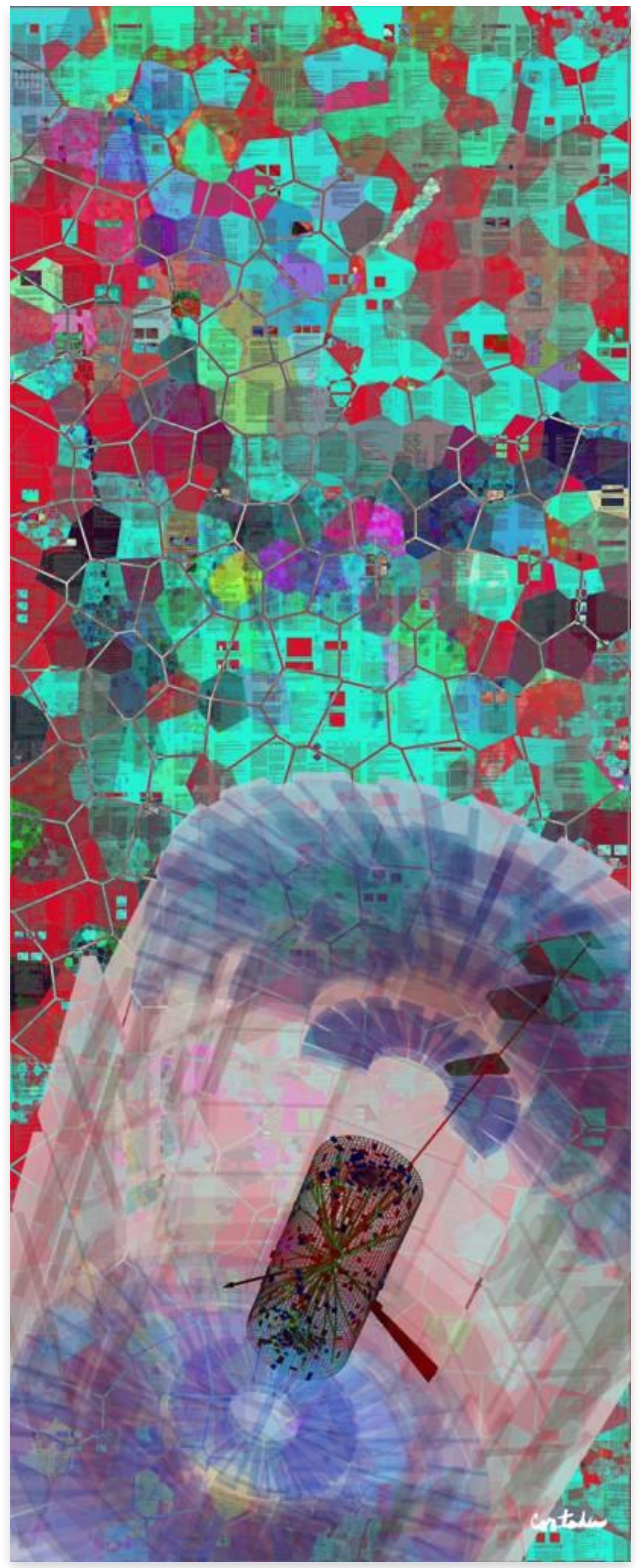
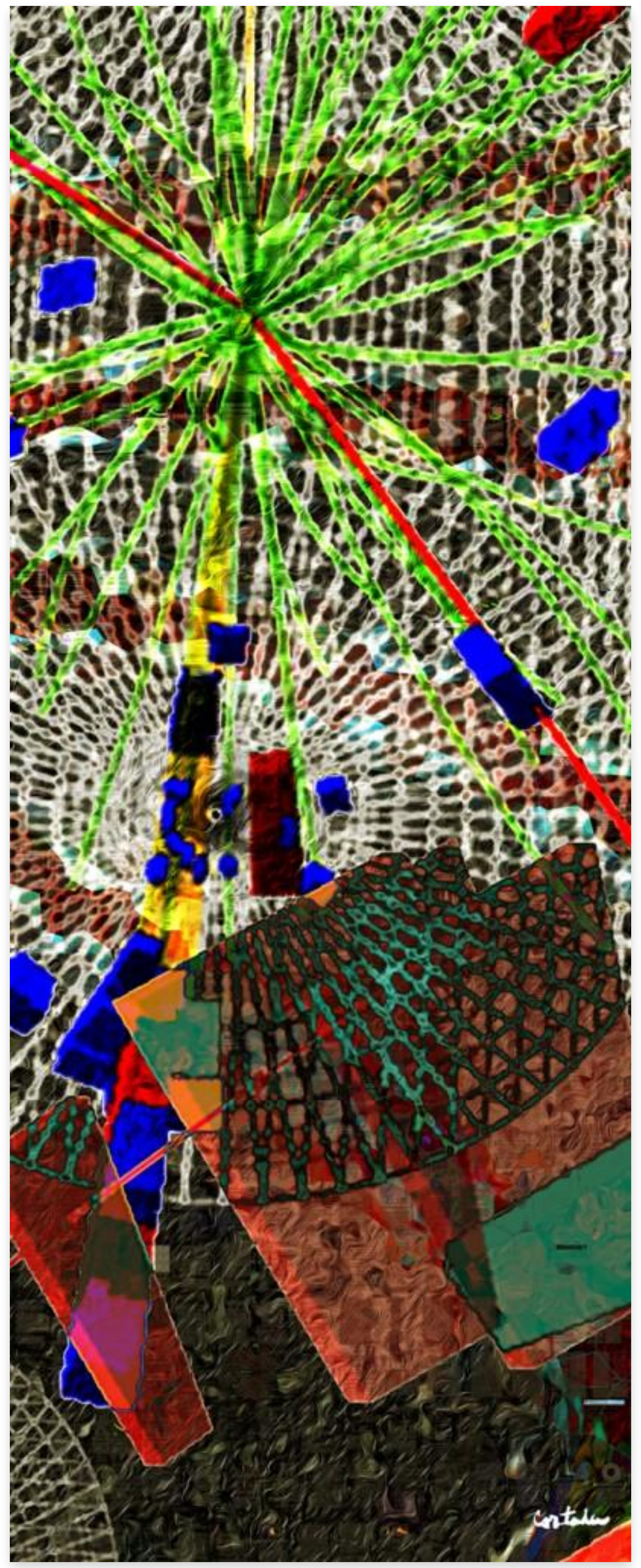
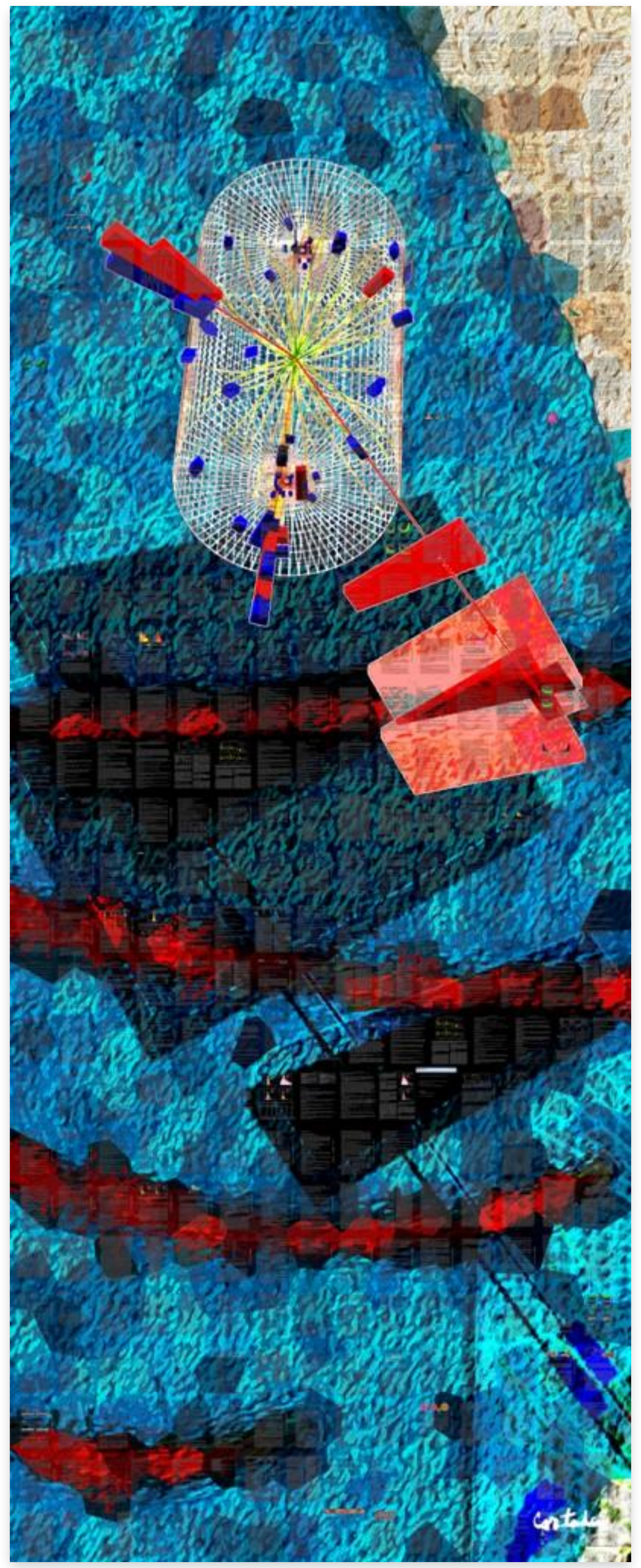
HL-LHC Computing

Oliver Gutsche (Fermilab)
U.S. CMS Undergraduate Summer Internship
July 5th, 2023

