ATLAS Computing Challenges Towards HL-LHC

Torre Wenaus (BNL)

US ATLAS / CSI Workshop

BNL

July 25 2018

BROOKHAVEN

CERN

CERN's Large Hadron Collider (LHC)

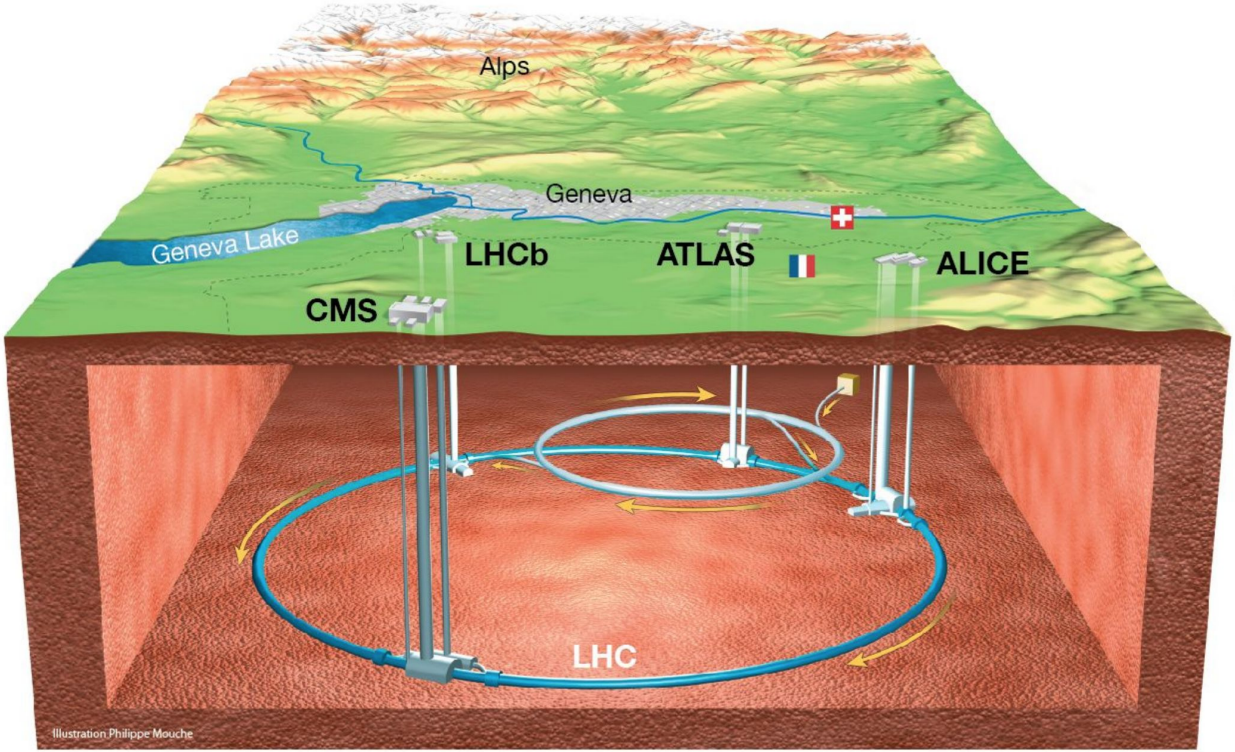
The 27km LHC tunnel has a history and a future:

1989-2000: LEP e+ e- collider up to energies of ~209 GeV, the 'Z factory'

2009-present: LHC p-p collider up to 14 TeV

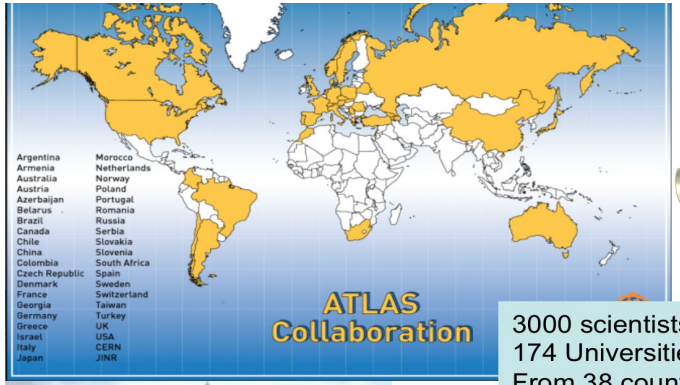
2026-2038+: High luminosity LHC (HL-LHC), civil construction just started, 10x integrated luminosity

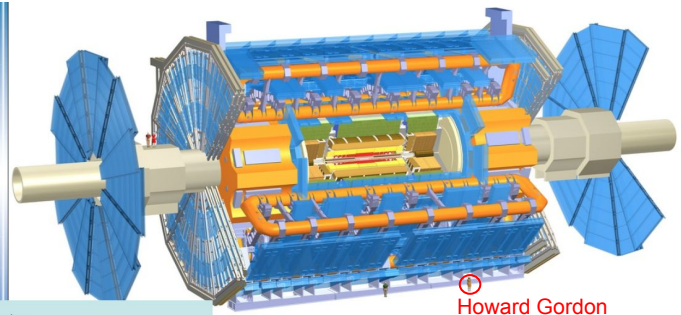
Beyond: High energy LHC up to ~27 TeV using new magnet technology, and ~4x luminosity increase



...and for the long term dreamers, a new tunnel: the Future Circular Collider, FCC, a 100km long tunnel extending under Lake Geneva

