The Worldwide LHC Computing Grid

and the role of Tier-2 sites in this ecosystem: an overview

Dagmar Adamova (NPI AS CR Prague/Rez)

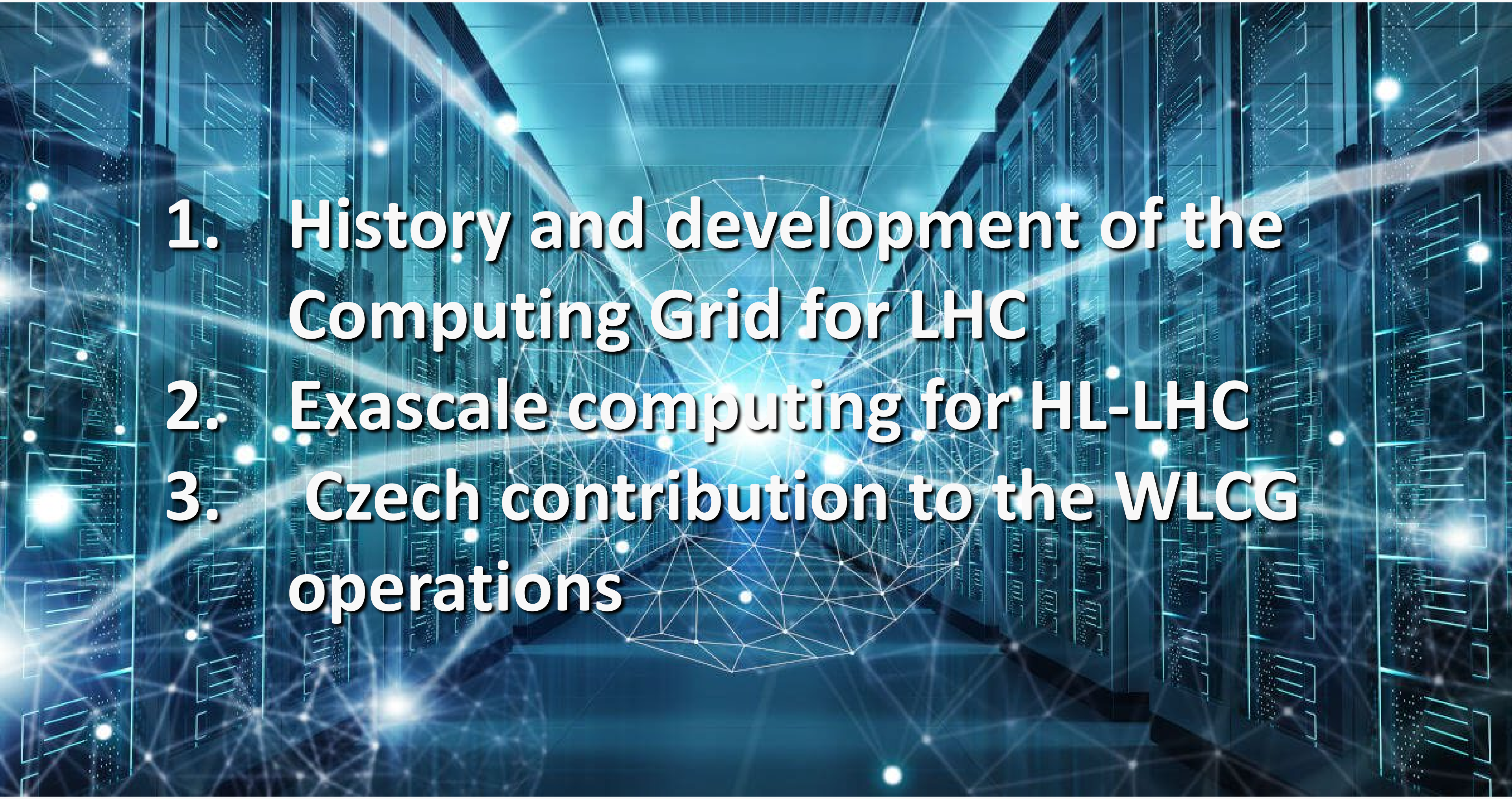
Yerevan, 14.09.2023

Disclaimer

Since I lack the information how much is a knowledge of computing infrastructures for data intensive Physics experiments distributed among the audience, I tried to keep the presentation without much technical details.

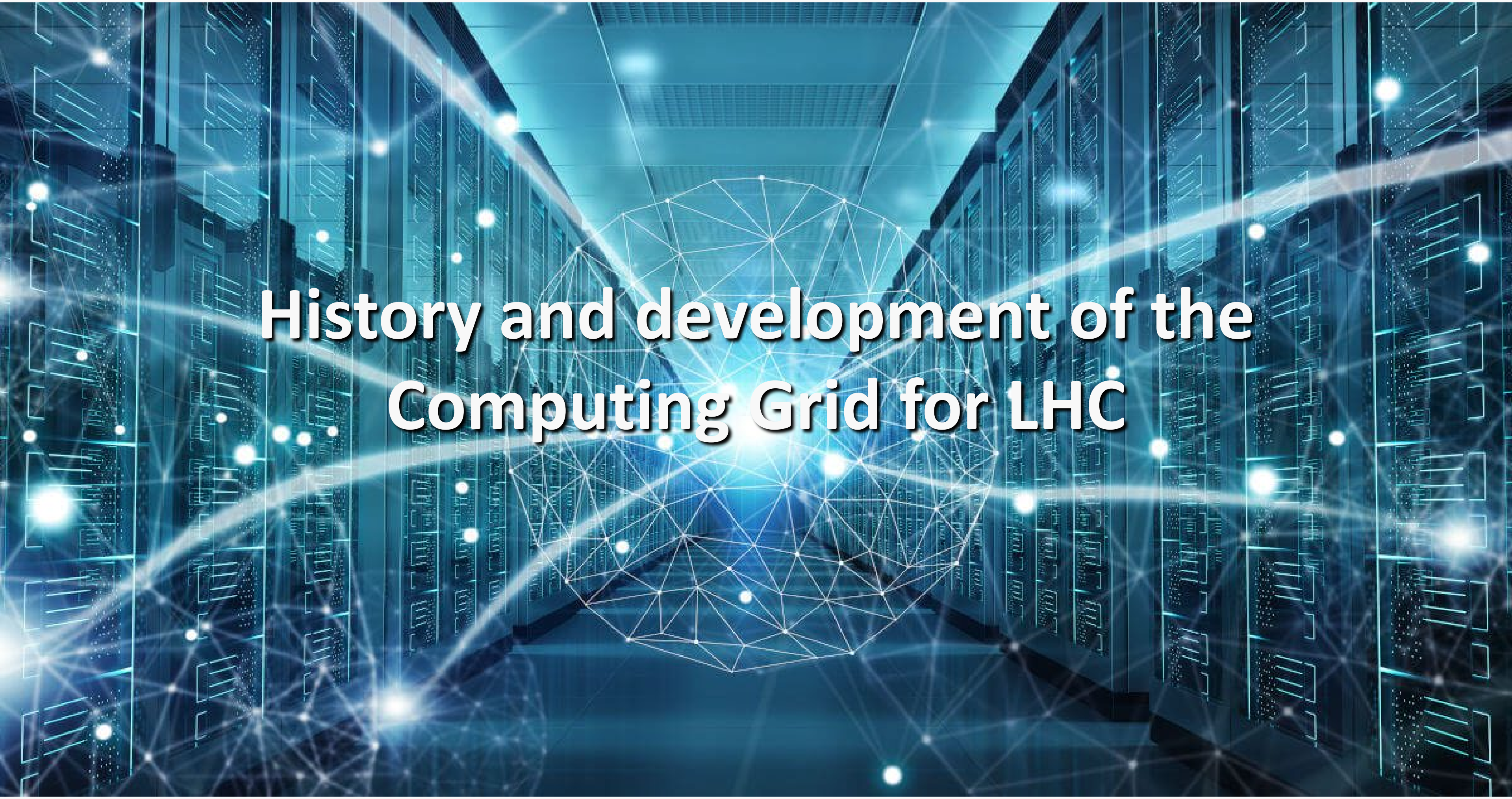
I apologize if someone will find it a bit boring.

OUTLINE

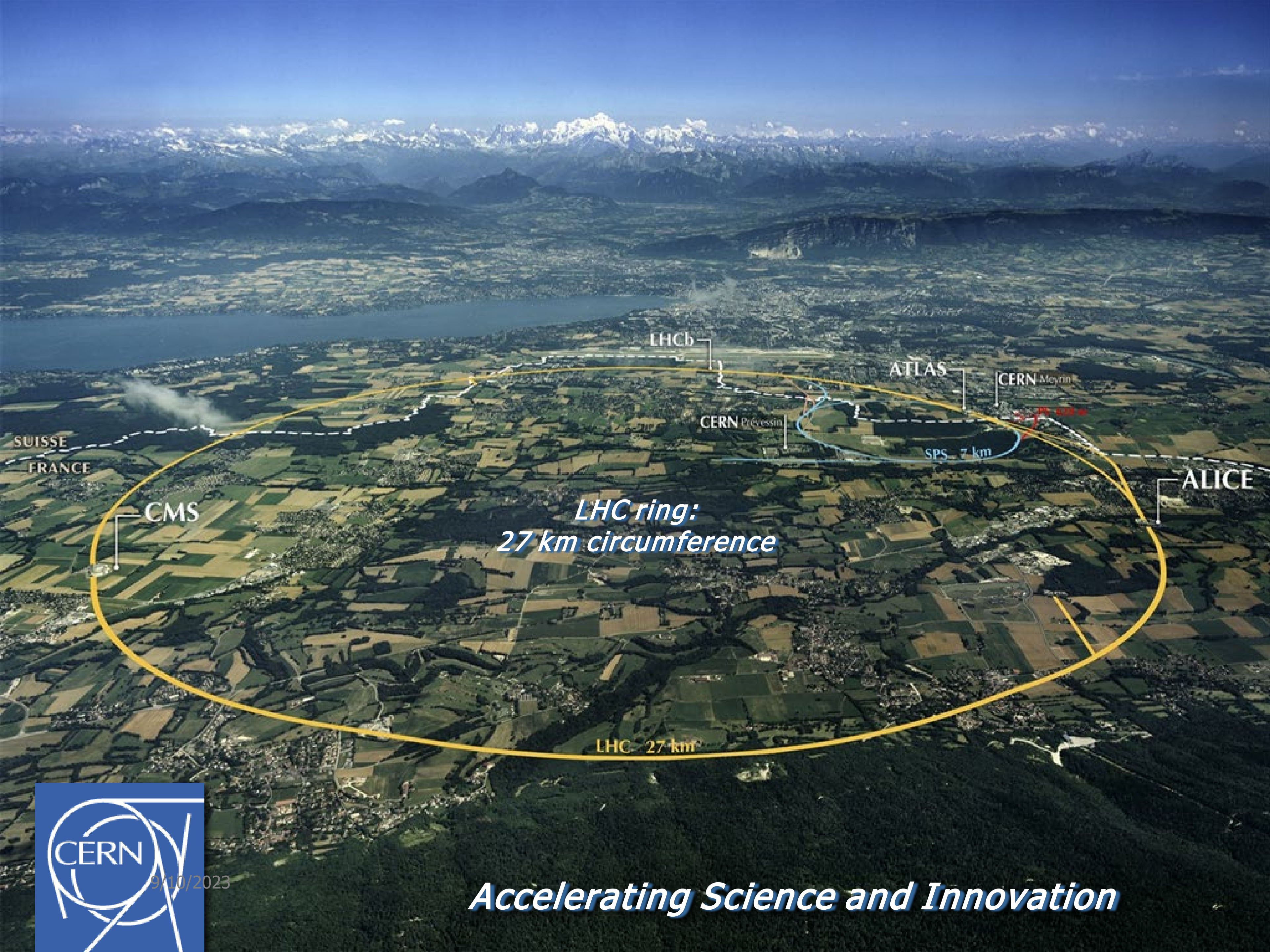


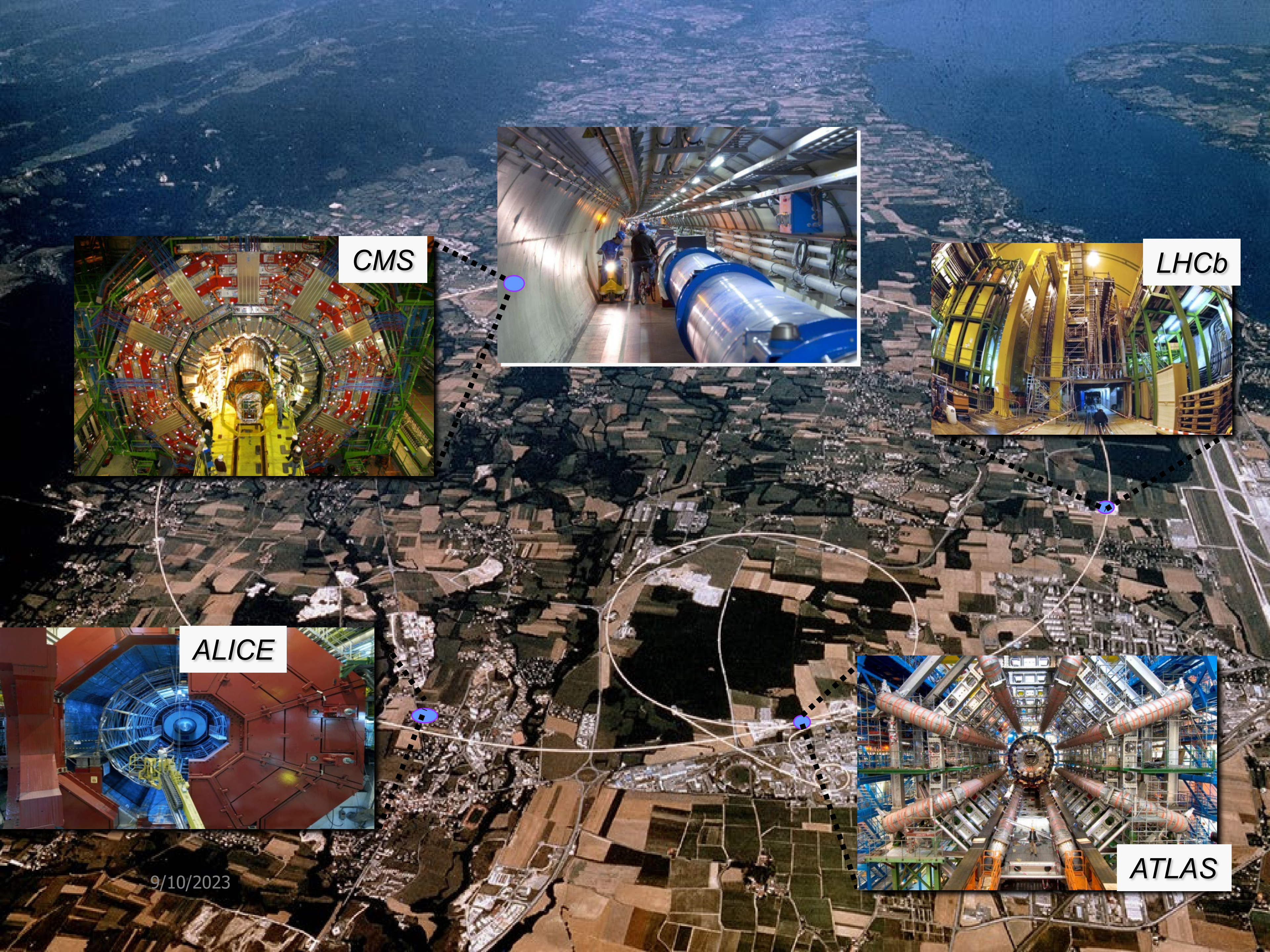
1. History and development of the Computing Grid for LHC
2. Exascale computing for HL-LHC
3. Czech contribution to the WLCG operations

