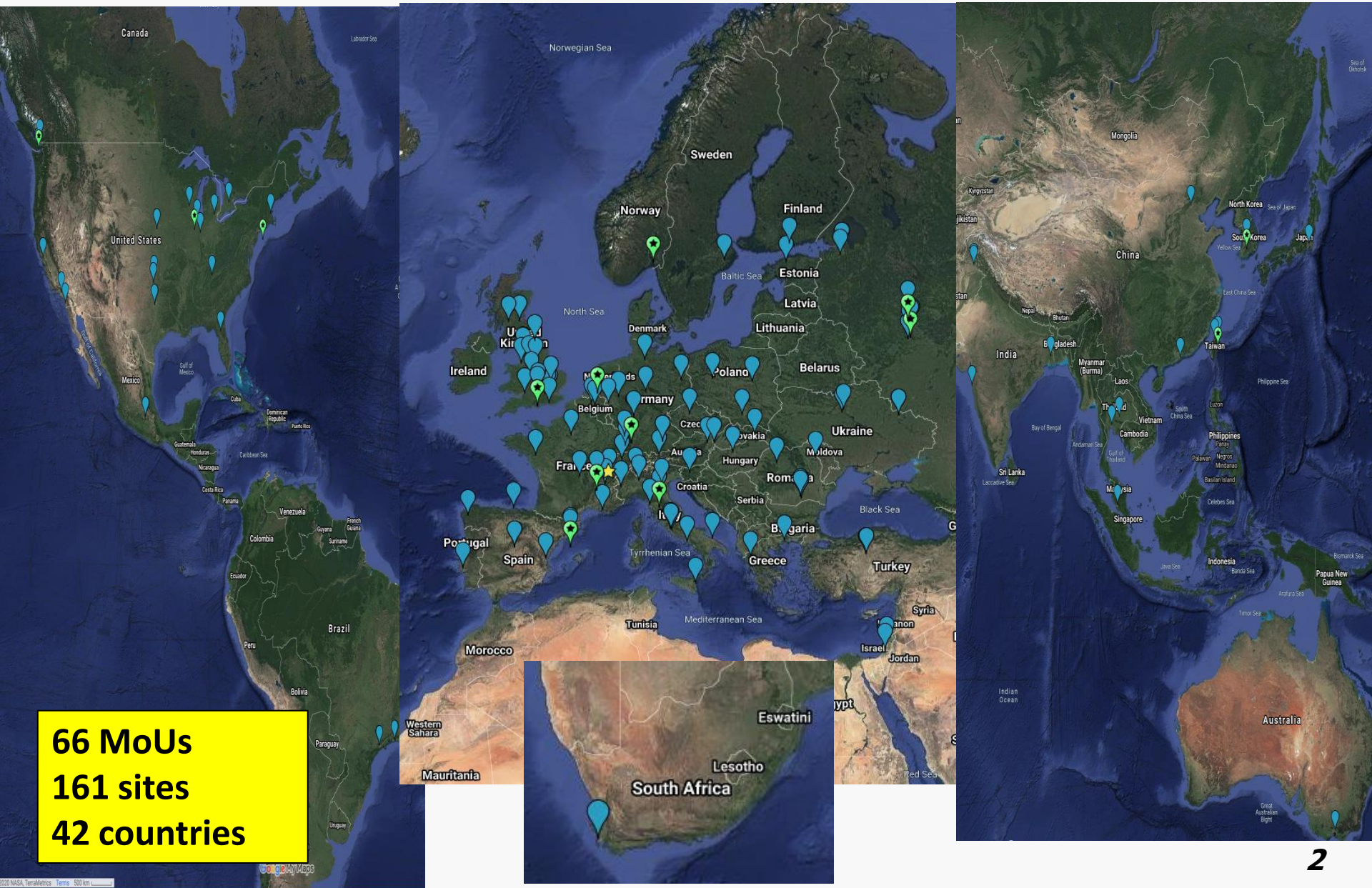
EXASCALE COMPUTING FOR HL-LHC

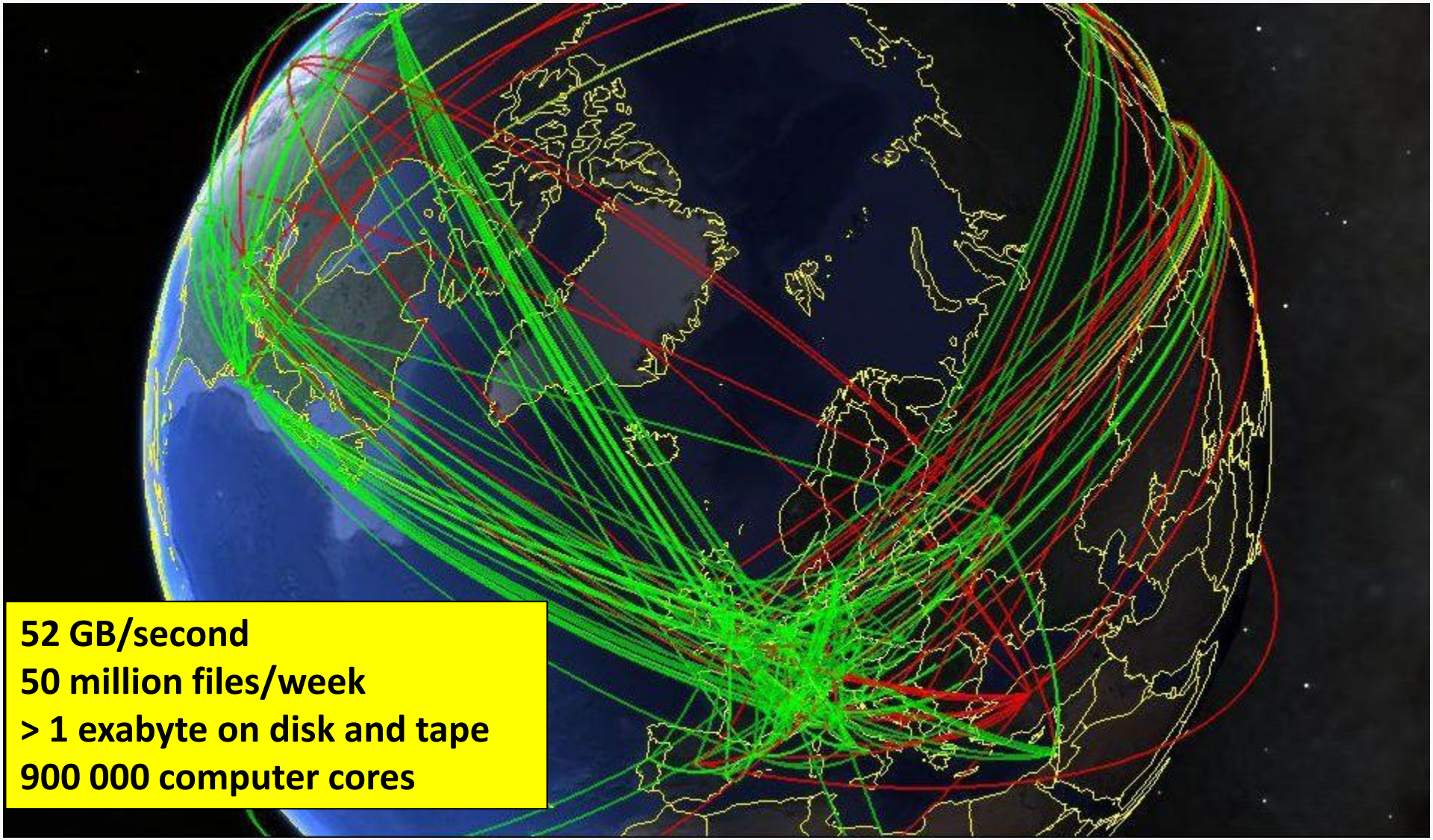
Dagmar Adamova (NPI AS CR Prague/Rez) and
Maarten Litmaath (CERN)

