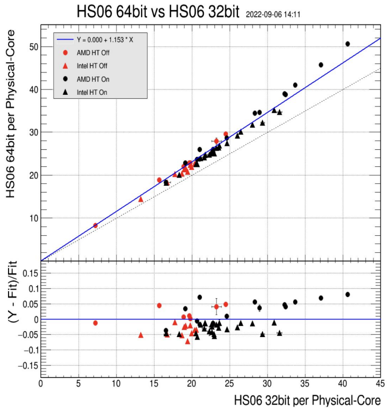
Facilities and Virtualization
Track 7 Summary

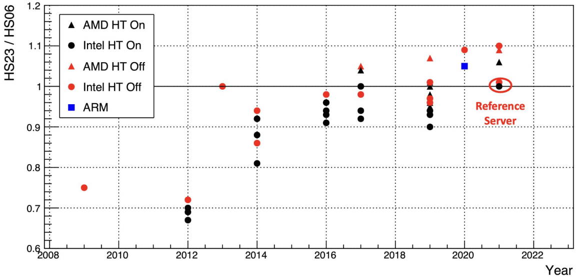
Tomoe Kishimoto (KEK)
Verena Martinez Outschoorn (University of Massachusetts Amherst)
Derek Weitzel (University of Nebraska Lincoln)
Arne Wiebalck (CERN)

Track 7 - Facilities and Virtualization Overview

- 7 sessions covering topics
  - Dynamic Provisioning and Anything-As-A-Service
  - Analysis Facilities
  - Computing Centre Infrastructure
  - Computing Centre Infrastructure and Cloud
  - Networking
  - HPC and Deployment
  - Deployment, Management and Monitoring
- 41 oral presentations and 21 posters

[HEPScore - new CPU benchmark for WLCG]

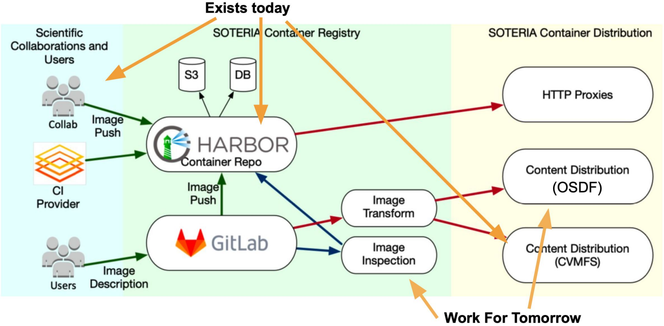


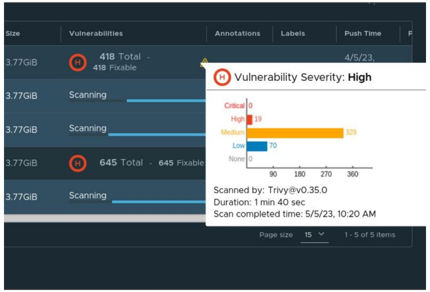


Virtualization & Containers

- Progress towards trustworthy container image distribution

SOTERIA for container registry, discoverability, visibility & traceability





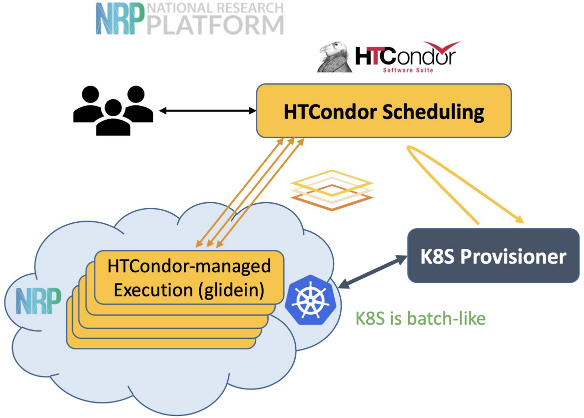
Apptainer Without Setuid

apptainer.org

Measurements of a HEP python-based container benchmark run times on a 16-core system (lower is better):

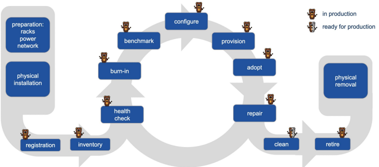
sandbox on local disk:	6:21
sandbox on lustre:	6:32 (only one node, not parallel launches)
kernel squashfs, sif on lustre:	6:33
standard squashfuse:	41:33
standard squashfuse_ll:	12:48
multithreaded squashfuse_ll:	6:29
sandbox on CVMFS:	6:50 (warm cache)

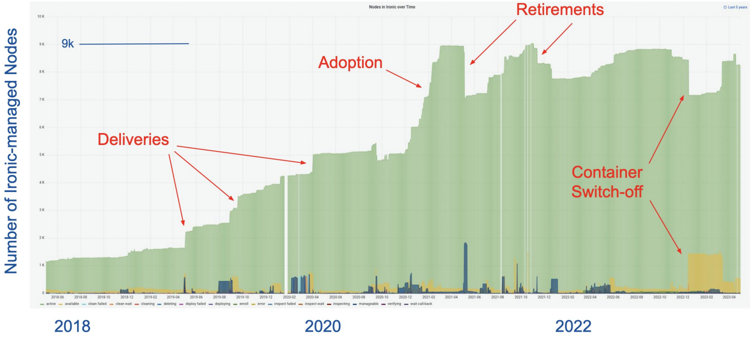
Demand-driven provisioning of k8s-like resource in OSG



