Distributed Computing Challenges at the LHC and HL-LHC

Marian Babik, CERN
4th GLOBAL RESEARCH PLATFORM WORKSHOP
Co-Located with IEEE International Conference On eScience 2023
October 9-10, 2023

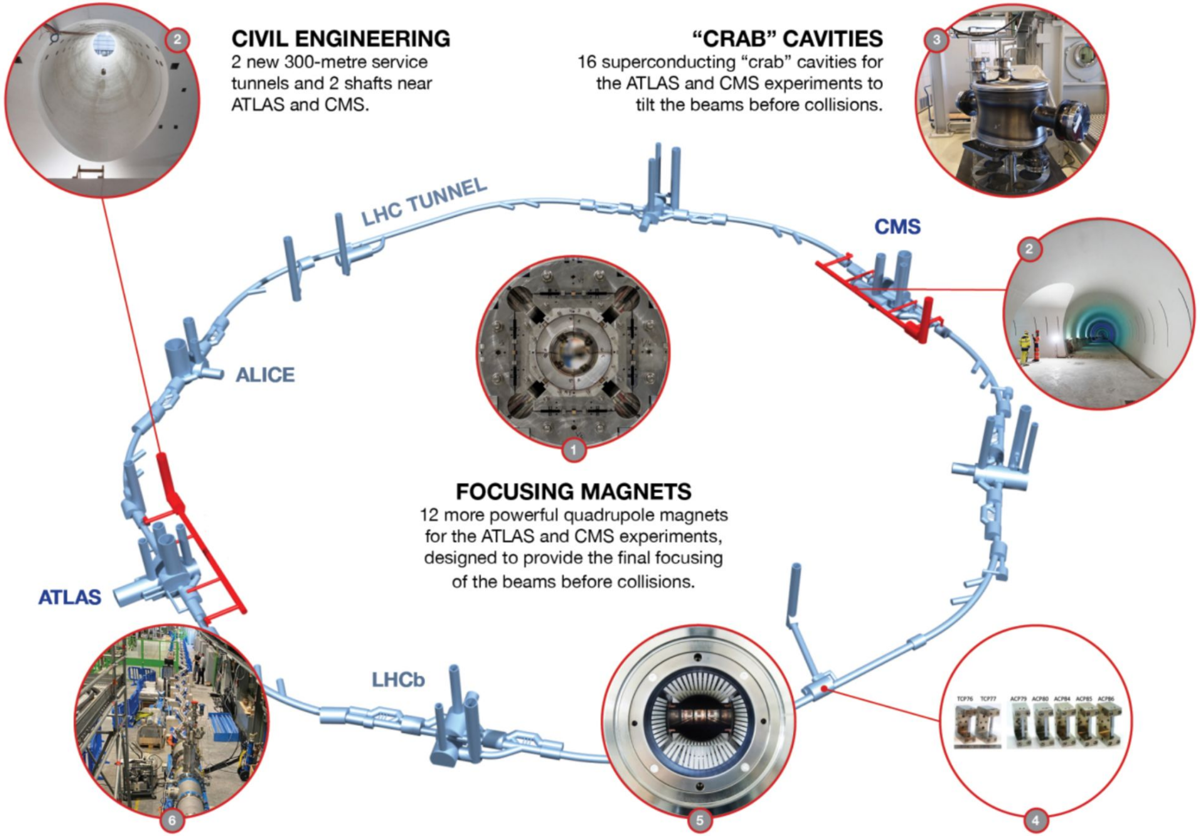
Introduction

For HEP software and computing the time horizon of future challenges is the next 15 years

The main contributor to those challenges is HL-LHC, both in terms of volume and complexity. The largest needs come from ATLAS and CMS

The LHC computing resources are provided by the WLCG infrastructure. Other HEP experiments will share a large part of such an infrastructure. Other sciences will use many of the same facilities

NEW TECHNOLOGIES FOR THE HIGH-LUMINOSITY LHC



CIVIL ENGINEERING
2 new 300-metre service tunnels and 2 shafts near ATLAS and CMS.

"CRAB" CAVITIES
16 superconducting "crab" cavities for the ATLAS and CMS experiments to tilt the beams before collisions.

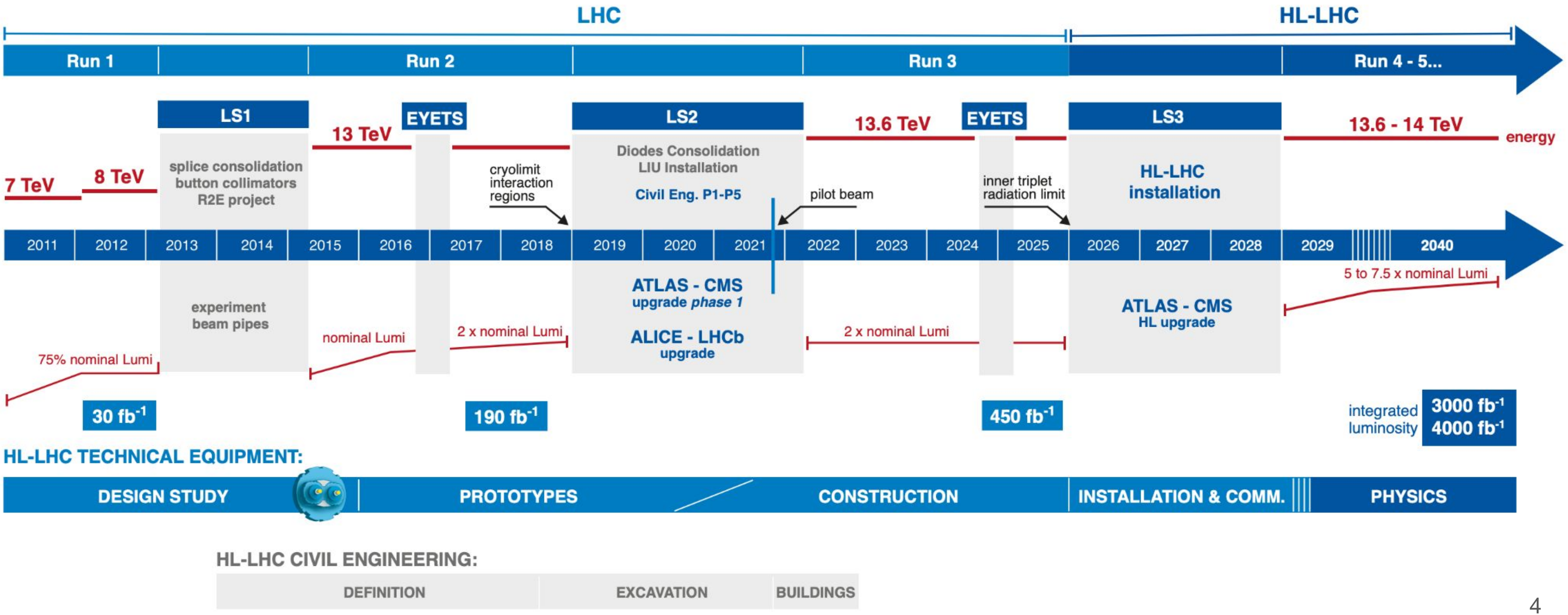
FOCUSING MAGNETS
12 more powerful quadrupole magnets for the ATLAS and CMS experiments, designed to provide the final focusing of the beams before collisions.

