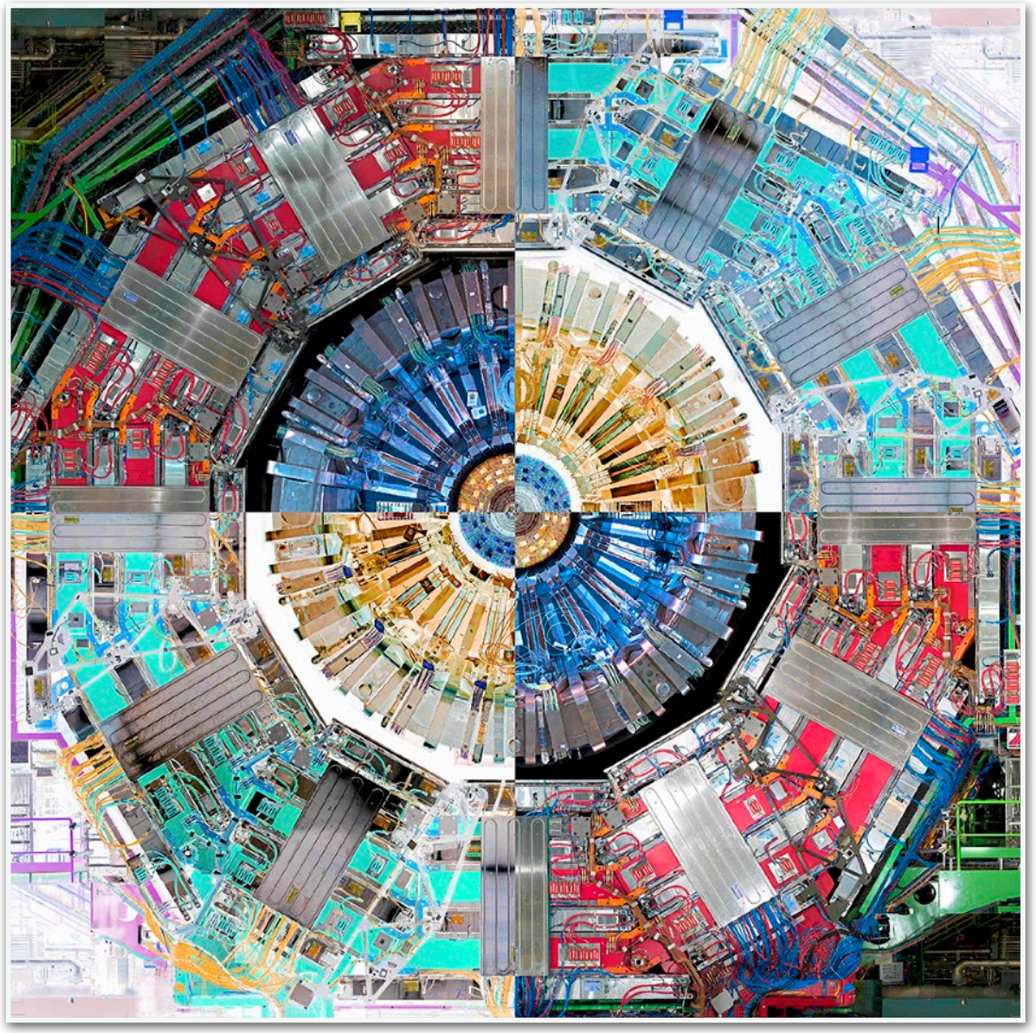
Fermilab

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

# Strategy towards HL-LHC Computing for U.S. CMS



Lothar A. T. Bauerdick

Informal DOE Briefing

June 13, 2019

# Where are we now?

several $\mathcal{O}$(N)

- Computing capabilities successfully scaled for Runs 1 & 2
  - **evading "breakdown of Moore's law":** effective use of multi-cores
    - ✅ ★ robust multi-threaded framework and use of heterogenous architectures
    - ✅ ★ order-of-magnitude efficiency gains by improving software / data formats
  - **distributed data federations:** efficient use of vastly improved networks
    - ✅ ★ transparent over-the-network access to data
    - ✅ ★ ESnet deployed high-throughput transatlantic networks
  - **sharing and on-demand provisioning** of computing resources
    - ✅ ★ HEPcloud, OSG: opened door to use HPC allocations, commercial clouds

- So far, computing has not been a limiting factor for CMS physics
  - ★ however, computing remains a significant cost driver for LHC program
  - ★ U.S. spent well above $200M on CMS computing (through Ops program)

