Computing

Oliver Gutsche

17th Hadron Collider Physics Summer School

19. August 2022

Who am I - Disclaimer

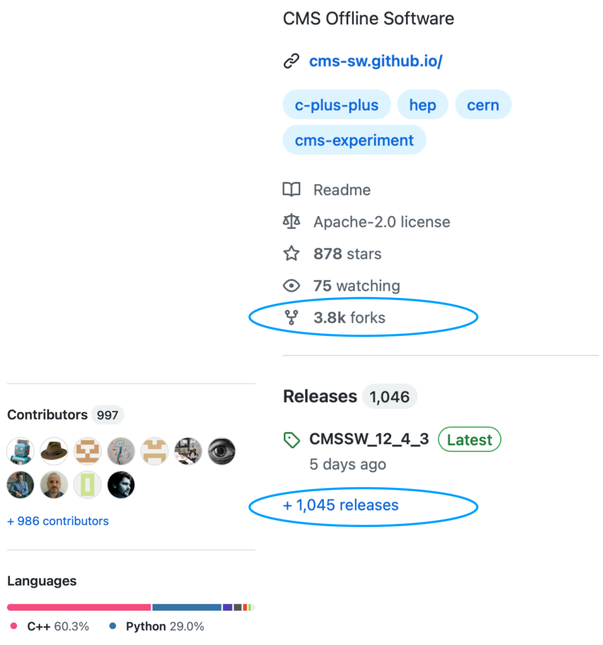
- Oliver “Oli” Gutsche
  - Senior Scientist at Fermilab
  - Particle Physics
  - CMS experiment - LHC
- Involved in HEP computing for a long time
  - Software Development
    - Event Display and Tracking Software
    - User Grid Analysis Tool
  - Operations
    - Operated analysis and Grid infrastructure
    - In charge for CMS Computing Operations leading up to the Higgs Discovery
  - Management (many …)
    - U.S. CMS Software and Computing Operations Program manager:
      - U.S. CMS contributions to resources and efforts for CMS (and beyond)

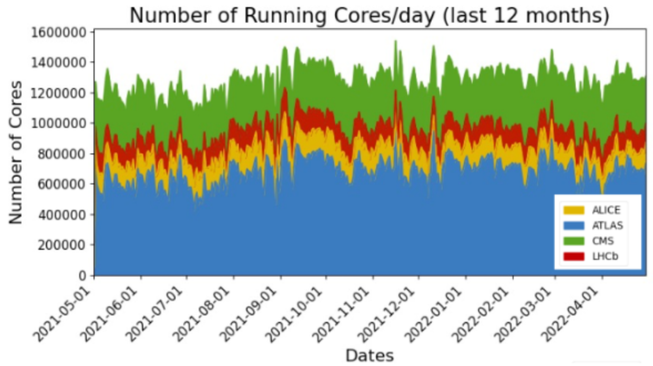
- Disclaimer
  - Biased towards LHC and CMS computing ➜ extrapolates to other LHC experiments and beyond
  - Not covering: Machine Learning & Quantum

Introduction

Introduction - Scale

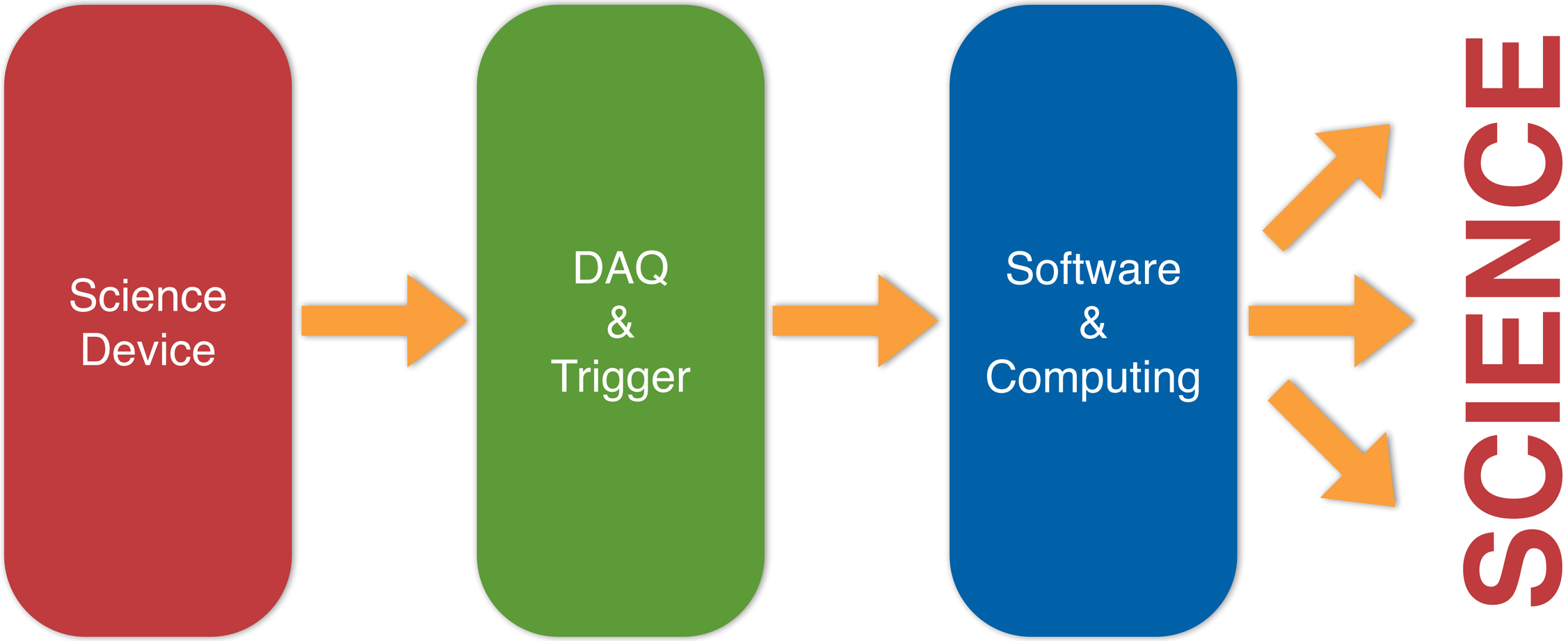
- High Energy Physics has a vast investment in software
  - Estimated to be around 50M lines of C++
  - Which would cost more than 500M$ to develop commercially
- S&C is a critical part of our physics production pipeline, from experimental design, to data collection (e.g. triggering), to data analysis/interpretation
- LHC experiments use about 1.4M CPU cores every hour of every day, we have around 1000PB of data with 1000PB of data transfers per year (10-100Gb links)
- This is a huge and ongoing cost in hardware and human effort and it should have broader impacts!





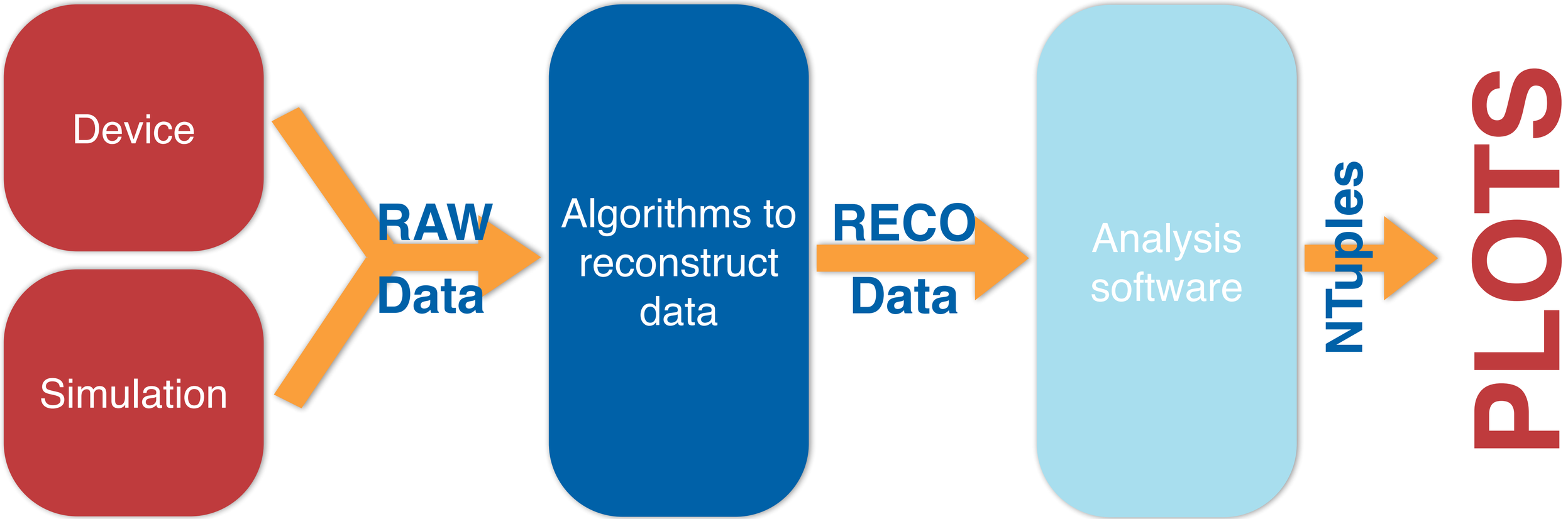
- From: Elizabeth Sexton-Kennedy (FNAL) @ Seattle Snowmass Summer Meeting 2022 - Collaborative initiatives across experiment boundaries

The Scientific Process



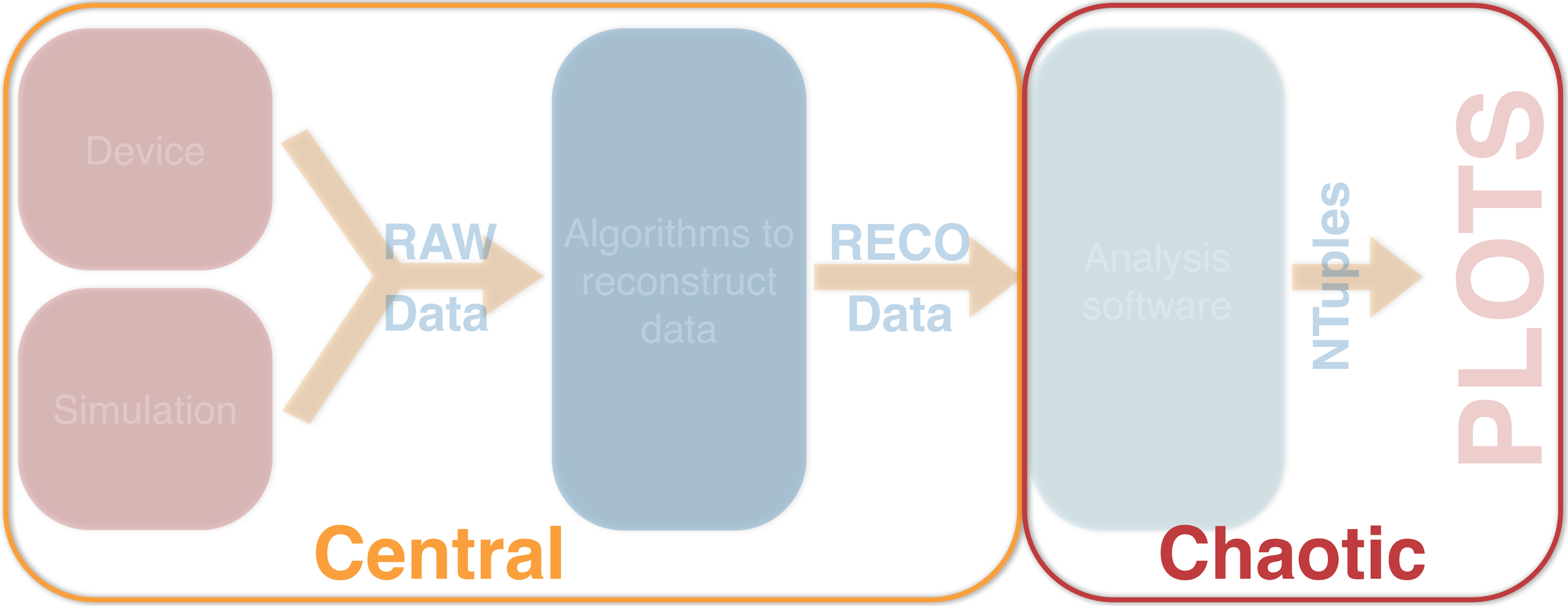
- Software & Computing is an integral part of the scientific process

