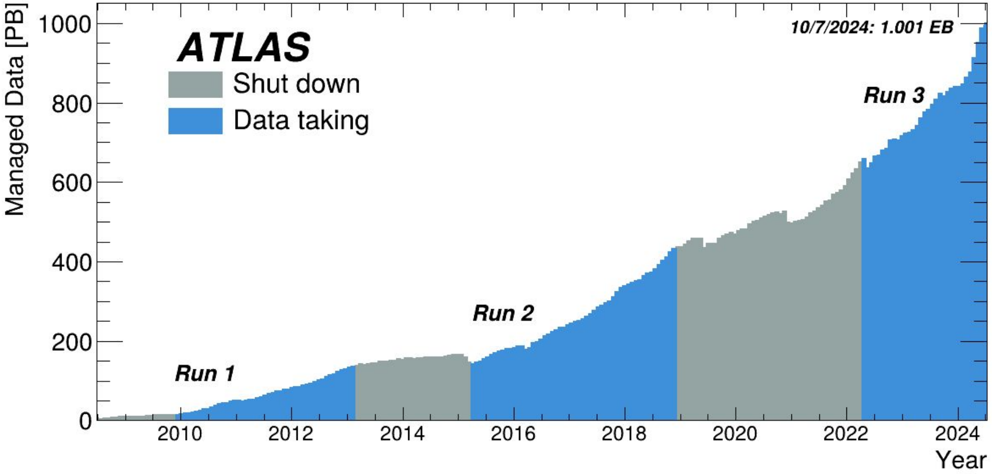
Computing @ ATLAS – in a glance



Rui Wang

Argonne National Laboratory

CPS Task Force Kick-Off Meeting: Data-Intensive Computing Task Force

Dec 19, 2024

U.S. DEPARTMENT OF ENERGY
Argonne National Laboratory is a
U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC.

