The role of HPC for HEP computing - Situation and Outlook

Torre Wenaus (BNL)

Scientific Computing Forum Meeting
CERN
September 20 2018

BROOKHAVEN

CERN

Outline

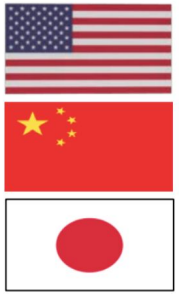
- The HPC landscape today and in the coming years
- HPC usage in the experiments
- HPC directed software development
- Towards exascale in 2021 (the first such machine arrives then), and ultimately HL-LHC

Caveat: Somewhat ATLAS-centric and US-centric, but I did get input (thanks!) from CMS, LHCb and ALICE

That said, ATLAS's longstanding emphasis on HPC utilization makes for a well informed perspective, and HPC issues are arguably the most challenging in the US

HPC global picture over the next few years

Global Picture HPC



- **USA**, 4 pre-exa and 3 exascale systems in 2018-2022
- **China,** exascale in 2021?
- **Japan**, exascale in 2022

2 pre-exascale by 2020 and two exascale systems by 2022/2023

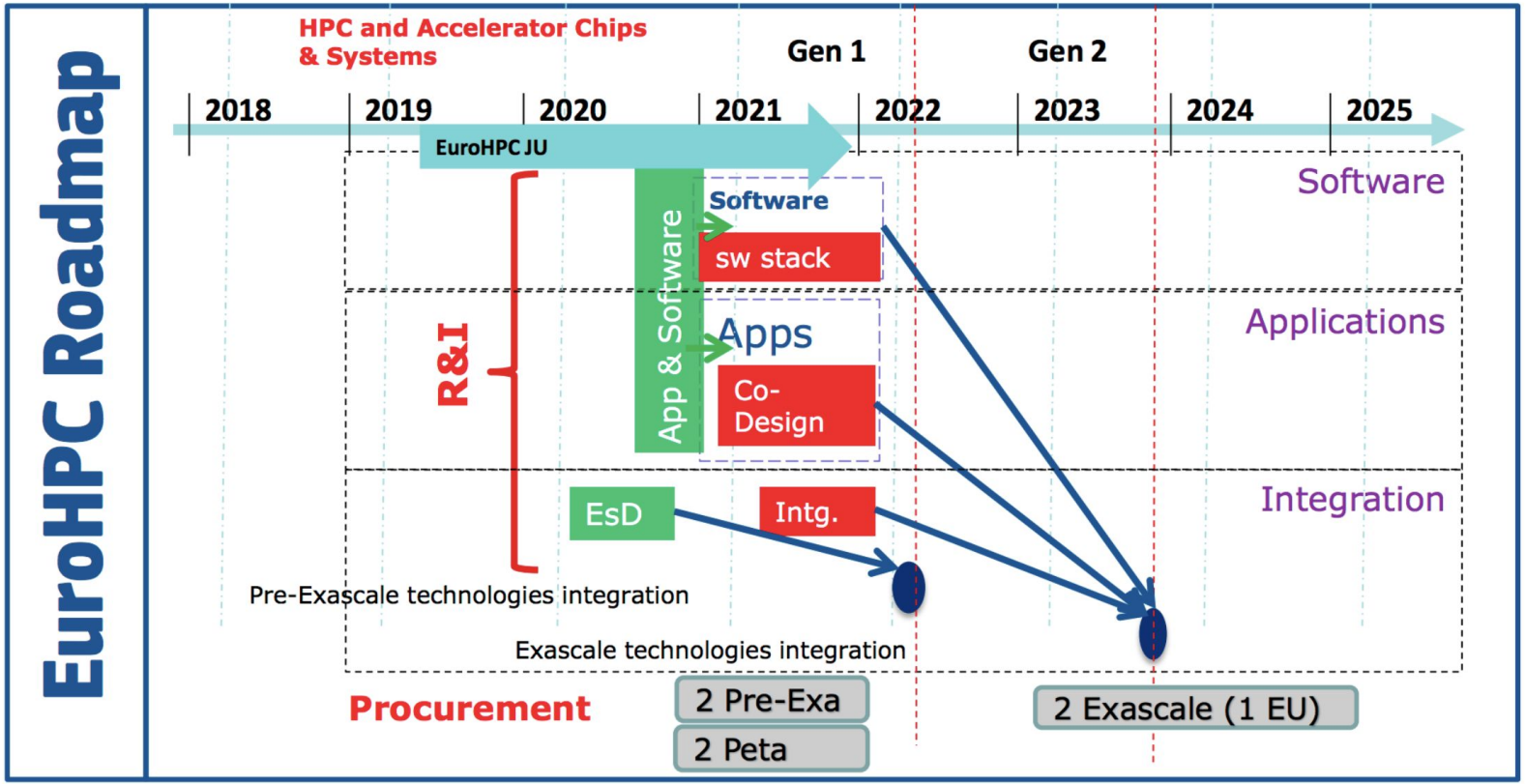
Hybrid HPC/Quantum infrastructure

emerging "computing architectures" (quantum/neuromorphic)

novel applications in key areas (Cybersecurity, AI)

Andrej Filipcic, June 2018, WLCG MB

EuroHPC timeline



Andrej Filipcic, June 2018, WLCG MB

ETP4HPC - 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.

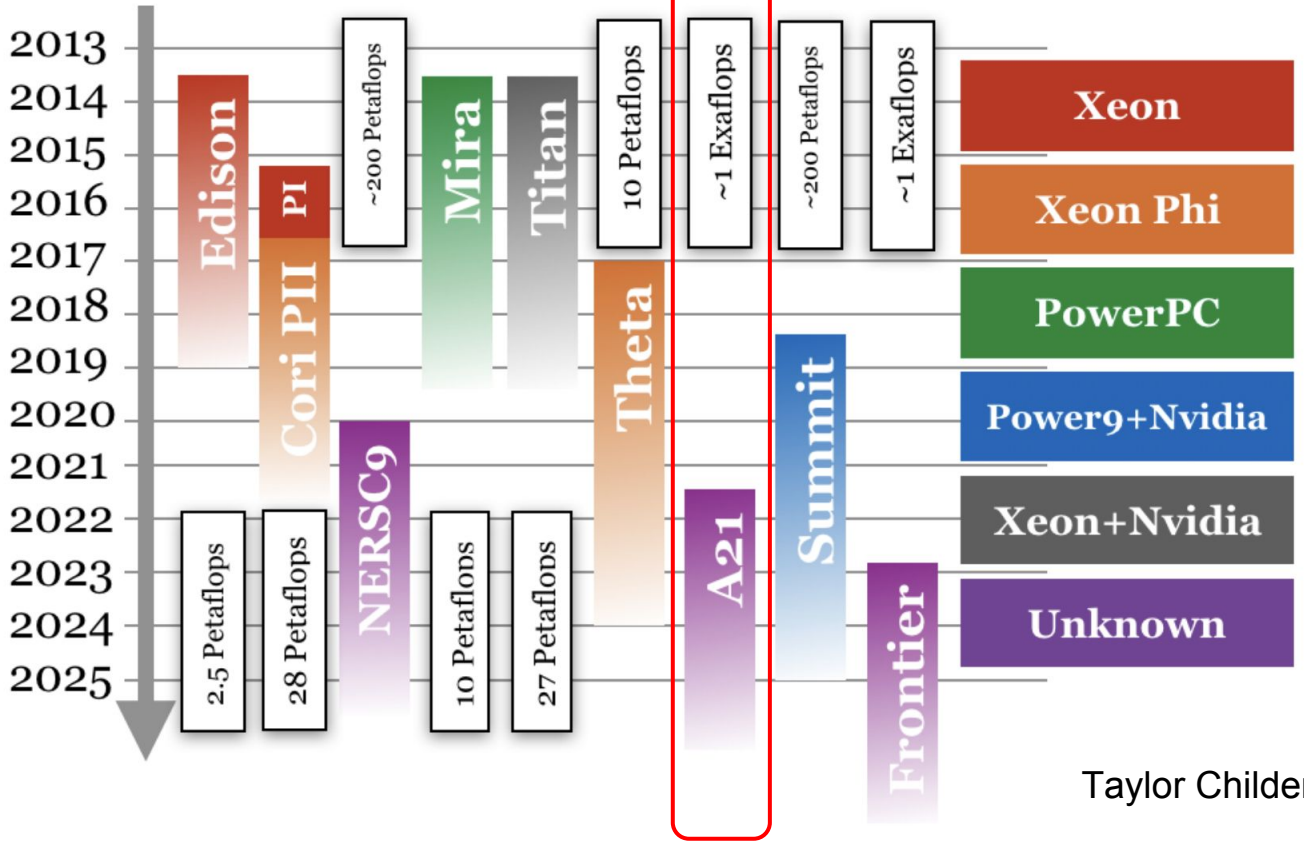
- CERN is one of top EU scientific endeavours
- But in the document, HEP requirements are not mentioned, apart from lattice QCD
- Future EU HPC centers will be extended to data processing facilities – eg ESiWACE needs large storage, transfers and remote processing (distributed systems)
- Most intensive-computing 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