Software & Computing



- **Software** is important for every step on the way to scientific results

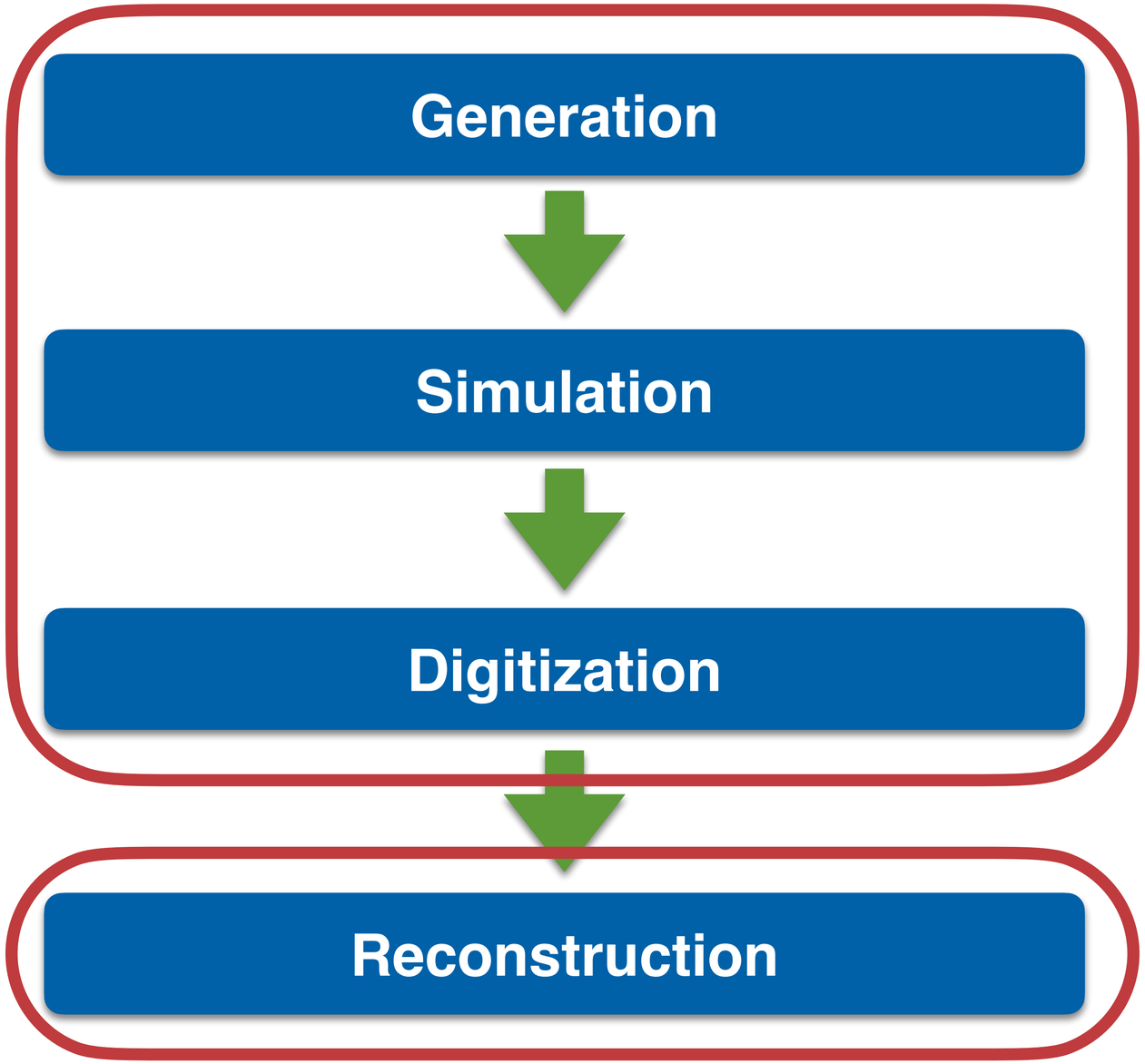
Software & Computing



- Computing resources (**Storage and Network**, **Processing**, …) are needed for all steps

Software

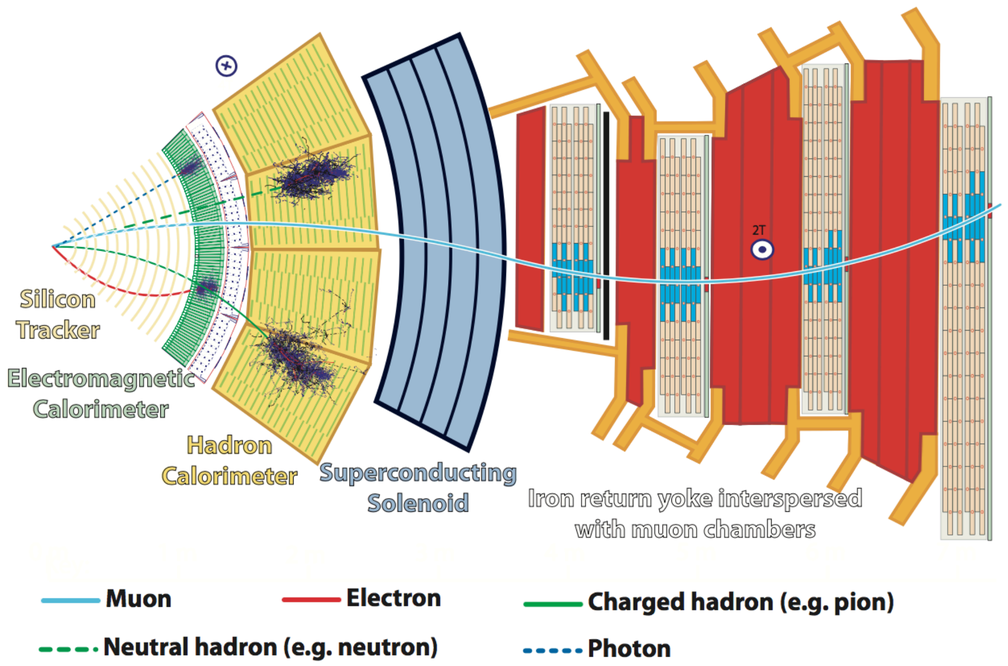
Processing steps



only for MC

both for data and MC

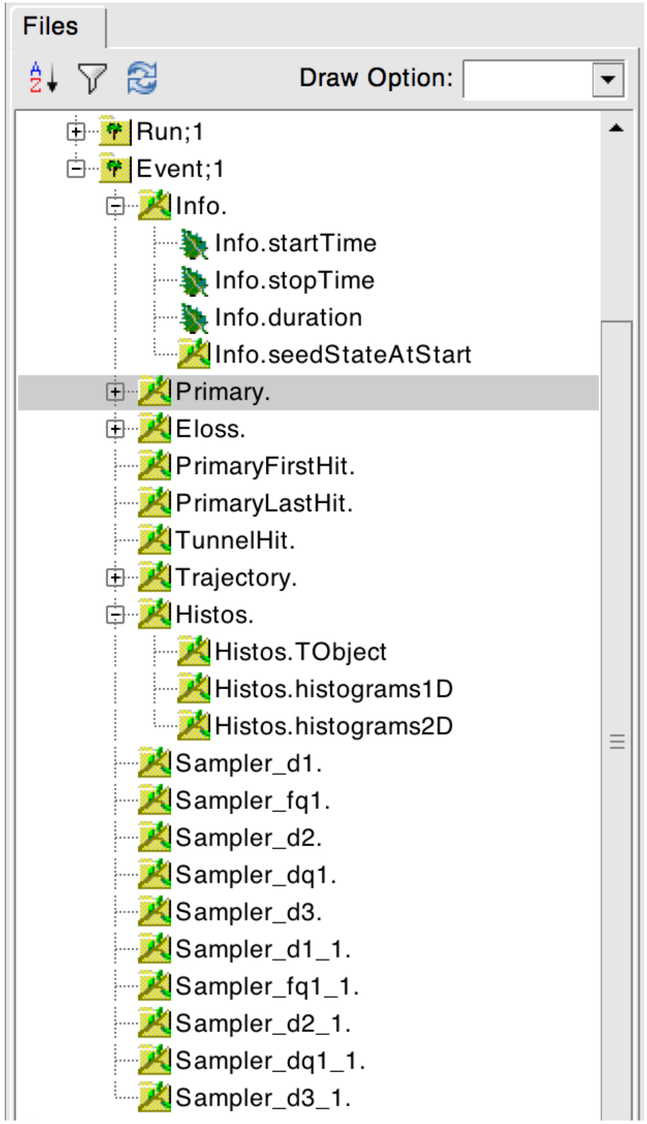
HEP event processing application



- HEP computing problem: HTC (High Throughput Computing)
  - Trivially parallelizable
- Individual applications access input data and produce output data
  - Principle: Input never changes
  - New information is generated and stored in output
  - Input information copied forward or discarded

Software Frameworks

- CMSSW in CMS
  - Object-Oriented software framework (other frameworks are available)
  - Data stored as C++ objects in memory
  - C++ objects read from files (PoolSource) and written to files (PoolOutputModule)
- Software is modular (EDProducer, EDAnalyzer, EDFilter)
  - Configuration defines schedule and dependencies of modules (“run hit reconstruction before track reconstruction”)
  - CMSSW schedules modules
- Principle: Modules don't manipulate objects
  - Modules read in data (consume) from the event in memory
  - EDProducers write new objects into the event (in memory, and later to a new file)
- Why: clean approach, traceability



Hardware Trends ➜ Software Necessities

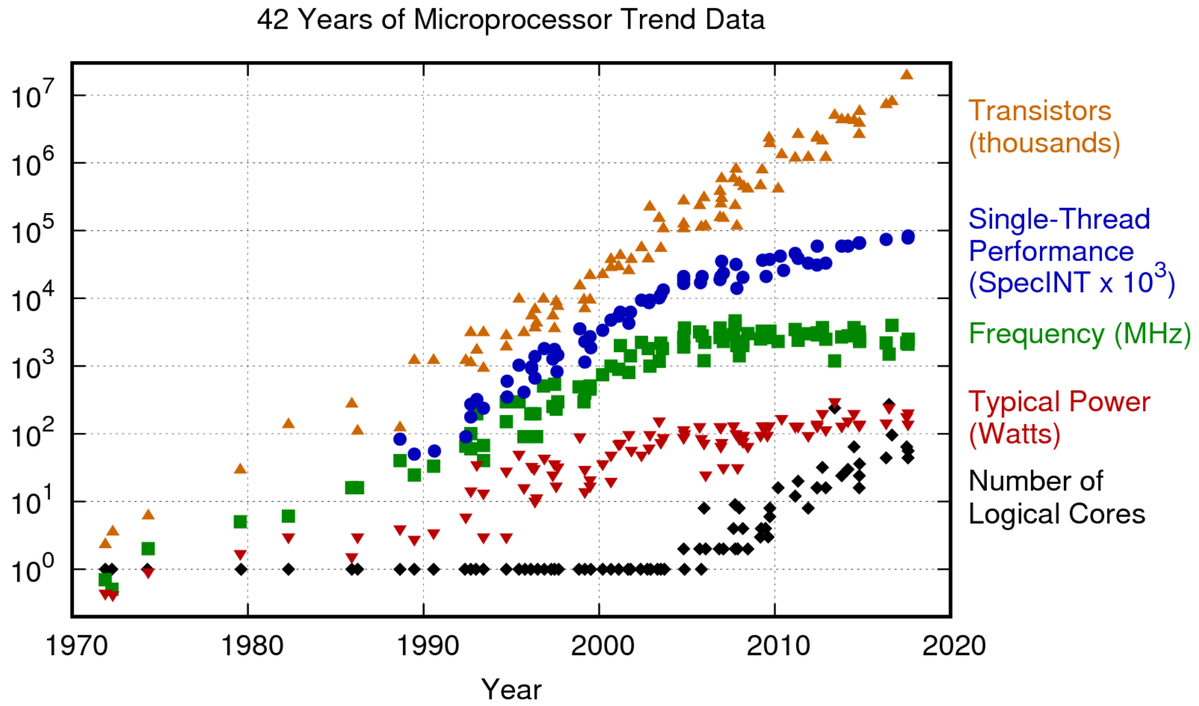
- Dennard Scaling: “as transistors get smaller, their power density stays constant”
  - “if the transistor density doubles, power consumption (with twice the number of transistors) stays the same.”
- Moore’s Law: “the number of transistors doubles about every two years”
- Koomey's Law: “performance per watt doubles about every 18 months (1.5 years)”
  - Combination of Dennard Scaling and Moore’s Law

- Breakdown of Dennard Scaling around 2006
  - Transistor counts in integrated circuits are still growing
  - But the resulting improvements in performance are more gradual
  - No speed-ups resulting from significant frequency increases

- New hardware architectures: many more cores
  - each less powerful
  - less memory per core
  - memory bandwidth is going to be the limiting factor

- Software needs to
  - Use more cores simultaneously ➜ multi-threading
  - Use SIMD architectures to keep power consumption low ➜ GPUs in Supercomputers

https://en.wikipedia.org/wiki/Dennard_scaling



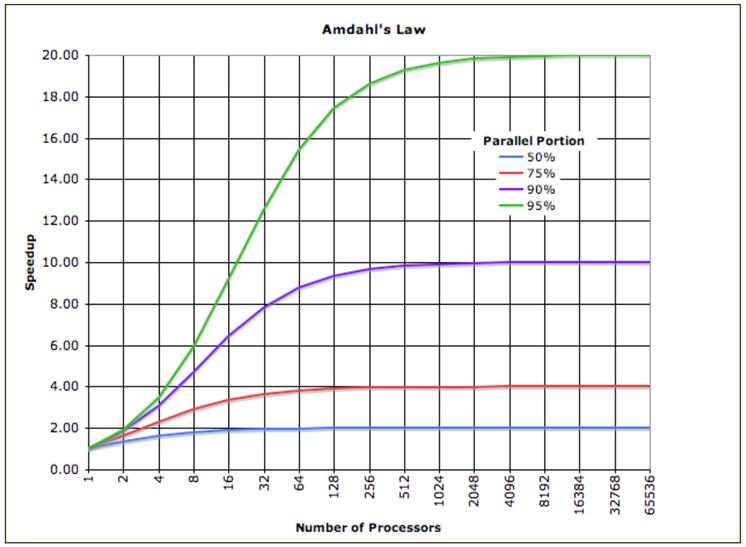
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/