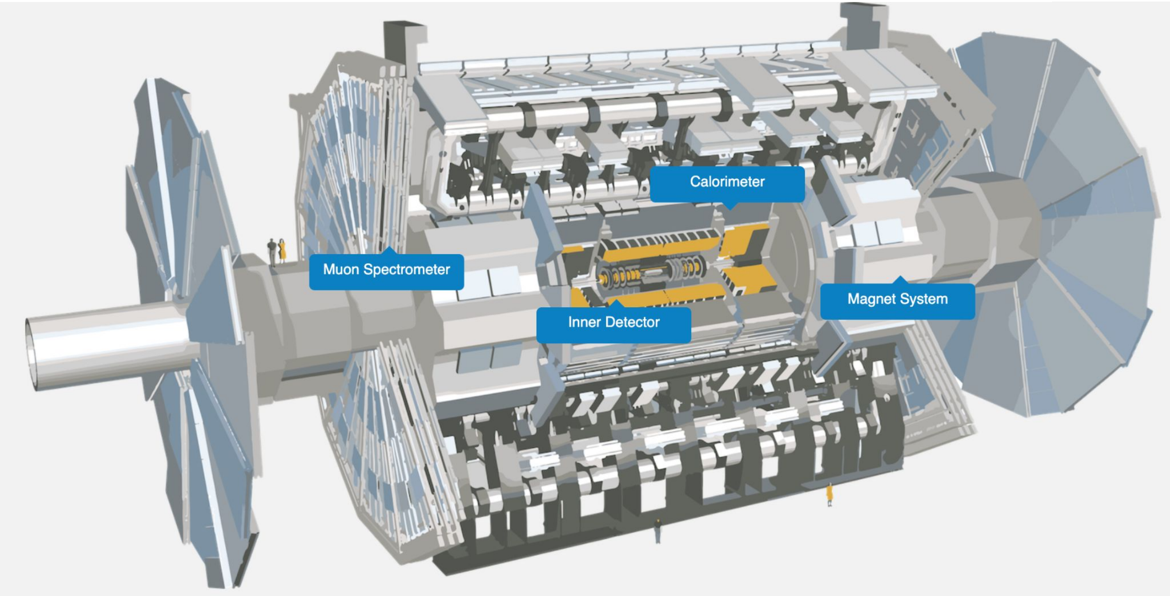
Introduction of the ATLAS experiment

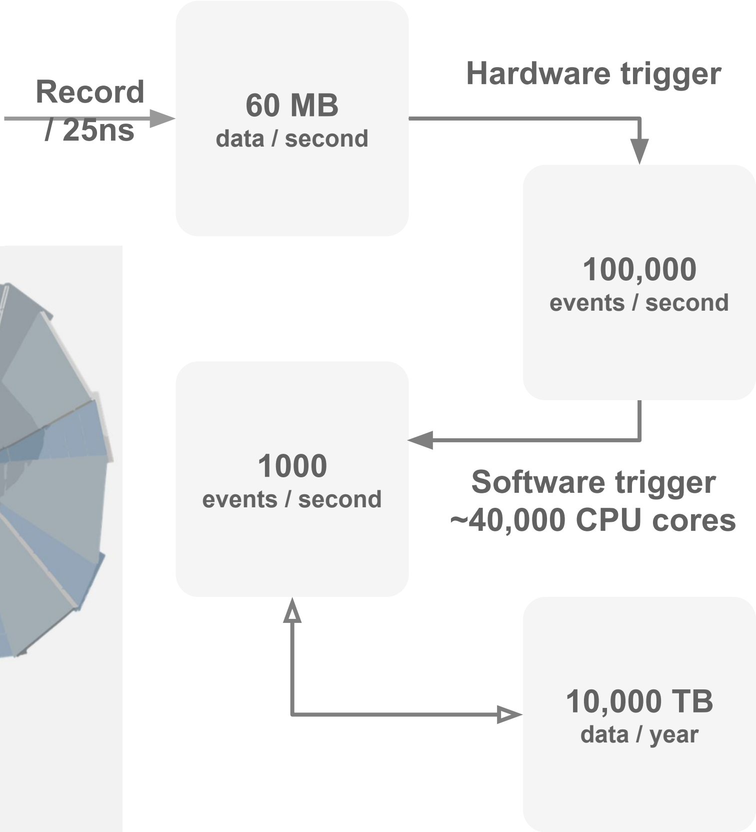
46
meters long & 25m of diameter

100
meters below ground

7000
tonnes

1000000000
collisions / second





Record / 25ns

60 MB
data / second

Hardware trigger

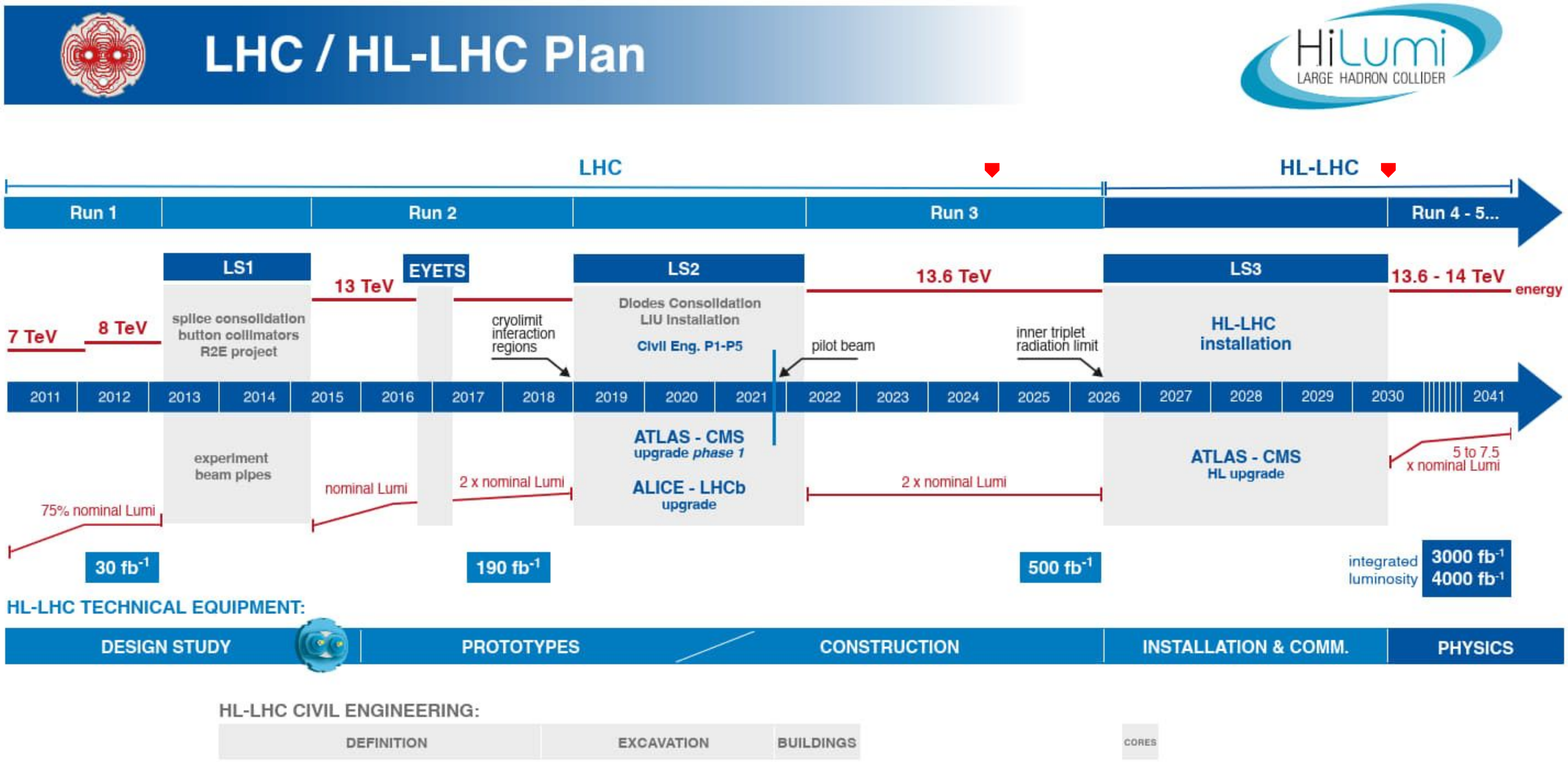
100,000
events / second

Software trigger
~40,000 CPU cores

1000
events / second

10,000 TB
data / year

Towards HL-LHC (x5 luminosity)



LHC / HL-LHC Plan

HiLumi
LARGE HADRON COLLIDER

LHC

HL-LHC

Run 1

Run 2

Run 3

Run 4 - 5...

LS1

EYETS

LS2

LS3

7 TeV

8 TeV

13 TeV

13.6 TeV

13.6 - 14 TeV

energy

splice consolidation
button collimators
R2E project

cryolimit
interaction
regions

Diodes Consolidation
LIU Installation
Civil Eng. P1-P5

pilot beam

inner triplet
radiation limit

HL-LHC
installation

2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026 2027 2028 2029 2030 2041

experiment
beam pipes

ATLAS - CMS
upgrade *phase 1*

ALICE - LHCb
upgrade

ATLAS - CMS
HL upgrade

75% nominal Lumi

nominal Lumi

2 x nominal Lumi

2 x nominal Lumi

5 to 7.5
x nominal Lumi

30 fb^{-1}

190 fb^{-1}

500 fb^{-1}

integrated
luminosity

3000 fb^{-1}
4000 fb^{-1}

HL-LHC TECHNICAL EQUIPMENT:

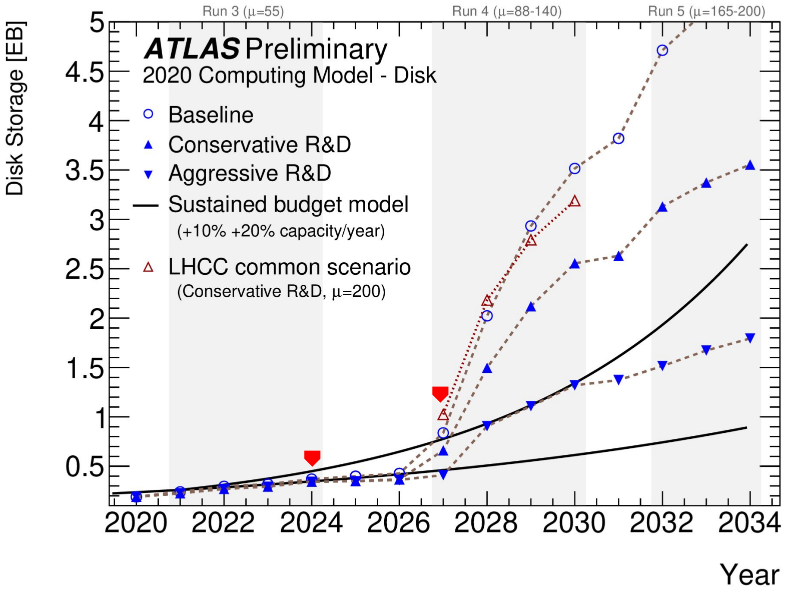
DESIGN STUDY | PROTOTYPES | CONSTRUCTION | INSTALLATION & COMM. | PHYSICS

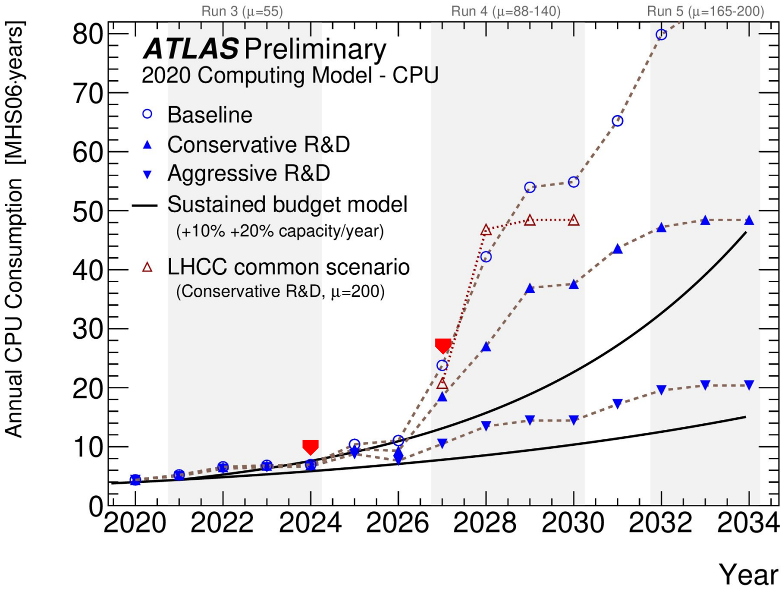
HL-LHC CIVIL ENGINEERING:

DEFINITION | EXCAVATION | BUILDINGS | CORES

Scale of ATLAS Computing

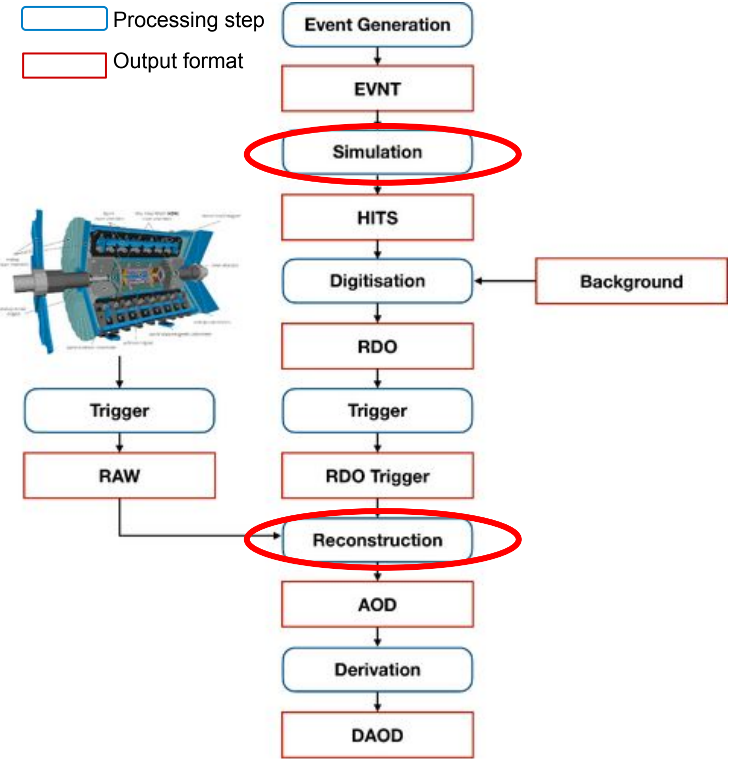
- 10x increase in data volume
- Greater event complexity
- Exabyte-scale storage