Computing Facilities: Automation & Improved Efficiency

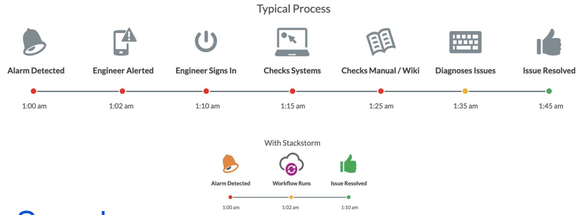
Automated Server Management for new Data Center at CERN



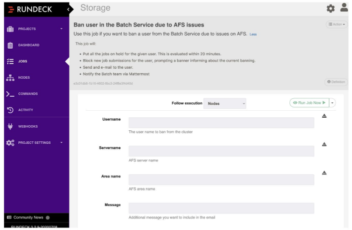


Managing Batch worker node lifecycle & maintenance at CERN

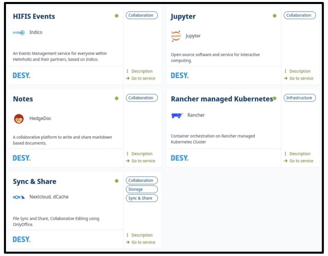
Stackstorm for automatic remediation



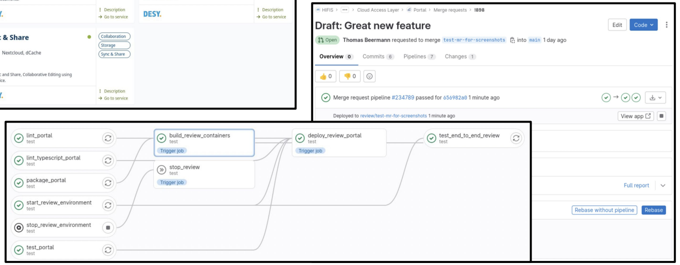
Prevent AFS overloads



OpenSearch search & analytics engine at CERN



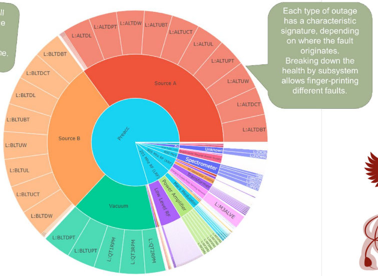
Management of Deployment of Cloud Services using GitOps at DESY



AI for Improved Facilities Operation

	Date_Time	System	Duration(in Minutes)	FaultType
◉	2021-06-26T00:35:00	RF	1.02	LRF2 Spark Trip
○	2021-06-25T12:30:00	RF	4.98	LRF Driver Trip and KRF 2 Reflected Power
○	2021-06-25T07:48:00	RF	1.98	LINAC KRF5/6 PFN ACIN BAD trip
○	2021-06-23T13:15:00	Diag/Inst	4.98	Linac's BLM IRM check out.
○	2021-06-21T21:31:00	RF	1.98	KRF7 Reflected Power Rad trip.

Outages are automatically assigned labels and the most recent ones are displayed.



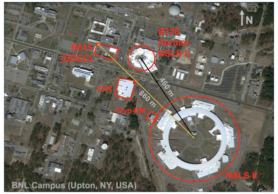
Checkpoint Restore in Userspace (CRIU) Tool

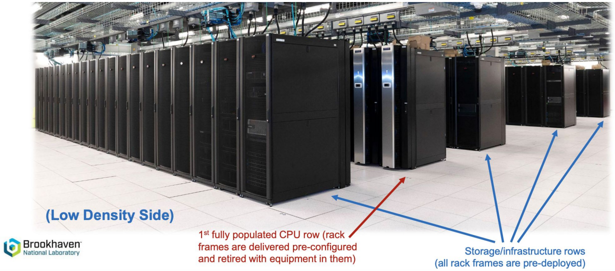
Checkpointing batch jobs

Computing Facilities & Infrastructure

- New facilities with improved infrastructure & capacity for expansion
- Dedicated solutions to meet experiment requirements of performance, availability & security

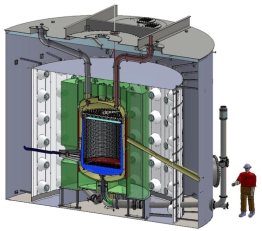
[Transition to New Data Center at BNL]

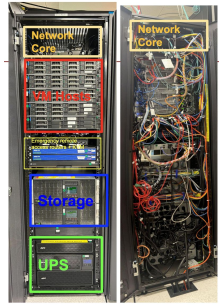




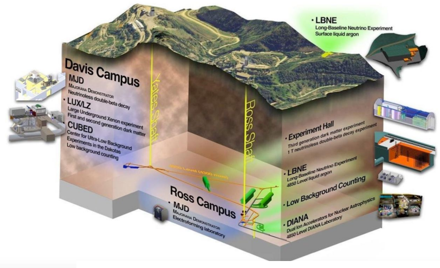
[Transition to New Data Center INFN-CNAF at Bologna Technopolo]

