network Analytics through Machine Learning

- Metrics traceroutes and throughput tests
- Topology dataset has holes due to routers not responding and private address
- First attempt Probabilistic approach to recostruct the topology
- secodn attempt try to use ML to interpolate the holes

- Data reconstruction





Data Center

➔ Max Dupuis

Prévessin Data Centre Powers Up

- Demands of HL-LHC require new data center at CERN, to be built in the Prevessin site
- Very modern design regarding cooling and efficiency
- PDC mostly for computing, Meryin DC will be mainly for storage

➔ Marcus Ebert

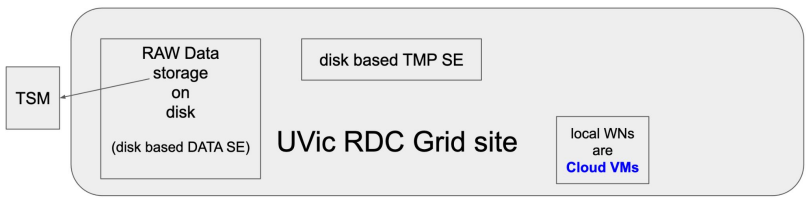
Canadian Belle-II Tier1 Raw Data Centre

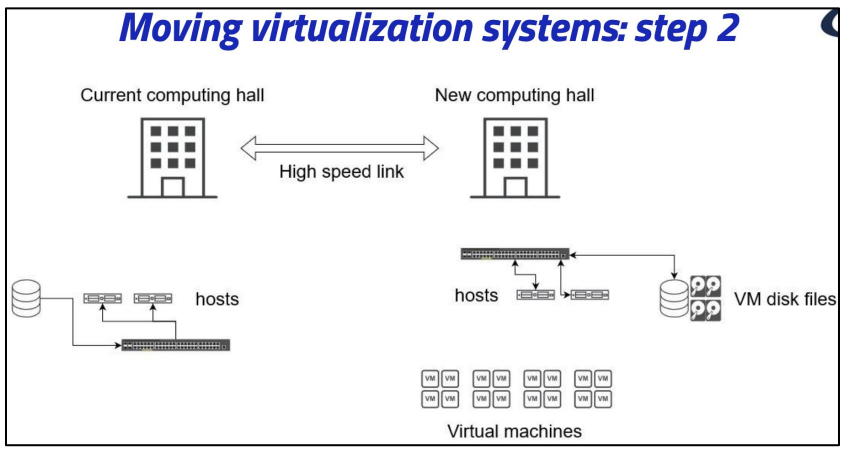
- Storage looks like disk only, but still with TSM backup
- HTConder WNs provisioned on local cloud infrastructure

➔ Daniele Lattanzio

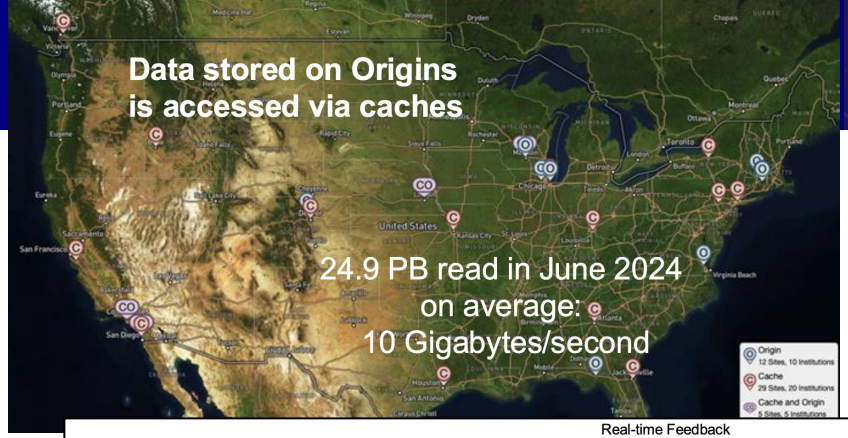
Moving a data center keeping availability at the top

- New data center building, hosting CINECA (incl. Leonardo HPC) and CNAF resources and services.
- To be moved services needed to set up in redundant mode first.
- Network setup is crucial.
- Whole enterprise is a matter of several months





