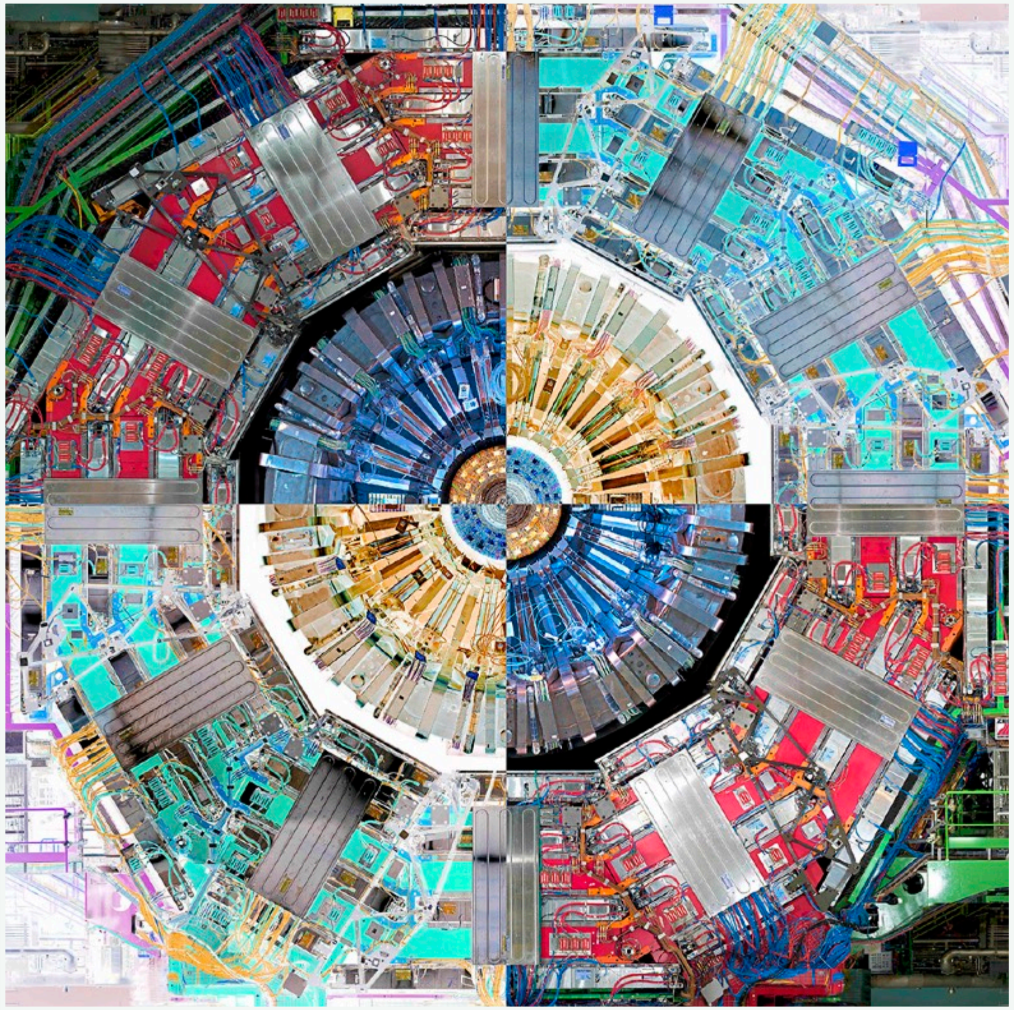
Fermilab

Managed by Fermi Research Alliance, LLC for the U.S. Department of Energy Office of Science