More cores - less powerful - less memory

The answer: Multi-Threading

- Keywords
  - Multi-threading
  - Vectorization

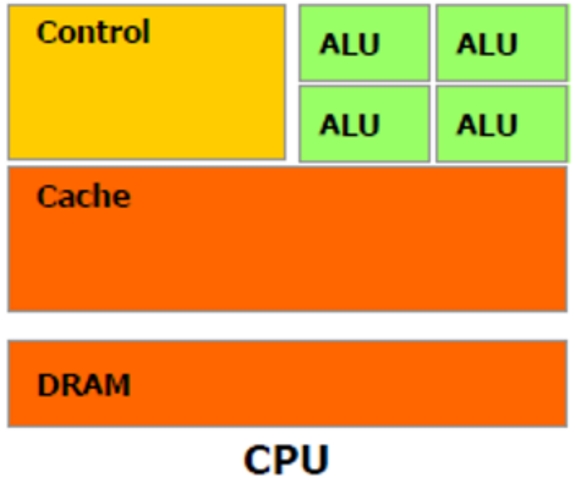


- CMSSW is using multi-threading since 2015
  - First framework of a major HEP experiment that uses true multi-threading in production workflows

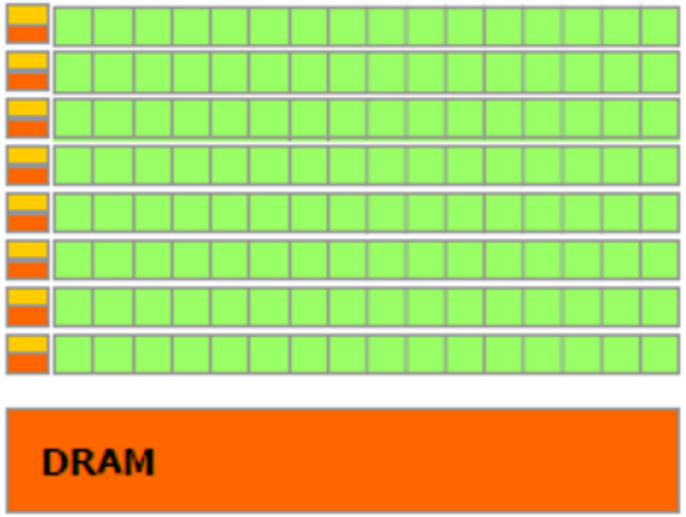
Multi-Threading is not Easy!

- Multi-threaded programming is different than what you are used to:
  - CMSSW designed specifically to hide complexity of multi-threaded programming
  - Only thing to remember: **don't use 'global' mutable data**

SIMD - Single Instruction Multiple Data



CPU



GPU

- CPU: multi-purpose processor optimized for serial execution with relatively narrow vectorization
  - For the experts: Up to 16 floats in AVX-512
- GPU: optimized for massive parallelism with vector-like mindset
  - Same kernel (of code) is executed on many different input data units
  - 1000s of parallel data units
- The big change is moving from current object-oriented data structures to data-oriented data structures
  - Current data structures are mainly arrays of structures (AoSs): objects with levels of inheritance
  - GPU programming: structures of arrays (SoAs): columns of simple data types
- A whole new programming paradigm
  - Worse than the transition from Fortran to C++ and single to multi-threaded C++
    - Multi-threading: very few changes to algorithms and data structures
    - SIMD: requires massive changes to data structures and new algorithms

Compute

CMS Computing Resources

- Over 70 sites world-wide
  - ~250,000 cores
  - ~200 Petabyte Disk
  - ~250 PB used tape space

CMS Computing Resources

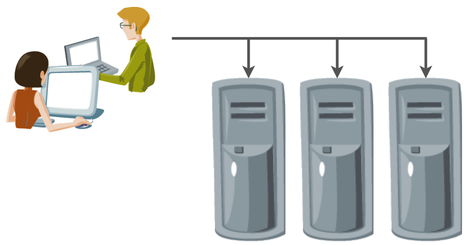
- HEP is a multi-national science discipline
  - Many nations contribute to our experiment(s)
  - Building a detector component, sending it to CERN, installing it and operating it is what we do!

- But how do we get access to all the computing resources we need?
  - Why don't we have one massive computing center at CERN?
    - Countries pledge resources to CMS
    - Countries prefer that these resources are located in their countries
      - Tax money is spent in respective countries

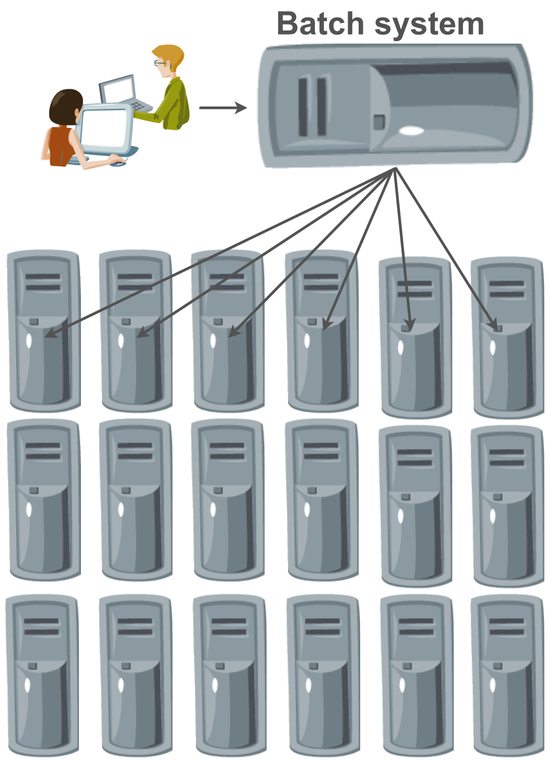
- The Grid
  - Countries provide access to resources
  - Everyone in CMS should be able to use them, no matter where the resources are or where the researcher is located!
  - The Grid allows for a globally distributed infrastructure to look like a single computing center

From Batch Systems to the Grid

- The Grid is all about **authentication** and **authorization**



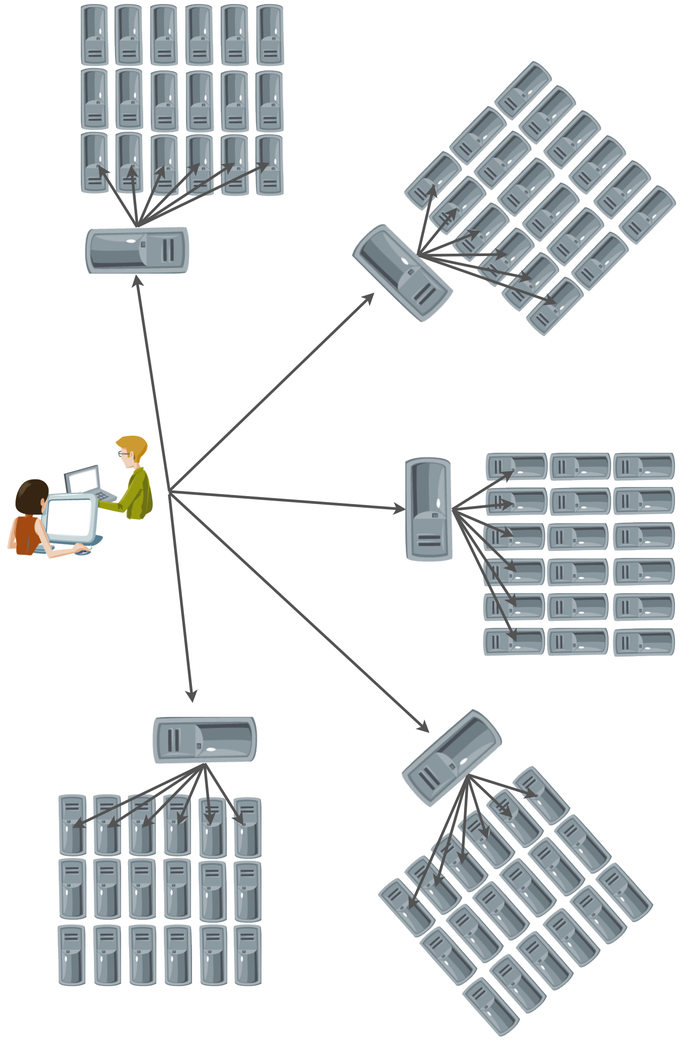
- Early times: several individual computers
  - User had to login to every single machine and start application
  - User accounts on all machines



- Batch system
  - Work queue
  - system that starts work on free machine ➜ user does not have to login to every single machine anymore
  - User account needs to be known to every machine in the cluster

From Batch Systems to the Grid

- Grid: interconnected batch systems
  - Problem: Synchronizing all user accounts to all sites not feasible
- Authentication: Grid certificates
  - Your Grid certificate is used to sign your jobs and all other Grid accesses
  - For security reasons, we use short term proxies of certificate
    - When a certificate gets compromised, it is only valid for a short period of time
- Virtual Organization
  - To identify different groups of users (for example different LHC experiments), we use virtual organizations (VOs)
  - CMS is a virtual organization
  - This is the reason why you need to create your proxy using
    - voms-proxy-init -voms cms