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

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

U.S. DEPARTMENT OF ENERGY | Office of Science

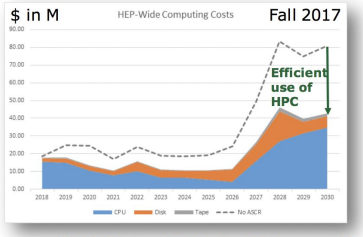
DOE HEP Status at HEPAP - May 2018 28

We must use them properly (use the accelerators)

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!**



U.S. DEPARTMENT OF ENERGY | Office of Science

DOE HEP Status at HEPAP - May 2018 29

Jim Siegrist, HEPAP meeting, May 2018

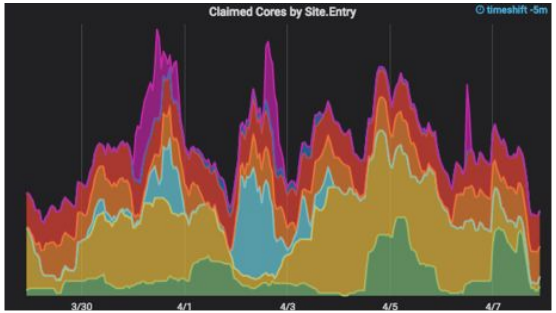
Similar themes in Europe and US

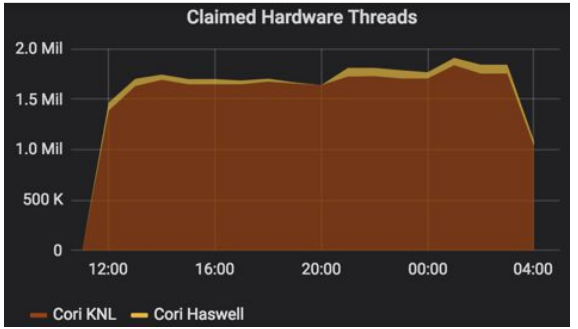
- The HPC planners acknowledge us and the importance of our compute intensive science
- They are building (soon!) **exascale facilities** that they expect us to use
- It is not clear that they are taking the **data intensive requirements** of our computing into consideration in system design
- The growth 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 (supplementary slide)
    - ***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?

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





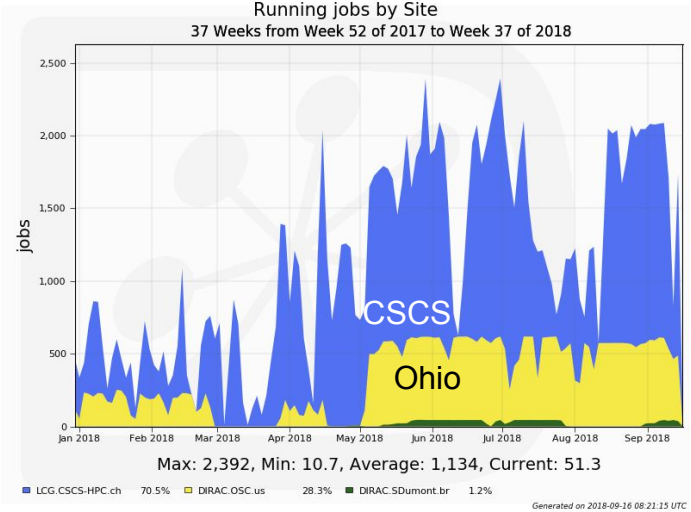
HPC usage in CMS: Europe

“Friendly” defined:
External connectivity
CVMFS for software installation
Virtualization present
x86 architecture, adequate memory
Workable security