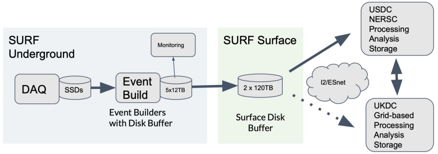
[IT Infrastructure for LZ at Sanford Underground Research Facility (SURF)]

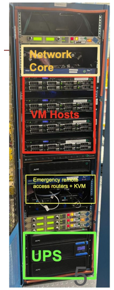




Surface Rack, front and back





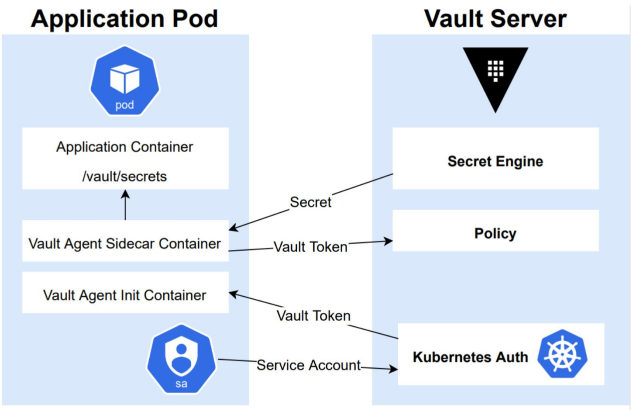


Underground Infrastructure & Network Rack

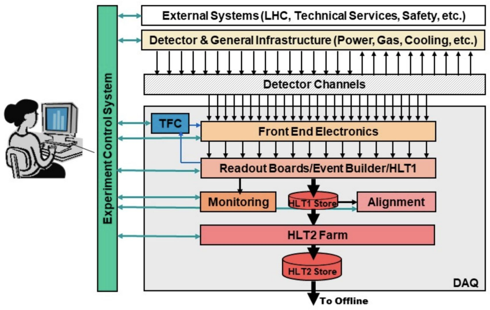
Critical Services & Control Systems

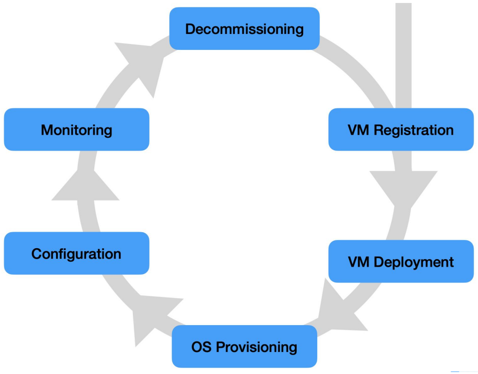
- Security, improved system management and disaster recovery

New Security Features in CMSWEB at CERN



Experiment Control System Infrastructure for LHCb





Modular toolsets for integrating HPC clusters in experiment control systems

Computing Facilities: Energy Efficiency & Carbon Footprint

- Various efforts to improve energy efficiency and carbon footprint
  - Not the same thing!
  - Energy becoming net zero, so embedded carbon footprint becoming more relevant
  - Input for when to replace / how long to run hardware, additional incentives to replace

IRISCAST Proposes

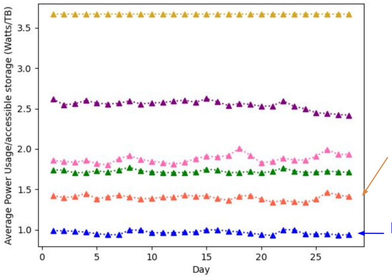
Measuring Carbon Footprint

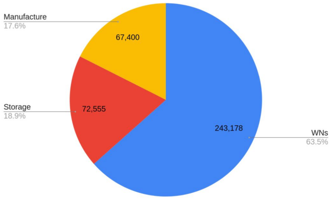
IRISCAST – Audit of Carbon Costs

High Level Feedback	Low Level Feedback		
Carbon Equivalent per month	Figure of merit per Job	→	Carbon Equivalent per Job

Environmental Impact Estimate of a Tier 2

wf: HP ProLiant SL2x170z G6 (2010) 17.0 W/core
wg: Dell PowerEdge R410 (2011) 14.2 W/core
wh: Supermicro X9DRT (2014) 12.3 W/core
wi: Supermicro X10DRT-P (2016) 10.2 W/core
wj: Dell PowerEdge R430 (2017) 10.1 W/core
wk: Dell PowerEdge R440 (2019) 10.9 W/core
wl: Supermicro H11DSU-iN (2020) 5.4 W/core
wm: Dell PowerEdge R6525 (2020) 6.1 W/core