PART 1



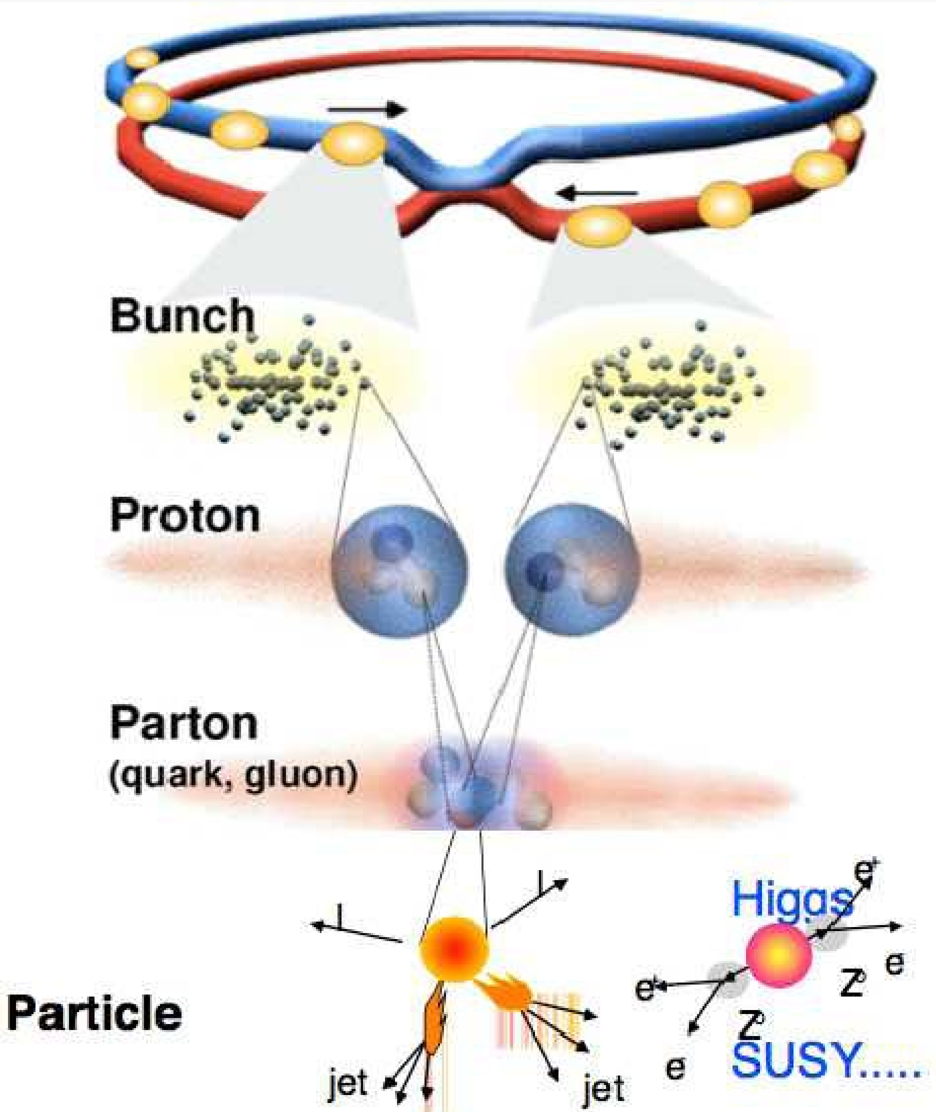
History and development of the Computing Grid for LHC


LHCb
ATLAS
CERN Meyrin
CERN Prévessin
SPS 7 km
SUISSE
FRANCE
CMS
ALICE
LHC ring:
27 km circumference
LHC 27 km
CERN
Accelerating Science and Innovation


CMS
LHCb
ALICE
ATLAS

The LHC tunnel view

Collisions in the LHC



Proton - Proton	2808 bunch/beam
Protons/bunch	10^{11}
Beam energy	7 TeV (7×10^{12} eV)
Luminosity	$10^{34} cm^{-2} s^{-1}$
Crossing rate	40 MHz
Collision rate ≈	10^{7}-10^{9}

New physics rate ≈ .00001 Hz

Event selection:
1 in 10,000,000,000,000

2022 Accelerator performance (Mike Lamont)

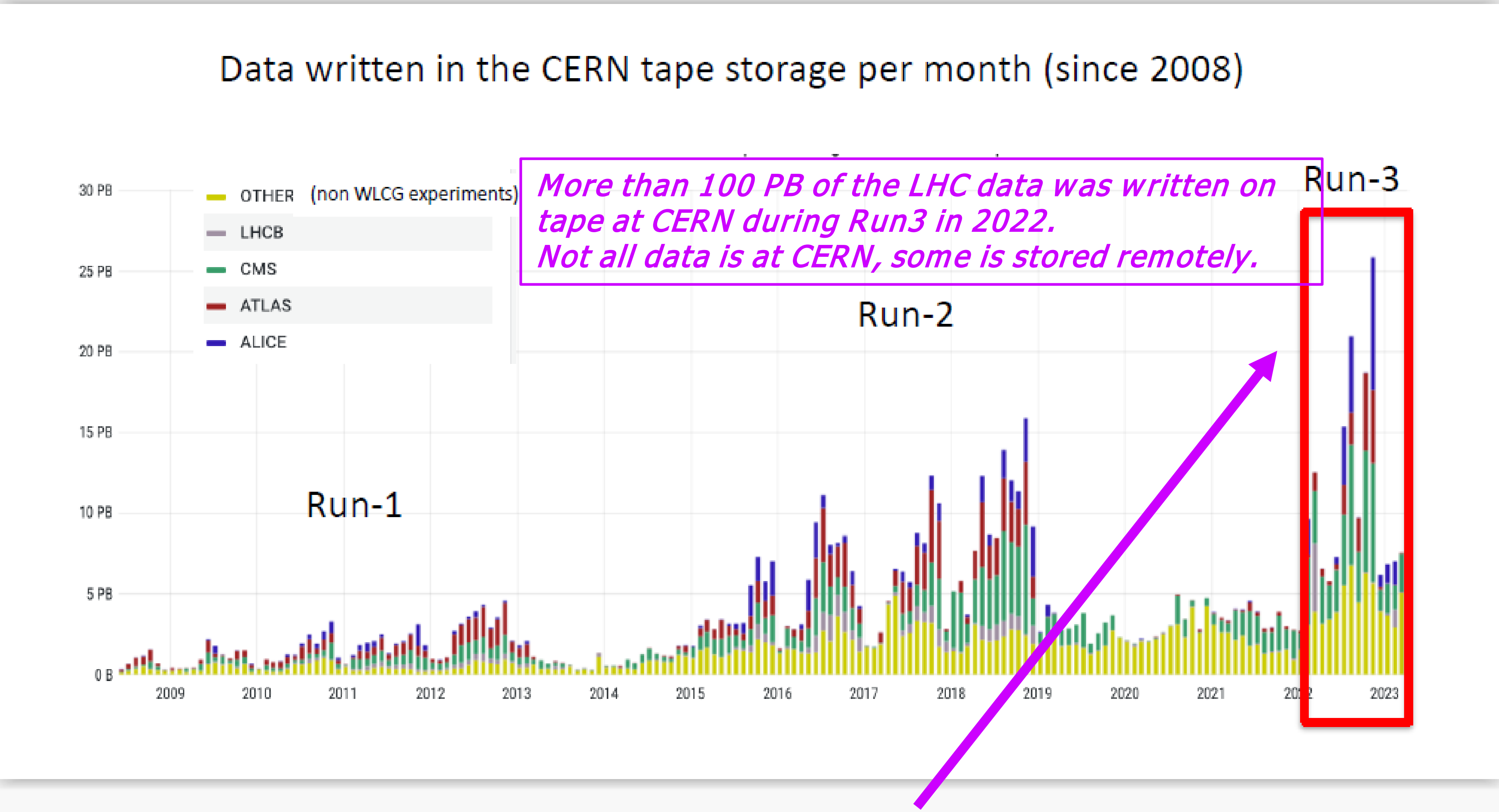
2022 GPD records (courtesy ATLAS)

2022 Beam Energy: 6.8 TeV

Peak Stable Luminosity Delivered	1.98×10^{34} $cm^{-2}s^{-1}$	Fill 8230	22/10/05 20:53
Maximum Average Events per Bunch Crossing	65.6	Fill 8301	22/10/21 18:53
Maximum Stable Luminosity Delivered in one fill	775.8 pb^{-1}	Fill 8274	22/10/15 22:23
Maximum Stable Luminosity Delivered in one day	1133.8 pb^{-1}	Sunday 23 October, 2022	
Maximum Stable Luminosity Delivered for 7 days	4.167 fb^{-1}	Monday 10 October, 2022 - Sunday 16 October, 2022	
Longest Time in Stable Beams for one fill	2 days, 9 hrs, 23 min	Fill 8178	22/09/24 00:21
Longest Time in Stable Beams for one day	1 day, 0 min	Sunday 25 September, 2022	
Longest Time in Stable Beams for 7 days	3 days, 14 hrs, 1 min	Thursday 29 September, 2022 - Wednesday 05 October, 2022	
Fastest ATLAS Ready from Stable Beams	0 min	Fill 7966	22/07/11 19:19
Fastest Turnaround to Stable Beams	1 hr, 51 min	Fill 8112	22/08/09 04:54
Maximum Colliding Bunches	2450	Fill 8267	22/10/14 22:28
Maximum Charge per Bunch Colliding	1.38×10^{11}	Fill 8299	22/10/21 13:12
Maximum Charge per Beam Colliding	3.33×10^{14}	Fill 8306	22/10/23 17:35
Maximum Total Charge per Beam	-	Fill 7920	22/07/05 16:47
Average Specific Luminosity	6.96×10^{30} $cm^{-2}s^{-1}(10^{11}\ p)^{-2}$	Fill 8216	22/09/30 22:47

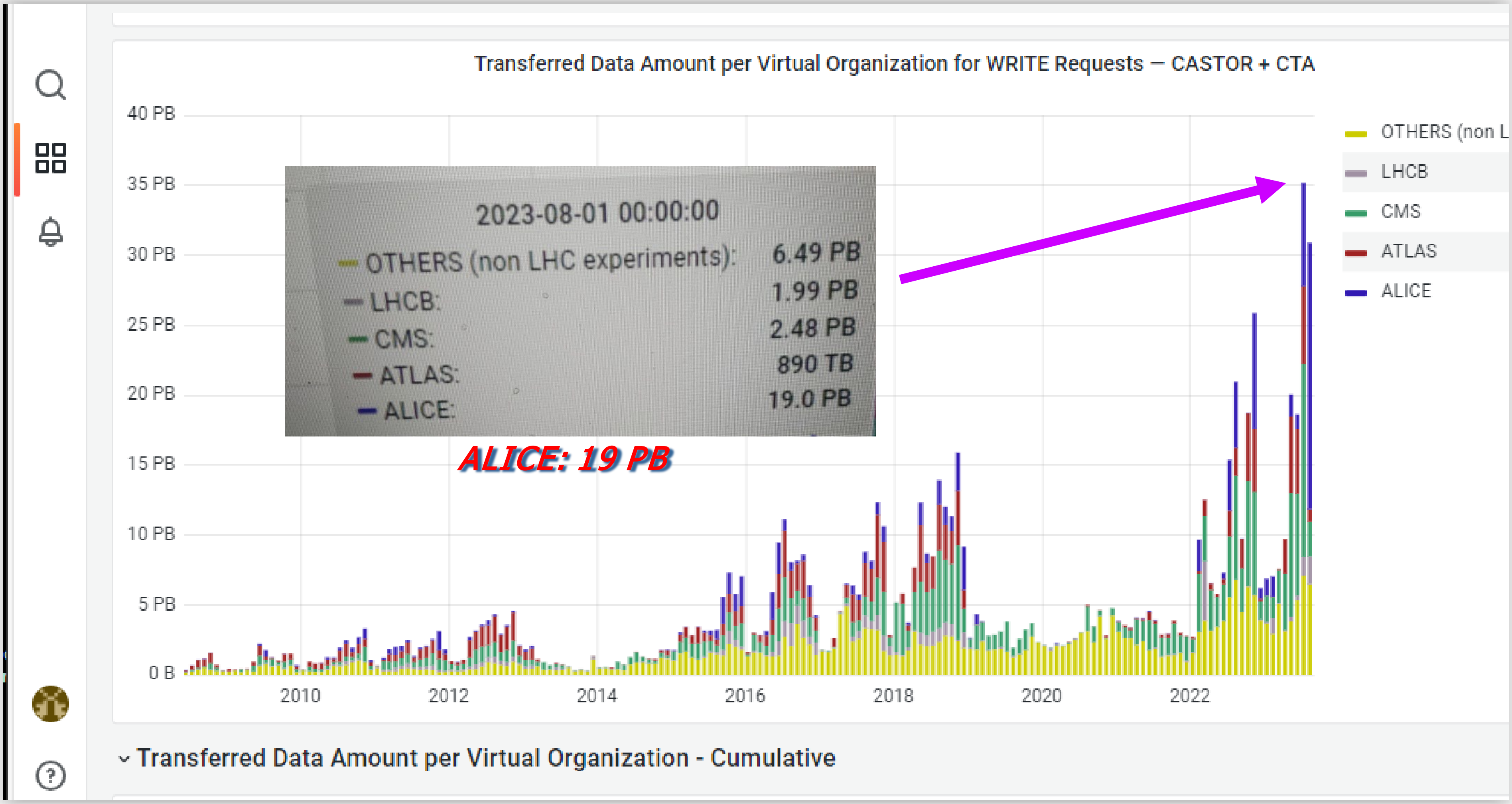
Stunning performance!

Run 3 data taking (2022) (CERN Grafana)



More than 26 PB of data written in November 2022 on the CERN tape storage.

Run 3 data taking (2023) (CERN Grafana)



31 PB of data written in August 2023 : running limited for a couple of days.

Putting things in context

An extremely large unit of digital data, **one Petabyte** is equal to 1,000 Terabytes. Some estimates hold that a Petabyte is the **equivalent of** 20 million tall filing cabinets or **500 billion pages of standard printed text**.