- **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

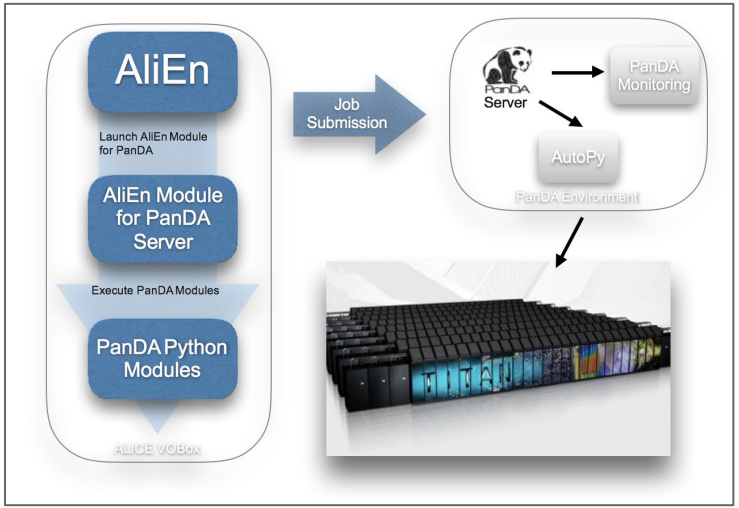
HPC usage in LHCb

- 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

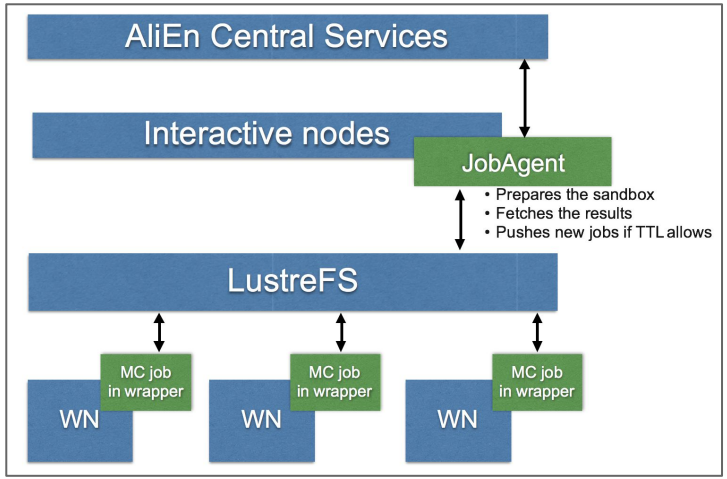


HPC usage in ALICE

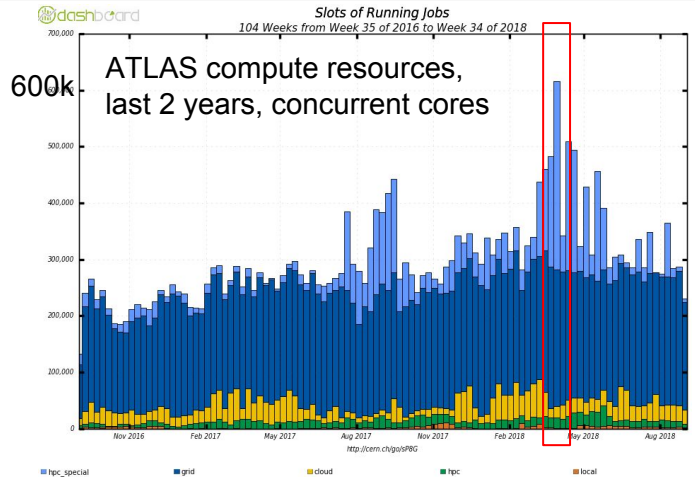
- Long time collaborators with ATLAS/PanDA on using Titan at Oak Ridge for HEP & NP
- AliEn - PanDA integration leverages PanDA services at Titan
  - ALICE jobs submitted via PanDA