CMS Computing for HL-LHC & Exascale Initiatives

Lothar A. T. Bauerdick
Fermilab PAC Meeting
July 18, 2018

LHC Computing

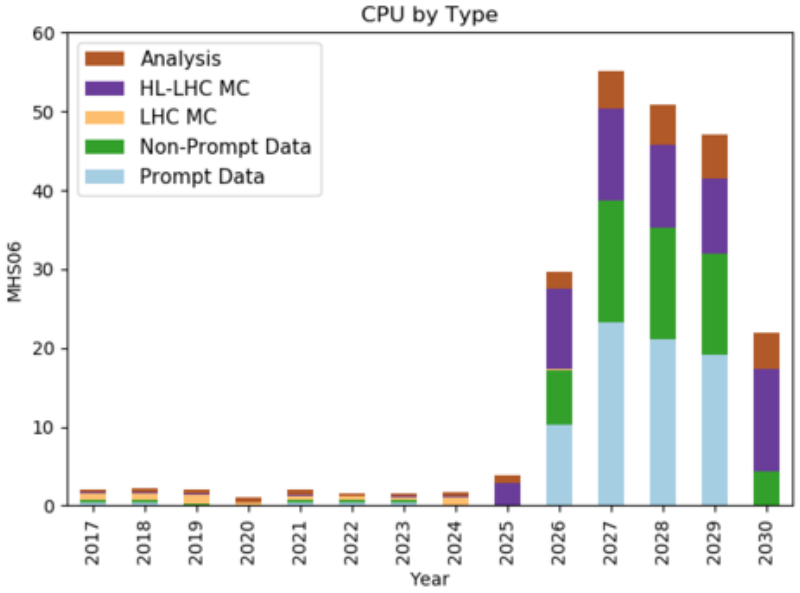
- Computing for LHC Runs 1 and 2 is a big success
  - pioneering a world-wide distributed computing model
    - ★ community shown to be very capable of building an adaptable and performant service, building on and integrating national and international structures
  - successful scaling up of computing capabilities by several orders of magnitude
- Significant upgrade of computing capabilities for Runs 2 & 3
  - **evading the "breakdown of Moore's law":** effective use multi-cores
    - ★ multi-threaded framework and improving software to gain >order of magnitude in computing efficiency
  - **enabling distributed data federations:** vastly improved networking, including transatlantic
    - ★ transparent over-the-network access to data
  - this makes possible the **wide sharing and on-demand provisioning of computing resources**
    - ★ opening the door to HPC allocations, commercial clouds through the HEPcloud project
- So far, computing has not been a limiting factor for LHC physics!
  - however, computing remains a significant cost driver for LHC program
    - ★ since its start in 2003, spent well above $200M on CMS computing in the US alone (USCMS Operations)
- So, what's left to do for HL-LHC?

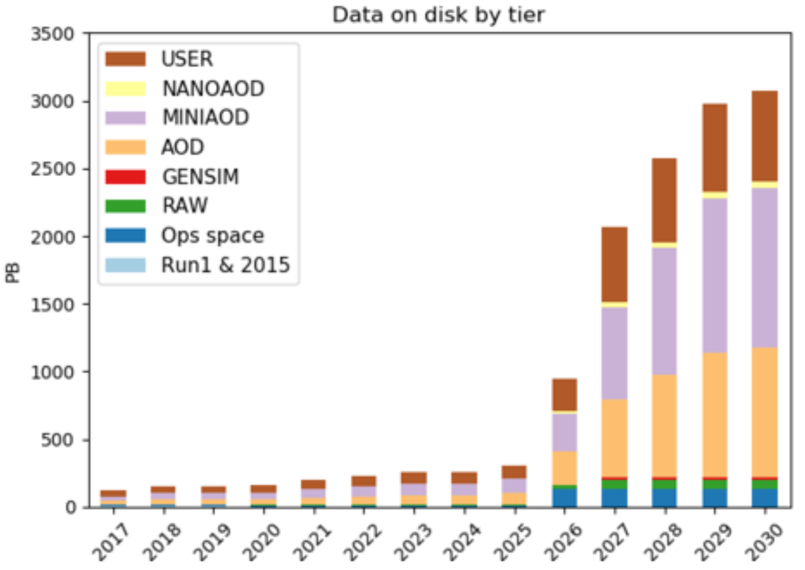
Extrapolating CMS Computing Resource Needs for HL-LHC