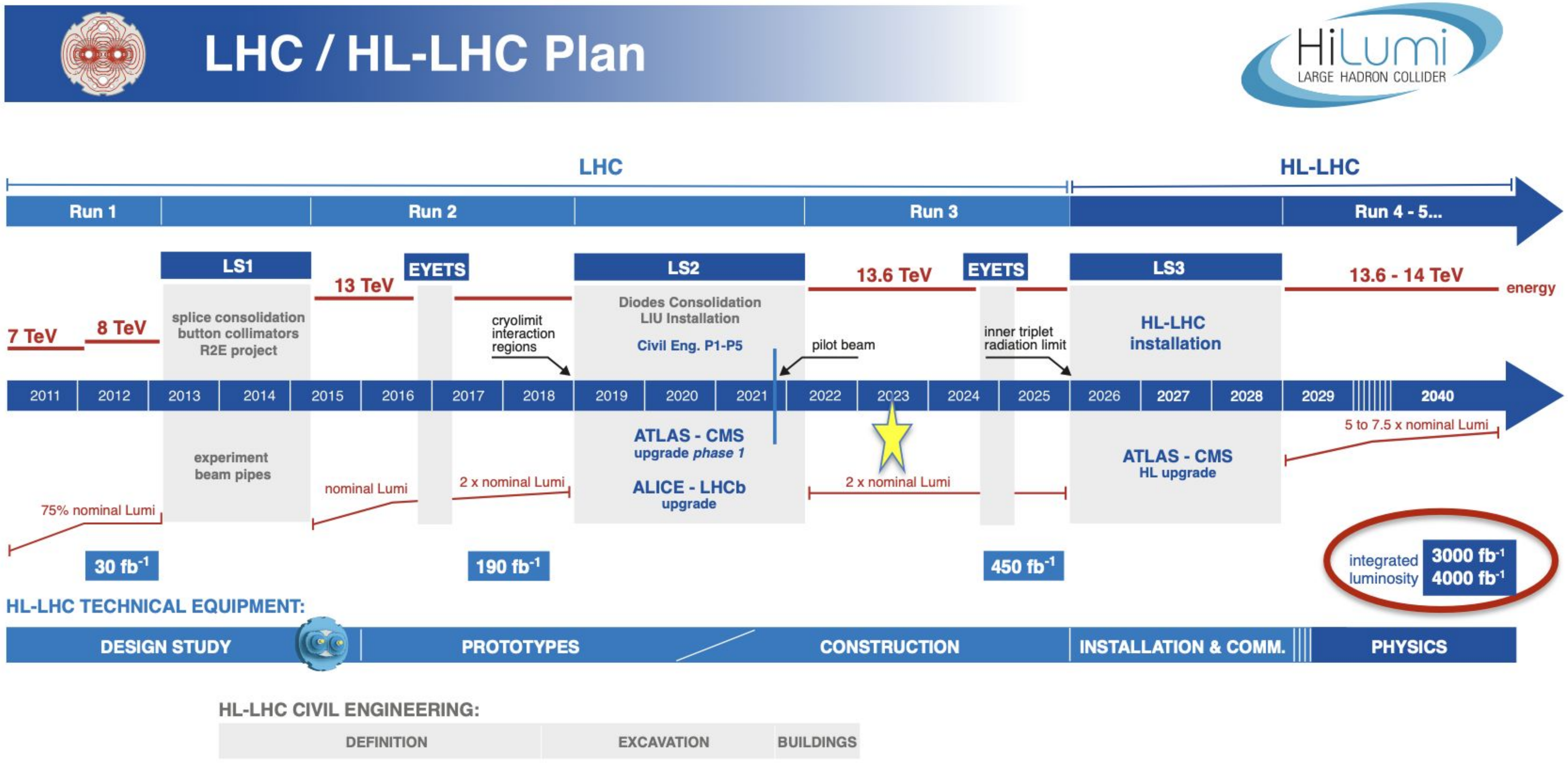



Oliver Gutsche

- Staff scientist at Fermilab ➜ Particle Physicist
  - Searching for Physics Beyond the Standard Model
  - Involved in Computing since the beginning of LHC
    - Managed Operations of Computing in the lead-up to the Higgs discovery
    - Then moved into managing the U.S. contribution to CMS Software and Computing: U.S. CMS Software & Computing Operations Program
  - Getting involved in Computing for future colliders
    - FCC, ILC, etc.
  - *More of an infrastructure person, but I know a lot about everything*

- After work, I explore Asia, Europe, America and all other parts of the world with my wife (btw., she is an astro-particle physicist looking at Galaxy Clusters)

The challenge!

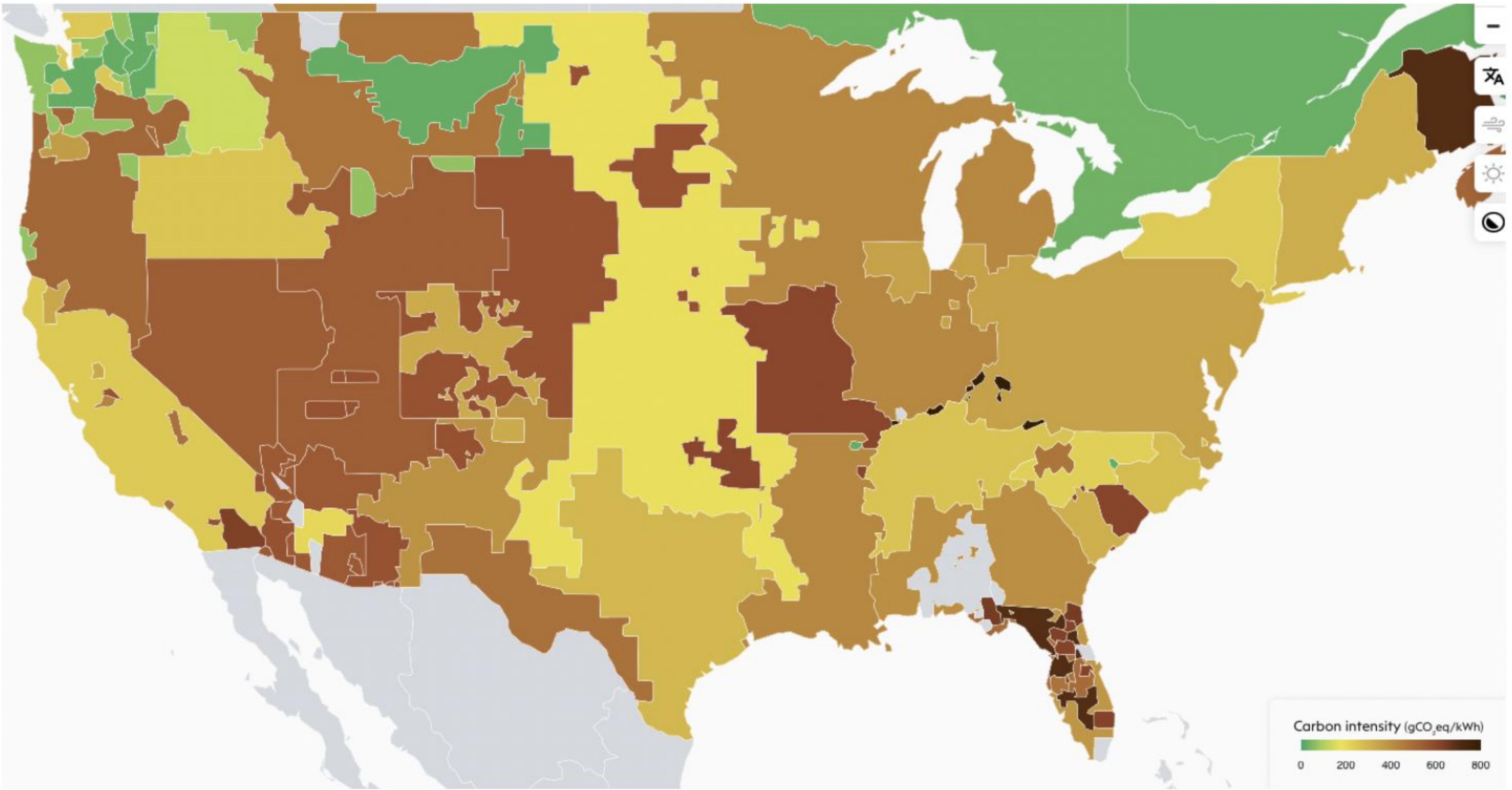


Note 95% of the total LHC data still to come (and be studied)!

From Tulika's Physics Overview talk in this talk series from June 5th.

Emissions from computing

- Data centers and computing contribute 2-4% of global GHG emissions, only expected to grow.
- Up-front considerations: where do we place computing facilities and how are they powered? Electricity emissions vary significantly across regions.
- But if electric grid is decarbonized, electricity supply might be biggest concern.