The ATLAS experiment at the LHC

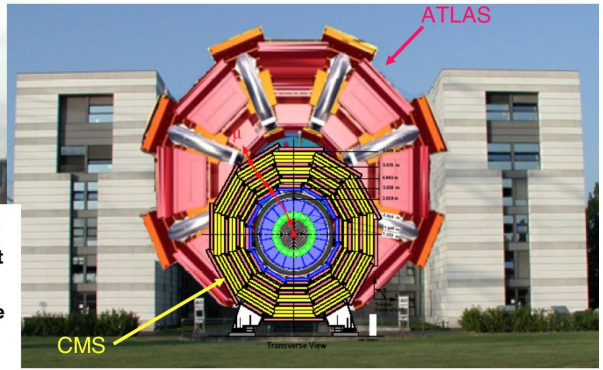


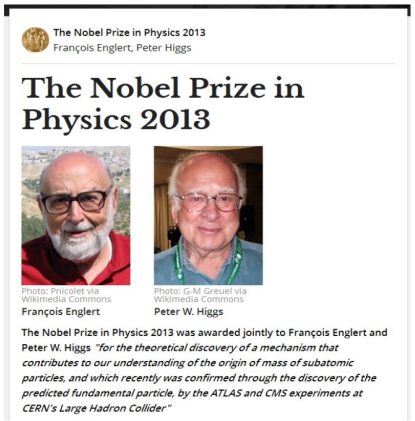


Howard Gordon

3000 scientists
174 Universities and Labs
From 38 countries
More than 1200 students

- **ATLAS has 44 meters long and 25 meters in diameter, weighs about 7,000 tons. It is about half as big as the Notre Dame Cathedral in Paris and weighs the same as the Eiffel Tower or a hundred 747 jets**





The Nobel Prize in Physics 2013
François Englert, Peter Higgs

The Nobel Prize in Physics 2013

Photo: Pricolet via Wikimedia Commons
François Englert

Photo: G-M Greuel via Wikimedia Commons
Peter W. Higgs

The Nobel Prize in Physics 2013 was awarded jointly to François Englert and Peter W. Higgs *"for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider"*

One of two (with CMS) general purpose detectors designed to reap the full discovery potential of proton-proton collisions to (at least) 14 TeV in the LHC

Friend of the Higgs boson and, so far, nemesis of supersymmetry

A discovery and precision physics experiment today and through at least the 2030s with the high luminosity HL-LHC upgrade in mid 2020s

The LHC data torrent

New physics rate ~ 0.00001 Hz

Event Selection :

1 in 10,000,000,000,000

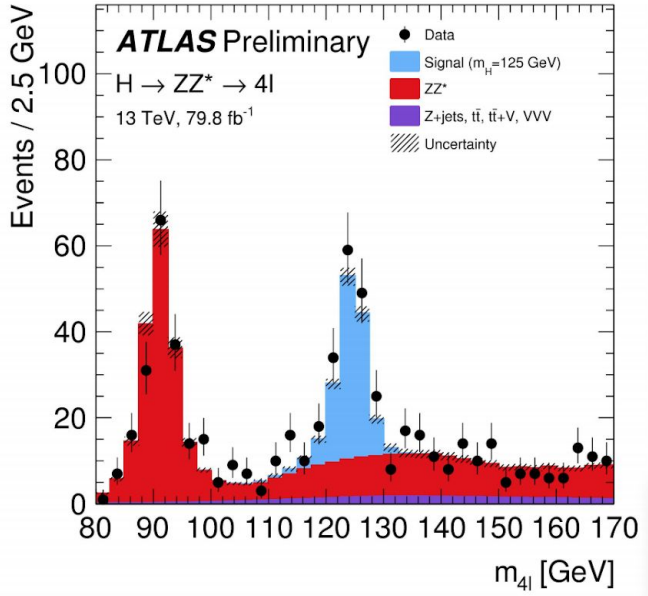
Like looking for a single drop of water from the Geneve Jet d'Eau over 2+ days

Pileup: proton interactions in the same bunch

ATLAS datataking today:
~1PB raw data/s off the detector filtered to 1-2 GB/s recorded (~1kHz)
Pileup: ~40

ATLAS at the HL-LHC:
~10x the event rate, larger and more complex events
Pileup up to 200

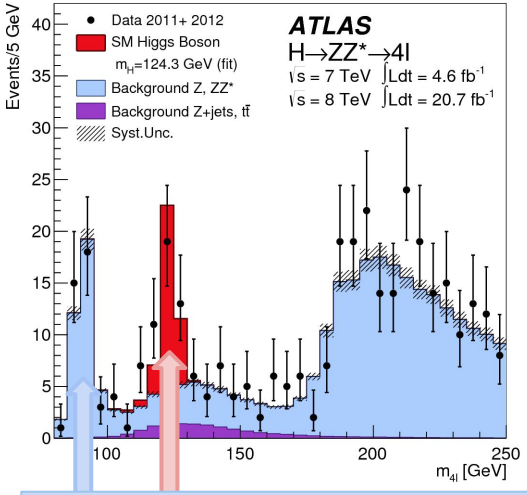
Computing at CERN and beyond

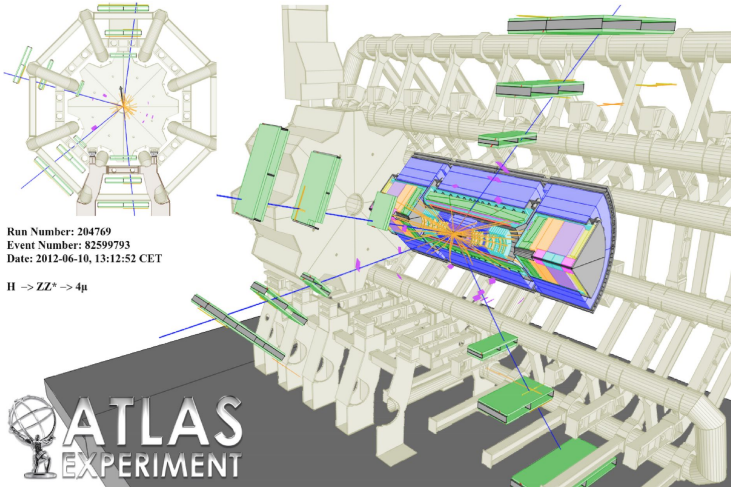


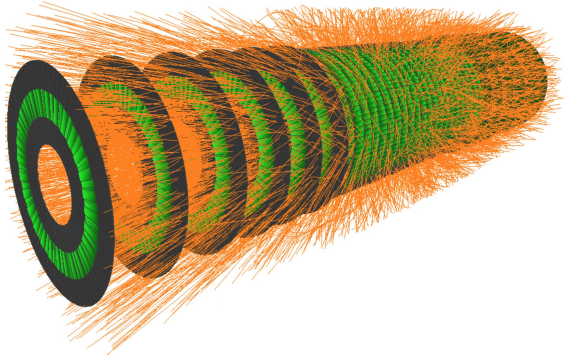
Inventing the web… global LHC computing hub… discovering the Higgs