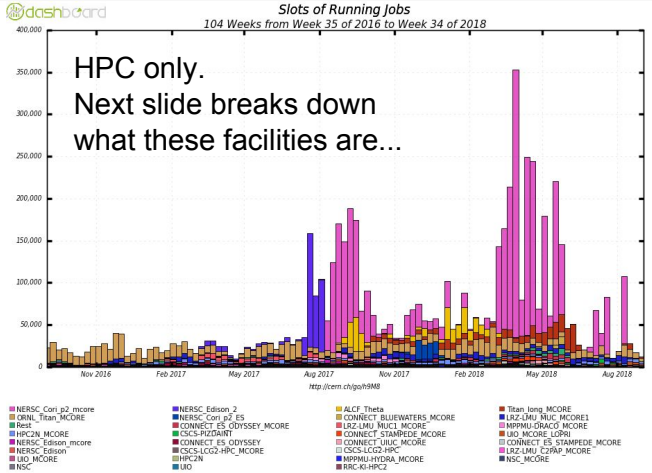
ALICE services for Titan



HPC usage in ATLAS



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

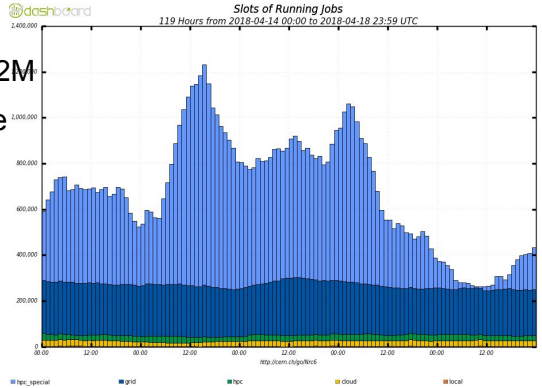


Light blue: "special" HPCs, where special means big, difficult to use, US DOE
Dark blue: the grid
Yellow: cloud resources including (dominantly) HLT
Green: "regular" HPCs, meaning easier to use, European or US NSF

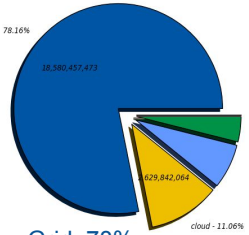
1.2M

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

Our workload management system is highly scalable!

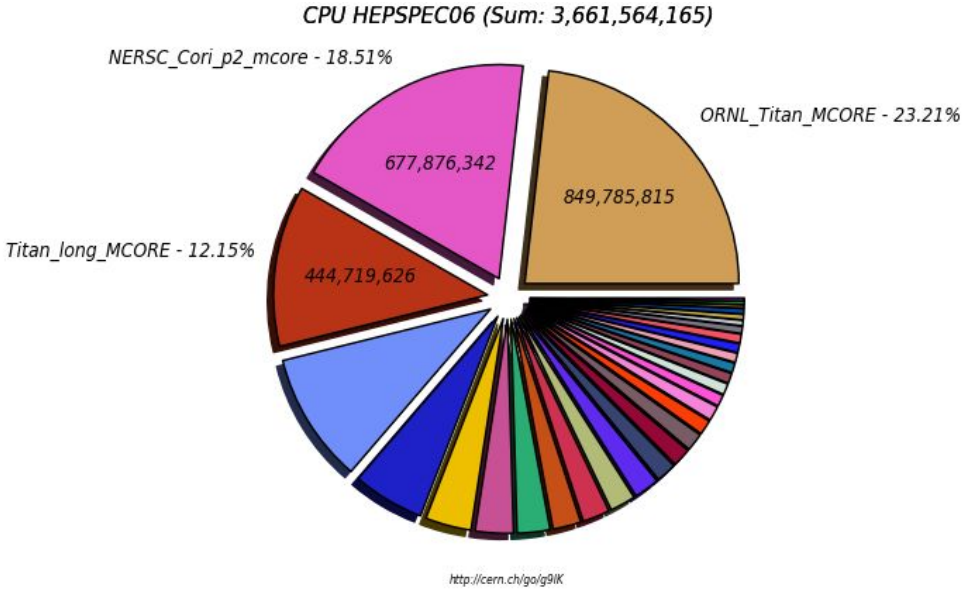


CPU HS06 shares, last year



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

HPCs in ATLAS



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

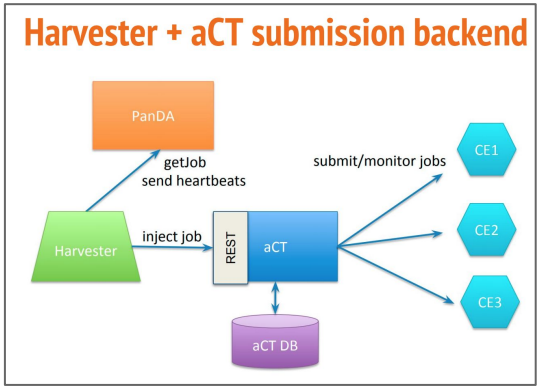
All in routine production, mostly Geant4 MC simulation

