

A brief roundup of topics of interest from recent Fall meetings: HEPiX/GDB/LHCOPN/DC24.

Hironori Ito, Ofer Rind

Dec 13, 2023



@BrookhavenLab

Links to the meetings

Links to the agendas of various meetings

- LHCOPN/ONE UVic Oct, 2023 <https://indico.cern.ch/event/1280363/>
- HEPiX UVic Oct 2023 <https://indico.cern.ch/event/1289243/>
- WLCG Management Board Oct 2023 <https://indico.cern.ch/event/1225423/>
- DC24 Workshop Nov 2023 <https://indico.cern.ch/event/1307338/>
- Pre-GDB Nov 2023 <https://indico.cern.ch/event/1225131/>
- GDB Nov 2023 <https://indico.cern.ch/event/1225118/>
- WLCG DOMA General Dec 2023 <https://indico.cern.ch/event/1350973/>

LHCONE/OPN

- CERN to ESnet
 - Connected by 2x400 Gbps in addition to the routes to/from London and Amsterdam
 - Used by BNL, FNAL and LHCONE (Tier2s)
- BNL to ESnet
 - Will be connect by 2x400 Gbps on Dec 19, 2023
 - Used by LHCOPN and LHCONE
- Dune joined LHCONE (Sept 2023)
 - WLCG, US ATLAS/CMS, Belle II, Pierre Auger Observatory, NOvA, XENON, JUNO, DUNE
 - How to control the exposure to different VOs
 - Multi VO VPNs. Very complicated
 - VO tagging
- Development
 - NOTED
 - SDN with FTS
 - Packet marking & pacing
 - SENSE
 - VPN with RUCIO and QoS

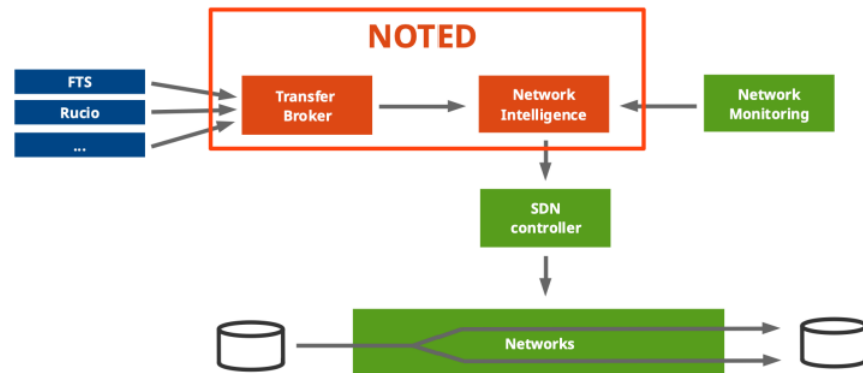
Open to other HEP collaborations



NOTED

NOTED: An intelligent network controller to improve the throughput of large data transfers in File Transfer Services by handling dynamic circuits

Architecture

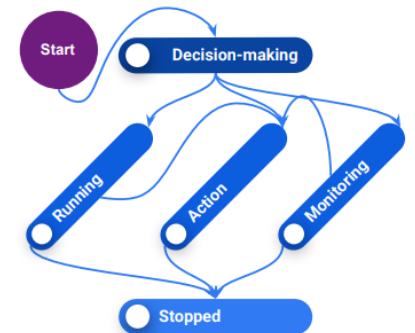


NOTED (Network Optimized Transfer of Experimental Data)

An intelligent network controller to improve the throughput of large data transfers in FTS (File Transfer Services) by handling dynamic circuits.

NOTED actions

- Decision-making: NOTED is making the network decision to potentially execute an action or not.
- Running: NOTED is running but there are no transfers in FTS so NOTED is waiting and running until the link-saturation alarm is cleared.
- Monitoring: NOTED is running and there are on-going FTS transfers, but they are below the defined bandwidth threshold that we establish.
- Action: NOTED is running and has triggered an SDN action to provide more bandwidth.
- Stopped: NOTED has stopped because there are no transfers in FTS and the link-saturation alarm has cleared.

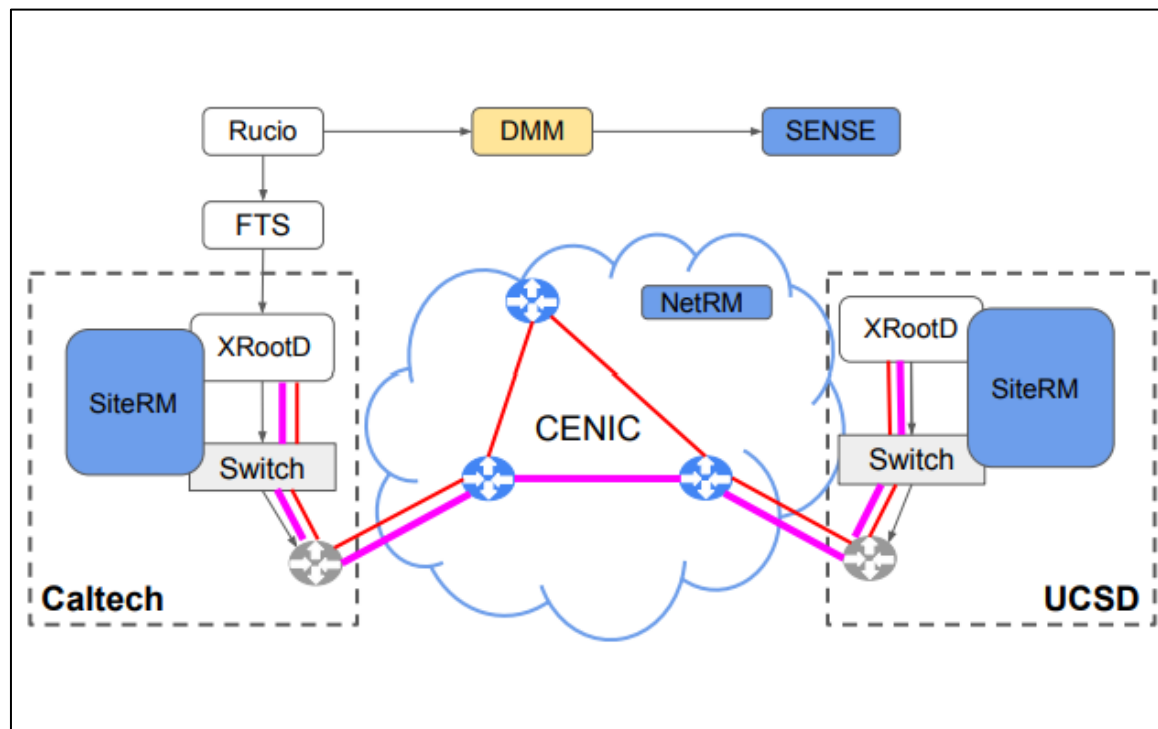


Rucio/SENSE

For every Rucio request, Rucio contacts DMM to ask for the endpoints (IP addresses) to use before contacting FTS