From individual, local computing to harnessing global resources

Distilling physics from the torrent







ATLAS HL-LHC tracker event

My 1990 thesis took place here at the 91 GeV Z resonance: Z decay to two leptons at the CERN LEP e+e- collider. Clean rich signal, easy to distill. Yesterday's signal is today's background…

Higgs to ZZ to four leptons at 125 GeV: the proverbial droplets within the torrent, buried in the complex and far from clean environment of proton-proton collider physics at the LHC

A Z resonance study in 1990 at LEP: done on one HP Apollo workstation. Entire ~10yr LEP dataset: O(TB)

Physics at the LHC:

Requires the largest distributed data intensive scientific computing infrastructure ever built

Worldwide LHC Computing Grid

170 Data centres

40 Countries

800'000 Cores

500 PB Disk

400 PB Tape

3 Tbps Network

Tiered Structure

Tier-0	CERN
Tier-1	Large data centres
Tier-2+3	Universities and Laboratories

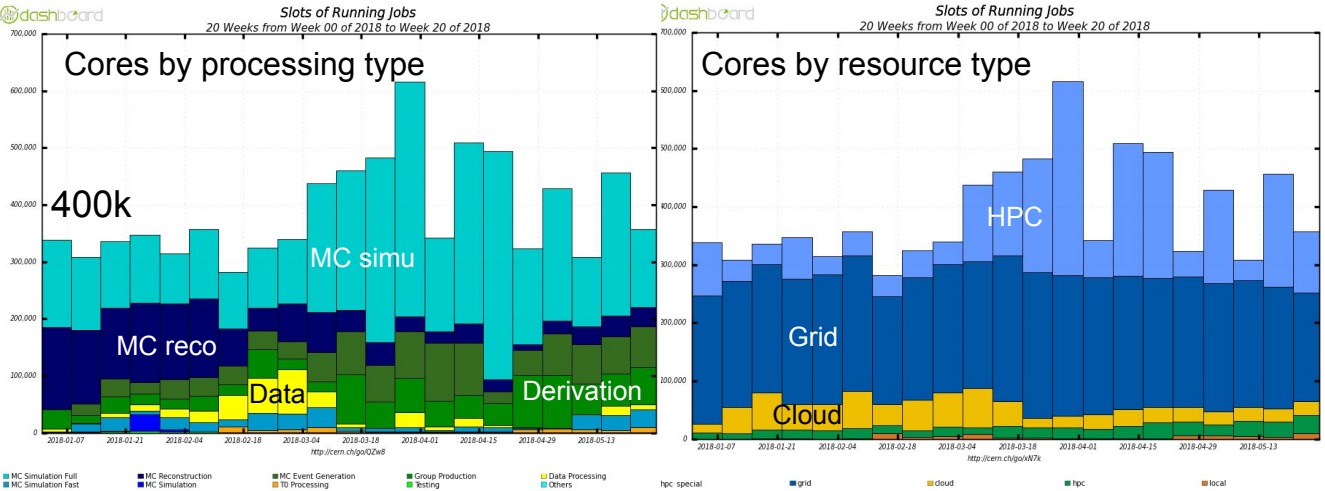
Heterogeneous Computing

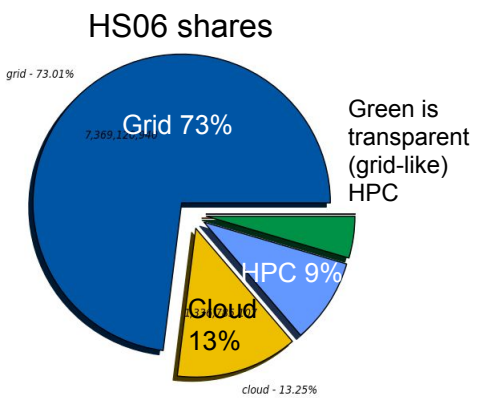
Data centres publicly funded by their countries

Host and support many different experiments

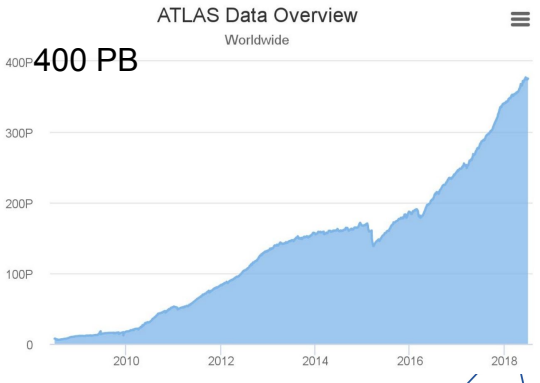
Pledge storage and compute based on an MoU