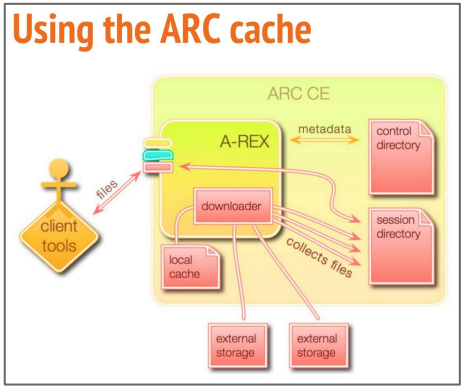
HPC directed software development

- HPC characteristics drive **HPC directed software development**
  - Often not exclusively, but with a strong HPC motivation
- Wide **variability in the "friendliness"** of HPC sites; the more friendly, the less ad hoc work
- The **importance of utilizing HPCs**, today but even more so in the future, motivates investment (unfriendly often correlates with big)
  - With **some of the effort contributed externally** ("use our HPCs" is not always an unfunded mandate, even if often an under-funded mandate)
- Without such development, HPC utilization would be too inefficient, too wasteful of the machine's capabilities, too operationally burdensome
- Examples
  - **Multithreaded frameworks:** a prime motivator and benefit of MT frameworks is enabling use of highly concurrent, accelerator-equipped architectures like HPCs
  - **I/O:** high concurrency, data intensive workflows, limited memory make for HPC specific challenges
  - **Nordugrid:** the longest lived grid infrastructure, all European HPCs (at least used by ATLAS) are integrated via the ARC software
  - **Event service** (ATLAS): high HPC utilization via fine grained workflows
  - **Harvester** (ATLAS): uniformity in interfacing to HPCs and grid resources
  - **Accelerator (e.g. GPU) utilization**: the outstanding problem to solve

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





ATLAS Event Service (AES)

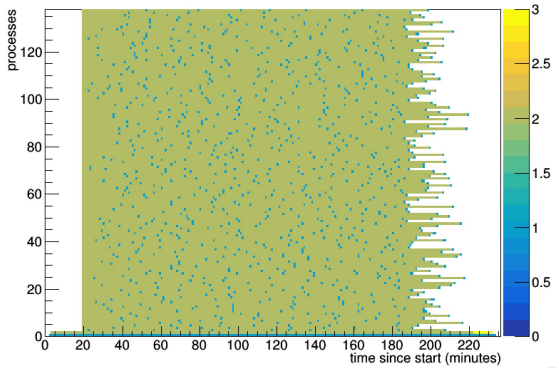
AES produces 18% of ATLAS MC events today, and growing

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

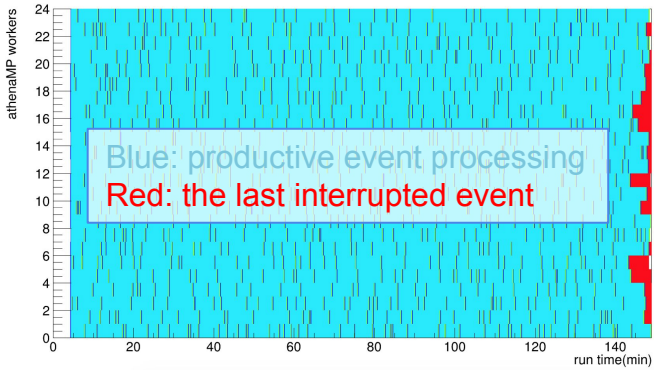
With event service, each core is allocated events to process until the scheduler slot ends.

