



US ATLAS Perspective on HL-LHC Software and Computing

Kaushik De, Paolo Calafiura

IRIS-HEP Kickoff Workshop, Chicago

October 31, 2018



Introduction

- ❖ The US ATLAS Software & Computing program supports HL-LHC R&D in parallel with continuing support for pre-HL-LHC data collection/analysis
- ❖ Going forward - US ATLAS partnership with IRIS-HEP needs to be a crucial cornerstone of a successful physics program at the HL-LHC
- ❖ Recent talks with more details by Torre Wenaus and Paolo Calafiura are in appendix, for reference.
- ❖ The rest of the talk is not about IRIS-HEP - we have the next few days for that - but about the US ATLAS program and its many ongoing and planned activities



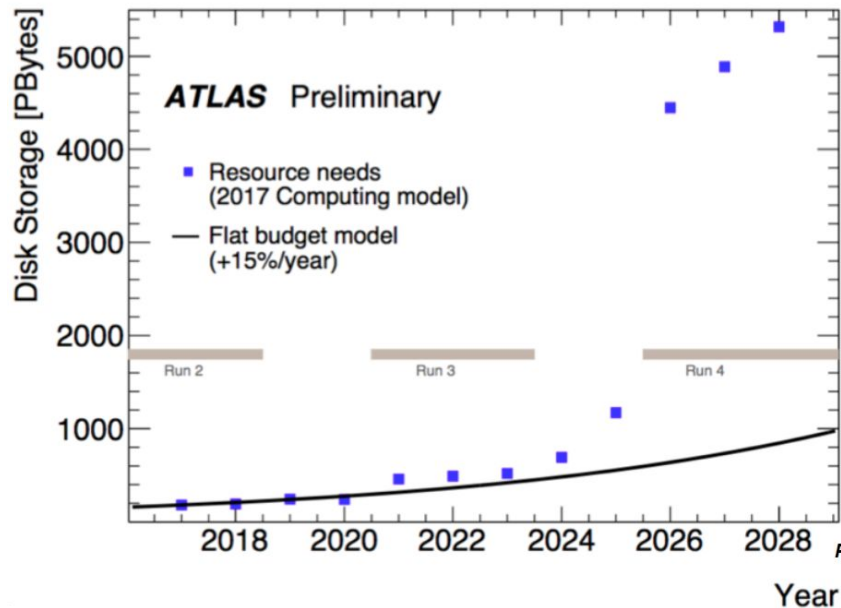
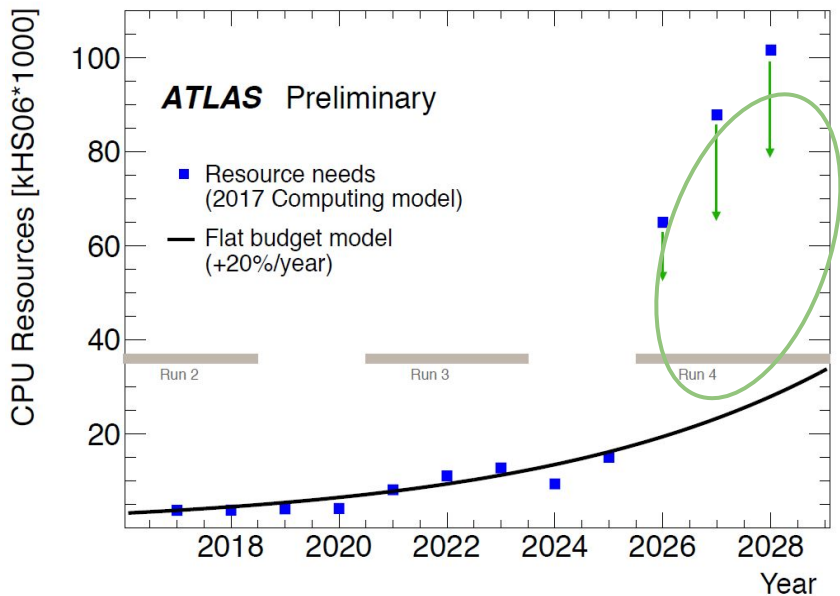
HL-LHC Computing R&D

Torre Wenaus (BNL)

US ATLAS Institutional Board Open Meeting
October 24, 2018

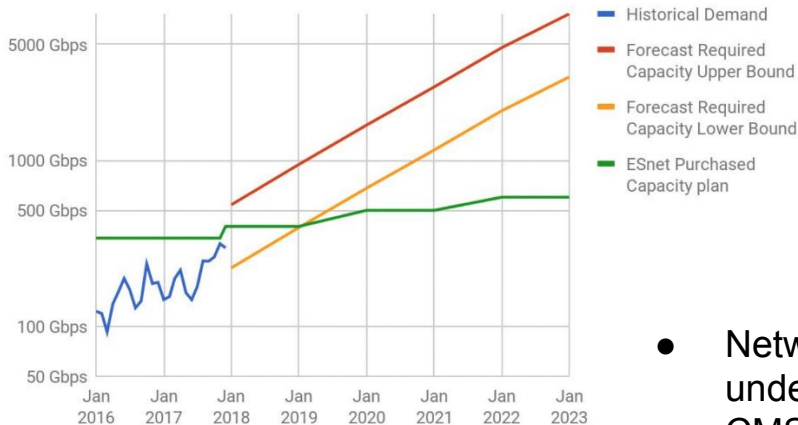


HL-LHC Challenges - Well Known



- ❖ x6 Event reconstruction time (x3 pileup)
- ❖ Dominated by Tracking

European Demand and Capacity Forecasts



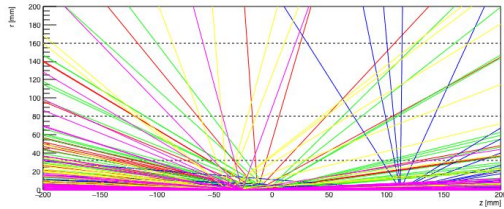
- ❖ x10 Storage (x10 Statistics)
- ❖ Dominated by Analysis data (AODs)

- Network growth projections are under joint studies - ESNET, CMS, ATLAS

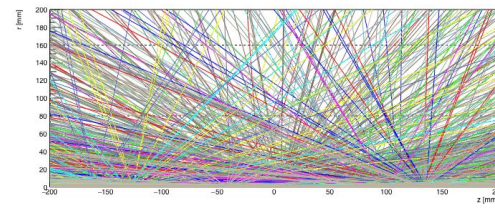


HL-LHC Challenges - Less Obvious

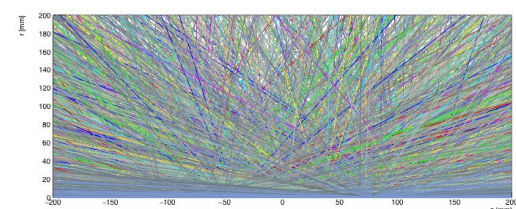
- ❖ While resource shortage (leading to funding shortage) for HL-LHC dominate conversations, other challenges await
- ❖ Event complexity will require new/smarter algorithms



2010, $\langle\mu\rangle=5$



2018, $\langle\mu\rangle=40$



2026, $\langle\mu\rangle=200$

- ❖ New discoveries/searches will also drive algorithms
 - Adoption of new Machine/Deep Learning techniques
 - New hardware - GPUs, FPGA, ARM, Quantum Computing
- ❖ New data formats, data access, data distribution
 - Exploration of other/non-Posix/new technologies



HL-LHC Challenges - the Unknown

- ❖ **Market forces**
 - Evolution of hardware, software and networking
 - We are not into predictions - but need to test/integrate new ideas
- ❖ **New computing architectures**
 - We already see opportunities - GPU/TPUs, FPGA farms, ARM
- ❖ **New software products**
 - Mostly opportunities, not restrictions



The Analysis Challenge

Analysis applications and data use roughly $\frac{1}{3}$ of US resources

- ❖ Developed/optimized mainly by the physics community

US ATLAS Ops provides physics groups with

- ❖ Analysis Centers
 - Peer-to-peer collaboration, connect big and small groups
 - Develop/gather/document best practices, train newcomers
- ❖ Analysis Facilities (aka shared T3) integrated with ATLAS grid
 - Support for distributed analysis workflows (e.g. “Trains”)
 - Support for new analysis platforms (e.g. JupyterLab)
- ❖ **IRIS-HEP Analysis Systems as “HL-LHC Analysis Center”**
 - Develop, optimize, disseminate HL-LHC Analysis Model(s)
 - Physicists occasionally listen to their peers



US ATLAS HL-LHC Strategies

- ❖ **US ATLAS HL-LHC R&D is moving along many fronts**
 - Setting up US ATLAS management structure and working teams
 - Mapping out the “space” for technical solutions
 - Building collaborations with our partners
 - In the next few slides we will expand on each of these three fronts
- ❖ **Management structure**
 - We added new high level US ATLAS WBS for HL-LHC
- ❖ **Build the team**
 - Make a list of missing people (while list is dominated by immediate needs, they are representative of the strategic direction for HL-LHC)
- ❖ **We need to focus on long term goals**
 - While we test our solutions during Run 3, over the next 5 years, we must keep focus on HL-LHC to be ready to meet its challenges
 - To find common projects between experiments & with CS experts, do not start from highest priority goals of today - those do not lend themselves well to common software development - think long term



HL-LHC Computing and Software

New US ATLAS WBS led by Heather Gray and Torre Wenaus

| | | |
|--------------|--|--|
| 2.4.1 | Software reengineering and algorithm development | Exascale application performance, accelerators, ML, Framework and I/O development targeting next-gen architectures, post-Moore computing. |
| 2.4.2 | Workflow porting and integration on new platforms | The work on HPC, exascale and opportunistic platforms to implement, optimize and commission new workflows for production, fully integrated with distributed computing. |
| 2.4.3 | Distributed computing development | Distributed computing development that is (ultimately) directed at HL-LHC. Data management and access, workflow management, analysis services, information services. |



Estimated Missing Effort in C&S Ops

| Task | FTE | WBS | Priority |
|---|-----------|-------|----------|
| Get a generator running on next-gen HPCs, e.g. Sherpa or Madgraph | 0.5 | 2.4 | 1 |
| Implement ATLAS framework support for offloading algorithms/tasks to GPU. Interface ML models to ATLAS framework. Support data science tools | 1 | 2.4 | 1 |
| New workflows integrating DDM and WFM like data streaming, intelligent caching and use of hierarchical storage (e.g. tape carousel); authentication/authorization | 1 | 2.4 | 1 |
| MC Reconstruction workflow on LCF class machines, includes Frontier-database issues and understand the I/O load on the data center and how to mitigate this | 1 | 2.4 | 1 |
| BigPanDA monitoring and its integration with the Elastic Search analytics tool. | 0.5 | 2.2.4 | 2 |
| Reengineer FastChain to run on exascale platforms | 1 | 2.4 | 2 |
| Development, management and tuning of Ceph based storage system and integration into a distributed storage solution for the US facility. Development & support for Data Transfer Nodes for DOE HPC interfaces with Globus-online. | 0.5 | 2.3.1 | 3 |
| ATLAS simulation workflows, fast and full, on the A21 machine and its precursors | 1 | 2.4 | 3 |
| Implementation of Derivation workflow on LCF class machines. In particular the I/O | 0.5 | 2.4 | 3 |
| ... four more rows ... | 3 | | 4 |
| Estimated Missing Effort | 10 | | |
| For next 2-3 years, per year | | | |

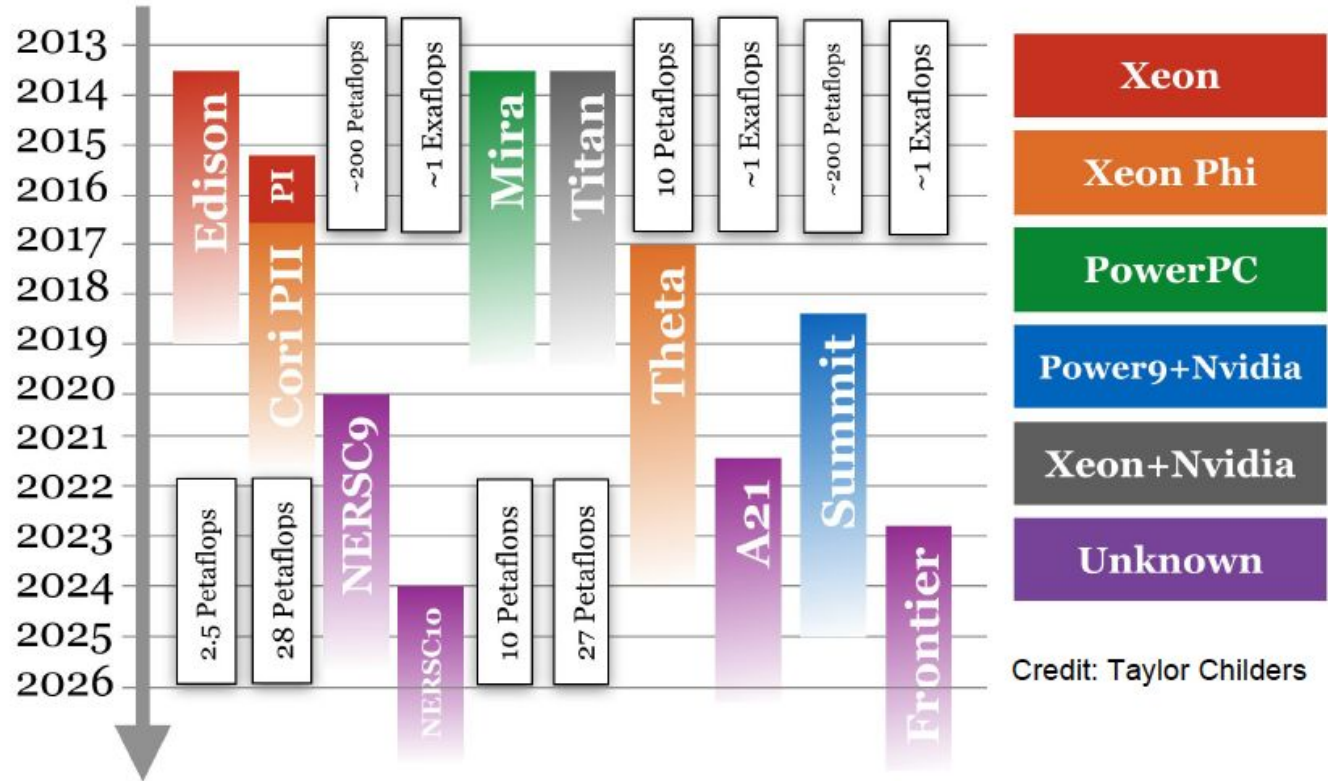


US ATLAS - Technical Strategies

- ❖ Resource shortage has driven most long term developments
 - CPU shortage - much of our focus so far
 - HPCs, Exascale, Coprocessors
 - Faster simulation, faster tracking, algorithm parallelism
 - New opportunities - ML, DL
 - Storage shortage - more difficult to mitigate
 - Hot/Warm/Cold storage - better archiving/unarchiving
 - Caching and streaming technologies, industrial partners
 - Data lake/ocean - reduce multiplicity
 - Network shortage - may be flying under the radar
 - Understand requirements, needs, usage
 - Cost matrix - how to optimize CPU vs storage vs network
- ❖ We also need to be strategic about new opportunities
 - For example: HPC, coprocessor architectures, non-HEP storage solutions, analysis technologies, computing model...
- ❖ US ATLAS evaluates progress and re-prioritizes regularly



Why is HPC Hard?



Credit: Taylor Childers

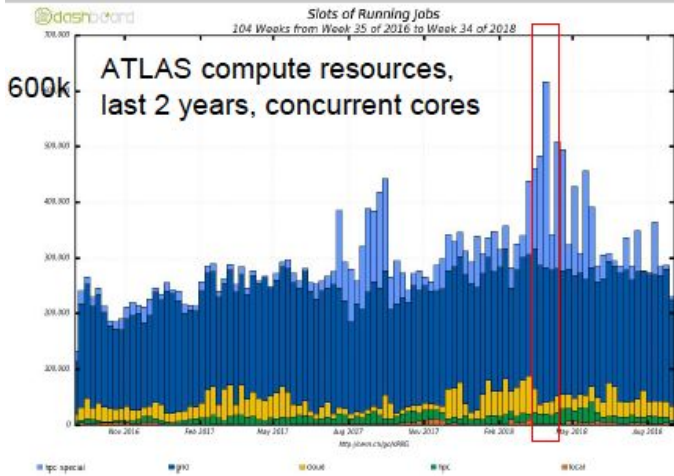
05/16/2018

PC - Inventory

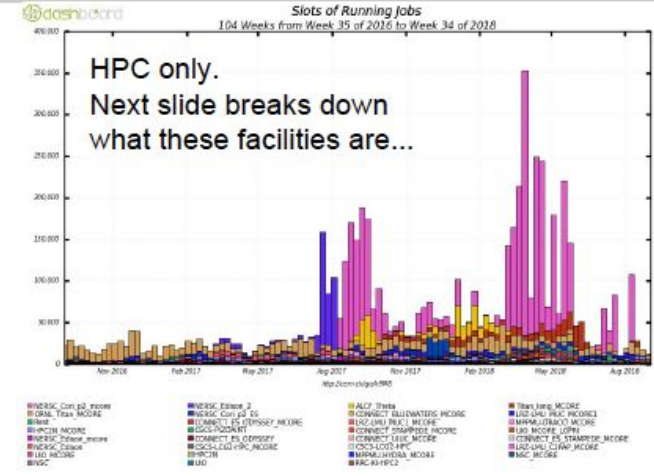
9

Slide from Paolo ~6 months ago

HPCs in ATLAS: deep experience & capability

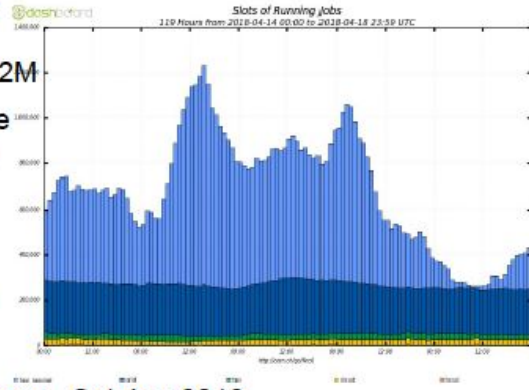


A long history but a new era in the last year: very large facilities, so far in the US



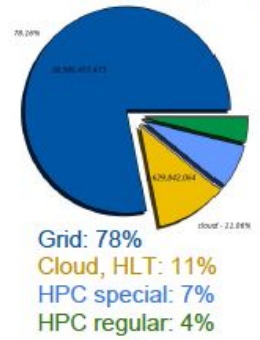
Light blue: "special" HPCs, where special means big, difficult to use, US DOE
 Dark blue: the grid
 Yellow: cloud resources including (dominantly) HLT
 Green: "regular" HPCs, meaning easier to use, operate like a grid site, European or US NSF

Zoom showing full size of scaling peak: 1.2M concurrent cores.



Our workload management system is highly scalable!