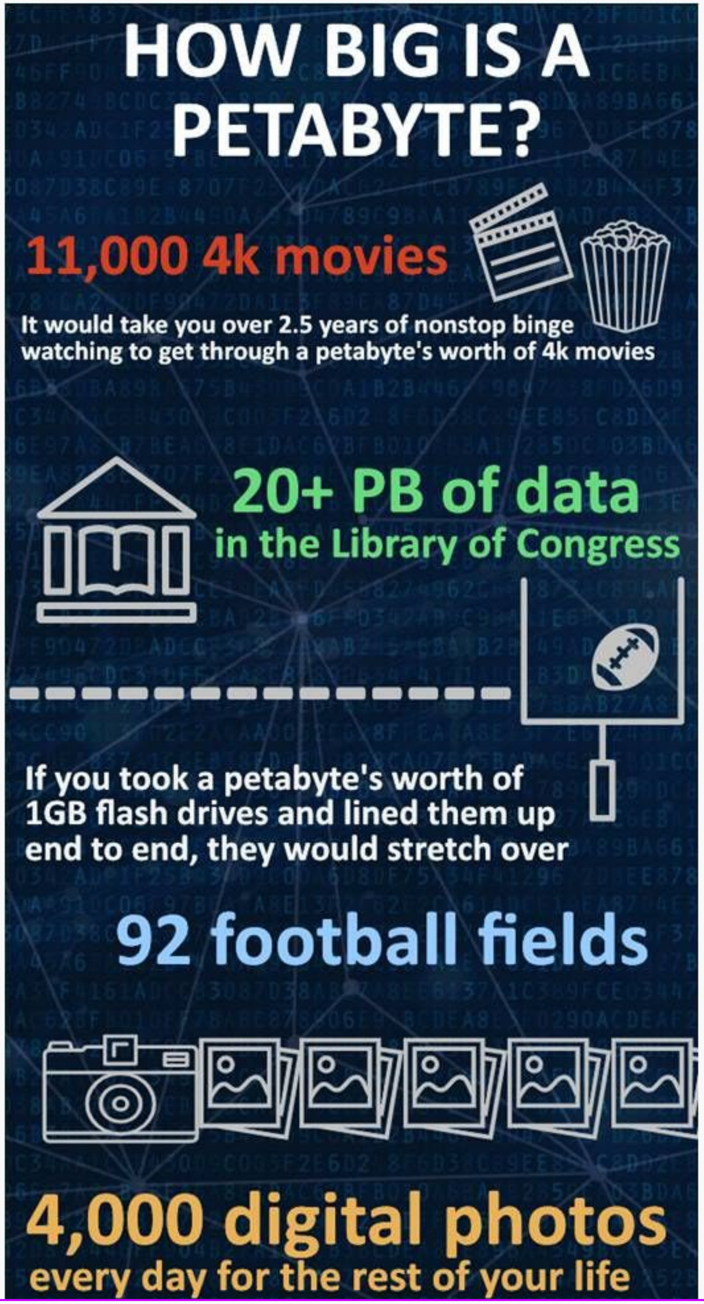
-The average **4k movie is 100GB of data**. This would mean **1 Petabyte of storage could hold 11,000** 4k movies. With an average run time of 2 hours, it would take you **over 2.5 years of nonstop watching** to get through a petabyte's worth of 4k movies.

- The **Library of Congress contains over 20 Petabytes of data**.

- If you took a **Petabyte's worth of 1GB flash drives** and line them up end to end, they would **stretch over 92 football fields**.

-If you stacked a **Petabyte's worth of 1TB SSD drives** on top of each other **in Madison Square Garden, they would reach from the court floor to the base of the score board Over two and a half times**.

- **1 petabyte**'s worth of data is **equal to taking over 4000 digital photos every day for the rest of your life**.



https://info.cobaltiron.com/blog/petabyte-how-much-information-could-it-actually-hold

1998: beginning of work on computing grid for LHC data

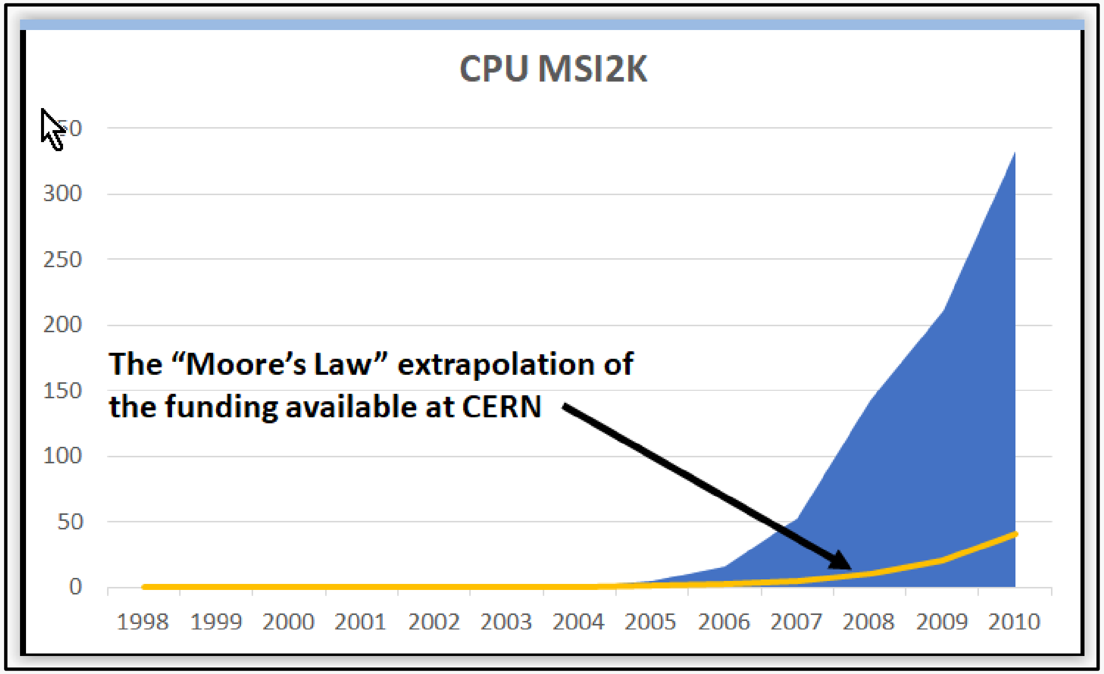
LHC – Construction had been approved in 1995 with a target date for first beams of 2005, (then 2007).

- The four experiment collaborations had already prepared initial estimates of the data rates, storage requirements and computing capacity that would be needed

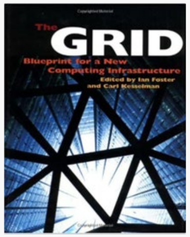
LHC detectors would produce much more data with more complex events than ever before

The LHC project budget had no line for computing at CERN

Most of the computing capacity required would have to come from outside of CERN



Solution: (Worldwide) LHC Computing Grid



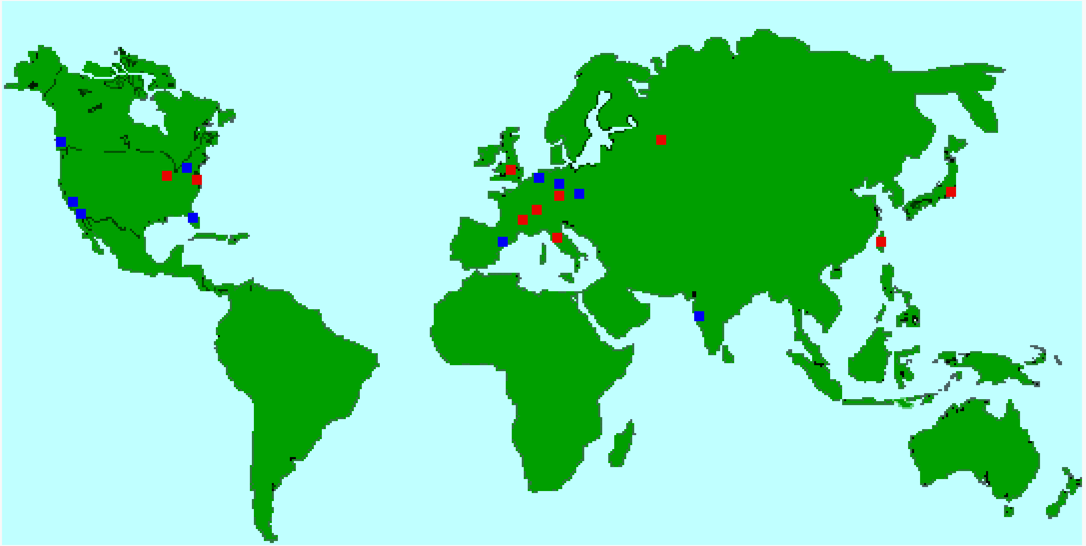
Computing Grid is a group of networked computing clusters/centers that work together as a virtual supercomputer to perform large tasks, such as analyzing huge sets of data. The system operations are managed with the GRID middleware.
1998: beginning of work on computing grid for LHC data.

The first campaign: LCG-1 Service Challenge in September 2003

- Agreement reached on principles for registration and security
- Certification and distribution process established and tested - June
- Rutherford Lab (UK) to provide the initial Grid Operations Centre
- FZK (Karlsruhe) to operate the Call Centre
- Pre-release middleware deployed to the initial 13 centres – July

- 13 PARTICIPANTS:

Taiwan, Brookhaven, CERN, Bologna, Fermilab, Karlsruhe, Lyon, Budapest, Moscow, **Prague**, Barcelona, Rutherford UK, Tokyo



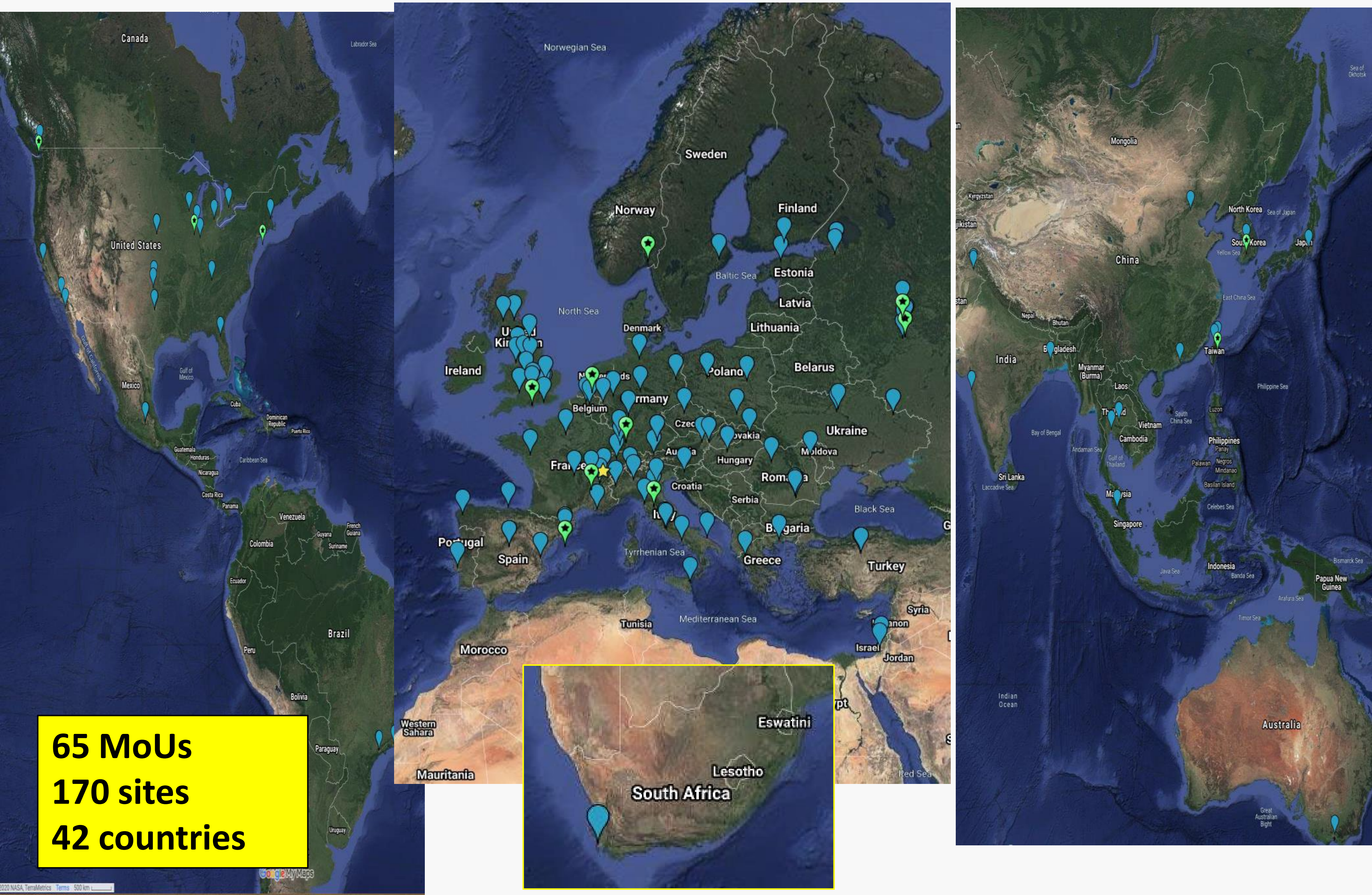
LHC Computing Grid development

- Between the first Service Challenge and the start of the LHC operations in 2009/2010, there were several Data and Service challenges to exercise all aspects of the service not just for data transfers, but workloads, support structures etc.
- e.g. DC04 (ALICE, CMS, LHCb) and DC2 (ATLAS) in 2004

- Ever since the start of the LHC operations in 2009/2010, the LHC Computing Grid provided processing, management and storage of the LHC data while expanding and undergoing upgrades. The infrastructure name is now Worldwide LHC Computing Grid (WLCG) .