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, no AES



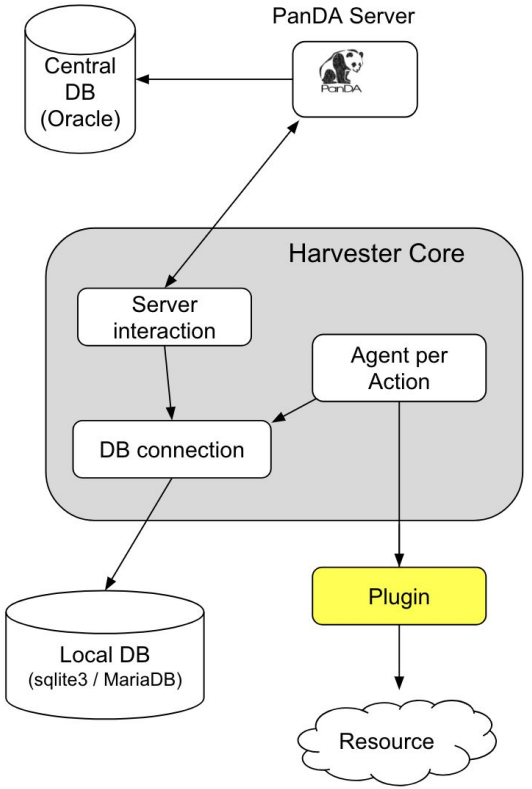
NERSC utilization per core with AES



AES elastically fills all cores for the full job lifetime

Harvester

- 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*



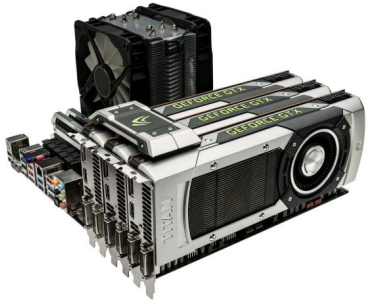
Accelerator utilization

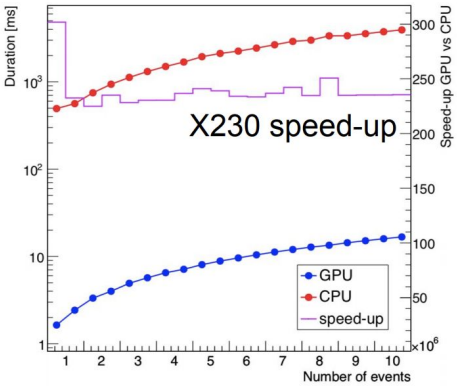
- A big topic today and will be bigger in the future
- An example of why: the DOE tells us (LHC computing) **we must utilize the accelerators if we're to be allowed onto exascale machines**
  - Not unreasonably; most of their power is in the accelerators
- We're **not in a position today to use accelerators at large scale in offline**
  - ATLAS has no offline production applications today that utilize GPUs
  - ALICE will have GPU based online track reco for Run-3
  - CMS is partially rewriting tracking and calo sw to use GPUs, and studying larger scale adoption in the computing model studies underway (ECoM2X)
- The DOE position prompted new activity, again an ATLAS example…
  - In July 2018 the new "HL-LHC Computing" activity area in US ATLAS organized a workshop that brought together HPC experts from BNL's Computational Science Initiative (CSI) and a team of senior ATLAS software developer / physicists
    - Look for **GPU and ML applications for exascale** and identify projects

How to use GPUs? … a recent HSF slide

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

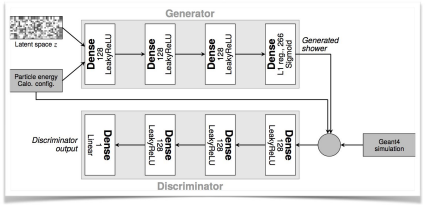
Leveraging Exascale for ATLAS

- **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

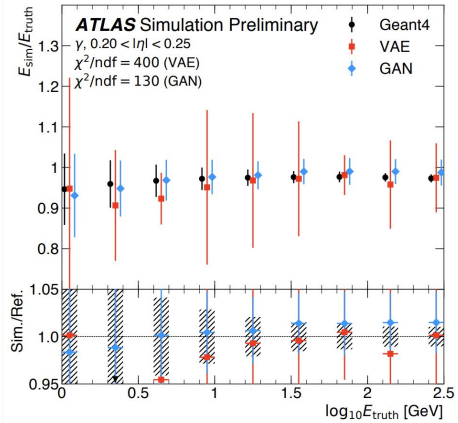
A possible work program towards exascale

- 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
  - Nearer term (Run-3), possible element of building a complete *fast chain* simulation + reco/digi that would (crucially) minimize storage needs
- 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*