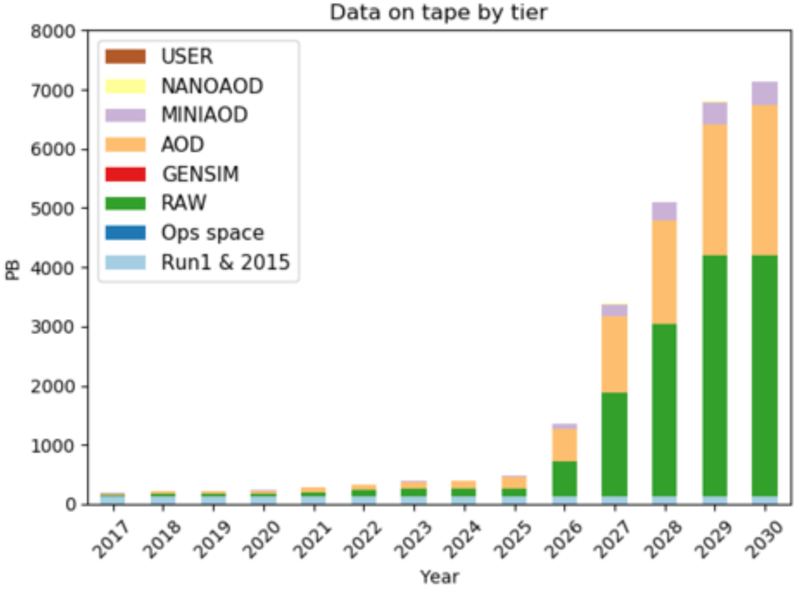
- How can we assess the CMS computing needs, >8 years out?
  - “naïve” extrapolations from Run2 and Run3 computing models to HL-LHC look daunting
- To gain an initial “upper limit” of HL-LHC computing resource needs:
  - estimates of event size and complexity due to increased event pile-up, the upgraded detector (high-granularity tracking, calorimetry, timing), and increased trigger rates
    - ★ up to 10 kHz trigger rates, and event pile-up increasing from 60 (Run3) to 200
    - ★ RAW event size estimated to increase ~x4, to ~ 7.4 MB
  - assume “known” (non-disruptive) advances in technologies and “traditional” software improvements as a reference — for R&D to enable experiments go beyond that baseline
    - ★ expect improving RECO times by ~10% per year, as we achieved during Run1 and Run2
  - assume radically smaller (and smarter) data formats for analysis, only kByte/event
    - ★ CMS defines a nanoAOD that could be effective for 50% of all physics analyses
  - assume an intelligent pile-up simulation in MC
    - ★ “pre-mixing” expected to make I/O load feasible, and to reduce the CPU need by a factor of 25

“Naïve” Extrapolations: Daunting!

- HL-LHC scales for CMS computing
  - Exa-byte scale disk and tape storage (x50 w/r to now)
  - CPU needs 5M cores (x20 w/r to now)
  - transfer of exa-byte-sized data samples across the Atlantic at 250-500 Gbps (ESnet now has allocated 40Gbps transatlantic for the LHC)
- These estimates got DOE’s attention…







Preparing for HL-LHC Software and Computing Needs

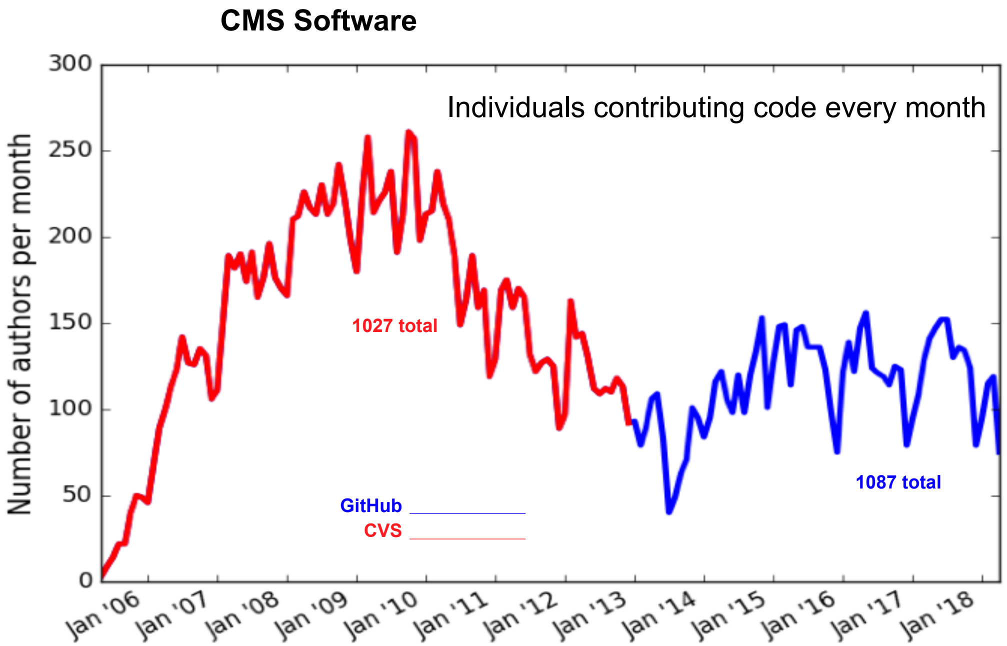
- In the HL-LHC era, computing could indeed become a limit to discovery
  - unless we can make significant changes that will help to moderate computing and storage needs, while maintaining the physics goals
- To develop and agree on a R&D program towards HL-LHC computing, last year the global HEP community produced a “Community White Paper”
  - WLCG submitted a Computing Strategy paper, to be reviewed by LHCC in the coming months
  - Experiments’ formal Computing Technical Design Reports are expected only much later (in 2022?)
- Fermilab and US CMS are vigorously participating in this process
  - SCD and US CMS R&D activities, funded by Ops Program, DOE SciDAC, and NSF grants
  - US CMS and Altas now have the outlines of a work program for both the labs and the universities, how to address the HL-LHC software and computing challenges
    - ★ this resulted in a conceptualization and then a proposal for a 5-year program of ~$5M/year, a NSF “Software Institute” (IRIS-HEP), that is expected to start this year
    - ★ and a complementary DOE sponsored program for Fermilab and its university partners, building on the strength of the SciDA, SCD, and USCMS activities and capabilities