SUPERCONDUCTING LINKS
Electrical transmission lines based on a high-temperature superconductor to carry the very high DC currents to the magnets from the powering systems installed in the new service tunnels near ATLAS and CMS.

COLLIMATORS
15 to 20 additional collimators and replacement of 60 collimators with improved performance to reinforce machine protection.

CRYSTAL COLLIMATORS
New crystal collimators in the IR7 cleaning insertion to improve cleaning efficiency during operation with ion beams.

LHC / HL-LHC Plan



HL-LHC Computing Roadmap

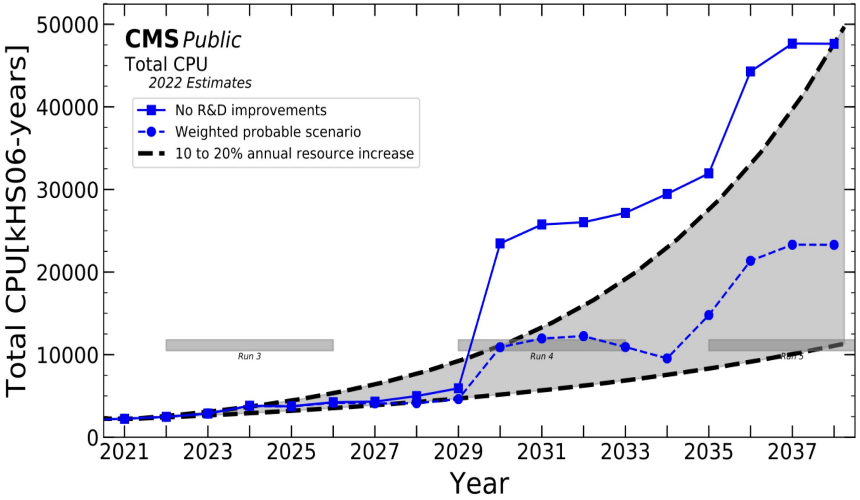
HEP Software Foundation Community Whitepaper: a bottom-up exercise. Identify the areas of work to address the HEP challenges of the 2020s

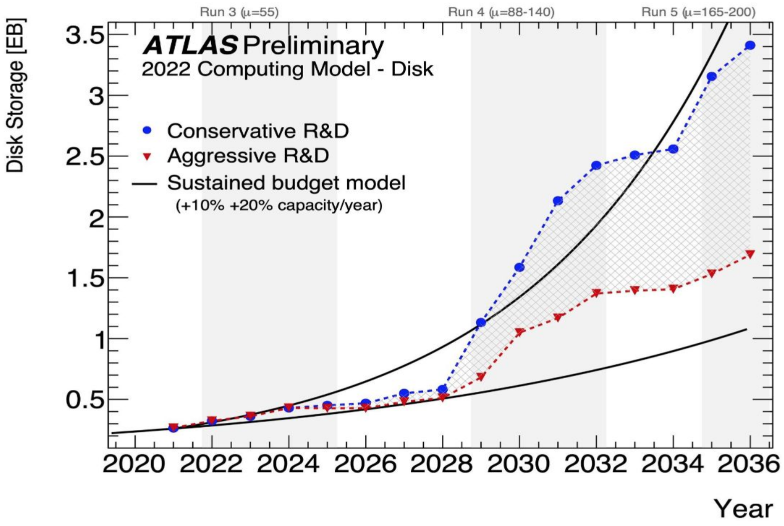
The first WLCG strategy toward HL-LHC document: a top-down high-level prioritization of the whitepaper, for the LHC needs

The LHCC review series of HL-LHC computing: a multistep process tracking the progress towards HL-LHC

- May 2020: review of ATLAS and CMS plans, Data Management (DOMA), offline software, the WLCG collaboration and infrastructure. Documents
- November 2021: update from ATLAS and CMS, common software activities (generators, simulation, foundation software, analysis, DOMA). Report

ATLAS and CMS needs for HL-LHC





The gap between available and needed resources is filling up, assuming the main R&D activities are successful

Investing in further (identified) R&D activities would fill this gap further. Need more effort

There are still large uncertainties

Networking in HL-LHC

Networking will play a central role in HL-LHC as enabler for HEP computing

- Support the core functions of WLCG (data acquisition/archival/processing)
- Provide more flexibility to the computing models, allowing to optimise

WLCG continues engaging with Funding Agencies and NRENs to ensure that enough capacity is made available and the LHC traffic does not get segregated below a critical level.

Several R&Ds were launched to study how to better leverage the network resources in the data and processing infrastructures for HL-LHC

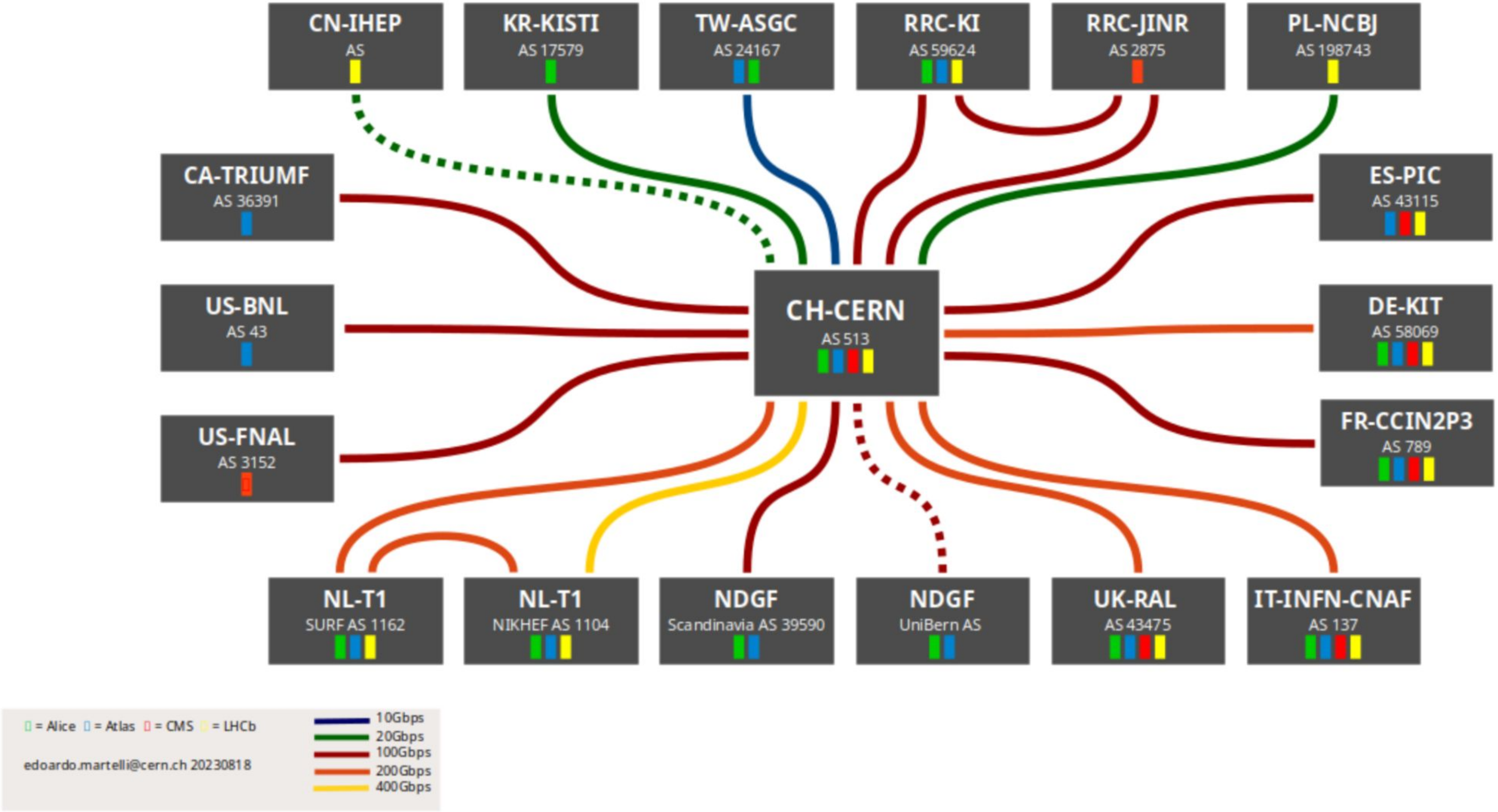
- Regularly discussed at the LHCONE/LHCOPN meetings