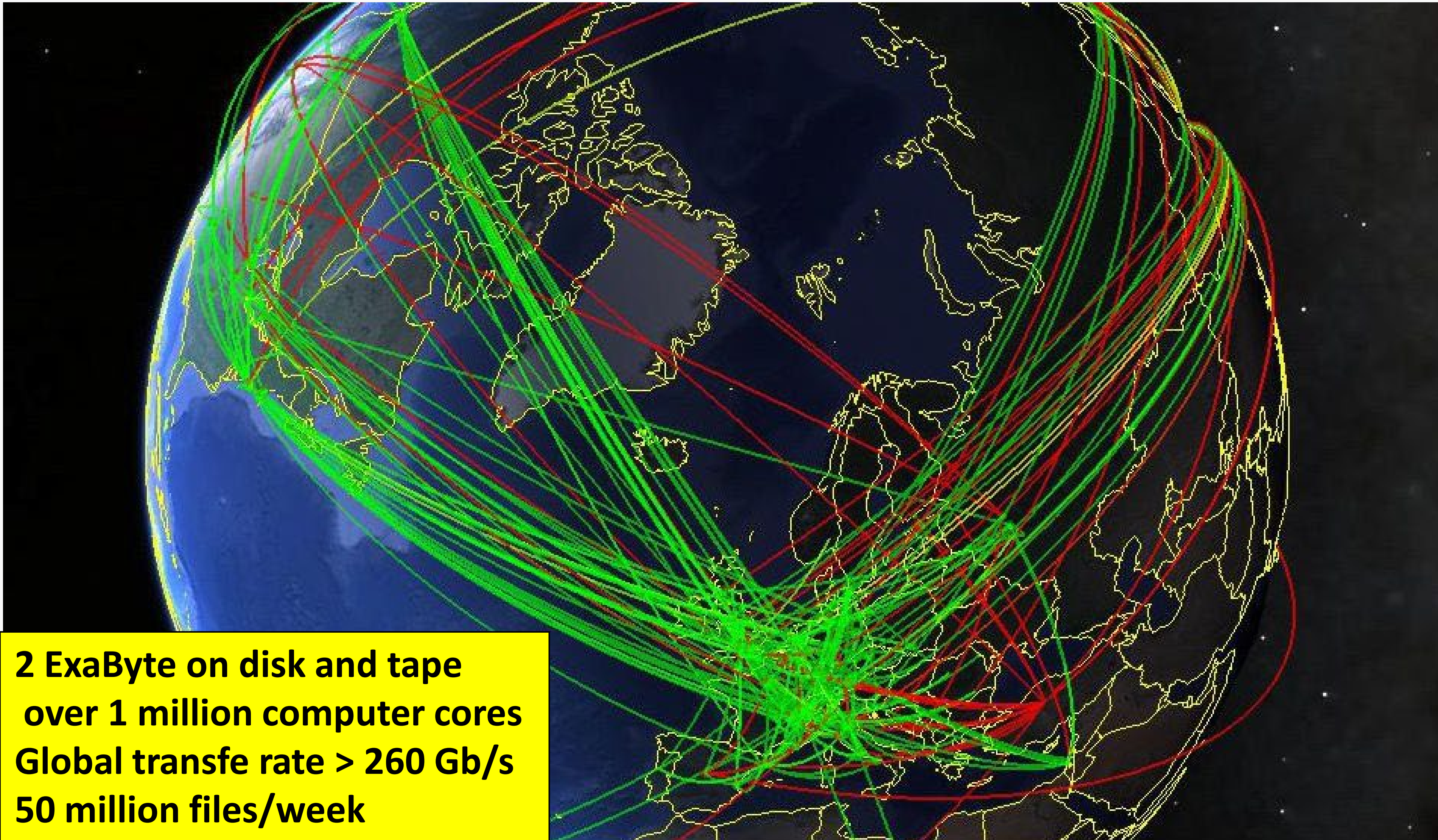
- WLCG Home Page: "*WLCG is a global computing infrastructure whose mission is to provide computing resources to store, distribute and analyze the data generated by the Large Hadron Collider (LHC), making the data equally available to all partners, regardless of their physical location.*
- WLCG is the world's largest computing grid*. It is supported by many associated national and international grids across the world."*

Worldwide LHC Computing Grid topology



65 MoUs
170 sites
42 countries

Worldwide LHC Computing Grid resources



2 ExaByte on disk and tape
over 1 million computer cores
Global transfe rate > 260 Gb/s
50 million files/week transferrd

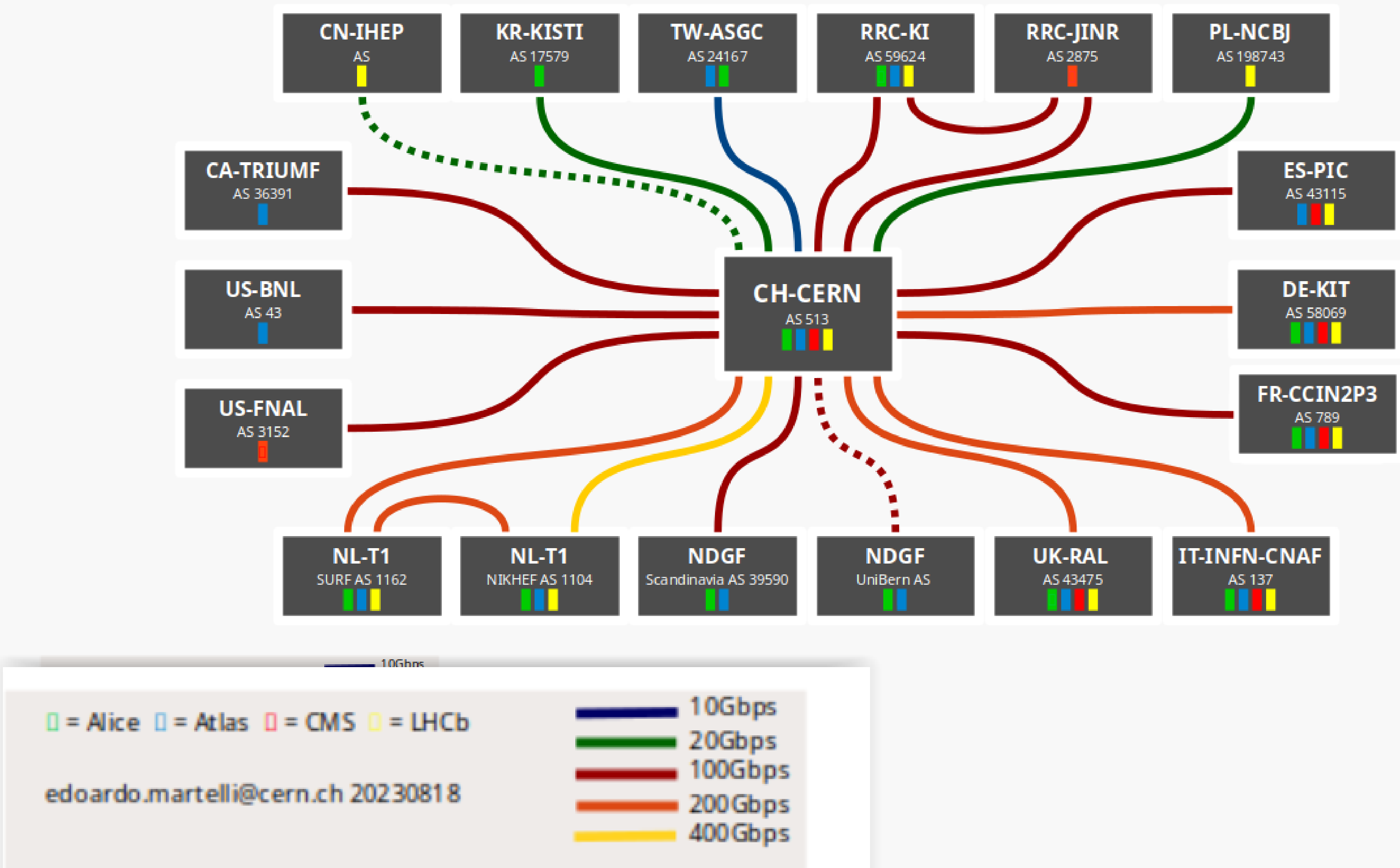
WLCG Tier structure

The WLCG ecosystem consists of sites ranked as Tier-0, Tier-1 and Tier-2. Tier-0 is CERN, 18 Tier-1s are large computing centers and ~150 Tier-2s are smaller size centers.

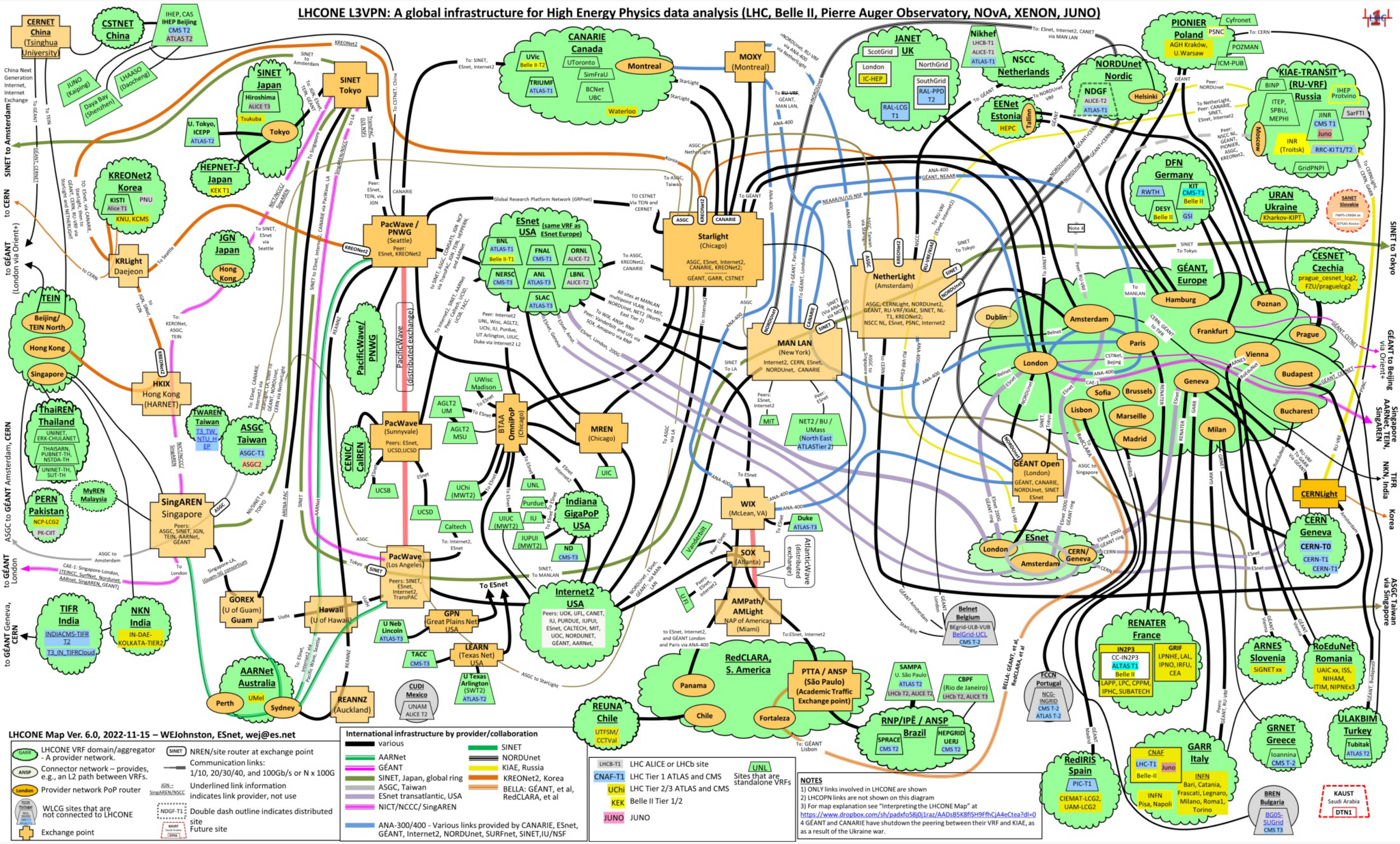
The role of Tier-2s in the WLCG ecosystem is crucial: they deliver resources comparable to Tier-1s.

LHC Optical Private Network

LHCOPN



LHCONE – LHC Open Network Environment topology



WLCG Fabric and software

- ***Computing***: mostly Intel or AMD CPUs with x86 instruction sets at multi-core servers. Some GPUs available at the main Grid sites and at HPC centers.

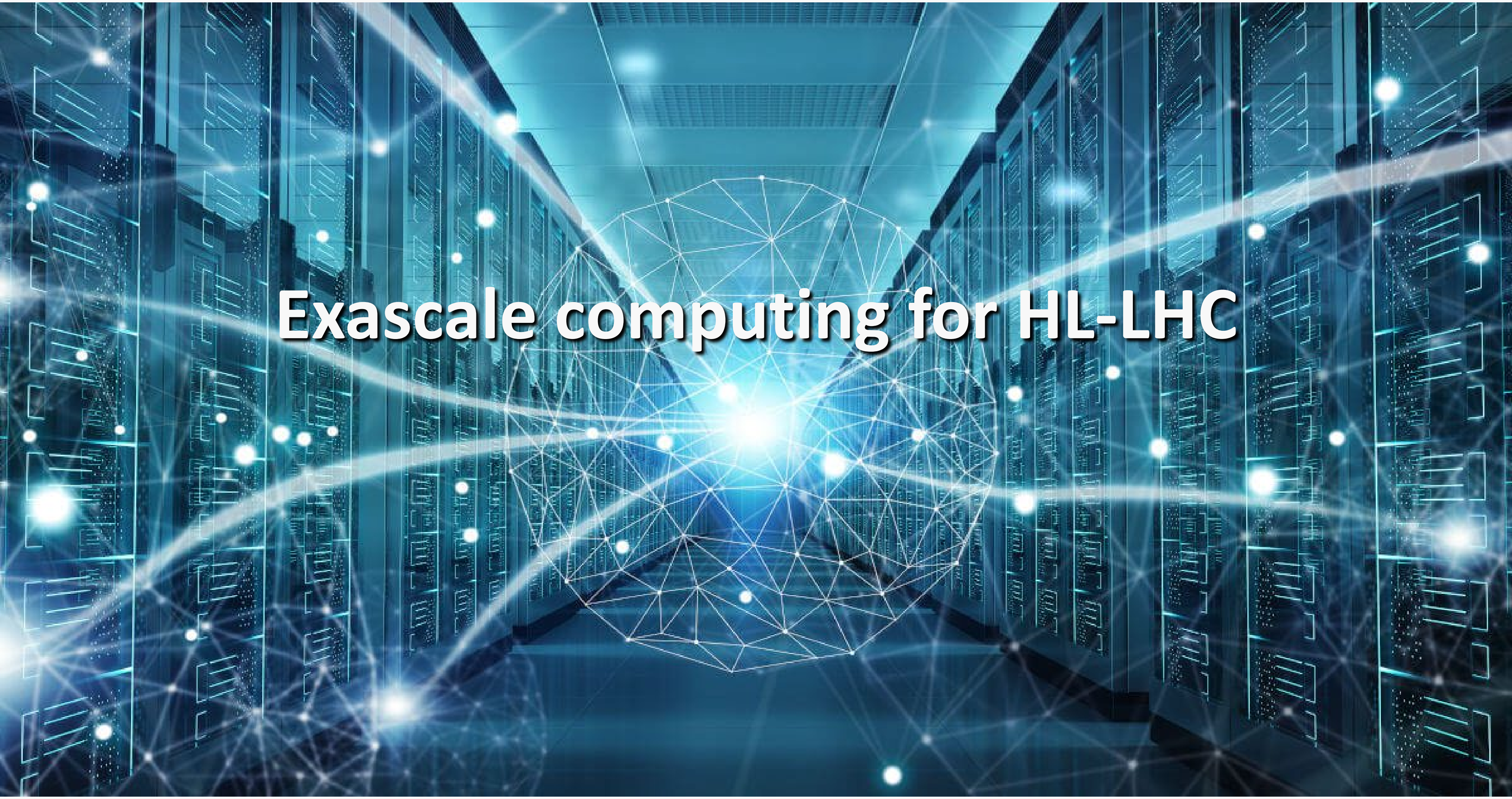
- ***Storage:*** mixture of tape and disk. Since 2018, the "Data Organization, Management and Access" (DOMA) project is active. R&D for "Data Lakes".