ATLAS data processing model

Software workflow



Processing step

Output format

Event Generation

EVNT

Simulation

HITS

Digitisation

Background

RDO

Trigger

RDO Trigger

Trigger

RAW

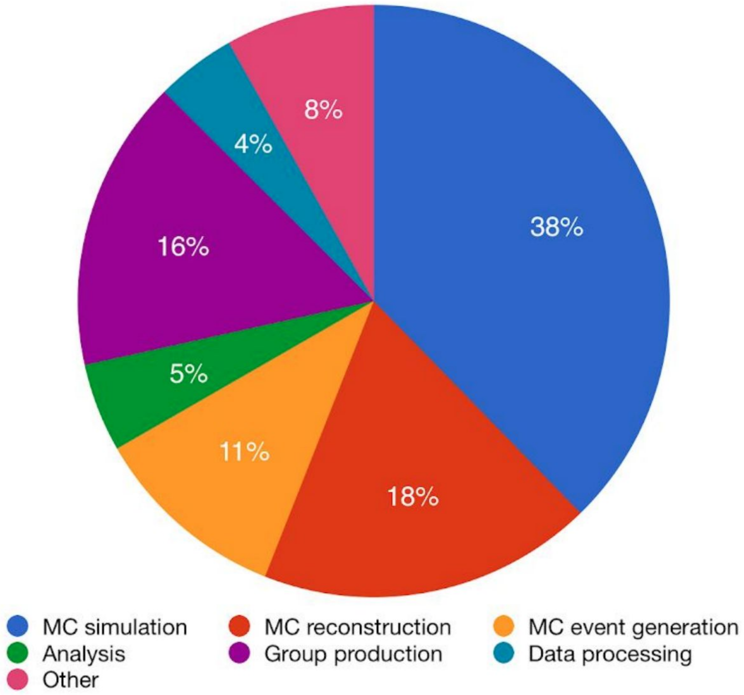
Reconstruction

AOD

Derivation

DAOD

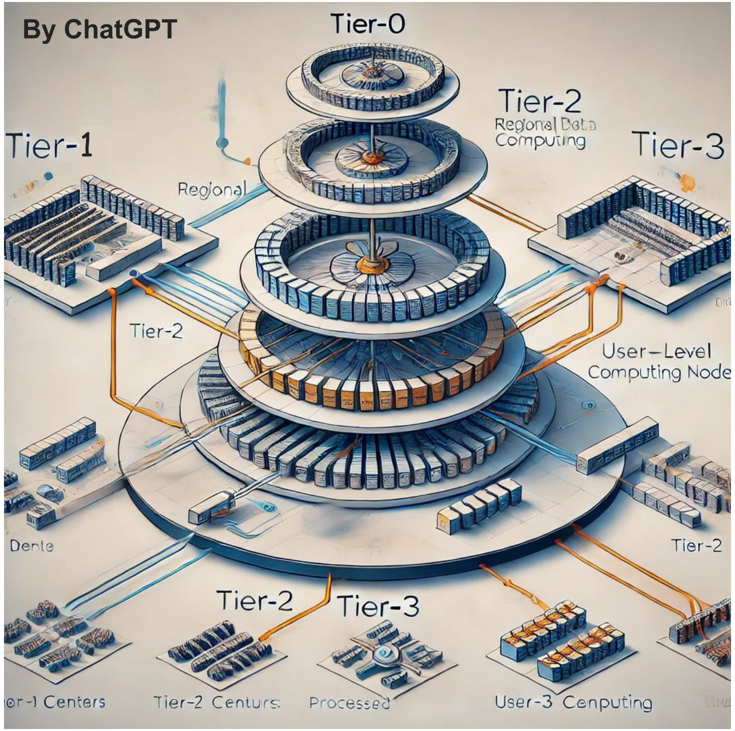
Wall clock consumption per workflow



ATLAS data processing model

Distributed Infrastructure

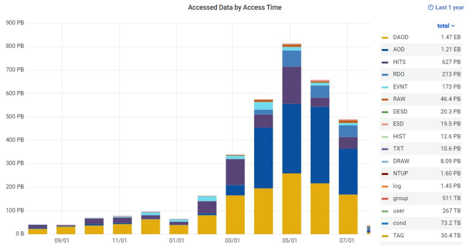
- **Tier-0 (CERN):**
  - First-pass processing of raw data
  - Distributes data to Tier-1 centers
- **Tier-1 Centers:**
  - Large-scale storage and reprocessing
  - Backup copies of critical data
- **Tier-2 Centers:**
  - Simulation and analysis tasks
- **Tier-3 Centers:**
  - Local user analysis and prototyping



By ChatGPT

Over 140 computing sites worldwide → processes 25 PB of data every week.