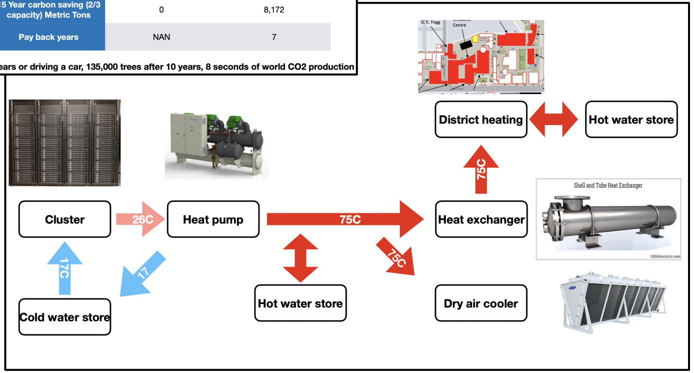


PUE = 1.3: **383,000 kg** = 445 sysadmins crossing the Atlantic in a plane*

Data Centre Refurbishment at QMUL

SCOPE	Minimum scheme: No Heat recovery	Full scope: Heat recovery & dry air cooler
Racks	26	39
Cooling Capacity	260KW	390KW
District Heating connection	No	Yes
APX cost		
15 Year carbon saving (2/3 capacity) Metric Tons	0	8,172
Pay back years	NAN	7

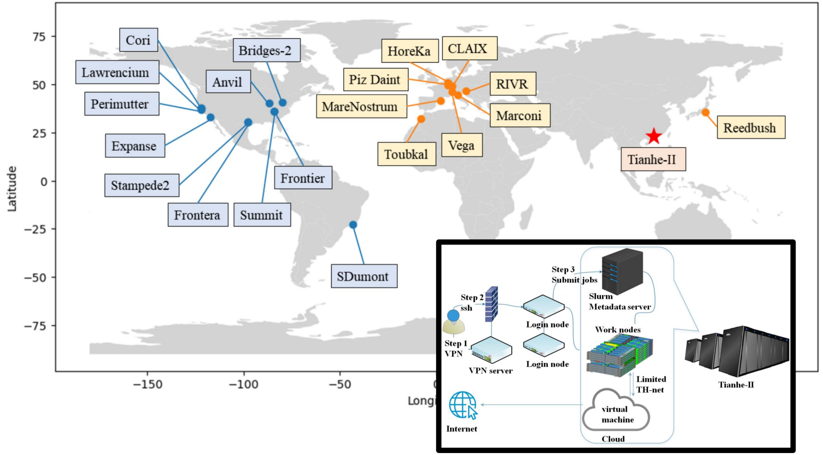
1,819 years or driving a car, 135,000 trees after 10 years, 8 seconds of world CO2 production



Computing Facilities: High Performance Computing (HPC)

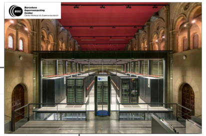
- Integration of HPC Centers into workflows
  - Overcoming limitations/policies of these centers, e.g. network connectivity
    - Reverse ssh, proxies

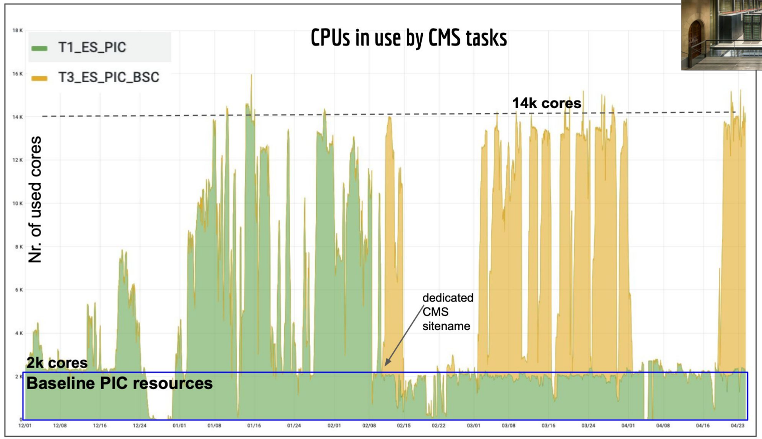
Tianhe-II Supercomputer for BESIII



3.12M CPU cores, 15PB

Barcelona Supercomputing Center for CMS





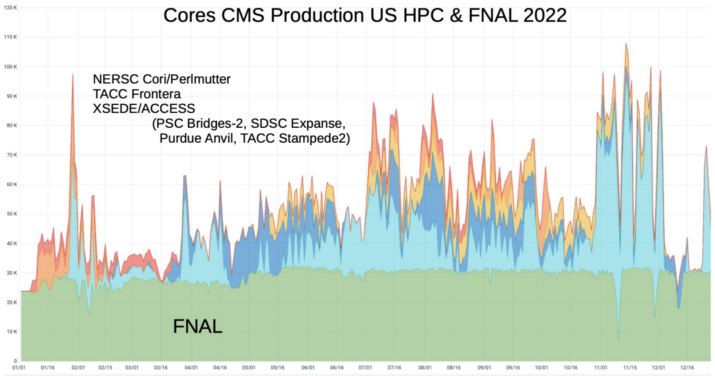
Computing Facilities: High Performance Computing (HPC)