Worldwide LHC Computing Grid topology



66 MoUs
161 sites
42 countries

Worldwide LHC Computing Grid resources



WLCG Fabric and software

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

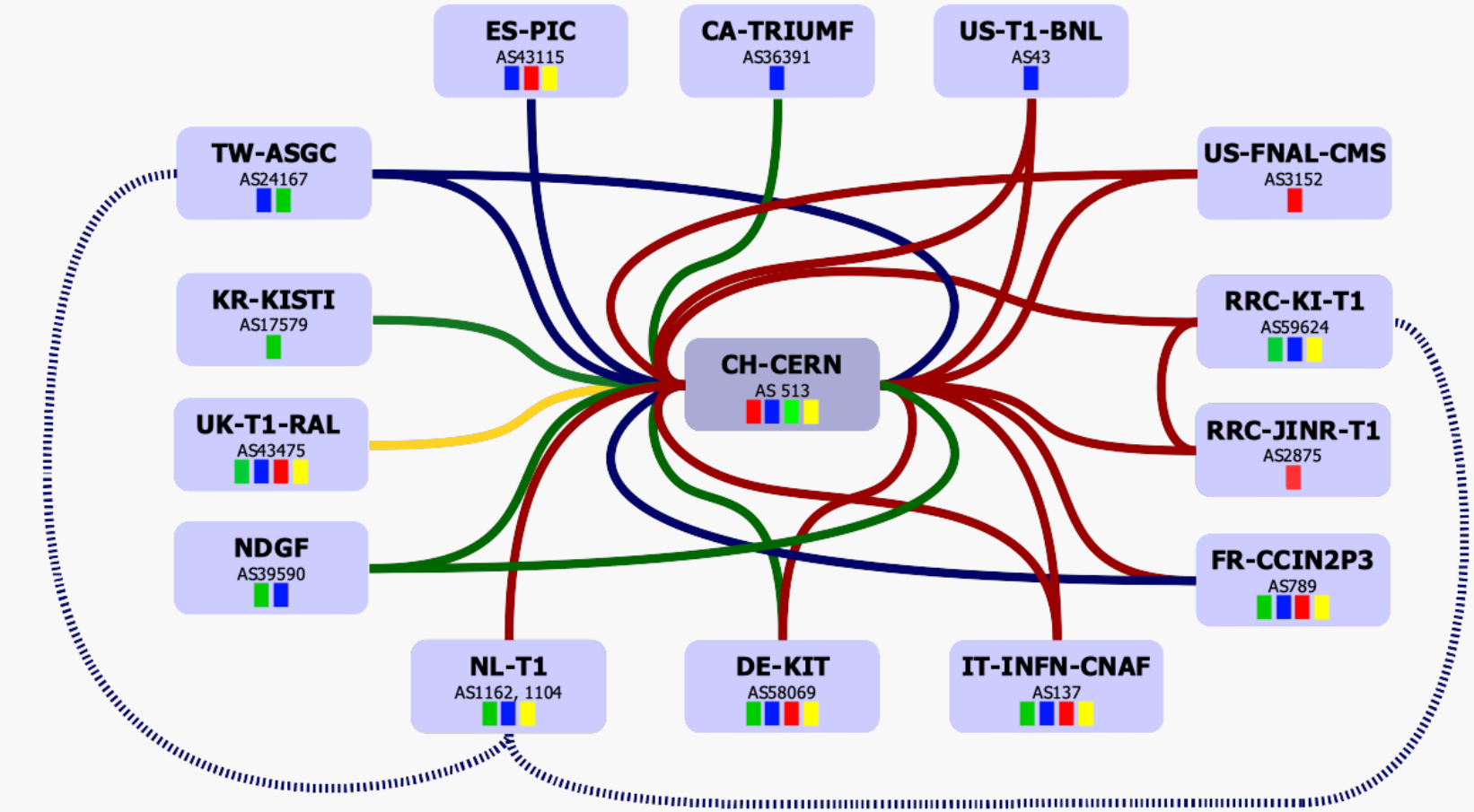
- ***Storage:*** mixture of tape and disk. Since 2018, the "Data Organization, Management and Access" (DOMA) project is active. R&D for "Data Lakes".

- ***Network:*** connection between CERN and T1s provided by a system of P2P connections of capacity of 10 Gb/s – 100 Gb/s, so called LHC Optical Private Network (LHCOPN). Most T0/1/2/3 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, originally single- but increasingly multi-threaded. Being ported e.g. to GPUs.

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

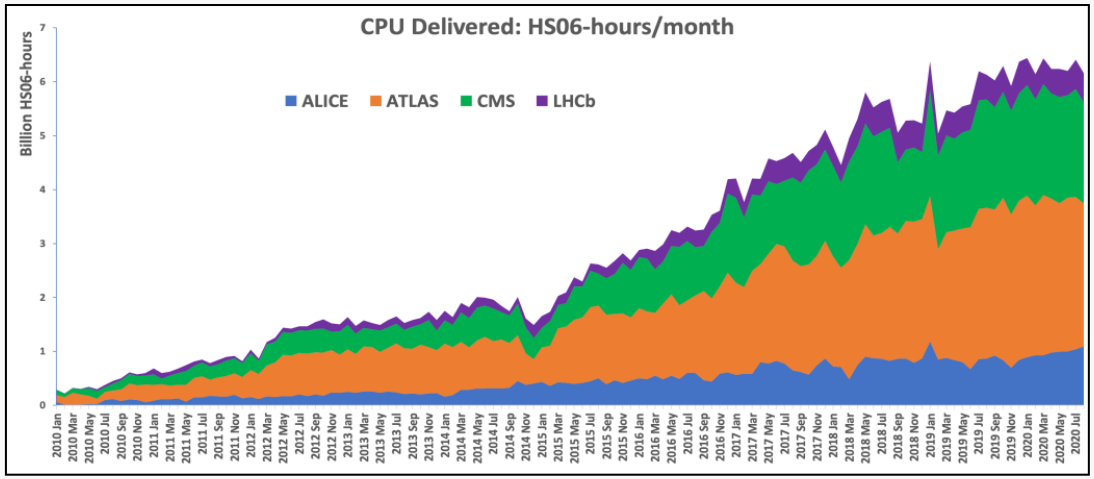
LHC Optical Private Network (LHCOPN)



WLCG usage of resources

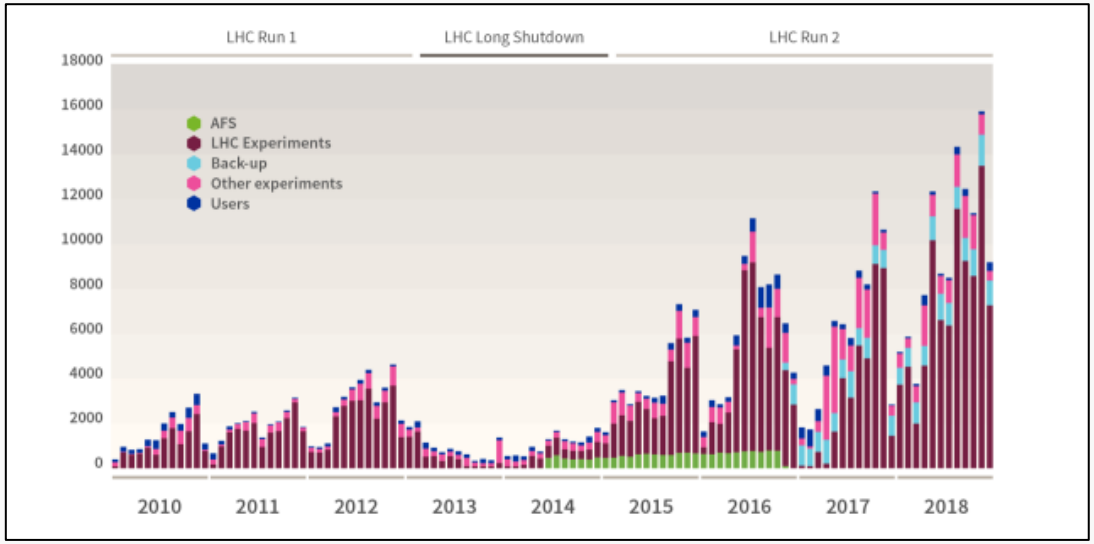
CPU use has been stable in 2020, higher than in 2018 (Run 2) and 2019 (start of LS 2)

Up to 900 000 cores running simultaneously.