For a regular request (red) DMM will return the IPv6 addresses selected for “best effort”

SENSE is only contacted by DMM in order to get the set of IPv6 addresses of the 2 sites involved in the transfer. This information is cached



For a priority Rucio request (pink) DMM picks a pair of free IPv6s and requests a bandwidth allocation on them to SENSE

DMM return the selected pair of IPv6 to Rucio

SENSE instructs SiteRM to implement specific routing and QoS on the given IPv6s at the site level

SENSE instructs NetworkRM to implement specific routing and apply QoS in CENIC nodes in between the 2 IPv6 endpoints

When the transfer is finished Rucio signals DMM which request the deallocation of the priority services

IPv6

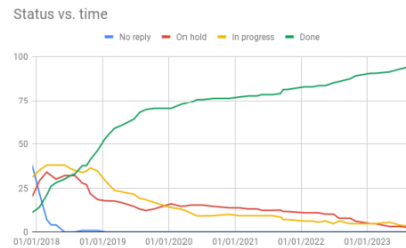
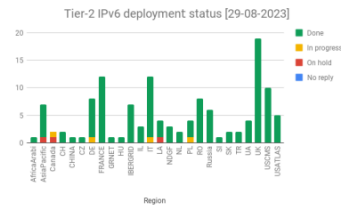
- End point is still IPv6-only services (IPv4 is “legacy” networking)
- *Message to new research communities - build on IPv6 from start*

Storage

Good news! - IPv6/IPv4 at Tier-1/2 sites

- Tier-1 complete
- Tier-2 deployment from Nov17
- ([status](#)) shows >94% T2 sites
 - 97% of Tier-2 storage dual stack

Experiment	Fraction of T2 storage accessible via IPv6
ALICE	91%
ATLAS	95%
CMS	100%
LHCb	100%
Overall	97%



D. Kelsey - Ensure use of IPv6



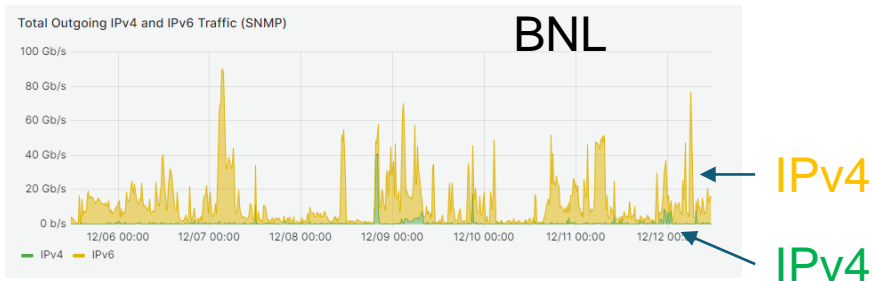
Worker nodes

- All WLCG sites should offer IPv6 connectivity on their compute services (CEs and WNs) by June 30, 2024
 - Switching off IPv4 is not requested nor recommended: sites wishing do to it must discuss it with the supported experiments
- Progress would be tracked by launching a GGUS ticketing campaign, exactly as it happened for the storage services and perfSONAR

At BNL

The testing of dual stack worker nodes has just begun.

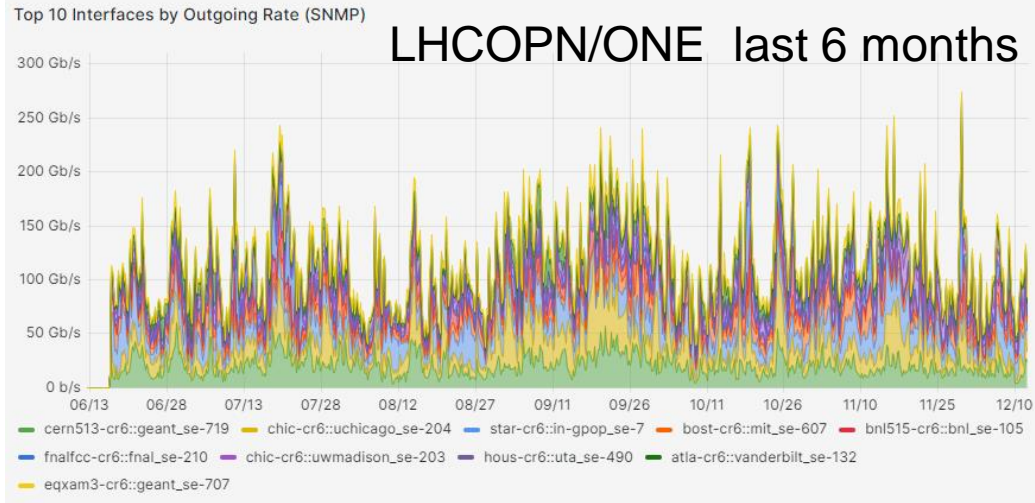
- Tests all clients: Panda, Rucio, CVMFS, etc...
- Run production jobs.