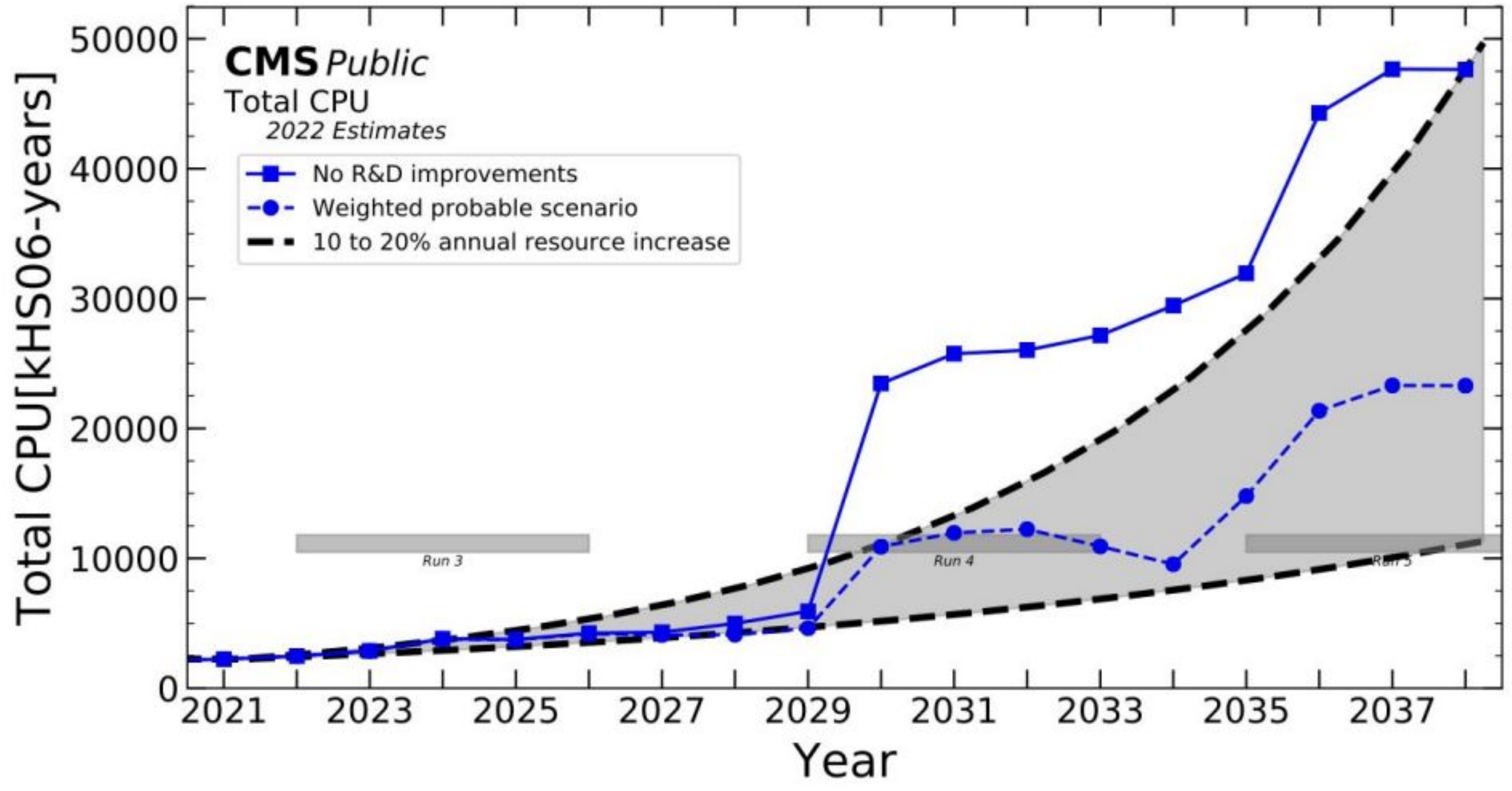


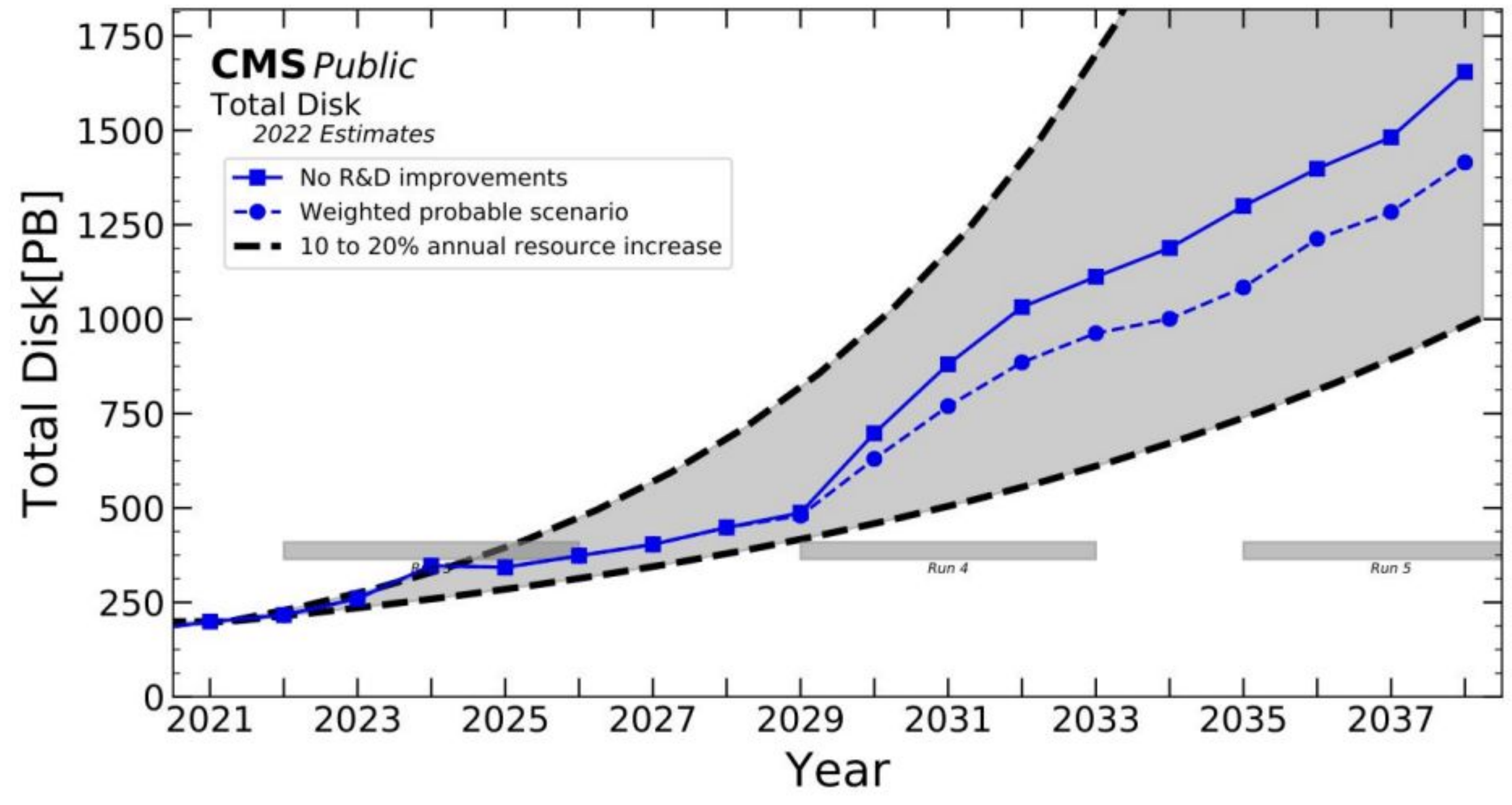
10 **electricitymap.org**

From Ken's "How to do Particle Physics in a Climate Emergency?" talk in this talk series from June 9th

High Luminosity LHC (HL-LHC)

- High Luminosity LHC (HL-LHC)
  - Next phase of the science harvest @ CERN: 2029-2042
  - 95% of the Integrated Luminosity of the LHC
  - Higher Intensity Proton-proton collisions
  - New CMS detector components with higher granularity and more channels
- Unprecedented Computing Needs compared to today
  - Number of events to be processed each year larger by **x3**: 150 Billion events
  - Size per event larger by **x5**: disk storage needs reaches ½ exabyte by 2030
  - Most data is active: needs to be held on quasi-randomly accessible storage systems to be processed by hundreds of simultaneous processing pipelines
  - Physics software: more than 4 million lines of highly specialized code with high algorithmic complexity and low computational intensity ➜ providing unique challenges to use accelerators.





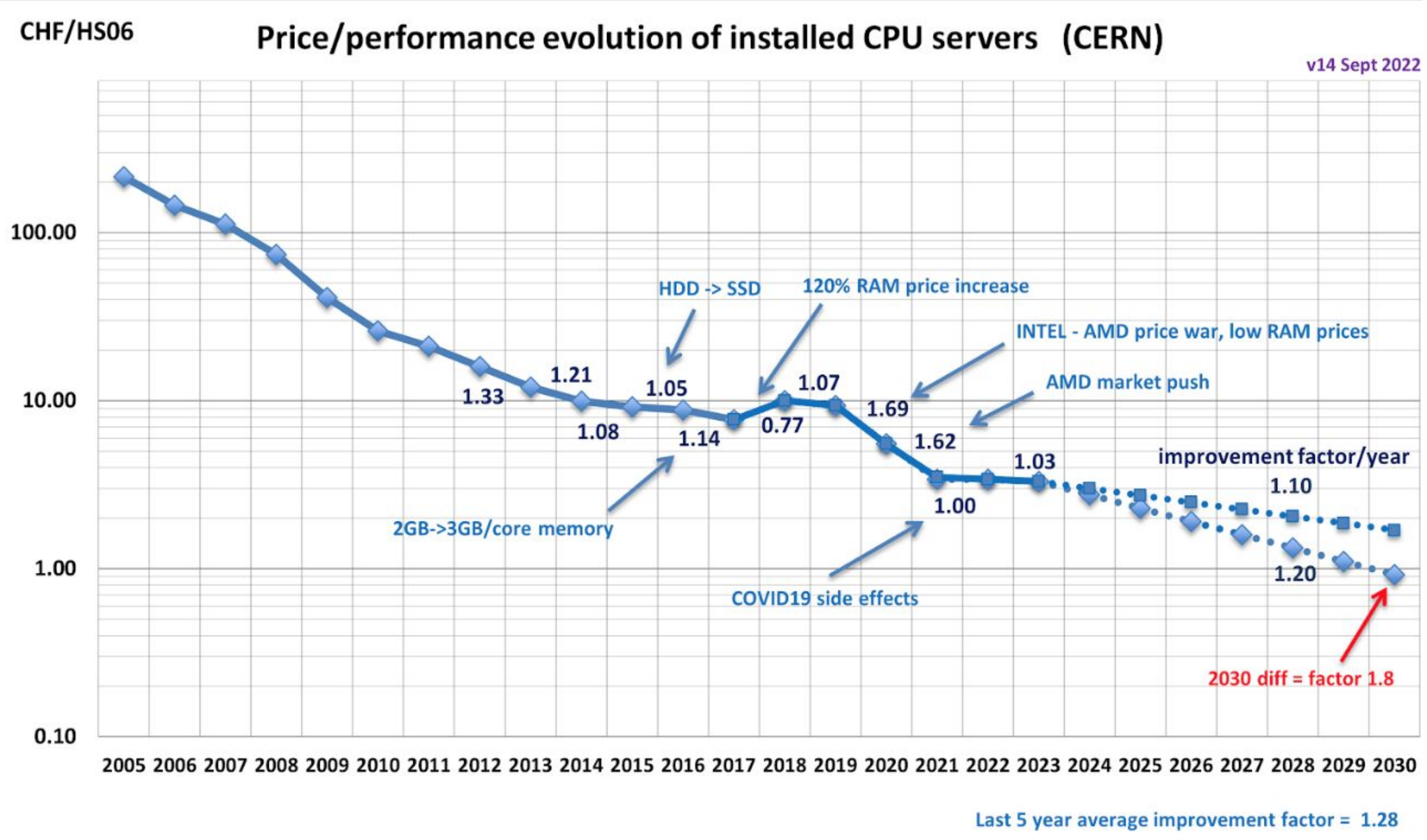
A (high level) Computing Model for HL-LHC

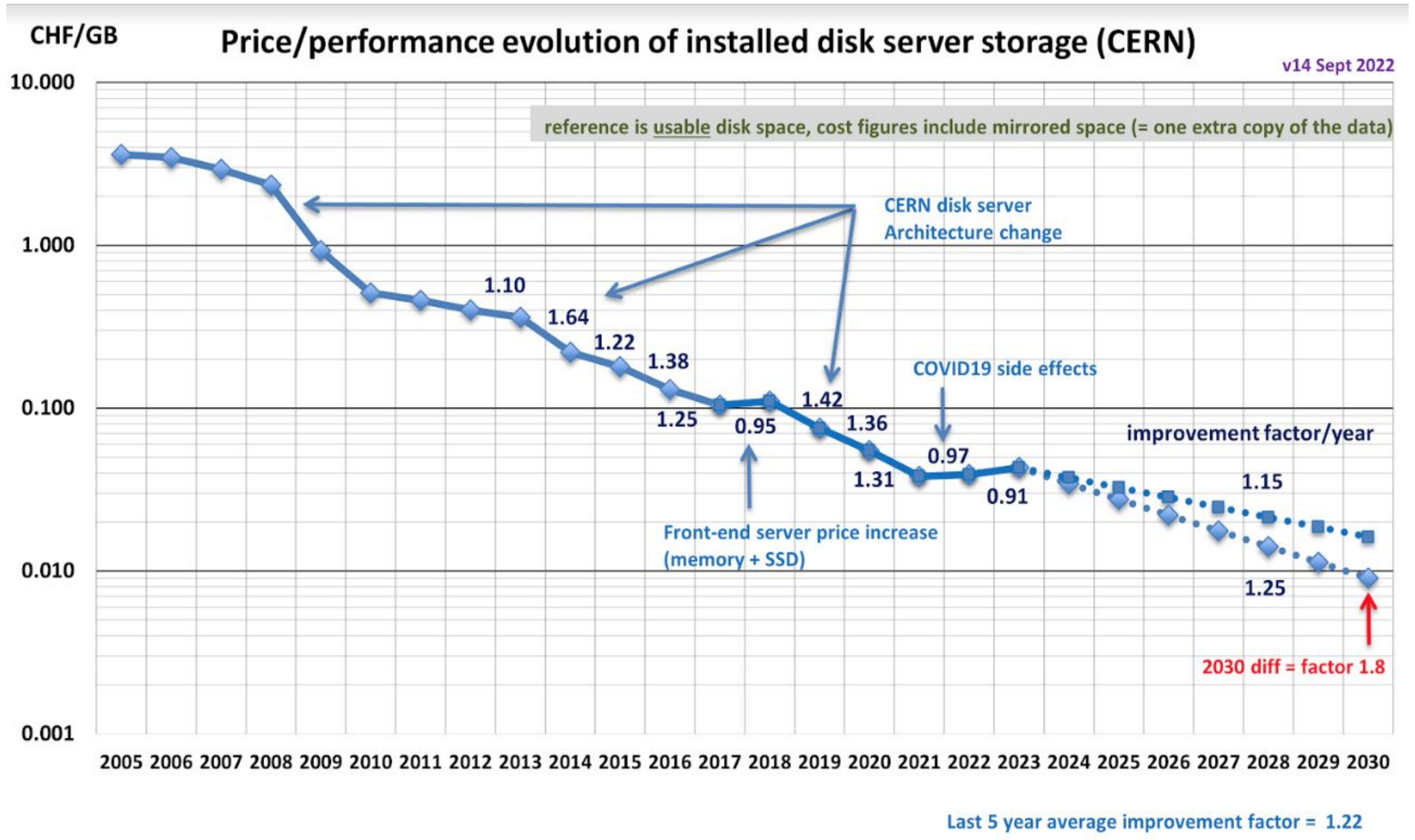
- It has to be technically viable (use technologies which will be available at the proper moment)
- It has to be financially viable (today, it is mostly translated with “cannot exceed current yearly budget from FAs” – a better translation is “asking for a budget in excess of today’s needs VERY STRONG motivations”)
- It needs to be operationally viable, with the manpower we think we can dedicate to it; where does the effort come from (research (grad students and postdocs) or professionals)?
- It has to allow for the core physics program of CMS, and possibly for “less core but interesting” programs
- (it has to match with trends and directions in the relevant venues, for example not be orthogonal to existing national / global roadmaps for scientific computing)

The current computing model

- Tiered computing center structure:
  - Worldwide LHC Computing Grid (WLCG)
  - Tier-0 + Tier-1s: CPU+Disk+Tape
  - Tier-2s: CPU+Disk
  - Pledges augmented by opportunistic (HPC, cloud, …)
- CPU: x86_64 (latest developments: PowerPC CPUs and NVidia GPUs in HLT)
- Disk: Spinning
- Network capacity is infinite (== the cost of data movement is not modelled)
- Central Operations
  - Central data processing and MC production
  - Central data placement on disk for processing input and analysis, on tape for long term storage (RAW stored on tape, 2 copies at Tier-0 and one Tier-1
  - No central ML training workflows
- Analysis
  - Grid jobs to access data
  - Produce user defined formats (NTuples) and use slimming/skimming (in some cases use directly nanos)
  - End-user analysis on interactive machines

Market trends (famous set of plot by B.Panzer on CERN procurement)