Distributed Infrastructure



➔ Fabio Andrijauskas

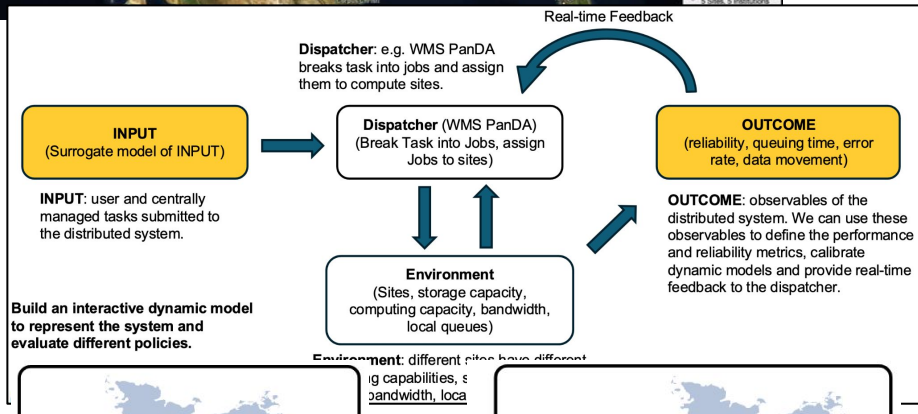
Benchmarking OSDF services to develop XrootD best practices

- OSG data federation
- Mostly build/deployed with Kubernetes
- Various setups being tested

➔ David Park

Towards an Introspective Dynamic Model of Globally Distributed Computing Infrastructures

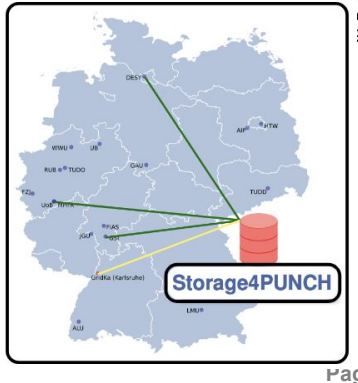
- Gathered PanDA matrics get evaluated
- ML methods trained on gathered metrics should improve "dispatcher" component.

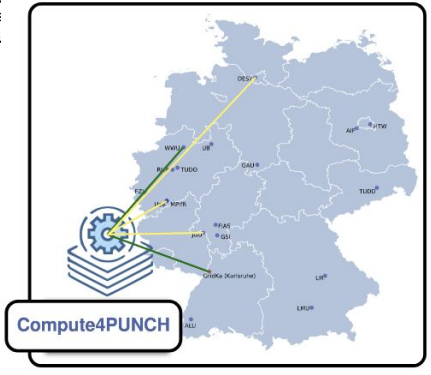


➔ Benoit Roland

Latest developments of the PUNCH4NFDI compute and storage infrastructures

- Federating compute and storage resources for astronomy, hadron and HEP community.
- All resource access handled via tokens, incl. automatic token prolongation via MyToken.
- REANA used as building block



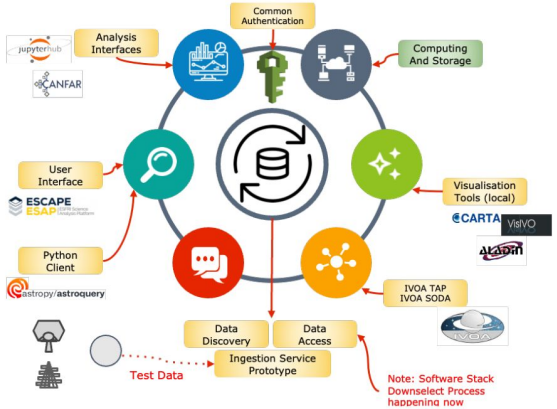


IPv6 + SKA + network advancements

➔ James William Walder –

Revolutionising Radio Astronomy: The UK's Role in SKA's SRCNet Deployment

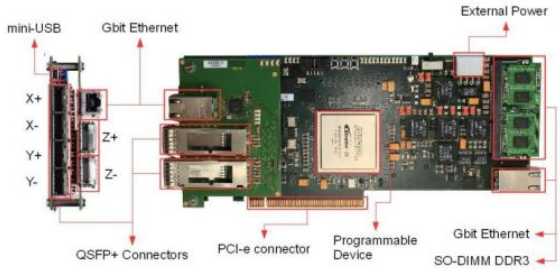
- SRCNet is a gateway for users to access SKA data
- UK Deployment will share some services with the WLCG Tier-1 (RAL)
- Will use Azimuth to create scientific platforms
- Uses perfSONAR to measure latency and bandwidth between different sites



➔ Carlotta Chiarini –

Hardware and software design of APEnetX: a custom FPGA-based NIC for scientific computing