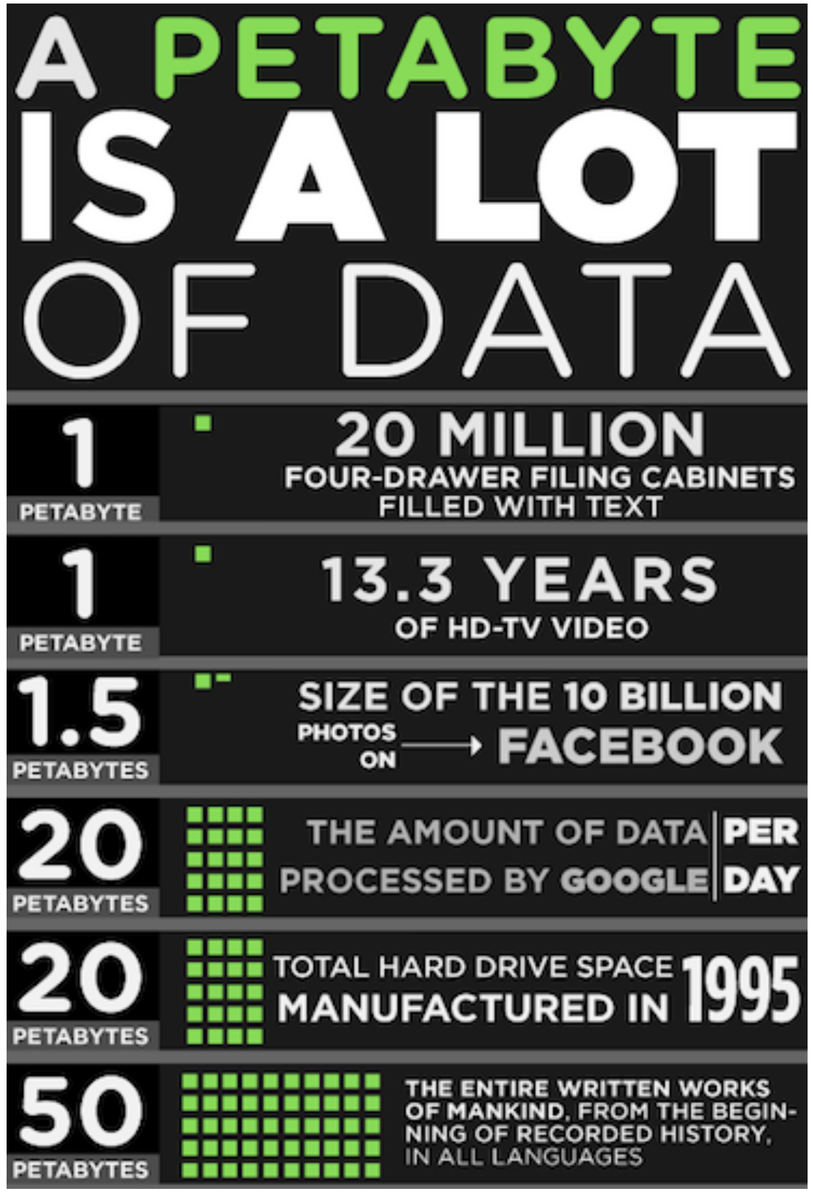


Tokens - the future

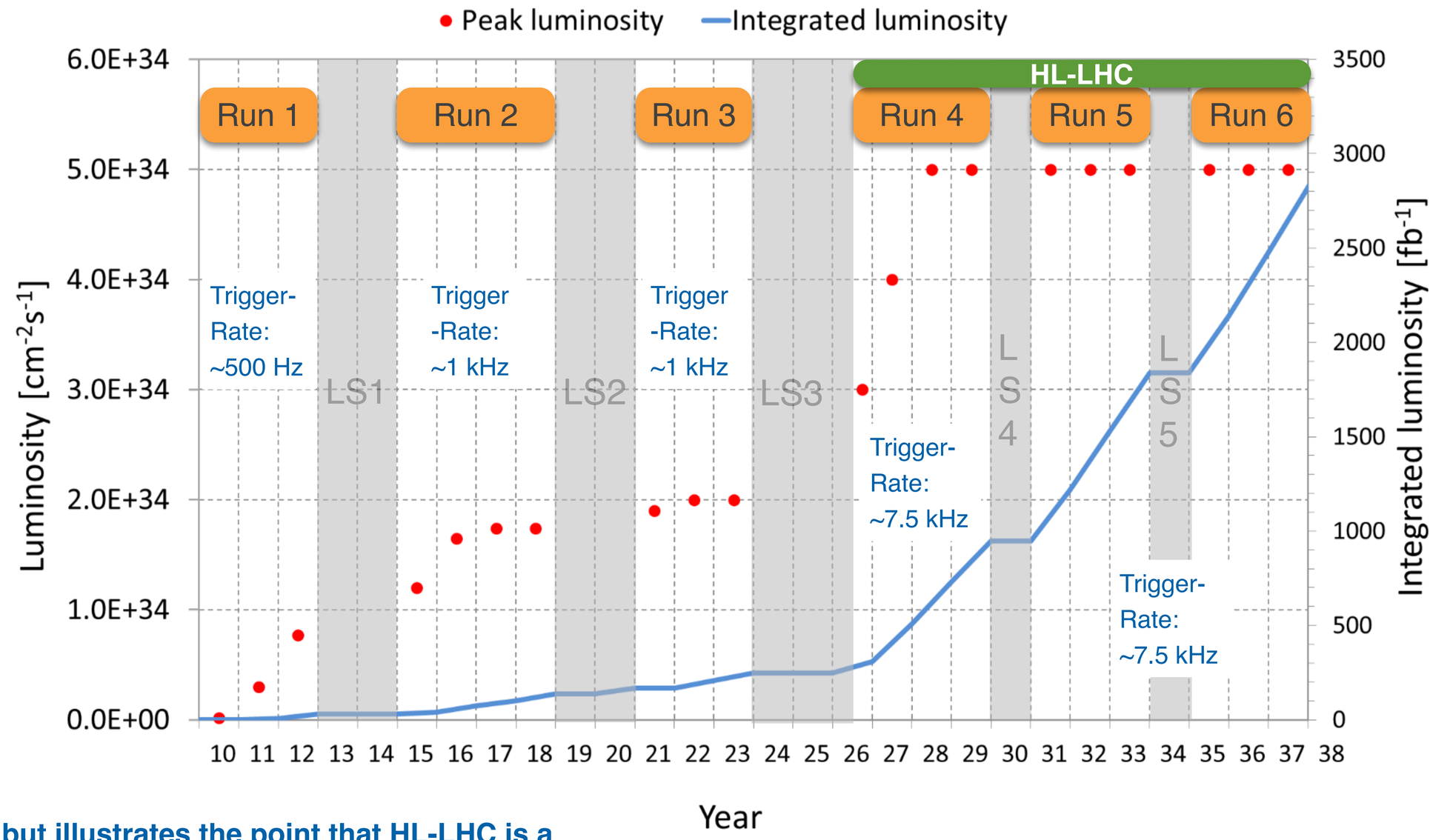
- Some basics
  - Authorization: Determining access rights or privileges to a given resource.
  - Authentication: establishing an identity for a remote entity.
  - Credentials: knowledge that establishes fact (identity).
- X.509 Certificates
  - certificate is issued by a ‘certificate authority’ (CA) and it contains both an identity and a cryptographic keypair
  - any action signed by the certificate’s keypair is assumed to be approved by the identity itself.
  - Anyone who has your credential has your power of attorney!
- Tokens: concept of capability-based authorization
  - least privilege: Each remote entity should only have enough authorization to perform its desired task – and no more!
  - A capability is an (unforgeable) token that authorizes a given action on an object.
    - bearer tokens; anyone who holds the token is assumed to be authorized to perform the action.
- Tokens are the internet standard: WLCG is transitioning to tokens in the next years
- Reference: https://sciauth.org

Storage

How Big is Data?

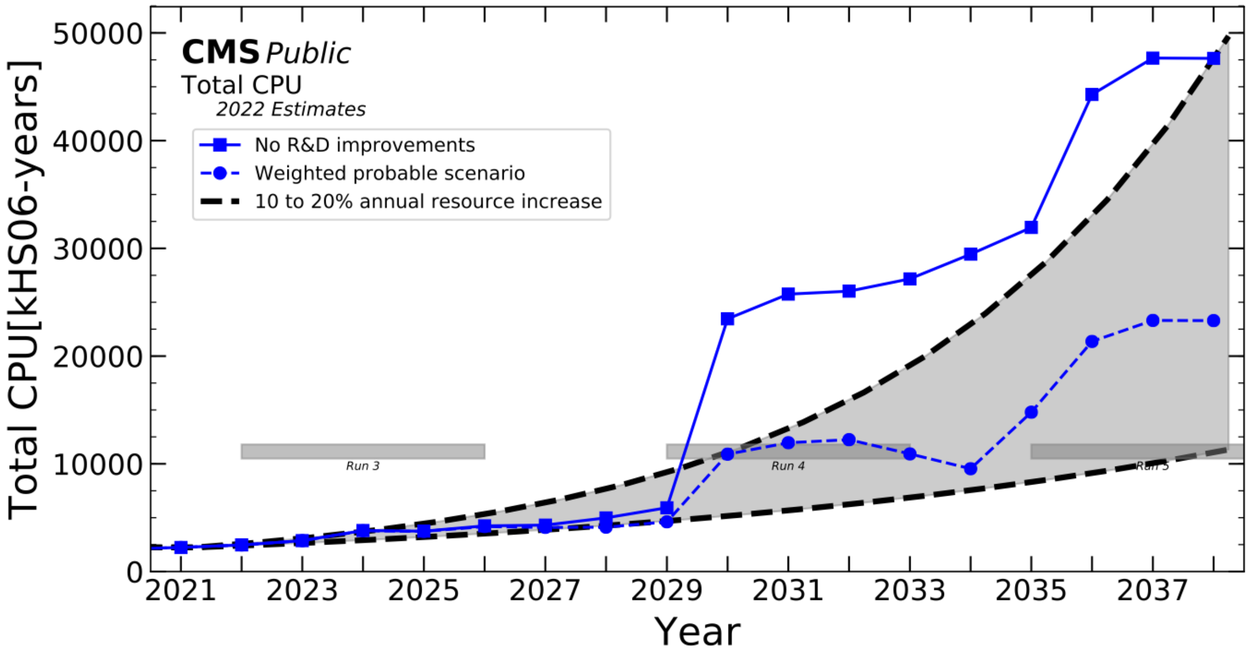


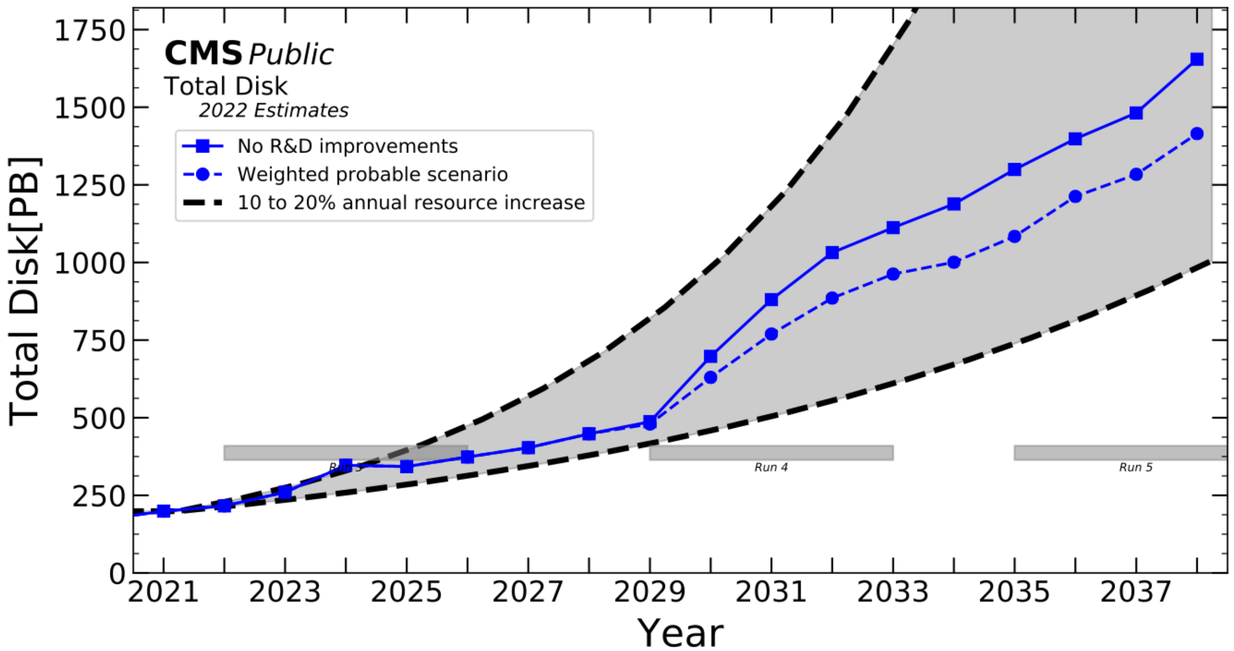
The Future - HL-LHC



Old schedule, but illustrates the point that HL-LHC is a significant step in data volume and complexity

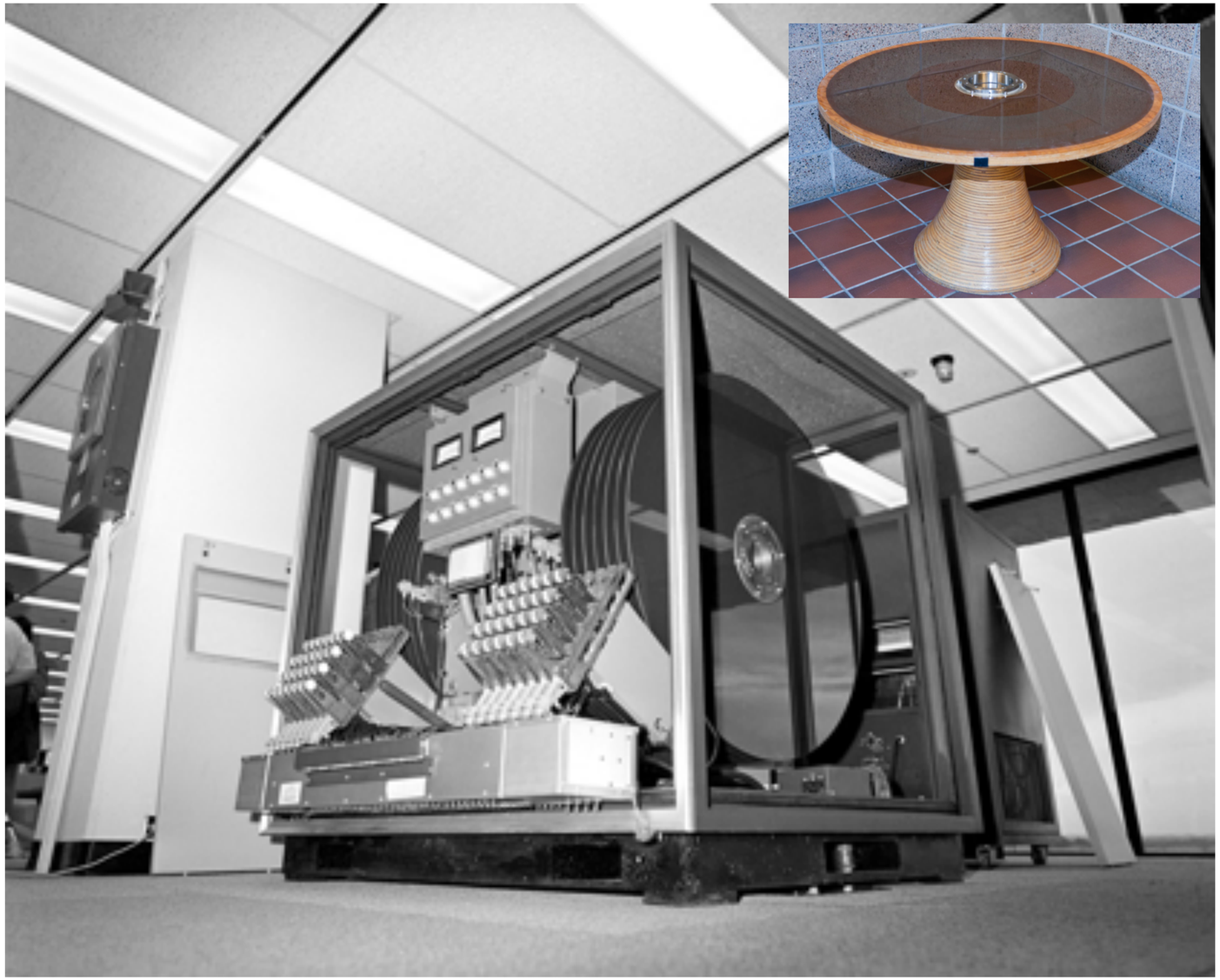
HL-LHC: Resource extrapolations





- Beginning of HL-LHC (2029):
  - CPU: 4-5 Million of today's cores
  - Disk: 500 PB, quickly climbing to > 1 Exabyte during Run-4

Disk in 1976



- Bryant Disk Drive (~1976)
  - 14 platter model 400 B
  - 100 MB capacity
  - 12 KVA
  - 3044 lbs weight

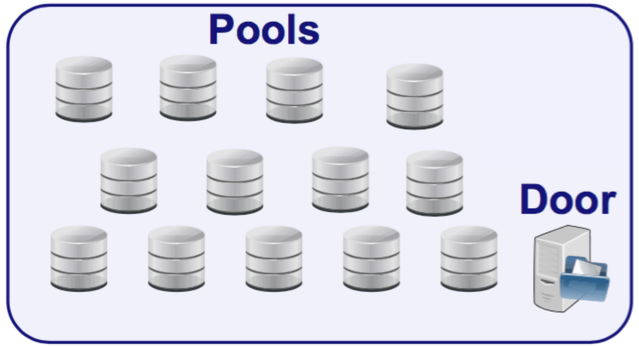
  - 1 TB:
    - 10,000 Bryant Disk Drives
    - 30 million lbs
    - $600M