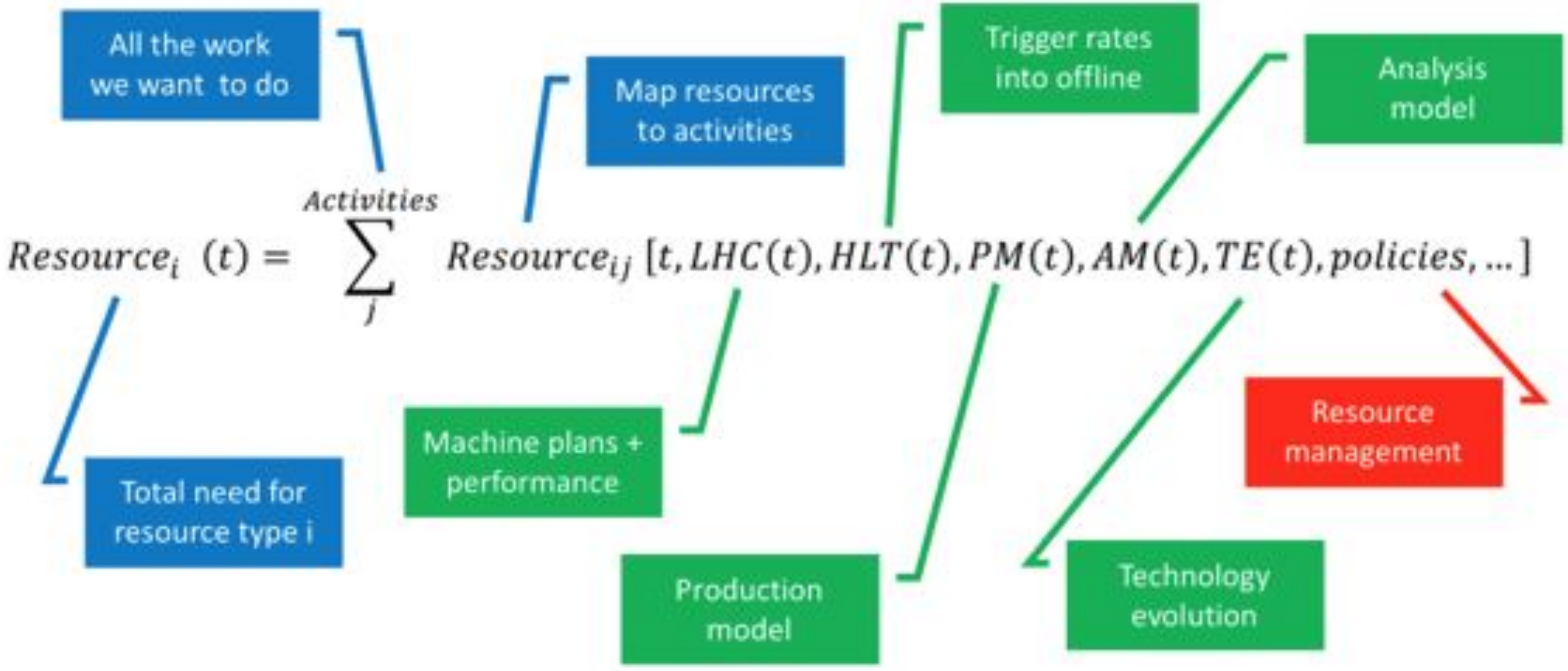


- What these plots hide is the "large increase" in 2021/2 due to the pandemic. Mostly solved by delaying procurement. For example, for CPUs
  - 2020: 3.6 CHF/HS06
  - 2021: no tender
  - 2022 (Q1 survey): 6.55 CHF/HS06
  - 2022 (actual tender): 3.6 CHF/HS06

General message from CERN: pandemic was a hiccup and not a long term change ➜ it "just" introduced a 1-2 years delay in price performance

DAQ TDR was expecting 1 CHF/HS06 in 2027-8, now it is 2029 – and HL-LHC schedule is now delayed

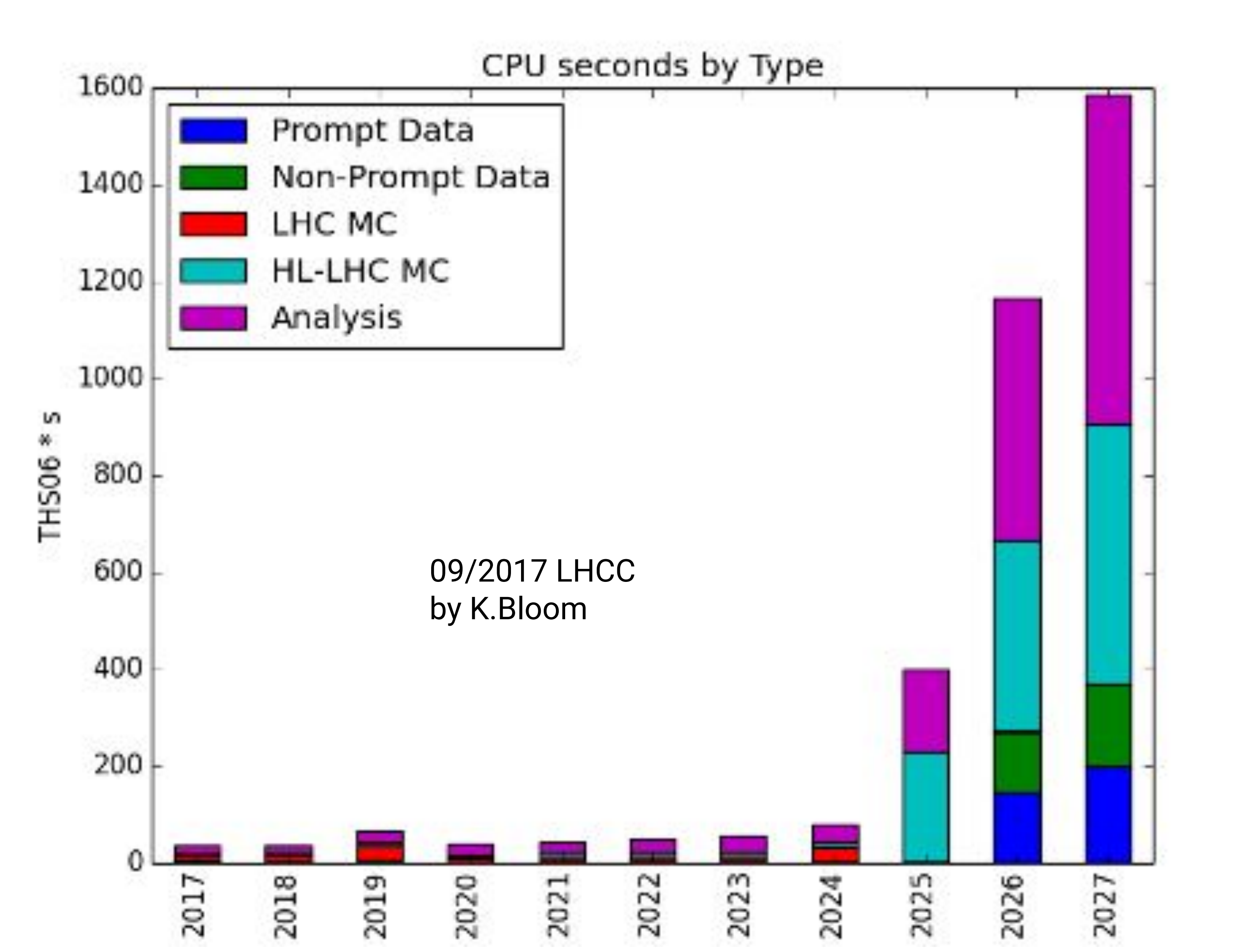
The model

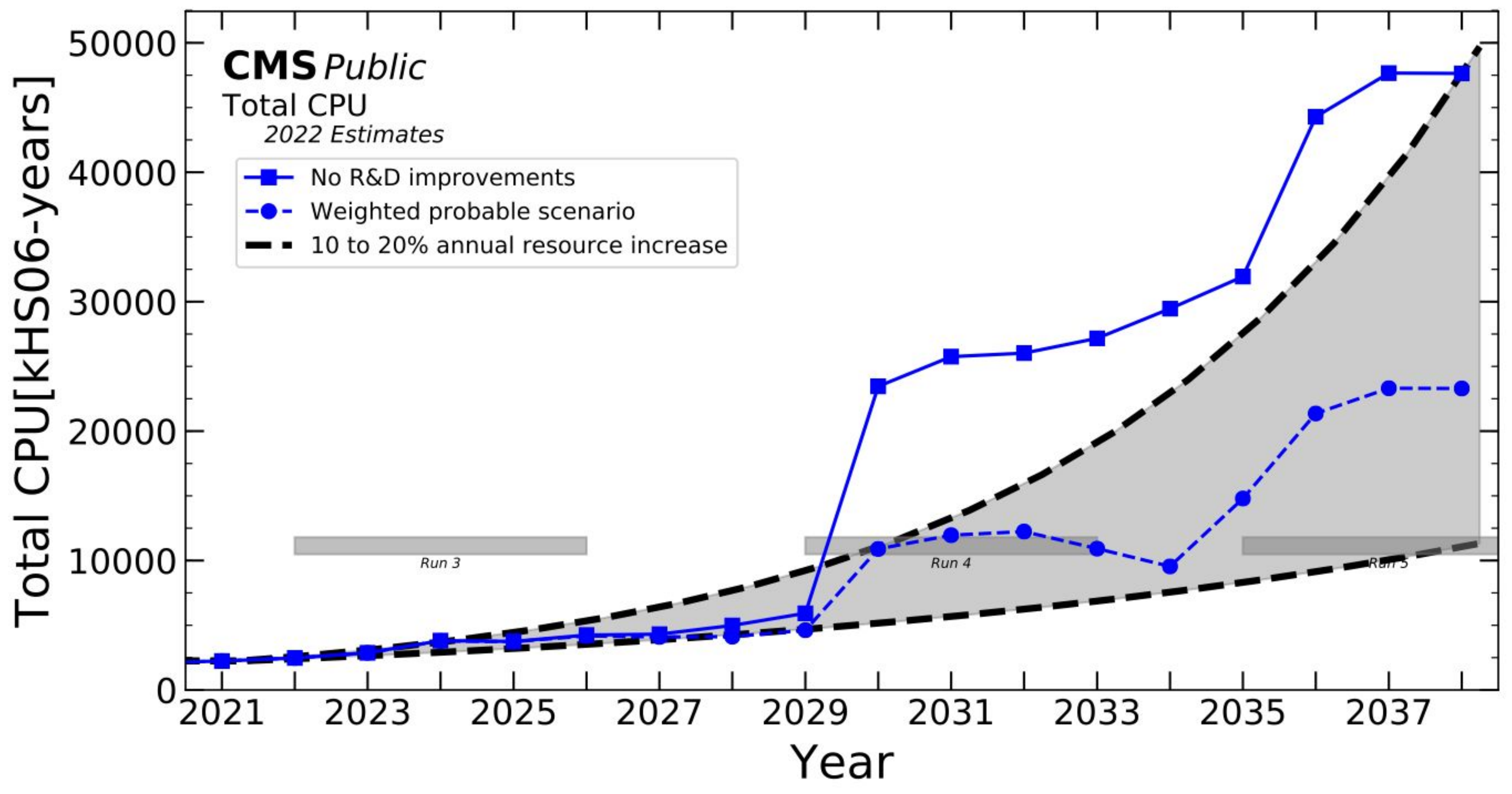


- CMS is building a model of needed computing resources with many inputs
  - Externally set (LHC # of live seconds, PU, Energy, …)
  - Set by our detectors (RAW data sizes, trigger capabilities, …)
  - (Physics Driven) CMS decisions (trigger rates, reprocessing steps, # of MC events, data tiers, parking / not parking, copies of distributed data …)
  - Externally set - again (money to execute all of this)

How to reduce costs (if not viable)?