Areas That Must be Addressed (WLCG Strategy)

- Modernizing Software
- Improving Algorithms
- Reducing Data Volumes
- Managing Operations Costs
- Optimizing Hardware Costs

Modernizing Software

- With today’s codes, performance is often very far from what modern CPUs can deliver. This is due to a number of factors, ranging from the construction of the code, not being able to use vector or other hardware units, layout of data in memory, and I/O performance
- We expect that code re-engineering might gain CMS physics software a moderate factor (x2) in overall performance. Our software is written > a hundred of authors and domain experts, so success in this area requires that the whole community develops a level of understanding of how to best write code for performance
- That also requires the appropriate support and tools, for example to satisfy the need to fully automate physics validation of software across different hardware types and frequent changes, to optimize the best use of opportunities
- The need for software modernization was a driver for creating the HEP Software Foundation, in which Fermilab and the US are heavily engaged



Improving Algorithms

- The level of pile-up anticipated for HL-LHC means that current reconstruction algorithms must be improved significantly to avoid **exponential computing time increases**. It is estimated that a considerable improvement could be obtained with some re-tuning of current algorithms, but new approaches are needed that could have larger benefits
- New CMS **detector technologies**, like very high-granularity calorimetry, tracking and timing require re-thinking of reco algorithms and particle flow
- Wider and deeper application of **Machine Learning** could lead to disruptive improvements and **change the scaling behavior of algorithms**, from triggers, pattern recognition, particle flow reco, to “inference-driven” event simulation
- All this requires expert effort PLUS engagement from the domain scientists, and Fermilab has unique opportunities due to its advantageous coupling of computing/software and physics expertise — e.g. at SCD and the LPC

Reducing Data Volumes and Managing Operations Costs

- A **key cost driver** is the amount of **storage** required.
- Investigating mechanisms for reducing data volume: removing or reducing the need for storing intermediate data products, managing the sizes of derived data formats, for example with “nanoAOD”-style even for some fraction of the analyses will have an important effect
- To **optimize operations cost**, investigating the opportunities with **storage consolidation** is a high priority. The idea of a “**data-lake**” where few large centers manage the long-term data, while needs for processing are managed through streaming, caching, and related tools, allows the cost of managing / operating large storage systems to be minimized, reduces complexity
- That gives opportunity to move common data management tools out of the experiments into a common layer. This allows better optimization of performance and data volumes, easier operations, and common solutions, including common solutions for workflows
- Storage consolidation can save cost on expensive managed storage, if we can hide the latency via streaming and caching solutions. This is feasible as many of our central workloads are not I/O bound, and data can be streamed to a remote processor effectively with the right tools

Optimizing Hardware Costs

- Storage cost can be reduced by more actively using “cold storage”. A highly organized access to tape or “cheap” low-performant disk could remedy the need to keep a lot of data on high-performance disk arrays
- The judicious use of virtual data (re-create samples rather than store) is another opportunity. This could save significant cost, but requires the experiment workflows to be highly organized and planned
- Moving away as far as possible from random access to data before the final highly refined analysis formats. Other considerations include the optimization of the amount of storage vs compute, and optimizing the granularity of data that is moved — dataset level vs event level
- Data analysis facilities could be provided as a centralized and optimized service, also allowing caching and collating data transformation requests