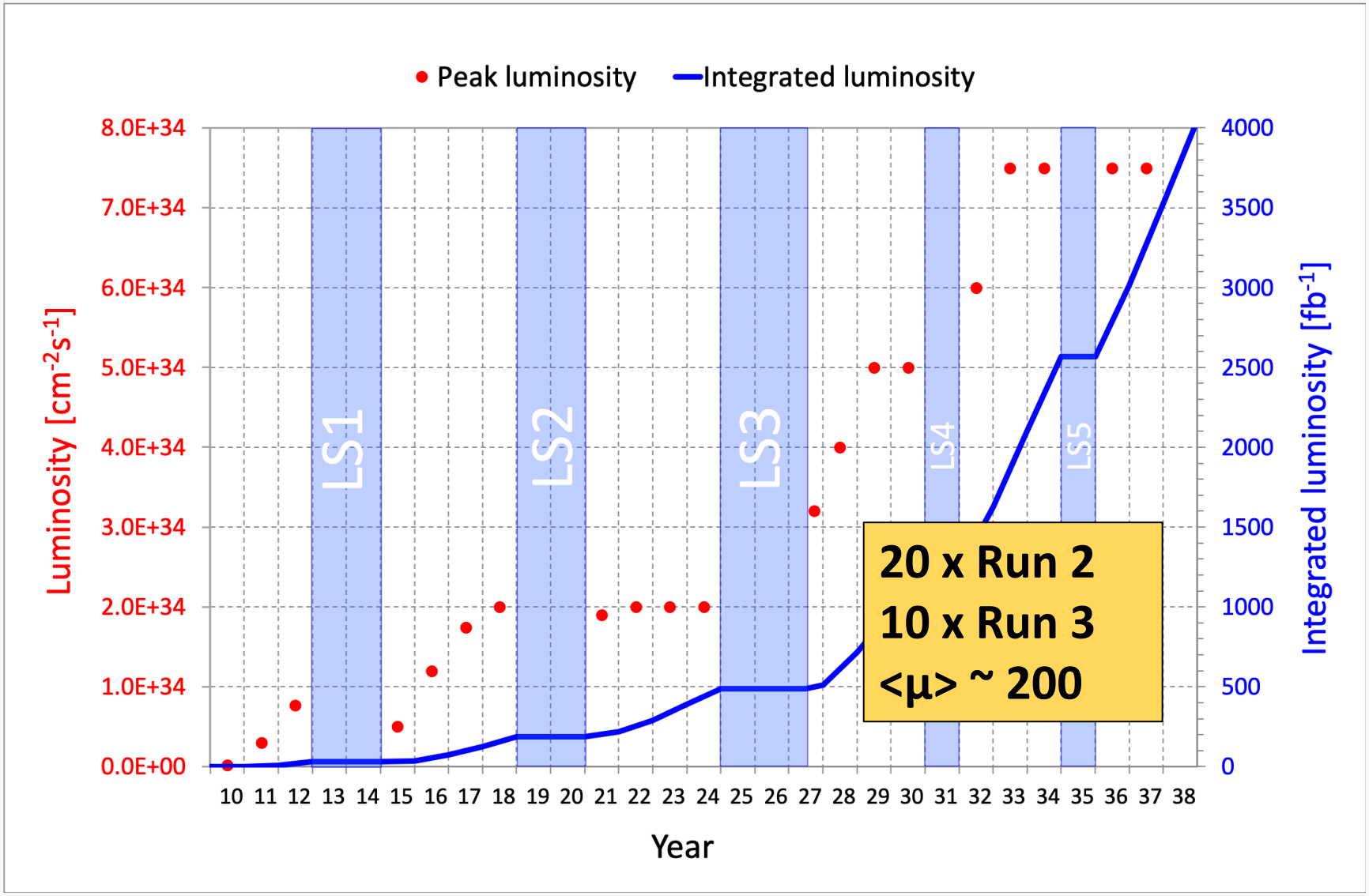


CERN computing: Data (in TeraBytes) recorded on tapes at CERN month-by-month (2010–2018)

In the end of 2018, the data stored on CASTOR reached 330 PBytes, of which *more than 200 PB was RAW data from LHC experiments*



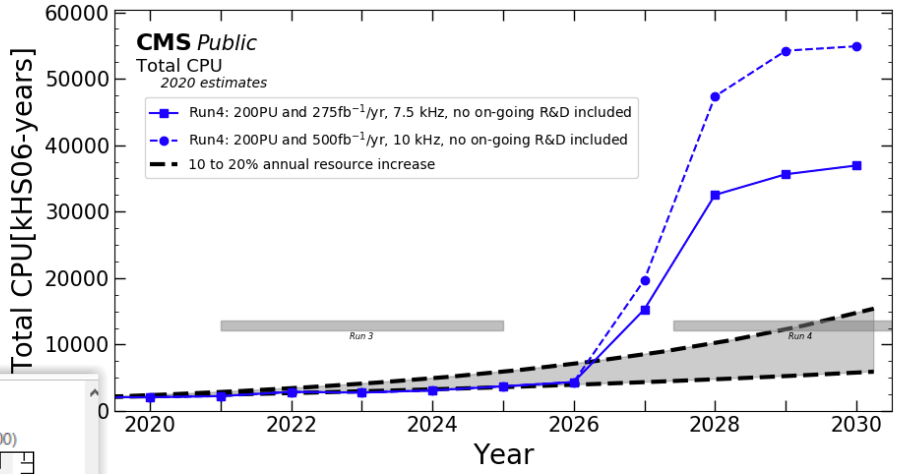
HL-LHC luminosity, current planning

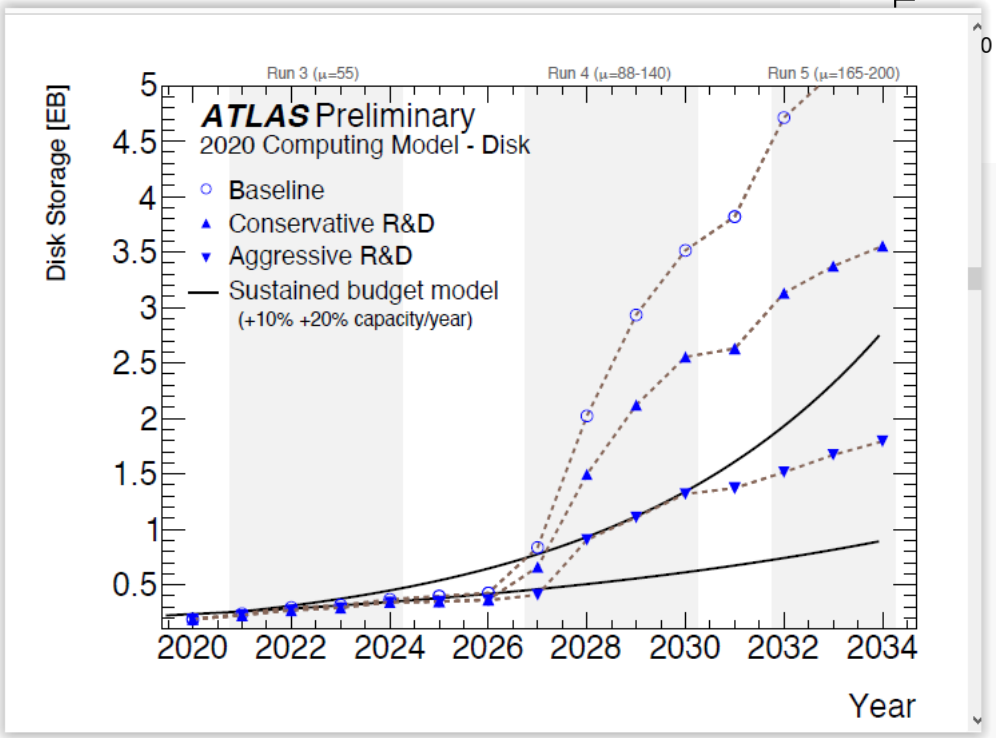


Note: ***LS2 now also includes 2021, Run 3 will be just 3 years 2022-2024***

ATLAS and CMS resource needs for HL-LHC

CMS: Anticipated growth in CPU resources towards Run 4.
Factor ~6 difference between "flat budget" and Physics needs.