- Initially (circa 2018) the predicted costs of HL_LHC computing were O(10x) larger than the “allowed flat budget”
- Handles to reduce costs:
  - Reduce needs
    - (fewer replicas, fewer reprocessing, less MC, smaller data formats, …, ML, …)
  - Use better $/performance solutions
    - (columnar analysis, GPUs, FPGA, …)
  - … wait more time, hardware gets cheaper ….
- To cut the long story short …
  - Last model iteration (2022) is compatible with Flat Budget via a “realistic” R&D scenario

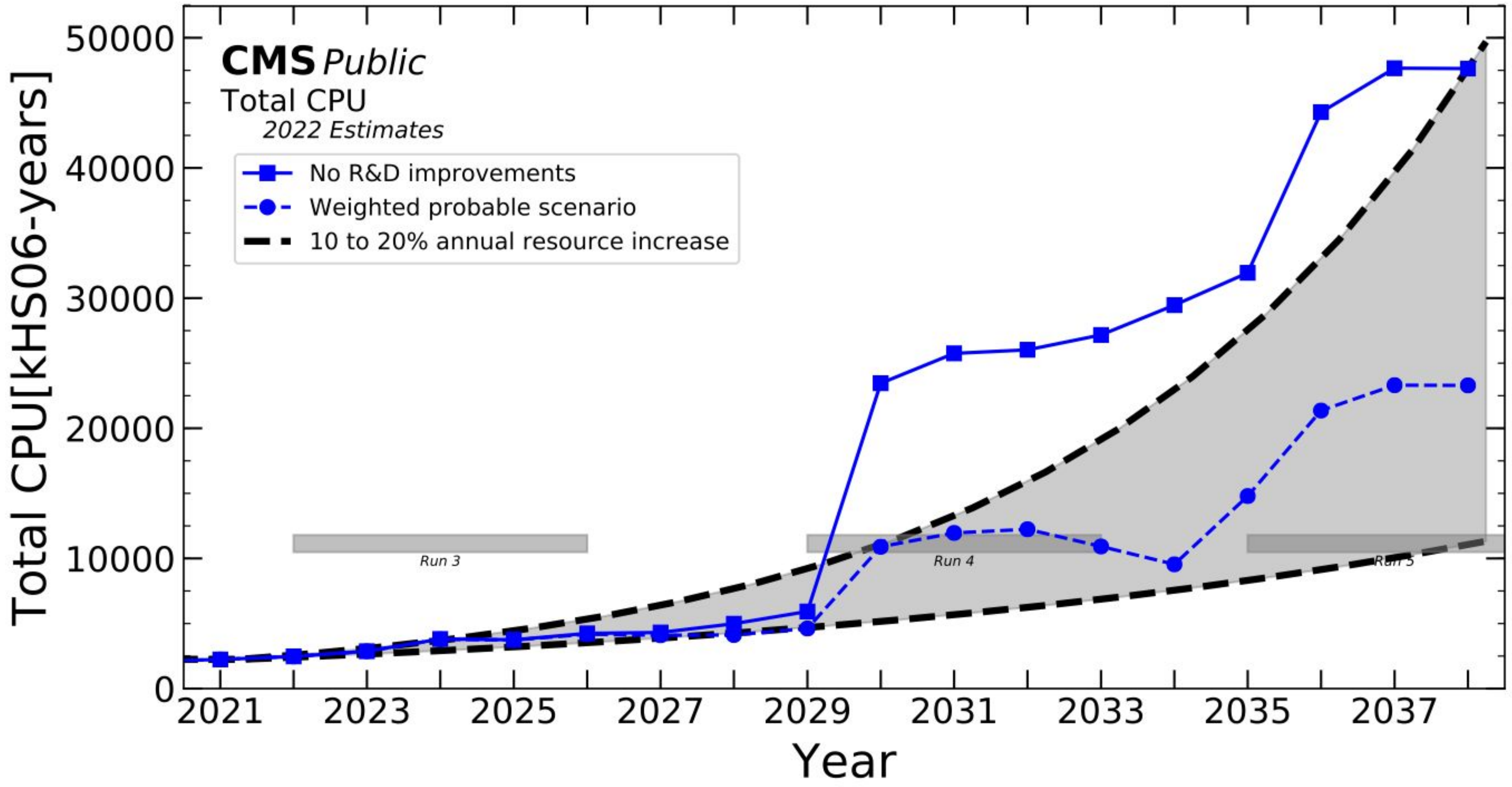


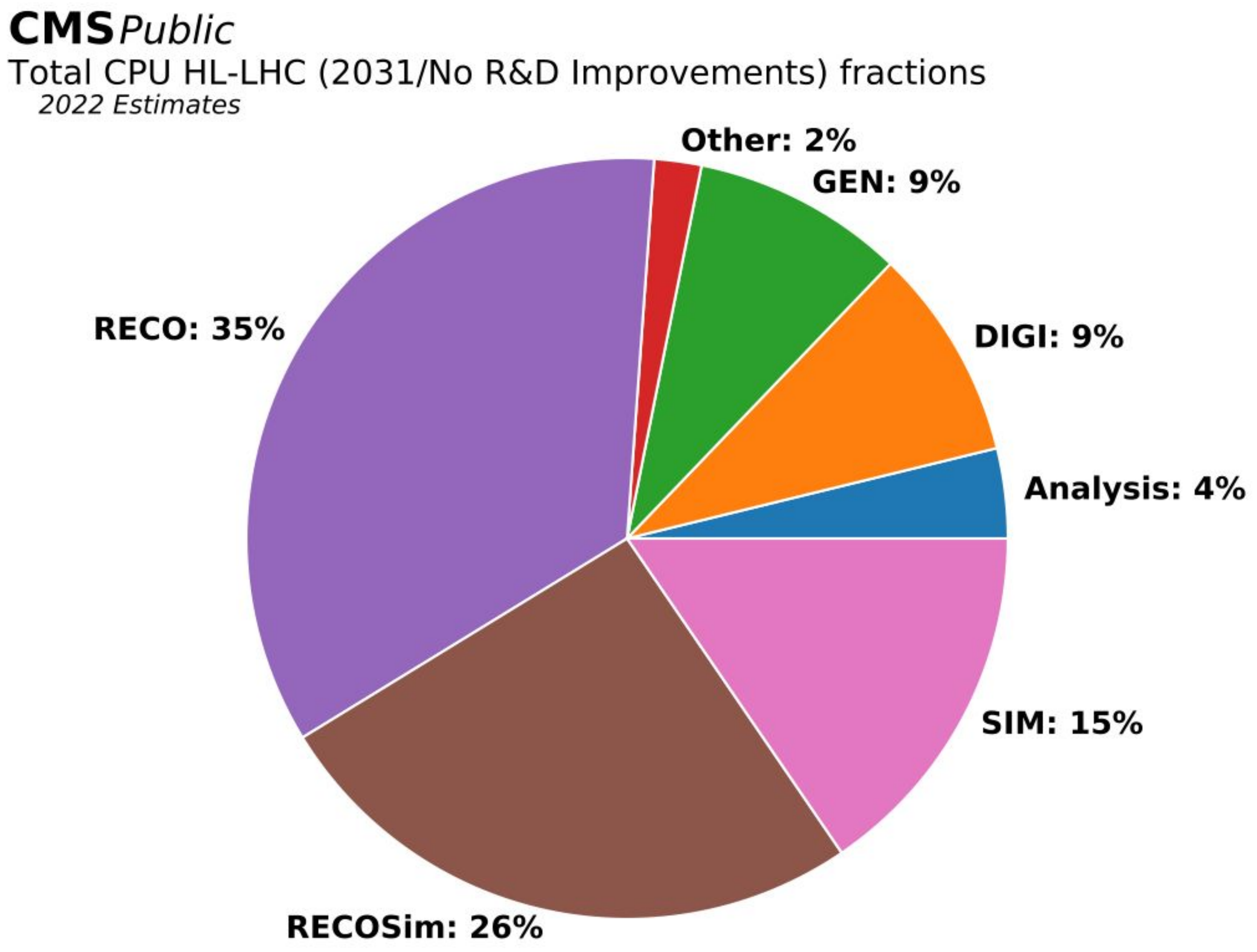


Current assumptions in HL-LHC Computing Model

- No parking/scouting
  - Current HLT rates match DAQ TDR: 5kHz in Run 4, 7.5kHz in Run 5
- The Run 2/3 practices for data processing continue
  - Prompt reco that keeps up with data taking, end-of-year rerecos, startup Monte Carlo, large Monte Carlo productions corresponding to the rerecos, miniAOD productions, and nanoAOD productions
    - Each year end-of-year rereco, a complete rereco of Run-4 in LS4
    - MC: 1-2 small productions (~4B events) and one large production (for end-of-year rereco pass) per year
- Heterogeneous compute (GPUs, FPGAs, ..) not explicitly modeled
  - For now, these resources enter as a cost reduction per unit compute
- Each data tier has a model of #replicas on disk vs time
  - NanoAOD(Sim) is small enough that we can afford many replicas
  - "Legacy" (eg, last good) versions kept on disk. Older versions are quickly migrated off of disk
    - AOD(Sim): 0.5 disk copies when produced, reduced further by 50% each year
    - MiniAOD(Sim): 2 disk copies when produces, stays at 2 as long as it is "legacy", reduced after being replaced
  - Small data tiers have more replicas than larger ones
- Tape:
  - All raw and all data from legacy data tiers kept indefinitely
  - Other data cleaned from tape after time for migration