U.S. and Fermilab are Central to CMS Computing

- For CMS, only the prompt reconstruction is entirely done at CERN, everything else touches Fermilab and the US
  - — both a liability and an asset for Fermilab and US CMS!
- Roles of Fermilab and US:
  - Software
    - ★ Core framework ➜ core expertise
    - ★ Physics algorithms ➜ domain expertise
    - ★ US role for software validation: nightly, and for releases
  - Infrastructure services
    - ★ Resource Provisioning, Workflow management
    - ★ Metadata, Data Transfer system, Data federation
    - ★ Distributed conditions database system
    - ★ Software distribution
  - Community support
    - ★ LPC, Tier-3s, Universities through CMSConnect
  - Contributions from projects external to CMS:
    - ★ Geant4, ROOT, Xrootd, HTCondor, glideinWMS, Frontier & Squids, CVMFS, HEPCloud, physics generators

- CMS software is written by **hundreds of domain experts** and a **small group of core experts**
  - 3 million lines of C++ code
  - 1 million lines of python (configuration)
- Software and Computing integral part of the Science Process
- Needs both **domain experts and core computing experts**

Resource Limitations Incentivize R&D Investments

- CMS has always been more resource limited compared to others
  - 40% of the collaboration are from countries that cannot contribute a Tier-1 center
  - Invest in computing R&D to optimize, and in social engineering of compromises (e.g. nanoAOD)!
  - CMS has to make more aggressive choices to reduce resource needs
- Physics:
  - CMS chose generators that need less resources (a few percent of the total CPU needs in Run2)
  - Full simulation is effective due to approximations like “russian roulette” and faster physics lists in Geant4 — we believe, without compromising physics performance!
- Data Volumes:
  - miniAOD is a success in social-engineering to convince collaborator to fully embark on the advantages of a smaller (compromise) data format — nanoAOD is expected to continue this trend
  - Pile-up simulation using “pre-mixing”, reduces I/O load and CPU time needed significantly
- Multi-threading
  - CMS is executing all workflows multi-threaded, reducing the memory needs per core
  - This enables to run on smaller core architectures (e.g. KNL and other many-core processors) ➜ R&D

Examples for Existing Fermilab and US CMS Activities

- Project on industry "Big Data" analysis technologies
  - for physics data analysis, one of the largest challenges of HL-LHC
    - ★ need to enable interactive analysis on significantly larger datasets than today
  - The LHC Big Data Project: Fermilab (SCD and a LDRD) partnering with CERN OpenLab, Intel and the NSF DIANA project
  - Goal is to reduce 1 PB to 1TB in 5 hours, using Apache Spark
    - ★ Already achieved throughput of 72 TB/hour!
- SciDAC-4 Project on HEP Reco on Parallel Architectures
  - Exploration into modern computing architectures to speed up reco
    - ★ goal: boost utilization of new computing architectures in HEP reconstruction:
      - – for LHC, and for neutrino experiments that are using LArTPC detectors
    - ★ use of advanced profiling tools and development techniques
      - – maximize throughput on leading parallel architectures (Xeon Phi, GPU)
    - ★ explore portable implementations for HPC and heterogenous platforms
  - SciDAC-4 funds work at Fermilab with U.Oregon — related activities:
    - ★ Cornell/Princeton/UCSD collaboration, NSF funded (see next slide)
    - ★ ASCR institutes: Rapids (Platform Readiness), other SciDAC projects: Hep.TrkX (tracking with ML), experiments: CMS, Atlas, DUNE, and SBN

CERN openlab

CERN OPENLAB IS SET TO ENTER ITS SIXTH THREE-YEAR PHASE AT THE START OF 2018. WE'LL BE LAUNCHING A NEW WEBSITE — COMPLETE WITH UPDATED BRANDING — FOR THIS NEW PHASE.