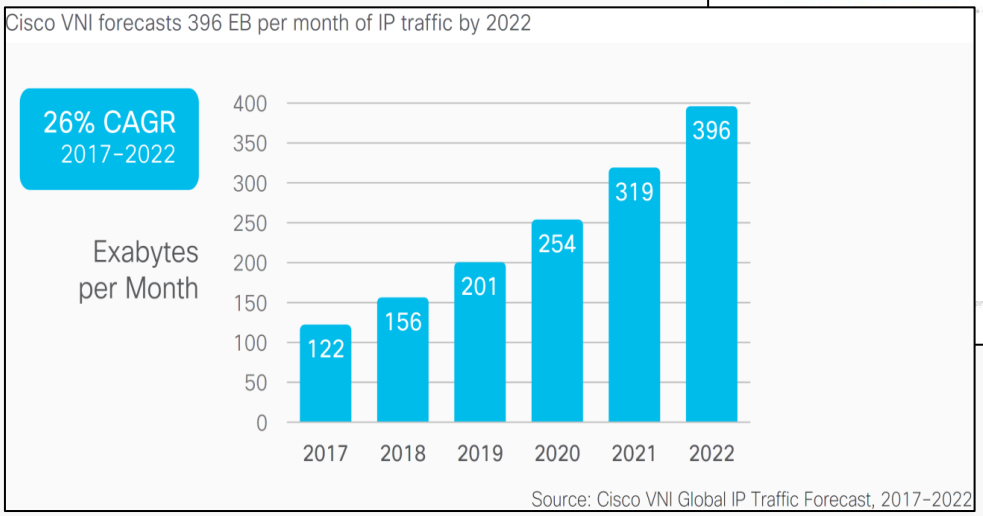


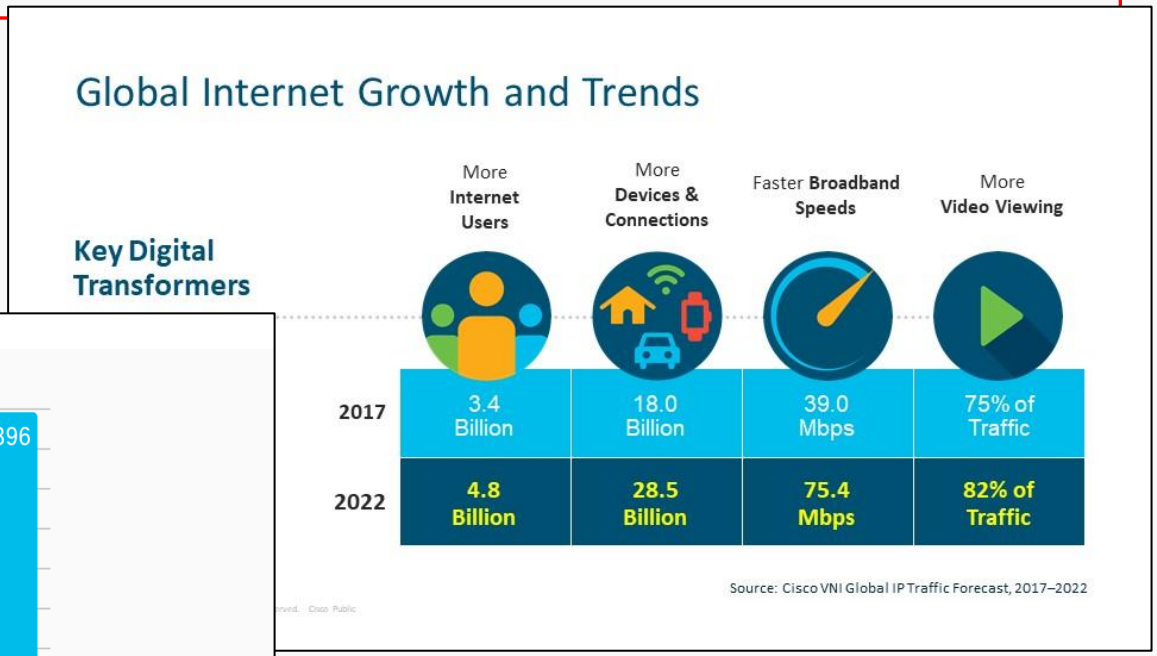
ATLAS: Anticipated growth in disk capacity towards Run 4.
Factor ~10 difference between "flat budget" and Physics needs.

A glance at the numbers in the business ecosystem

- Google, Facebook, Microsoft, and Amazon store at least ***1200 PetaBytes*** of information ***per day*** at present.
- By 2025, the amount of data generated each day is expected to reach
- ***463 ExaBytes globally.***
- The entire digital universe is expected to reach ***44 ZettaBytes by 2020.***
- If this number is correct, *it will mean there are 40 times more bytes than there are stars in the observable universe.* *(World Economic Forum)*

CISCO forecasts 396 ExaBytes per month of IP traffic by 2022.
(Source: Cisco Visual Networking Index)





Actions required to be prepared for HL-LHC

- ***Computation:*** need to have enough compute power to process raw and provide enough simulated data. Necessary to make the best of available technology. Need to be able to use concurrency. This means portability, to allow for HEP code to run on different kind of resources without being re-written. Need to be able to use external facilities which were not built just for HEP, e.g. HPC centers.

- ***Storage:*** need to find enough storage capacity and deliver processed data fast to analysts. Control demand e.g. by reducing the size of the analysis formats.

- ***Network:*** not a real problem these days. With the explosion of the network traffic due to the use of social media, the global network infrastructure became very robust and the global throughput continues to rise.

- ***Software:*** need to provide portability and work on shooting up performance.

Portability and Machine Learning

- To speed up the existing code: either optimize or use concurrency. Use of multi threading rather extensive. Done on existing x86 architecture.

- More speed-up using GPUs. Number of portability languages developed - to access heterogenous hardware - no need to re-write the existing code to be portable.

- HPCC: In July 2020, CERN together with three other research organisations formed a collaboration to deal with the challenges related to the use of high-performance computing (HPC) to support large, data-intensive science projects like WLCG. The members of the collaboration are CERN, SKAO, GÉANT and PRACE (the Partnership for Advanced Computing in Europe).

- Machine learning:
  - has been used in the HEP community since the LEP days.
  - recently huge advances in deep learning – rely on large neural networks
  - software for building complex neural networks available.
  - deep learning used in HEP simulation and analysis.

Simulation

Detector simulation is the largest consumer of CPU time in HEP.

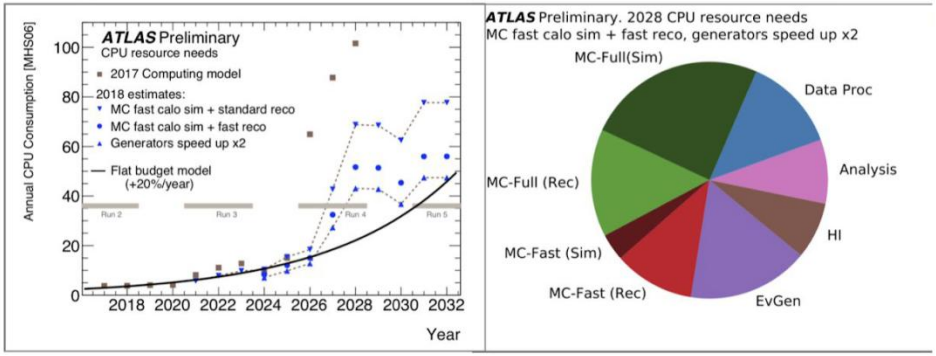
Monte Carlo simulations are highly parallelizable, which makes them a great target for GPU computation.