The LHCC sees the strategic role of networks and asked for regular updates.

Networking R&D roadmap

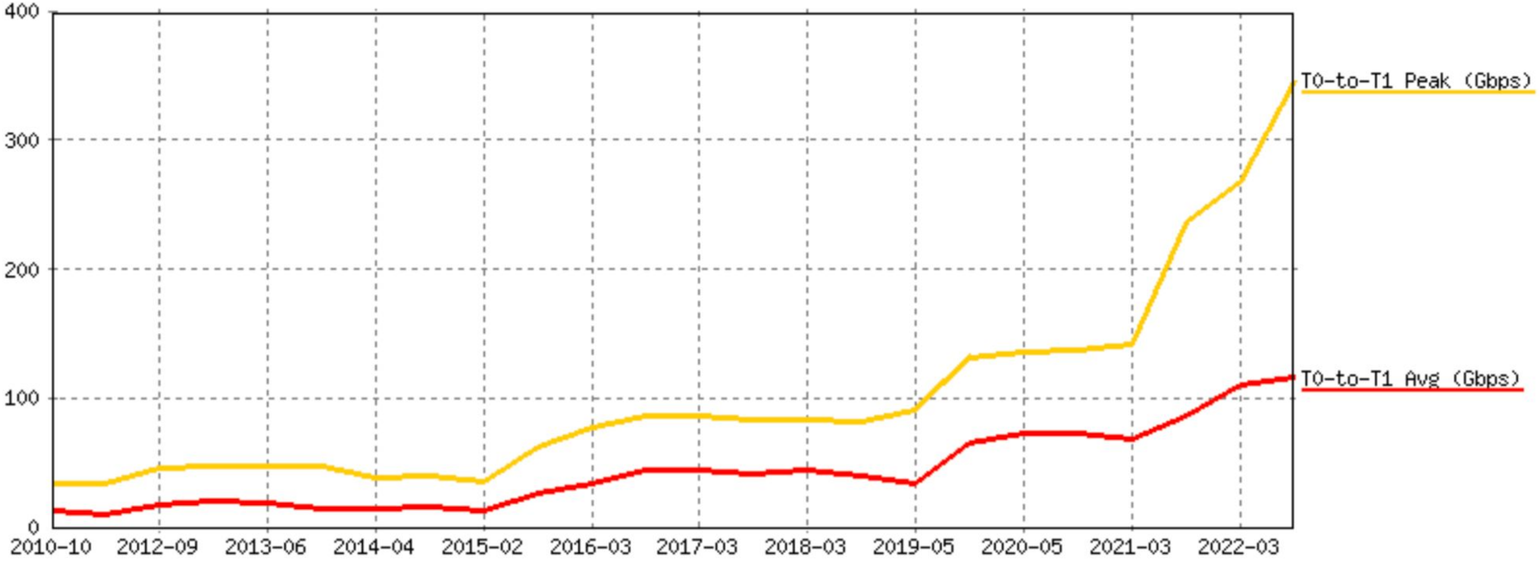
- HEPiX Network Functions Virtualisation Working Group
  - Working Group Report was published at the end of 2019 with three chapters
    - Cloud Native DC Networking
    - Programmable Wide Area Networks
    - Proposed Areas of Future Work
- LHCOPN/LHCONE workshop (spring 2020)
  - Requirements on networks from the WLCG experiments
- Research Networking Technical Working Group
  - Formed after the workshop in response to the requirements discussion
  - 98 members from ~ 50 organisations have joined
  - Three main areas of work:
    - **Network Visibility & Analytics**
    - **Network Performance** - software improvements (pacing, congestion algorithms)
    - **Network Orchestration** - followed up by NOTED, GNA-G, SENSE and FABRIC

LHCOPN


CN-IHEP
AS
KR-KISTI
AS 17579
TW-ASGC
AS 24167
RRC-KI
AS 59624
RRC-JINR
AS 2875
PL-NCBJ
AS 198743
CA-TRIUMF
AS 36391
ES-PIC
AS 43115
US-BNL
AS 43
CH-CERN
AS 513
DE-KIT
AS 58069
US-FNAL
AS 3152
FR-CCIN2P3
AS 789
NL-T1
SURF AS 1162
NL-T1
NIKHEF AS 1104
NDGF
Scandinavia AS 39590
NDGF
UniBern AS
UK-RAL
AS 43475
IT-INFN-CNAF
AS 137
= Alice = Atlas = CMS = LHCb
edoardo.martelli@cern.ch 20230818
10Gbps
20Gbps
100Gbps
200Gbps
400Gbps

LHCOPN Long-term Growth

LHCOPN network traffic from CERN Tier0 to all the aggregated Tier1s



Run1: 2010-12 LS1:2013-14 Run2: 2014-2018 LS2: 2019-21 Run3: 2022

Y-Axis: Gbps (Giga bit per second)