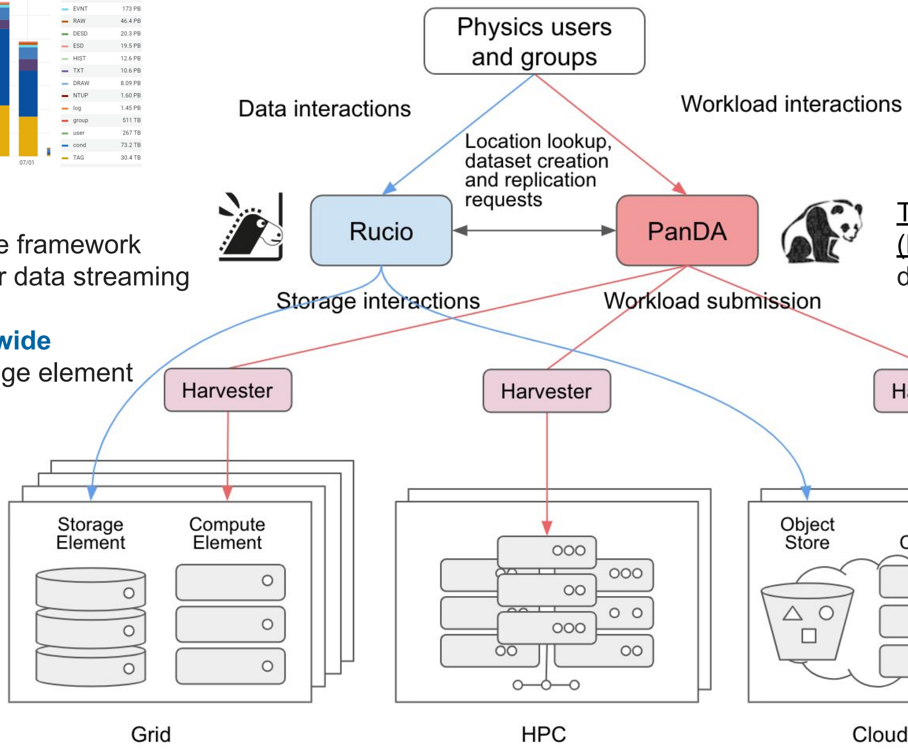
ATLAS Computing model



Rucio
Data management software framework
Based on xrootd service for data streaming

>130 storage sites worldwide
One site can have >1 storage element

- Production disk
- User disk
- Group disk



The Production and Distributed Analysis (PanDA)
data-driven workload management system

>140 compute sites worldwide
One site can have >1 compute element

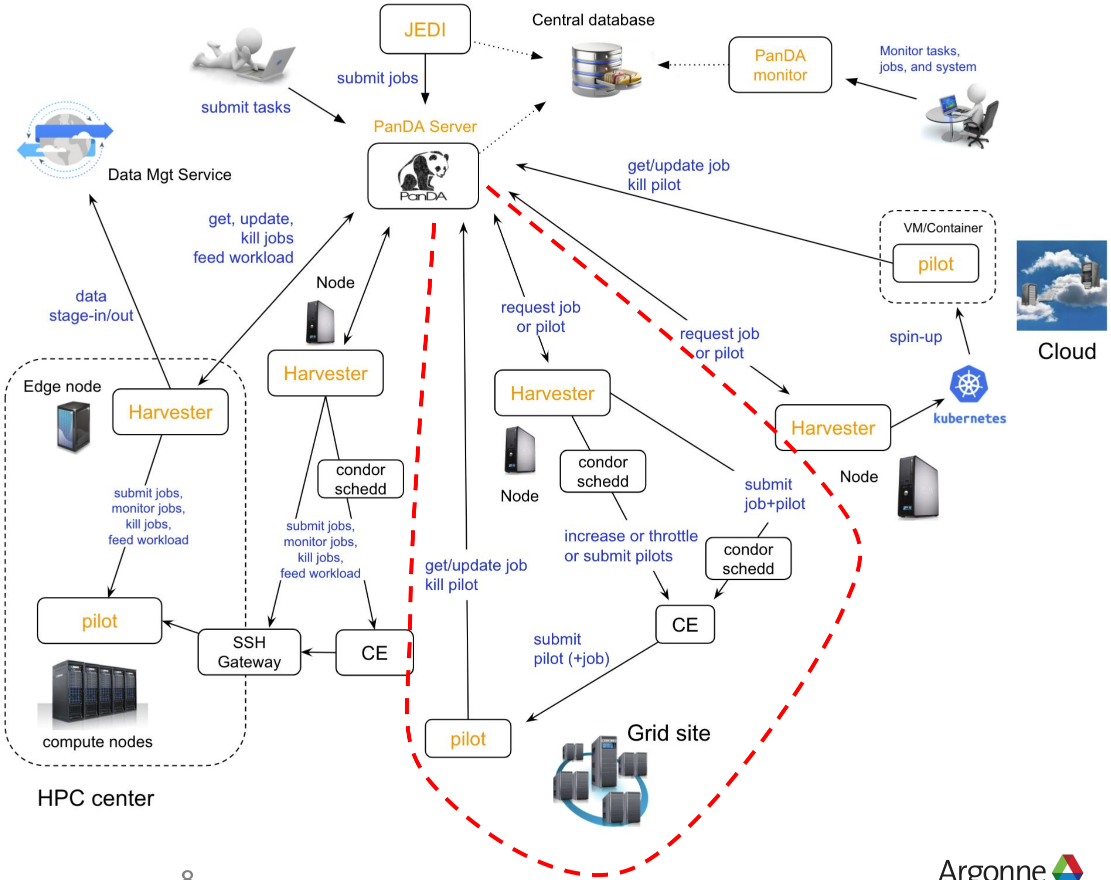
- Production
- Analysis
- Low memory (2GB/core)
- High memory (6GB/core)

PanDA WFMS

PanDA server + Harvester(&Pilot)

- **JEDI** – high-level engine to tailor workload for optimal usages of heterogeneous resources
- **<u>Harvester</u>** – a resource-facing service between WFMS and the collection of pilots for resource provisioning and workload shaping
- **Pilot** – a transient agent to execute a job on a worker node & reporting metrics