ATLAS processing today

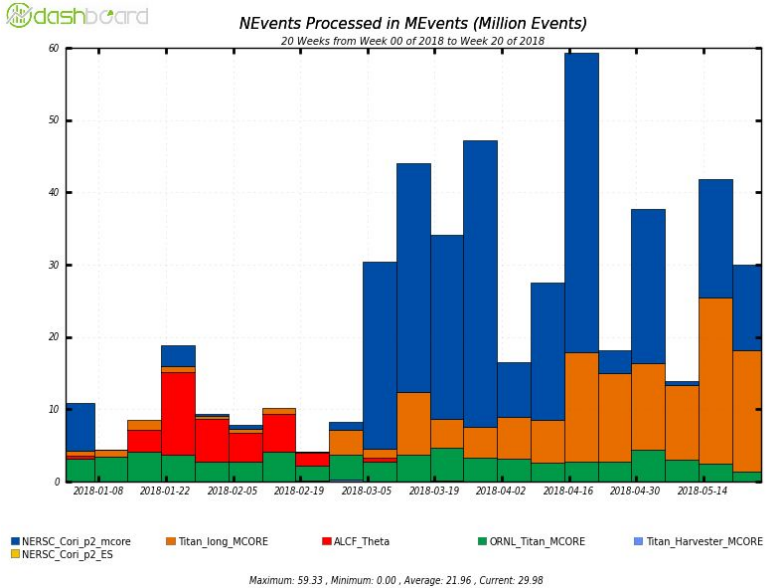




- Routinely running 24x7 on ~300-350k cores across the LHC grid
- A range of workflows: Monte Carlo simulation, reconstruction of MC and real data, derivation of compact analysis formats
- A range of resources: Grids, HPCs, clouds
- HPCs drive peaks above 1M cores (cores 5-10x weaker than grid)
- Moving >1 PB, >20 GB/s, 1.5-2M files per day, ~400PB total data set



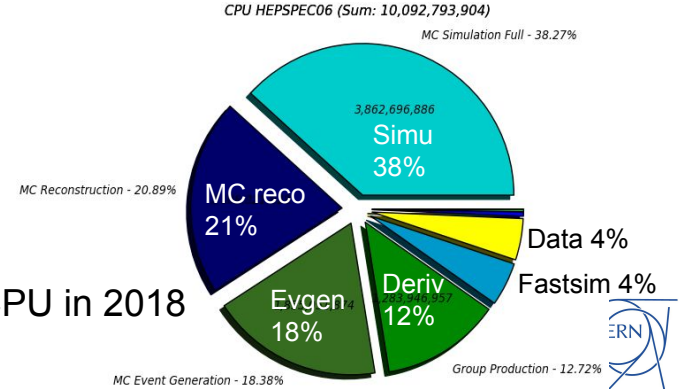
HPCs increasingly important



- Event production in the 'hpc_special' (complex HPC) category, requiring custom infrastructure (not grid-like)
  - NERSC Cori at LBNL
  - Titan at Oak Ridge
  - Theta at Argonne
- Powerful machines but highly non-transparent to use, requiring dedicated expertise, effort, software tools, porting…
- Commissioning a new resource provisioning component, Harvester, to bring greater uniformity across resources

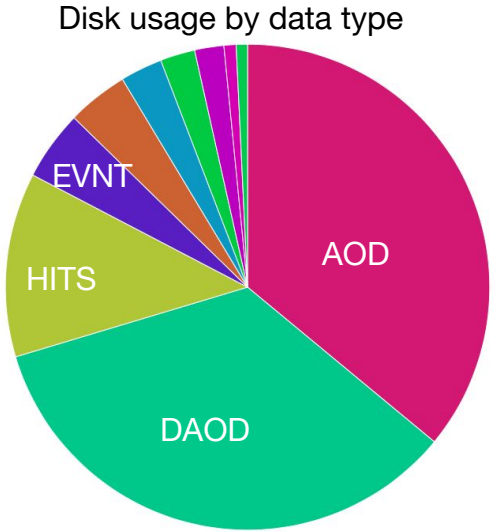
These HPCs used today solely for MC simulation, the largest component of our CPU usage

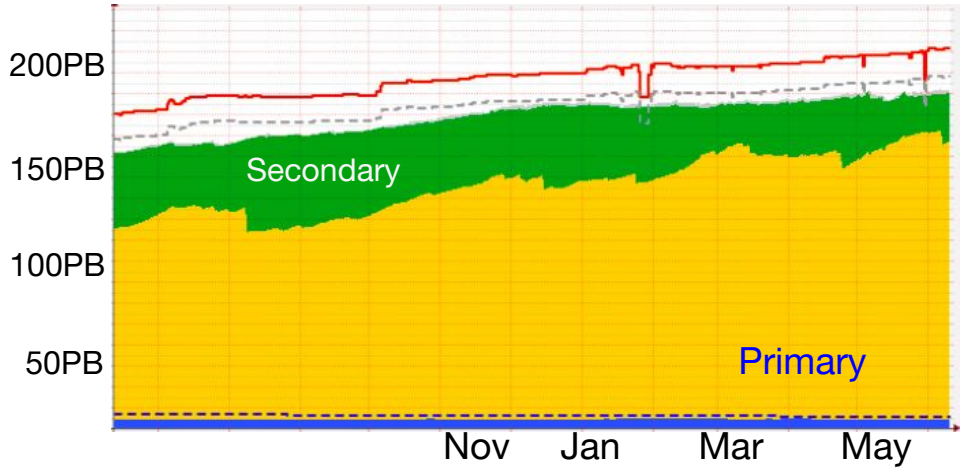
CPU in 2018



ATLAS data storage in 2018

- ~200 PB of disk perpetually ‘full’
  - Dotted grey line is operational maximum, red line is actual space
- Primary data: pinned, not deletable until unpinned
- Secondary data: deletable as necessary
- Lifetime model and obsolescence rules drive deletion (with exceptions possible)
- Largest use of disk: analysis formats (AOD, DAOD)





The future of ATLAS computing

Today:

- ATLAS is compute-limited in its science
  - CPU: ATLAS aggressively harvests all the cycles it can gather from grids, HPCs, clouds, clusters, volunteer computing
    - Leveraging sophisticated workflow management systems that make operating on diverse heterogeneous resources practical
  - Storage: ATLAS keeps 100% of disk storage in use with careful active management to remove data promptly when no longer of prime interest

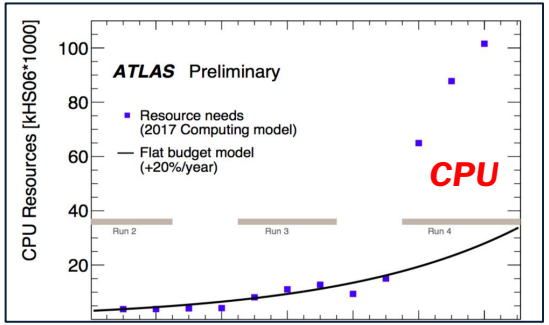
Tomorrow:

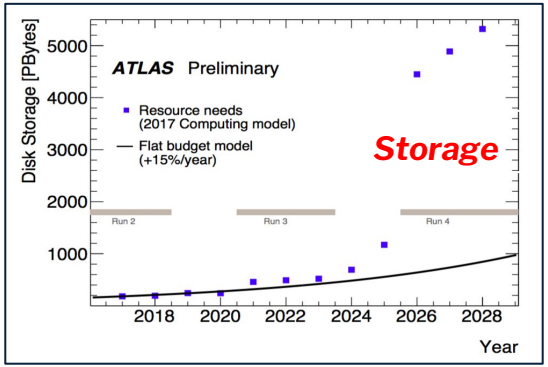
- It gets much worse!