CPU HS06 shares, last year



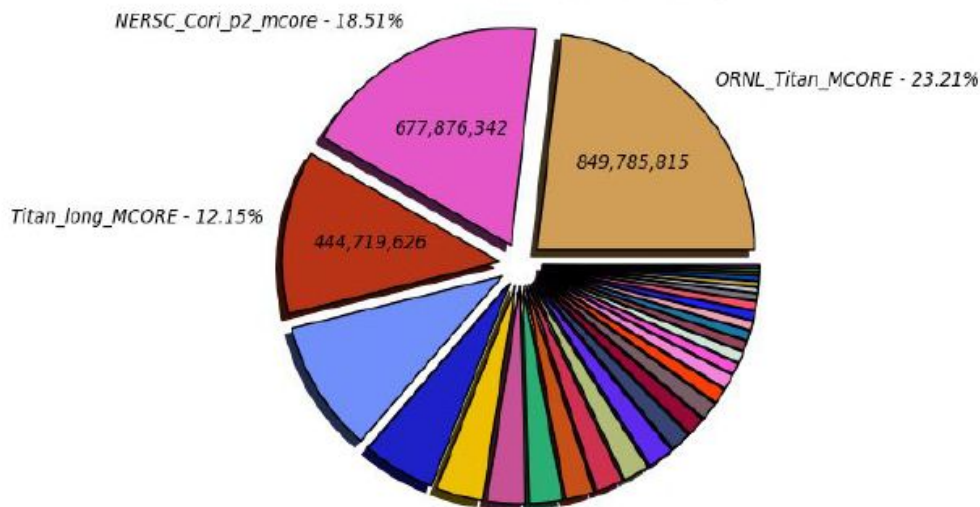
T. Wenaus October 2018



Slide from Torre few weeks ago



CPU HEPSPEC06 (Sum: 3,661,564,165)



http://cern.ch/gog9IK

| | |
|--|---|
| ORNL_Titan_MCORE - 23.21% (849,785,815) | NERSC_Cori_p2_mcore - 18.51% (677,876,342) |
| Titan_long_MCORE - 12.15% (444,719,626) | HPC2N_MCORE - 0.84% (30,210,322) |
| LRZ-LMU_MUC_MCORE - 5.41% (198,226,732) | ALCF_Theta - 3.47% (126,994,896) |
| UIO_MCORE - 2.85% (104,855,229) | Rest - 2.47% (90,576,620) |
| UIO_MCORE_LOPRI - 2.12% (77,473,702) | CONNECT_ES_ODYSSEY_MCORE - 2.04% (74,741,607) |
| HPC2N - 1.93% (70,849,251) | NERSC_Edison_2 - 1.91% (69,799,172) |
| NSC_MCORE - 1.52% (55,773,020) | CONNECT_ES_ODYSSEY - 1.30% (50,717,683) |
| CSCS-LCG2-HPC_MCORE - 1.34% (48,934,003) | CONNECT_STAMPEDE_MCORE - 1.11% (40,600,424) |
| MPPMU-DRACO_MCORE - 1.10% (40,256,724) | LRZ-LMU_C2PAP_MCORE - 0.86% (31,398,295) |
| CSCS-LCG2-DRACO - 0.83% (30,870,556) | |

Breakdown of the HPC facilities in the previous plots

- US DOE HPCs (all in the “special” category)
 - Titan at Oak Ridge
 - Cori at NERSC (successor to Edison)
 - Theta at ANL (successor to Mira)
- Nordugrid
 - Several of their facilities are HPCs, including HPC2N, #4
- European HPCs
 - LRZ (SuperMUC), MPPMU, CSCS, ...
- US NSF HPCs
 - Sites with ‘CONNECT’ in their name

All in routine production, mostly Geant4 MC simulation



CPU is *not* the biggest HL-LHC computing challenge!



- The DOE exascale mandate is driving particular attention to CPU, not without reason as it's a challenge, but storage is a greater challenge
 - Extrapolating today's computing to HL-LHC gives a 3x deficit in CPU and a 6x deficit in storage
- US ATLAS is a leader in advancing R&D to reduce storage needs, and collaborates closely with a WLCG R&D effort we've helped to create
- Examples of storage-directed activities driven by US ATLAS
 - Advanced xrootd based caching for efficiently distributing hot data
 - US ATLAS co-leads the WLCG R&D in this area
 - xrootd's creator and project leader is in US ATLAS
 - Tape carousel workflows serving data from tape, with the potential to reduce disk needs dramatically
 - BNL Tier-1 has longstanding expertise in this from RHIC
 - Leverages the US-developed workload manager PanDA's tight coupling with data management to orchestrate the workflow
 - Event streaming service for fine-grained, optimized data delivery
 - Next step in the US ATLAS driven development of the event service

Event caching and streaming is important



HL-LHC R&D Collaborations: IRIS-HEP

❖ Let's work on it this week...



❖ US ATLAS collaborates directly with IRIS-HEP through Gordon, Heather, Rob, Kyle, Mark and others involved



HL-LHC R&D Collaborations: DOE

- ❖ **Geant ECP/CCE (HEP)**
 - Identify Geant physics models suitable to run on ECP architectures. Quantify gains (Evans/Canal)
- ❖ **Tracking CCE (ATLAS/CMS)**
 - Quantify gains/parallelization opportunities for select tracking algorithms (Pagan-Griso/Kortelainen)
- ❖ **HEP.TrkX ASCR/CCE (ATLAS/CMS)**
 - Exploration of data-driven tracking algorithms (PC/Spentzouris/Spiropoulou)
- ❖ **NESAP for Data (HEP)**
 - Parallel distributed ROOT I/O (with ALCF)
- ❖ **Aurora Early Science**
 - Simulating and Learning in the ATLAS Detector at the Exascale
 - Early access to A21 architecture/software platform



HL-LHC R&D Collaborations: BNL CSI

Organized first US ATLAS HL-LHC Computing activity workshop together with BNL CSI (Computational Science Initiative),

- brought in ATLAS software experts to work with CS specialists in ML, GPUs and HPC utilization, 3 days in July 2018
- We established two working groups that have been active since the workshop:
 - **Fast simulation on accelerators**
 - **Distributed ML training**
- These working groups have established initial objectives and are working collaborations between ATLAS members and CSI personnel

Turn this into template on how to engage CS specialists from outside HEP

- ❖ Collaborations with OLCF, ALCF, NERSC, SLAC
- ❖ Developing shared language requires non-trivial (expert) effort from both sides



HL-LHC R&D Collaborations: Google

- ❖ **Google-ATLAS Proof of Concept demonstration project**
 - Data transfers and PanDA jobs shown to work transparently between Google cloud and WLCG sites
 - Results presented at NEXT 2018, CHEP 2018 and many talks at CERN
- ❖ **Expanded R&D projects started in 5 new working groups**
 - Track 1: Data Management across Hot/Cold storage
 - Track 2: Machine learning and quantum computing
 - Track 3: Optimized I/O and data formats
 - Track 4: Worldwide distributed analysis
 - Track 5: Elastic computing for WLCG facilities
- ❖ **All 5 groups co-led by US ATLAS and Google members**
- ❖ **Active interest and participation from international partners**
 - CERN IT, CERN OpenLab, WLCG, Tokyo U, UK & EU institutions...



Appendix 1: Reference talk by Torre Wenaus



HL-LHC Computing R&D

Torre Wenaus (BNL)

US ATLAS Institutional Board Open Meeting
October 24, 2018

Outline



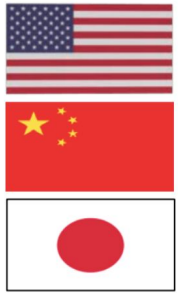

- The HPC landscape today and in the coming years
 - and what this has to do with HL-LHC R&D
- Towards HL-LHC computing in ATLAS
- US ATLAS WBS 2.4: HL-LHC computing
- A work program towards exascale (= supercomputer ≥ 1 Exaflop)

Distilled in part from previous talks, including a longer talk given at CERN's Scientific Computing Forum that covered also (lightly) other LHC experiments, and ATLAS tools important to HPC utilization. It's all in the supplementary slides.

For more on HPCs in ATLAS see the recent [opportunistic resources mini-TIM](#) that took place during ATLAS S&C Week.

HPC global picture over the next few years



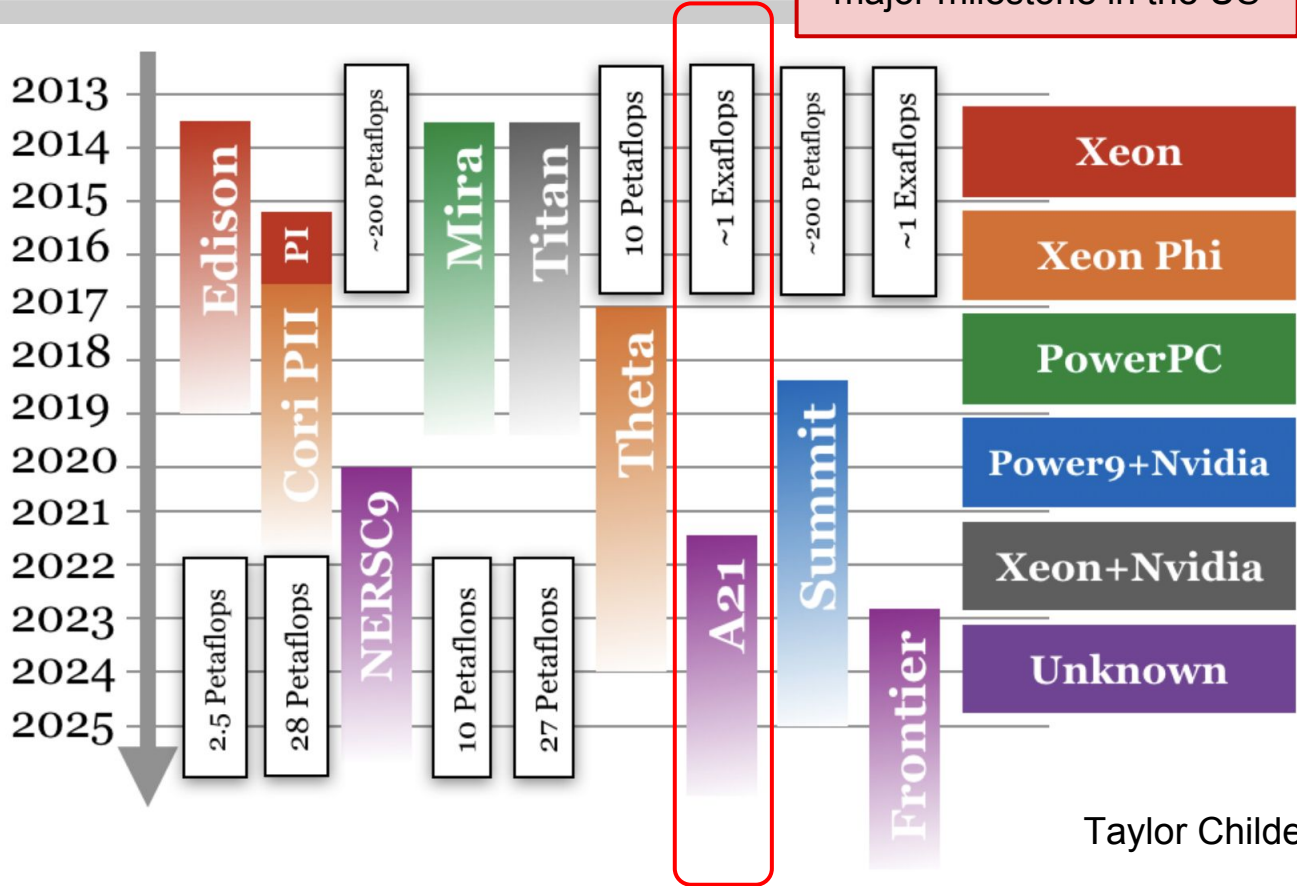
| | | |
|---------------------------|---|--|
| Global Picture HPC |  | <ul style="list-style-type: none">• USA, 4 pre-exa and 3 exascale systems in 2018-2022• China, exascale in 2021?• Japan, exascale in 2022 |
| |  | <p>2 pre-exascale by 2020 and two exascale systems by 2022/2023</p> <p>Hybrid HPC/Quantum infrastructure</p> <p>emerging "computing architectures" (quantum/neuromorphic)</p> <p>novel applications in key areas (Cybersecurity, AI)</p> |

[Andrej Filipcic, June 2018, WLCG MB](#)

HPC evolution in the US



First exascale in 2021: a major milestone in the US



Taylor Childers, ANL

HPCs in HEP: US DOE view

Similar views from HEPAP panel
(supplementary slide)

What We've Learned So Far

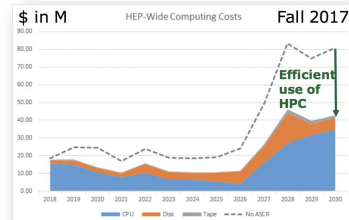
- ▶ HPC architectures will continue to evolve, but moving to vectorized, multithreaded codes tailored to I/O-bound systems will result in higher efficiency codes
- ▶ Engaging HPC experts to analyze code has helped identify algorithm alternatives and data flow bottlenecks, in some cases resulting in spectacular speedups (e.g. 600x). Continued engagement is therefore essential!
- ▶ Need to identify which codes could benefit the most
- ▶ Using Exascale machines badly (e.g. by ignoring the GPU/accelerator) will result in a factor-of-40 penalty in performance that will not be tolerated. HEP will lose its allocations if it does this.
- ▶ Engaging Exascale Computing Project (ECP) experts early and often will result in faster adoption of best practices for exascale machines, and influence ECP design choices to HEP's benefit. HEP needs a coordinated interface to both ECP & the Leadership Computing Facilities.
- ▶ Need to identify which codes could benefit the most
- ▶ LQCD regularly rewrites its code, has reaped significant speedup benefits every time
- ▶ Reinforced that multiyear NERSC allocations & better metrics for pledges are needed
- ▶ End-to-end network data flow models are needed to support tradeoff analysis of storage vs. CPU vs. network bandwidth on a system-wide and program-wide basis
- ▶ Greater sharing of the underlying data management software layer may also be beneficial

We must use HPCs properly
(ie use the accelerators)

And we must use them heavily

Updated HEP Computing Model

- ▶ In preparation for the Inventory Roundtable, the largest HEP experiments from all three frontiers were asked to provide a **more detailed estimate** of their expected computing needs
- ▶ CPU, storage, network, personnel, and HPC portability
- ▶ **Cost estimates for all experimental frontiers:**
- ▶ "Business as usual" (minimal additional HPC use): **\$600M ± 150M**
- ▶ With effective use of HPC resources this reduces to: **\$275M ± 70M**
- ▶ By 2030 cost share by frontier is estimated to be:
 - ▶ ½ Energy Frontier
 - ▶ ¼ Intensity Frontier
 - ▶ ¼ Cosmic Frontier
- ▶ **A strategy encompassing all HEP computing needs is required!**



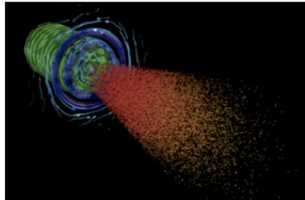
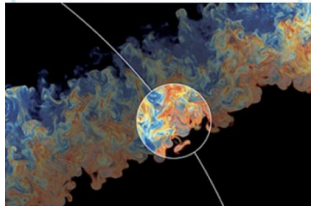
[Jim Siegrist, HEPAP meeting, May 2018](#)

US DOE's ECP program



National security

Next-generation, full-system stockpile stewardship codes
Reentry-vehicle-environment simulation
Multi-physics science simulations of high-energy density physics conditions



Energy security

Turbine wind plant efficiency
Design and commercialization of SMRs
Nuclear fission and fusion reactor materials design
Subsurface use for carbon capture, petroleum extraction, waste disposal
High-efficiency, low-emission combustion engine and gas turbine design
Carbon capture and sequestration scaleup
Biofuel catalyst design

Economic security

Additive manufacturing of qualifiable metal parts
Urban planning
Reliable and efficient planning of the power grid
Seismic hazard risk assessment



Scientific discovery

Cosmological probe of the standard model of particle physics
Validate fundamental laws of nature
Plasma wakefield accelerator design
Light source-enabled analysis of protein and molecular structure and design
Find, predict, and control materials and properties
Predict and control stable ITER operational performance
Demystify origin of chemical elements

Earth system

Accurate regional impact assessments in Earth system
Not us! L-QCD.
Analysis and catalytic conversion of biomass-derived alcohols
Metagenomics for analysis of biogeochemical cycles, climate change, environmental remediation

Health care

Accelerate and translate cancer research



[ECP: Exascale Computing Project](#)

“Accelerating delivery of a capable exascale computing ecosystem”
10-year project led by six DOE and NNSA laboratories and executed in collaboration with academia and industry



Similar themes in US and Europe

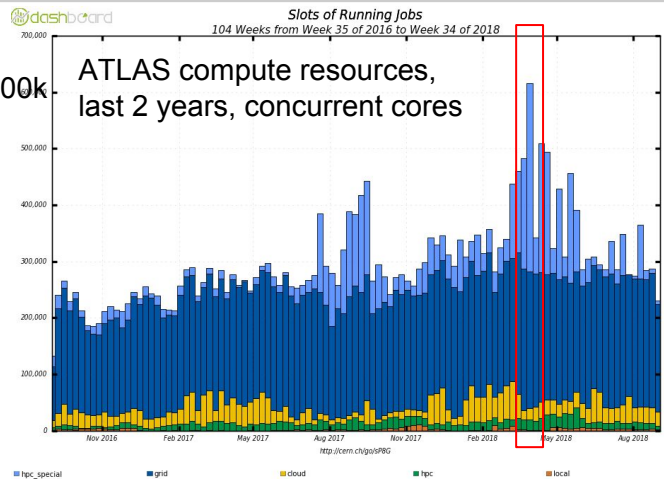


EU HPC strategy highlights in a supplementary slide

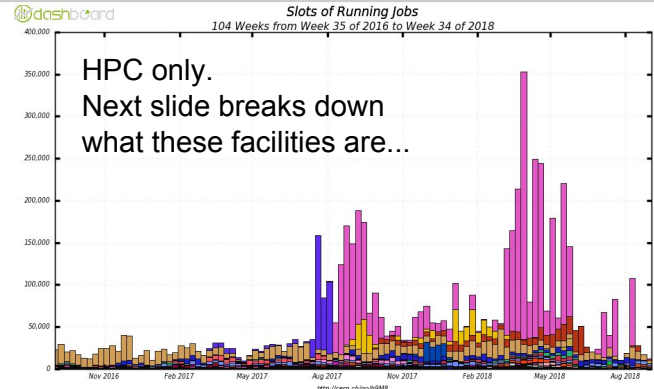
- The HPC planners acknowledge us and the importance of our compute intensive science
- They are building (soon!) **exascale facilities** that they expect us to use
- They don't seem to be taking the **data intensive requirements** of our computing into consideration in system design
- The growth of HPC facilities will **complement and may temper the growth of LHC-dedicated** computing facilities
- We must learn to use these machines: develop **payloads and workflows that exploit them** effectively at manageable development and ops levels
 - There is recognition **we need help**, e.g. [Exascale Computing Project](#) in US is a 10 year billion dollar program by DOE to build an exascale ecosystem
 - ***But at least at present, experimental HEP is not included***
- We must live with their requirements and limitations, but see next point
 - They rely on accelerators, so **we must use the accelerators**
 - Data intensive computing with them may be a challenge
- We need to win **our place at the table for future design** of the HPC landscape
- Some promising signs of more attention to Big Data (HTC) on HPCs, e.g. from EuroHPC
 - And **on HTC we are the experts**. Can we market our expertise?