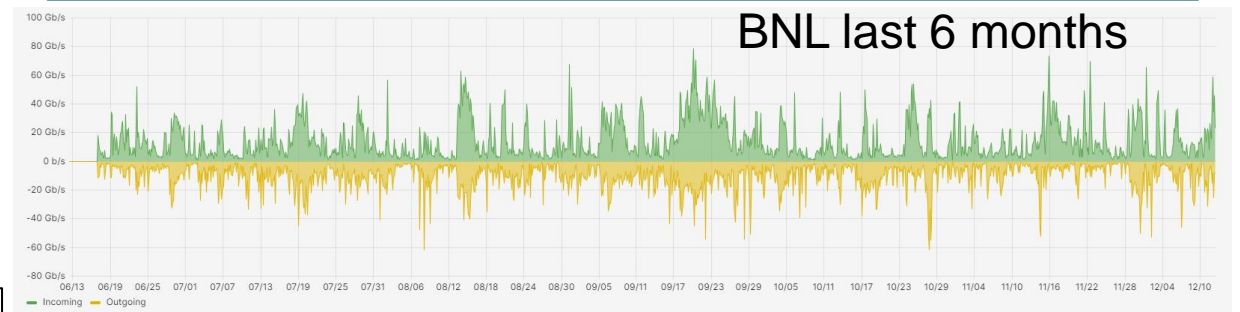
Data rate during the last 6 months

External Network: some numbers

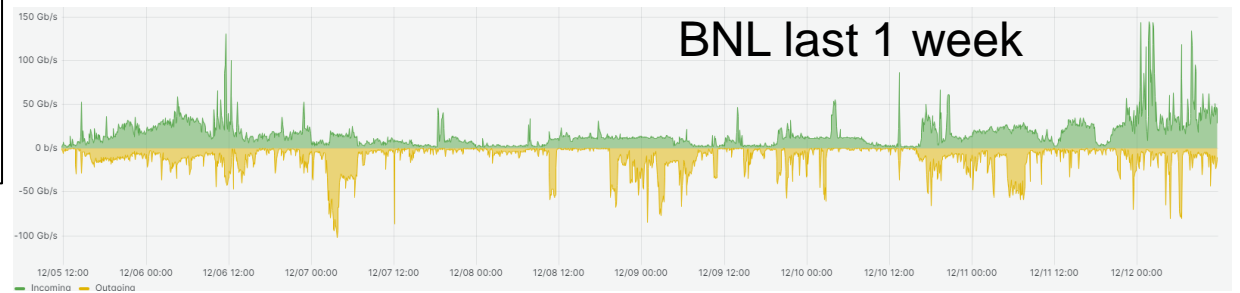
- LHCONE capacity: 1.2 Tbps
- LHCOPN capacity: 2.1 Tbps
- Internet capacity: 1 Tbps



BNL: 2x100Gbps --> 2x400 Gbps Dec 19, 2023



At BNL
Average: 20~30Gbps
Maximum: over
100Gbps regularly



DC24

DC24 Timetable

- Dates: **February 12th (Mon) to February 23rd (Fri)**
- Proposal to distribute different exercises over the challenge days, e.g.
 - Day 1-3: T0 export
 - Day 4-5: Reprocessing like traffic
 - Day 6,7 (weekend): Keep things running...
 - Day 8-9: MC like traffic
 - Day 10-11: Increase to flexible scenario
 - Day 12: Repeat things e.g. with adjusted setting
 - Day 13-14 (weekend): Hope that nothing completely broke

The target rate is well within the commonly observed rate seen at BNL as shown earlier.

Tier-0-Tier1s (Higher rates)

NOTICE: These are link rates from/to CERN-PROD and the different Tier1s
 [1] <https://twiki.cern.ch/twiki/bin/view/LHCOPN/OverallNetworkMaps>
 [2] MONIT link: <https://monit-grafana-open.cern.ch/d/000000523/home?orgid=16&viewPanel=1>

TO	Tier1s														Sum total	Sum total GB/s			
CERN-PROD source (Write rates)	TW-ASGC	RRC-KI	ES-PIC	DE-KIT	FR-CCIN2P3	IT-INFN-CNAF	UK-RAL	NDGF (CH-LHEP)	NDGF (Scandinavia)	NL-T1(Nikhef, SA PL-NCBJ)	CN-IHEP	RRC-JINR	CA-TRIUMF	US-BNL	US-FNAL	RC-KISTI	Sum total	Sum total GB/s	
ALICE																			
ATLAS (injected + prod)	0.1	0.8	13	38.4	43.5	27.7	43.5	0	24.4	18.9	0	0	0	28.6	67.4	0	306.3	38.2875	
CMS	0	0	4.38	23.54	13.14	17.61	34.02	0	0	9.88	8.76	8.93	0	0	0	0	120.26	15.0325	
LHCb	0	0	36.38	111.94	105.64	109.31	113.52	0	27.4	29.78	8.76	8.93	68	28.6	67.4	177	2	895.56	111.945
Network Capacity[1]	100Gbps	100Gbps	100Gbps	200Gbps	100Gbps	200Gbps	200Gbps	100Gbps	400Gbps	20Gbps	20Gbps	100Gbps	100Gbps	200Gbps	100Gbps	40Gbps			
DUNE																			
Belle II (from KEK via LHCONE)	0	0	0	1.9	2.8	3.7	0	0	0	0	0	0	0	0	5.6	0	14	1.75	
CERN-PROD destination (Read rates)	TW-ASGC	RRC-KI	ES-PIC	DE-KIT	FR-CCIN2P3	IT-INFN-CNAF	UK-RAL	NDGF (CH-LHEP)	NDGF (Scandinavia)	NL-T1(Nikhef, SA PL-NCBJ)	CN-IHEP	RRC-JINR	CA-TRIUMF	US-BNL	US-FNAL	RC-KISTI	Sum total	Sum total GB/s	
ALICE				n/a	n/a	n/a	n/a	n/a	n/a	n/a							0	0	
ATLAS (injected + prod)	0	0.08	1.64	6.36	6.57	4.19	6.16	0	3.27	3.42	0	0	0	5.94	10.86	0	48.49	6.06125	
CMS	0	0	15	36	36	45	28	0	0	0	0	41	0	0	0	254	455	56.875	
LHCb	0	0	3.44	14.26	10.31	13.74	20.62	0	0	6.87	7.65	5.41	0	0	0	0	82.3	10.2875	
Total	0	0.08	20.08	56.62	52.88	62.93	54.78	0	3.27	10.29	7.65	5.41	41	5.94	10.86	254	0	585.79	73.22375
Network Capacity[1]	100Gbps	100Gbps	100Gbps	200Gbps	100Gbps	200Gbps	200Gbps	100Gbps	400Gbps	20Gbps	20Gbps	100Gbps	100Gbps	200Gbps	200Gbps	40Gbps			
DUNE																			
Belle II				1			1				1				3.5	3.5			