intel

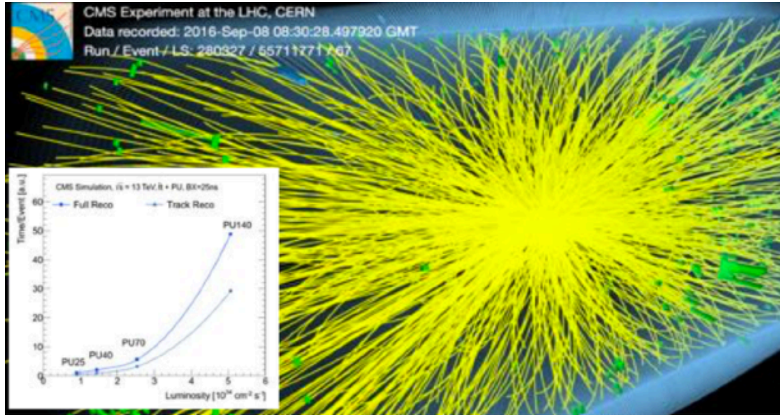
Fermilab

APACHE Spark™

dianahep

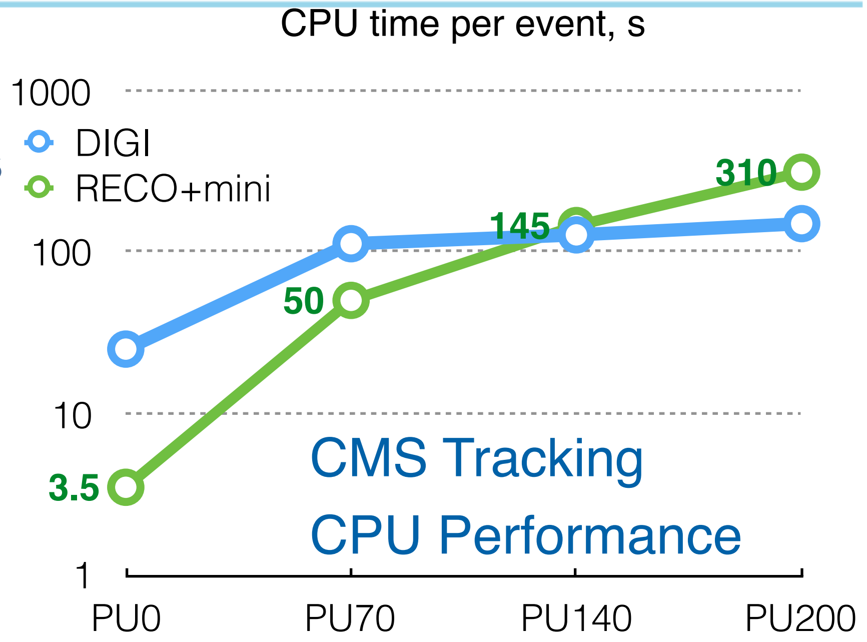
SciDAC

Scientific Discovery through Advanced Computing

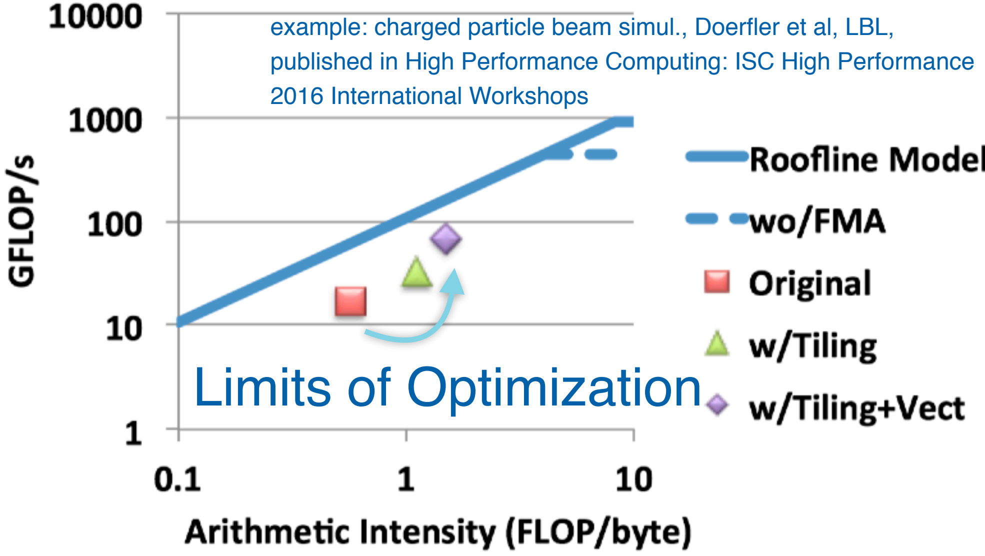


Example: R&D in Optimizing Tracking Pattern Recognition

- In CMS, tracker pattern recognition dominates CPU
  - project to speed up, parallelize, vectorize Kalman filter algorithms
  - Partnership of NSF PIF project with DOE SciDAC-4
- Successful vectorization of KF fit and track finding
  - Convert event data to structures suitable for vectorization
  - Keep algorithm development general and adaptable to rapidly (and unpredictably!) changing HW landscape
    - ★ Intel MIC, then KNL, now testing on Skylake
  - measured CPU timing: ttbar at 70 PileUp, single thread
    - ★ Vectorized pattern recognition x10 faster w/r to current CMS tracking implementation, at comparable physics performance
  - Working on completing the physics validation
- Optimizing algorithms for GPUs is ongoing
  - Challenge: **track finding** combinatorics of identifying next hit
    ➜ large amount of data per FLOP, moving in and out of GPU
    ➜ lower ratio of FLOP/byte, leading to lower performance



CMS Tracking CPU Performance

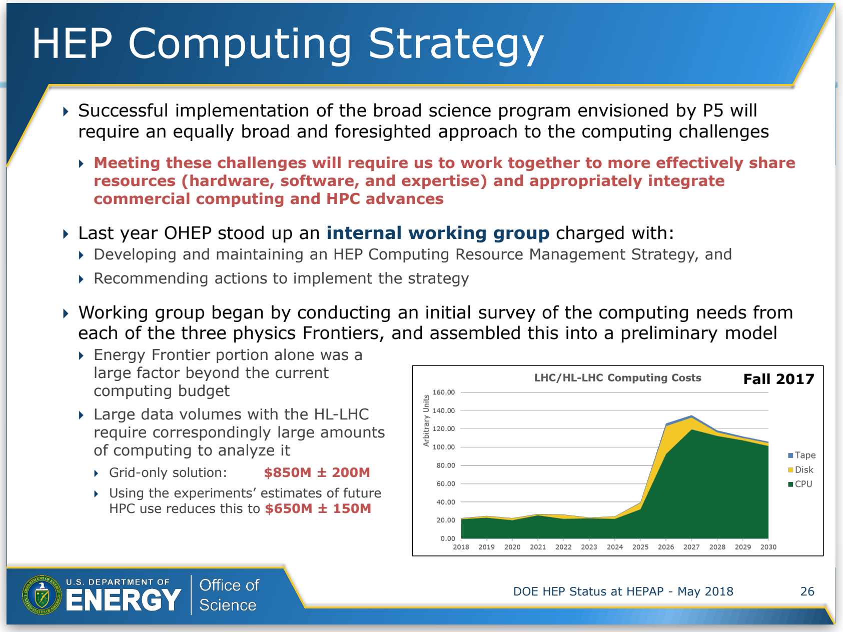

example: charged particle beam simul., Doerfler et al, LBL, published in High Performance Computing: ISC High Performance 2016 International Workshops

Limits of Optimization

see e.g. http://trackreco.github.io

Bringing non-HEP Resource to Bear

- J.Siegrist at HEPAP:
  - “Successful implementation of the broad science program envisioned by P5 will require an equally broad and foresighted approach to the computing challenges
  - “Meeting these challenges will require us to work together to more effectively share resources (hardware, software, and expertise) and appropriately integrate commercial computing and HPC advances
- CMS is fully embracing the use of HPC for all production workflows
  - we directly went for running full simulation + reconstruction on HPC
    - ★ running just physics generators or Geant simulation alone would not benefit CMS
- With HEPcloud, Fermilab has already demonstrated integration of commercial computing and HPC, at very large scales
  - with HEPcloud, we solve the challenges of accessing these resources:
    - ★ Data access (network, I/O performance), Collaboration access (authentication, authorization), Software access (certification), Time access (turn around)
- Architectures of future HPC will heavily rely on “accelerators”: GPUs, FPGAs, TPUs, etc — J.Siegrist:
  - “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.
    - ★ “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… HEP needs coordinated interface to ECP & the Leadership Computing Facilities
    - ★ “Need to identify which codes could benefit the most, studies of selected HEP codes