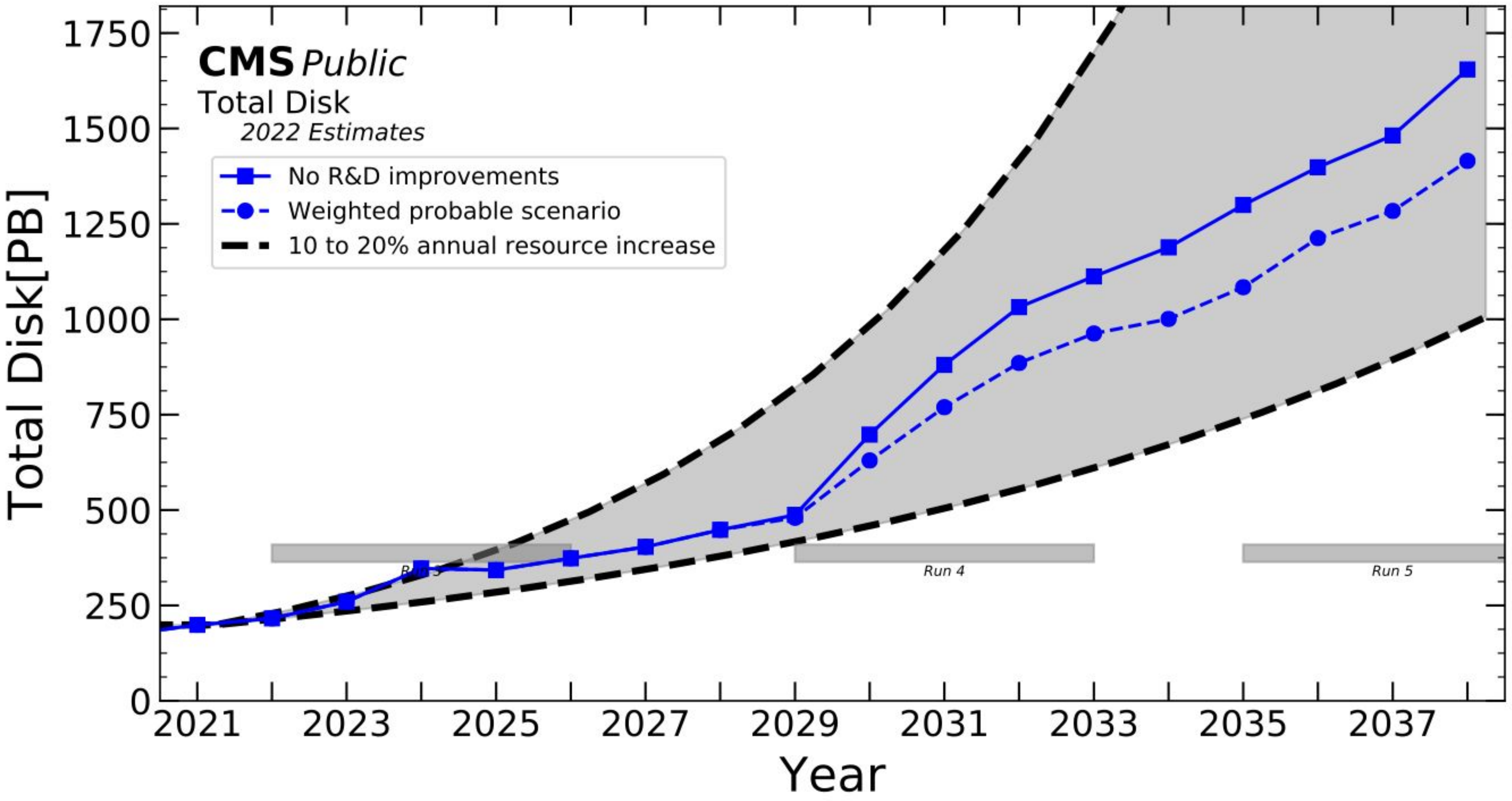
CPU

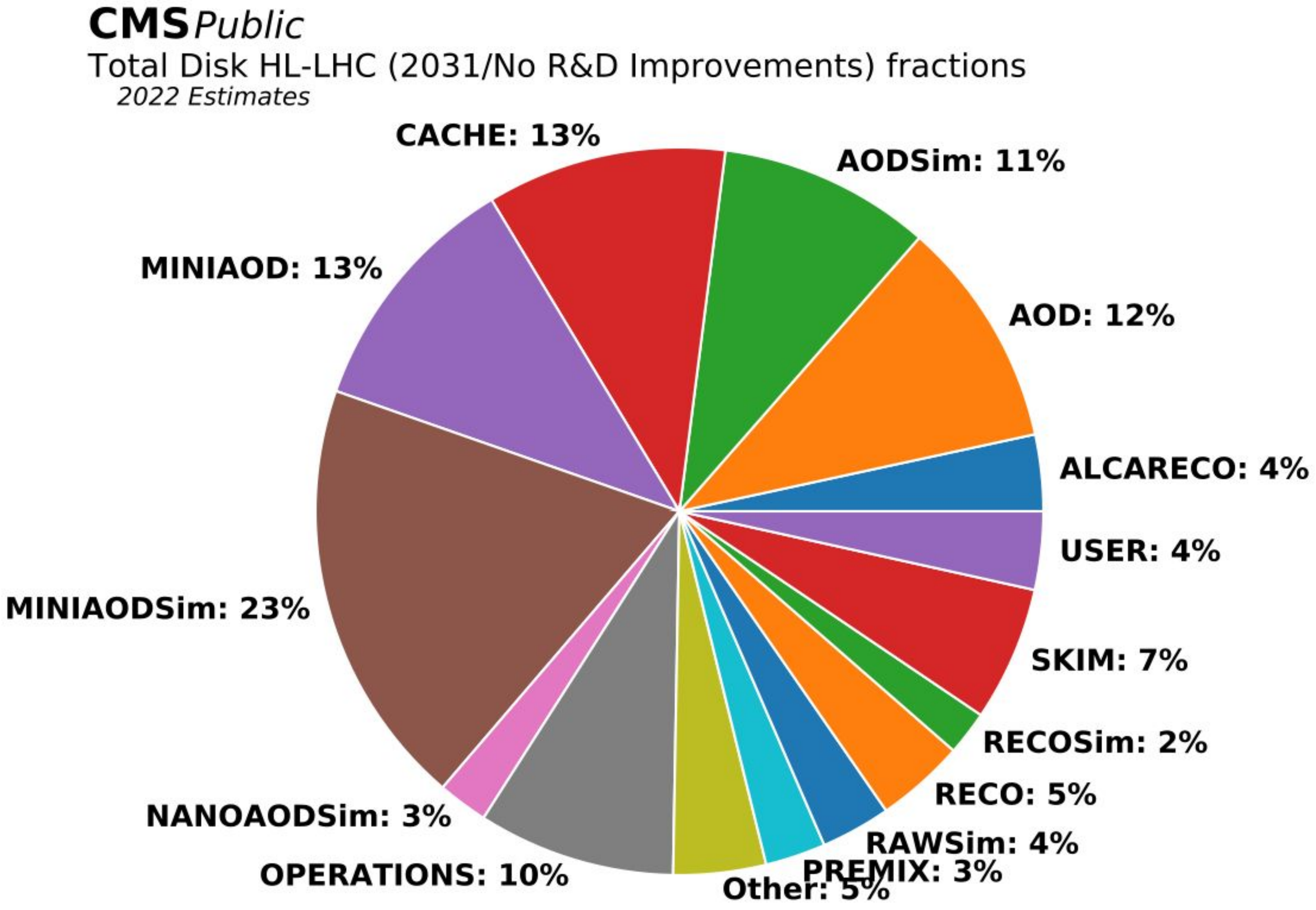




https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSOfflineComputingResults

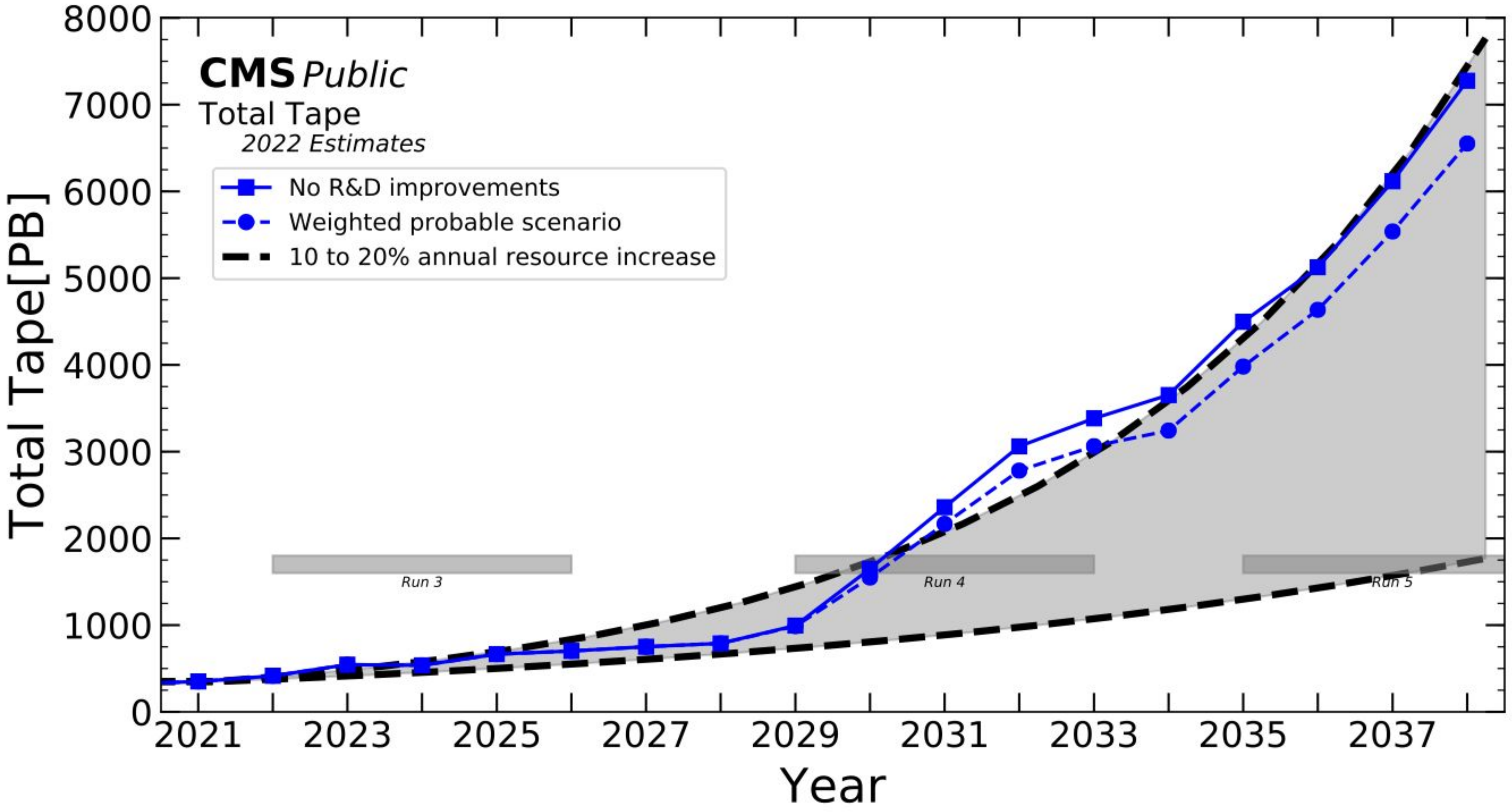
Disk

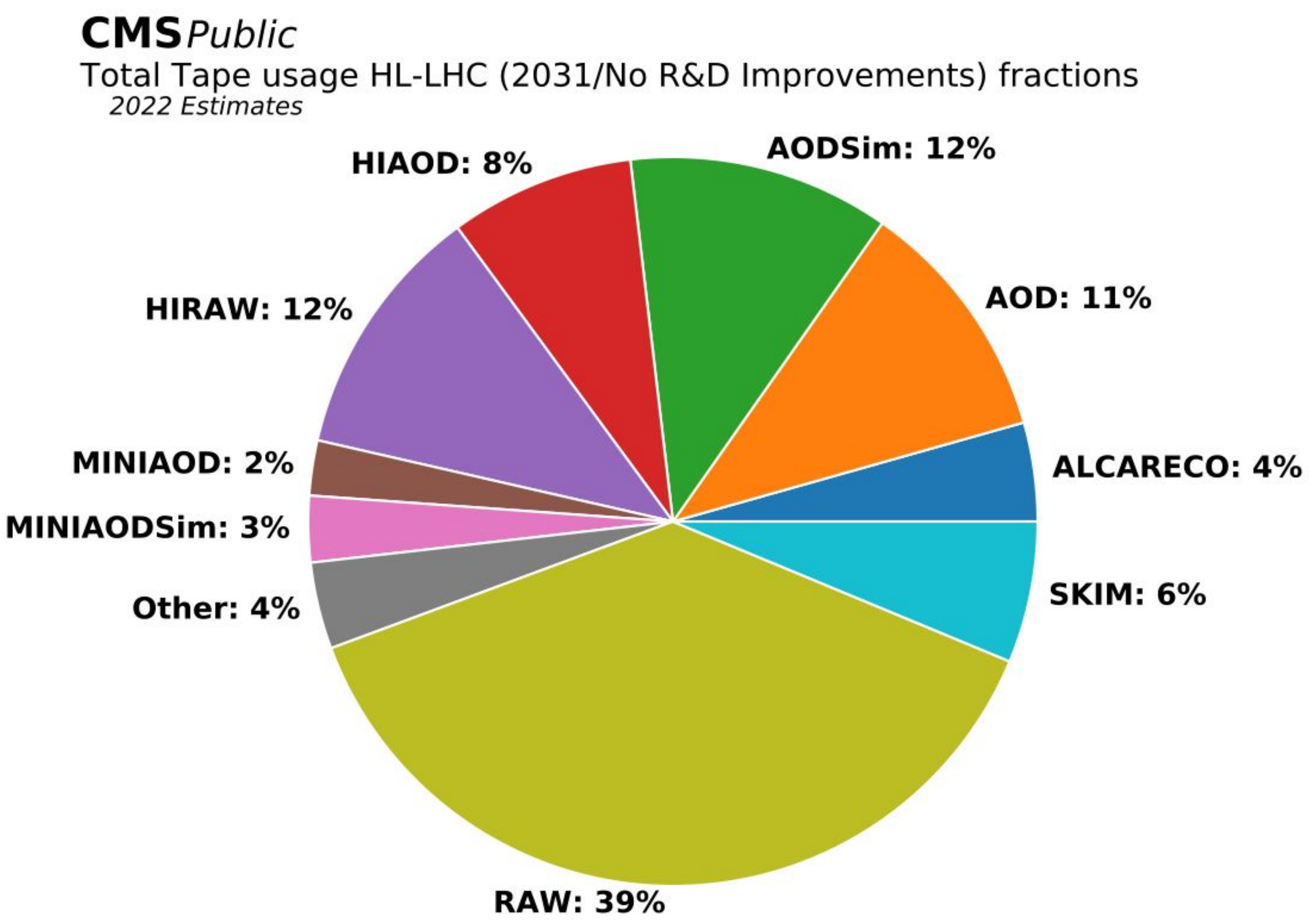




https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSOfflineComputingResults