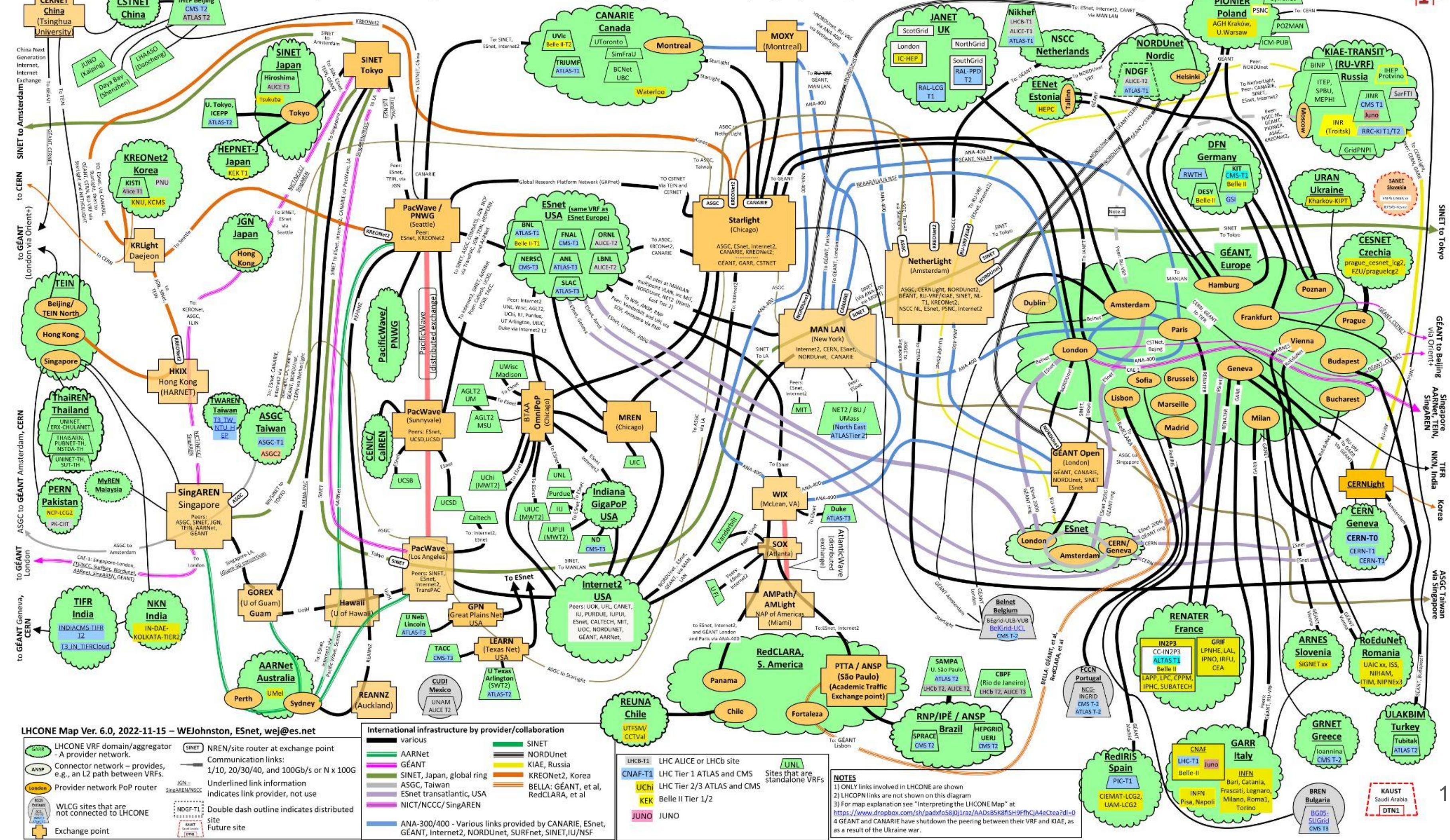
Out: direction Tier0 to all Tier1s

Avg: average network bandwidth on the previous 12 months

Peak maximum peak network bandwidth on the previous 12 months

LHCONE

LHCONE L3VPN: A global infrastructure for High Energy Physics data analysis (LHC, Belle II, Pierre Auger Observatory, NOvA, XENON, JUNO)



WLCG Data Challenges

The data challenges are an incremental **process** to prepare for the HL-LHC network needs, through a regular dialog between the network providers, the experiments and the facilities.

We identified the main use cases at HL-LHC in terms of network use (RAW data export and reprocessing), for the 4 LHC experiments

We estimated the network needs including contingency and considering different scenarios

We set metrics and intermediate targets to be progressively challenged

The challenges offer the possibility to bring in production many network R&D activities

Data Challenge 2021: Data rate table

ATLAS & CMS T0 to T1 per experiment

- **350PB RAW**, taken and distributed during typical LHC uptime of 7M seconds / 3 months (50GB/s aka. 400Gbps)
- Another 100Gb/s estimated for prompt reconstruction data (AOD, other derived output)
- In total approximately 1Tbps for CMS and ATLAS together

ALICE & LHCb

- 100 Gbps per experiment estimated from Run-3 rates

Minimal model

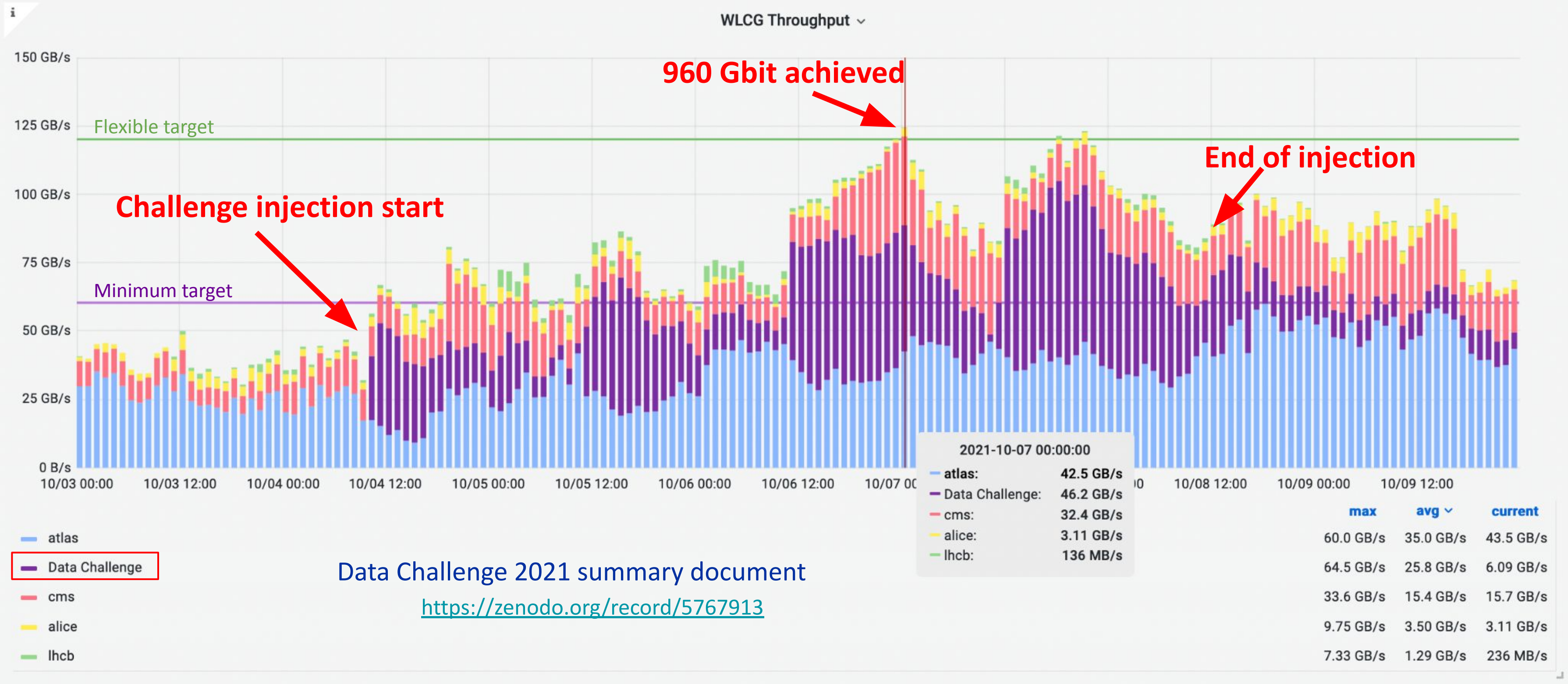
- ∑ (ATLAS,ALICE,CMS,LHCb) *2 (for bursts) *2 (overprovisioning) = **4.8Tbps**

Flexible model

- Assumes reading of data from above for reprocessing/reconstruction within 3 months
- Means doubling the Minimal Model: **9.6Tbps**
- However data flows from the T1s to T2s and T1s!