HPCs in ATLAS: deep experience & capability



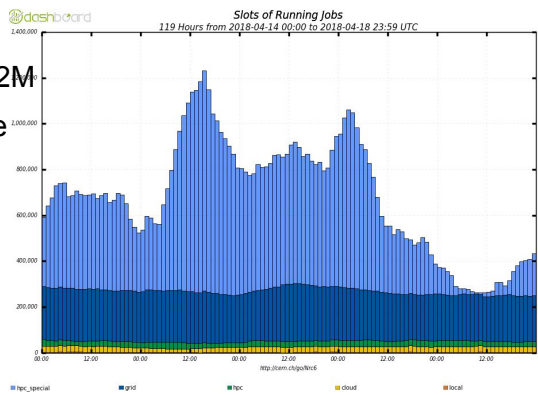
A long history but a new era in the last year: very large facilities, so far in the US



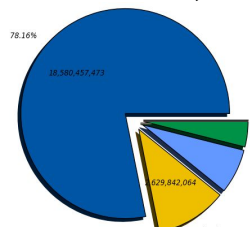
- NERSC Cori p2_mcore
- ORNL Titan_MCORE
- Intel
- HPCIM_MCORE
- NERSC Edison_mcore
- NERSC Edison
- UIO_MCORE
- NSC
- ALCF Theta
- CONNECT BLUEWATERS_MCORE
- LRZ-LMU MUC1_MCORE
- LRZ-LMU MUC2_MCORE
- CONNECT ES ODYSSEY_MCORE
- CSCS-PUIZZANTI
- CONNECT ES ODYSSEY
- NERSC Cori p7 ES
- CONNECT STAMPEDE_MCORE
- CONNECT UIUC_MCORE
- CSCS-LCG2-HPC_MCORE
- HPCIM
- LRZ-LMU MUC3_MCORE
- LRZ-LMU MUC4_MCORE
- LRZ-LMU MUC5_MCORE
- CONNECT ES STAMPEDE_MCORE
- LRZ-LMU E2PAP_MCORE
- NSC_MCORE
- Titan_Iong_MCORE
- LRZ-LMU MUC_MCORE1
- LRZ-LMU MUC_MCORE2
- LRZ-LMU MUC_MCORE3
- CONNECT ES STAMPEDE_MCORE
- LRZ-LMU E2PAP_MCORE
- NSC_MCORE

Light blue: “special” HPCs, where special means big, difficult to use, US DOE
 Dark blue: the grid
 Yellow: cloud resources including (dominantly) HLT
 Green: “regular” HPCs, meaning easier to use, operate like a grid site, European or US NSF

Zoom showing full size of scaling peak: 1.2M concurrent cores.
 Our workload management system is highly scalable!



CPU HS06 shares, last year



Grid: 78%
 Cloud, HLT: 11%
 HPC special: 7%
 HPC regular: 4%

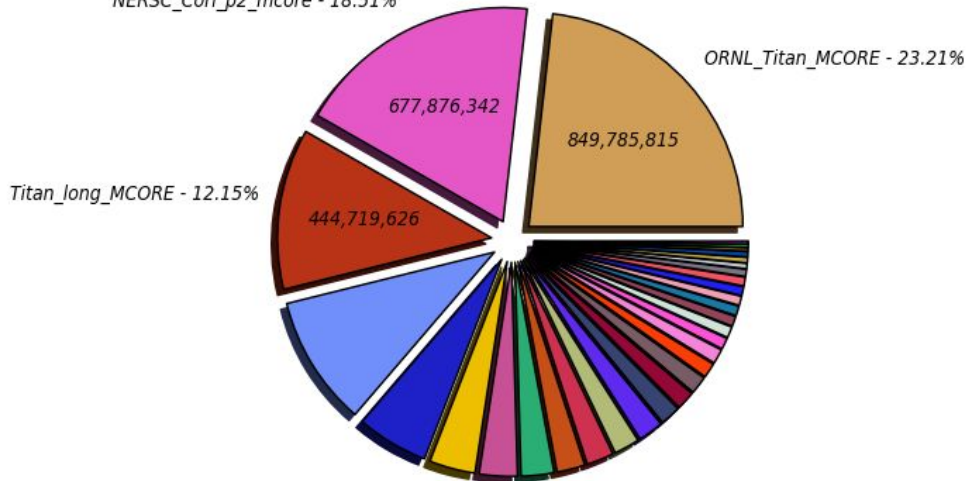


HPCs in ATLAS



CPU HEPSPEC06 (Sum: 3,661,564,165)

NERSC_Cori_p2_mcore - 18.51%



<http://cern.ch/go/g9lIK>



Breakdown of the HPC facilities in the previous plots

- US DOE HPCs (all in the “special” category)
 - Titan at Oak Ridge
 - Cori at NERSC (successor to Edison)
 - Theta at ANL (successor to Mira)
- Nordugrid
 - Several of their facilities are HPCs, including HPC2N, #4
- European HPCs
 - LRZ (SuperMUC), MPPMU, CSCS, ...
- US NSF HPCs
 - Sites with ‘CONNECT’ in their name

All in routine production, mostly Geant4 MC simulation

Towards HL-LHC Computing in US ATLAS



- Triggered in particular by the DOE position re: making HPCs and exascale a major part of the HL-LHC computing strategy...
- Srini and Jim in June 2018 established the new US ATLAS Operations Program
WBS area 2.4, “HL-LHC Computing”
 - (We settled on that name, but today “R&D” is implicit in the title)
 - They asked Heather Gray and TW to set it up and manage it
- A specific, high level activity area that conveys
 - we are listening to and working on DOE’s mandate
 - we are giving greater attention to re-engineering our software for coming HPC generations and, in particular, the first exascale machine
 - we have an organizational home for new development effort applied to the challenges of HL-LHC computing and exascale
- Of course HL-LHC computing presents more challenges than meeting a DOE mandate to use exascale HPCs
 - The WBS breakdown (on a coming slide) reflects that

How WBS 2.4 “HL-LHC Computing” fits in



- Broadly, how do we distinguish “HL-LHC Computing R&D” from the ongoing development and operations work, when by design we keep future-directed development closely coupled to present-day need and application?
 - ie no R&D ivory towers
- We make the distinctions between WBS 2.4 and related WBS areas (Facilities, Software, Physics Support) as seamless and transparent as possible
 - the same managers manage both the “now” and the future R&D
 - the same developers appear in the future R&D as in the “now”; distinction is a matter of FTE fractions assigned to each
- At the same time, we wave the flag that the HL-LHC Computing WBS has unmet effort needs and we make the argument for new effort
 - Embodied in a prioritized list of needed new effort put together by Paolo and Kaushik with input from L2 managers (see supplementary slide)



- 2.4.1 Software reengineering and algorithm development
 - Reengineering for heterogeneous platforms (accelerators), algorithm development (evgen, simu, reco, analysis), exascale-targeted applications, software performance, machine learning, framework and I/O development targeting next-gen processors
 - Co-managers *Ed Moyse, Vakho Tsulaia*
- 2.4.2 Workflow porting to new platforms
 - The work on HPC, exascale and opportunistic platforms to port new workflows for production, fully integrated with distributed computing. Operations of the newly ported systems lies under Facilities
 - Co-managers *Doug Benjamin, Taylor Childers*
- 2.4.3 Distributed computing development
 - Distributed computing development that is (ultimately) directed at HL-LHC. Data management and access, workflow management, analysis support as a service
 - Co-managers *Alexei Klimentov, Wei Yang*

Setting up the activity



- We're in the process of fleshing out the detailed activities and the participants & FTE fractions, together with the L3 managers
- We had a useful [workshop](#) at BNL in July
 - Many senior (US) ATLAS S&C people discussing HPC utilization, machine learning, GPU usage with experts from BNL's Computational Science Initiative (CSI)
 - ATLAS has no offline production applications today that use accelerators
 - Established working groups (next slides) to look for **GPU and ML applications for exascale** and initiate projects
- Planning to broaden similar contacts to other labs, pools of expertise
 - Will have a meeting collocated with the Supercomputing conference in November
- Collaborating in wider contexts: [IRIS-HEP](#), [HSF](#), [WLCG R&D projects](#)
- Fully integrated with ATLAS S&C

Leveraging exascale for ATLAS

Amir Farbin gave a talk on this in the recent ATLAS ML workshop

- **Training deep learning neural networks** as an exascale/accelerator use case is where the BNL workshop landed...
 - The workshop concluded that a promising route for ATLAS to exploit exascale in 2021 -- including, crucially, the use of accelerators -- is via ML applications, in particular
 - **Fast simulation**, and particularly **fast chain** (fast all the way to analysis outputs)
 - **Tracking**, in which there are a number of ML efforts
 - And, **scaling ML applications** to utilize large scale resources in order to minimize turnaround time in network development and tuning
 - **Distributed training** is of interest to achieve fast turnaround
- Presents the possibility of bringing ATLAS workload management tools to bear (PanDA)
- Large scale orchestration of parallel processing, with management of associated data flows and metadata
- Accordingly, the workshop convened fast simulation, distributed training and tracking working groups (*) that have started to develop specific goals and work programs

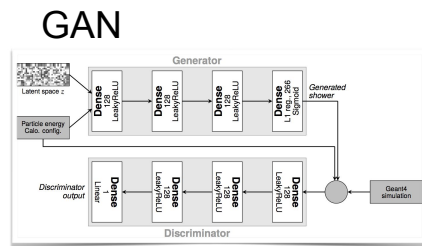


(*) Meetings organized on the open mailing list usatlas-hllhc-computing-l@lists.bnl.gov. See the [info page](#) for the mailing list to sign up.

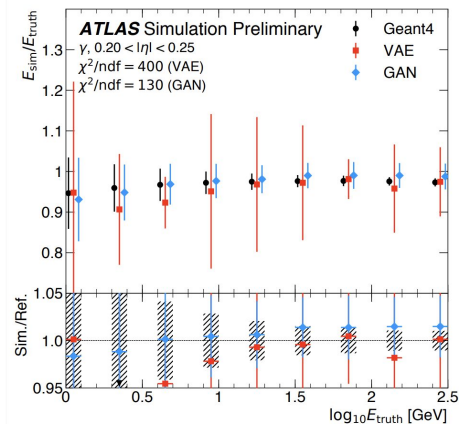
A possible work program towards exascale



- Simulating events in ATLAS is the largest CPU consumer: about 50%
- ATLAS Run-3 objective is to **use fast simulation for most simulation needs**
 - Uses parameterised models of detector response (in particular calorimetry) to achieve a 10x speedup
 - **ML, particularly GANs**, well suited to developing high quality detector response models, with projects now in development, e.g. [CaloGAN](#)
- **GPU/ML based tracking**: innovation to address HL-LHC pileup combinatorics
- Developing, tuning and (re)training of networks for these applications will be a compute intensive process that could be well-suited to exascale
 - Leverages the scale of the machine to minimize turnaround time
 - **Spiking for fast turnaround** rather than steady state for large throughput
 - *Will the demands of training be enough to benefit from exascale?*
- Can we benefit from exascale for fast simulation proper as well as training?
 - *Will ML inference in a fast chain workflow use enough GPU to benefit from exascale? Would enable steady state, large throughput usage*



Early results



CPU is *not* the biggest HL-LHC computing challenge!



- The DOE exascale mandate is driving particular attention to CPU, not without reason as it's a challenge, but storage is a greater challenge
 - Extrapolating today's computing to HL-LHC gives a 3x deficit in CPU and a 6x deficit in storage
- US ATLAS is a leader in advancing R&D to reduce storage needs, and collaborates closely with a WLCG R&D effort we've helped to create
- Examples of storage-directed activities driven by US ATLAS
 - Advanced xrootd based caching for efficiently distributing hot data
 - US ATLAS co-leads the WLCG R&D in this area
 - xrootd's creator and project leader is in US ATLAS
 - Tape carousel workflows serving data from tape, with the potential to reduce disk needs dramatically
 - BNL Tier-1 has longstanding expertise in this from RHIC
 - Leverages the US-developed workload manager PanDA's tight coupling with data management to orchestrate the workflow
 - Event streaming service for fine-grained, optimized data delivery
 - Next step in the US ATLAS driven development of the event service

Every experiment is exploring ML in calo and tracking



Generative Models @ LHC

- Every Experiment is Exploring: ATLAS, CMS, LHCb, ALICE

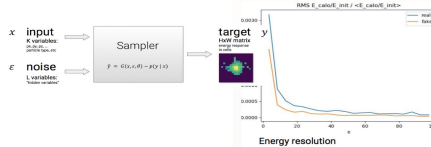
Generative models for fast cluster simulation @ALICE

Amir Farbin, July 2018

Most computational expensive step in simulation is the **particle propagation**
 => avoiding the step using generative models

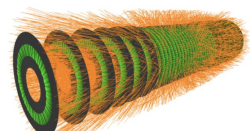
| Method | MSE(mm) | speedup |
|--------------------|---------|-----------------|
| GEANT3 | 0.085 | 1 |
| Random (estimated) | 186.155 | N/A |
| GAN-MLP | 55.385 | 10 ⁴ |
| GAN-LSTM | 54.395 | 10 ⁴ |
| VAE | 37.415 | 10 ⁴ |
| DCGAN | 26.18 | 10 ² |
| cVAE | 13.33 | 10 |
| proGAN | 0.88 | 30 |

Fast calorimeter simulation @ LHCb



TrackML Particle Tracking Challenge
 High Energy Physics particle tracking in CERN detectors
 \$25,000 Prize Money
 CERN · 656 teams · a month ago

Throughput phase on codalab now, needs more participation!
<https://competitions.codalab.org/competitions/20112>
[See this talk for info](#)



HEP.TrkX

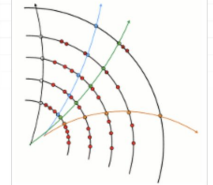
Cross experiment, DOE supported
<https://heptrkx.github.io/>

about HEP advanced tracking algorithms with cross-cutting applications (Project HEP.TrkX)

summary This is an HEP/ASCR DOE pilot project to evaluate and broaden the range of computational techniques and algorithms utilized in addressing HEP tracking challenges. Specifically the project will provide a framework to develop and evaluate new algorithms for track finding and classification, that will be demonstrated by applying advanced pattern recognition techniques to track candidate formation. For example, an optimized track formation algorithm that scales linearly with LHC luminosity, rather than quadratically or worse, may lead by itself to an order of magnitude improvement in the track processing throughput without affecting the track identification performance, hence maintaining the physics performance intact in the LHC upgrades.

A Common Tracking Software (Acts)

<http://acts.web.cern.ch/ACTS/>



<https://indico.cern.ch/event/742793/>

CTD/WIT 2019
 Connecting the Dots and Workshop on Intelligent Trackers
 IFIC, València, Spain
 2nd - 5th April 2019
 Connecting The Dots / Intelligent Trackers 2019

Common ground for collaboration:

- [IML machine learning forum](#) across the LHC experiments
- Community wide [HSF software forum](#)

T. Wenaus October 2018



Thank you



- Thank you to all in US ATLAS and international ATLAS who contributed materials, discussions and most of all work in advancing R&D towards HL-LHC computing

Some related activities & materials



- [GPU hackathon series](#) latest one this week at BNL (DOE sponsored)
- [ANL Aurora A21 early science program](#), ANL HEP selected for participation
- [ATLAS / CSI workshop on development towards exascale](#), BNL, July 2018
 - [Simulation software: fast and full](#), Heather Gray
 - [Proposals](#), Amir Farbin
 - [Scaling DNNs using HPCs](#), Abid Malik
- [Data intensive science at LCFs](#), Jack Wells (ORNL), June 2018
- [BigPanDA for Titan and Summit early science program](#), A. Klimentov, July 2018
- [Connecting the Dots workshop series](#) on advanced tracking
- [Kaggle TrackML](#) ML tracking challenge
 - Ongoing [throughput phase](#)
- [The WBS 2.4 breakdown](#) as of the August scrubbing, Heather Gray
- [WBS 2.4 planning googledoc](#)



HEP computing: US HEPAP panel view



The panel strongly encourages U.S. ATLAS and U.S. CMS to pursue an aggressive “advanced computing” R&D program. In view of the critical role of data handling and processing to the success of these programs, this challenge should not be underestimated.

Thirty years ago, the recognition of the peculiar, event structured, data in particle physics, permitted the use of multiple modest, even commodity, computers in large numbers at significantly lower cost than mainframes. The scale of the future needs for Run 3 of the LHC and particularly for the high luminosity phase, HL-LHC, probably demands an analogous change of approach. What is recognized is the need to use diverse and heterogeneous architectures and to exploit high performance computing facilities, cloud services and data center facilities. The experiments should not underestimate the resources needed to ensure success in this new environment. A paradigm shift in the manner in which the analyses are performed, to enhance the productivity of the experiments, could perhaps be envisaged.

It is important that additional effort be directed towards a new computing model, including a cost model for funding agencies, which ensures data processing and efficient analysis throughput in the HL-LHC running period. In particular, newly emerging computer architectures should be studied and their impact on the performance of the existing code base should be evaluated. Additional burdens for the funding agencies should be identified early and carefully assessed.

[Hugh Montgomery, HEPAP meeting, May 2018](#)

EU HPC strategic research agenda



“A roadmap for the achievement of exascale capabilities by the European High-Performance Computing (HPC) ecosystem”

2.1

THE VALUE OF HPC

2.1.1

HPC as a Scientific Tool

Scientists from throughout Europe increasingly rely on HPC resources to carry out advanced research in nearly all disciplines. European scientists play a vital role in HPC-enabled scientific endeavours of global importance, including, for example, CERN (European Organisation for Nuclear Research), IPCC (Intergovernmental Panel on Climate Change), ITER (fusion energy research collaboration), and the newer Square Kilometre Array (SKA) initiative. The PRACE Scientific Case for HPC in Europe 2012 – 2020 [PRACE] lists the important scientific fields where progress is impossible without the use of HPC.

<http://www.etp4hpc.eu/sra-2017.html>

- CERN is one of top EU scientific endeavours
- But in the document, HEP requirements are not mentioned, apart from lattice QCD
- Future EU HPC centers will be extended to data processing facilities – eg ESiWACE needs large storage, transfers and remote processing (distributed systems)
- Most intensive-computing communities are participating in EuroHPC, HEP is left out for now
- Increased funding from both EC and Member state can result in lower funding of dedicated WLCG infrastructure
- It should be discussed how to ensure HEP presence in future design of EuroHPC landscape

[Andrej Filipcic, June 2018, WLCG MB](#)

US ATLAS HPC resource allocations



US DOE has ASCR Leadership Computing Challenge (ALCC).

For many years we have gotten awards.

In 2016 we were awarded 13M hours at NERSC and 93.5M hours at ALCF (ANL)

In 2017 OLCF was added. 58M hrs at ALCF, 58 Mhrs at NERSC and 80M hrs OLCF (Titan) . We also got 10M hrs at NERSC through ERCAP program.