S&C challenges for the HL-LHC

- 5-10x the event rate: much more data, and more complex events
- Pile-up of ~200 proton-proton interactions per beam crossing
  - Conventional tracking approaches lead to a combinatorial explosion
- Flat-budget hardware improvements of ~6x fall far short of requirements
  - Leaves us a factor of ~3 short in CPU and ~6 short in storage
- Far more data than we can pay for using conventional storage and data management approaches
  - Must evolve our computing and data management approaches
  - Feed our applications with data efficiently at scale in the distributed environment
  - I/O bound workloads make this particularly challenging
- HEP software typically executes 1 instruction at a time per thread
  - Major re-engineering beyond multithreading required to maximize benefit from modern CPUs (vectorization, pipelining)
  - Accelerators like GPUs are even more challenging
- **We cannot afford to buy our way out of the problem with hardware -- we must innovate in software**

HL-LHC computing needs far exceed flat budget growth
CPU: ~3x Storage: ~6x





Strategy towards HL-LHC computing

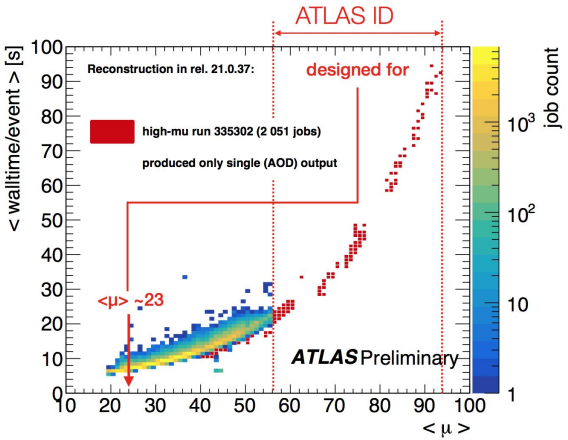
- Five main themes:

WLCG Strategy towards HL-LHC

  - Software performance
    - Performance analysis and reengineering, data layout & I/O, compatibility with accelerators, efficient use of memory
  - Algorithmic improvements and changes
    - Generators, fast simulation, reconstruction, leveraging ML
  - Reduction of data volumes, particularly disk
    - Compression, slimming, greater use of cold storage (tape)
  - Managing operations cost
    - Greater automation and system intelligence
    - Greater use of organized production (already dominant in ATLAS)
  - Optimizing hardware costs
    - The 20% performance gain/yr at flat cost in our present estimates may be an over-estimate based on recent market survey
    - Developing detailed cost model for quantitative assessment of solutions, alternatives
    - Leveraging resources paid for by others, most notably HPCs

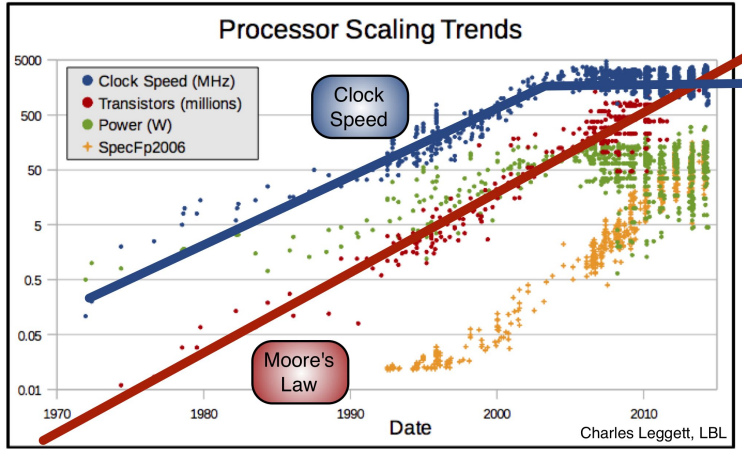
Scaling to HL-LHC: CPU

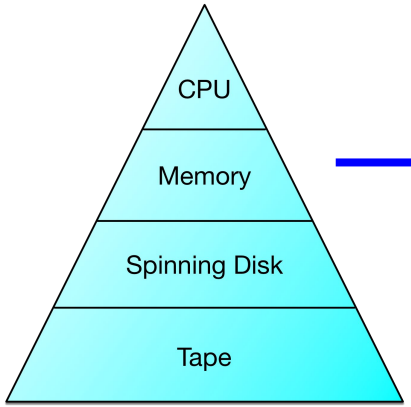
- CPU shortfall has shrunk in last 18mo thanks to tracking speedups; current estimate is ~3x (down from ~6x)
- Ongoing work will bring further improvements
  - Performance optimization
  - Further reconstruction optimizations
  - Use of truth info in MC track reconstruction
  - Pre-mixing of pileup events, reducing digitization time
  - More fast simulation
- Expect **substantial use of HPCs**, with *data intensive* requirements from our use cases
  - ATLAS apps are good HPC citizens: fully parallel MPI applications, keeping every core busy for the full job duration
  - Demands reengineering sufficient to utilize them efficiently, *including their accelerators*