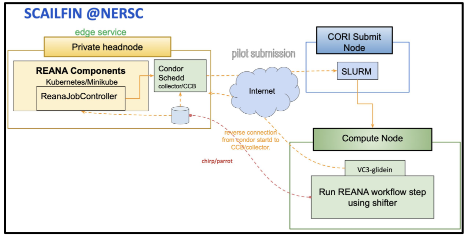
- Integration of HPC Centers into workflows
  - Exploring possibilities to use of heterogeneous architectures (ARM, Power, GPUs)
  - Scalable infrastructure, more complex workflows for AI/ML

NERSC National Energy Research Scientific Computing Center

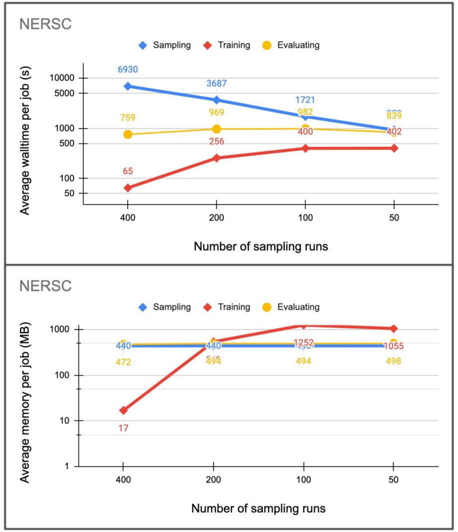
US HPC resources for CMS - Commissioning GPU resources

Large-scale HPC deployment of Scalable CyberInfrastructure





reana

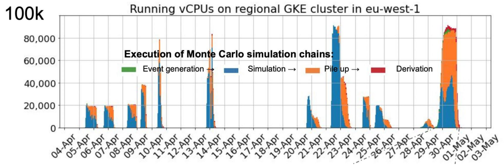


Cloud Resources

- Expanding experience using cloud resources
  - Opportunities for new ideas and evolution, complementary resources, elastic usage, access to new architectures, etc

Commercial Clouds in distributed computing for ATLAS

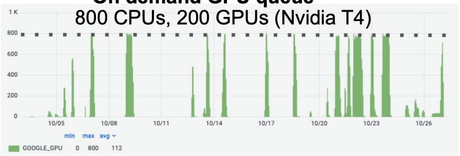
Elastic processing on Google Cloud



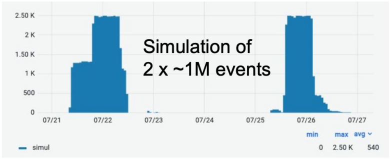
O(10^5) vCPUs
O(10^4) Pods
O(10^3) Nodes
1 managed K8S cluster
<1 Engineer

On demand GPU queue

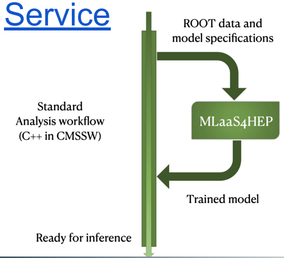
800 CPUs, 200 GPUs (Nvidia T4)

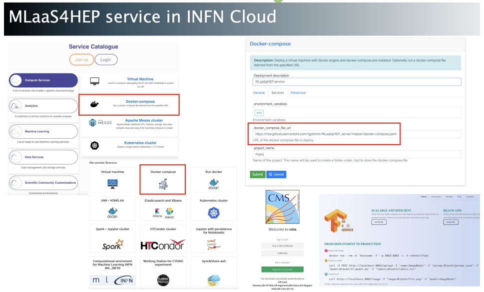


First ATLAS tasks on ARM: Amazon Gravitron 2 processors

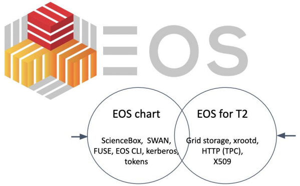


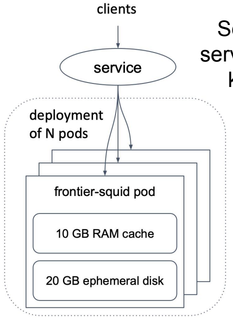
Cloud-native solution for ML as a Service