Disk

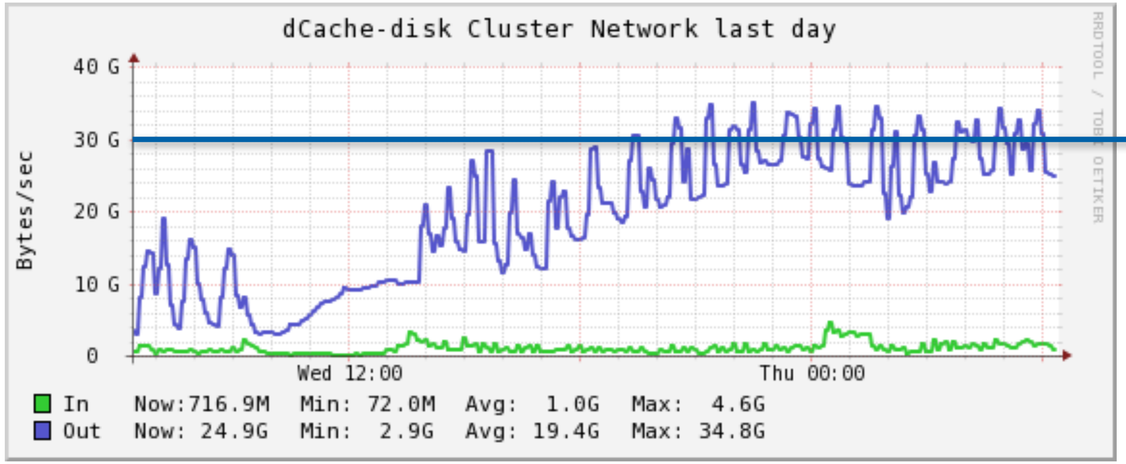
Harddrive

Diskpool

Storage System



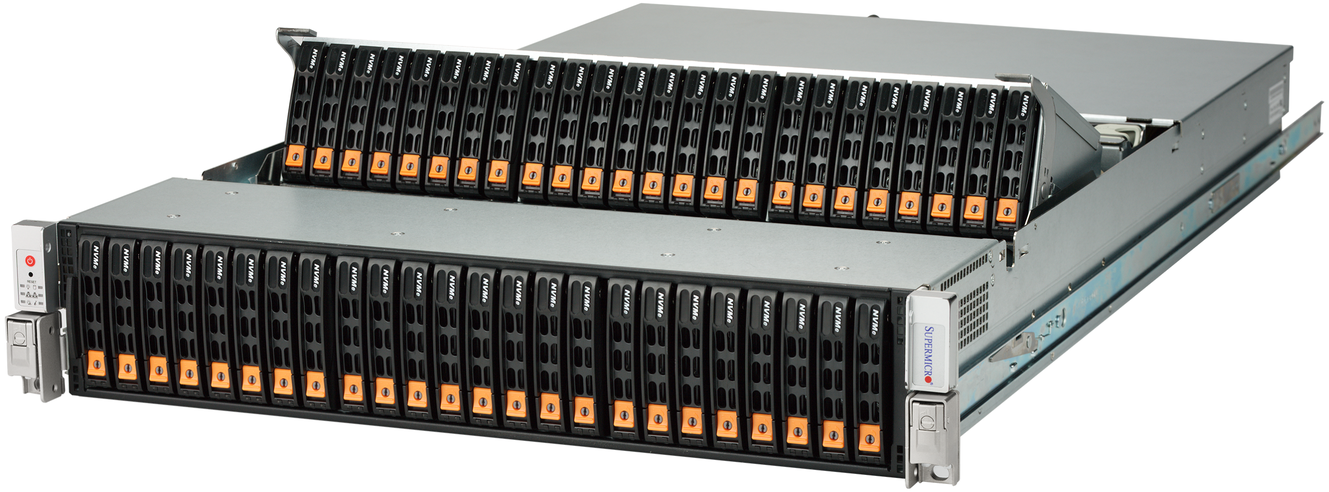
- Access to a lot of data from a lot of CPUs is provided through disk systems
  - Hard drive ➜ Disk pool ➜ Storage Systems
- We operate massive installations of many Petabytes that look like a single hard drive to the user
  - Storage systems handle access, provide replication
  - Access is not like you access as file on your hard drive (POSIX) but through special protocols (scaling reasons)
  - Using community systems like dCache, EOS, Ceph, but also industry systems like HDFS from Apache Hadoop



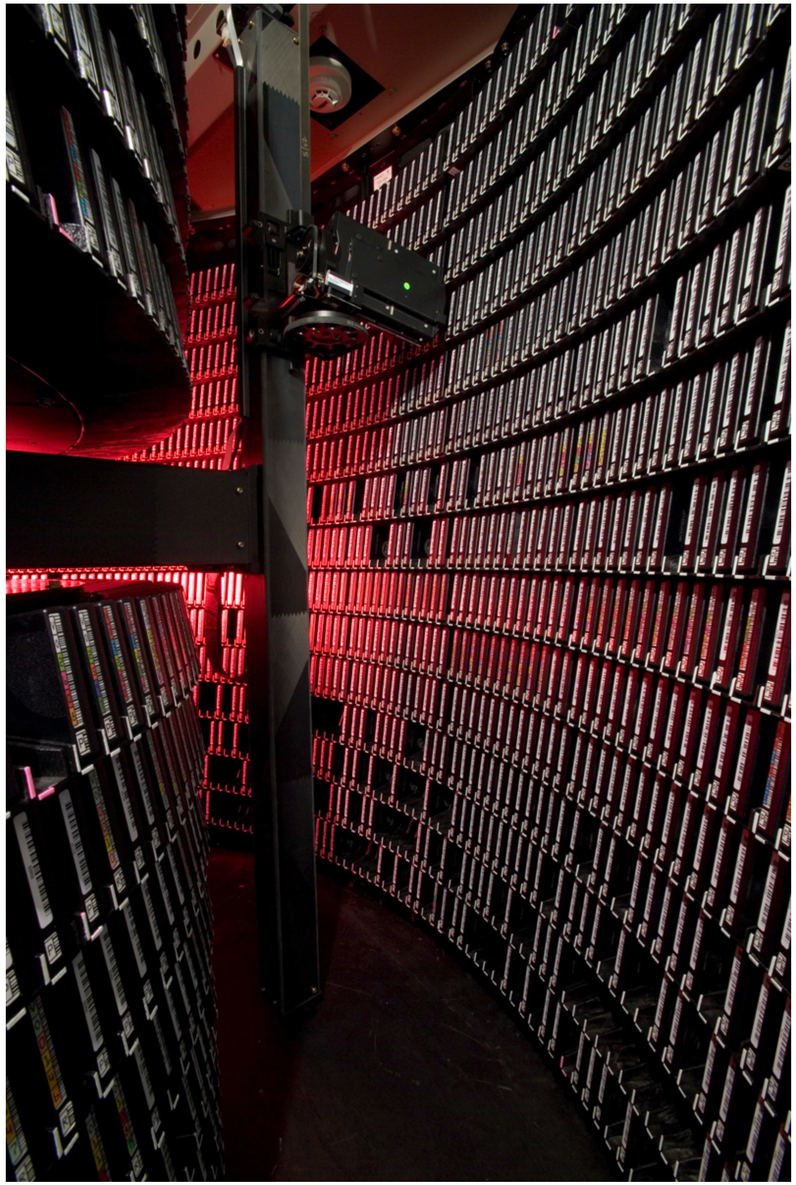
FNAL dCache
30 GB/s
(240 Gbps)

NVME - QoS

- Newer, faster forms of storage hardware
  - SSD arrays and NVME arrays
  - Provide a lot more I/O capabilities (throughput, IOPS, …)
- Although significantly more expensive than spinning disk, they provide advanced capabilities
  - Time to rethink storage setup and provide different hardware types for different Quality-of-Service needs? Design hierarchical storage setups that use faster storage for files that are accessed more frequently and more heavily?



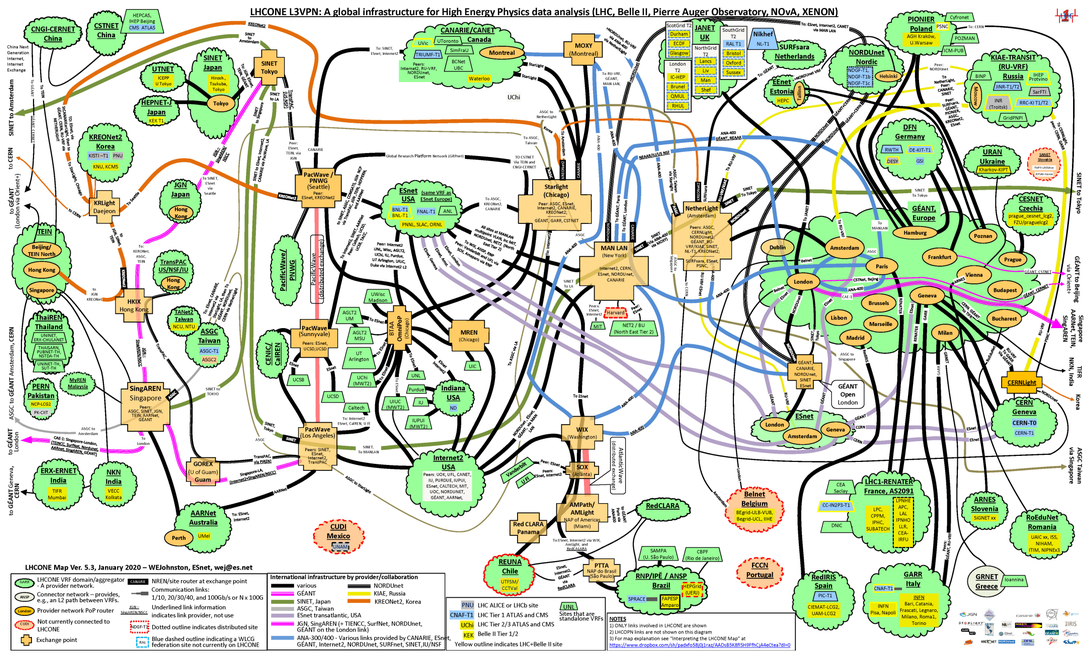
Tape

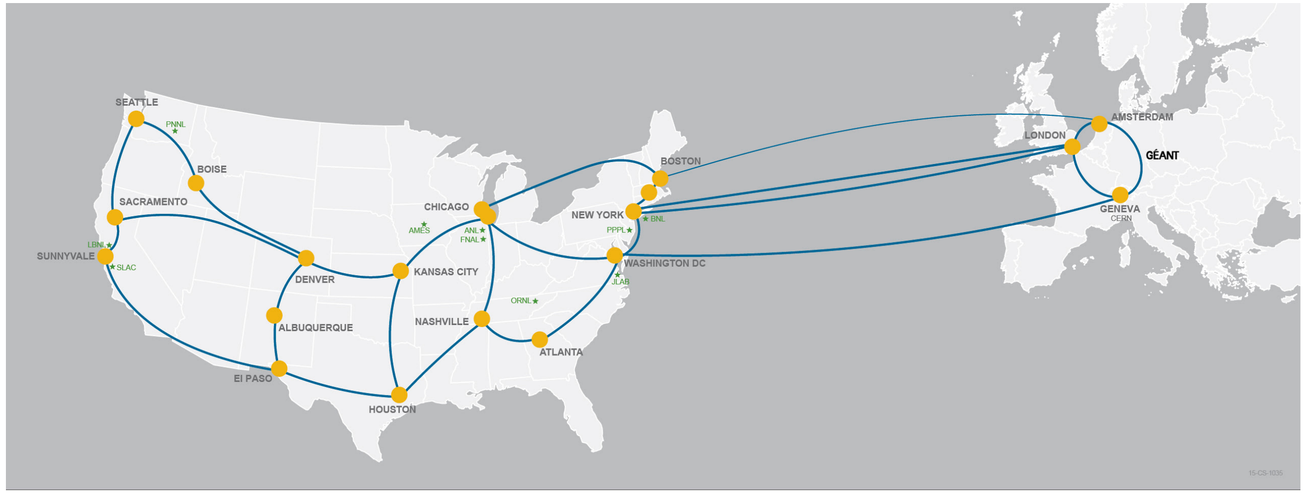


- Longterm storage is handled by magnetic tape
  - Still the cheapest way of storing Petabytes of data ➜ if you don't need to access it
  - For access, data has to be copied ("staged") to disk
- Individual cartridges can currently store more than 10 TB
- Individual robots have 10k - 15k cartridge slots