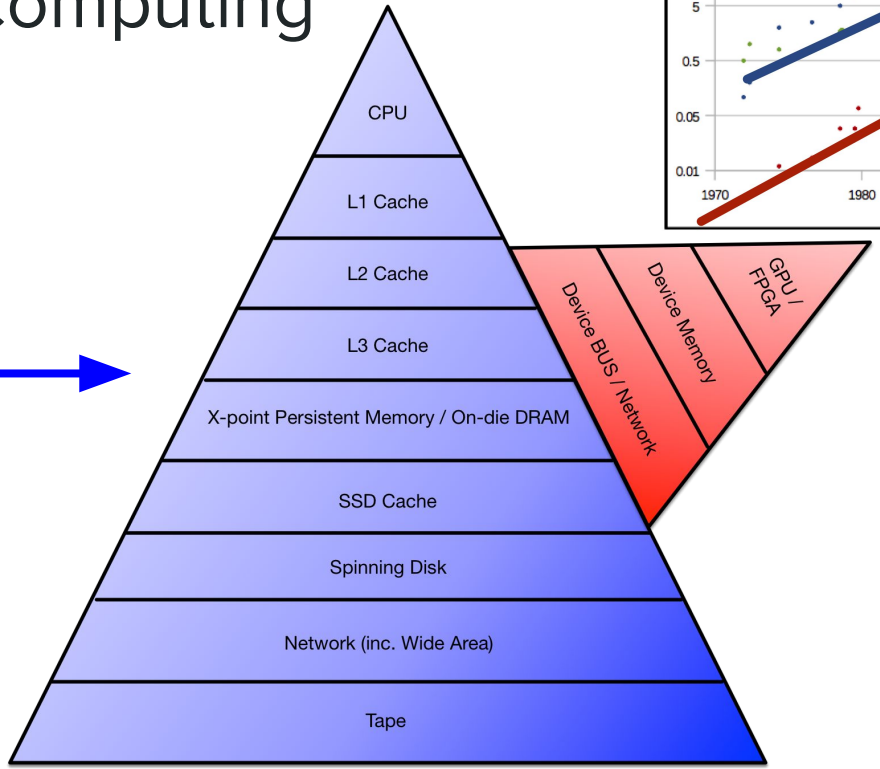
ATLAS reconstruction time dependence on pileup (mu), as measured in data, showing exponential rise due to inner detector (ID) tracking combinatorics

Shifting landscape for end-to-end computing





The Good Old Days



The Brave New World

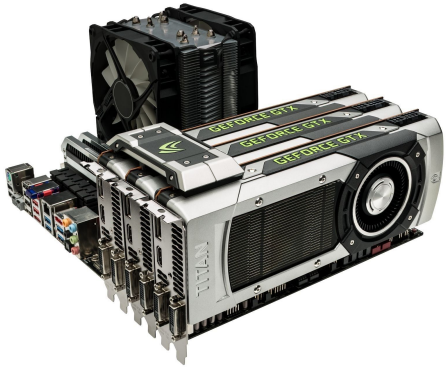
Graeme Stewart, CERN

Software performance - experience so far

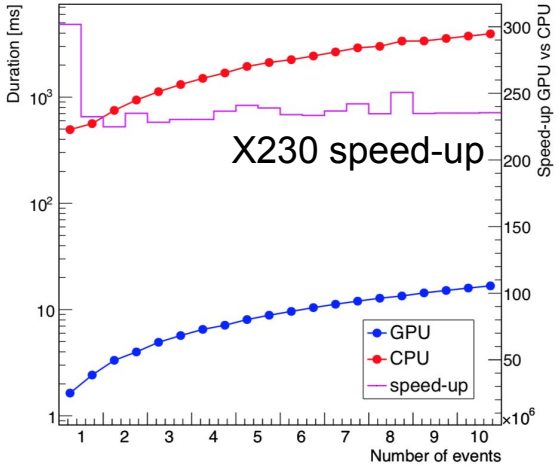
- Compiler/code/build based improvements
  - Investigating (typically unused) advanced compiler features
  - Reducing shared library overheads by building large libs instead of loading many small ones
  - AutoFDO feedback guided optimization
  - Looked at Geant4, reconstruction code, NLO generators (the newer slower ones)
  - Several 15-20% CPU improvements (not all can be combined)
    - HEP code very sensitive to compiler switches; requires substantial physics validation
- Code profiling shows up to 25% of time spent in memory management
  - Allocation and de-allocation of small, very short lived objects
    - 10% can be saved with refactorization
- Hardware counter based analysis: HEP code ~1 op/cycle, HPC code ~4, vector instructions up to 8
  - 100% improvement should be feasible, leveraging vector units, with considerable work: substantial code changes, high level skills
  - Overlap with GPU utilization work
  - ALICE HLT tracking code re-design for GPUs also produced huge gains on CPUs with vector units
- Overall estimate of potential sw performance gains without changing paradigms & algorithms: 200%
  - 50% at moderate cost

from Markus Schulz and the cost model working group

Accelerated Computing



- GPUs superb at delivering floating point operations
  - Often x10-20 higher than CPUs
  - But difficult to program against in many cases
    - Don't deal well with branchy code
  - GPGPU cards not cheap, not easy to measure efficiency of use
- Excel at *training* deep learning neural networks
- Data ingestion can be limiting factor for other uses
  - Need sufficient calculation (i.e. a lot) to amortize the cost of data ingestion
- Some cases where they can help analysis a lot
  - Goofit and Hydra minimiser
  - Typical case: analysis with large numbers of toy models varying parameters to understand systematics
- *How can we put them to greater use?*



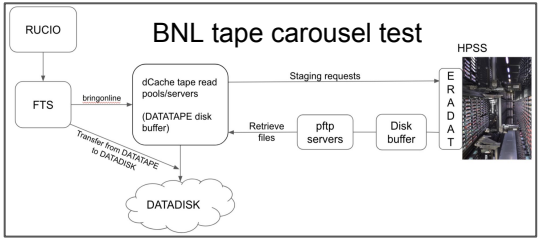
Phase space generator speed-up with Hydra