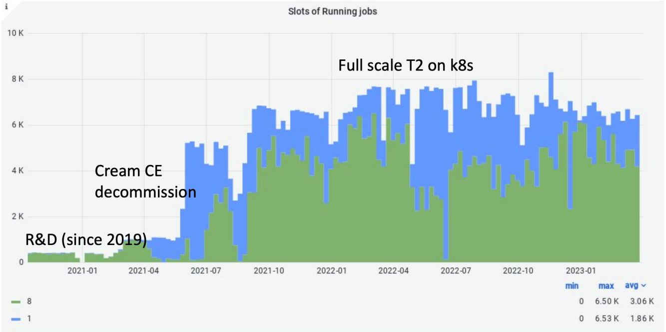




Fully cloud-native T2 in ATLAS





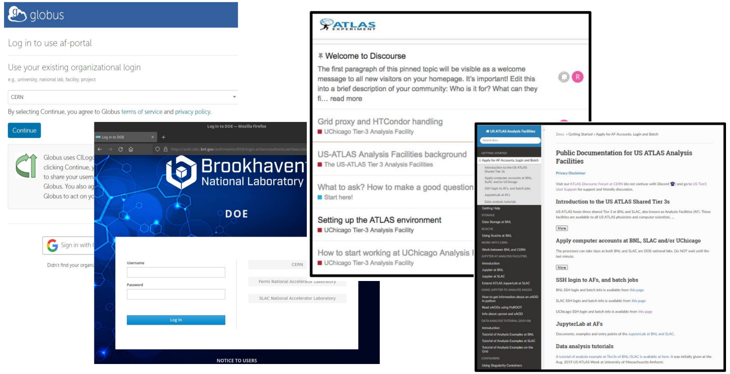


Financial study of cloud resources

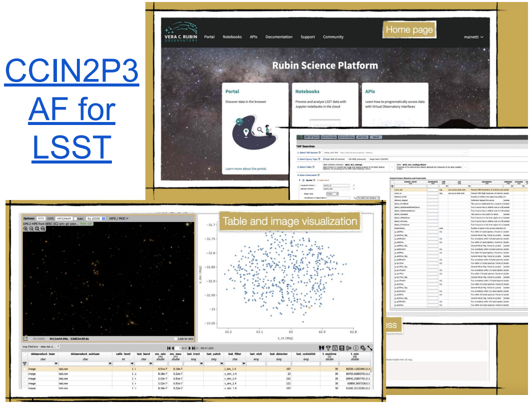
Analysis Facilities (AF)

- Variety of approaches to support analysis workflows "... with integrated data, services, and computational resources"
  - What is an AF actually? Local storage vs remote, batch vs interactive, …
  - Moving analysis from one AF to another, aiming at consistent & convenient user experience,
    - Goal is to abstract the underpinning infrastructure

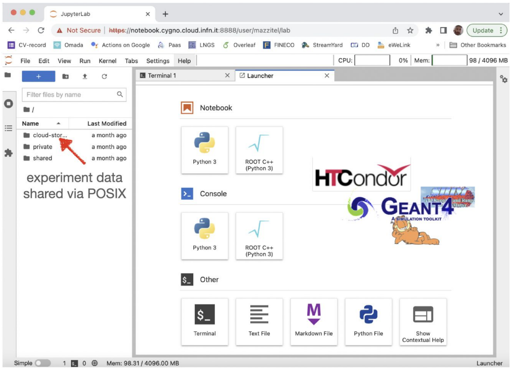
[US Shared AFs for ATLAS]



[CCIN2P3 AF for LSST]



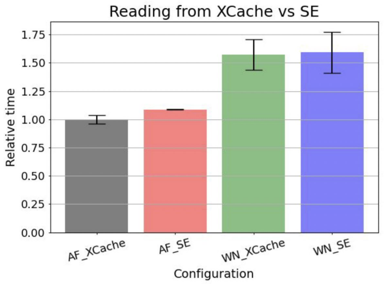
[INFN Cloud for CYGNO]

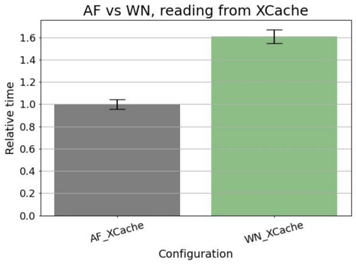


Analysis Facilities (AF)

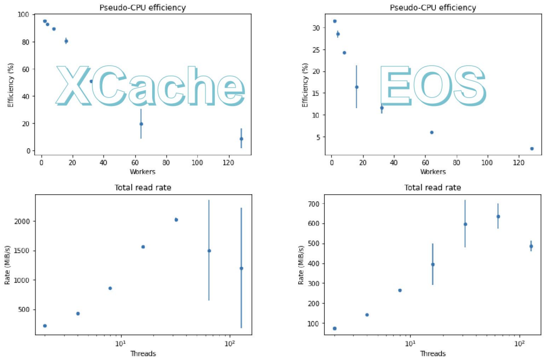
- Variety of approaches to support analysis workflows “... with integrated data, services, and computational resources”
  - What is an AF actually? Local storage vs remote, batch vs interactive, …
  - Scalability and turn-around time
  - XCache is a common approach to reduce I/O latency