- ***Network:*** connection between CERN and T1s provided by a system of P2P connections of capacity of 100 to 400 Gb/s, so called LHC Optical Private Network (LHCOPN). Most T0/1/2 sites interconnected via LHC Open Network Environment (LHCONE). L3VPN service over research and education networks.

- ***Software:*** complex system of experiments' dedicated frameworks written in a mixture of C++ and Python. Rely on many external packages from within and outside the field. Many millions of lines of code. Generally written for x86, increasingly multi-threaded. Being ported e.g. to GPUs.

- ***Analysis :*** software quite various, but moving towards the Python ecosystem and particularly to notebooks.

PART 2

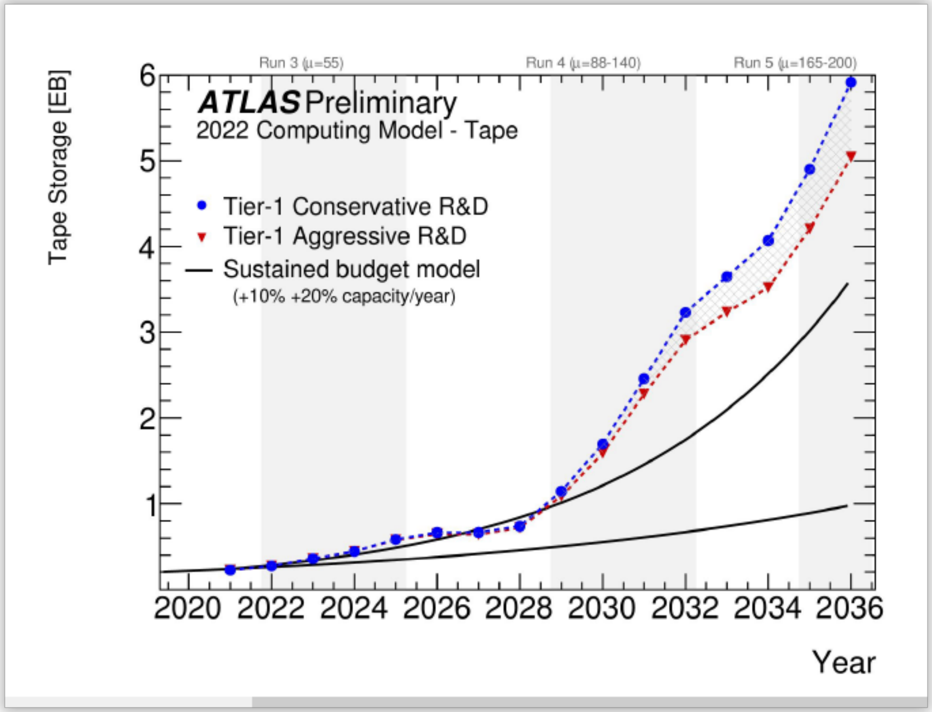


Exascale computing for HL-LHC

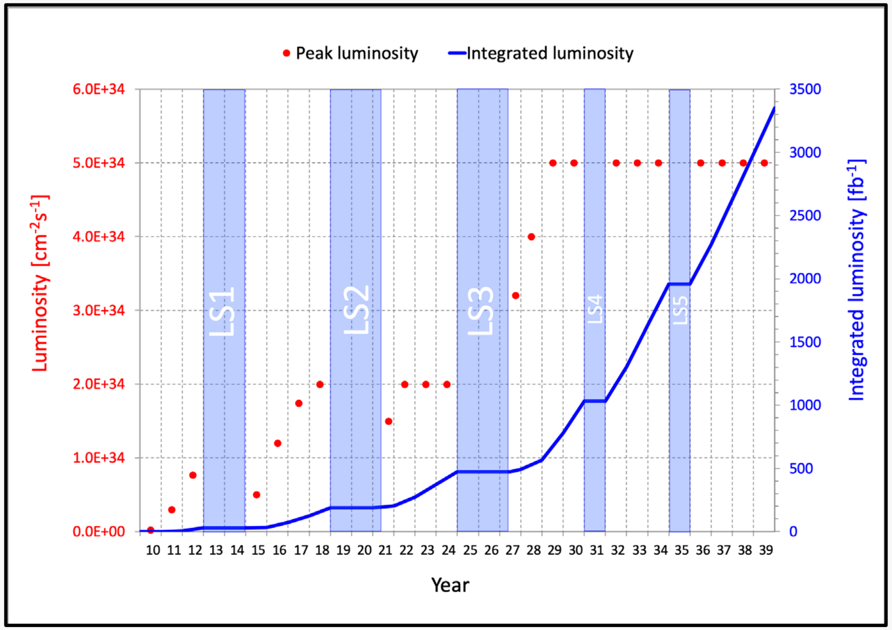
WLCG resources

- Every institute – member of the LHC experiments is obliged to provide a contribution to computing and storage resources of WLCG
- Two times a year the institutes deliver pledges to the WLCG management board.
- The basic condition is so called flat budget contribution which should ensure a 10% - 20% increase in the resource performance.

- The flat budget policy was basically sufficient for covering the needs of Run-1 and Run-2, during Run-3 experiments already relied partly on external non-WLCG resources.
- With the High Luminosity LHC (HL-LHC) campaigns approaching, the experiments evaluated their future needs and a large discrepancy occurred.
- A number of R&D projects are in progress to evolve the WLCG infrastructure into a high performance ecosystem adequate for HL-LHC.



HL-LHC luminosity, current planning (Run 4 and beyond)

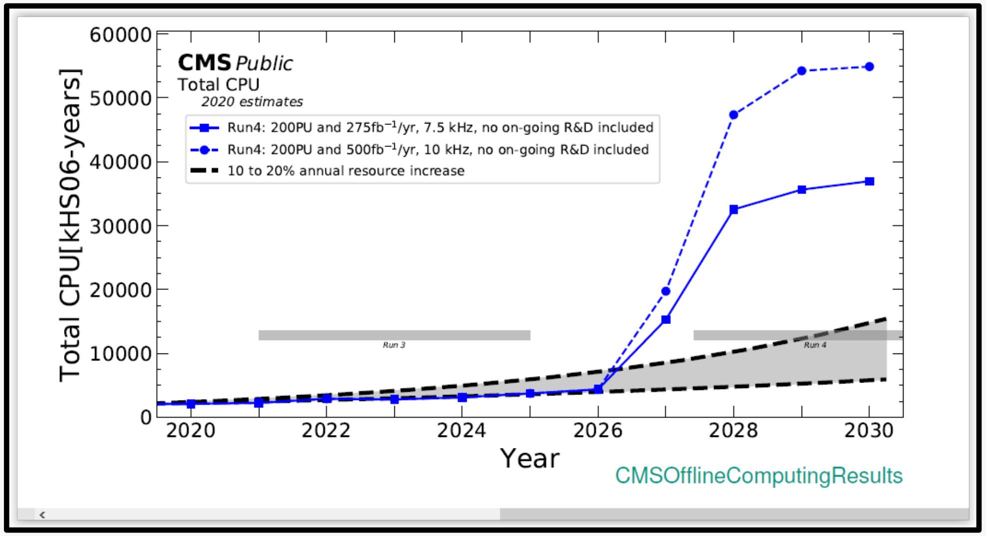


2027: commissioning year
Run-4 production years: 2028 to 2030
LS4: 2031
Run-5 production years: 2032 to 2034

-The data volumes anticipated in HL-LHC time span will be dramatically larger than those currently managed, supposed to reach the multi-Exabyte scale.

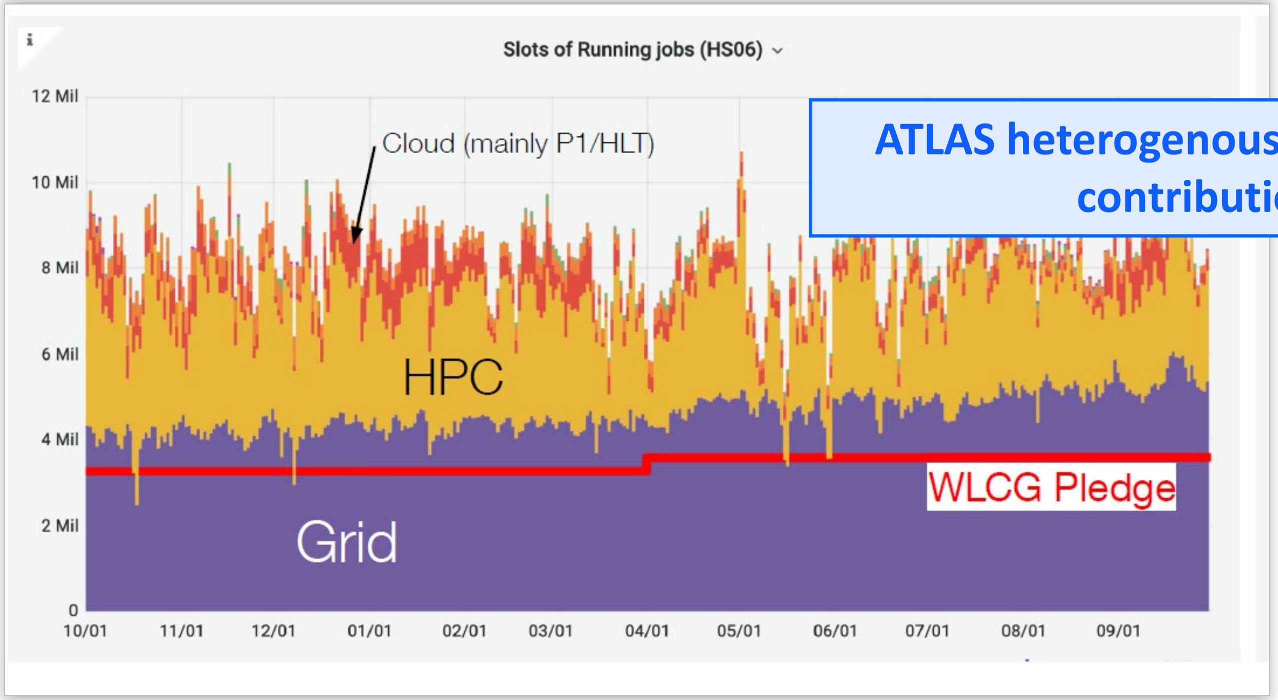
-The estimated amount of resources needed for the management of this data will be about 6 - 10 times larger than allowed by the flat budget policy.

CMS: Anticipated growth in CPU resources needed towards HL-LHC.
Factor ~6 difference between "flat budget" and Physics needs.



Use of non-Grid, external, "opportunistic resources"

- To match the needs for computing power, experiments employ additional, non-Grid external heterogenous resources: HPCs, (commercial) clouds, HLT farms.
- In the future this heterogeneity will expand.
- BUT: one cannot rely on such resources: the possibility they disappear cannot be ignored
- Pledged Grid resources remain vital to the experiments.



ATLAS heterogenous resources: a significant contribution from HPCs

In different strategy documents, a number of R&D projects is proposed that will contribute to the development of a high performance ecosystem for HL-LHC.

Ever growing exploit of HPCs – not without obstacles, e.g.:
No network access from the worker nodes
No persistent storage
HEP software is viewed as poor standard and un-trusted

So, why bother?

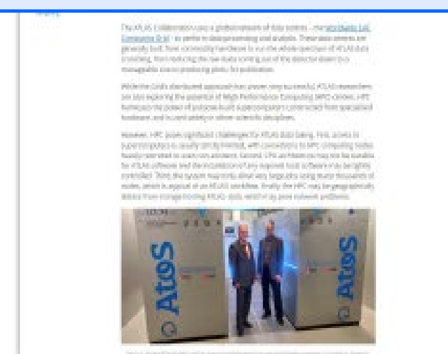
Frontier, #1
By OLCF at ORNL - https://www.flickr.com/photos/olcf/52117623843/
CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=[illegible]

Top 500 HPCs in 2022: https://top500.org/lists/top500/2022/06/highs
Total combined performance of all 500 exceeded the Exaflop barrier

Fugaku, #2
https://www.r-ccs.riken.jp/en/fugaku/

The HPC Vega, the Slovenia's first and only peta-scale supercomputer. Collaboration with ATLAS.

- A great untapped rapidly growing resource
  - More than 100x WLCG (1 million 3GHz cores x 10 Flops/cycle = 30Pflop/s)
- A substantial part of national computing infrastructure investment now and in the future
- Potential for allocations or "free" opportunistic computing
- Interesting and motivating R&D and PR
- Improving flexibility of experiment workloads and services



ATLAS briefing on Vega HPC, June 2022

The European High Performance Computing Joint Undertaking (EuroHPC JU)

The European High Performance Computing Joint Undertaking (EuroHPC JU) is a legal and funding entity, created in 2018 and located in Luxembourg to lead the way in European supercomputing.

The EuroHPC JU allows the European Union and the EuroHPC JU participating countries to coordinate their efforts and pool their resources to make Europe a world leader in supercomputing. This boosts Europe's scientific excellence and industrial strength, support the digital transformation of its economy while ensuring its technological sovereignty.
The EuroHPC JU was created in 2018 and recently reviewed by means of Council Regulation (EU).

European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

The Mission

Establish a single collaborative cluster of next generation European Strategy Forum on Research Infrastructures (ESFRI) facilities in the area of astronomy- and accelerator-based particle physics in order to implement a functional link between the concerned ESFRI projects and European Open Science Cloud (EOSC).

THE OUTPUT

ESCAPE Software Repository: the repository of scientific software services of the research infrastructures concerned by the ESCAPE.

ESCAPE ESFRI Science Analysis Platform: a flexible science platform for the analysis of open access data.

ESCAPE Virtual Observatory: astronomical high-level products archive and related services.

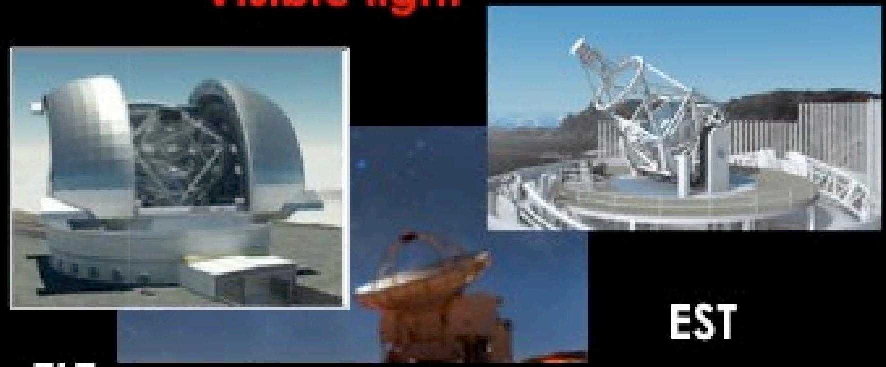
ESCAPE Data Infrastructure for Open Science: a scalable federated data infrastructure as the basis of an open access data service for the ESFRI projects within ESCAPE and concerned by Exabyte-scale data volumes.