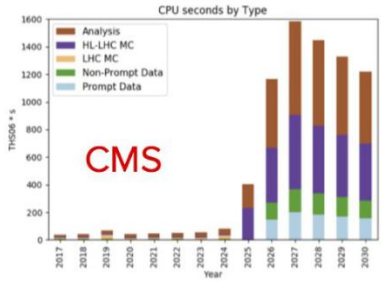
Forecast Simulation Needs

Many physics and performance studies require large datasets of simulated events

ATLAS

- Geant4 is highly CPU-intensive
- Already lacking statistics -- increasing luminosity poses greater challenges





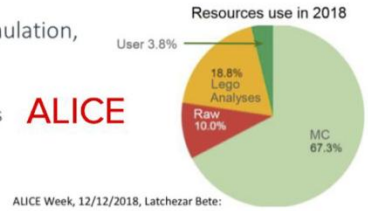
- Simulate more events to keep up with HL-LHC data volumes: 10×(Phase1)
- May also need to improve accuracy of physics lists to simulate HGCal
- Reconstruction will take longer due to high pileup and granular detectors

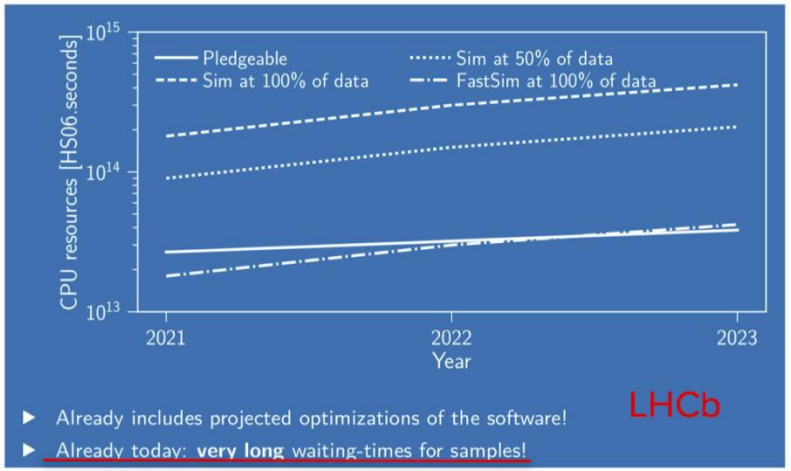
➢ Need more events, more accuracy, in more complicated geometry… w/ relatively smaller fraction of total CPU usage

- 2/3 of the computing resources are dedicated to MC simulation, all full sim
  - fast sim not used in production yet
  - fully parametrised fast simulation approach for upgrade studies

ALICE

- expected 10-100 times more data in Runs 3 and 4
  - cannot cover that with current usage of full sim





Simulation, Fast simulation and ML

To speed-up simulation:

- Optimization of the current Geant4 code to run faster
- Adapt simulation to heterogenous architectures like GPUs
- Plug trained network back into simulation via Geant4 ML models

Fast simulation is usually

- Experiment dependent
- Simplified, automatic easy way to extract parameters

Improve Fast simulation:

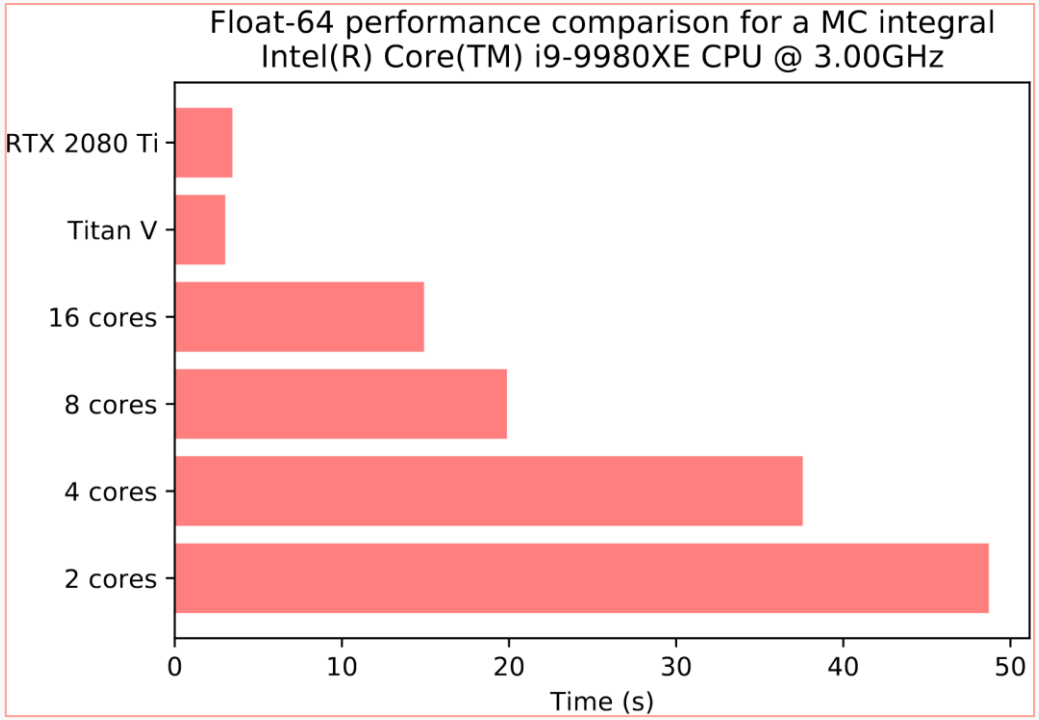
- Use of Deep Learning
- Use of Library of deposits
- Improve Physics fidelity of Fast simulation

Fast Simulation Parametric and Machine Learning

- GFlash - historical parameterization of calorimeter energy deposition.
  Working to improve both the speed and accuracy of the implementation.
- Auto-regressive neural networks training on different calorimeter data.

Parton-level Monte Carlo generators

- Currently, the MC generators are not dramatic CPU consumers but this will change in the time of HL-LHC due to huge pile-up.
- Behind most predictions for LHC phenomenology lies the numerical computation of basic integrals computed numerically using MC methods.
- ***GPU computation can increase the performance of the integrator by more than an order of magnitude.***