CIEMAT AF for CMS

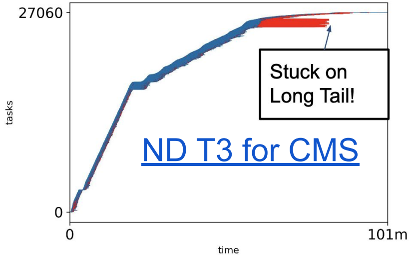


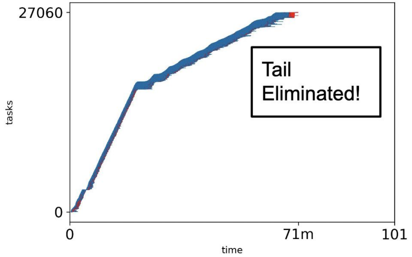


CERN IO Performance for Analysis



ND T3 for CMS

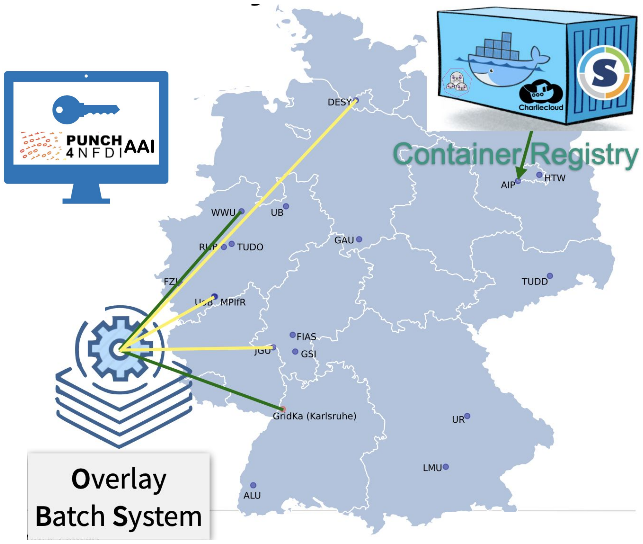




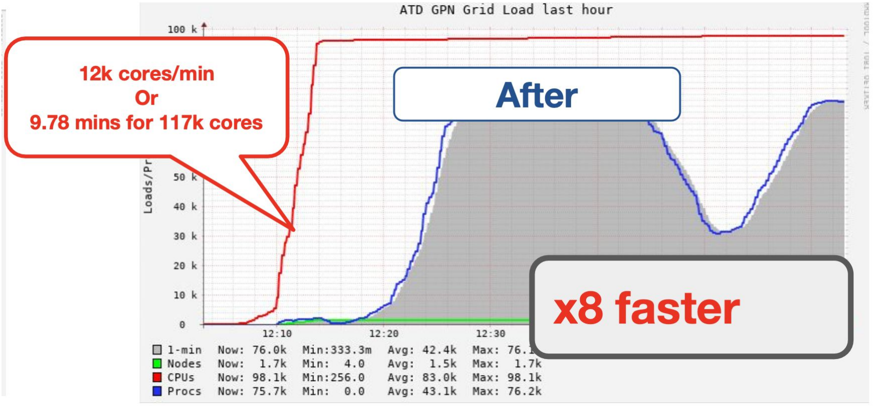
Heterogeneous & Opportunistic Resources

- Optimization of utilization of opportunistic resources
- Federation of heterogeneous resources

[Federated heterogenous compute & storage for PUNCH4NFDI Consortium]



[Opportunistic use of High-Level Trigger farm for ATLAS]



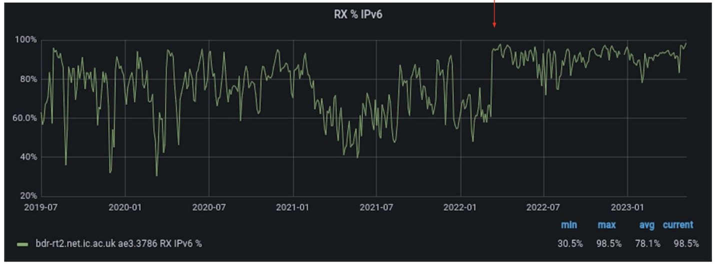
Networking

- Completion of the transition to IPv6, monitoring and network traffic analysis

[IPv6 on WLCG](#)

Experiment	Fraction of T2 storage accessible via IPv6
ALICE	90%
ATLAS	90%
CMS	96%
LHCb	100%
Overall	93%

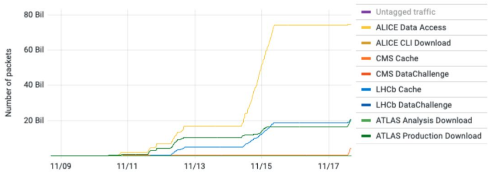
dCache storage preference set to IPv6



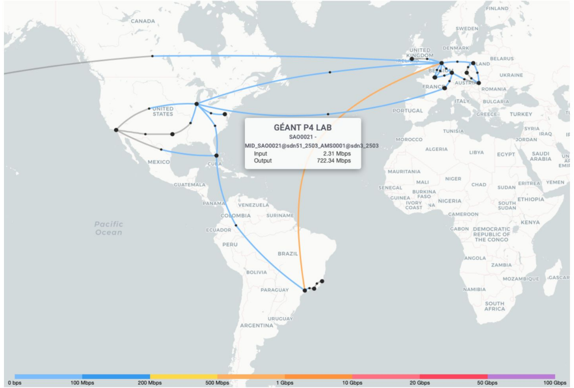
Since Feb 2022
~90% IPv6

[P4flow for network traffic analysis](#)