Tape





https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSOfflineComputingResults

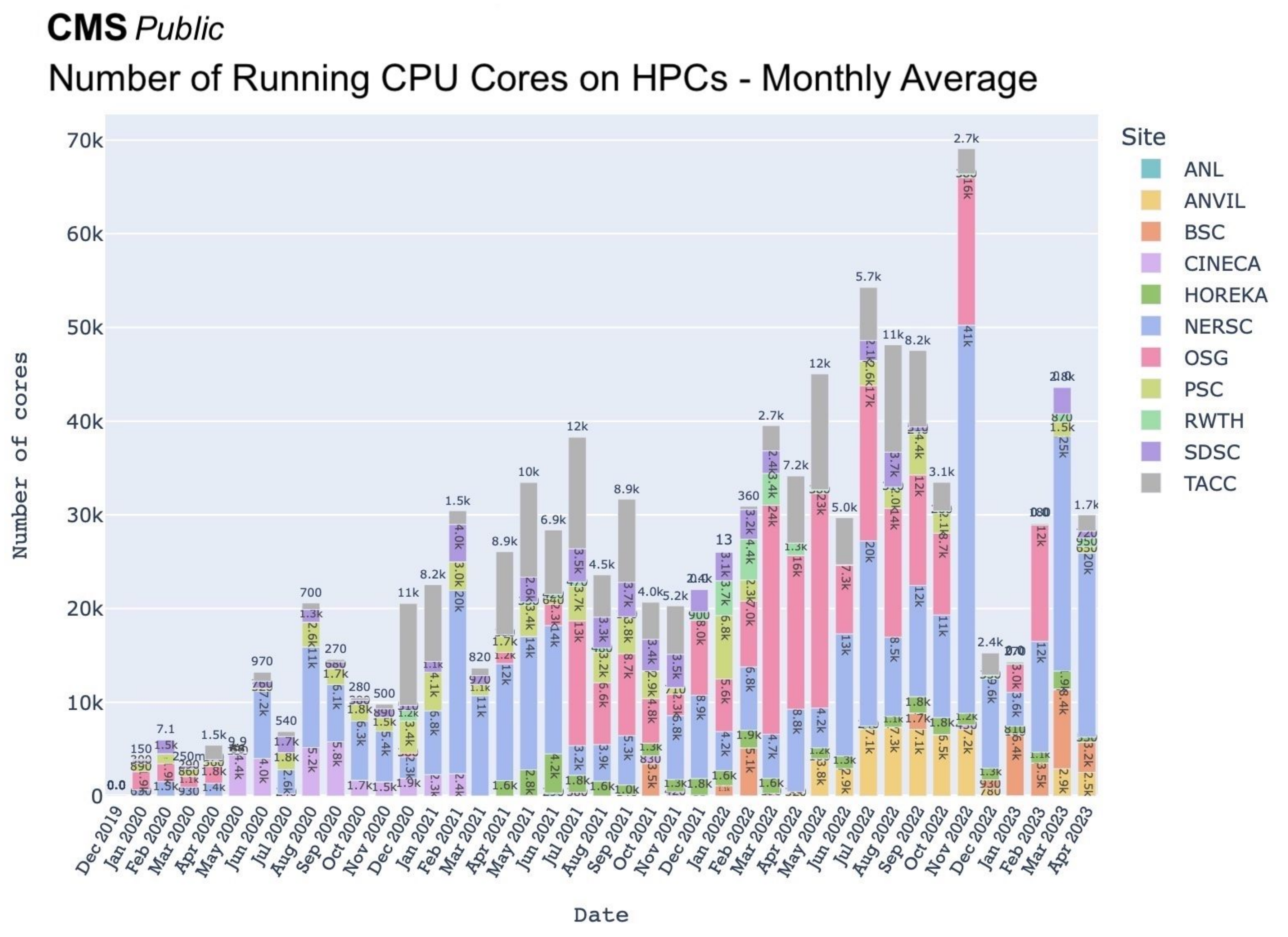
What keeps me up at night?

- Run-3 data taking: ~5 kHz data taking rate (prompt+parked) already at HL-LHC planning level
- More realistic planning scenario for HL-LHC
  - A total trigger rate of 20 kHz;
  - A prompt fraction of 20\%, processed after 48 hours;
  - The MC scaling factor with luminosity increased to 0.4;
  - 1.5 average copies of RAW on tape.
  - A 200 kHz HLT scouting rate (event size at 4 kB/event), with no need for an additional MC production;
  - An end of year processing of the full 20 kHz rate.
- And variations of that:
  - The total rate and prompt fraction are varied;
  - The MC scaling factor with luminosity is varied;
  - The average number of RAW tape copies is varied;
  - Different scouting settings (rates and event sizes);
  - The length of the prompt processing delay is changed and we change the end-of-year processing strategies (longer prompt processing delay and no end-of-year processing)

Some topics to consider for R&D

Heterogeneous Hardware

- x86-only hardware era comes to an end
- Heterogeneous hardware has to be used
  - GPUs, FPGAs, ARM, RISC-V, …
  - Industry is leading the way with demanding more capable AI hardware at less and less power consumption
  - We will need to follow industry trends
- SuperComputers will be part of the resource mix
  - Grid will not be able to provide for all of our needs
  - SuperComputers (High Performance Computing (HPC)) optimized for large applications spanning multiple nodes/cpus and utilize GPUs extensively to minimize power consumption
    - Our application is HTC (High Throughput Computing) where applications easily fit on a fraction of a CPU
- We need to write software for a heterogeneous hardware architecture world!

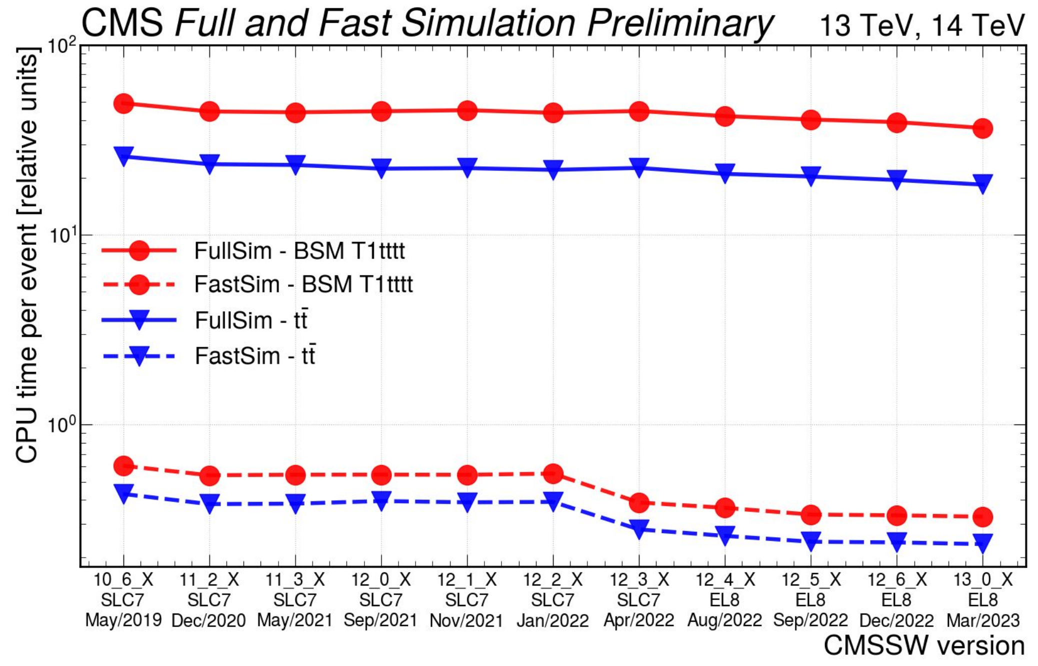


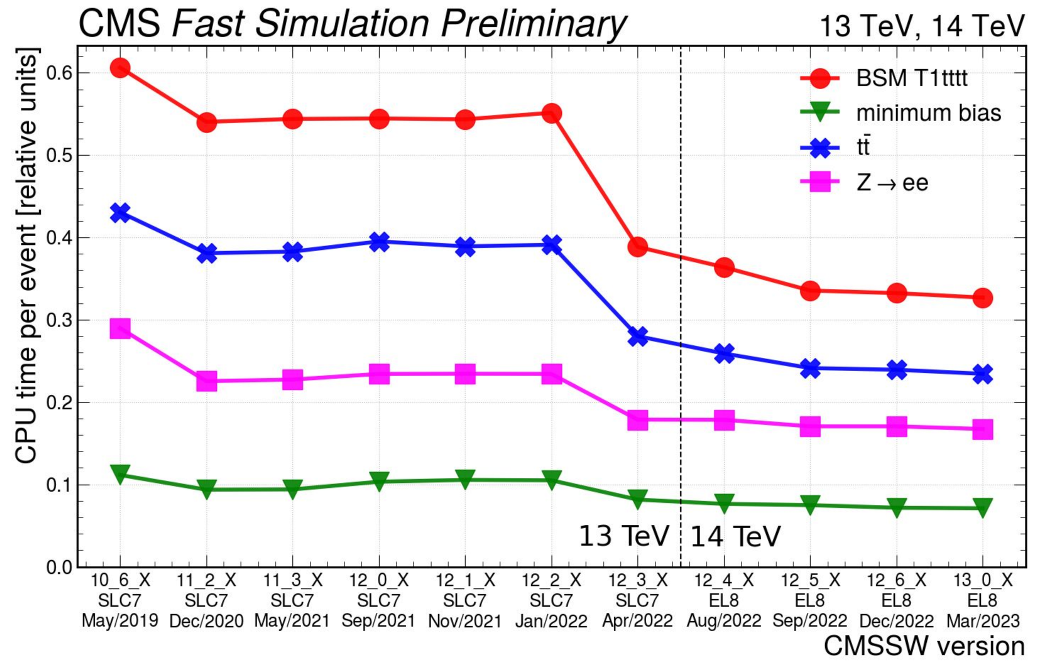
Software for heterogeneous architectures

- CMS reconstruction: Over 4 Million lines of C++ code
  - Optimized for x86 CPU architecture
- Algorithms need to be re-architected and ported so that they can run on GPUs and other massively vectorized hardware architectures
- Currently, every vendor of GPUs (NVidia, AMD, Intel) has their own programming interfaces ➜ write the same algorithm several times
  - Portability solutions will help writing an algorithm once and re-compiling it for different GPUs
  - Not standardized yet unfortunately, CMS decided to go with Alpaka
- The switch from CPU to GPU is as big as the switch from Fortran to C++, if not bigger
  - Sociological problem: domain experts (physicists building detectors) will not be able anymore to develop efficient reconstruction algorithms for their detector components
  - Need GPU programming experts to support them

Evolution of Offline Software

- Core software "architecture"
  - Address the high number of functionalities/architectures/microarchitectures/ML engines supported.
  - Better usage of HPC is an open point. How to improve here?
  - All this flexibility may bring some costs: "backward" compatibility (especially for MC), validation, reproducibility.
- Generation
  - Negative weights have been the main concern. This may need an effort beyond software improvements and that touches analyses strategies and demands.
  - GPU offload is touching also generators and it's involving both CMS and cross-experiment efforts.
- Simulation
  - Full Sim on GPU will be a challenge. How to address it?
  - FastSim and FullSim are complementary but for HL-LHC. Do we need a paradigm shift (only "sociological")?
  - ML methods are showing promising results. How do we see them in the picture?
- Reconstruction
  - A steady CPU usage reduction. We are now in the era of fine tuning but still every year delivering ~8-10% improvement
  - GPU readiness is moving from online reco to offline reco (with portability and validation).
  - ML methods are showing promising results. How do we see them in the picture?