No MC production flows nor re-creation of derived data in the 2021 modelling!

DC21 goal: 10% of HL-LHC



Data Challenge 2021 summary document

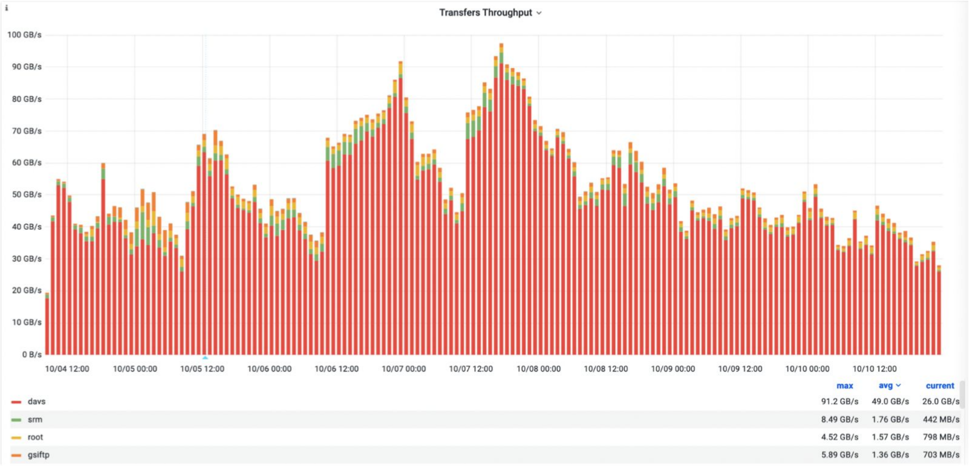
https://zenodo.org/record/5767913

DC21: Technologies

The data challenges are not just about throughput but also functionality

In 2021 they provided an opportunity to commission new features that are now in use during Run-3.

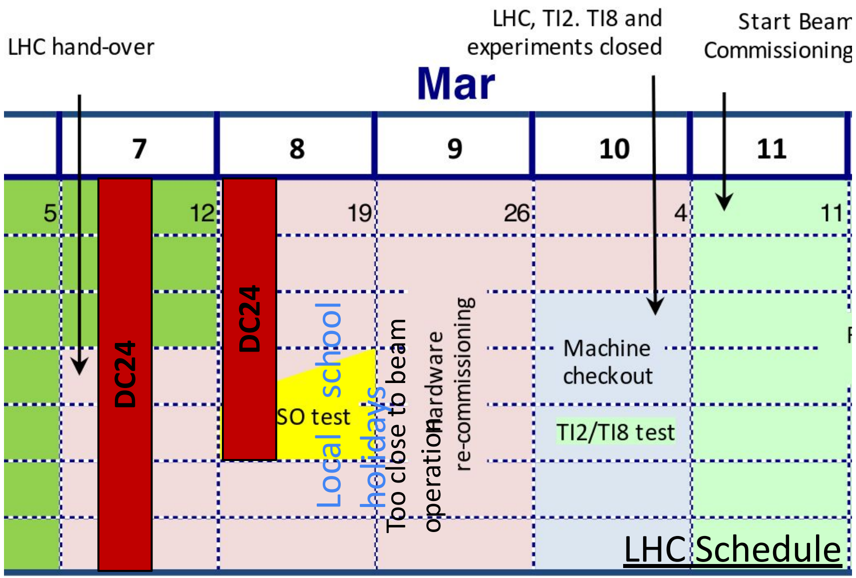
- For example the HTTP protocol (replacing gridFTP) for asynchronous transfers



Traffic mostly through HTTP (**RED**)

Data Challenge 2024

- Final DC24 dates: **February 12-23** (approved by WLCG MB)
  - Original DC24 proposal for MB (25% target)
  - Planning documents with DC24 proposals
- Organization and communication
  - WLCG DOMA General meetings
  - DC24 Workshop @ CERN
    - November 9-10 (Thu, Fri), 2023
      - register before November 3rd
    - after (pre-)GDB focused on tapes



DC24 ATLAS Rates

ATLAS DC24 transfer rates

(preliminary version: 20230926)

Final T2 ingress/egress depends on number of participating T2 sites and might be in given range

rows in red color: sites must explicitly ask be included in DC24 (details will be sent to all-clouds list)

Deletion rates are calculated from ingress bandwidth assuming 3GB average filesize)

Table: DC24 (src: ingress / egress)			Site WAN (Gb/s)		DC24 minimal scenario				DC24 flexible scenario			
			Total (Gb/s)	Usable by ATLAS	T0 Export	Total Gb/s & bandwidth		Space [TB/24h] (deletions/hour)	T0 Export	Total Gb/s & bandwidth		Space [TB/24h] (deletions/hour)
Site	Tier	Cloud				∑ ingress	∑ egress			∑ ingress	∑ egress	
CERN-PROD	T0	CERN	2100	911	270.0	27.9	291.3	0 (0k)	270.0	93.1 - 112.2	363.1	884 (13k)
T0 summary					270.0	27.9	291.3		270.0	93.1 - 112.2	363.1	
BNL-ATLAS	T1	US	400	400	60.0	82.2	60.0	764 (11k)	60.0	107.5 - 119.6	120.0	1089 (15k)
FZK-LCG2	T1	DE	400	162	32.0	61.7	32.0	431 (6k)	32.0	86.3 - 100.3	64.0	911 (13k)
IN2P3-CC	T1	FR	200	93	33.0	53.3	33.0	413 (6k)	33.0	**81.6 - 95.8**	**66.0**	861 (12k)
INFN-T1	T1	IT	300	81	24.0	39.5	24.0	319 (5k)	24.0	54.8 - 64.0	48.0	588 (8k)
NDGF-T1	T1	ND	200	157	16.0	30.7	21.8	151 (2k)	16.0	77.9 - 96.6	32.0	842 (12k)
SARA-MATRIX	T1	NL	400	291	15.0	30.4	15.0	192 (3k)	15.0	54.4 - 66.0	30.0	604 (9k)
pic	T1	ES	200	89	13.0	21.4	13.0	170 (2k)	13.0	29.1 - 34.4	26.0	319 (5k)
RAL-LCG2	T1	UK	400	196	39.0	60.6	39.0	464 (7k)	39.0	88.5 - 100.1	78.0	861 (12k)
RRC-KI-T1 (no active T0 export)	T1	RU	200	79	8.0	13.4	8.0	109 (2k)	8.0	15.1 - 17.2	16.0	160 (2k)
TRIUMF-LCG2	T1	CA	100	100	30.0	45.9	30.0	403 (6k)	30.0	60.8 - 69.7	60.0	643 (9k)
T1 summary					270.0	439.3	275.8		270.0	655.9 - 763.8	540.0	
T2 summary						213.1	107.2			574 - 759	420 - 732	
Summary						680.4	674.2			1323 - 1635	1323 - 1635	