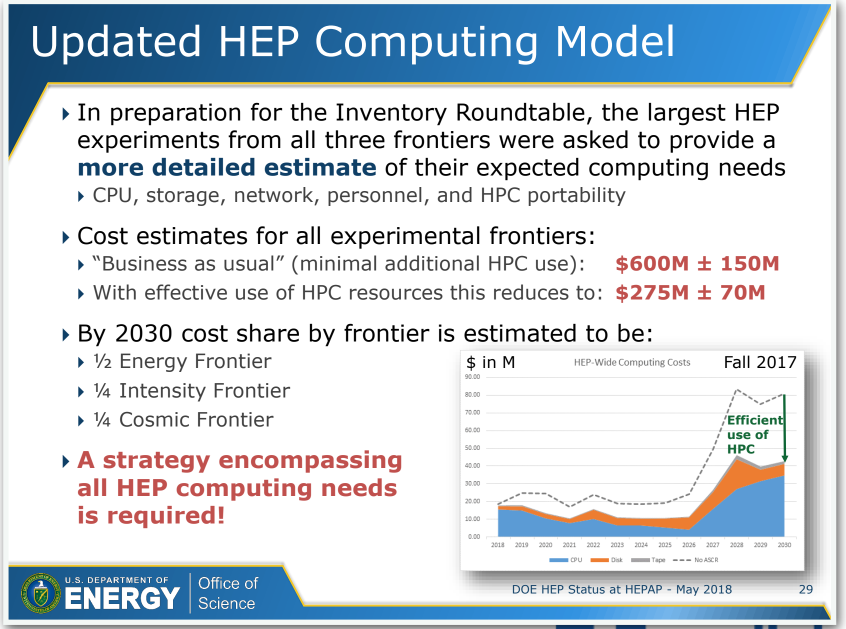


HEP Computing Strategy

- Successful implementation of the broad science program envisioned by P5 will require an equally broad and foresighted approach to the computing challenges
  - **Meeting these challenges will require us to work together to more effectively share resources (hardware, software, and expertise) and appropriately integrate commercial computing and HPC advances**
- Last year OHEP stood up an **internal working group** charged with:
  - Developing and maintaining an HEP Computing Resource Management Strategy, and
  - Recommending actions to implement the strategy
- Working group began by conducting an initial survey of the computing needs from each of the three physics Frontiers, and assembled this into a preliminary model
  - Energy Frontier portion alone was a large factor beyond the current computing budget
  - Large data volumes with the HL-LHC require correspondingly large amounts of computing to analyze it
    - Grid-only solution: **$850M ± 200M**
    - Using the experiments’ estimates of future HPC use reduces this to **$650M ± 150M**

U.S. DEPARTMENT OF ENERGY | Office of Science

DOE HEP Status at HEPAP - May 2018 26



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

Initial Pilot Collaboration with ECP

- The Exascale Computing Project
  - “accelerating delivery of a capable exascale computing ecosystem for breakthroughs in scientific discovery, energy assurance, economic competitiveness, and national security”
    — also for HEP and HL-LHC?
- Fermilab is starting a pilot project with ECP, building on our very successful GeantV R&D
  - Result of discussions around a recent ECP workshop, with one of the funded nuclear physics applications
- Develop a ECP “proxy application” for Geant simulations
  - Standalone app that captures essential elements of architecture and tasks needed by actual app
  - Use “proxy” to characterize and document opportunities for performance improvement in ECP environment
  - Setup the stage for a subsequent ECP projects to realize these improvements in the HEP software stack
  - Allows us to work with ECP researchers with related domain expertise
  - Access to ECP experts and know-how, will allow us to further evolve architecture and algorithmic solutions
- A very important first step
  - Its success will help us engage ECP to other projects of interest to us

Closing Remarks

- HL-LHC computing presents a formidable and quite costly challenge
- The core to solving HL-LHC computing lies in modernizing the physics software, algorithms and data structures, to allow cost effective computing solutions based on industry trends and emerging science infrastructures
- Fermilab and our collaborating universities are central to addressing these challenges, in particular given the special role of Fermilab and the US for CMS
- Fermilab and US CMS already are part of a broad eco-system of R&D, which also includes the neutrino program — we can bring to bear the lab's computing core competencies, SCD capabilities and leadership, and a unique opportunity for close interactions between physicists and computing experts
- DOE encourages us to look outside the field, for more computing resources and for expertise to efficiently use future computing architectures,
  and we're ready to take on these challenges

The End