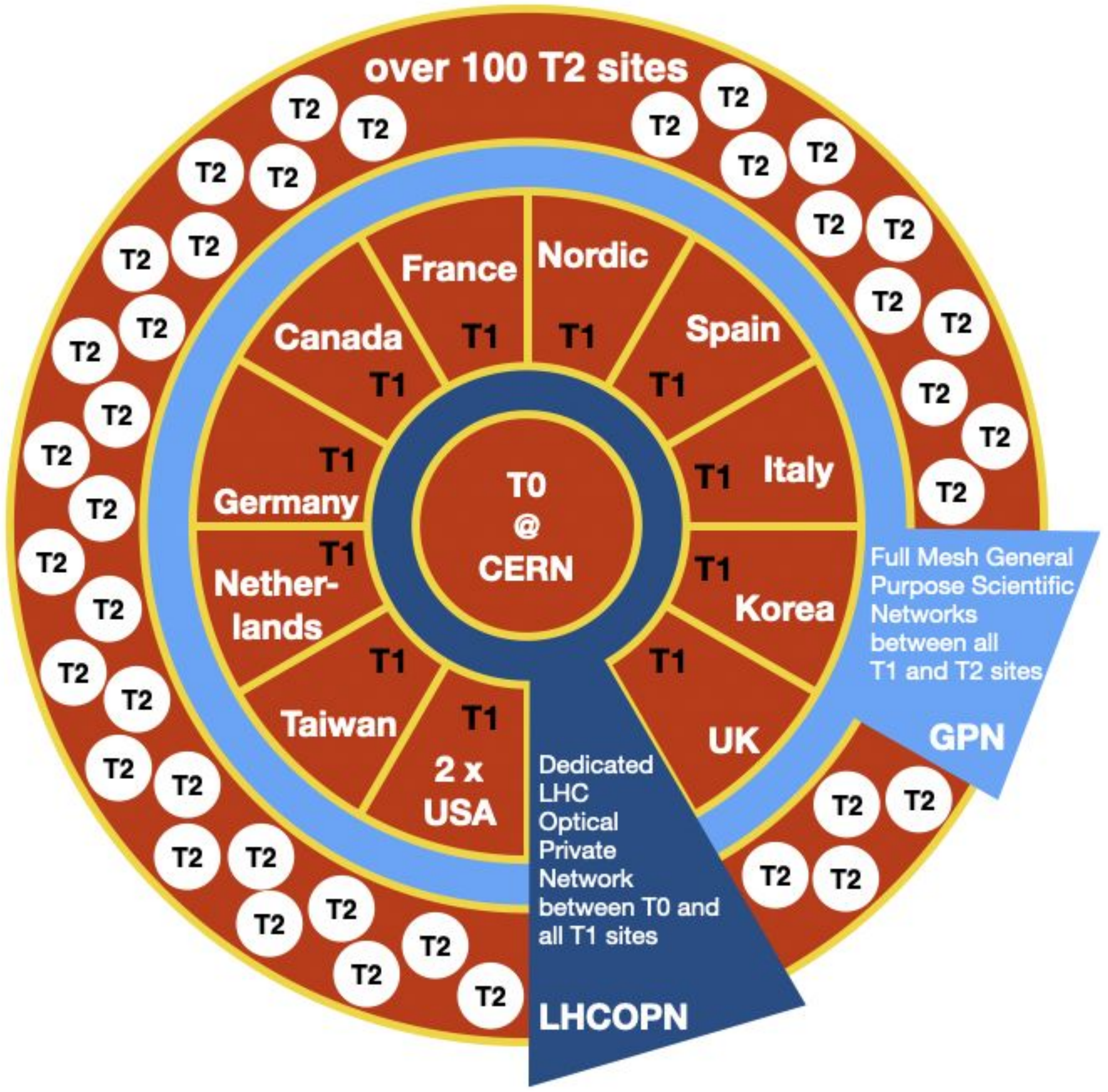


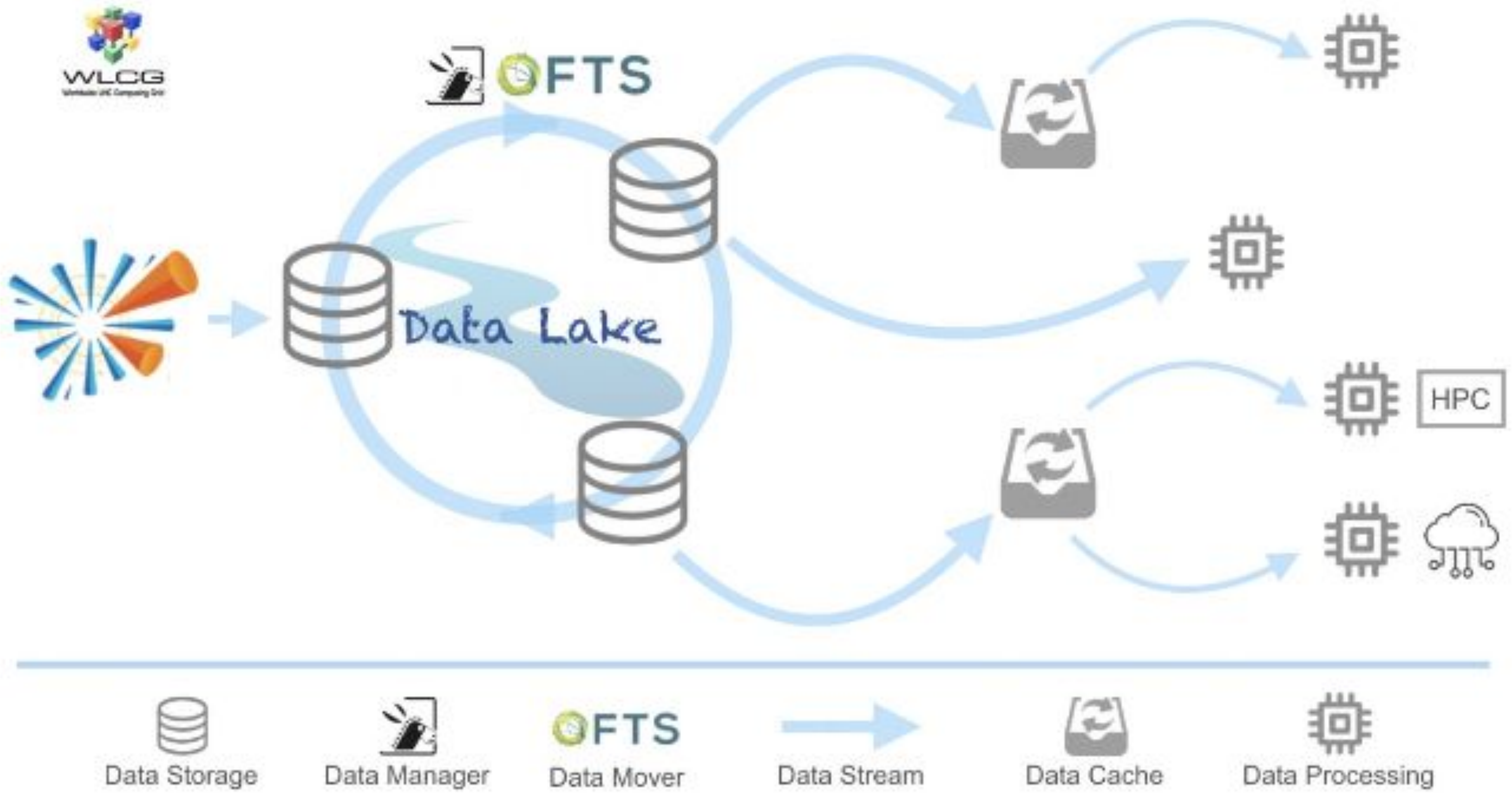


Analysis Facilities

- New concepts are needed to analyze even more events in a reasonable time
  - LHC will (mostly) transition from discovery machine to high statistics analysis machine in HL-LHC
  - Industry toolkits are well advanced (different than at the beginning of LHC where ROOT was all there is)
- Analysis Facility concepts
  - Try to bridge HEP specifics to the industry toolkit
  - Try to provide integrated solution with data handling and optimized processing
  - Try to be columnar and declarative to be able to optimize the backend independently
- What we still need to solve
  - Running on one analysis facility with hundreds of users
  - Providing access to lower level event information (MINIAOD, AOD) without having people recreate the analysis format (NANOAOD) ➜ columnar service and object stores (CEPH)

Infrastructure: Grid vs. Lake?





- The Grid has served us well
  - It was developed when nobody else was doing scientific computing at scale
  - These days, many more science disciplines are having massive computing requirements
- National infrastructures are developing to support all sciences simultaneously
  - U.S.: Openscience Grid (OSG), National Research Platform (NSF), Integrated Research Initiative (DOE-ASCR)
- Do we need to rethink how we provision resources?

- consolidate to less regional/national entry points,
- Often referred to as "Data Lakes"
- Data/workflow management by experiment becomes
- more high level
- More fine-grained data and job flux within a data lake
- Required major changes/enhancements of middleware
- Different operations model

Storage

- Users' Perspective
  - Unlimited storage quota
  - No I/O limitations - fast access
  - Endless network capacity
  - Find data when needed at wanted location
- Computing Model Perspective
  - Resources are limited
    - Trade access speed (IO ops) vs. capacity
  - Trade number of replicas vs. network capacity
  - Optimizing the parameters is a complex process
- Current state of affairs
  - Site storage organized around two quality of service (QoS) tiers
    - Tape: primary archival storage
      - Inexpensive (per byte) and durable o High latency (requires staging)
    - Disk: primary access medium
      - More expensive but lower latency

Storage: Needed R&D

- Archival storage
  - Modeling expected tape recall rates
    - Even with an incremental chain model, bandwidth needs are large
    - Tape capacities increasing much faster than drive rates
  - Can tape software in use today cope with these demands?
    - CERN Tape Archive (CTA) is designed for Run 4 requirements
- Caches
  - Establishing common Quality of Service (QoS) tiers
  - Developing/deploying caching infrastructure
  - Role of high-speed (IOPS) storage in
    - Analysis facilities
    - HPC computing
  - Role of object stores and other non-block storage
    - Analysis-specific storage
- Networks: Software Defined Networks (SDNs)
  - Networks expected to become increasingly dynamic
  - Bandwidth as a scheduled resource: Technology: Software Defined Networks (SDNs)
  - Middleware needs to take advantage of SDNs
    - Packet marking and flow labeling for monitoring
    - Network orchestration via projects such as SENSE and NOTED

U.S. CMS HL-LHC R&D Strategic Plan

- To do physics in HL-LHC, we need to
  - Archive multiple-hundreds of PB of RAW data on tape
  - Process all the events and produce even more simulations
  - Utilize accelerators and advanced computing architectures efficiently
  - Integrate AI/ML on unprecedented scale
  - Provide access for analysis: more events analyzed in less time ➜ high statistics analyses

- The U.S. CMS S&C Operations Program defined 4 “Grand Challenges” (GC) that are tackling high priority areas and are embedded in the overall CMS effort:

  (1) Modernize physics software and improve algorithms
  (2) Build infrastructure for exabyte-scale datasets
  (3) Transform the scientific data analysis process
  (4) Transition from R&D to operations

U.S. and International Partners

- U.S. CMS is part of the community's ecosystem for computing and software related research and developments.
  - Research partnerships
    - Host National Lab: Fermilab
    - 7 U.S. Tier-2 institutes and additional U.S. institutes
    - Other National Labs
    - CERN
  - National and international consortia
    - Open Science Grid (OSG)
    - HEP Software Foundation (HSF)
  - Joint and collaborative projects
    - IRIS-HEP
    - HEP-CCE
  - Community efforts
    - Joint Blueprint activities with U.S. ATLAS, OSG, ESnet, and IRIS-HEP Snowmass Computational Frontier

Summary & Conclusions

- HL-LHC is an unprecedented challenge for Software & Computing (and many other parts!)
  - We heard that before: before the LHC start, Software & Computing was an unprecedented challenge
- I am confident that we're going to make it (somehow).
  - But we need you to contribute and think about solutions for these hard problems
- We are not alone
  - This is different than before the start of LHC
  - Big data sciences have emerged from Genomics to Astro Physics to Light Sources to Nuclear Physics to …
  - All will have to share computing infrastructures and will have to use common software solutions
- And don't forget, software & computing are good examples for transferable skills to industry