DC24: Networking R&D

NOTED

- Monitor link saturation and predict the behaviour of the applications
- When NOTED detects that the link is going to be congested provides a dynamic circuit using AutoGOLE/SENSE
- Ongoing work in decision making, improving the forecasts, monitoring integration, FTS integration

AutoGOLE/SENSE

- End-to-end service to dynamically procure VPNs between routers to enforce a given path
- Implement network QoS to prioritise transfers at the router level
- More details later today in Session 3: Orchestration Among Multiple Domains (J. Mambretti)

ALTO/TCN

- Application-Layer Traffic Optimization provides means to to obtain network information
- Exploit this network information in higher-level long-term schedules (FTS / Rucio)
- More details later today in Session 3: Orchestration Among Multiple Domains (R. Yang)

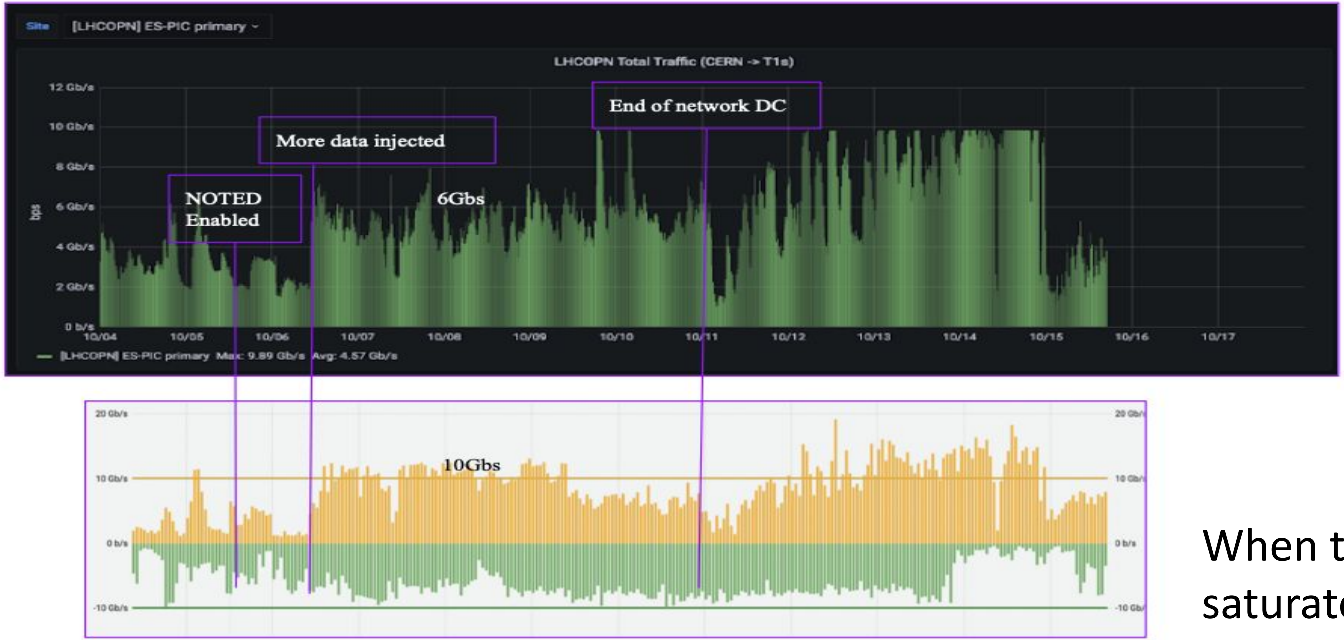
Packet marking/SCITAGS (network visibility)

- Identify traffic at the network layer (experiment and activity)
- More details tomorrow in Session 4: High-Fidelity Data Flow Monitoring, etc.

Network throughput studies

- Packet pacing (BBRv3, TC)
- Jumbo frames

NOTED: CERN-PIC



NOTED is a Software Defined Network R&D project to share network traffic between different paths

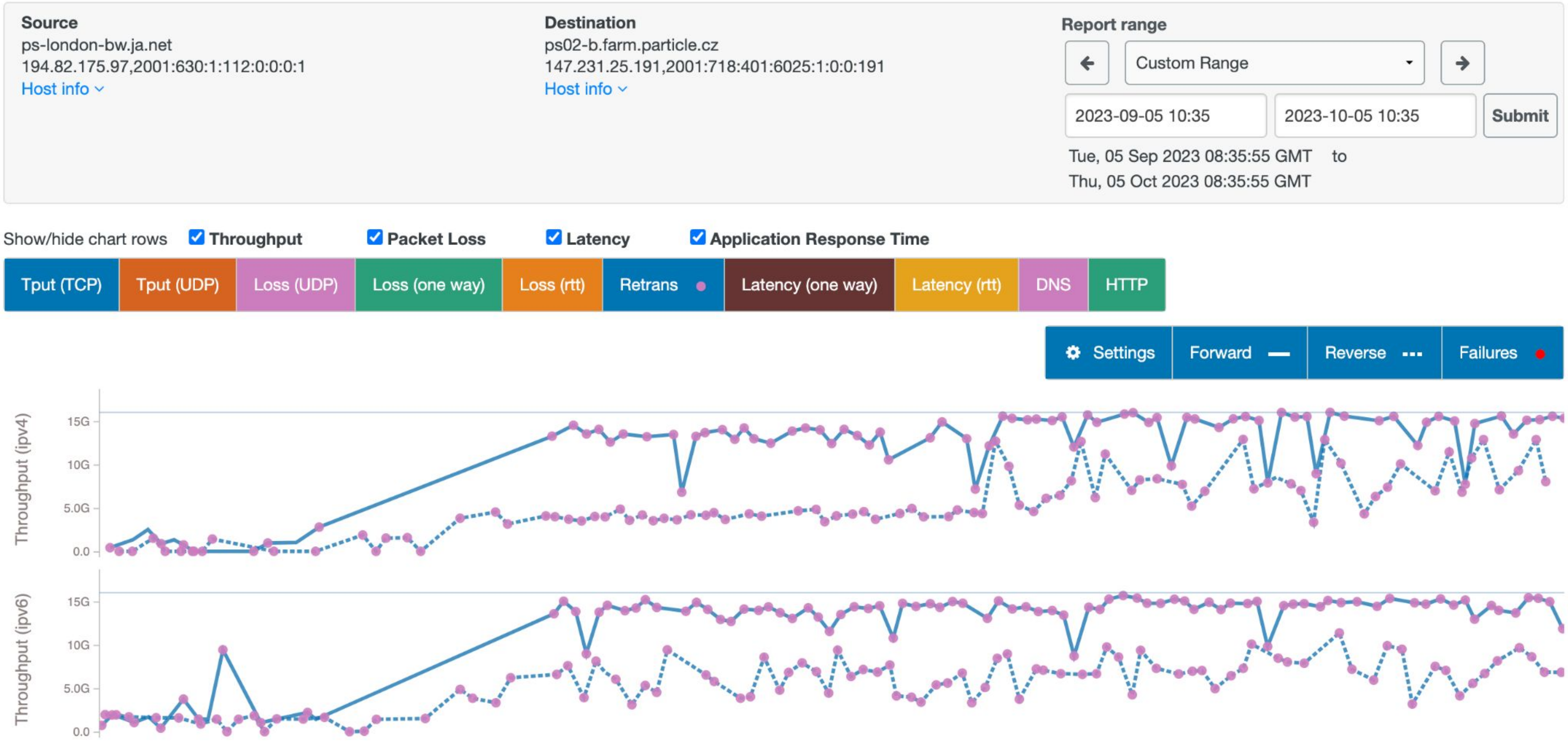
Enabled during the data challenge between CERN and PIC

When the 6 Gbps LHCOPN link saturated, NOTED added the LHCONE link to complement it.