GAN



Early results



Every experiment is exploring ML in calo and tracking

Generative Models @ LHC

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

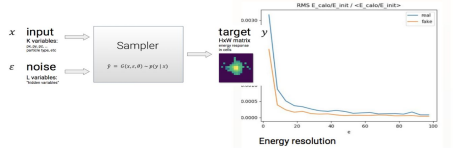
Amir Farbin, July 2018

Generative models for fast cluster simulation @ALICE

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)	166.155	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



Just concluded
https://www.kaggle.com/c/trackml-particle-identification

HEP.TrkX

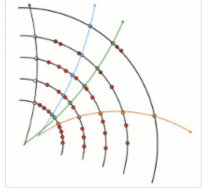
Cross experiment, DOE supported
https://heptrkx.github.io/

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

Finally

- **HPCs are here** in HEP computing, here to stay and grow
  - They require **dedicated investment** of effort
  - We require **stable allocations**, not just backfill, to make the investments pay; resource acquisition model is important
- They bring **accelerators** like GPUs with them, which we can't leave idle
- Particularly crucial for ATLAS and CMS: on an HL-LHC timescale, major funding agencies are mandating a **very high profile for HPCs**
  - Beginning with a mandate to use the first exascale machine in the US in 2021
- We've done the preparatory work for using accelerators in simu/reco, building multithreaded frameworks, but we seem **far from applications that are exascale ready**
- Are **machine learning applications** -- or at least their training component -- the most achievable path to apps for the first exascale machine in 2021?
  - Can we sketch out now more ambitious objectives for Run-4?
- Many questions to answer: we must **boost our development efforts and enlist CS experts** to help answer them

Thank you

- Thank you to all who contributed materials, discussions and -- most of all -- work to get productivity out of HPCs
- Especially…
  - Doug Benjamin, Tommaso Boccali, Brian Bockelman, Concezio Bozzi, Paolo Calafiura, Simone Campana, Taylor Childers, Davide Costanzo, Kaushik De, Alessandro Di Girolamo, Johannes Elmsheuser, Andrej Filipcic, Rob Gardner, Heather Gray, Wen Guan, Alexei Klimentov, Markus Klute, Eric Lancon, Tadashi Maeno, Abid Malik, Nicolo Magini, Danila Oleynik, Sergey Panitkin, Srini Rajagopalan, Stefan Roiser, Rod Walker, Jack Wells, Wei Yang

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 (just concluded)

Supplemental

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

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

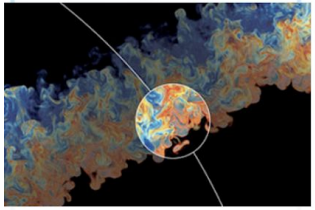
...and got 80M hours each at ALCF and OLCF from ALCC in 2018

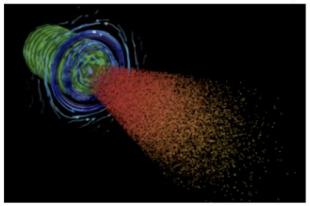
Then there is the backfill time on Titan 195 Mhrs were used in 2017

Doug Benjamin, March 2018

US DOE's ECP program

National security	Energy security	Economic security	Scientific discovery	Earth system	Health care
Next-generation, full-system stockpile stewardship codes	Turbine wind plant efficiency	Additive manufacturing of qualifiable metal parts	Cosmological probe of the standard model of particle physics	Accurate regional impact assessments in Earth system	Accelerate and translate cancer research
Reentry-vehicle-environment simulation	Design and commercialization of SMRs	Urban planning	Validate fundamental laws of nature	Not us. L-QCD.	
Multi-physics science simulations of high-energy density physics conditions	Nuclear fission and fusion reactor materials design	Reliable and efficient planning of the power grid	Plasma wakefield accelerator design	analysis and catalytic conversion of biomass-derived alcohols	
	Subsurface use for carbon capture, petroleum extraction, waste disposal	Seismic hazard risk assessment	Light source-enabled analysis of protein and molecular structure and design	Metagenomics for analysis of biogeochemical cycles, climate change, environmental remediation	
	High-efficiency, low-emission combustion engine and gas turbine design		Find, predict, and control materials and properties		
	Carbon capture and sequestration scaleup		Predict and control stable ITER operational performance		
	Biofuel catalyst design		Demystify origin of chemical elements		





[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