PanDA monitor – web-based monitoring of tasks and jobs + a common interface for end users, central operations team and remote site administrators



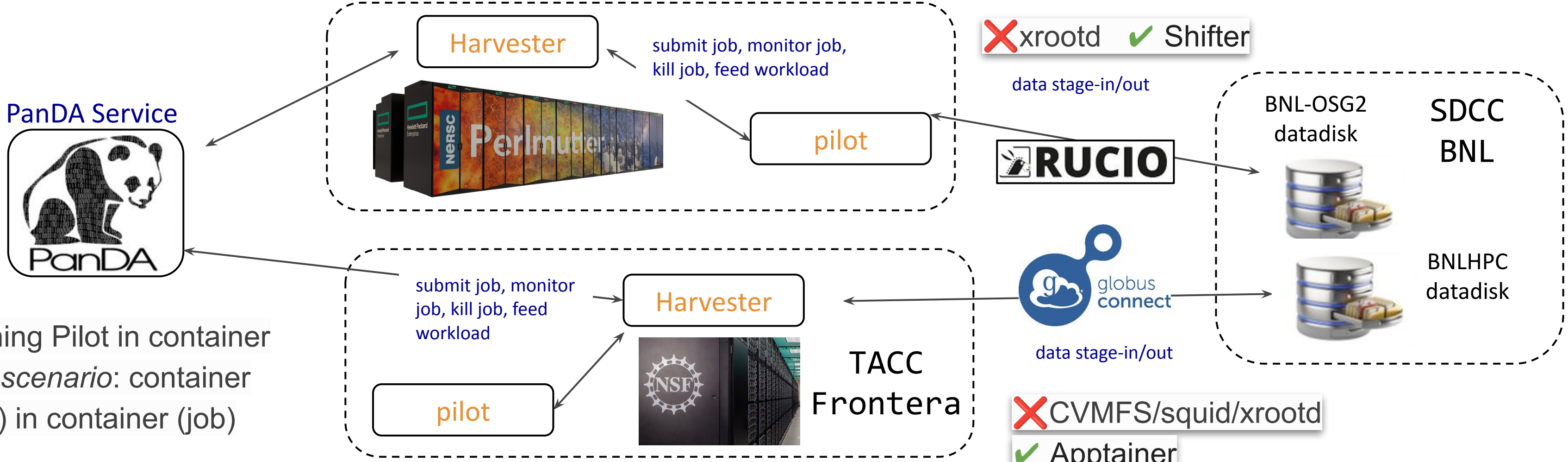
HPC model

European HPCs (Vega)

- Same environment as the Gird sites (Grid mode) -> All kinds of workflow
  - CVMFS (distributed software, config and condition data) + squid(Frontier) for caching
  - Xrootd service for data streaming

US HPCs (Perlmutter & Frontera @ TACC)

- Highly customized (each HPC treat separately) -> MC simulation initially + Event Generation (high mem)



❌xrootd ✔ Shifter

❌CVMFS/squid/xrootd

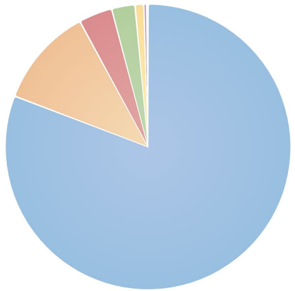
✔ Apptainer

Running Pilot in container

Best scenario: container (pilot) in container (job)

CPU jobs on US HPCs

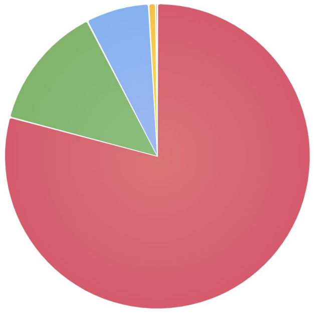
Submitted jobs



~40M jobs submitted in Nov. 2024

	Value	Percent
GRID	31.1 Mil	81%
hpc	4.31 Mil	11%
cloud_special	1.44 Mil	4%
cloud	989 K	3%
hpc_special	349 K	1%
Tier-0	95.3 K	0%
gpu	29.1 K	