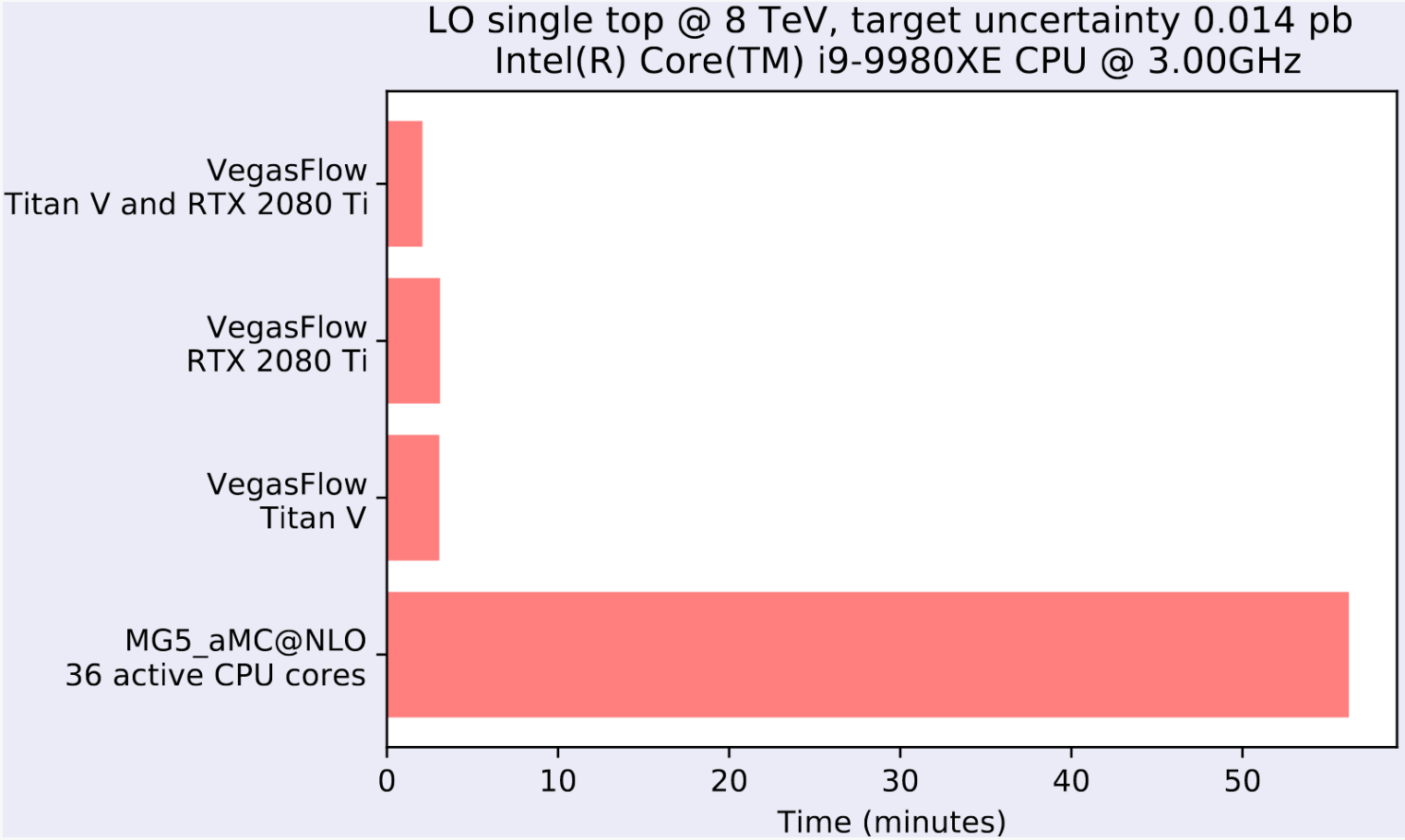


Monte Carlo integration of
n-dimensional gaussian function

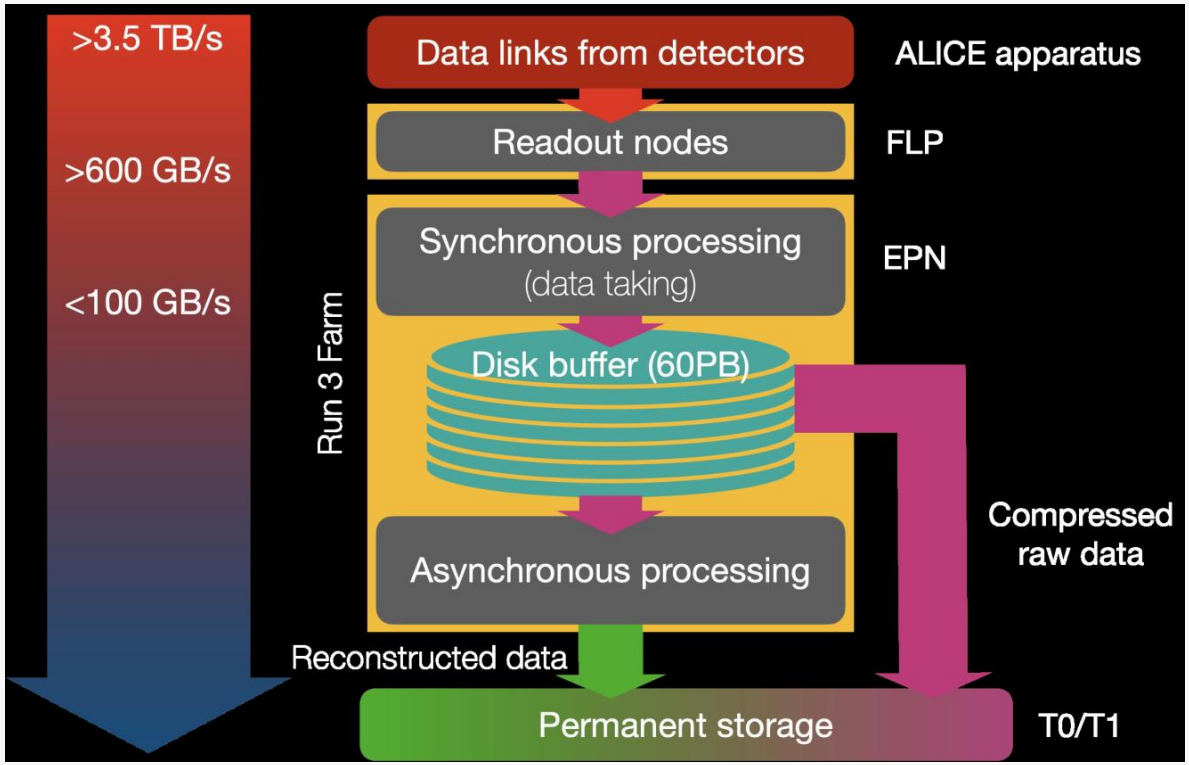
$$I = \int \mathrm{d}x_1 \ldots \mathrm{d}x_n \, e^{x_1^2 + \cdots + x_n^2}$$

Generators: Running on GPUs with VegasFlow

- VegasFlow: Framework for evaluation of high dimensional integrals based on MC algorithms.
- ***For Leading Order: ported an old Fortran code to GPU. No GPU-specific optimization.***



ALICE Data Processing – Run3 and Run4



Synchronous processing

Goal of synchronous reconstruction is to reach **factor 35 of compression**.

Most relevant detector is TPC: **from 3.4 TB/s to 70 GB/s.**

Efficient usage of accelerators allows to **trade between 40 and 150 CPU cores for a single graphics card.**

Online reconstruction on the EPN farm

Event processing nodes equipped with GPUs (up to a few thousand)

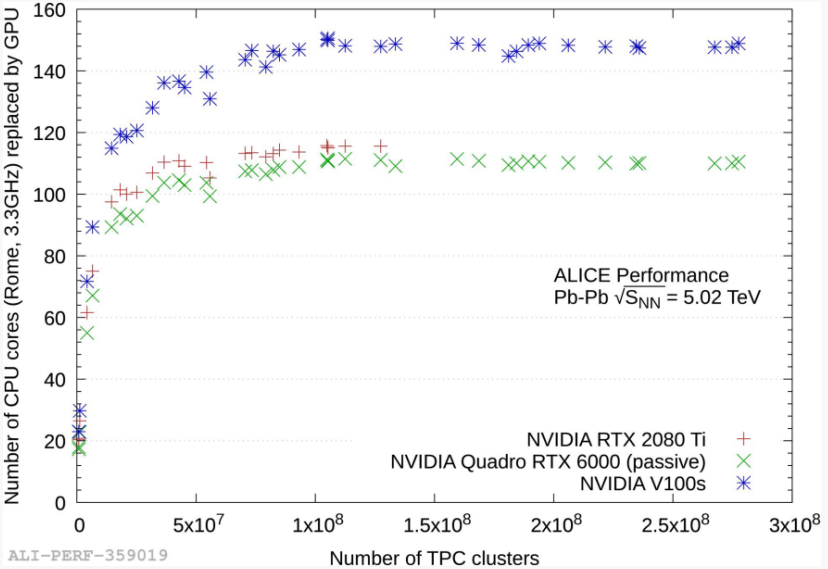
Synchronous processing: during data taking, main user: TPC, 100% TPC standalone tracking on GPUs.