In 2018 - We received 100M hrs at NERSC through ERCAP program. We have submitted an ALCC proposal for 100 Mhrs ALCF, 70M hrs NERSC, 80 Mhrs OLCF
...and got 80M hours each at ALCF and OLCF from ALCC in 2018

Then there is the backfill time on Titan 195 Mhrs were used in 2017



From Paolo and Kaushik

Prioritized missing effort estimates in US ATLAS ops

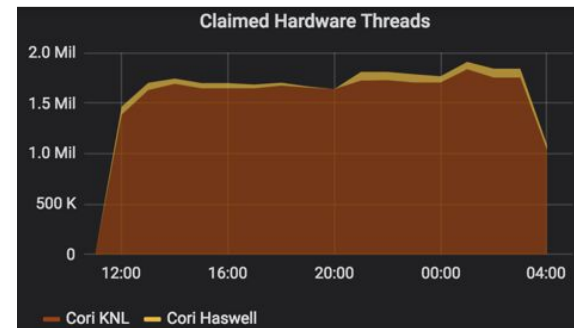
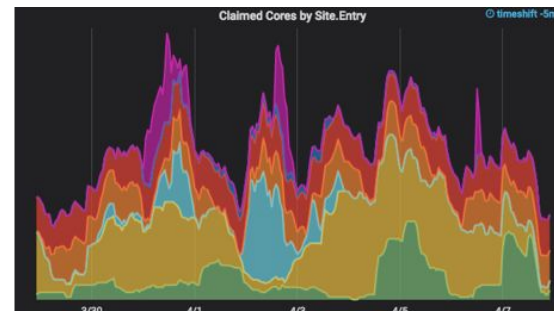
| Task | FTE | WBS | Priority |
|---|-----------|-------|----------|
| Get a generator running on next-gen HPCs, e.g. Sherpa or Madgraph | 0.5 | 2.4 | 1 |
| Implement ATLAS framework support for offloading algorithms/tasks to GPU. Interface ML models to ATLAS framework. Support data science tools | 1 | 2.4 | 1 |
| New workflows integrating DDM and WFM like data streaming, intelligent caching and use of hierarchical storage (e.g. tape carousel); authentication/authorization | 1 | 2.4 | 1 |
| MC Reconstruction workflow on LCF class machines, includes Frontier-database issues and understand the I/O load on the data center and how to mitigate this | 1 | 2.4 | 1 |
| BigPanDA monitoring and its integration with the Elastic Search analytics tool. | 0.5 | 2.2.4 | 2 |
| Reengineer FastChain to run on exascale platforms | 1 | 2.4 | 2 |
| Development, management and tuning of Ceph based storage system and integration into a distributed storage solution for the US facility. Development & support for Data Transfer Nodes for DOE HPC interfaces with Globus-online. | 0.5 | 2.3.1 | 3 |
| ATLAS simulation workflows, fast and full, on the A21 machine and its precursors | 1 | 2.4 | 3 |
| Implementation of Derivation workflow on LCF class machines. In particular the I/O | 0.5 | 2.4 | 3 |
| ... four more rows ... | 3 | | 4 |
| Estimated Missing Effort | 10 | | |

HPC usage in CMS: US



- Using US HPC resources [NERSC (Cori), TACC (Stampede), PSC (Bridges)] through Fermilab HEPCloud and Open Science Grid to execute full workflows (generation, simulation w/ pileup, digitization, reconstruction)
 - requires additional attention compared to grid sites
 - Targeting both low-scale (steady-state) and large bursts
- HEPCloud demonstrated running on HPCs at scale, > 2M hardware threads
- Adding in provisioning support for Leadership Class Facilities (ALCF, OLCF) - nodes have no internet access

CMS is preparing a document with minimal requirements and strategies to approach HPC centers; they will be happy to share it with WLCG



HPC usage in CMS: Europe

- **CH:** Strong collaboration with CSCS
 - Support for HEP workflows out-of-the-box
 - Grid integration via ARC-CE
 - “Friendly”: CVMFS, outbound networking
 - Pursuing use as a “detached Tier-0”, performance similar to CERN, test at 10k core scale imminent
- **IT:** PRACE/CINECA collaboration
 - CVMFS yes, Singularity yes, Outbound connectivity yes(-ish)
 - Testing phase of CMS (and not only) sw on the KNL partition (20 Pflops)
 - Going to apply for a PRACE grant together with the other LHC Experiments in Italy; resources to be seen via T1-CNAF
- **ES:** Use of HPC facilities for HPC workflows at scale is under discussion
 - Successful end-to-end integration of Mare Nostrum (BSC) in ATLAS WMS, relies on ATLAS mechanism to cope with no outbound connectivity (Harvester)
 - ATLAS also got 200k hours on Mare Nostrum

“Friendly” defined:

External connectivity

CVMFS for software installation

Virtualization present

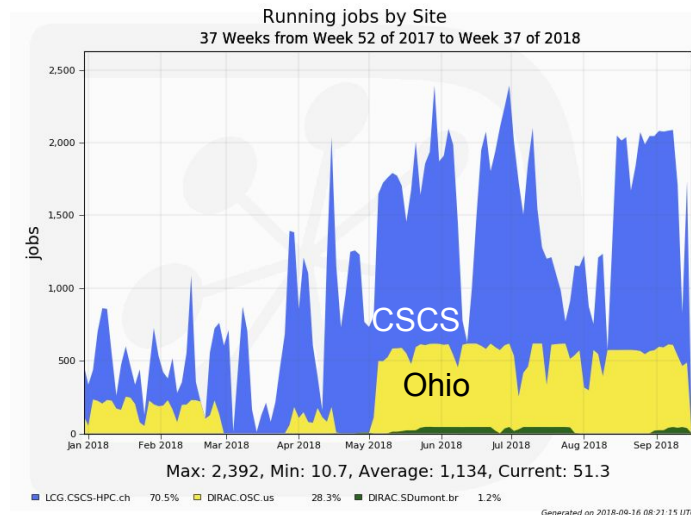
x86 architecture, adequate memory

Workable security

HPC usage in LHCb



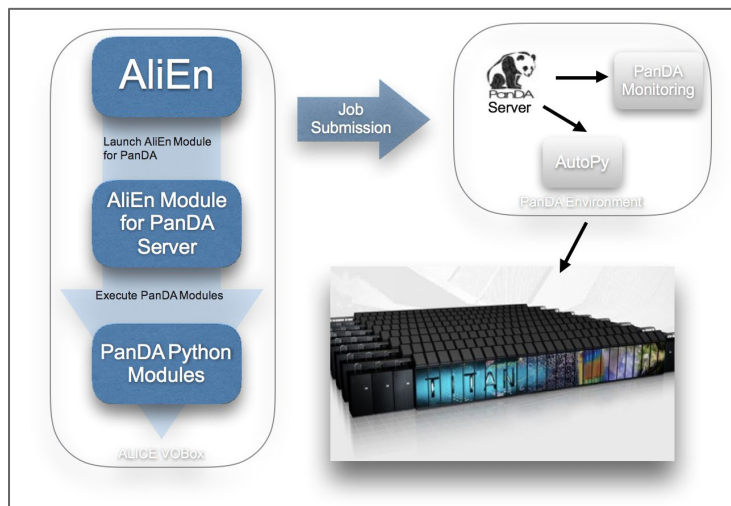
- Used for MC production, almost seamlessly when friendly (same definition)
- Friendly HPCs in use: CSCS, OSC Ohio (1.5% of jobs)
- Some work (and it takes work) on less friendly facilities but no scaling up yet (expected soon at e.g. Santos Dumont HPC center in Rio)
- Ongoing effort to use Knights Landing, testing at CERN and Bologna
 - Simulation 7-10x slower than typical grid; consistent with ATLAS findings
- Multi-threaded Gaudi coming for Run-3, will reduce many-core memory consumption



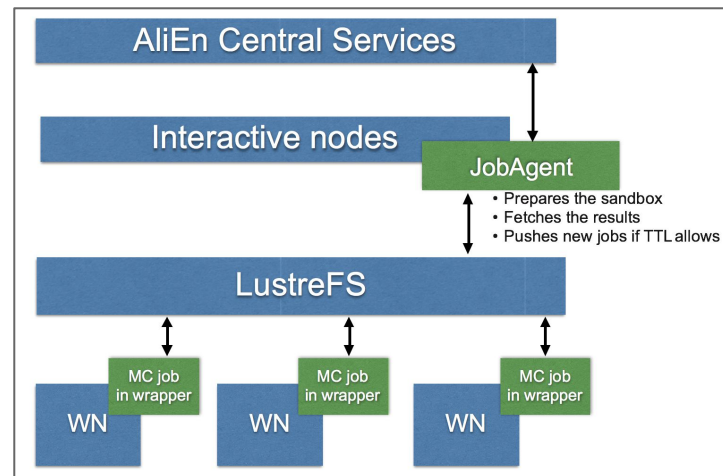
HPC usage in ALICE



- Long time collaborators with ATLAS/PanDA on using Titan at Oak Ridge for HEP & NP
- AliEn - PanDA integration leverages PanDA services at Titan
 - ALICE jobs submitted via PanDA



ALICE services for Titan

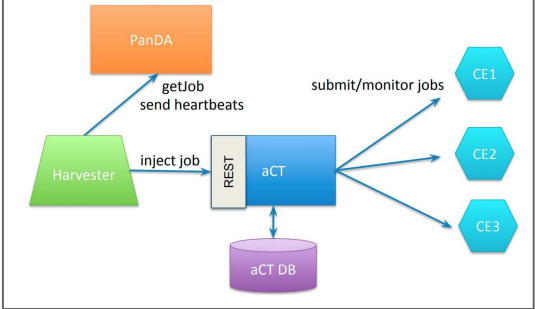


Nordugrid's ARC software

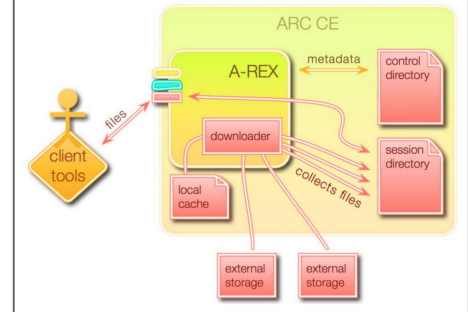


- Has long been the backbone of HPC integration in Europe
- Considered using ARC for US DOE HPCs but not seriously enough to make it happen
- Integrates workload and data management
- Isolates internal details from external users
- Integration requires little manpower for each system
 - But some policy dependence, friendly = easier
- Integrated with Event Service
- Integrating with Harvester to support advanced, dynamic workflows

Harvester + aCT submission backend



Using the ARC cache

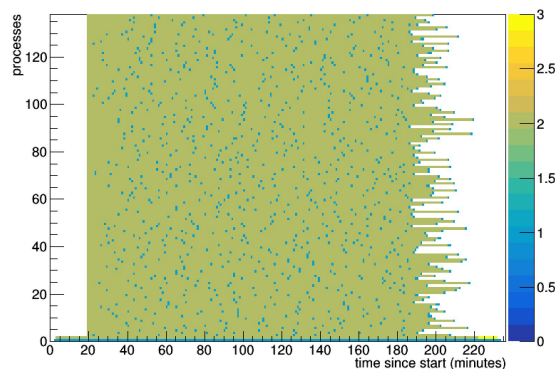


ATLAS Event Service (AES)

AES produces 18% of ATLAS MC events today, and growing

Without event service, each core processes N events. Once a core has finished its allocation, it idles (white)

NERSC utilization per core, no AES

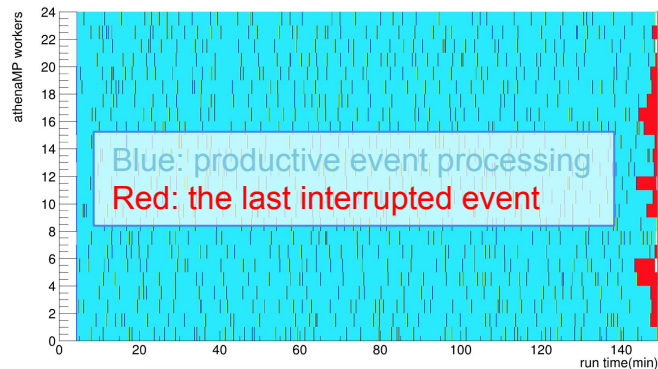


With event service, each core is allocated events to process until the scheduler slot ends.

Largely a US development

If the job is suddenly killed by preemption, all processed events are preserved except the last few minutes (all are lost in a conventional job)

NERSC utilization per core with AES

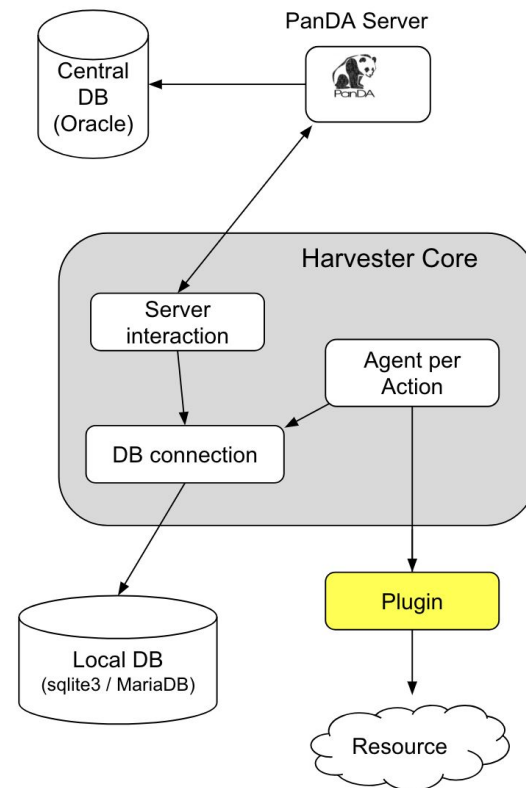


Harvester

Largely a
US development



- A new interface, common across resource types, between resource and workload manager (PanDA or other)
- Particularly useful for the “special” HPCs, each is special in its own way: no uniformity in interfaces, scheduling policies, internal data handling, remote access, external data flows, security, ...
- Use Harvester to provide uniformity by encapsulating the heterogeneity in an edge service
 - A plugin for each unique machine
- Plugins allow to independently optimise each system according to its policies, capabilities, limitations
- Plugins implement data management and data ingress/ingress for the machine also, which is highly sensitive to the data characteristics of the site
 - cf. the fact they aren't built to be data intensive -- treat them with special handling via Harvester
- *Expensive in effort, but that's the nature of these systems*





Appendix 2: Reference talk by Paolo Calafiura



Computing and Software News

Paolo Calafiura, Kaushik De
May 16, 2018
US ATLAS IB



HL-LHC Shared R&D

Summary of SciDAC-4 and Cross-cutting activities: HEP Roadmap Impact

The world-wide HEP community, through the development of a community-wide white paper, has identified a number of important software and computing challenges that need to be addressed over the next decade, largely in the context of HL-LHC. This document highlights how DOE research projects will impact these challenges.

| R&D Areas | SciDAC-4 projects | CCE/CompHEP projects |
|--|--|---|
| Data Analysis Systems (*) | HEP Data Analytics (EFIF): HPC storage and workflow mgmt tools applied to analysis. | Big Data (CMS/IF): Spark and distributed python on HPC. |
| Reconstruction and Trigger Algorithms (*) | HEP Event Reco (EFIF): Parallelization and vectorization of tracking algorithms. | HEP-TKX (HL-LHC): advanced data-driven tracking algorithms |
| Data Organization, Management and Access (*) | HEP Data Analytics (EFIF): full online dataset storage, permitting analysis using direct access to physics objects. | Big Data, Data Transfer, Edge Services, Burst Buffers |
| Applications of Machine Learning (*) | | HEP-TKX (HL-LHC): Deep Learning algorithms for tracking |
| Physics Generators | Generators (EF/IF) provide a new framework for the theoretical particle physics community for MC integrators HEP Data Analytics (EF) Tuning of generators to data using advanced optimization techniques on HPCs | Generator Scaling (EF) |
| Data-Flow Processing Framework | HEP Data Analytics (EFIF): incorporate event selection and reconstruction workflows into the analysis workflow, combining framework applications into full-scale workflow. | Frameworks (EF/CMS): vectorized algorithms; co-processor, many-core (GPU), and multicore scheduling; parallel I/O; multi-language features. |
| Facilities and distributed computing | HEP Data Analytics (EFIF): demonstrate new workload scheduling and data storage capabilities for near-real-time analysis (new model for analysis) | Data Transfer, Edge Services, Burst Buffers |
| Detector Simulation | | Geant (EFIF): vectorized geometry (VecGeant) and particle transport, multithreading within experiment frameworks, parallel random number generation, many-core (GPU, KNL) support, hadronization models |
| Software V&V, Development, Deployment | | SpackDev (EF/IF): Modernizing HEP build release with HPC toolkits |

**SCIDAC
COMP-HEP**

For each of the CWP target R&D focus areas, the table lists the SciDAC-4 projects and the CCE/CompHEP projects that will impact relevant future computing challenges. The SciDAC projects included are: Physics Generators (Hoerche), HEP Event Reconstruction with Cutting Edge Computing Architectures (Cernat), and HEP Data Analytics on HPC (Kowalkowski). The CCE/CompHEP projects included are the Big Data on HPC, Data Transfer, HPC Edge Services, I/O & Storage Hierarchy (Burst Buffers), Scalable Physics Simulations (Generator Scaling), Frameworks for Advanced Architectures, Geant, and SpackDev. Notes: (1) R&D Focus Areas in boldface font and marked with (*) have been called out as initial strategic S202 areas, and are expected to be the first areas addressed by the Institute, and (2) Focus areas of Visualization and Data and Software Preservation within the CWP are not included in the table because no CompHEP-associated DOE projects are contained in this document.

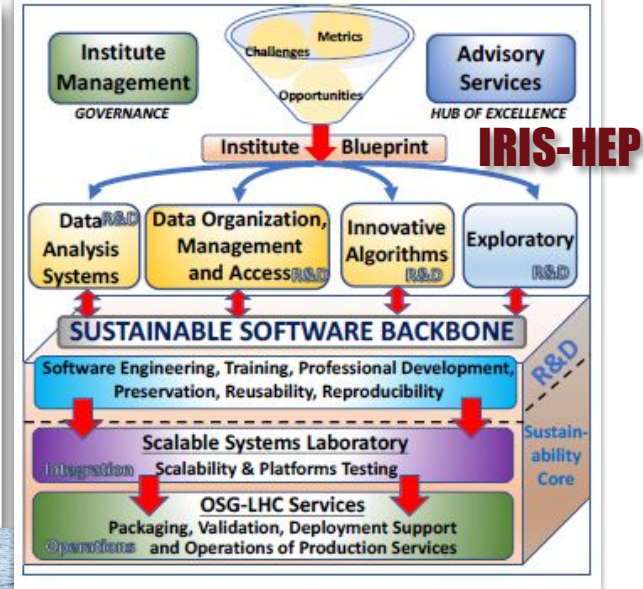
arXiv:1712.06982v3 [physics.comp-ph] 11 Feb 2018

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

HSF

HEP Software Foundation¹

ABSTRACT: Particle physics has an ambitious and broad experimental programme for the coming decades. This programme requires large investments in detector hardware, either to build new facilities and experiments, or to upgrade existing ones. Similarly, it requires commensurate investment in the R&D of software to acquire, manage, process, and analyse the sheer amounts of data to be recorded. In planning for the HL-LHC in particular, it is critical that all of the collaborating stakeholders agree on the software goals and priorities, and that the efforts complement each other. In this spirit, this white paper describes the R&D activities required to prepare for this software upgrade.