Submitted jobs

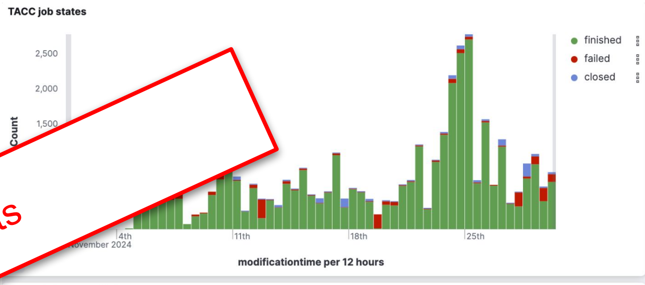


	Value	Percent
NERSC_Perlmutte[illegible]RE	277 K	79%
TACC-FRONTERA-UCORE	46.1 K	13%
TACC-FRONTERA-SCORE	23.2 K	7%
TACC-FRONTERA-TEST	2.31 K	1%

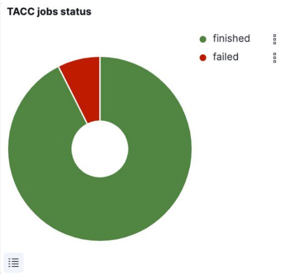
Internal plots

~100K SU used in 2024@TACC

TACC job states



TACC jobs status



TACC timing and efficiency

Export

jobstatus: Descending	Count	Queue time	execution	Total time	s
finished	36,918	11,081.58	2,421.665	4 hours	-
failed	2,969	8,543.823	35.661	5 hours	(
closed	1,104	2,670.164	0	45 minutes	(

TACC input-output

14.62TB input size	**40,908** input files
37.66TB output size	**39,085** output files

Job throughput

workload balance: flat is preferred

I/O limitation: job exceeds 200GB

Memory limitation: Sherpa Generation can up to >10GB/core

GPU & ARM

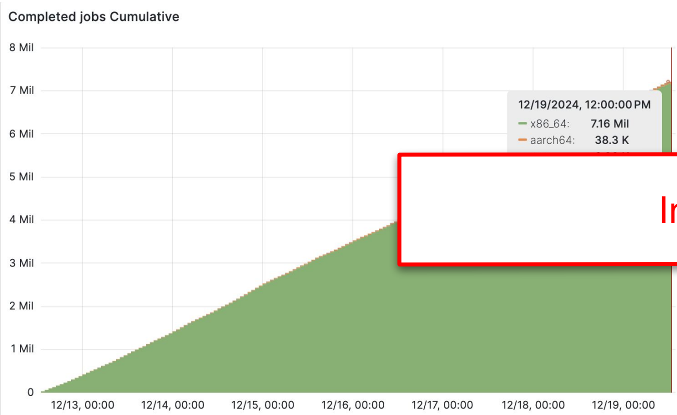
Simulation production on ARM

Software R&D to utilize GPU

- Geant4, celeritas, ACTS, etc.

Celeritas

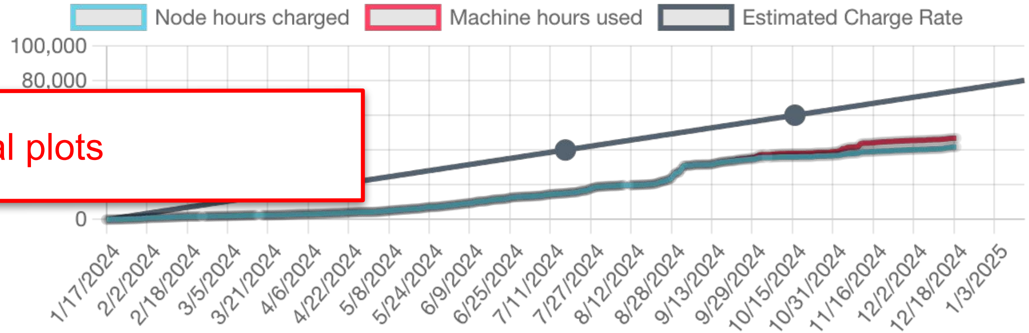
GPU-accelerated High Energy Physics Monte Carlo simulation



User's AI/ML job on GPU

- **HPC: users added to collaboration allocation**
  - Perlmutter is open for all ATLAS users, investigating Aurora
  - Integrating GPU queue via PanDA
- **Analysis Facility (Tier3)**
  - Fully environment & software support for ATLAS users

GPU usage on Perlmutter (ERCAP)



Internal plots

Summary

- ATLAS distributed computing model supports a wide range of resources
  - Capable for tens of PB data processing / week
  - HL-LHC demands significant scaling and R&D effort
- Moving Forward
  - Utilizing available HPC resources
    - Simplify user access and environment configuration under collaboration allocation
    - Remote management: Globus compute
  - Software R&D: GPU & ARM & FPGA
  - Improve the resource estimation and utilization (CPU/memory/IO/disk/etc.)
    - Workload balance
    - Reduce data write to the Tap
  - And more…!

HEP-CCE

More publications and plots can be found: https://atlas-swc.web.cern.ch/