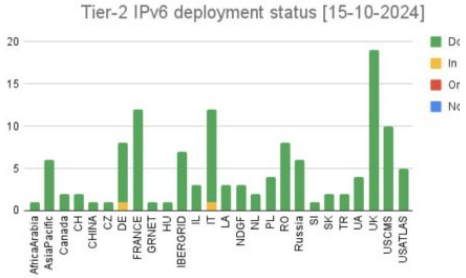
- Torus interconnect
- Network is based on own-designed, FPGA-based NICs
- Fast-send for small packets optimization (by-pass QDMA)
- Supports packets prioritization
- * per queue credit system

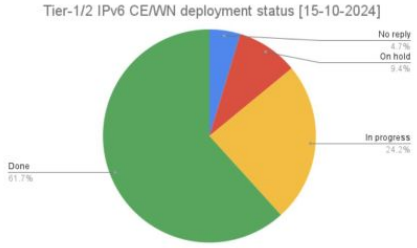


➔ David Kelsey –

Towards an IPv6-only WLCG: more successes in reducing IPv4

- Campaign to deploy IPv6:
- 99% of T2s storage is dual-stack
- 62% of T2s worker nodes is dual-stack
- IPv6 is required to tag the packets
- Still around 20% of transfers between CERN and KIT fallback to IPv4



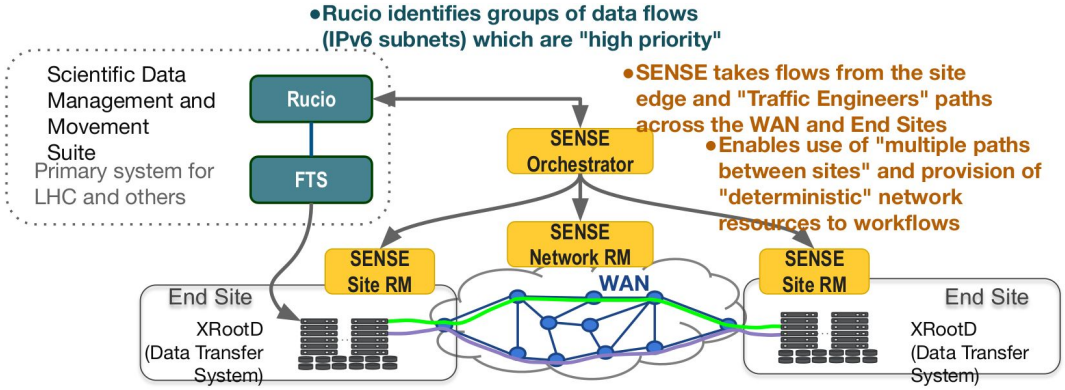


IPv6 + SKA + network advancments

➔ Justas Balcas –

Software defined network control for LHC Experiments

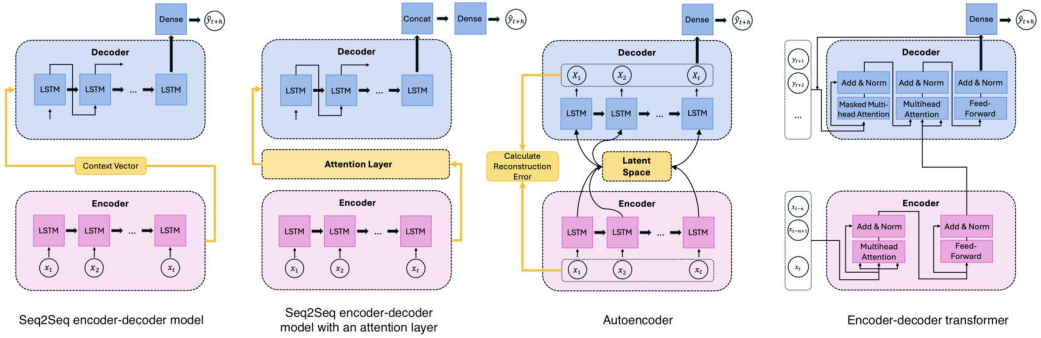
- The Internet is a blackbox and we have to change it
- SENSE can provide a model of the whole network
- No site changes, management can be done at the network provider level
- SENSE allows to choose between different paths
- Software Router for SENSE/Rucio



➔ Maria Del Carmen Misa Moreira –

Diving into large-scale congestion with NOTED as a network controller and machine learning-based traffic forecasting

- Large transfers can saturate some particular path leaving other idle
- FTS (current and future transfers) and CRIC are the sources of information
- NOTED provides automatic link selection (load balancing) when traffic is high



Seq2Seq encoder-decoder model | Seq2Seq encoder-decoder model with an attention layer | Autoencoder | Encoder-decoder transformer

- thanks to all speaker for wonderful talks,
- the organizing committees for the good arrangements
- the audience for the active listening and the sharp questions

…

hope to see you at next chep in 2026