HL-LHC: Role of DOE Projects

DOE HEP-CCE and SCIDAC-4 supporting a number of “cross-frontiers” R&Ds including:

- ❖ HEP.TrkX (ASCR/CompHEP)
- ❖ Generators on HPCs (two SCIDAC)
- ❖ Geant(-V) (CompHEP)
- ❖ Data Analysis on HPCs (SCIDAC+LBL LDRD)

Plus other DOE-ASCR projects benefiting ATLAS directly

- ❖ BigPanDA
- ❖ VC³

Summary of SciDAC-4 and Cross-cutting activities: HEP Roadmap Impact

The world-wide HEP community, through the development of a community-wide white paper, has identified a number of important software and computing challenges that need to be addressed over the next decade, largely in the context of HL-LHC. This document highlights how DOE research projects will impact these challenges.

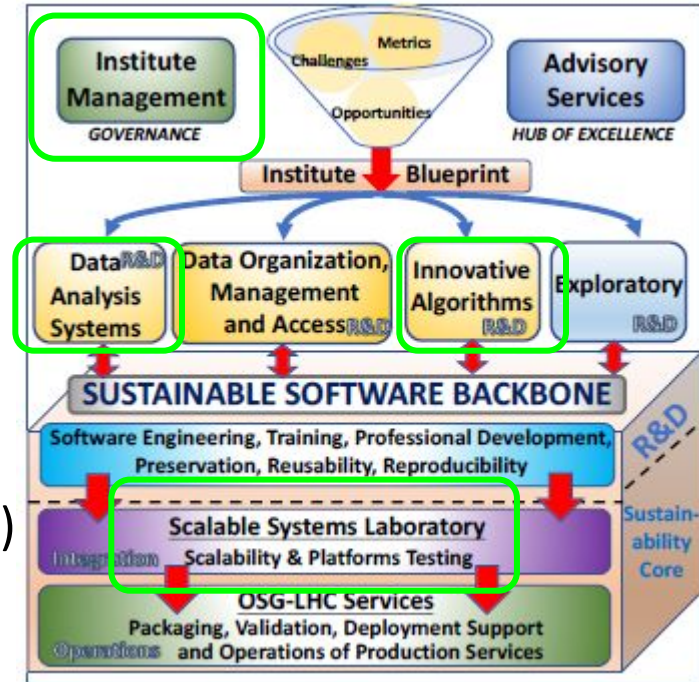
| R&D Areas | SciDAC-4 projects | CCE/CompHEP projects |
|--|--|--|
| Data Analysis Systems (*) | HEP Data Analytics (EF/IF): HPC storage and workflow mgnt tools applied to analysis. | Big Data (CMS/IF) Spark and distributed python on HPC. |
| Reconstruction and Trigger Algorithms (*) | HEP Event Reco (EF/IF): Parallelization and vectorization of tracking algorithms. | HEP.TrkX (HL-LHC): advanced data-driven tracking algorithms |
| Data Organization, Management and Access (*) | HEP Data Analytics (EF/IF): full online dataset storage, permitting analysis using direct access to physics objects. | Big Data, Data Transfer, Edge Services, Burst Buffers |
| Applications of Machine Learning (*) | | HEP.TrkX (HL-LHC): Deep Learning algorithms for tracking |
| Physics Generators | Generators (EF/IF) provide a new framework to the theoretical particle physics community for MC Integrators HEP Data Analytics (EF) Tuning of generators to data using advanced optimization techniques on HPCs | Generator Scaling (EF) |
| Data-Flow Processing Framework | HEP Data Analytics (EF/IF) : incorporate event selection and reconstruction workflows into the analysis workflow, combining framework applications into full-scale workflow. | Frameworks (IF/CMS) : vectorized algorithms; co-processor, many-core (GPU), and multicore scheduling; parallel I/O; multi-language features. |
| Facilities and distributed computing | HEP Data Analytics (EF/IF) : demonstrate new workload scheduling and data storage capabilities for near-real-time analysis (new model for analysis) | Data Transfer, Edge Services, Burst Buffers |
| Detector Simulation | | Geant (EF/IF) : vectorized geometry (VecGeom) and particle transport, multithreading within experiment frameworks, parallel random number generation, many-core (GPU,KNL) support, hadronization models |
| Software V&V, Development, Deployment | | SpackDev (EF/IF) : Modernizing HEP build/release with HPC toolkits |

For each of the CWP target R&D focus areas, the table lists the SciDAC-4 projects and the CCE/CompHEP projects that will impact relevant future computing challenges. The SciDAC projects included are: Physics **Generators** (Hoeche), **HEP Event Reconstruction** with Cutting Edge Computing Architectures (Cerati), and **HEP Data Analytics** on HPC (Kowalkowski). The CCE/CompHEP projects included are the **Big Data** on HPC, **Data Transfer**, **HPC Edge Services**, I/O & Storage Hierarchy (**Burst Buffers**), Scalable Physics Simulations (**Generator Scaling**), **Frameworks** for Advanced Architectures, **Geant**, and **SpackDev**. Notes: (1) R&D Focus Areas in boldface font and marked with (*) have been called out as initial strategic S2I2 areas, and are expected to be the first areas addressed by the institute, and (2) Focus areas of **Visualization** and **Data and Software Preservation** within the CWP are not included in the table because no CompHEP-associated DOE projects are contained in this document.



IRIS-HEP Proposal

- ◆ NSF HEP Software Institute conceptualization award led to Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP) **proposal now submitted**
 - \$5M/year, 5 years proposal
 - Peter Elmer (PI), Gordon Watts, Brian Bockelmann (co-PIs)
 - US ATLAS area managers: Kyle Cranmer (Analysis), Heather Gray (Algorithms), Rob Gardner (SSL)



Targeting FY18 start



HEP Inventory of Computing Needs

HEP Inventory of Computing Needs
 Roundtable Meeting
 May 7-8, 2018
 Cambria Hotel & Suites, Rockville, MD

Room: Sinequa A/B final 5/4/18

| Day 1 : May 7 | | |
|---|---|------------------------|
| 8:30 AM | Welcome, Motivation, Objectives for the Meeting | Jim Siegrist |
| 8:50 AM | The Challenges Ahead | Tom LeCompte |
| 9:30 AM | HPC Examples: HACC | Salman Habib |
| 9:45 AM | HPC Examples: WarpX | Jean-Luc Vey |
| 10:00 AM | HPC Examples: LQCD | Ruth Van de Water |
| 10:15 AM | <i>Break</i> | |
| Lab-by-Lab Overviews of Computing Resources and Strategy | | |
| 10:30 AM | ANL & ALCF+OLCF | Salman Habib |
| 11:25 AM | Brookhaven | Kirsten Kleese van Dam |
| 12:00 PM | <i>LUNCH</i> | |
| 1:50 PM | Fermilab | Liz Sexton-Kennedy |
| 2:35 PM | LBNL & NERSC+ESNet | Peter Nugent |
| 3:30 PM | SLAC | Richard Dubois |
| 3:55 PM | <i>Break</i> | |
| 4:10 PM | Summary of Needs from Data Call | Tom LeCompte |
| 4:45 PM | Discussion | ALL |
| 5:30 PM | Adjourn | |

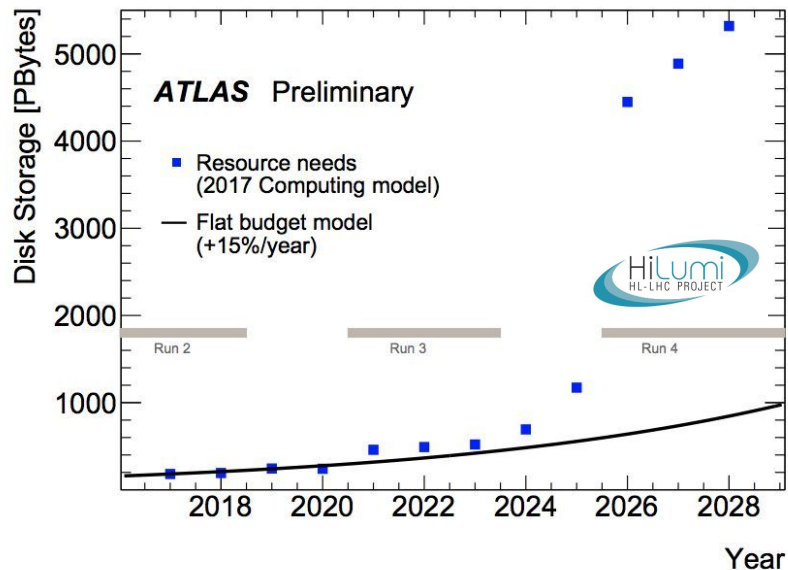
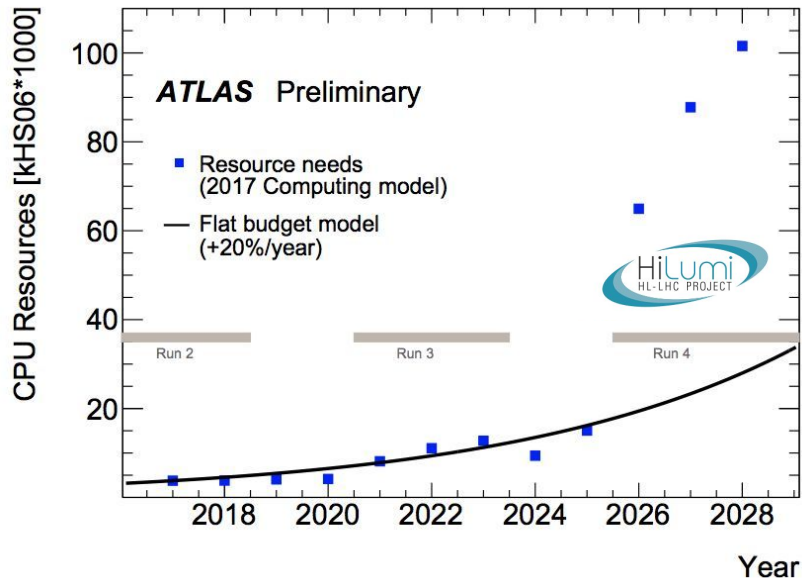
| Day 2: May 8 | | |
|--|---|------------------------|
| Experiment-by-Experiment Overviews of Computing Needs | | |
| 8:30 AM | Objectives for the Second Day | Tom LeCompte |
| 8:45 AM | ATLAS | Paolo Calafiura |
| 9:15 AM | CMS | Oliver Gutsche |
| 9:45 AM | IF Experiments at Fermilab | Panagiotis Spentzouris |
| 10:15 AM | <i>Break</i> | |
| 10:30 AM | LSST/DESC | Katrin Heitmann |
| 10:50 AM | DESI | Peter Nugent |
| 11:05 AM | Cosmological Simulations and Archiving (incl. CMB-S4) | Salman Habib |
| 11:30 AM | Closing, Next steps | Jim Siegrist |
| 11:45 AM | Discussion | ALL |
| 12:30 PM | Adjourn | |

Editorialized Goals:

- ❖ “Quantify” HL-LHC resource shortage
- ❖ Can Energy Frontier learn how to run on HPCs from other HEP communities?
- ❖ Determine role of national labs, Exascale Computing Process
- ❖ Find opportunities to optimize by sharing hardware and people



ATLAS HL-LHC Computing Needs



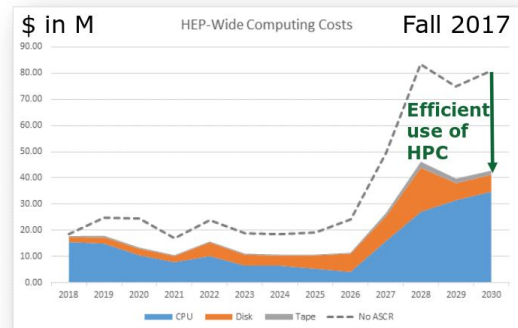
**US ATLAS Run 4 resource gap O(\$100M)
(projections with large uncertainties, current computing model)**



Jim Siegrist @ HEPAP

Updated HEP Computing Model

- ▶ In preparation for the Inventory Roundtable, the largest HEP experiments from all three frontiers were asked to provide a **more detailed estimate** of their expected computing needs
 - ▶ CPU, storage, network, personnel, and HPC portability
- ▶ Cost estimates for all experimental frontiers:
 - ▶ "Business as usual" (minimal additional HPC use): **\$600M ± 150M**
 - ▶ With effective use of HPC resources this reduces to: **\$275M ± 70M**
- ▶ By 2030 cost share by frontier is estimated to be:
 - ▶ ½ Energy Frontier
 - ▶ ¼ Intensity Frontier
 - ▶ ¼ Cosmic Frontier
- ▶ **A strategy encompassing all HEP computing needs is required!**



[Jim Slides](#)

DOE message
is clear:
Use HPCs!





Jim Siegrist @ HEPAP

Homework for
in-depth DOE
workshop
(~three months)

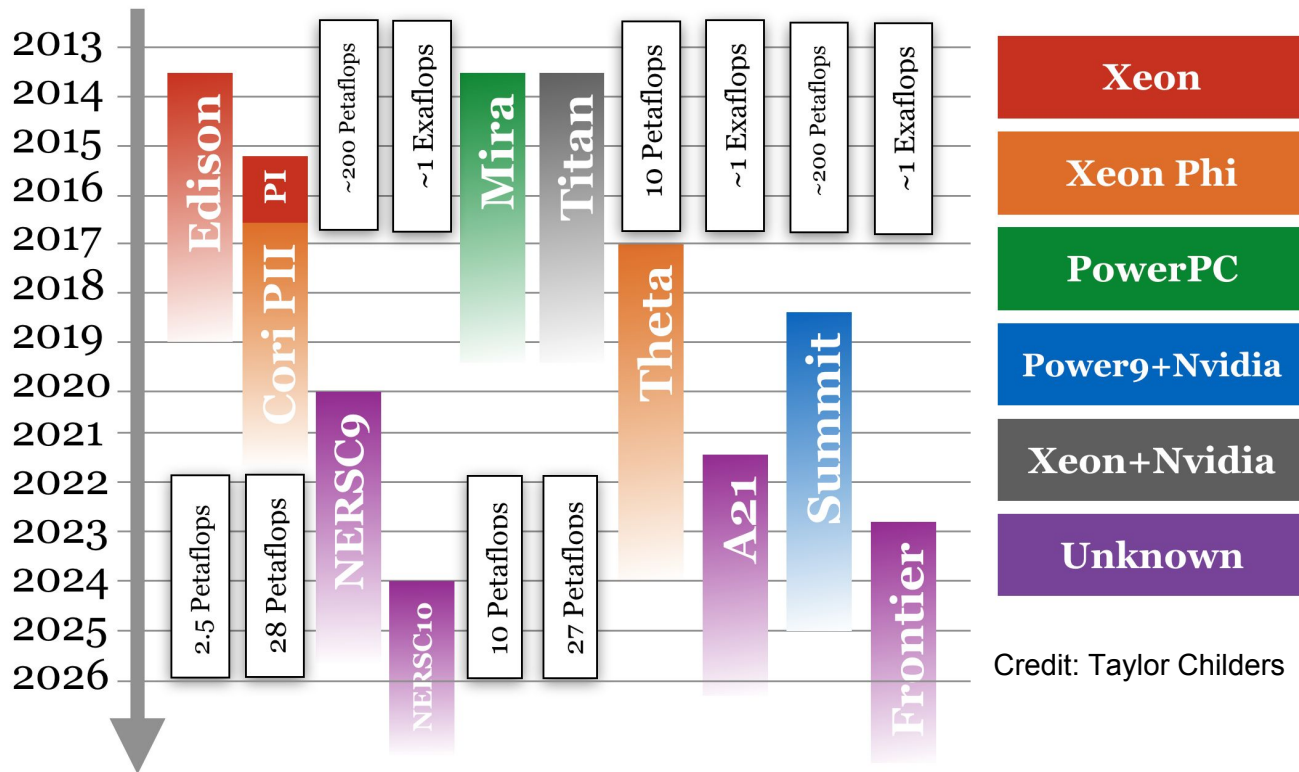
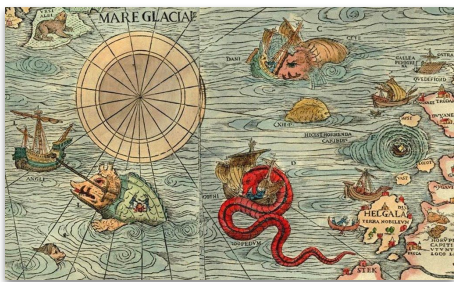
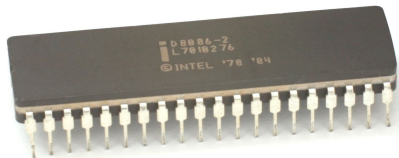
Next Steps

- ▶ OHEP exploring a process to enable multiyear allocations at NERSC
 - ▶ **Studies of selected HEP codes**
 - ▶ In-depth analysis of 1-2 critical codes to identifying resource bottlenecks and opportunities for speedup (both general and GPU-accelerated), drawing on expertise at NERSC, the LCFs, and the ECP
 - ▶ Discussions with the broader community to assess the potential for vectorization and efficient CPU/GPU utilization of the most resource-intensive codes in use; “dissect-a-thons” to triage codes
 - ▶ Identification of recurrent kernels and themes in HEP software
 - ▶ **Identification of common areas where efficiencies of scale can be jointly explored**
 - ▶ Data processing and storage models optimized for current and anticipated CPU/storage/network costs
 - ▶ Shared best programming practices

Community input is important — please work with your experiment’s computing leads to provide input



Why is HPC Hard?



Credit: Taylor Childers



Development Latency

- ❖ Took five years to run massively parallel jobs in production
 - Distributed task farm, running concurrently on each node.
 - Production system manages **single-core to 200K-core jobs**
 - Incremental changes to physics algorithms
 - **Smaller (2.5x less memory), more flexible workflows**
- ❖ Moving to SIMD/MIMD heterogenous architectures considerably more work, not incremental
 - New approaches, new skills needed for physics algorithms
 - **Will heterogeneous systems ever become mainstream?**
 - Will software performance and usability improve?



Next Steps

1. Create new US ATLAS Ops project dedicated to HL-LHC Computing
2. Define priority areas jointly with ATLAS (per Jim S request)
 - Study athena performance, and model heterogeneous architectures
3. Add HL-LHC focus to US ATLAS technical meeting and scrubbing (SLAC, August 28-29)
 - Reconcile operational duties with R&D needs
 - guesstimated 15 additional FTEs, mostly shareable with CMS



Thanks

Thanks to Ken Bloom, Rob Gardner, Oli Gutsche, Eric Lancon, Liz Sexton, Taylor Childers, Doug Benjamin, Wahid Bhimji, Simone Campana, Torre Wenaus, Vakho Tsulaia, Ben Nachman, Michela Paganini, Kaushik De, James Catmore, Bill Murray, Steve Farrell, Charles Leggett, Andrea Dotti, Ale Di Girolamo, David Strom, Jim Kowalkovski...