Identified FR-CCIN2P3 to pass the link capacity.
 They plan to update to 200Gbps link for DC24

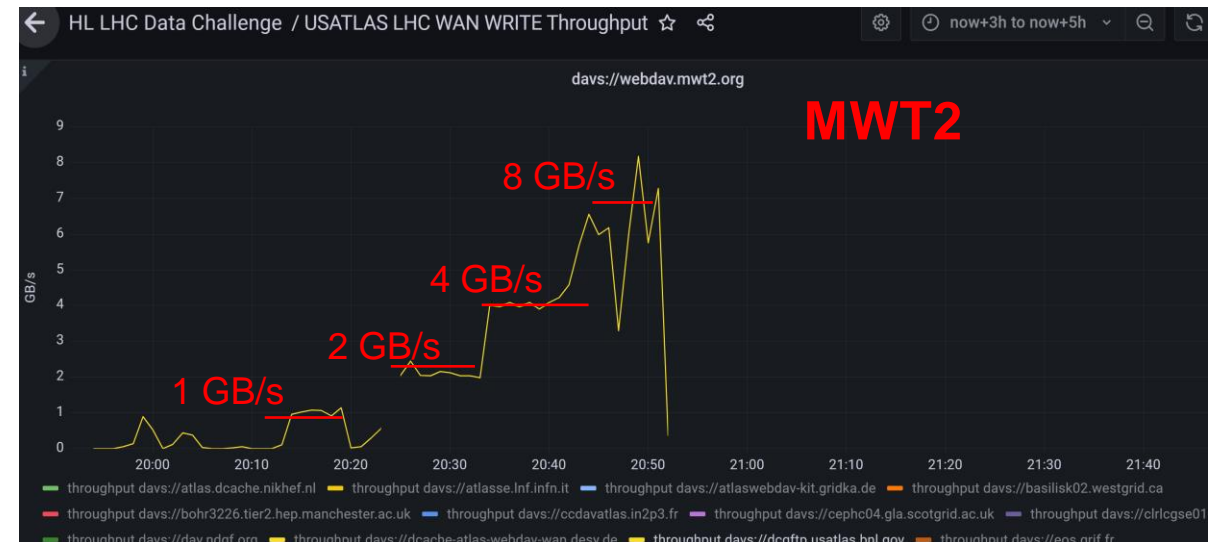
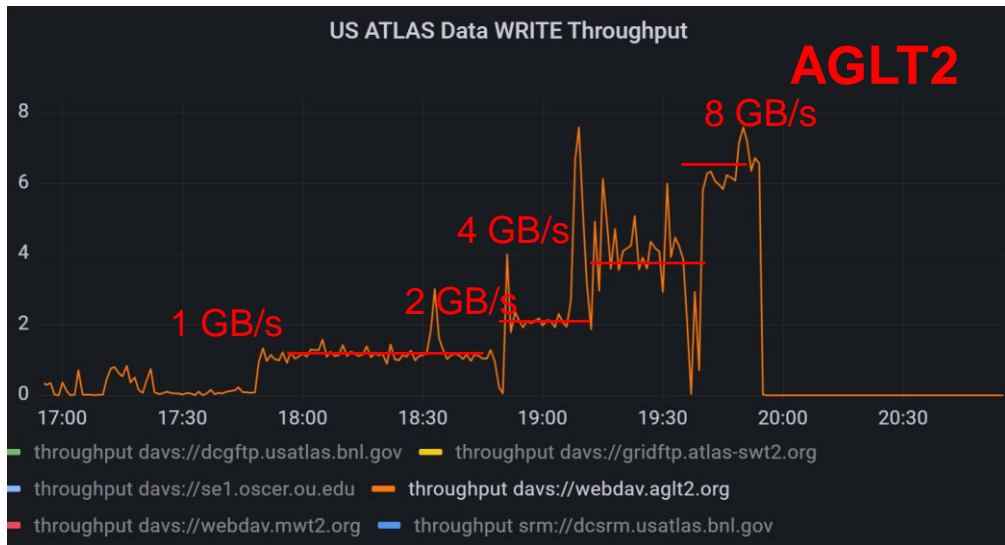
US ATLAS Network Testing

Planning for USATLAS Site Load Testing to identify bottlenecks and performance issues in advance of DC24 (Feb 12-23, 2024)