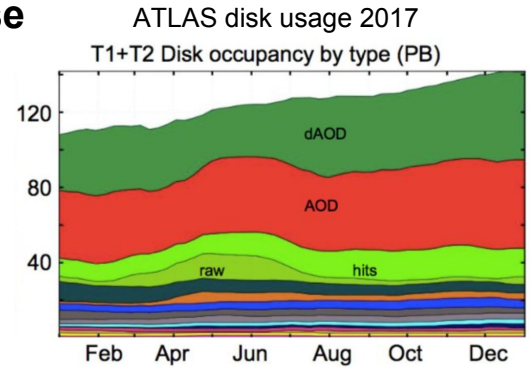
Machine learning in HEP

- Wide and growing use in HEP and ATLAS, albeit less so in production code so far
  - You'll hear about examples in ATLAS
- Applications coincide with computing domains that can give a large payoff in software reengineering for better resource utilization and leveraging HPC/exascale
  - e.g. fast simulation
- Training is computationally demanding, but how much computing will training consume?
- Will the memory demands of NNs sufficiently deep to be effective tax our memory constraints?
- Promising avenue to utilize accelerators?

Scaling to HL-LHC: Storage

- ~6x shortfall by today's estimate, a level that has held ~steady
- 'Opportunistic storage' basically doesn't exist
- Working on format size reductions, but hard to achieve large gains
  - 30% reduction is target of newly empaneled study group
- Replica counts already squeezed, hard to achieve large gains
- **Storage shortfall is our biggest problem. We need new approaches!**
- ATLAS disk usage is currently ⅔ analysis formats and ⅓ everything else
- Within the '⅓ everything else' are samples that reside mostly on tape, staged onto disk cache when needed for processing
- A way to **dramatically reduce our storage footprint** is to **grow the use of tape** (it looks like our cheap storage will remain tape)
  - Use a '**tape carousel**' approach for the analysis formats
  - A moving window of say ~10% staged to disk at any one time
  - BNL has longstanding experience in this, thanks to PHENIX
- **This is hard:** tape is slow and complicates workflow orchestration
  - Analysis workflows are time critical and already complex
- Tape is geographically limited, while processing happens everywhere
- **Fertile ground for R&D...**

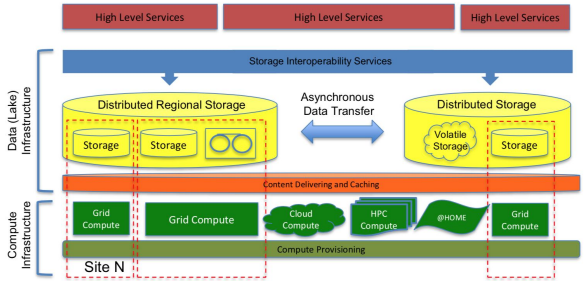




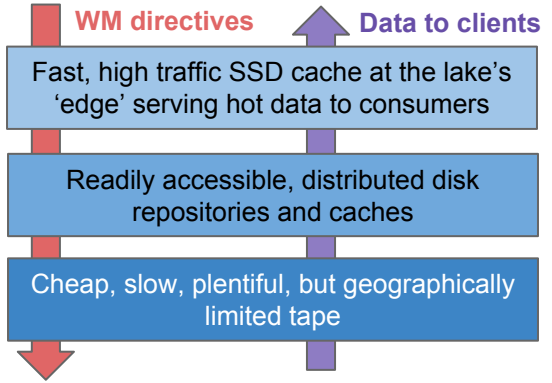
Data lakes and workload management

- Our sites are linked with (ever higher) high-bandwidth networking
  - We can expect **~100x bandwidth growth** by HL-LHC
- **Data lakes**: integrated consolidation of distributed storage (and compute) facilities, leveraging high-bandwidth networks
- Data lake encompasses facilities with several levels of storage
  - **Tape**, at a relatively limited number of sites
  - **Standard disk**, at large storage repositories and smaller caches
  - Fast SSD '**edge cache**' for the hottest data
  - Should be able to **place data optimally** based on (dynamic) need
- Workload management knows the hot popular data in use
  - Use that knowledge to drive preparing data in the lake, asynchronously to the processing, e.g.
    - tape staging in a **carousel workflow**
    - placing hot data in SSD cache **'close' to available CPU**
    - **transforming/marshaling data** optimally for client delivery
  - Requires APIs supporting WM directives
- **Instead of 1.8 replicas on disk today, WM + data lake manages dynamic availability of actively used data with replica count <<1**

Data lake schematic



Data lake interactions



ATLAS computing upgrades

2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4
Run 1: 7-8 TeV, 0.7×10^{34} (μ≈20), 25 fb^{-1}			**LS1**			**Run 2:** 13 TeV, 1.9×10^{34} (μ≈55), 150 fb^{-1}			**LS2**

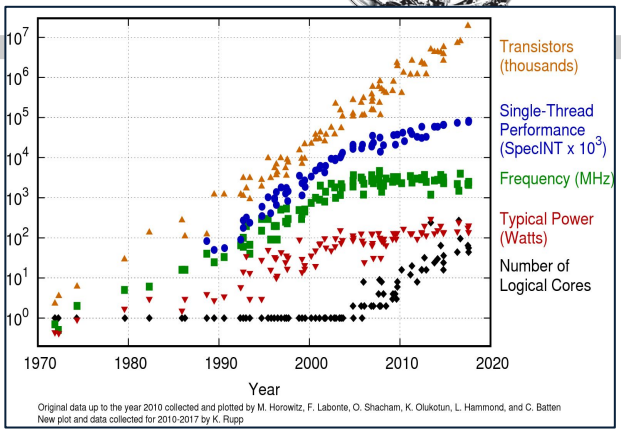
2020	2021	2022	2023	2024	2025	2026	2027	2028	2029
Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4
LS2	**Run 3:** 14 TeV, 3×10^{34} (μ≈80), 300 fb^{-1}			**LS3**			**Run 4:** 14 TeV, 7.5×10^{34} (μ≈200), 3000 fb^{-1}		