Radio

JIVE-VLBI

SKA

Visible light

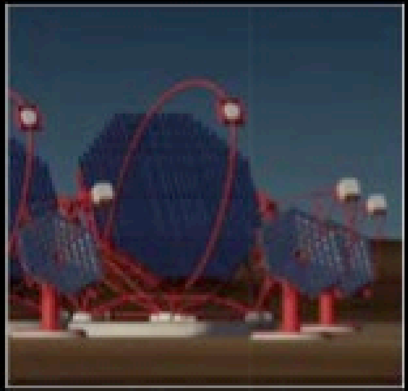


ELT

ESO

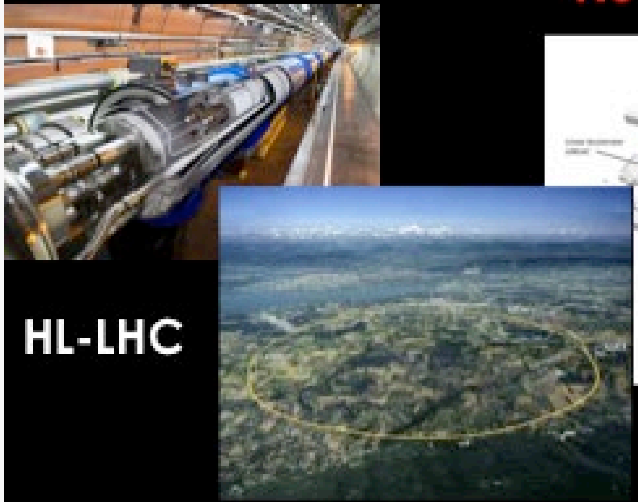
EST

Gamma rays



CTA

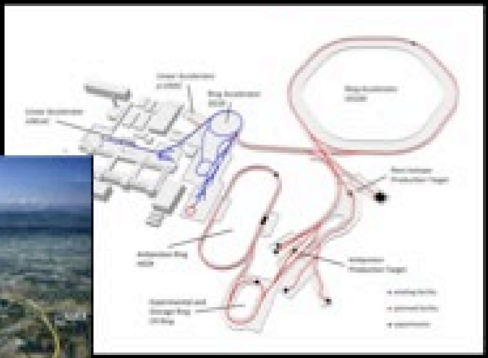
Accelerator-based **Particle Physics**



HL-LHC

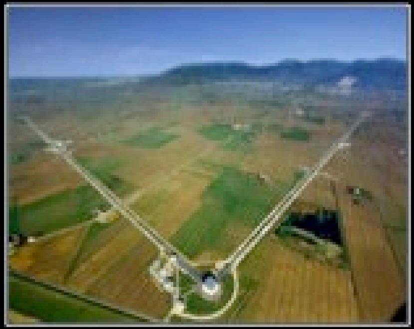
CERN

Accelerator-based **Nuclear Physics**



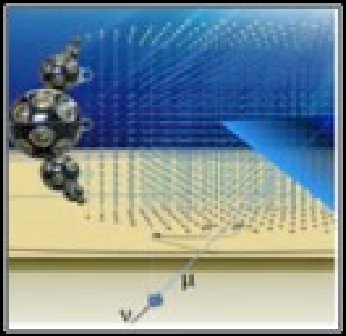
FAIR

Gravitational **Waves**



EGO-VIRGO

Cosmic-rays **Neutrinos**



KM3NeT

ESFRI and other large RIs in ESCAPE

Conclusions (WLCG)

- *WLCG infrastructure has been running smoothly during Run 3 and before*
- COVID-19 has had only little impact on operations
- The HEP community has *a number of challenges* to address with respect to computing and software *before the HL-LHC era*:

  Computation, Portability, Storage and Data Delivery, Analysis.
- Crucial for maintaining the data transfer performance: ongoing WLCG Data Challenges and WLCG Network Data Challenges. The upcoming one in February 2024.
- Funding agencies and institutes must realize that *computing and software are as important for physics as detector development and construction*.

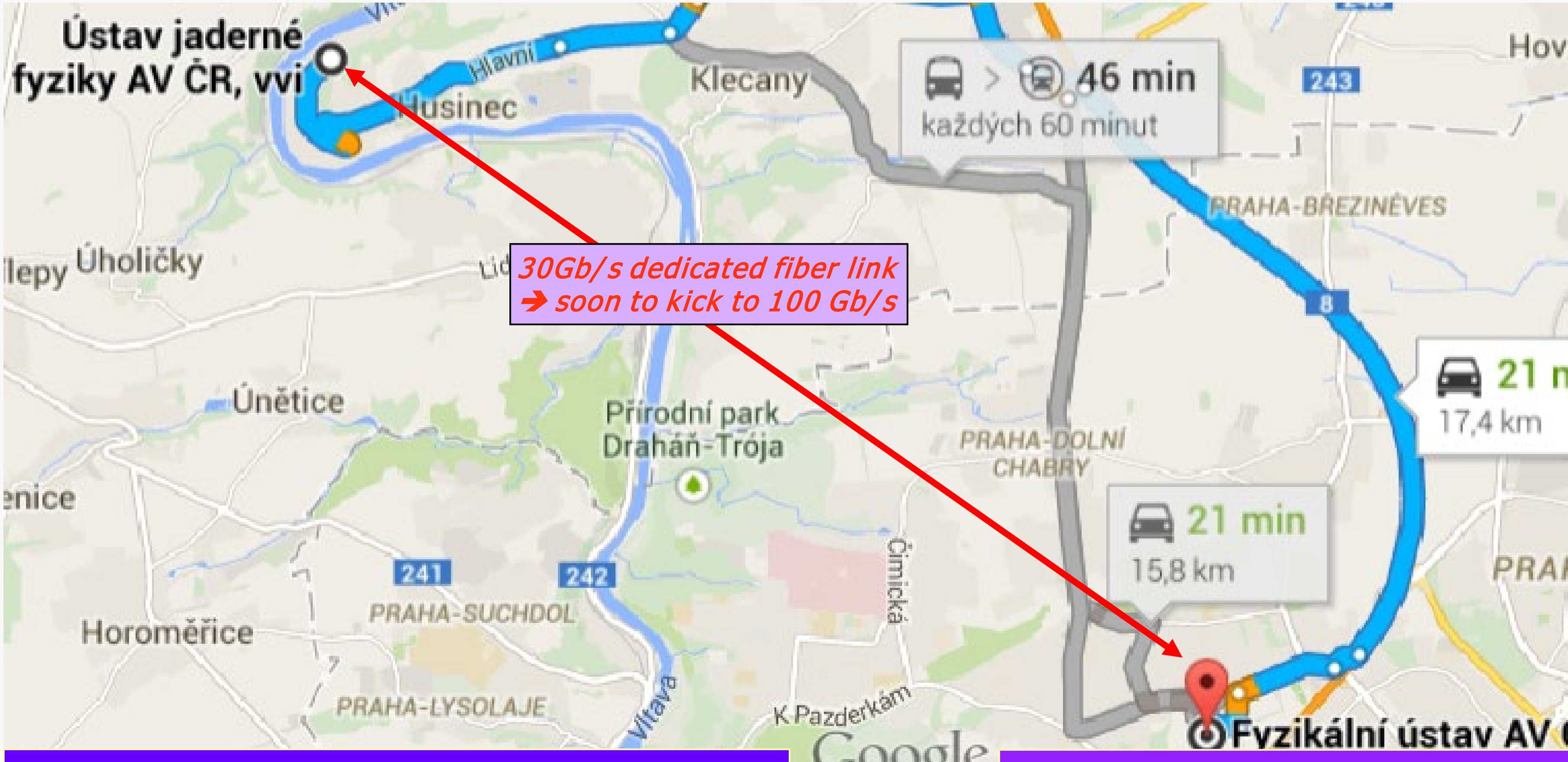
- *A policy for the future: federation of resources like within the ESCAPE project. WLCG engaged with other HEP experiments* (DUNE, Belle 2) *and communities* (astronomy) *to* collaborate on development of the infrastructure to be shared and covering for the challenges of the ExaByte science projects.

PART 3

Czech contribution to the WLCG operations

Czech republic Tier-2 center - geographical layout

Nuclear Physics Institute AS CR (NPI)
A large ALICE group, no ATLAS involvement
Operates ALICE XRootD storage servers

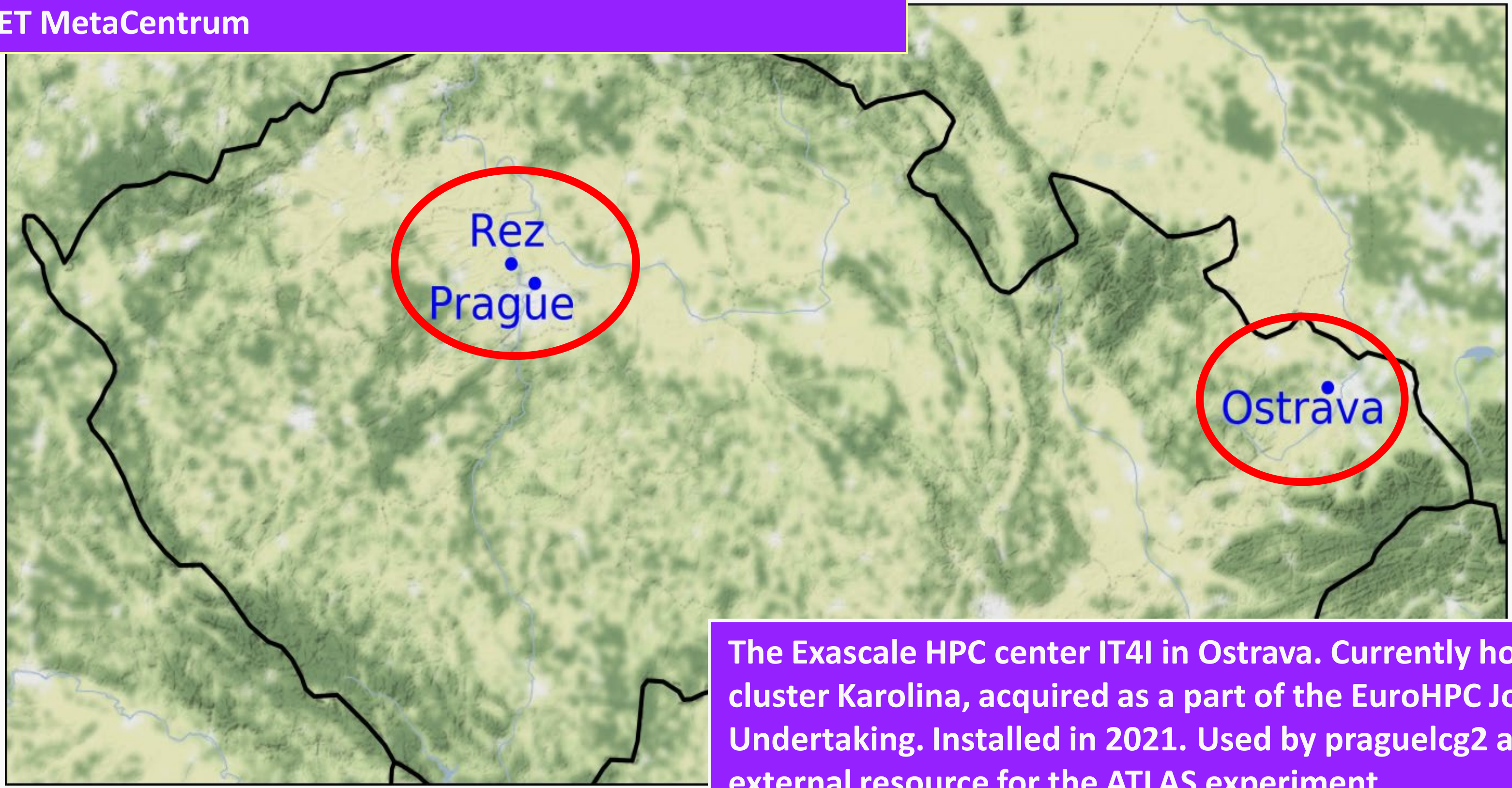


Since the beginning, the xRootD storage servers for ALICE were only at NPI. Only for a limited period, XRootD servers were also in operation at FZU.

Institute of Physics AS CR (FZU)
Regional computing center
WLCG Tier-2 site praguelcg2
All the CPU resources for ALICE and ATLAS
Quite small ALICE group, much larger ATLAS community

Extended geographical layout

Other participating institutions – in Prague
Charles University (CU), Faculty of Mathematics and Physics
Czech Technical University (CTU), Faculty of Nuclear Sciences and Physical Engineering
CESNET MetaCentrum



The Exascale HPC center IT4I in Ostrava. Currently hosting a cluster Karolina, acquired as a part of the EuroHPC Joint Undertaking. Installed in 2021. Used by praguelcg2 as an external resource for the ATLAS experiment.

HEP Computing in Prague: WLCG site praguelcg2
(a.k.a. the farm GOLIAS)

- A national computing center for processing data from various HEP experiments
- Distributed resources with all central services and most of hardware at FZU
- Basic infrastructure already in 2002.
- *One of the 13 sites participating in the first LCG campaign in 2003*

- *April 2008, WLCG MoU signed* by Czech Republic (ALICE+ATLAS)
- *Certified as a Tier2 center of LHC Computing Grid (praguelcg2)*
- Collaboration with various other Grid projects

- *Very good network connectivity* provided by CESNET / e-INFRA CZ.
- Multiple dedicated 10 – 100 Gb/s connections to collaborating institutions, 2*100 Gb/s connection to LHCONE.

- Provides computing services for ATLAS + ALICE, Auger, NOVA, Fermilab, Astrophysics ...