A bit more challenging target in US

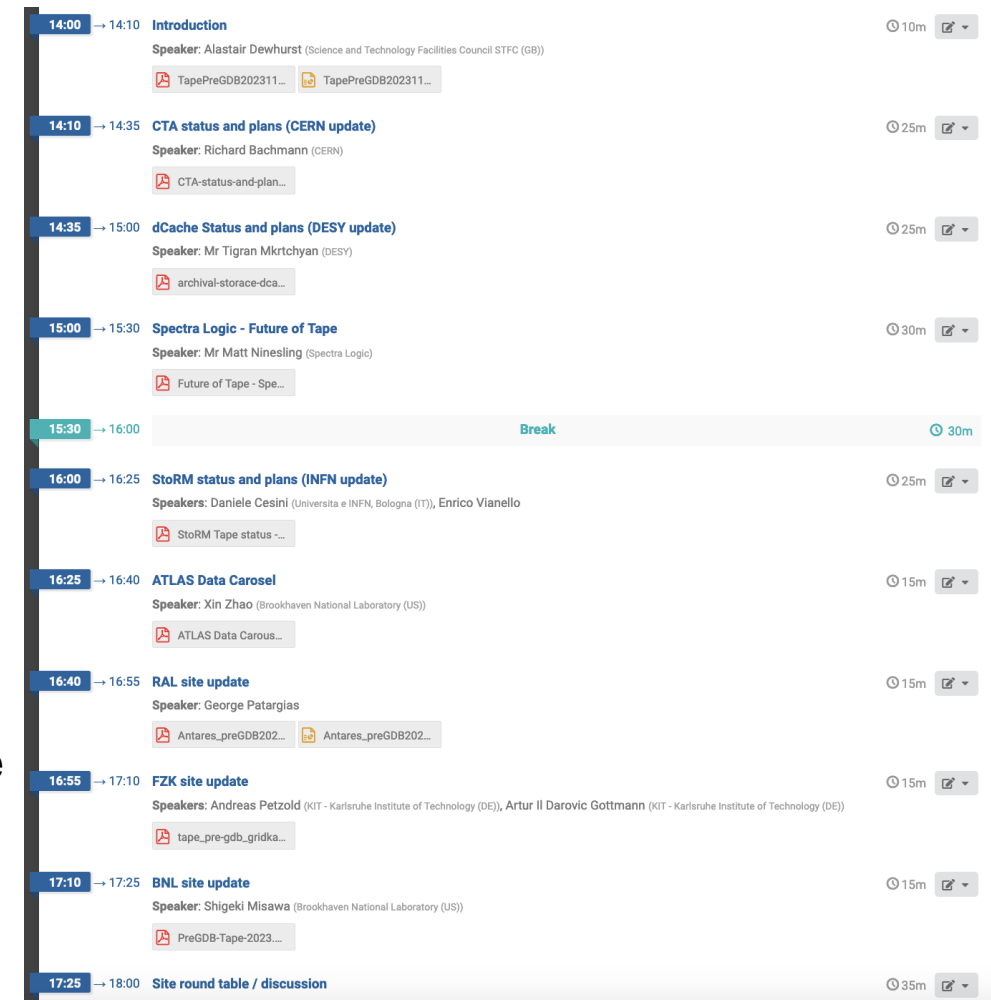
Table: DC24 (src: ingress / egress)			Site WAN (Gb/s)		DC24 minimal scenario				DC24 flexible scenario			
Site	Tier	Cloud	Total (Gb/s)	Usable by ATLAS (Gb/s)	T0 (Gb/s) Export	Total Gb/s & bandwidth		Space [TB/24h] (deletions/hour)	T0 Export	Total Gb/s & bandwidth		Space [TB/24h] (deletions/hour)
						∑ ingress	∑ egress			∑ ingress	∑ egress	
BNL-ATLAS	T1	US	400	400	60.0	81.8	60.0	764(11k)	60.0	111.8-120.2	120.0	1099(13k)
AGLT2	T2	US	200	125		9.9 - 11.4	7.0 - 7.0	56 (1k)		47.0 - 56.5	49.3 - 49.3	531 (8k)
MWT2	T2	US	200	150		24.2 - 28.4	9.9 - 9.9	155 (2k)		59.6 - 70.2	67.2 - 67.2	596 (8k)
NET2	T2	US	10	10		0.0 - 0.0	0.0 - 0.0	0 (0k)		0.0 - 0.0	0.0 - 0.0	0 (0k)
OU_OSCER_ATLAS	T2	US	100	25		1.2 - 1.4	0.6 - 0.6	8 (0k)		5.3 - 6.4	4.8 - 4.8	60 (1k)
SWT2_CPB	T2	US	100	80		9.7 - 11.0	8.5 - 8.5	48 (1k)		58.4 - 70.4	60.7 - 60.7	674 (10k)

Load Generator



November pre-GDB

- Topical [meeting](#) on Tape Evolution organized by Alastair Dewhurst, et. al. and held at CERN on Nov. 7th
- ~25 participants in person, ~35 on Zoom
 - Reports from sites, storage software providers, and industry (recordings uploaded)...including Data Carousel update (Xin) and BNL Site Report (Shigeki)
- Some highlights:
 - [New drives and media from IBM](#) (the last remaining company that develops tape technology)
 - 50 TB tapes (Strontium Ferrite), 400 MB/s read/write (as previous generation)
 - More stringent environmental specs (humidity)
 - Roadmap will continue exponential capacity increase
 - Adding archive metadata to [CTA](#) and porting to Alma 9
 - [dCache-CTA integration](#) with upcoming deployments
 - Extensive update on [Storm Tape-REST API](#) plus migration to the new INFN-CNAF Data Center