The hidden enabler: Scientific Networks

Scientific Networks are the most reliable resource form





ESnet
ENERGY SCIENCES NETWORK

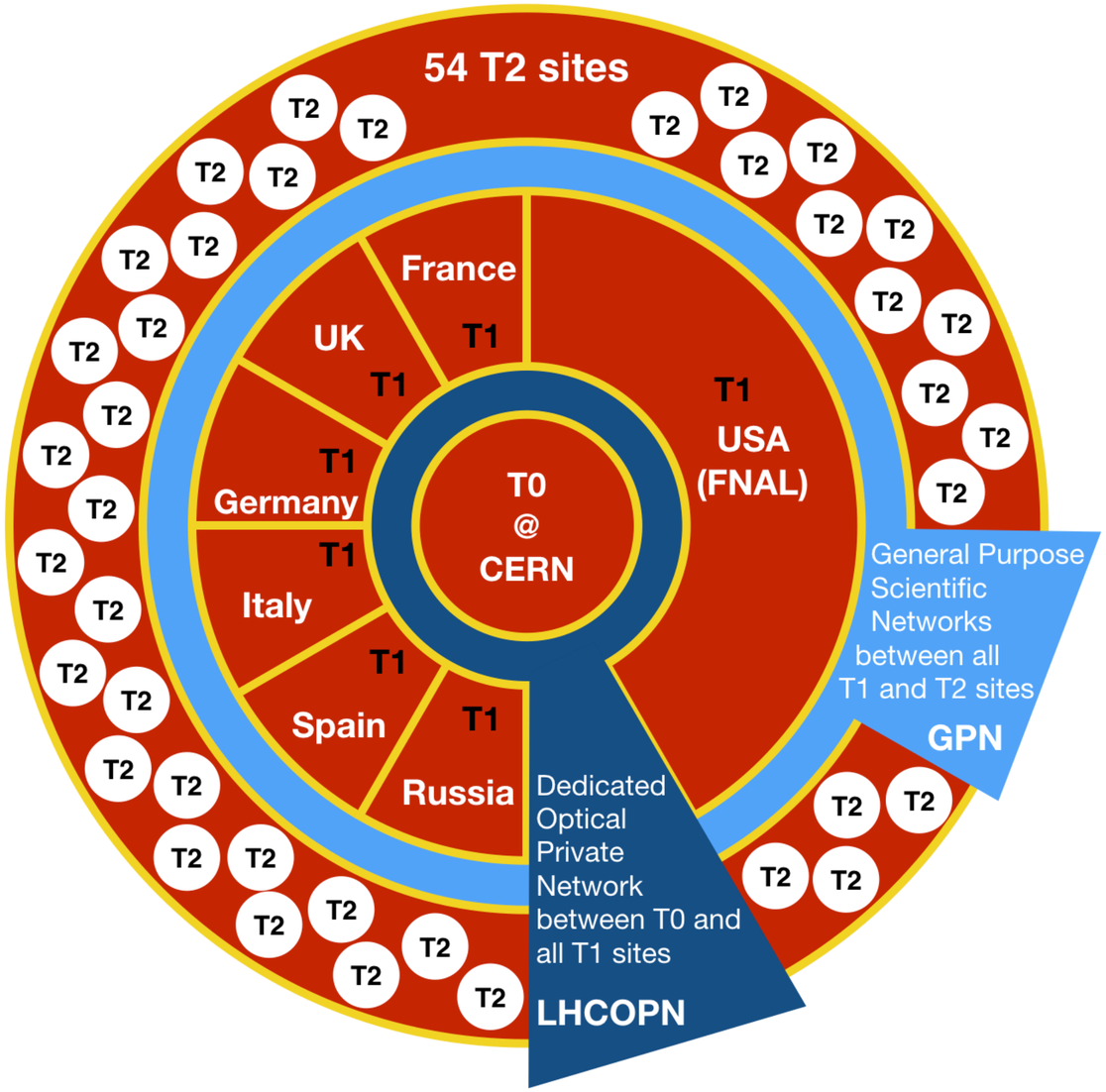
★ Department of Energy Office of Science National Labs

Ames	Ames Laboratory (Ames, IA)
ANL	Argonne National Laboratory (Argonne, IL)
BNL	Brookhaven National Laboratory (Upton, NY)
FNAL	Fermi National Accelerator Laboratory (Batavia, IL)
JLAB	Thomas Jefferson National Accelerator Facility (Newport News, VA)
LBNL	Lawrence Berkeley National Laboratory (Berkeley, CA)
ORNL	Oak Ridge National Laboratory (Oak Ridge, TN)
PNNL	Pacific Northwest National Laboratory (Richland, WA)
PPPL	Princeton Plasma Physics Laboratory (Princeton, NJ)
SLAC	SLAC National Accelerator Laboratory (Menlo Park, CA)

- Scientific networks are much more reliable than CPU and disk
  - Beginning of the LHC: topology was planned to be resilient to networking problems ➜ Today: full mesh: all sites connected to all sites
- Scientific networks are the key to global science
  - Interconnect continents and research institutes and universities, enabling large data science
- They will play an even bigger role in the future!

Data

CMS Grid Infrastructure



- Tier 0 (T0) at CERN (20% of all CMS computing resources)
  - Record and prompt reconstruct collision data
  - Calculate prompt condition and alignment constants
  - Store RAW data on tape (only archival copy, no access)
- Tier 1 (T1): regional centers in 7 countries (40% of all CMS computing resources)
  - Store one distributed copy of all data on tape (served copy)
  - Archive MC on tape
  - Skim data
  - Re-reconstruct data with newer software and conditions/alignment constants
- Tier 2 (T2): local computing centers at Universities and Laboratories (40% of all CMS computing resources)
  - Simulate MC events
  - User access to data for analysis
- Tier 3 (T3)
  - It's complicated!

Data Organization

- Event
  - real data (collision events)
  - MC events
- Lumi section
  - 23 seconds of data taking
  - Variable number of events for MC
    - number is calculated to be able to digitize-reconstruct all events in a lumi section in 12-24 hours
- Run
  - Run number from during data taking
  - MC has arbitrary run number
- Every data event is uniquely identifiable by
  - RUN:LUMI:EVENTNR
- Files
  - only complete lumi sections
  - necessary for correct book keeping
  - all uniquely named (with random name)
- Blocks
  - group of files
  - designed to fit on a tape cartridge to optimize tape staging (tape mount overhead and searching on tape is significant)
- Dataset
  - Group of blocks

The book keeping problem

- Data storage atomic unit: block

- The problem: How do we keep track of blocks at all the CMS sites:
  - Every site has their own storage setup and has different file paths (like a mount point in your linux distribution)
    - Site A: /mnt/eos/my-directory/my-root-file
    - Site B: /big-storage-system/my-other-directory/my-other-root-file
    - This is called the **Physical File Name (PFN)**
  - CMS created its own namespace which is the same at all sites
    - maps the file of a dataset into a **Logical File Name (LFN)**
    - /store/data/<era>/<primary dataset name>/<data tier name>/<processed dataset name>-<version>/<counter>/<filename>
  - Every site has a mapping between LFN and PFN
    - $CMS_SITE/SITECONF/local/PhEDEx/storage.xml
    - CMSSW and other systems can use these translations automatically
      - specifying an LFN in a CMSSW job is sufficient to open a file
  - **This is the reason why CMS data storage and distribution scales to many sites!**

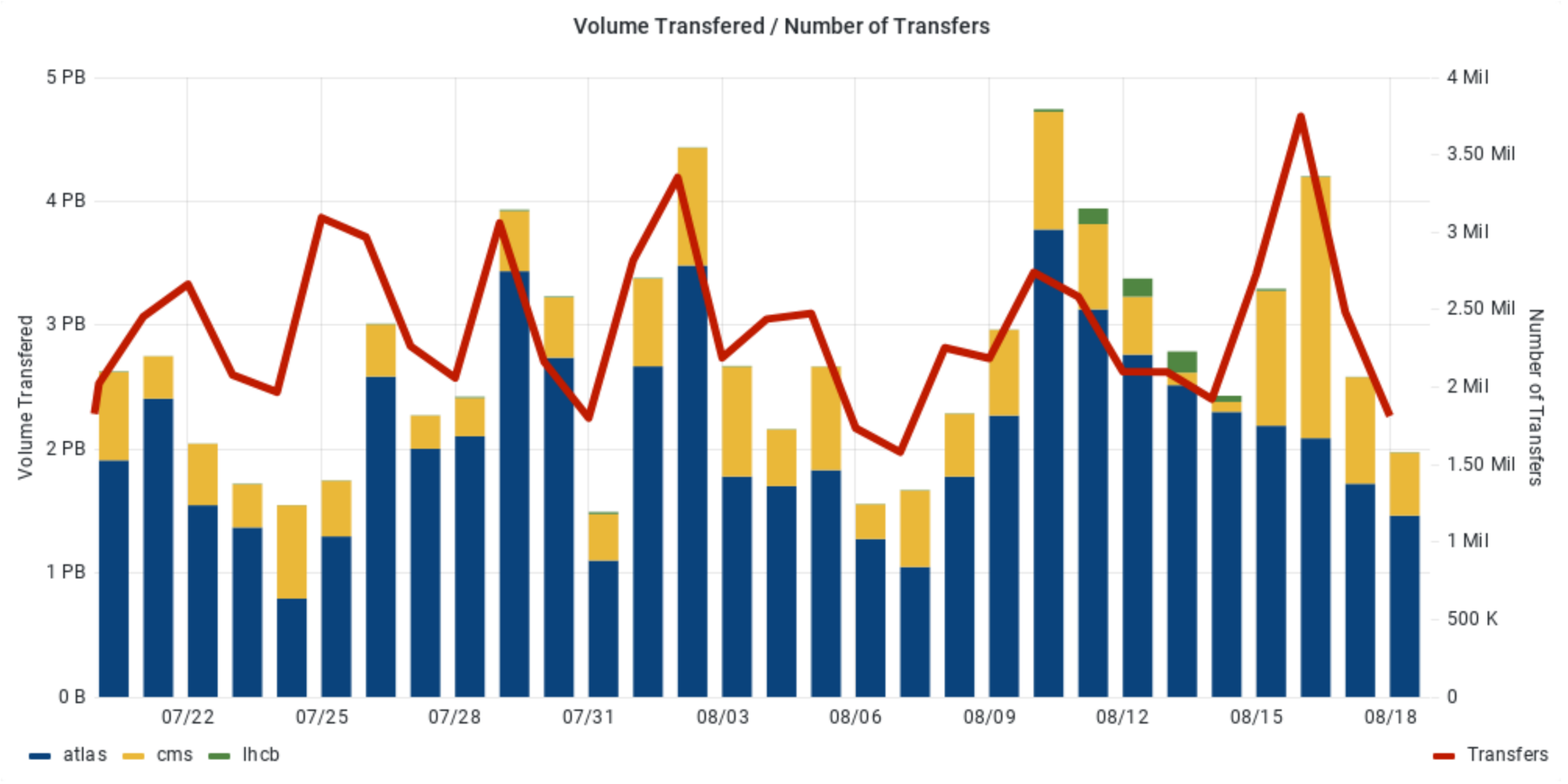
Data Transfers and Bookkeeping - Example systems

- Rucio: http://rucio.cern.ch
  - Manages files across distributed sites and orchestrates transfer of files between them
  - Scales to Millions of files and hundreds of Petabyte across many hundreds of sites
  - Rule based data distribution

- Data Bookkeeping System (DBS)
  - Database at CERN
  - all metadata of all CMS data

- DAS
  - Aggregation system that cross-correlates information from different sources (DBS, PhEDEx, …)

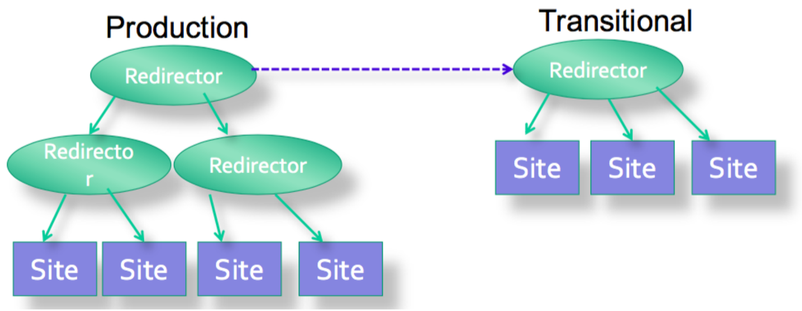
LHC transfers

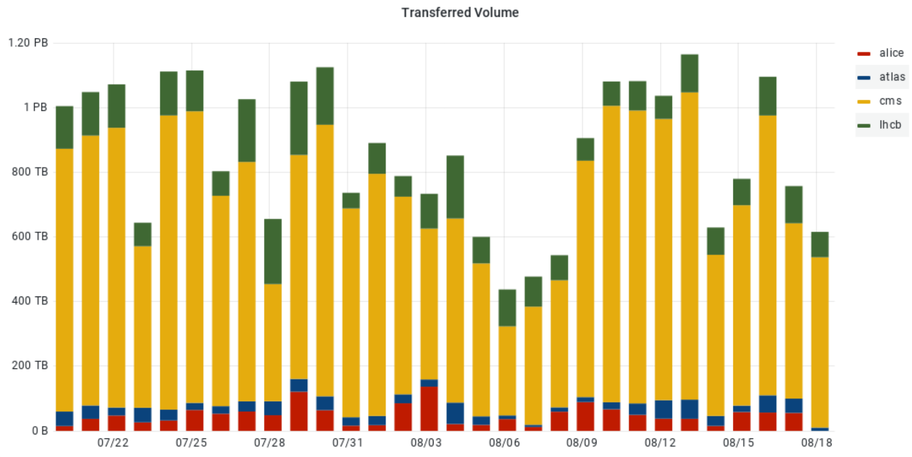


- All of LHC experiments: order of 100 PB transferred per month

XRootD: Any Data, Any Time, Anywhere (AAA)

- Based on opening a file through the wide area network
  - Example: file is stored at CERN but the application opens the file at Fermilab
  - Latency is significantly reduced (almost unnoticeable) through clever caching and pre-fetching in CMSSW