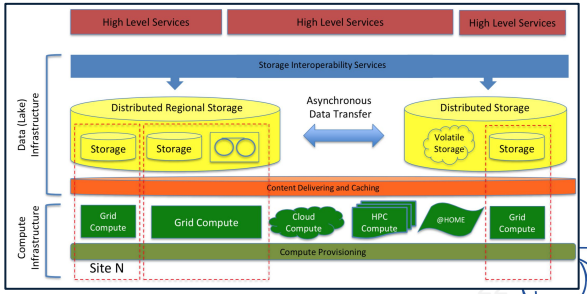
Upgrade	Shutdown	LHC Luminosity Target	Main ATLAS S&C Changes
Phase-I	2019-21 ***LS2***	2-3 × design*	• **Multithreaded framework AthenaMT in production** • **Updated analysis model: study group active** • **New forward muon detectors** • **Full integration of upgrade reco into main software** • **Expanded role for fast simulation** • **First exascale HPC experience**
HL-LHC (Phase-II)	2024-26 ***LS3***	5-7.5 × design*	• **New tracking software capable of pileup 200** • **Further expanded role for fast simulation** • **Wholesale integration of ML, GPUs, … toBeInvented** • **Completely reworked data + workflow management** • **Fast HPC-optimized generators**

* Original Luminosity Target = 1 x 10^{34} cm^{-2}s^{-1}

Community collaboration

- Increasingly important agent for collaboration is the HEP Software Foundation (HSF)
  - Created in 2015 to facilitate cooperation and common efforts in HEP S&C internationally, particularly *concurrency & reengineering*
  - HSF led the development of the HEP S&C Roadmap (end 2017)
  - HSF will lead the common software efforts towards HL-LHC, in tandem with WLCG for the computing aspects
- ATLAS believes in open source: ATLAS offline software becoming public
  - Two year process to turn Athena into open source community software, facilitating common projects and supporting sw careers
  - Technically ready and will open it after Run-2
  - Pre-releasing it now for key collaborations
    - e.g. with *BNL CSI for software reengineering studies*
- ATLAS is a lead collaborator on the most prominent HL-LHC community R&D effort to date, DOMA (Data Organization, Management and Access)
  - Coordinated by WLCG, following the WLCG Strategy towards HL-LHC
  - Integrates data management and associated workflow management R&D





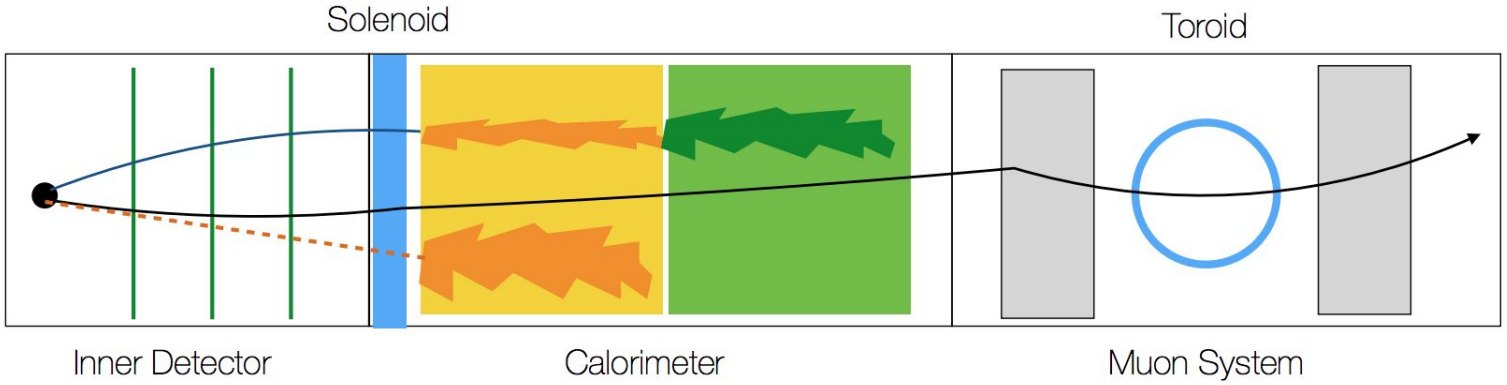
HEP Software Foundation Roadmap for Software and Computing R&D in the 2020s

HSF-CWP-2017-01
December 15, 2017

- HSF established in 2015 to facilitate *coordination* and *common efforts* in software and computing across HEP in general
- Charged by WLCG to address R&D for the next decade
- 70 page document on arXiv (1712.06982)
- **13 topical sections** summarising R&D in a variety of technical areas for HEP Software and Computing
  - Backed by topical papers with more details also (e.g. 50-page detailed review about Detector Simulation)
- **1 section on Training and Careers**
- **310 authors** (signers) from 124 HEP-related institutions
- Feature article in CERN Courier
- More details on the HSF web site

Contents

Onward



Onward to hearing about our primary software domains and the prospects and challenges they present in readying ourselves for new processor generations, exascale and HL-LHC

Finally

- This timely workshop coincides (not by accident) with growing attention to understanding how we can meet the S&C challenges of HL-LHC
  - HEP S&C roadmap and WLCG strategy document for HL-LHC laid out a rough path and identified R&D areas
    - R&D now underway and starting to flesh out ideas and plans
  - US DOE has highlighted the importance of experimental HEP utilization of HPCs and, very soon, exascale
    - ATLAS is a HEP leader in leveraging HPCs, efficient in CPU utilization but not coprocessors, critical for exascale
  - ML is on a steep growth path (the key DNN paper that started the boom came the same year as the Higgs discovery!), to what extent is it a game changer and enabler?
- What should we target first?
- Plenty to discuss this week with the benefit of CSI expertise

Acknowledgements

Many thanks to Michael Begel, Ian Bird, Simone Campana, Kaushik De, Alexei Klimentov, Eric Lancon, Mario Lassnig, Andreas Salzburger, Markus Schulz, Graeme Stewart,

and the many in the ATLAS, WLCG, and HSF communities building HEP software & computing and its future.