10Gbps target reached

20

Network Throughput: BBRv3



From packet pacing meeting last week - https://indico.cern.ch/event/1329666/

DC24: Technologies and Collaboration

Token based authentication for data transfers

- Decide about porting features of GsiFTP to Http/WebDAV (e.g. multi-stream) if necessary
- Coordinate timeline with WLCG AuthZ working group

Tape REST API

- Roll out plan for all T1s

WLCG data transfer monitoring

- Focus Xrootd monitoring deployment initially at CERN and FNAL

Collaboration beyond LHC experiments

- Foster exchange with "close" projects, Belle-2, DUNE, SKA

Collaboration with other HEP experiments

WLCG presented a joint paper with DUNE and Belle-2 to the Snowmass 2021 process

The paper presents the strategic directions to address the computing challenges of the experiments in the next decade. It complements the WLCG contribution to the European Strategy for Particle Physics in 2019

Snowmass 2021

- Consolidation of the WLCG scientific computing infrastructure
- Evolution of such an infrastructure to integrate modern technologies and facilities
- **Broadening the scope of the WLCG collaboration to create partnership with other HEP experiments**

Physics > Computational Physics

[Submitted on 14 Mar 2022]

HEP computing collaborations for the challenges of the next decade

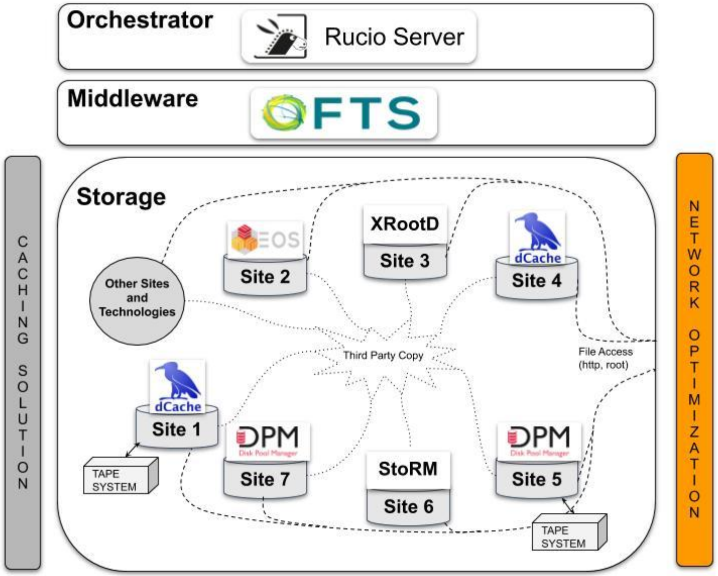
Simone Campana, Alessandro Di Girolamo, Paul Laycock, Zach Marshall, Heidi Schellman, Graeme A Stewart

Large High Energy Physics (HEP) experiments adopted a distributed computing model more than a decade ago. WLCG, the global computing infrastructure for LHC, in partnership with the US Open Science Grid, has achieved data management at the many-hundred-Petabyte scale, and provides access to the entire community in a manner that is largely transparent to the end users. The main computing challenge of the next decade for the LHC experiments is presented by the HL-LHC program. Other large HEP experiments, such as DUNE and Belle II, have large-scale computing needs and afford opportunities for collaboration on the same timescale. Many of the computing facilities supporting HEP experiments are shared and face common challenges, and the same is true for software libraries and services. The LHC experiments and their WLCG- partners, DUNE and Belle II, are now collaborating to evolve the computing infrastructure and services for their future needs, facilitated by the WLCG organization, OSG, the HEP Software Foundation and development projects such as HEP-CCE, IRIS-HEP and SWIFT-HEP. In this paper we outline the strategy by which the international HEP computing infrastructure, software and services should evolve through the collaboration of large and smaller scale HEP experiments, while respecting the specific needs of each community. We also highlight how the same infrastructure would be a benefit for other sciences, sharing similar needs with HEP. This proposal is in line with the OSG/WLCG strategy for addressing computing for HL-LHC and is aligned with European and other international strategies in computing for large scale science. The European Strategy for Particle Physics in 2020 agreed to the principles laid out above, in its final report.

Comments: contribution to Snowmass 2021
Subjects: **Computational Physics (physics.comp-ph)**
Cite as: arXiv:2203.07237 **[physics.comp-ph]**
(or arXiv:2203.07237v1 **[physics.comp-ph]** for this version)
https://doi.org/10.48550/arXiv.2203.07237

Today DUNE,Belle-2 and JUNO are WLCG “observers” and share many services with WLCG (including some LHCOPN/LHCONE networks)

Collaboration with other sciences

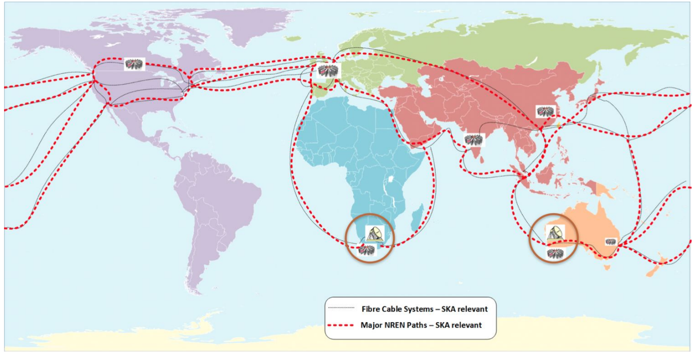


The ESCAPE project implemented a prototyped a data infrastructure prototype across Europe

- based on many of the WLCG building blocks and on top of many WLCG facilities

Examples of prototyped applications:

- SKA: data delivery from Perth and Cape Town to Europe and access through the data lake services



Experiments such as SKA will in future need a similar bandwidth as HL-LHC, sharing many of the network paths

It is now the opportunity to discuss how LHC, other HEP experiments and other sciences will cohabit

Acknowledgments

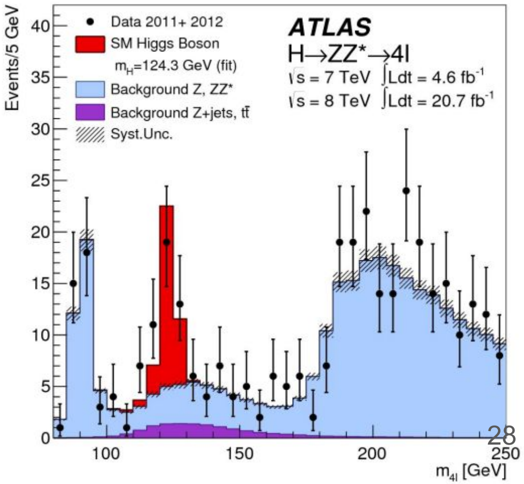
I'd like to thank Simone Campana (CERN), Christoph Wissing (DESY), Mario Lassnig (CERN) and Petr Vokac (CVUT, Prague) for their help and support in preparing this talk.