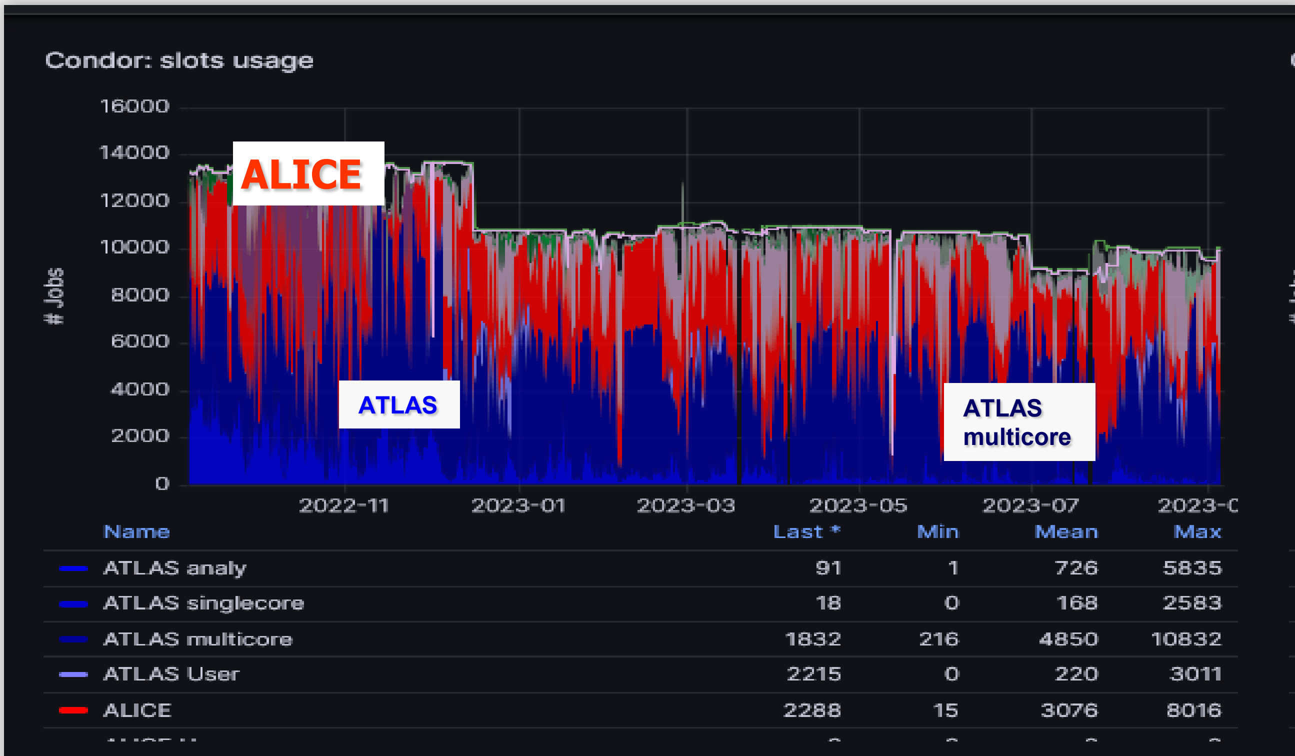
Current resources

- Dedicated server room 62 m^2 (2015)
- 1 batch system (HTCondor)
- 2 main WLCG VOs: ALICE, ATLAS
- ~ 10000 job slots on site + ~1500 job slots at Charles University
- ~ 11 PB in total on disk storage on site and at NPI (dCache, XRootD) plus ~1PB
- on NFS servers
- Regular upscale of resources on the basis of various financial supports,
- mainly the academic grants.

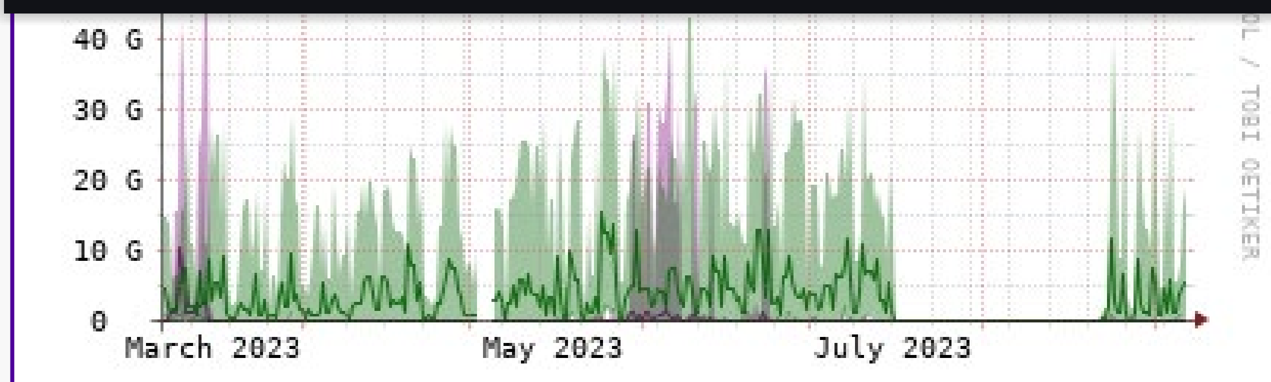
- Monitoring: Grafana
- Configuration management by Puppet
- Provisioning and SW management by Foreman

- ALICE XRootD storage at NPI – 3.8 PB of disk space

Performance monitoring examples

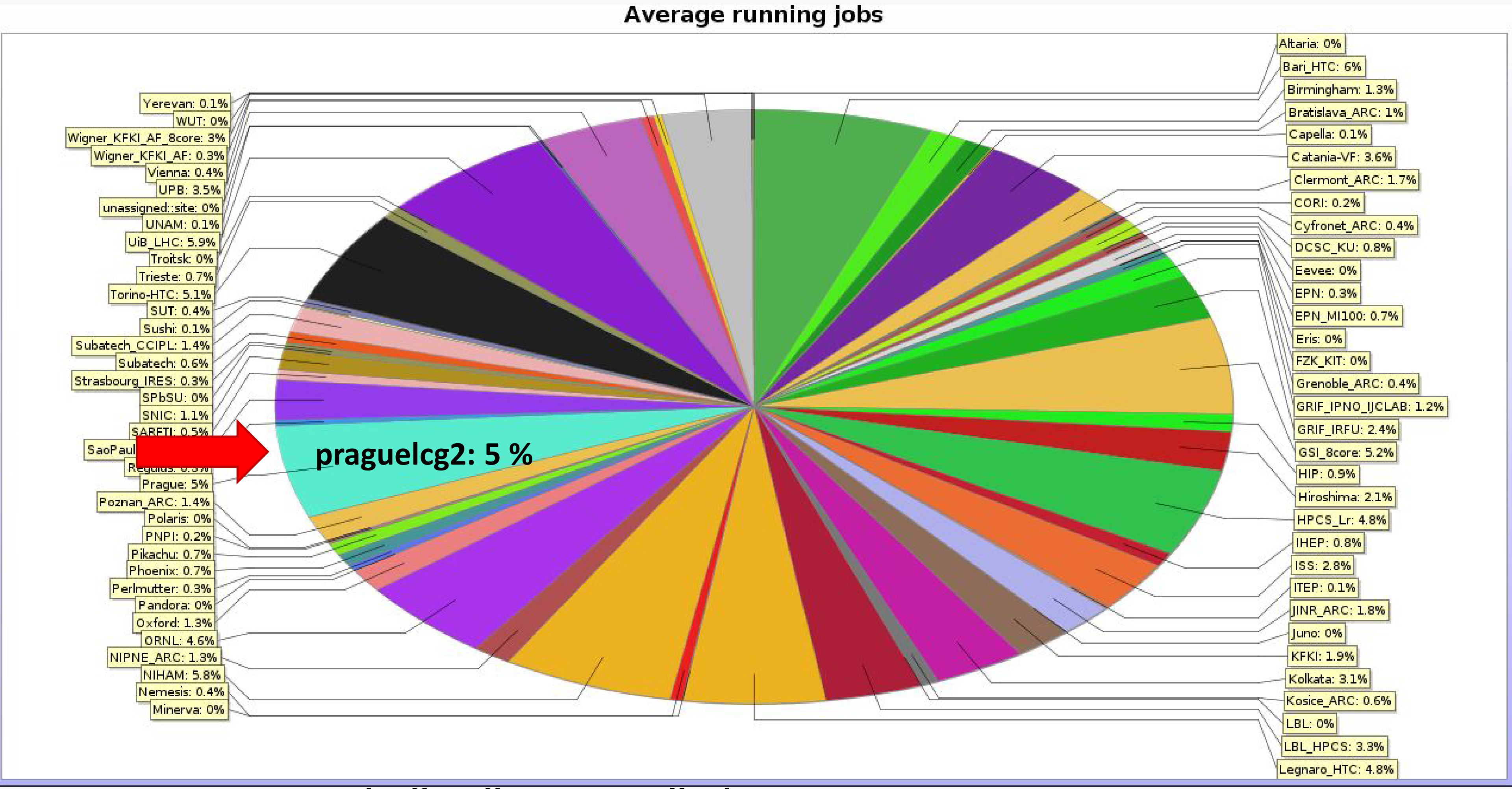


Running jobs profile during last year, by the local monitoring at the central site in Prague. ***Main CPU consumers are ALICE and ATLAS.***



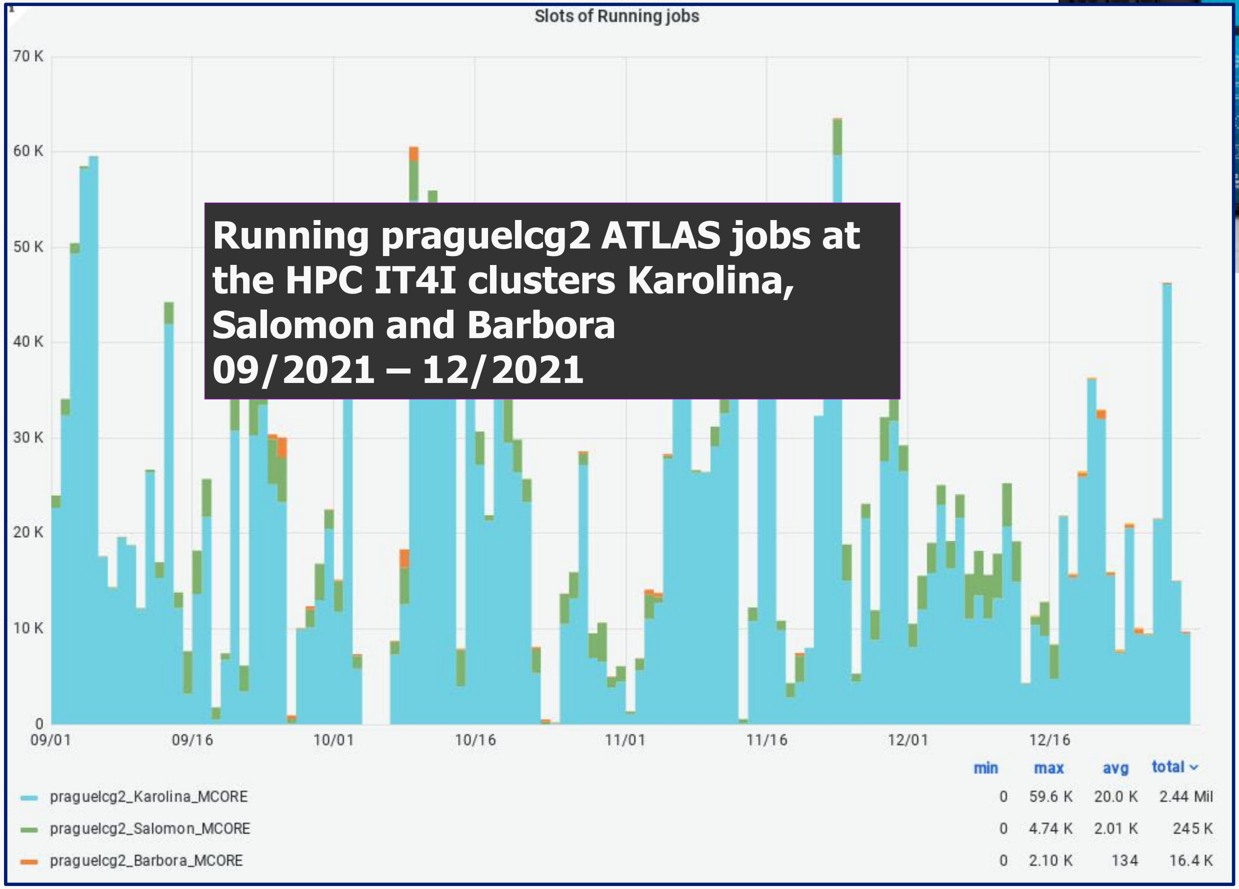
Network traffic between the main site of praguelcg2 and NPI storage cluster during last 6 months. ***Maximum throughput 43.2 Gb/s.***

CPU delivery - ALICE Tier-2 share: 09/2022 – 09/2023

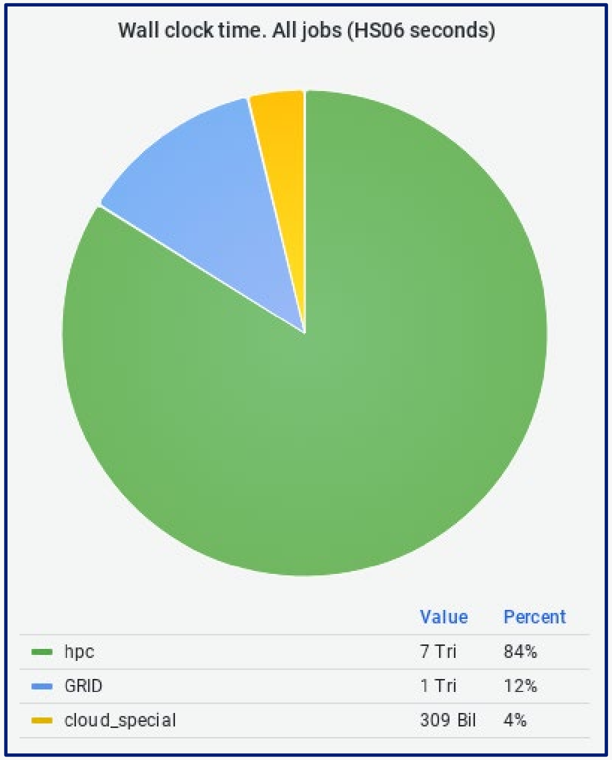


Czech republic share during 09/2022 – 09/2023 was 5%.

Use of the resources of the HPC center in Ostrava (IT4I), by ATLAS



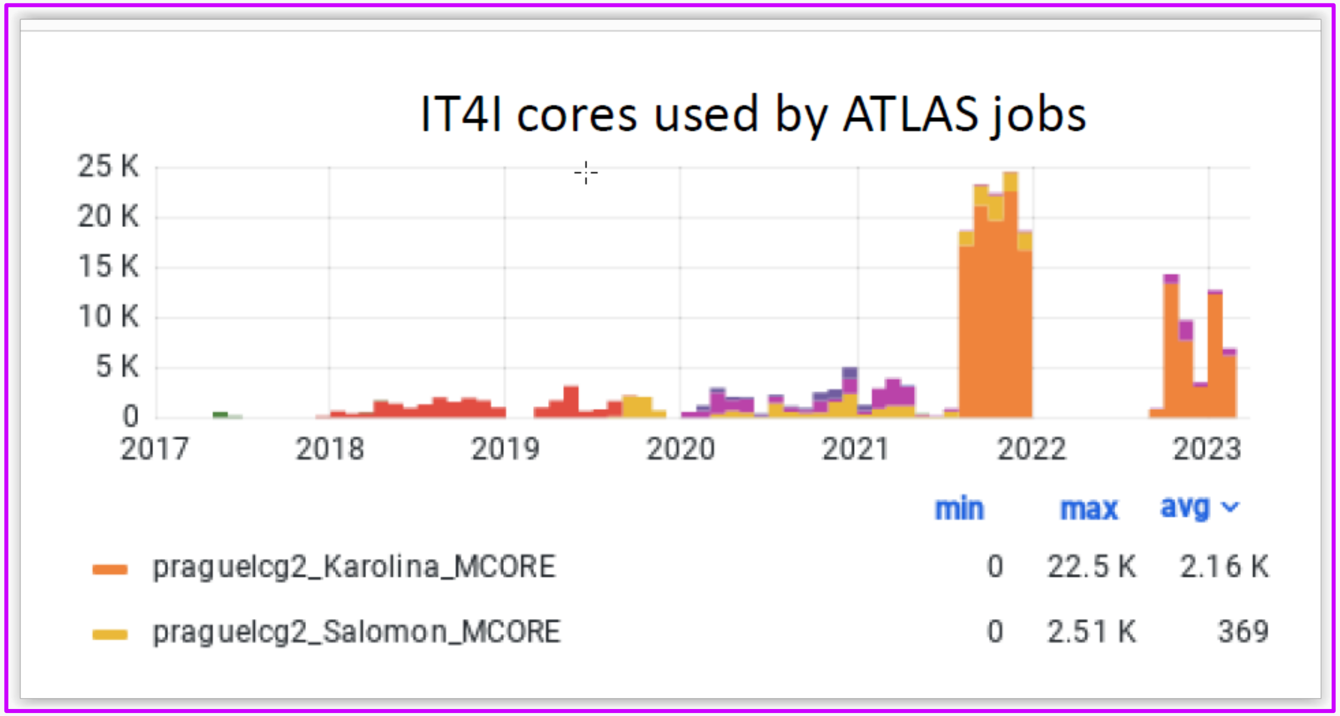
Running praguelcg2 ATLAS jobs at the HPC IT4I clusters Karolina, Salomon and Barbora 09/2021 – 12/2021



Contributions from the external resources to the praguelcg2 ATLAS operations
In 2021: HPC: 84%, Grid:12%, BOINC:4%.
In 2020: HPC: 47%, Grid:52%, BONIC:1%.