❖ All mistakes and misappropriations are mine



Backup



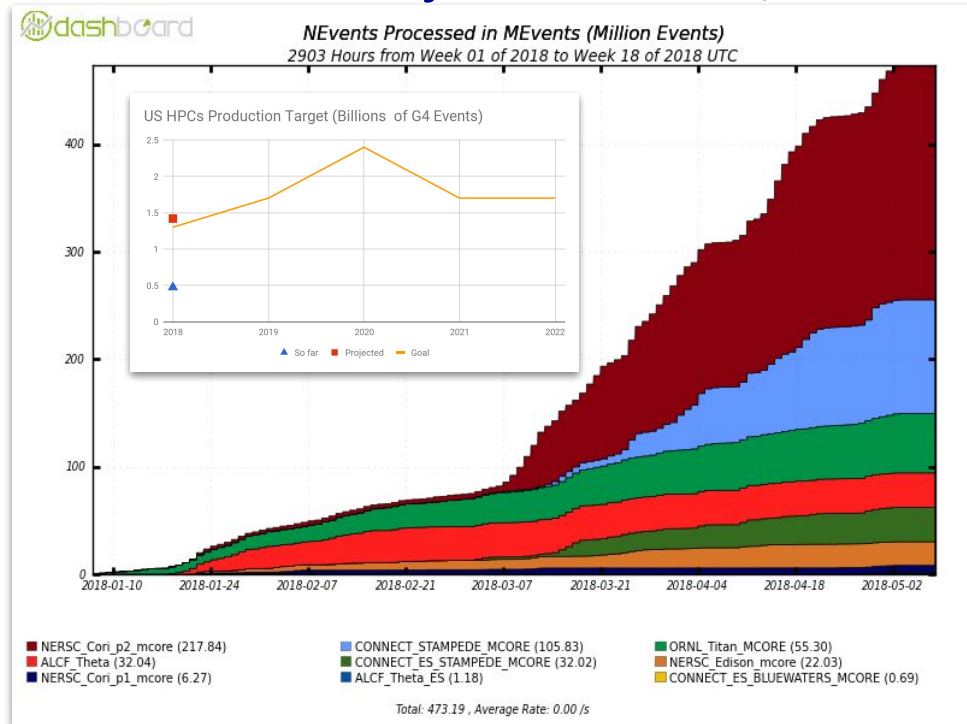
US ATLAS Simulation on HPCs

US ATLAS goal for 2019+ is to run 75% of its G4 jobs on HPCs, 50% in 2018 (1.3 B events).

❖ 470 M G4 events since Jan

- 40% on NERSC cori KNL
- 25% on NSF HPCs
 - running via OSGConnect

❖ Submitted ALCC request with multi-year allocation forecast.





- ▶ There have been many attempts to make use of GPUs for HEP
- ▶ In general, GPUs are not really practical to use as just another regular compute engine
 - excel at some tasks, fail horribly at others
 - code/kernels need to be rewritten to take advantage of hardware / memory layout
 - work best with SIMD style processing
- ▶ Some types of HEP code are well suited for GPUs
 - some pattern recognition (eg tracking)
 - G4 EM and neutral physics
 - Calorimeter clustering
- ▶ Some things don't
 - anything branchy, sorts, etc
 - lots of code, little data
- ▶ Just because we have one way of doing something now that's designed for a CPU, doesn't mean we need to do it the same way on a GPU
 - track fitting w/ kalman -> machine learning for pattern reco

[Charles Leggett talk in Naples](#)



HL-LHC Shared R&D

Summary of SciDAC-4 and Cross-cutting activities: HEP Roadmap Impact

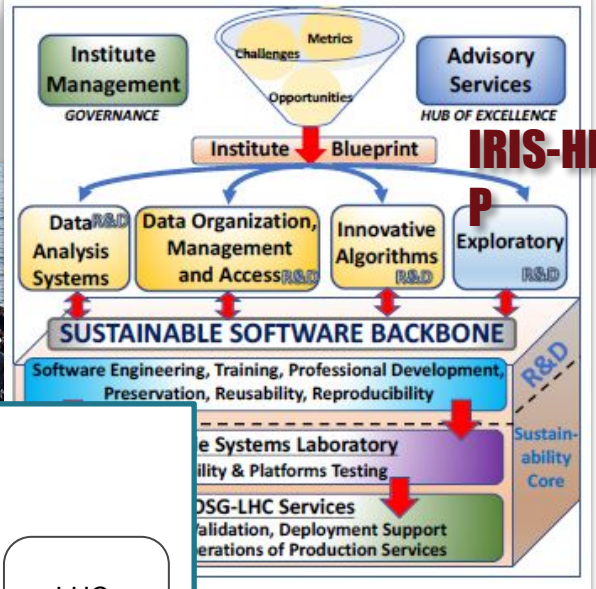
The world-wide HEP community, through the development of a community-wide white paper, has identified a number of important software and computing challenges that need to be addressed over the next decade, largely in the context of HL-LHC. This document highlights how DOE research projects will impact these challenges.

| R&D Areas | SciDAC-4 projects | CCE/CompHEP projects |
|--|---|---|
| Data Analysis Systems (*) | HEP Data Analytics (EFIF): HPC storage and workflow mgmt tools applied to analysis. | Big Data (CMS/IF): Spark and distributed python on HPC. |
| Reconstruction and Trigger Algorithms (*) | HEP Event Reco (EFIF): Parallelization and vectorization of tracking algorithms. | HEP-TKX (HL-LHC): advanced data-driven tracking algorithms |
| Data Organization, Management and Access (*) | HEP Data Analytics (EFIF): full online dataset storage, permitting analysis using direct access to physics objects. | Big Data, Data Transfer, Edge Services, Burst Buffers |
| Applications of Machine Learning (*) | | HEP-TKX (HL-LHC): Deep Learning algorithms for tracking |
| Physics Generators | Generators (EF/IF) provide a new framework to the theoretical particle physics community for MC integrators HEP Data Analytics (EF) Tuning of generators to data using advanced optimization techniques on HPCs | Generator Scaling (EF) |
| Data-Flow Processing Framework | HEP Data Analytics (EFIF): incorporate event selection and reconstruction workflows into the analysis workflow, combining framework applications into full-scale workflow. | Frameworks (IF/CMS): vectorized algorithms; co-processor, many-core (GPU), and multicore scheduling; parallel I/O; multi-language features. |
| Facilities and distributed computing | HEP Data Analytics (EFIF): demonstrate new workload scheduling and data storage capabilities for near-real-time analysis (new model for analysis) | Data Transfer, Edge Services, Burst Buffers |
| Detector Simulation | | Geant (VecGen, multithreaded framework, generative support). |
| Software V&V, Development, Deployment | | SpackDev build system |

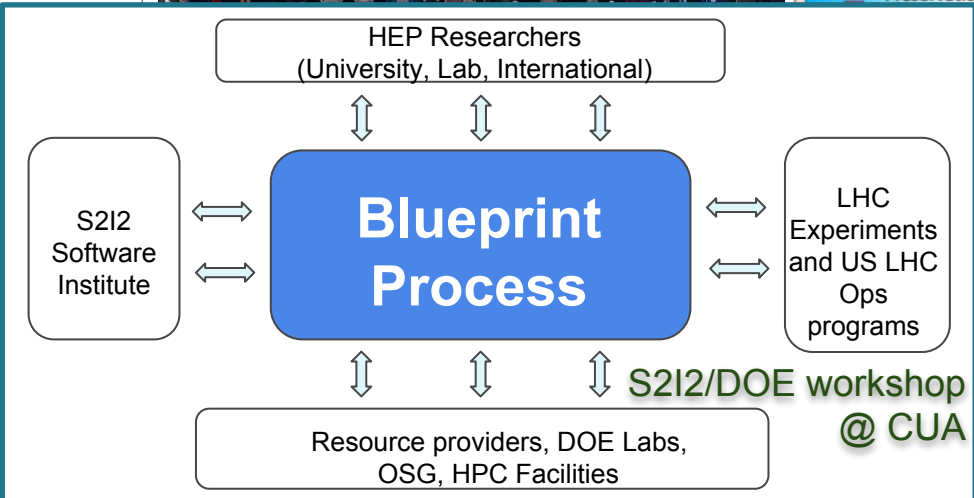
For each of the CWP target R&D focus areas, the table lists the SciDAC-4 projects and that will impact relevant future computing challenges. The SciDAC projects included are: HEP Event Reconstruction with Cutting Edge Computing Architectures (Ceraat), and HE (Kowalkowski). The CCE/CompHEP projects included are the Big Data on HPC, Data T & Storage Hierarchy (Burst Buffers), Scalable Physics Simulations (Generator Scaling), Architectures, Geant, and SpackDev. Notes: (*) R&D Focus Areas in boldface font and out as initial strategic S2I2 areas, and are expected to be the first areas addressed by the Visualization and Data and Software Preservation within the CWP are not included in the CompHEP-associated DOE projects are contained in this document.

11 Feb 2018

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



IRIS-HE



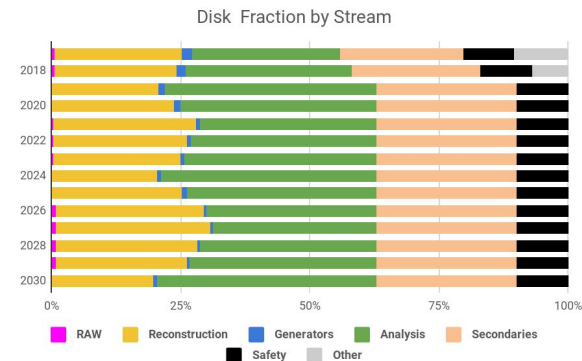
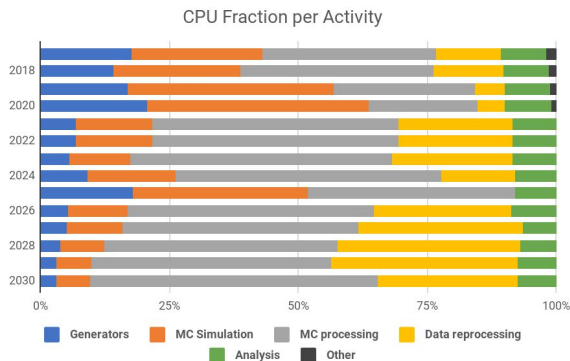


Analyzing the resources needed

Run 2

Run 3

Run 4



Four optimization priorities:

Simulation (Event Generation, G4),

Reconstruction (Tracking),

Analysis Streams,

Data Organization and Delivery



Bridging the Gap: Storage

Optimize analysis streams:

- ATLAS currently has dozens of derived analysis streams 10-100KB/evt each
 - can't afford for Run 4
 - many streams will be merged and/or become **virtual** (produced on demand)
- Physics-aware compression (ALICE, LHCb)

Data delivery: increase the role of tape (5x cheaper than disk)

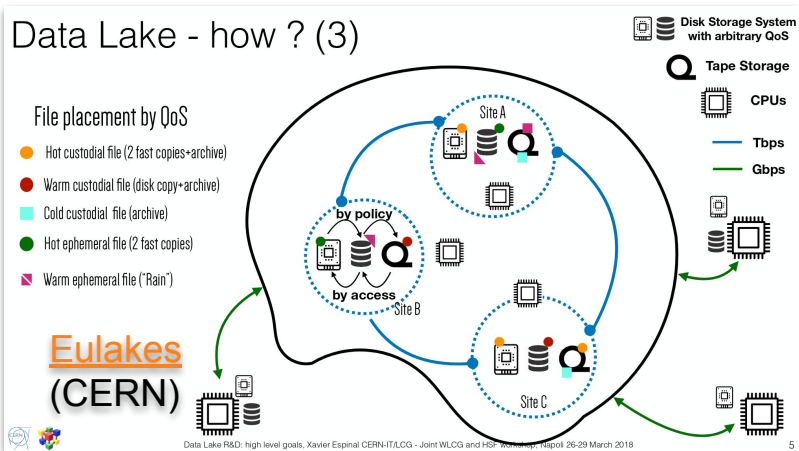
- ATLAS looking to stream data straight from tape for many production workflows



Data Lakes

- Exabytes of heterogeneous, distributed storage connected by ultrafast networks, presented as unified storage to sites/applications. Data transformed on-the-fly from storage-optimized to client-optimized

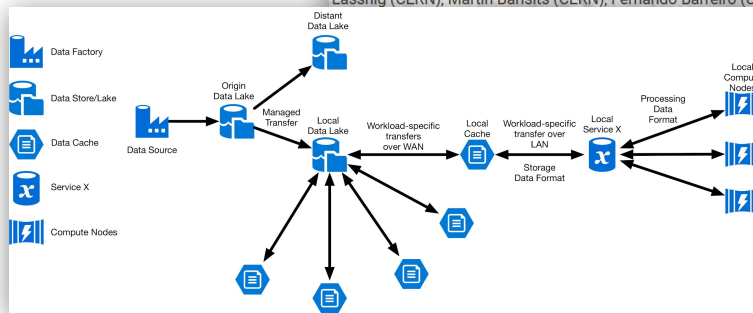
Data Lake - how ? (3)



Cloud Atlas: Oceans of Data

Google Cloud and CERN/Atlas Collaboration

Karan Bhatia (Google), Andy Murphy (Google), Alexei Klimentov (BNL), Kaushik De (UTA), Mario Lassnig (CERN), Martin Barisits (CERN), Fernando Barreiro (UTA), Thomas Beermann (CERN), ... in (BNL), Tobias Wegner heuser (BNL)

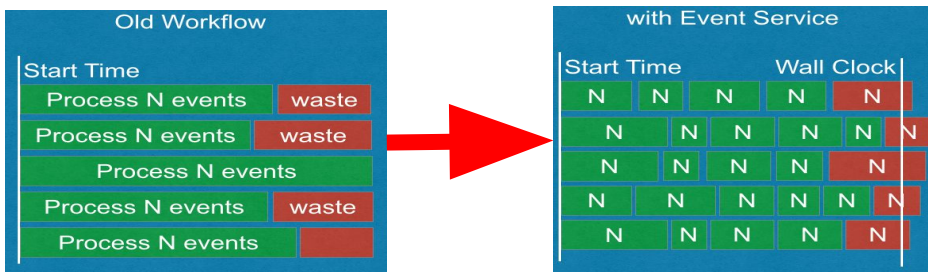
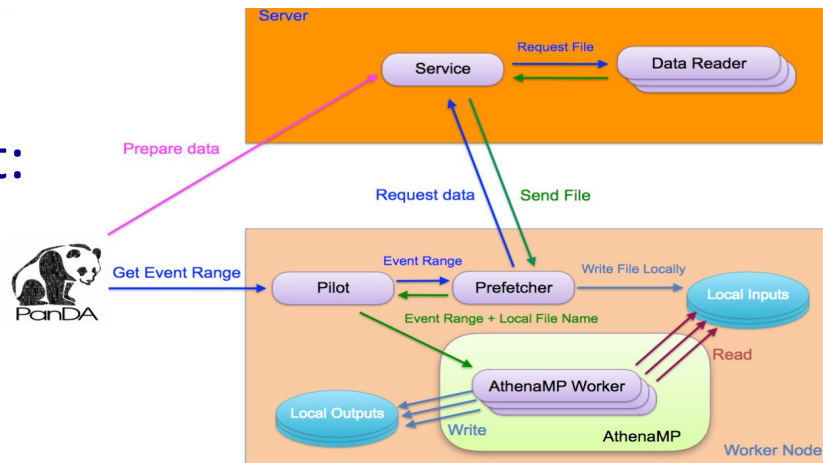


Organizing Data Lakes (Chicago)



Optimizing Workflows

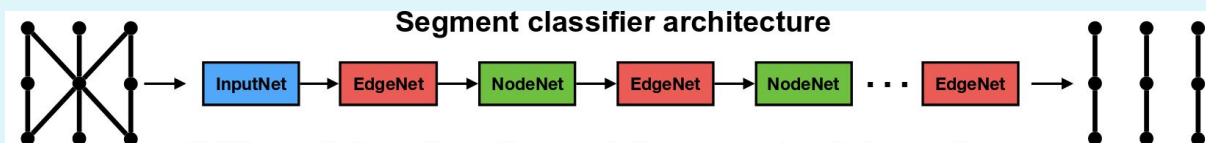
ATLAS EventService workflow concept:
schedule events rather than files.
Efficient, flexible paradigm to run
single-core to 250K-core workflows.
O(10%) CPU savings expected



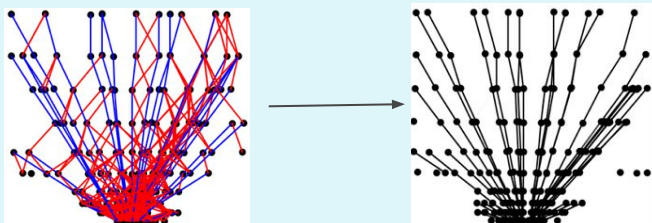
Almost stateless, so well-suited for running opportunistically on HPCs, clouds, on shared and volunteered resources.

New Ideas to Parallelize Tracking

HEP.TrkX: find $O(N)$ tracking NNs that run “trivially” on GPUs



Geometric DL: learn about objects and relationships on graphs and manifolds



Segment classifier Graph NN.

Less than 7K parameters! Trigger applications (FPGAs)

Accuracy 99.5%, Purity 99.5%, Efficiency 98.7%

Other approaches to parallelization:

TrickTrack (cellular automata), Parallel KF (data structures), ...

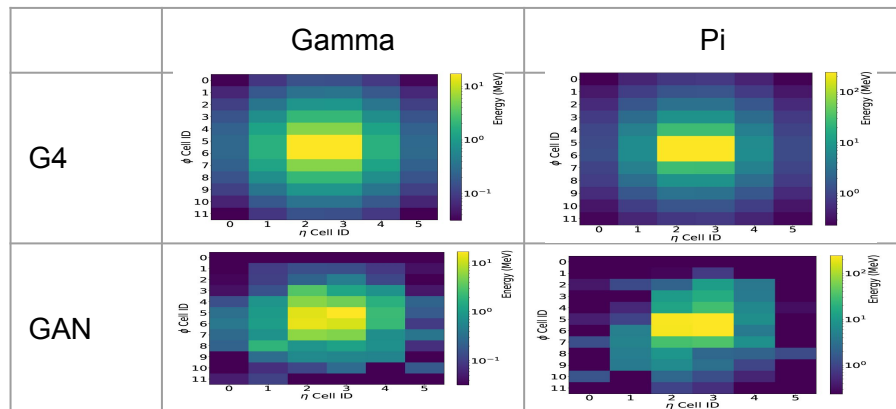
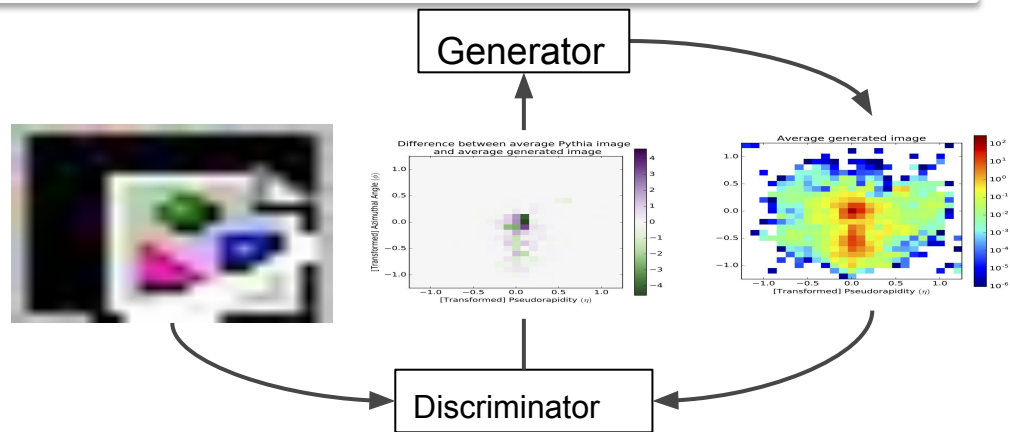


New Ideas: Simulation GANs

Fast Simulation can be used today only for a subset of analyses due to its physics performance.