Questions ?

Backup slides

The Large Hadron Collider (LHC)@CERN

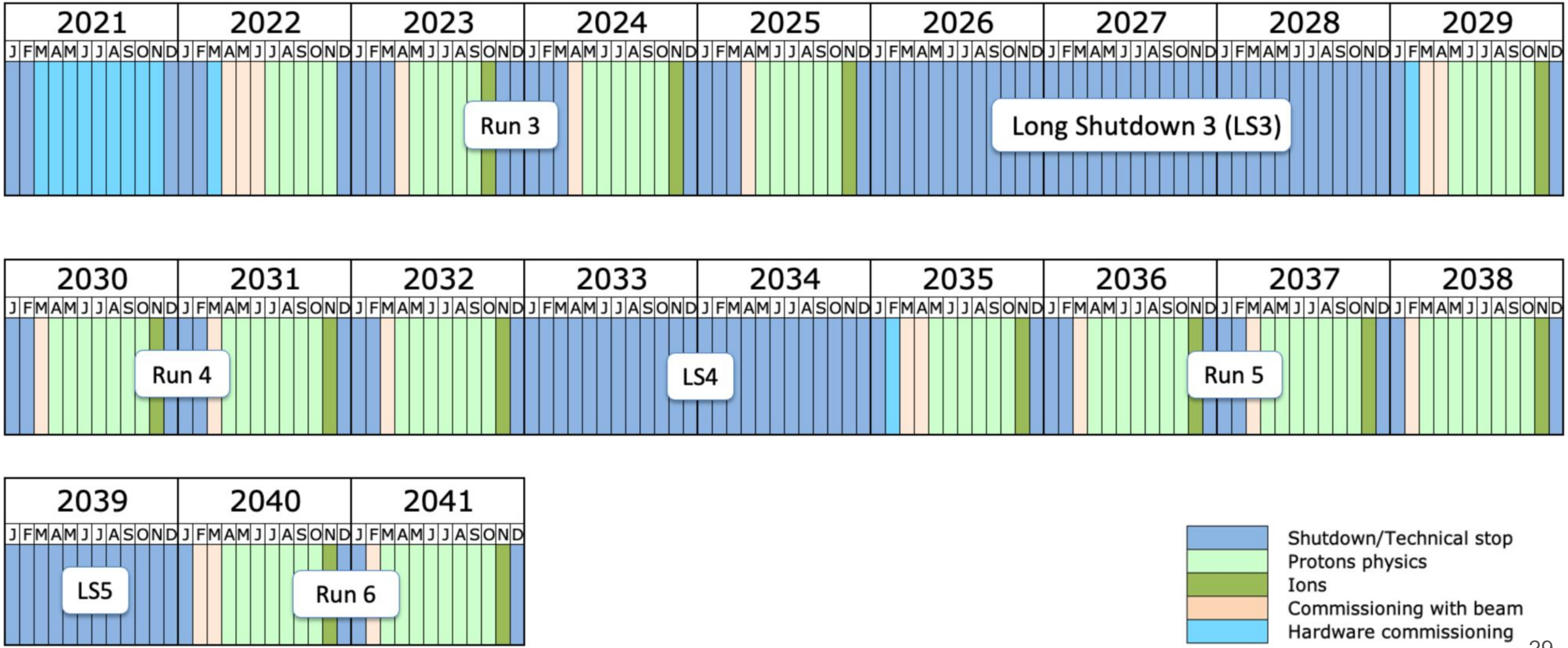
- pp (or Pb-Pb) collisions
- 4 experiments (ATLAS, CMS, LHCb, ALICE)
- Discovery of Higgs boson
  - Nobel prize for physics 2013



LHC Schedule

Longer term LHC schedule

In January 2022, the schedule was updated with long shutdown 3 (LS3) to start in 2026 and to last for 3 years. HL-LHC operations now foreseen out to end 2041.



Last update: April 2023

DC24 Rates

T1 Sites (T0 export / T1→T2 reco)	HL-LHC Minimal Scenario [Gbps]	HL-LHC Flexible Scenario [Gbps]	DC27 (100%) [Gbps]	DC26 (~~60~~→50%) [Gbps]	DC24 (25%) [Gbps]	DC24 ATLAS [Gbps]	DC24 CMS [Gbps]	DC24 Alice [Gbps]	DC24 LHCb [Gbps]	~~DC23 (30%) [Gbps]~~	DC21 (10%) [Gbps]
CA-TRIUMF	200	400	100	60	30	30	0	0	0	~~30~~	10
DE-KIT	600	1200	300	180	80	32	26	11	11	~~90~~	30
ES-PIC	200	400	100	60	30	13	13	0	3	~~30~~	10
FR-CCIN2P3	570	1140	290	170	70	33	21	7	9	~~90~~	30
IT-INFN-CNAF	690	1380	350	210	90	24	35	14	16	~~100~~	30
KR-KISTI-GSDC	50	100	30	20	10	0	0	10	0	~~10~~	0
NDGF	140	280	70	40	20	16	0	4	0	~~20~~	10
NL-T1	180	360	90	50	20	15	0	1	4	~~30~~	10
NRC-KI-T1	120	240	60	40	20	8	0	8	4	~~20~~	10
UK-T1-RAL	610	1220	310	180	80	39	21	1	18	~~90~~	30
RU-JINR-T1	200	400	100	60	30	0	30	0	0	~~30~~	10
US-T1-BNL	450	900	230	140	60	60	0	0	0	~~70~~	20
US-FNAL-CMS	800	1600	400	240	100	0	100	0	0	~~120~~	40
(transatlantic link)	1250	2500	630	380	160	60	100	0	0	~~190~~	60
Sum	**4810**	**9620**	**2430**	**1450**	**640**	270	**246**	**56**	**65**	~~**730**~~	**240**

Non-LHC Participation in DC24

Interest in the wider HEP community to join DC24
- Belle II, DUNE, JUNO
- Perhaps SKA (radio astronomy)
- Involved sites are often supporting also LHC experiments

Overall traffic from non-LHC expected to be small compared to LHC
- Parts of the traffic going through LHCONE networks
- Direction often in the opposite direction, e.g.
  - LHC RAW data: From Europe to US and Asia
  - DUNE: From US to Europe (and Asia)
  - Belle II & Juno: From Asia to Europe and US

Monitoring
- Good common(!) monitoring of LHC traffic already challenging
- Common dashboard with non-LHC experiments would be great, but quite some effort
- However (low level) monitoring of network providers should show these activities

Data rate complexity

Data rate experience from DC21

Higher complexity of data flows than assumed has become evident

Include feedback from the experiments and the network community

Mixing of ingress/egress values was very confusing

More complex setup has three major data flows

RAW export, prompt reconstruction/derivation export …	Tier-0 to Tier-1	Unidirectional
Reconstruction, Reprocessing, Simulation, Derivations, …	Tier-1+2 to Tier-1+2	Bi-directional
Data consolidation, recovery operations, …	Tier-1+2 to Tier-1+2	Bi-directional

Assume the following steps

2021 → 10%
2024 → 25%
2026 → 50%
2028 → 100%