Use of the resources of IT4I, current status, by ATLAS

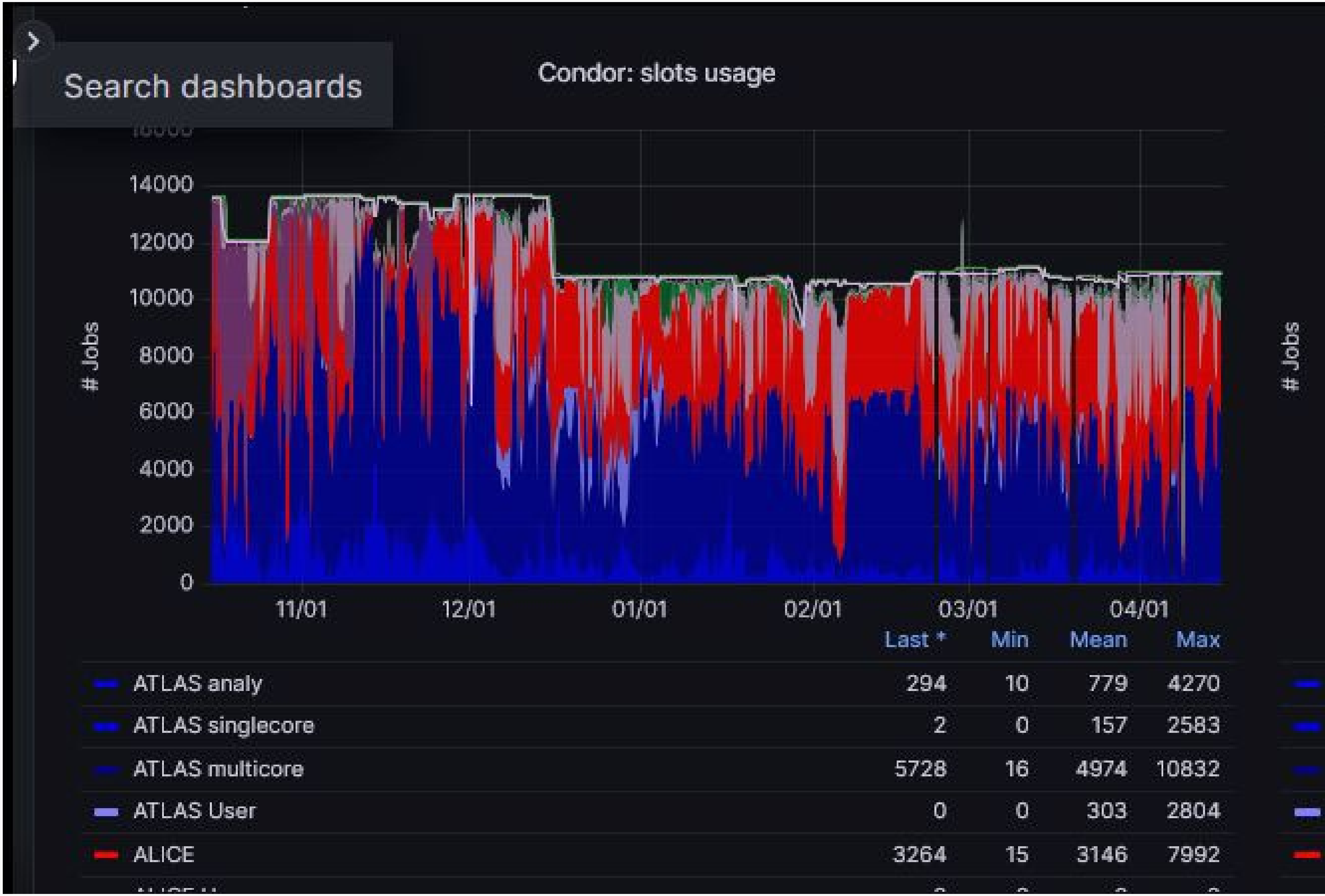
- 2023: Standard project with highest allocation this year
- 25.6M corehours
- allocation almost used up already within 2 months
- equivalent to 66000 FZU cores used for 2 months



Conclusions (CR computing)

- Czech Tier-2 center is successfully delivering CPU and storage resources to the Grid and local users with minimal outage.
- Czech republic has delivered in excess the pledged computing resources and participates in various WLCG infrastructure projects
- Keeps operating cost reasonable with periodic hardware refresh cycle
- Estimated requirements for LHC experiments in 2023 - 2026: main funding contribution for LHC computing reduced to nearly half
- But, we will do our best to keep up the reliability and performance level of the services and deliver the maximum we can.

BACKUP


Search dashboards
Condor: slots usage
Jobs
14000
12000
10000
8000
6000
4000
2000
0
11/01
12/01
01/01
02/01
03/01
04/01
Last * Min Mean Max
ATLAS analy 294 10 779 4270
ATLAS singlecore 2 0 157 2583
ATLAS multicore 5728 16 4974 10832
ATLAS User 0 0 303 2804
ALICE 3264 15 3146 7992
Jobs

To put this talk in context …

Work on the LHC computing grid started 24 years ago – Remember 1998?

- Bill Clinton was impeached over his affair with Monica
- Google was launched in September
- The Nokia 6110 was the best mobile phone available – it would be another 9 years before the iPhone appeared

- A good home network connection was 64 kbits/sec
  A high-speed inter-site network was 622 Mbits/sec (if available)
- A big hard disk held 12GB in a hefty 5.25” package
  An IBM tape cartridge held 20GB
  – less than an Apple Watch today

- Amazon had been selling books online for three years but it would be a further 8 years before Amazon Web Services announced the S2 (storage) and EC2 (Elastic Computing) services – the first cloud

1998: beginning of work on computing grid for LHC data

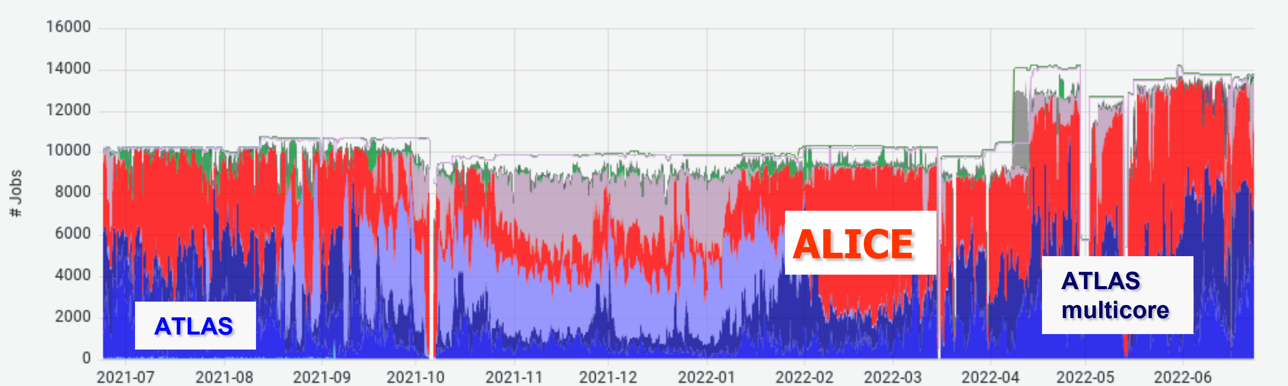
LHC –

- Construction had been approved in 1995 with a target date for first beams of 2005
- The four experiment collaborations had already prepared initial estimates of the data rates, storage requirements and computing capacity that would be needed

HEP Computing –

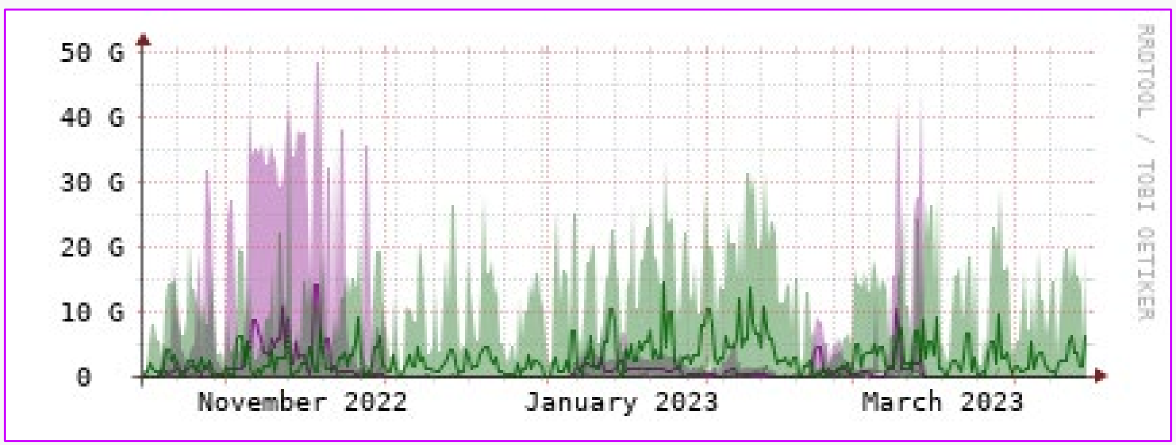
- PC clusters with Linux was the "standard"
  With remote job submission via the internet
- Major sites had mass storage management systems using tape robots
- There was very good TCP/IP network expertise for efficient and reliable data access

Performance monitoring examples



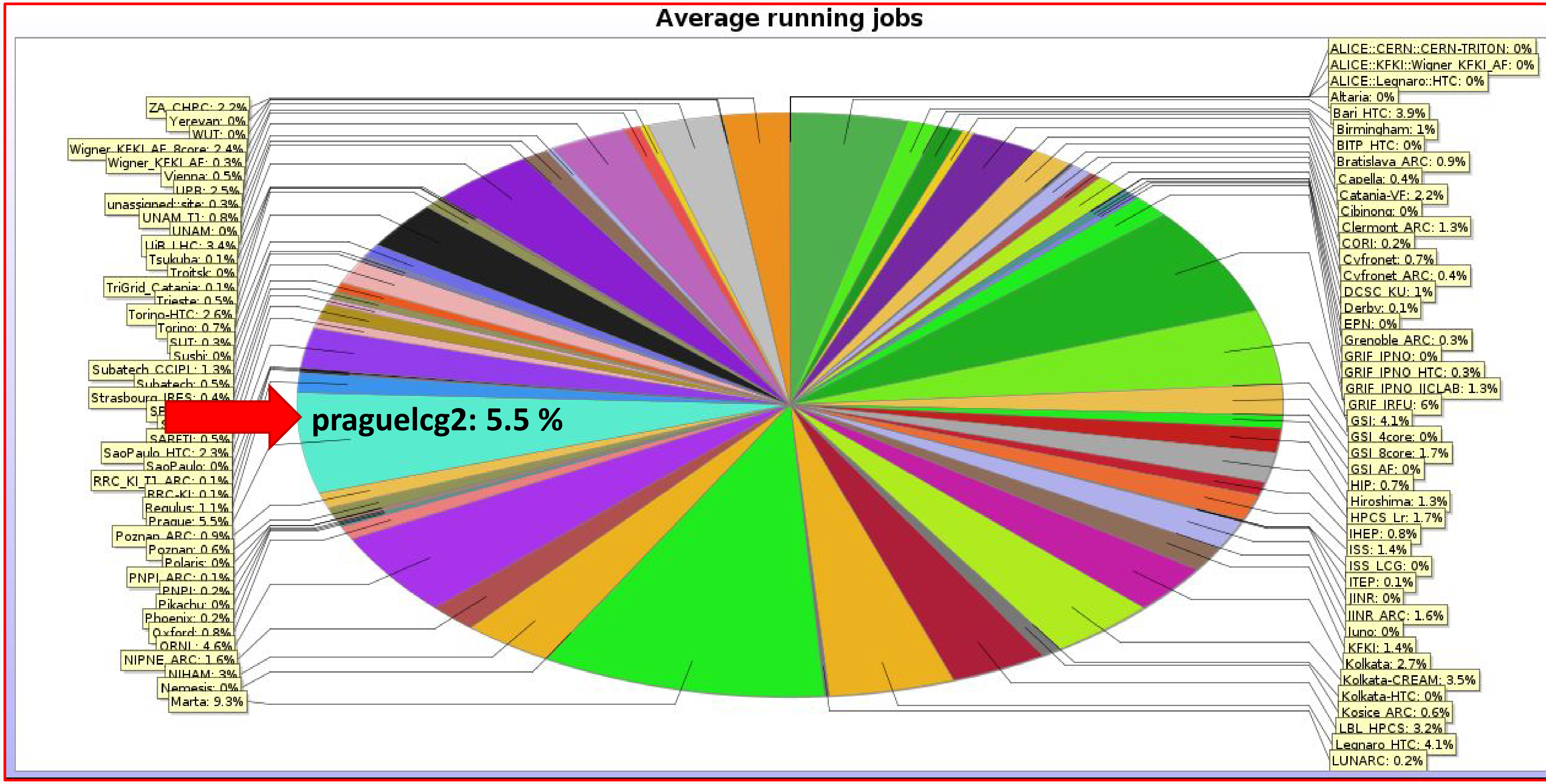
	Last *	Min	Mean	Max
ATLAS analy	4461	0	1294	5701
ATLAS singlecore	143	0	392	4493
ATLAS multicore	2960	0	1968	8280
ATLAS User	1	0	1390	8106
ALICE	4029	0	3751	8074
ALICE User	0	0	0	0

Running jobs profile for the period 1/2021 – 6/2022, by the local monitoring at the central site in Prague. ***Main CPU consumers are ALICE and ATLAS.***



Network traffic between the main site of praguelcg2 and NPI storage cluster during last 6 months. ***Maximum throughput 39.2 Gb/s.***

CPU delivery - ALICE Tier-2 share: 1/2021 – 8/2022



Czech republic share during 1/2021 – 8/2022 was ~ 5.5%.