Counters of an access-list in bits/s



Counters of an access-list in number of packets

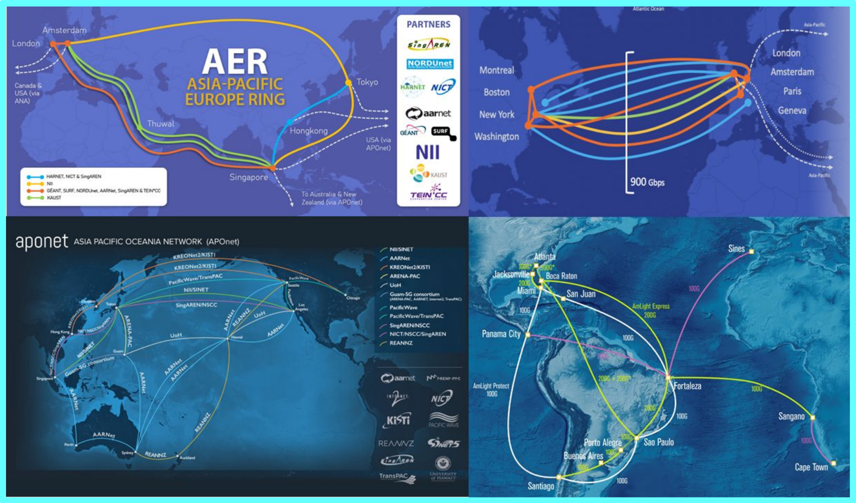


WLCG traffic routing from São Paulo directly to Amsterdam.

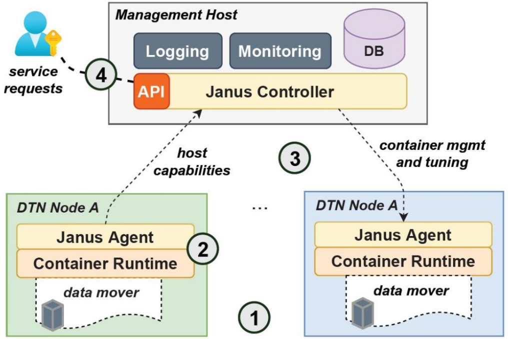
Networking

- Developments towards a next-gen network-integrated system
    - Global coordination of first-class resource – flexible, dynamic that balances and makes best use of available resources
    - Several ongoing R&D projects: regional caches, intelligent control & optimizations

Global Network Advancement Group
Next Generation System



In-network data caches

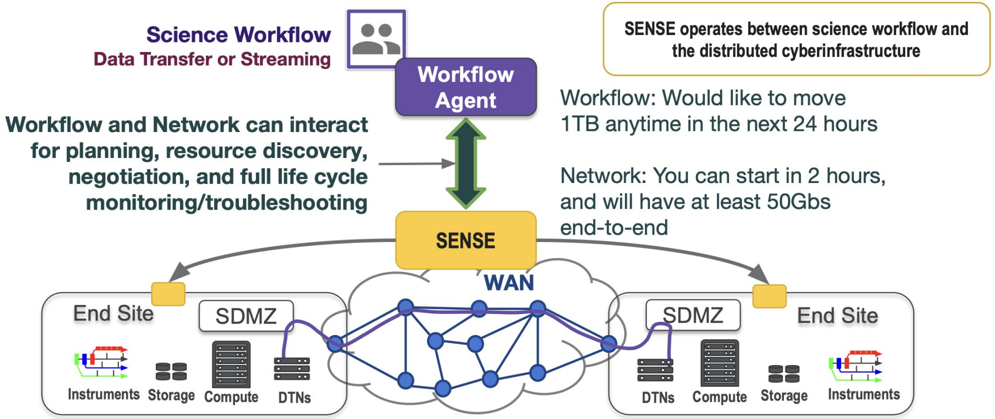


data sharing among users in the same region
in-network opportunity to better dictate usage

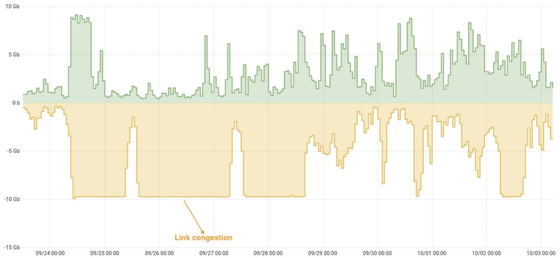
Networking

- Developments towards a next-gen network-integrated system
  - Global coordination of first-class resource – flexible, dynamic that balances and makes best use of available resources
  - Several ongoing R&D projects: regional caches, intelligent control & optimizations

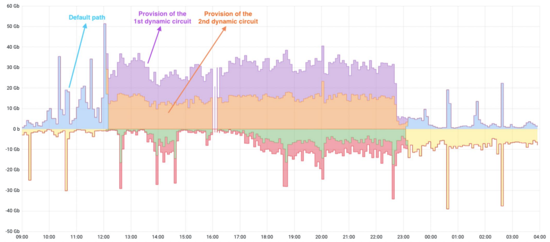
SENSE End-to-End Network Path Control with QoS Capabilities



NOTED intelligent network controller



NOTED demo for SC22



Track 7 - Facilities and Virtualization Conclusions

- 7 sessions covering topics
    - Dynamic Provisioning and Anything-As-A-Service
    - Analysis Facilities
    - Computing Centre Infrastructure
    - Computing Centre Infrastructure and Cloud
    - Networking
    - HPC and Deployment
    - Deployment, Management and Monitoring

- 41 oral presentations and 21 posters
- **Excellent contributions, stay tuned for the proceedings!**