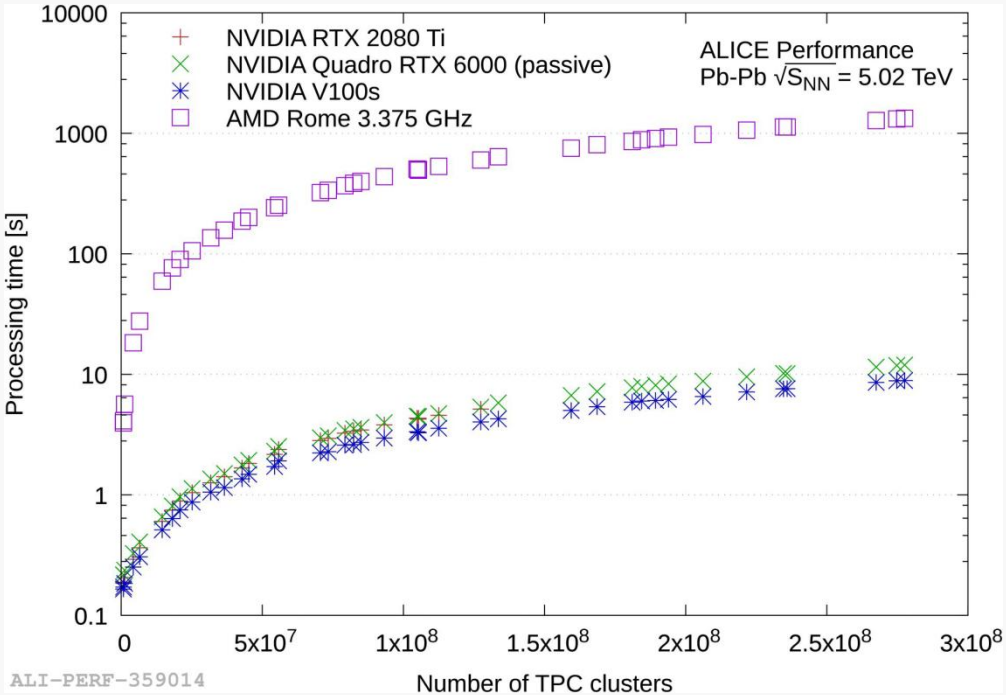
Asynchronous: during no-beam periods and pp collisions

ALICE GPU tracking: performance

40-150 CPUs replaced by one GPU

TPC tracking speeded up by factor of 50-100



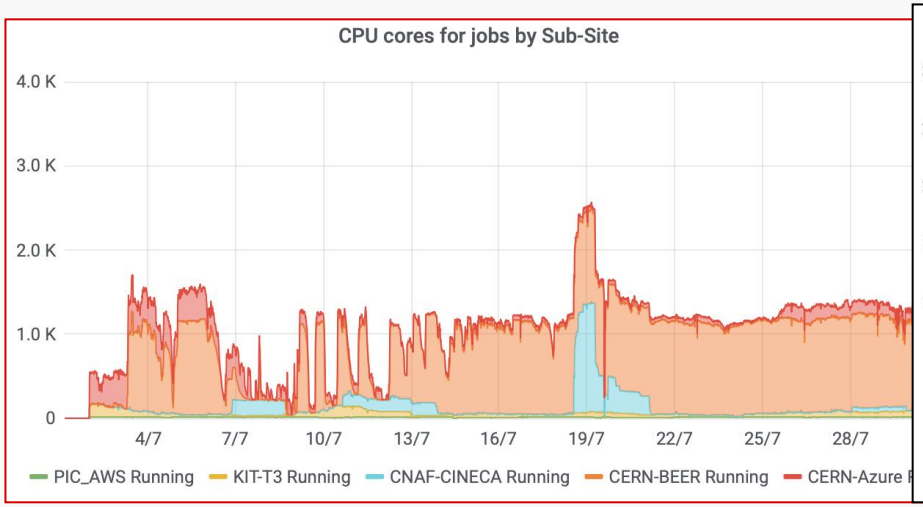


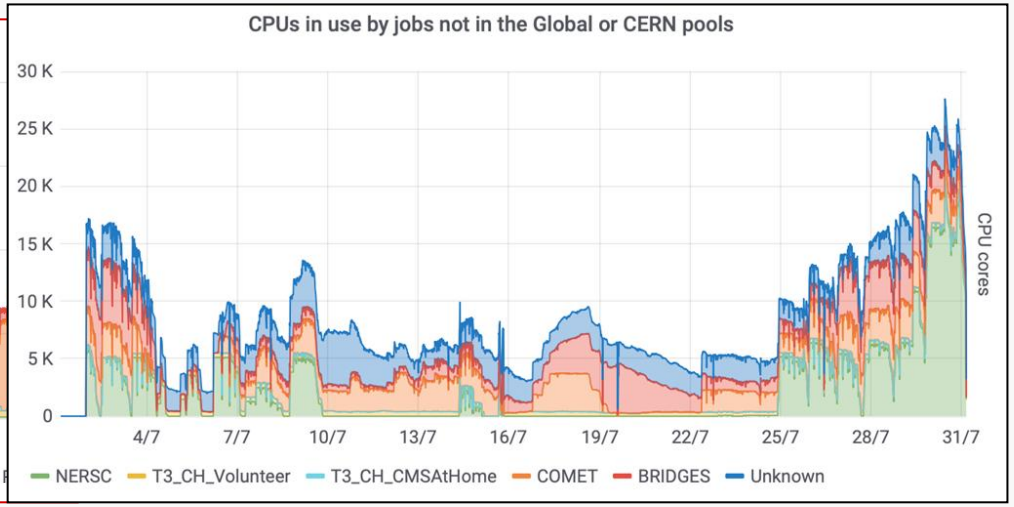
CMS: Dynamic heterogenous resources integration

Recent progress in the integration of new resources into the CMS Global Pool and for CMS use:

- ***HPC***: via GlideinWMS pilot submission (CINECA) or integrated to HEPCloud (NERSC, etc)
- ***Cloud***: as extension of Grid sites (CERN_Azure and PIC_AWS)
- ***Opportunistic use of clusters***: CERN_BEER and Research or University campus (e.g. at Purdue)

Non-standard resources require enhanced workload-to-resource matchmaking: working on an expanded description of jobs and resources for flexible and efficient scheduling (e.g. select no-input data tasks etc.)





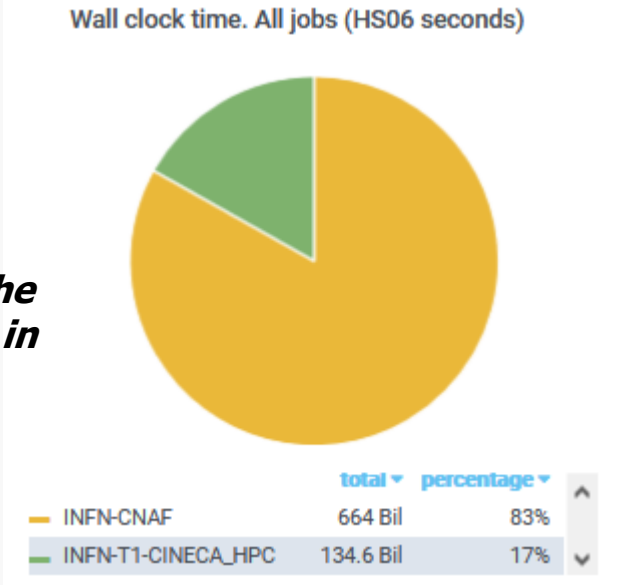
Integration of the HPC resources into WLCG ecosystem

Motivation **for the use/integration of HPC resources:**

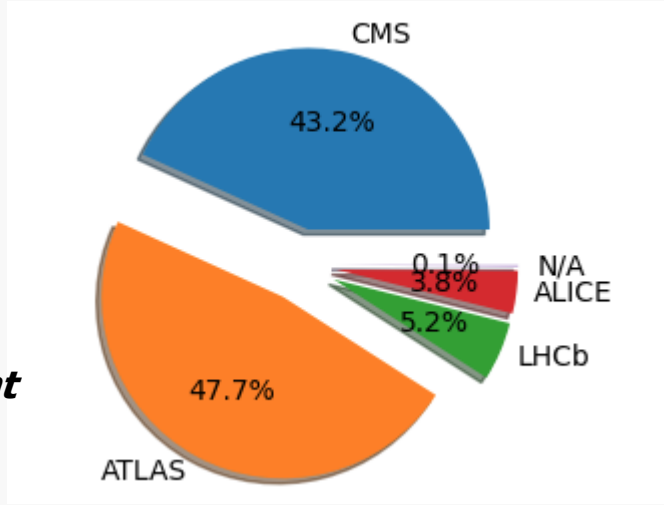
- **Growing funding in HPC** infrastructures looking onwards to deploying Exascale machines
- Countries/Funding agencies **pushing HEP communities to make use of these resources**
- Interest in HEP experiments **to access best technologies available**, usually employed at HPC sites
- HPC contribution in the future regarded as **integral part of WLCG strategy towards HL-LHC**

Integration of CNAF (WLCG T1) HPCC CINECA (the largest Italian HPCC provided by PRACE) into a virtual facility, transparent (as much as possible) to the experiments and WLCG.