Generative Adversarial Networks like [CaloGAN](#) promise fastsim-like performance with much better accuracy.

- Event Generator GANs, and full detector GANs also being investigated
- Stability and precision need study
- Non-trivial effort needed to integrate with rest of simulation framework.
- Potential for 10-20% **total** CPU savings.
- GANs run “natively” on GPUs, TPUs, etc





US ATLAS Missing Expertise

| Goal | Project | Extra Effort | Shared | Related Projects |
|--|--|--------------|-----------|---|
| Sustain HPC production | HPC Operations | ~2 FTE | Maybe | Tier 2 Operations |
| Optimization, Heterogeneous Computing and Exascale | Performance Engineering: CPU, memory, storage, networking | ~2 FTE | Yes | ESnet, A21 Early Science, ECP, NESAP, Summit Early Science |
| | Data Science | ~2 FTE | Partially | ALCF Data Sciences, NERSC DAS |

HPC centers and national labs are a natural long-term source of this kind of expertise



US ATLAS R&D Goals and Needs

| Goal | Project | Extra Effort | Shared | Related Projects |
|--|--------------------------------|---------------|-----------|--|
| Sustain HPC production | HPC Integration | ~10FTE*years | Maybe | A21 Early Science, ECP, NESAP, Summit Early Science |
| CPU Optimization, Heterogeneous Computing and Exascale | Fine-grained workflows | ~10 FTE*years | Partially | BigPanda, Harvester, yoda, EventService |
| | Parallel Reconstruction | ~20 FTE*years | Yes | aCTS, HEP.TrkX, Parallel KF, TrickTrack, IRIS-HEP? |
| | Parallel Simulation | ~20 FTE*years | Yes | CaloGAN, GeantV, G4 GPU |
| Storage Optimization | Data Organization and Delivery | ~20 FTE*years | Yes | Rucio, XCache, Cloud ATLAS, EULakes, ServiceX, IRIS-HEP? |
| Analysis Optimization | Data Analysis Systems | ~10 FTE*years | Yes | LBL LDRD, IRIS-HEP?, SCIDAC-4 |

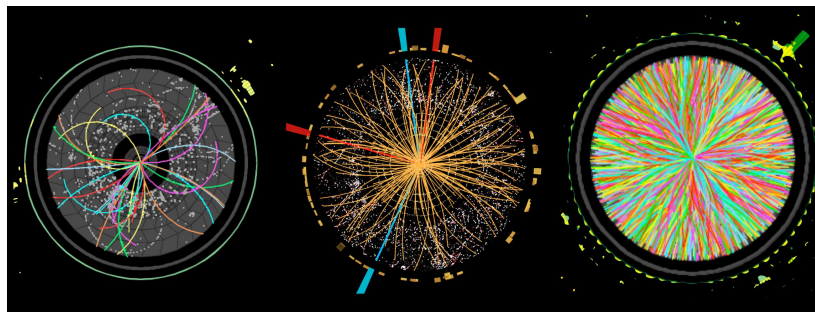
The **rough** “Extra Effort” **estimates** try to take into account the contributions of ATLAS Ops and other R&D projects. New architectures do not come for free.



Computing Extrapolation Ingredients

Most relevant LHC/TDAQ ingredients for ATLAS and CMS

| | Run 2 | Run 3 | Run 4 |
|----------------------------------|----------------------------|------------|--------------|
| Trigger Rate | 1KHz | 1KHz | 10KHz |
| $\langle u \rangle$ (colls/xing) | 35 (20-60) | 80 (60-80) | 200 |
| B Events/year | 7.8 (<i>nominal 5.5</i>) | 7.8 | 75 |





Q5: More Ingredients

- ❖ Resources per event (memory, storage, CPU)
 - estimated from upgrade studies for TDRs, software optimization efforts, and linear extrapolations where appropriate
- ❖ List of production workflows, processing frequency, Data/MC, Fast/Full simulation
 - **Naively assumed to be unchanged**
- ❖ Budget (assumed flat)
- ❖ Price/performance evolution
- ❖ ...



Where does the Gap Come from?



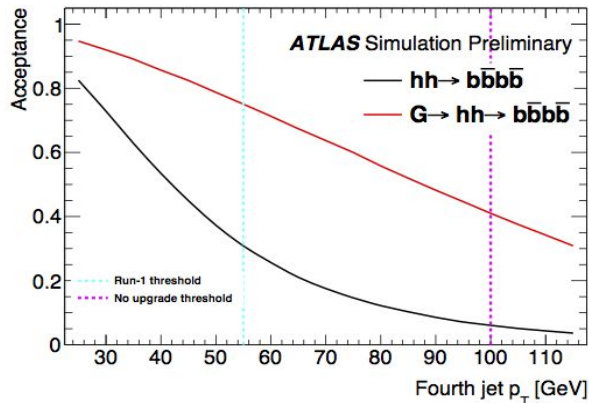
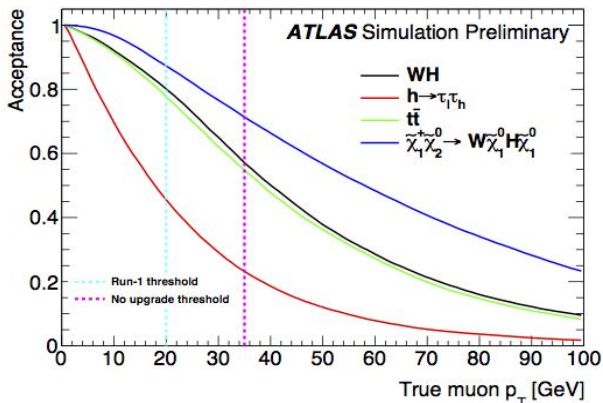
| | <i>Run 3/ Run 2</i> | <i>Run 4/ Run 2</i> | |
|--------------|-------------------------|-------------------------|-------------------------|
| No of Events | 1x | 10x | Trigger rate |
| Event Size | 3x | 5x | Pileup, #Channels |
| Reco Time | 3x | 5x | Pileup, Detector Layout |
| | | | |
| CPU | 2x | 6x | 20%/yr flat budgets |
| Storage | 1.5x | 4x | 15%/yr flat budgets |



The 10KHz Question

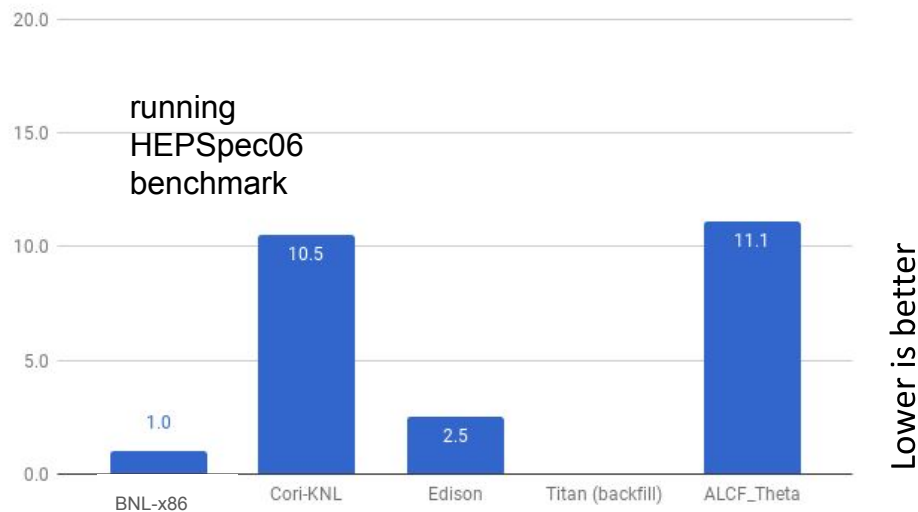
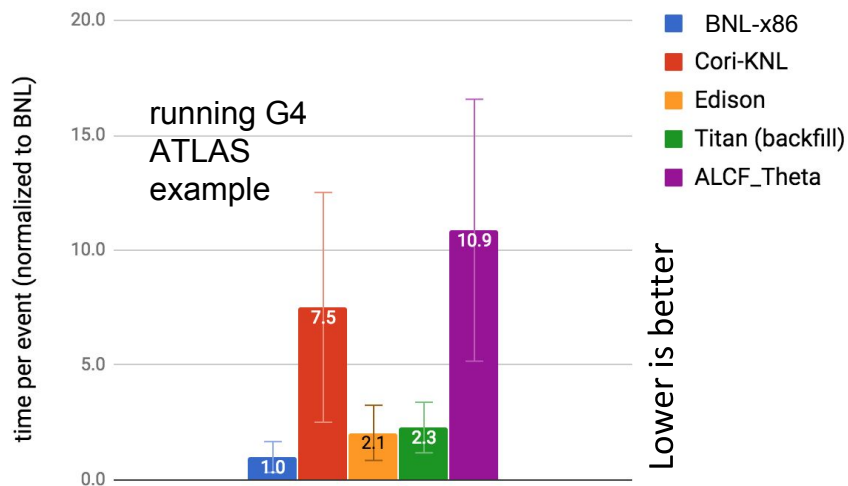
HL-LHC: Higgs physics and searches at Electroweak scales

- ❖ Ability to detect trigger on objects at $p_T \approx M_W/2$
- ❖ Raising p_T thresholds to reduce trigger rates would negate $\sim 50M\$$ investment on trigger systems.
 - Might as well keep running at Run 3 intensities





Efficiency: the Xeon Phi (KNL) Gap

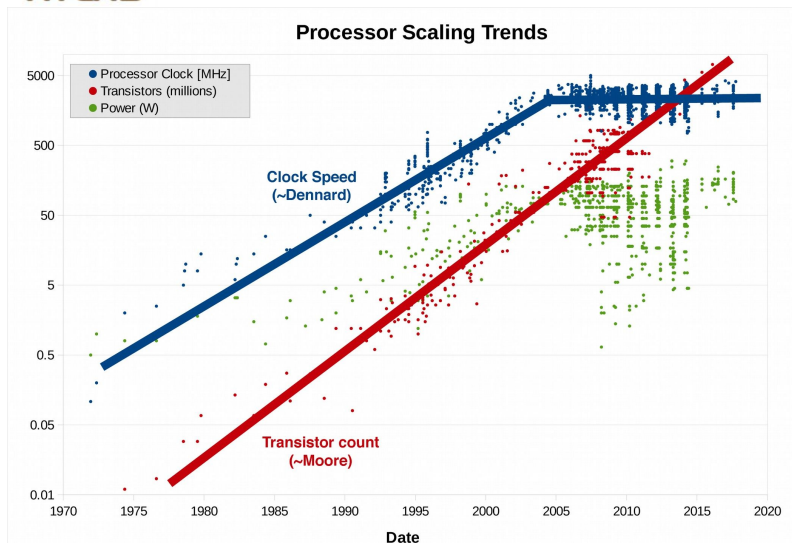


Xeon Phi (KNL) cores **~2x slower than anticipated.**

More resources, or significant vectorization effort needed

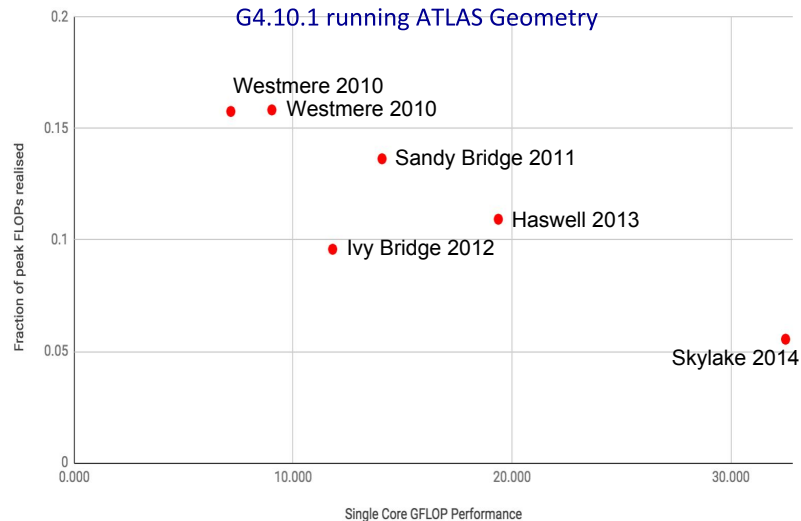


Efficiency: the Vectorization Gap



Geant4 Simulation

Private test (Graeme Stewart)
G4.10.1 running ATLAS Geometry



- ❖ Vendors kept clock speed constant since Run 1
- ❖ Exponentially more cores
- ❖ More parallelism per core (vector units)
- ❖ Cores getting simpler and simpler, Xeon Phi being the extreme example

Useable FLOPS/core **constant** 2010-2014

- ❖ Not yet able to use vector units
 - Luckily we know how to make good use of many cores
 - ATLAS routinely running 50K cores jobs on cori
 - O(25) FTE*years that went into ATLAS concurrency support paid off!

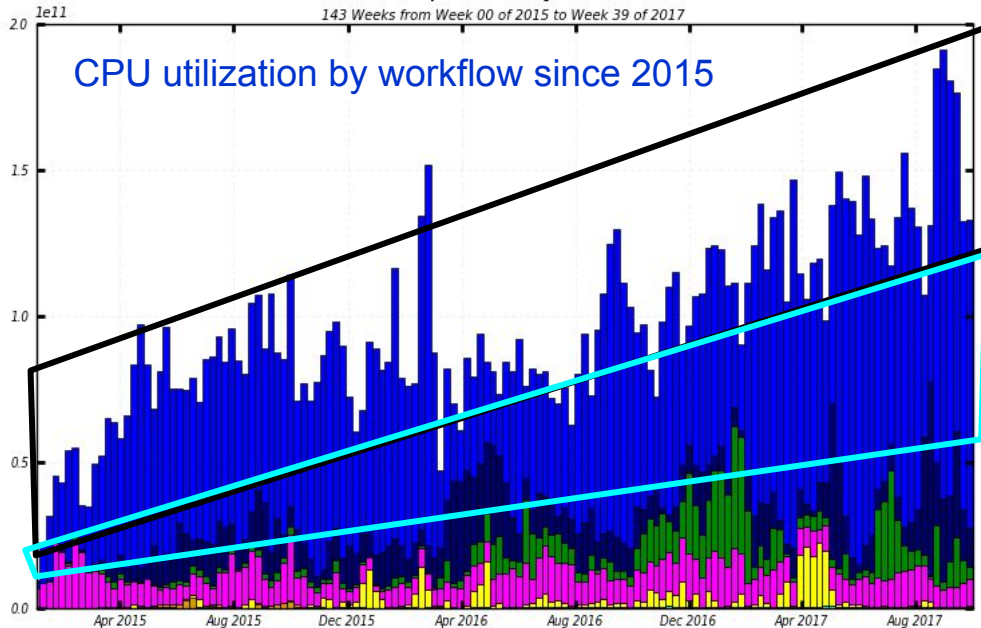


Finding new Resources: CPU Tiers



CPU consumption Good Jobs in seconds

143 Weeks from Week 00 of 2015 to Week 39 of 2017



Maximum: 191,444,990,254 , Minimum: 0.00 , Average: 95,234,981,502 , Current: 132,999,664,550

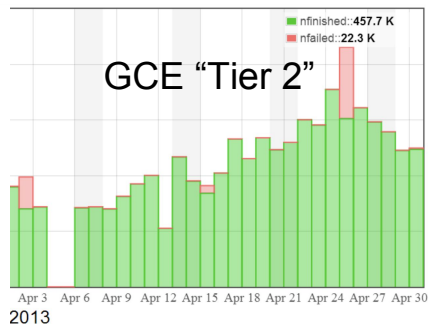
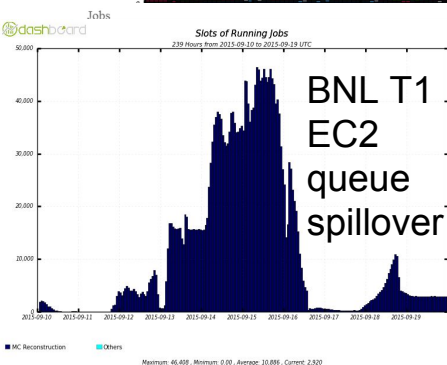
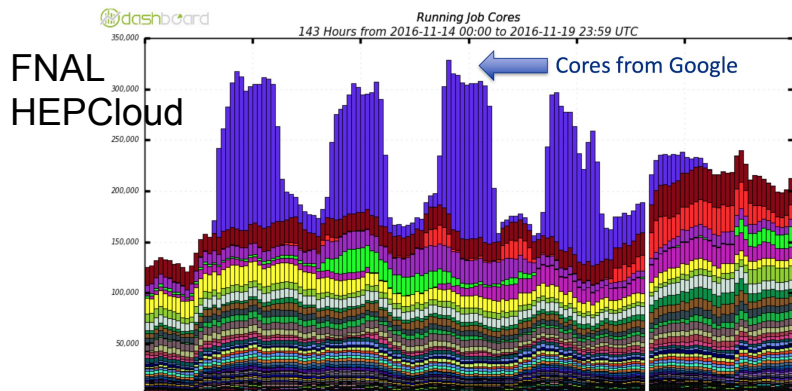
Utilization spikes. May be cost effective to offload to on-demand resources, such as Clouds