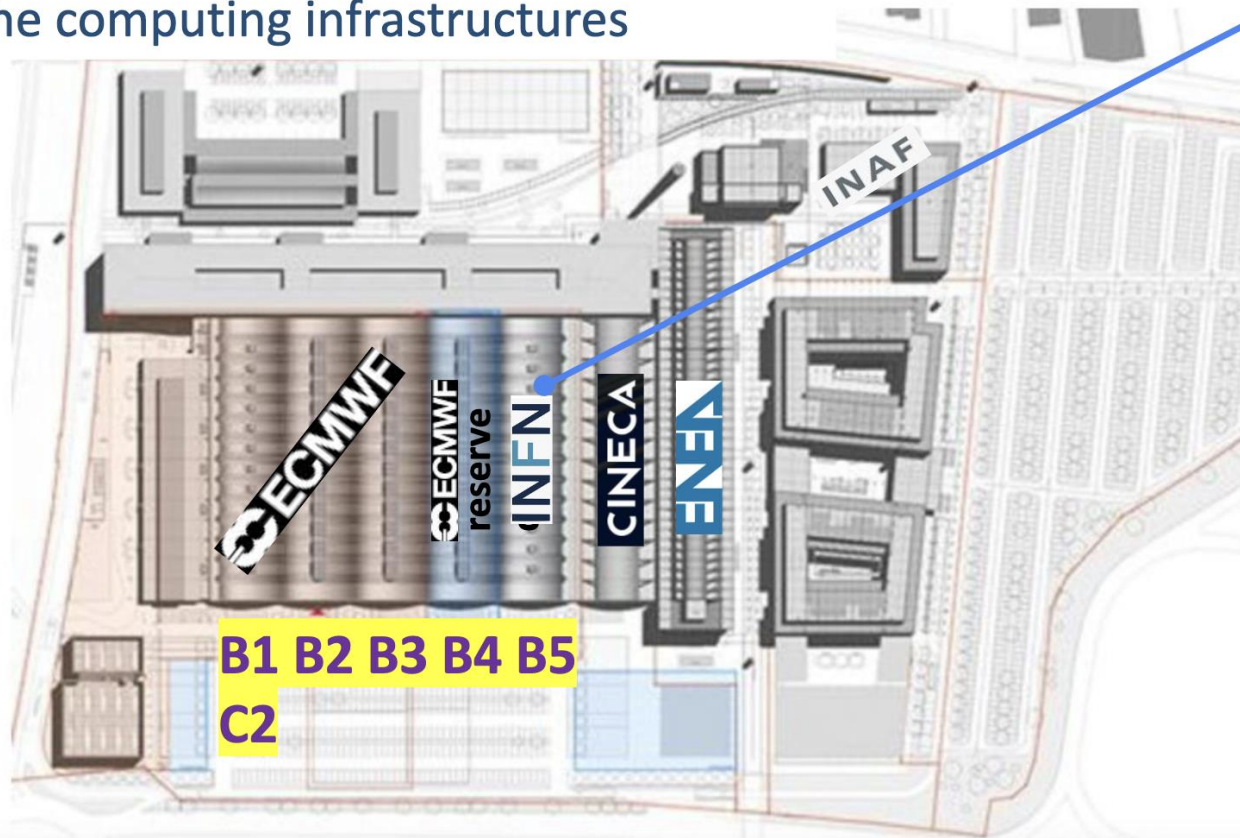


Time	Topic	Speaker	Duration
14:00	Introduction	Alastair Dewhurst (Science and Technology Facilities Council STFC (GB))	10m
14:10	CTA status and plans (CERN update)	Richard Bachmann (CERN)	25m
14:35	dCache Status and plans (DESY update)	Mr Tigran Mkrtchyan (DESY)	25m
15:00	Spectra Logic - Future of Tape	Mr Matt Ninesling (Spectra Logic)	30m
15:30	Break		30m
16:00	StoRM status and plans (INFN update)	Daniele Cesini (Universita e INFN, Bologna (IT)), Enrico Vianello	25m
16:25	ATLAS Data Carousel	Xin Zhao (Brookhaven National Laboratory (US))	15m
16:40	RAL site update	George Patargias	15m
16:55	FZK site update	Andreas Petzold (KIT - Karlsruhe Institute of Technology (DE)), Artur Il Darovic Gottmann (KIT - Karlsruhe Institute of Technology (DE))	15m
17:10	BNL site update	Shigeki Misawa (Brookhaven National Laboratory (US))	15m
17:25	Site round table / discussion		35m

INFN – An audience with the Pope

What can the Tecnopolo host?

The computing infrastructures



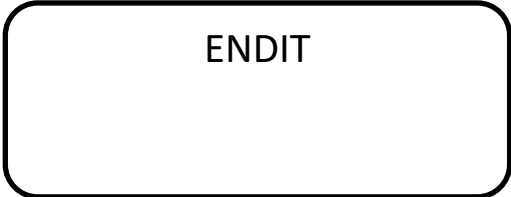
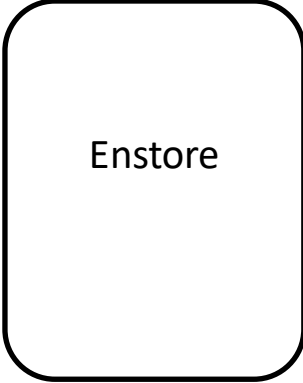
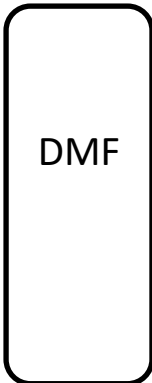
Each of the 6 “botti” (barrels) is
~5000m² of usable IT space



Same architect and design of the
“Sala Nervi” in the Vatican

36

Tape Summary (2023)



Tape Media: LTO or Enterprise

CERN, RAL

DESY

BNL

IN2P3

FZK

SARA

FNAL, PIC JINR

Triumf

NDGF

CNAF



Source: A. Dewhurst pre-GDB Summary Talk

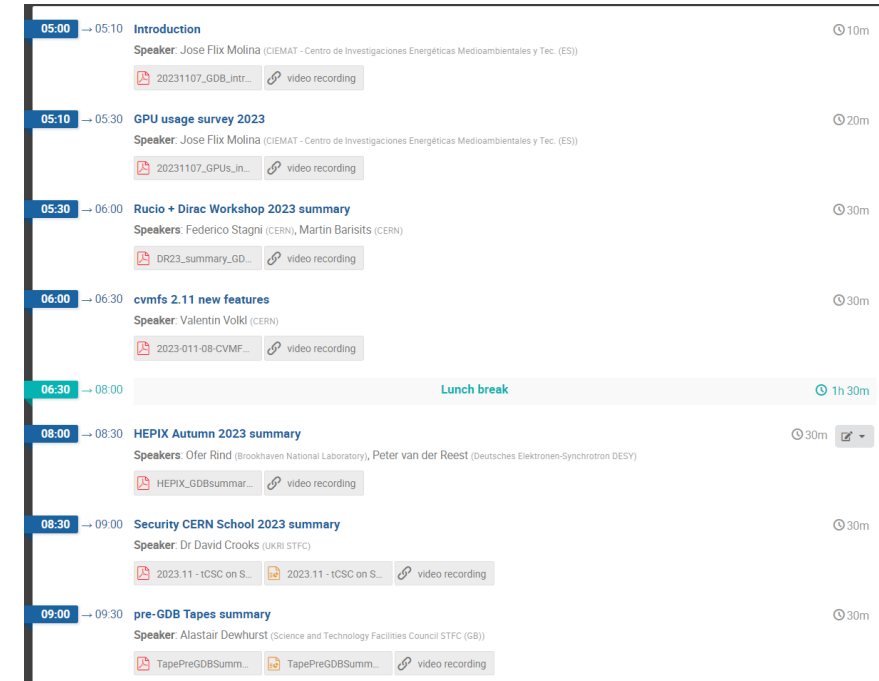
November GDB

- Monthly meeting of the WLCG Grid Deployment Board
- Summary talks on HEPiX, DIRAC & Rucio Workshop, CERN School on Computing Security, and Tapes pre-GDB
- Presentation of results from the 2023 survey on GPU usage

(main) conclusions derived from answers



- Outside HLT farms, **GPU usage is marginal**. No short-term plans to include GPUs at scale from WLCG sites (in general), but **sites offering GPUs help ongoing R&D activities**
- **FPGAs only for online**, no plans for offline yet
- No benchmarks for GPU atm, which affect the accounting → **Benchmark WG on it!**
- Good to start now developing a **GPU pledge framework** within WLCG (related to previous item)
- All of these resources treated as **opportunistic**, for the moment, waiting for guidance from WLCG, though the usage is not yet at scale
- The survey will be **conducted again in 1 year**, to know what's changed