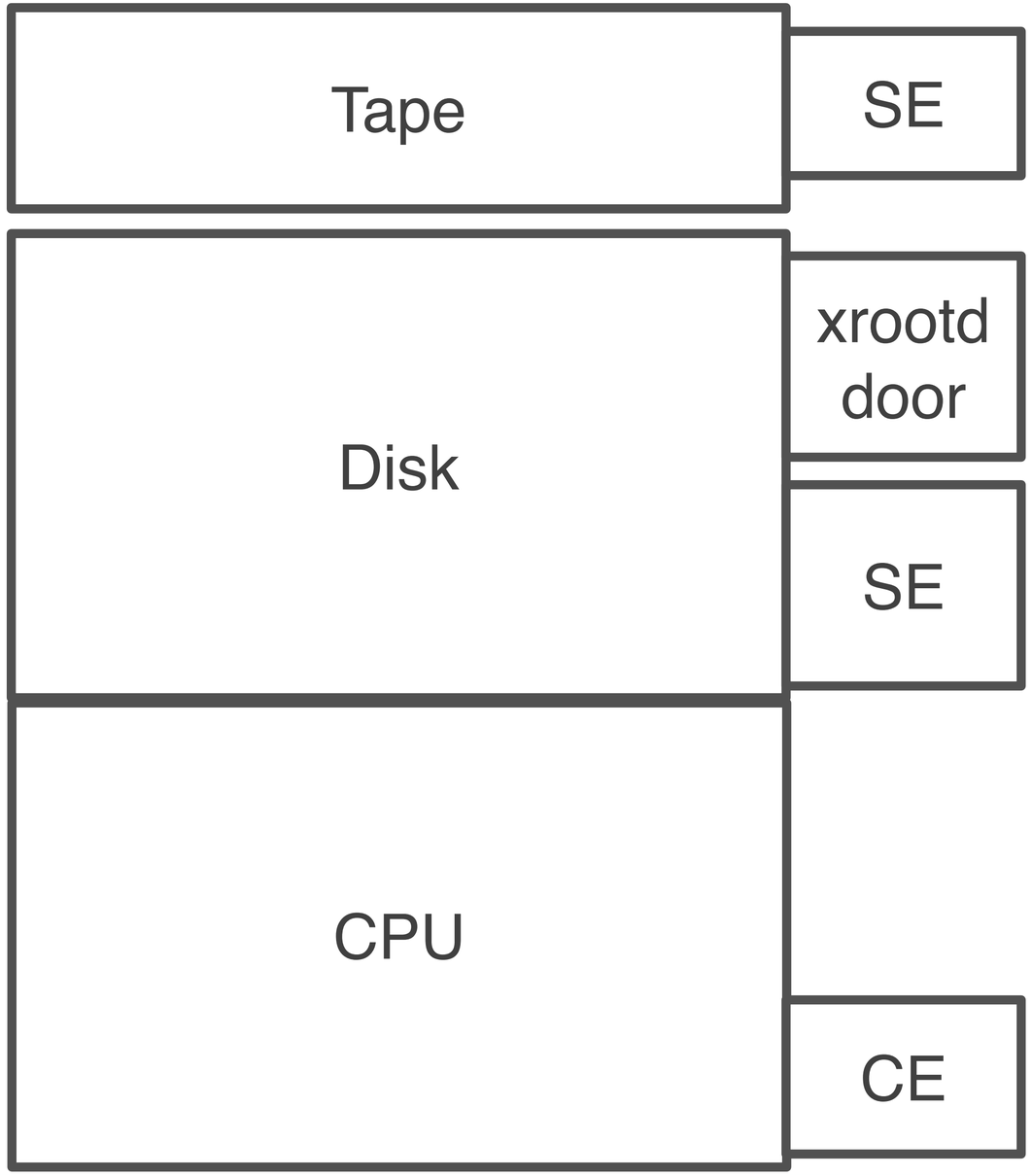


- How it works
  - If CMSSW cannot find a PFN, the request for the corresponding LFN is passed to a re-director ➜ Fall-back
  - The re-director sends UDP packages to all sites, the site who has the file and responds the quickest is asked to server the file
  - We have a redirector hierarchy to prefer local access (first read the file from a US site before going over the Atlantic link)

- Currently: of the order of 30 PB per month

What is a Grid site

Grid site



- CE: Compute Element
  - Grid door to CPU resources
  - Receives Grid job
  - Performs authentication through proxy and VOMS
  - Passes job to local batch system
- SE: Storage Element
  - Grid door into storage resources
  - Same as CE only for Grid file copy jobs
- xrootd door
  - Receives data federation requests for files from redirectors, responds and serves files over wide area network (streaming)

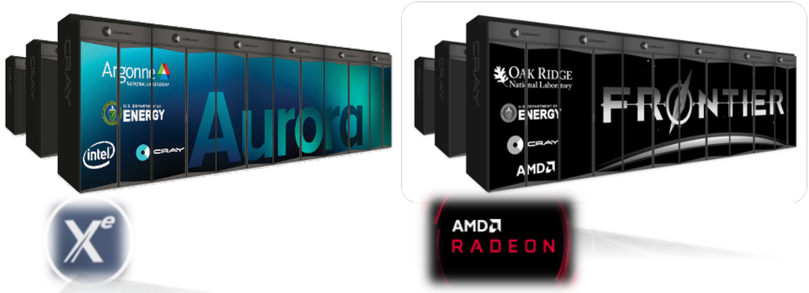
Beyond the Grid: HPC / cloud

Supercomputers

- Supercomputers are built for High Performance Computing (HPC)
  - HTC: many small processes, not data exchange between processes (HEP use case)
  - HPC: large single process spanning many cores with fast interconnects for data exchange (climate modeling, LatticeQCD, etc.)
- Because of the size of the installations, optimizing on FLOPS per Watt
  - Latest machines are based on GPUs
- National research infrastructures are investing in HPC machines
  - If the software is able to use HPC architectures, HEP can also use Supercomputers

		Accelerators				
		Intel	NVidia	AMD	FPGA	Other
CPU	Intel	Aurora	Cori Piz Daint Tsukuba MareNostrum		Tsukuba	
	AMD		Perlmutter	Frontier El Capitan		
	IBM		Summit Sierra MareNostrum			
	Arm		Wombat			
	Fujitsu					Fugaku

AMD CPU+NVIDIA GPU

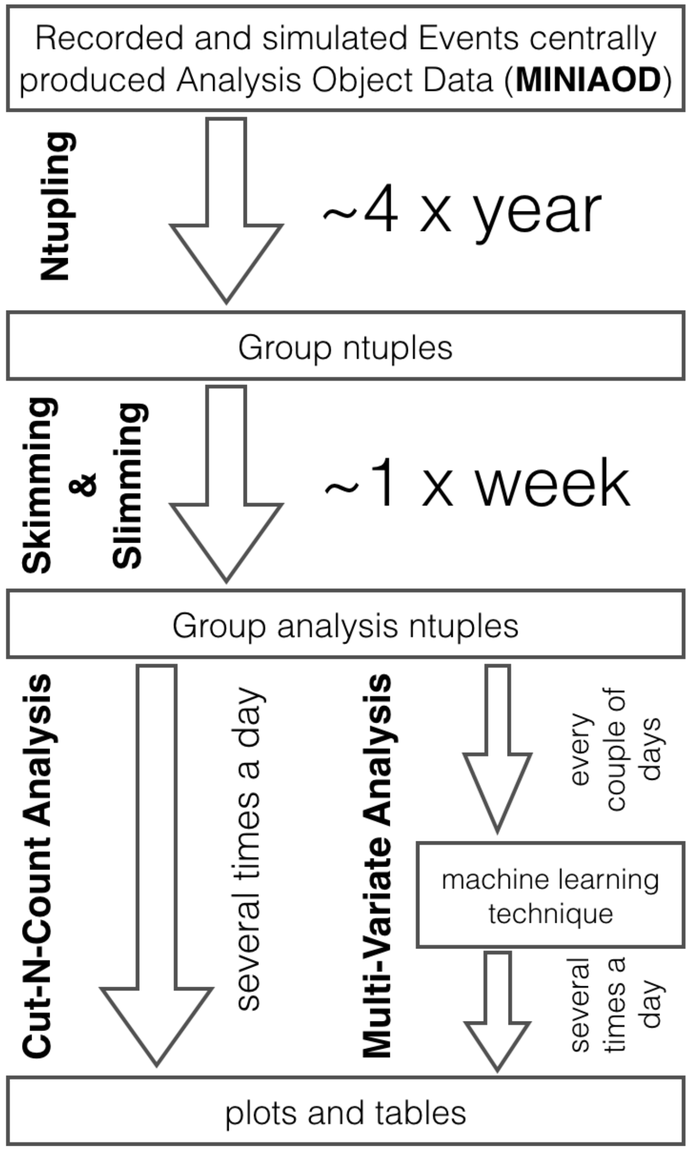


Commercial Clouds

- Rent virtual machines from Hyperscalers like Amazon Web Services, Microsoft Azure, Google Cloud, Oracle Cloud, etc.
  - Advantage:
    - Elasticity, because hyperscalers have many millions of cores
    - Newest hardware: cloud provider carry the risk of inventory and hardware replacement, also exotic hardware like the latest GPUs can be rented
  - Disadvantage:
    - Can be expensive, more than Grid sites
    - Data volume is expensive, especially if data has to be copied out of the cloud
- HEP is exploiting clouds when “the price is right”