CPU-intensive (simulation)
Well suited for HPC

Data Intensive Processing
Best run on grid

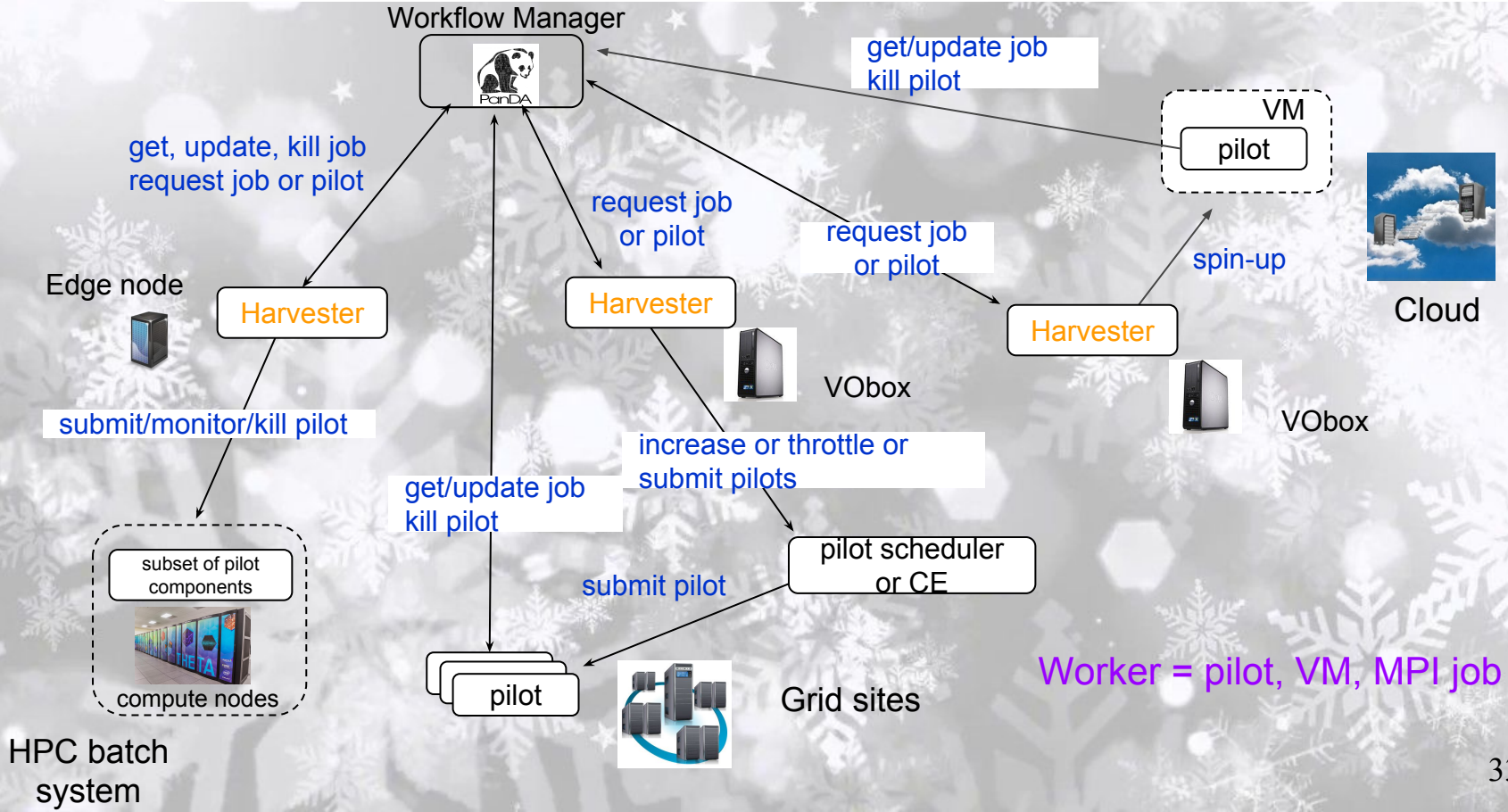


Spikes and Commercial Clouds



- ❖ ATLAS and CMS demonstrated elasticity and stability running on google GCE and amazon EC2
- ❖ Cloud resources fully integrated in their production systems
 - ATLAS Tier 0/HLT run offline as private Cloud
- ❖ Production-ready (in principle)

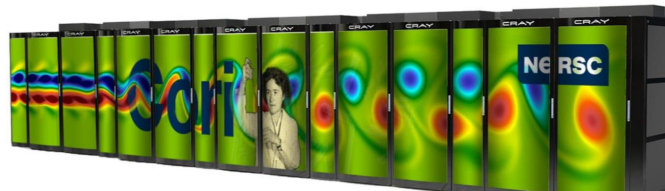
Harvester Resource Broker





HENP Analysis @ NERSC

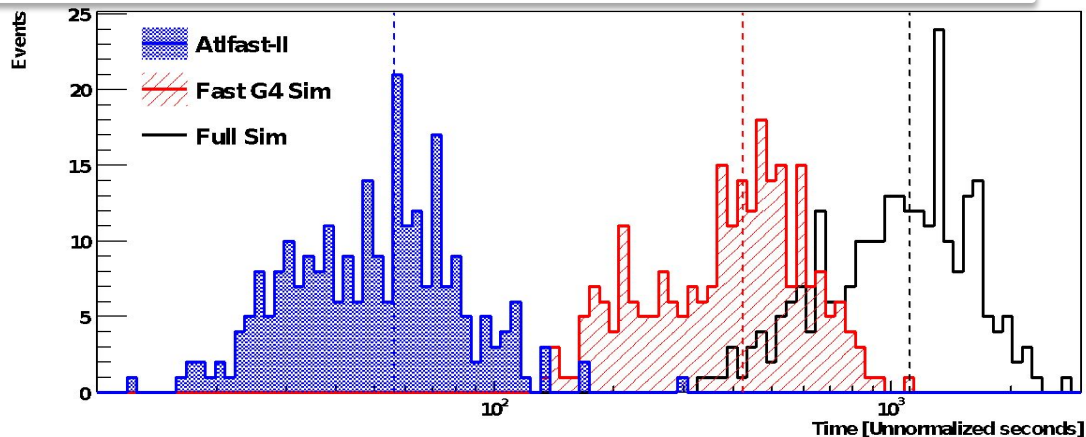
- ❖ Next-Generation Data-Intensive Analysis Framework for High Energy and Nuclear Physics on HPCs
 - Multi-year, multi-division LBL LDRD – Zachary Marshall PI
 - Involvement from ATLAS, ALICE, LUX, LZ, Daya Bay
- ❖ Goal:
Characterizing performance of data intensive workflows;
ensuring we can run them on HPCs and broader resources
 - Exploit cutting-edge HPC at NERSC (NERSC-9, with cori as a test bed)
 - Timeline for NERSC-9 matches ATLAS/ALICE Run 3



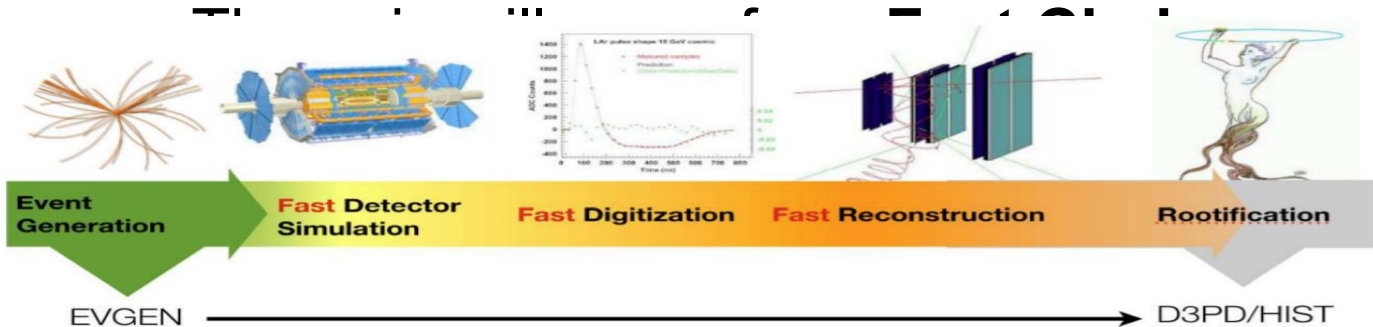


Fast Simulation and Fast Chain

Fast Simulation in Run-2
x10 faster than Full
Simulation (G4)



Fast Simulation alone, not enough particularly for HL-LHC.





ATLAS Simulation Optimization

30% of CPU dedicated to G4 Simulation

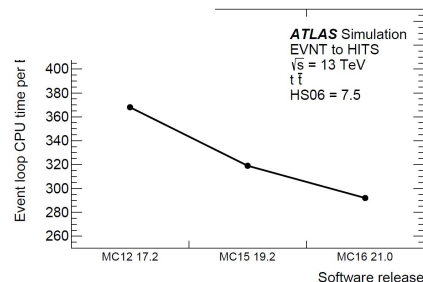
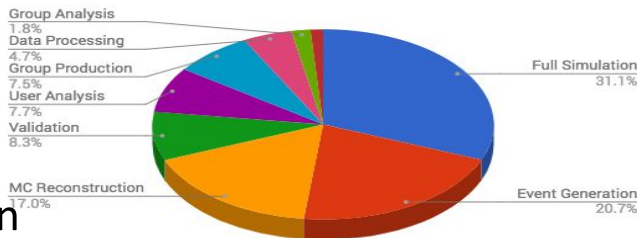
❖ G4 Technical optimizations
(5-10% gain from each one)

- Hotspot analysis & optimization
- Code inlining + refactorings
- Build one (large) statically linked G4 library
- Profile Guided Optimization (PGO)

❖ Next up are physics optimizations (10-20% each):

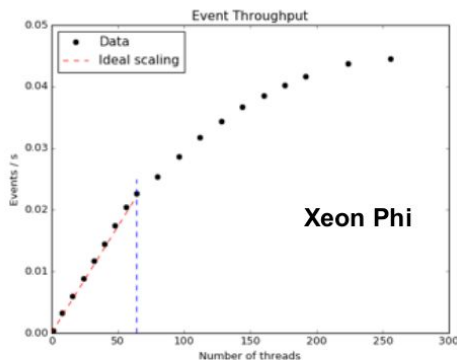
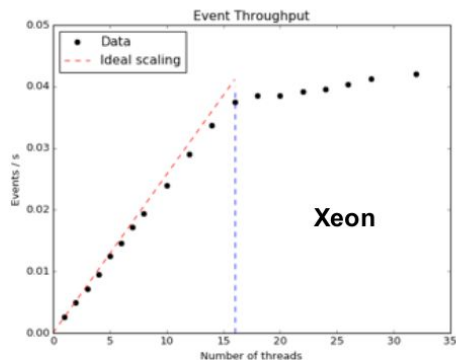
- “Russian Roulette”
 - Discard N-1 particles of a category (e.g. low energy photons in a shower), keep Nth particle with weight N
 - Introduced by CMS, will it work with ATLAS calorimeters?
- “Parallel Universes”
 - Propagate different particle category in different geometries

US ATLAS Wall Clock CPU - 2016



ATLAS MT simulation on KNL

- ATLAS simulation is being migrated to multi-threading
 - Event-level parallelism based on Geant4 and AthenaMT
 - Nearly complete full simulation configuration (G4AtlasMT) now ready
- Intel's new Knights Landing generation of Intel Xeon Phi processors is a good target for this type of application
 - Highly parallel architecture for CPU-heavy code
- G4AtlasMT shows good scaling performance on both Xeon and Xeon Phi architectures

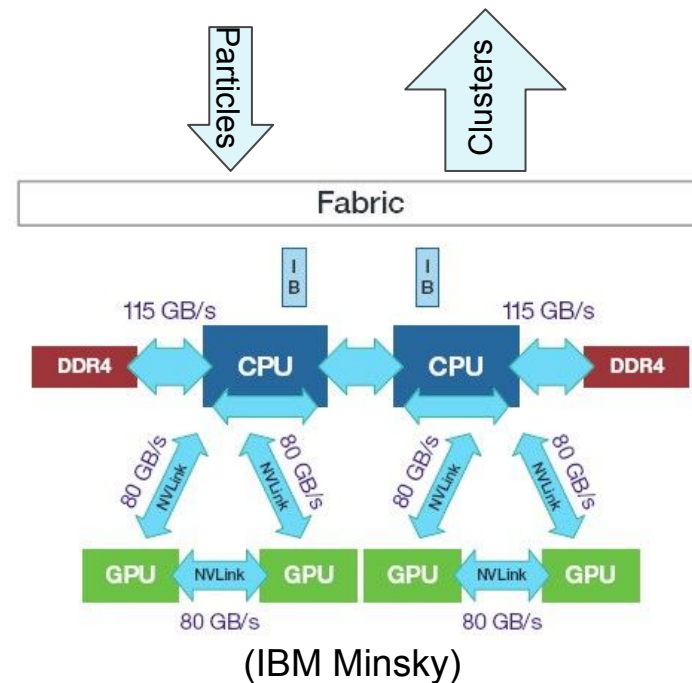




Next-Gen HPC Case Study

Can we use profitably ORNL summit:
4000 nodes: 2 IBM POWER9, 6 NVIDIA
Volta/node.

- Run G4 simulation on CPU, offload calorimeter simulation to GAN running on GPU
 - may take 100s events to offload efficiently to one Volta, which will then run GAN <1s.
 - Power 9 cores will take $O(1000)$ s to run the rest G4 simulation for the 100 events. Not fast enough to keep GPU busy.
 - How about using the CPU to stream input data coming from e.g. BNL T1 to , and stream GAN-simulated clusters back?
 - Will be hard to keep load balanced
 - **Worth giving it a try**



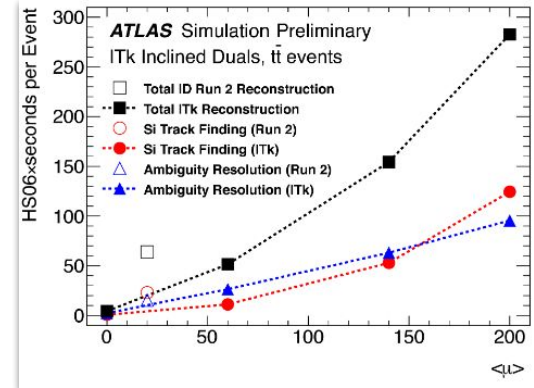
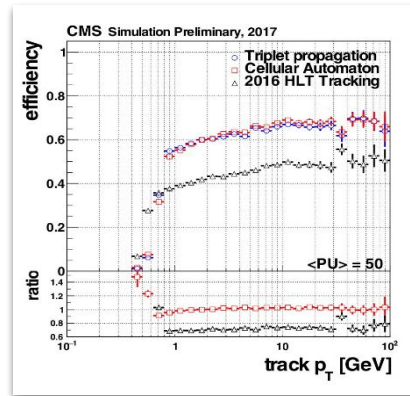
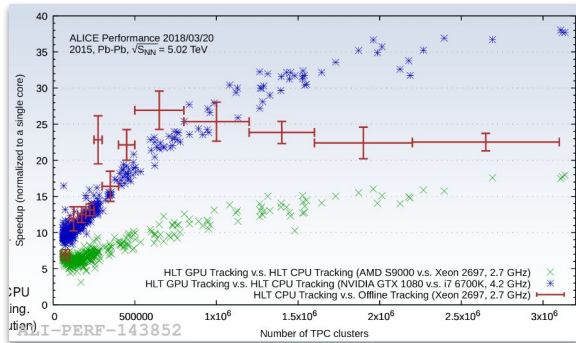


Optimizing Tracking

Tracking CPU scales $> O(N^2)$ with intensity.

- ❖ Likely limiting factor for HL-LHC precision physics
- Iterative, branchy, highly optimized algorithms.

Trying new approaches, including detector “co-design”





More Storage Optimization Ideas

ServiceX (Chicago)

- ❖ Transform data on the fly from storage optimized (e.g. Ceph-friendly) to client-optimized (e.g. ROOT) and vv

Physics-aware compression (ALICE, LHCb)

- ❖ Drop information (clusters, tracks, truth,...) from analysis streams on an event-by-event basis
 - May be combined with virtual data concept to produce missing information on demand



HSF Community White Paper

arXiv:1712.06982v3 [physics.comp-ph] 11 Feb 2018

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

HEP Software Foundation¹

ABSTRACT: Particle physics has an ambitious and broad experimental programme for the coming decades. This programme requires large investments in detector hardware, either to build new facilities and experiments, or to upgrade existing ones. Similarly, it requires commensurate investment in the R&D of software to acquire, manage, process, and analyse the shear amounts of data to be recorded. In planning for the HL-LHC in particular, it is critical that all of the collaborating stakeholders agree on the software goals and priorities, and that the efforts complement each other. In this spirit, this white paper describes the R&D activities required to prepare for this software upgrade.

arXiv

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

In 2017 the HEP Software Foundation produced a [roadmap white paper](#) on the software and computing challenges that will be faced during the next decade.

The CWP Roadmap can still be signed by members of the community who endorse it. Please use contact the [CWP Ghost Writers](#) to add your name. We very much encourage you to do this, to show the breadth of the community's support for the roadmap.

Community White Paper Reports

The roadmap summarised reports from fourteen working groups who studied the challenges in their sub-domains. All of the reports produced during the Community White Paper process are listed below. Working groups are in the process of [finalising](#) and [uploading](#) their work to arXiv.

| Paper | Report Number | Link |
|---|-----------------|---|
| CWP Roadmap | HSF-CWP-2017-01 | arXiv |
| Careers & Training | HSF-CWP-2017-02 | ShareLaTeX |
| Conditions Data | HSF-CWP-2017-03 | Google Doc |
| Data Organisation, Management and Access | HSF-CWP-2017-04 | Overleaf |
| Data Analysis and Interpretation | HSF-CWP-2017-05 | Dropbox |
| Data and Software Preservation | HSF-CWP-2017-06 | Google Doc |
| Detector Simulation | HSF-CWP-2017-07 | Summary Google Doc ; Google Doc |
| Event/Data Processing Frameworks | HSF-CWP-2017-08 | Google Doc |
| Facilities and Distributed Computing | HSF-CWP-2017-09 | Google Doc |
| Machine Learning | HSF-CWP-2017-10 | ShareLaTeX |
| Physics Generators | HSF-CWP-2017-11 | Overleaf |
| Security | HSF-CWP-2017-12 | See section 3.13 of roadmap |
| Software Development, Deployment and Validation | HSF-CWP-2017-13 | arXiv |
| Software Trigger and Event Reconstruction | HSF-CWP-2017-14 | Summary Google Doc ; Google Doc |
| Visualisation | HSF-CWP-2017-15 | Google Doc |



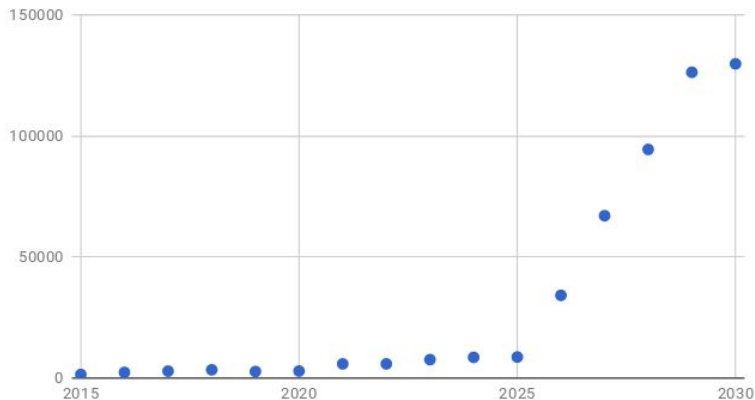
Most Recent Model Predictions

Model used for spreadsheet differs from ATLAS preliminary plots

❖ Main differences

- 60 days of running in 2026
- Kept constant 1.3 Sim/data ratio throughout
- Fast/full sim ratio == 1 in Run 3 and Run 4

T1 + T2 CPU (kHS06)



T1 + T2 Disk [PB]