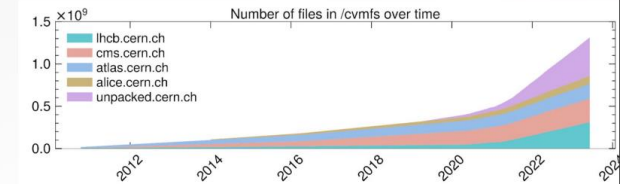


05:00	→ 05:10	Introduction	10m
		Speaker: Jose Flix Molina (CIEMAT - Centro de Investigaciones Energéticas Medioambientales y Tec. (ES))	
		20231107_GDB_intr... video recording	
05:10	→ 05:30	GPU usage survey 2023	20m
		Speaker: Jose Flix Molina (CIEMAT - Centro de Investigaciones Energéticas Medioambientales y Tec. (ES))	
		20231107_GPUs_in... video recording	
05:30	→ 06:00	Rucio + Dirac Workshop 2023 summary	30m
		Speakers: Federico Stagni (CERN), Martin Barisits (CERN)	
		DR23_summary_GD... video recording	
06:00	→ 06:30	cvnfs 2.11 new features	30m
		Speaker: Valentin Volk (CERN)	
		2023-011-08-CVMF... video recording	
06:30	→ 08:00	Lunch break	1h 30m
08:00	→ 08:30	HEPiX Autumn 2023 summary	30m
		Speakers: Ofer Rind (Brookhaven National Laboratory), Peter van der Reest (Deutsches Elektronen-Synchrotron DESY)	
		HEPIX_GDBsummar... video recording	
08:30	→ 09:00	Security CERN School 2023 summary	30m
		Speaker: Dr David Crooks (UKRI STFC)	
		2023-11-tCSC on S... 2023-11-tCSC on S... video recording	
09:00	→ 09:30	pre-GDB Tapes summary	30m
		Speaker: Alastair Dewhurst (Science and Technology Facilities Council STFC (GB))	
		TapePreGDBSumm... TapePreGDBSumm... video recording	

November GDB

- Monthly meeting of the WLCG Grid Deployment Board
- Summary talks on HEPiX, DIRAC & Rucio Workshop, CERN School on Computing Security, and Tapes pre-GDB
- Presentation of results from the 2023 survey on GPU usage
- Nice talk on new features in CVMFS 2.11
 - Few new features, but improvements in logging and performance, as well as bug fixes
 - Interesting collaboration with Jump Trading (new “streaming” cache manager mode bypassing cache except for catalog)
 - Numerous improvements coming to unpacked.cern.ch in CVMFS 2.12

Outlook: unpacked.cern.ch



- Very useful bridge to container deployment model
 - And lower-barrier entry to cvmfs publishing
- Many improvements that will be included in 2.12, following successful summer student project
 - REST API
 - Major refactor
- Can possibly free up some space by garbage-collection campaign

Outlook on possible new features (2.12)

- File Bundles
 - Groups downloads of files that are accessed together
 - Can improve interactive access
- Container tools and ephemeral write shell
 - Helm charts
- Zstd compression

HEPiX Autumn 2023

The [HEPiX forum](#) brings together worldwide information technology staff, including system administrators, system engineers, and managers from High Energy Physics and Nuclear Physics laboratories and institutes, to foster a learning and sharing experience between sites facing scientific computing and data challenges.

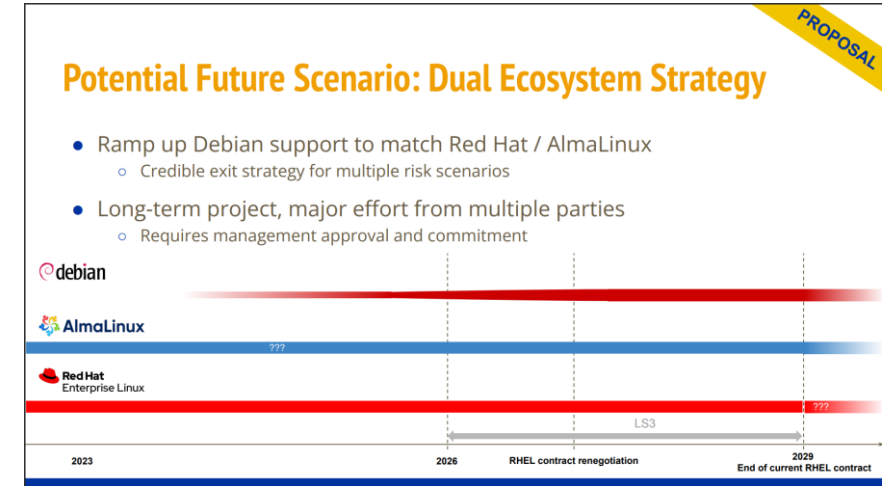


- Oct. 16-20 at Univ. of Victoria
- ~70 attendees, plus over 40 online
- Co-located with LHCOPN/LHCONE
- Presentations organized along seven tracks, all plenary ([SDCC Site Report](#))
 - “A workshop, not a conference”
- Four invited talks on local projects
 - [The Digital Humanities - Open Social Scholarship](#)
 - [Ocean Networks Canada - Multidisciplinary Data from the Deep](#)
 - [The P-One ocean-based neutrino detector](#)
 - [Scaling Digital Research Infrastructure for SKA Astronomy in Canada](#)
- A few vendor presentations from [Weka DDN](#) and [Hypertec](#)
- Discussion on future of Linux distros

Future of Linux Distros



- Community concerns prompted by June 21st RedHat announcement about changes to source availability for RHEL8/9
- CentOS 7 EoL June 30th, 2024
- CERN moving to Alma Linux in addition to RHEL, with Debian likely for some special cases
 - RH site license expires end of May 2029
- DESY, which also supports photon science experiments, currently looking at supporting some combination of Alma, Ubuntu LTS and Debian
 - Trying to identify RHEL dependencies and how to deal with them in the long term – concerns about client apps such as GPFS
- Experience with Alma Linux seem to have been quite positive so far (including at FNAL)
- HEPiX can potentially play an important role supporting community-wide Linux efforts
 - Looking to have Alma developers participating in the Spring HEPiX