ATLAS: CINECA provided ~ 17% of the CNAF HS06 in 3 months

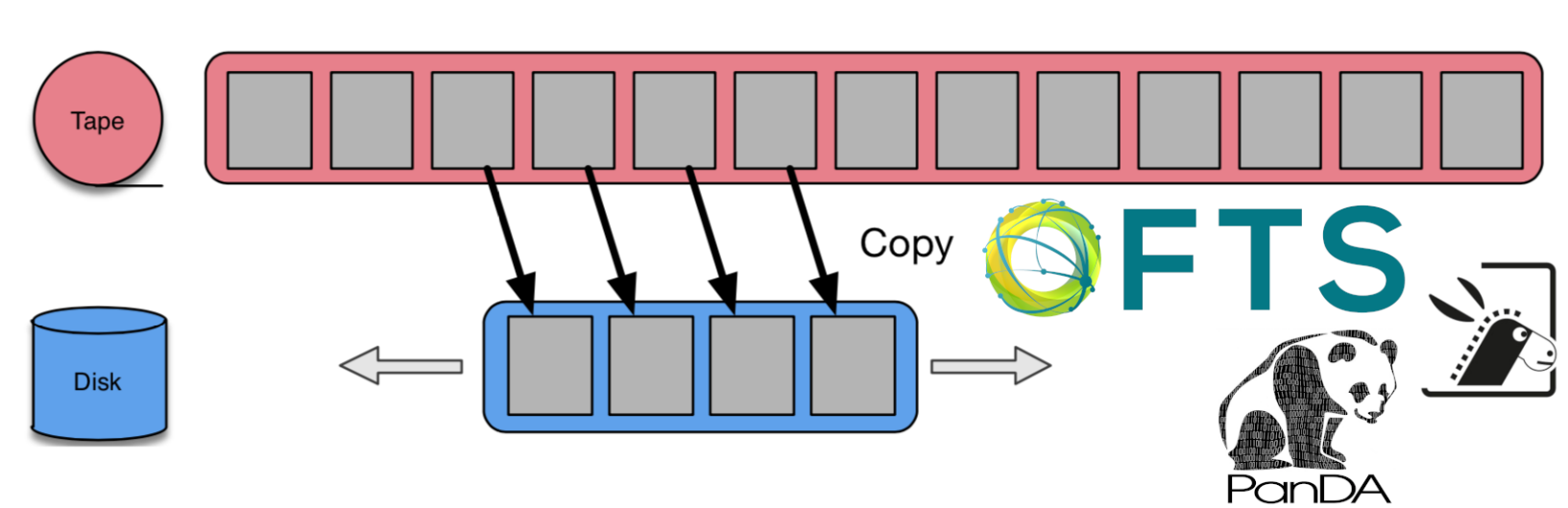


CINECA utilization per experiment



ATLAS data treatment: DataCarousel

- ***Objective:*** *use tape as input for I/O intensive workflows to reduce the expensive storage of data on disks*
- It is a sliding window approach to orchestrate *data processing with the majority of data resident on tape storage*
- The processing is executed by *staging the data onto disk storage and promptly processing them*
- *Only the minimum required input data are located on disk at any time*
- Tested on full Run2 RAW data reprocessing (18 PB staged over several weeks)



Conclusions

- *WLCG infrastructure is running smoothly in LS2*
- COVID-19 has had only little impact on operations so far

- 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.
- Fortunately, there are *tools available for us* to deal with the challenges .
- Funding agencies and institutes must realize that *computing and software are as important for physics as detector development and construction*.
- *Flat budget* planning is appreciated very much but *may not be sufficient* to take full advantage of the HL-LHC potential.
- Computing systems for HEP now require detailed project planning, management and sustainability over coming years.
- *WLCG engaged with other HEP experiments* (DUNE, Belle 2) *and communities* (astronomy) *to* collaborate on evolving the infrastructure according to the challenges ahead.

REMINDER

"The amount of data that experiments can collect and process in the future will be limited by affordable software and computing, not by physics"

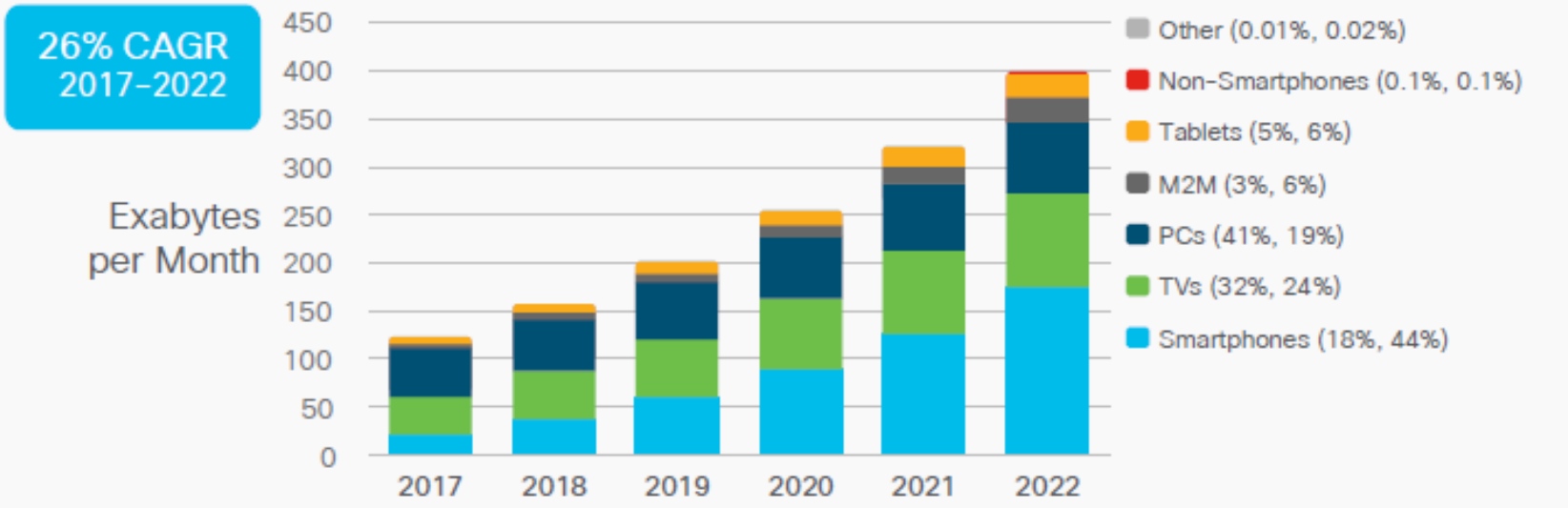
Cited from "Community White Paper" of the HEP Software Foundation : http://hepsoftwarefoundation.org/activities/cwp.html

BACKUP SLIDES

Future of global network traffic

Global IP traffic by devices

Figure 4. Global IP traffic by devices



* Figures (n) refer to 2017, 2022 traffic share

Source: Cisco VNI Global IP Traffic Forecast, 2017–2022

Future of global network traffic

Global IP traffic will more than triple

Global IP traffic is expected to reach 396 ExaBytes per month by 2022, up from 122 ExaBytes per month in 2017. That's 4.8 ZettaBytes of traffic per year by 2022.

By 2022, the busiest hour of internet traffic will be six times more active than the average. Busy hour internet traffic will grow by nearly five times (37 percent CAGR) from 2017 to 2022, reaching 7.2 PetaBytes per second by 2022. In comparison, average internet traffic will grow by nearly four times (30 percent CAGR) over the same period to reach 1 PetaBbyte by 2022.