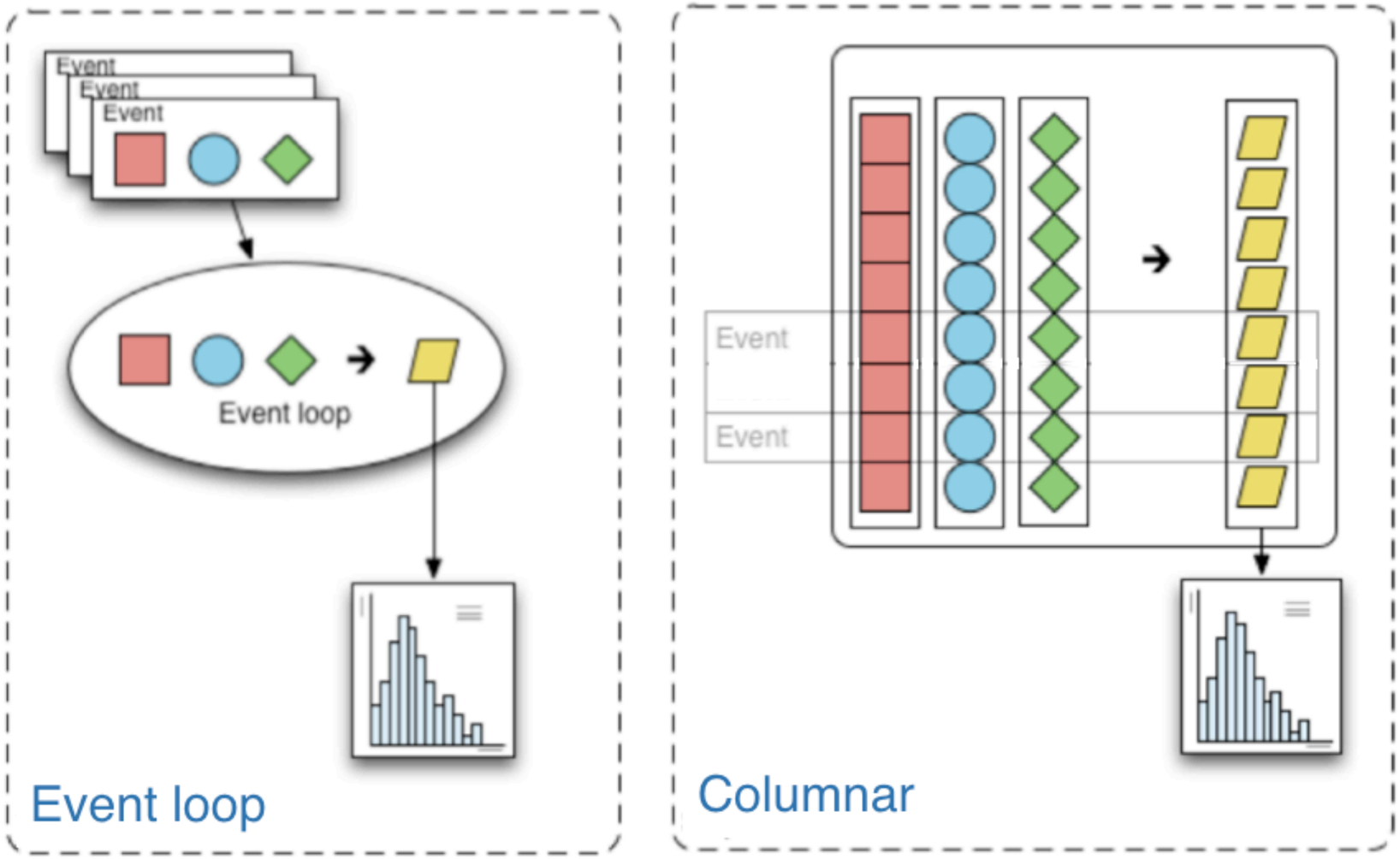
Analysis

Technically, what is HEP analysis right now? ➜ ROOT



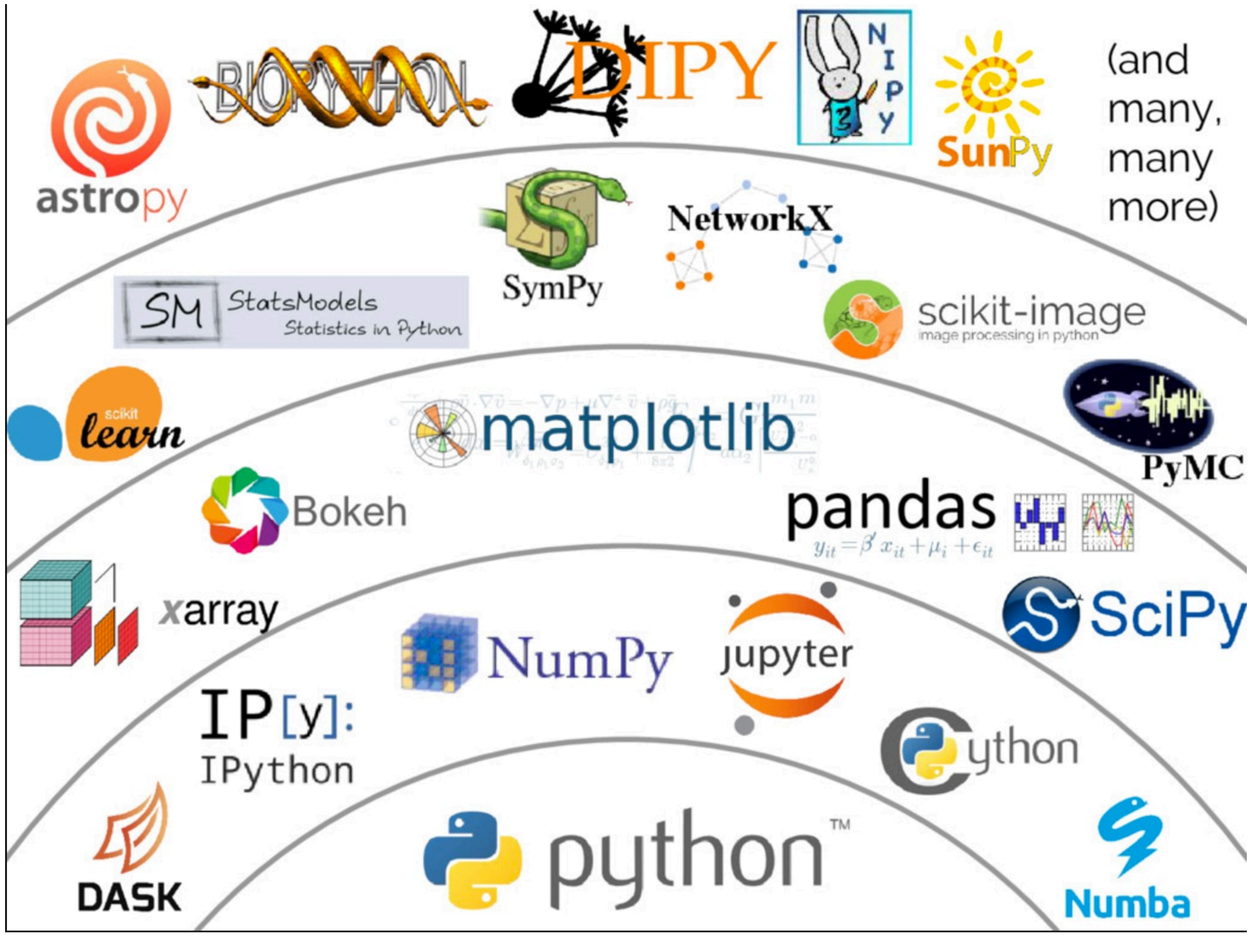
- ROOT based analysis: https://root.cern.ch/
- AOD and AODSIM were designed to be used directly for analysis to make plots
  - Objects are browsable and readable from bare root
  - Data reduction through central or user skims
- In LHC Run 2, we created the MINIAOD and MINIAODSIM data tiers
  - Factor ~10 smaller than AOD and MINIAOD
  - Designed to satisfy the needs of 80% of all analyses
- Still, most analysts and analysis groups convert the official data tiers into ntuples (flat files with vectors of simple types)
- Multi-step process
  - Convert data and simulation into ntuples
  - Skimming & Slimming
  - Plots & Tables

New: Columnar Analysis



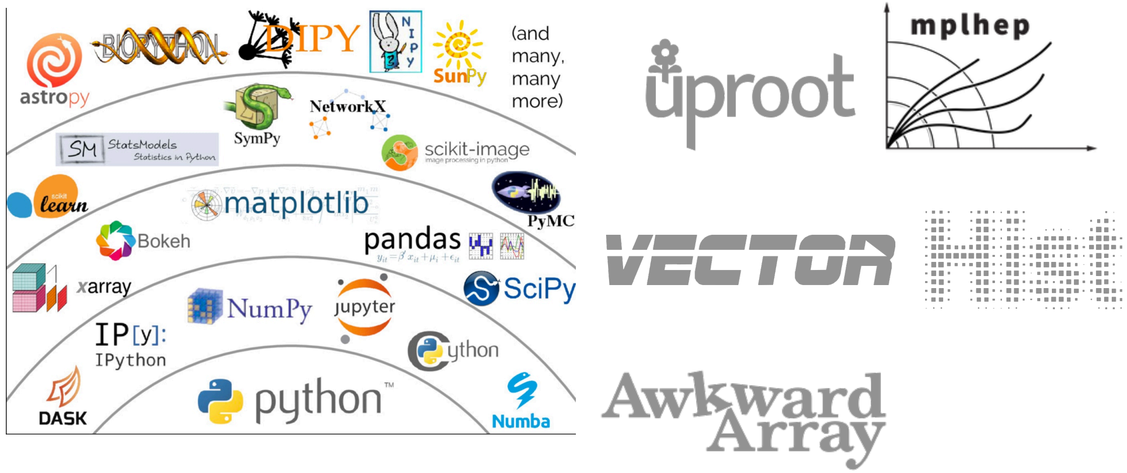
- Traditionally, data is analyzed through an event loop
- Columnar analysis: analyze same properties across events (functional programming)

Python Ecosystem



- Data science tools based on python are commonly used in industry
  - Much larger user and developer base than science ➜ science benefits from ecosystem

COFFEA



- COFFEA - Columnar Object Framework For Effective Analysis
  - https://coffeateam.github.io/coffea/
  - Serves as incubator for rapid prototyping of **missing pieces** in our ecosystem

Stronger Together

Our partners - we are not alone

Who?

IRIS-HEP, NSF USA - 2018

17 universities and research centers are participating in the project:

ErUM-DATA, Helmholtz Institute Germany - 2019

HEP-CCE, DOE USA - 2019

2 hardware projects (HEP&NP)
SciDAC, ECP (ASCR)

170 collaboration members - 2001

7 universities

PATH, NSF USA - 2020

EP R&D, CERN - 2020

19 universities + RAL

SWIFT-HEP, STFC 2020 and
ExCALIBUR-HEP, UKRI UK - 2021

Celeritas - ECP, DOE USA - 2021

37 universities + 8 industry participants

AIDAInnova, European Commission EU -2021

Plus lots of single PI level programs and:
DOE SciDAC
NSF Institutes
DOE Quantum Centers

7

- Elizabeth Sexton-Kennedy (FNAL) @ Seattle Snowmass Summer Meeting 2022 - Collaborative initiatives across experiment boundaries

Coordination

- After the CWP many of the groups that collaborated on writing the paper continued on as HSF & WLCG working groups.

WLCG Archival Storage Group
- WLCG Archival Storage Group

WLCG Network Throughput WG
- WLCG Network Throughput WG

WLCG Security Working groups
- WLCG Traceability Working Group
- WLCG Authorisation Working Group
- WLCG Federated Operations Security Working Group

WLCG Containers Working Group
- WLCG Containers Working Group

Kubernetes WG
- Kubernetes

WG for Transition to Tokens and Globus Retirement
- WG for Transition to Tokens and Globus Retirement

WLCG
Worldwide LHC Computing Grid

HSF
HEP Software Foundation

- Data Analysis
- Detector Simulation
- Frameworks
- Physics Generators
- PyHEP - Python in HEP
- Reconstruction and Software Triggers
- Software Developer Tools and Packaging
- Training
- Differentiable Computing
- Season of Docs
- Google Summer of Code
- intelligent Data Delivery Service
- Licensing
- Quantum Computing
- Reviews
- Visualisation
- Analysis Facilities

- Elizabeth Sexton-Kennedy (FNAL) @ Seattle Snowmass Summer Meeting 2022 - Collaborative initiatives across experiment boundaries

Introduction - HSF

- Reflecting the global nature of HEP, we have a global umbrella software organization in the HEP Software Foundation, HSF.
- HSF does not itself seek funding, but supports bids to funding agencies for projects that align with the goals documented in the Community White Paper, CWP:

A Roadmap for HEP Software and Computing R&D for the 2020s

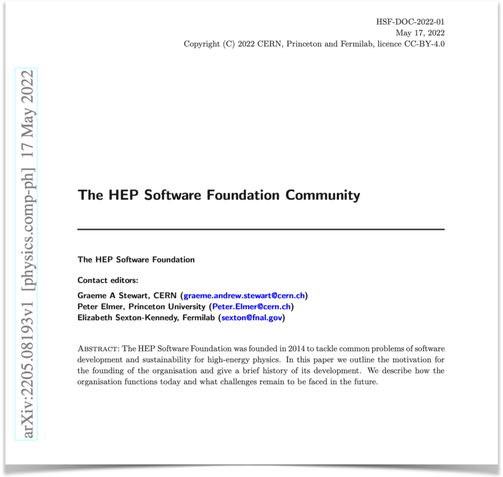
HEP Software Foundation Collaboration • Johannes Albrecht (Dortmund U.) Show All(310)
Dec 18, 2017

49 pages
Published in: *Comput.Softw.Big Sci.* 3 (2019) 1, 7
Published: Mar 20, 2019
e-Print: 1712.06982 [physics.comp-ph]
DOI: 10.1007/s41781-018-0018-8 (publication)
Report number: HSF-CWP-2017-01, HSF-CWP-2017-001, FERMILAB-PUB-17-607-CD
View in: OSTI Information Bridge Server, HAL Archives Ouvertes, CERN Document Server, ADS Abstract Service

pdf links cite 141 citations

arXiv:1712.06982v5 [physics.comp-ph]

3


HSF-DOC-2022-01
May 17, 2022
Copyright (C) 2022 CERN, Princeton and Fermilab, licence CC-BY-4.0

arXiv:2205.08193v1 [physics.comp-ph] 17 May 2022

The HEP Software Foundation Community

The HEP Software Foundation

Contact editors:
Graeme A Stewart, CERN (graeme.andrew.stewart@cern.ch)
Peter Elmer, Princeton University (Peter.Elmer@cern.ch)
Elizabeth Sexton-Kennedy, Fermilab (sexton@fnal.gov)

Abstract: The HEP Software Foundation was founded in 2014 to tackle common problems of software development and sustainability for high-energy physics. In this paper we outline the motivation for the founding of the organisation and give a brief history of its development. We describe how the organisation functions today and what challenges remain to be faced in the future.

arXiv:2205.08193v1 [physics.comp-ph]

- Elizabeth Sexton-Kennedy (FNAL) @ Seattle Snowmass Summer Meeting 2022 - Collaborative initiatives across experiment boundaries

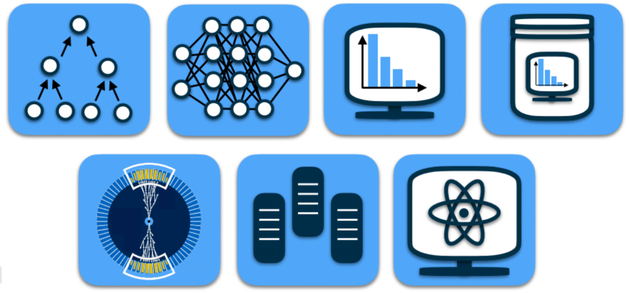
For more information

Some reading

DPF Community Planning Exercise

The Particle Physics Community Planning Exercise (a.k.a. "Snowmass") is organized by the Division of Particles and Fields (DPF) of the American Physical Society. Snowmass is a scientific study. It provides an opportunity for the entire particle physics community to come together to identify and document a scientific vision for the future of particle physics in the U.S. and its international partners. Snowmass will define the most important questions for the field of particle physics and identify promising opportunities to address them. (Learn more about the history and spirit of Snowmass here "How to Snowmass" written by Chris Quigg). The P5, Particle Physics Project Prioritization Panel, will take the scientific input from Snowmass and develop a strategic plan for U.S. particle physics that can be executed over a 10 year timescale, in the context of a 20-year global vision for the field.

Report of the 2021 US Community Study on the Future of Particle Physics organized by the APS Division of Particles and Fields



Conveners: V. D. Elvira (> Fall '21), S. Gottlieb, O. Gutsche (< Fall '21), B. Nachman

Topical Group Conveners: S. Bailey (> Fall '21), W. Bhimji, P. Boyle, G. Cerati, M. Carrasco Kind, K. Cranmer, G. Davies, V. D. Elvira (< Fall '21), R. Gardner (< Fall '21), K. Heitmann, M. Hildreth (< Fall '20), W. Hopkins, T. Humble, M. Lin (> Fall '21), P. Onyisi, J. Qiang, K. Pedro (> Fall '21), G. Perdue, A. Roberts, M. Savage, P. Shanahan, K. Terao, D. Whiteson, F. Weurthwein

Topical groups

Name
CompF1: Experimental Algorithm Parallelization
CompF2: Theoretical Calculations and Simulation
CompF3: Machine Learning
CompF4: Storage and processing resource access (Facility and Infrastructure R&D)
CompF5: End user analysis
CompF6: Quantum computing
CompF7: Reinterpretation and long-term preservation of data and code

Report Drafts

Please note that these are draft reports. We welcome your feedbac

- CompF Frontier Report (Google doc for feedback): v0

- CompF1: v0
- CompF2: v0
- CompF3: v0
- CompF4: v0
- CompF5: v0
- CompF6: v0
- CompF7: v0

- Draft reports are now available: https://snowmass21.org/computational/start#report_drafts

The Future

“It’s all about data and the science we do with it!”