CERN IT Linux Strategy, A. Iribarren

What the DESY Linux distro future could look like:

- short-term: Need to upgrade the CentOS Linux 7 machines!!!
 - In some cases, no alternative to RHEL or clones at the moment.
 - RHEL? AlmaLinux? ... ?
- mid/long-term:
 - Migrate everything that has no hard RHEL dependencies to Ubuntu or Debian
 - The photon science, accelerator R&D and theory community probably are OK, and might even welcome this step
 - Where there are dependencies on RHEL:
 - Can they be solved in a secure and sustainable way using containers, without subscriptions?
 - Can we minimize the use of subscriptions to a minimum, e.g. just some portal machines
 - ... probably the LHC analysis code is the hardest part here
 - Moving away from the RedHat world will be a change, not only technical:
 - Long-term-supported, enterprise distros were/are much appreciated

Linux at DESY, Y. Kemp

Some Other Highlights - 1



- More and more Kubernetes: [Deploying dCache](#), [Centralized log management at Diamond Light Source](#), [UNL Analysis Facility](#) and even using it as a platform for running an entire Tier-2 site at [UVic](#) and [NET2](#).
- Deploying and testing ARM ([Testing for WLCG](#)) and updates from the [HEPiX benchmarking working group](#) including new features like addition of time series plugin for recording energy consumption (among other metrics), progress on GPU workloads and monitoring performance via Panda

GridPP
UK Computing for Particle Physics

Schedule

- ATLAS are finishing up, for now.
- CMS are setting up; will have access to ~1500 cores for 3-4 weeks.
- ALICE will follow after that.
- LHCb will have the opportunity by the end of the year.
- After that? Target as requested, though in the long-term we are mainly an ATLAS site.
- Once all four experiments have evaluated their initial runs (~1million core-hours) we propose a joint meeting; possibly a pre-GDB (in February?).
- The important discussion is about whether ARM resources can be pledged in September 2024 to deliver resource in 2025.
- Our aim is to provide the experiments with the opportunity to evaluate *using* heterogenous resources on the Grid; and to assess how to minimise the additional effort needed to *provide* heterogenous resources.

David Britton, University of Glasgow

Arm for WLCG, D. Britton

NET2 is a US-ATLAS site with a pure OKD cluster

ATLAS EXPERIMENT

Not a traditional system

- No CE, Condor, PSB, or Slurm
- OKD: the community distribution of Kubernetes that serves as the upstream project for Red Hat OpenShift.
- Direct submissions to Kubernetes API;

Now in production

- Request for new queue: Sep/20
- First HammerCloud jobs: Sep/27
- Different types of jobs: Sep/28
- Began production: Sep/29

From opening the queue to production in less than 2 weeks.

NET2 operates completely as an OKD cluster. No subclusters!

Oct 17th, 2023

Eduardo Bach - HEPiX Autumn 2023

NET2: a first example of OpenShift/OKD for Tier 2 provisioning and cluster management in US ATLAS, E. Bach

Studies on Grid Jobs

- Ran HEPscore23 via PanDA to [continuously measure grid site performance](#)
- Was utilized to determine some underperforming sites, including one where HT was incorrectly disabled
 - The group worked with site admin on correcting: resulted in a 66% performance improvement

HEPscore23 via PanDA

- We are running on 134 different Panda Resources (Queues)
- INFN, CERN, CA-VICTORIA, DESY-FAH, JINR, Vega...
- Infrastructure:
 - PanDA, HammerCloud, Rucio, ActiveMQ, Elasticsearch, Grafana, Kibana...

HS23 statistics before and after changes

66% performance improvement of site



HEPiX Benchmarking Working Group Report, C. Hollowell

Some Other Highlights - 2

- A couple of talks looking at new computational techniques
 - Using [quantum-assisted generative models to speed up calorimeter simulation](#)
 - A nice [review](#) of interactive AI/LLM tools and their uses for code development and on local data at Saclay
- A few talks on tools and techniques for enhancing site security, as well as an update on [establishing trust and security policies for research infrastructures](#)
- Storage and Filesystems track investigations of [HPSS disk cache performance tuning](#) at KIT, [improvements to ENDIT](#) at NDGF, and an interesting look at the [status of Ceph in 2023](#) by Dan van der Ster from Clyso
- Plus nine talks on facility infrastructure, including improving energy efficiency, supporting EoL experiments, and the myriad challenges of day-to-day operations of data centers.

Results

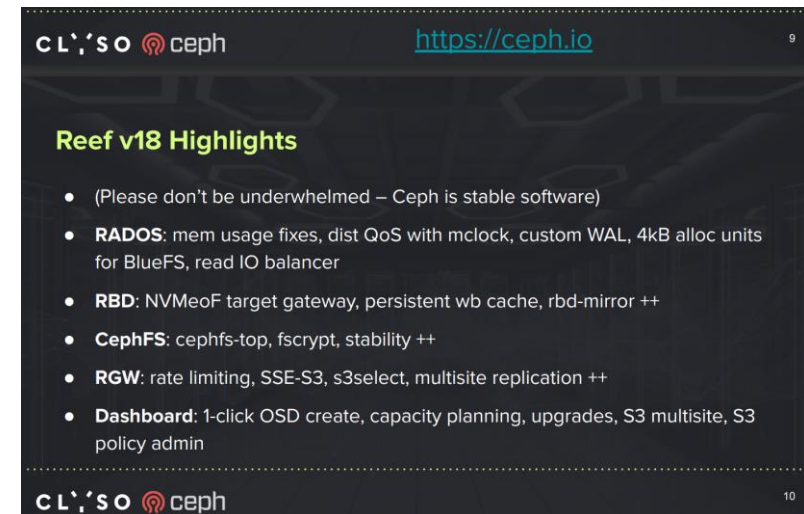
Wall time to generate 1024 samples	
Geant4	$\sim 1000\text{ s}$
GPU A100	$2.19 \pm 0.14\text{ s}$
QPU	$\sim 0.180\text{ s}$



QPU $\sim 12\times$ faster than GPU

QPU $\sim 10^4\times$ faster than Geant4

Quantum Assisted Calorimeter Simulation, J. Quetzalcoatl Toledo Marín



Ceph in 2023 and Beyond, D. van der Ster