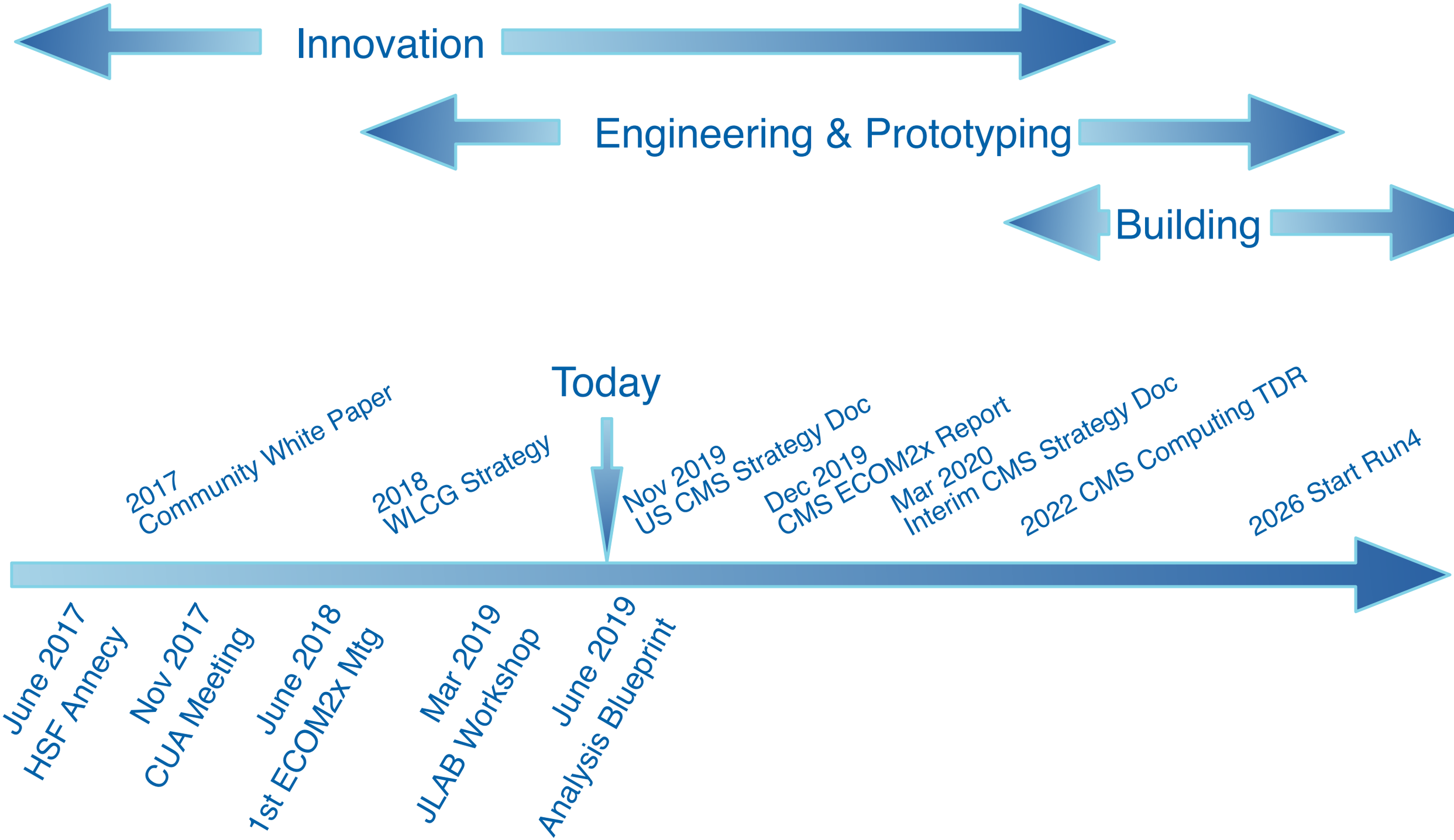
# HL-LHC: Significant Progress since "Naïve" Extrapolation of 2017

- For HL-LHC, computing could indeed become a **limit to discovery**, unless we can make **significant changes** that will help moderate computing and storage needs, while maintaining physics goals:
  - HL-LHC requires Exa-scale computing: x50 storage, x20 CPU, >250Gbps networks

- Roadmap towards HL-LHC computing Technical Design is clarified
  - "Community White Paper" and WLCG "Computing Strategy Paper"
- Fermilab and US CMS are vigorously participating in this process
  - R&D activities in SCD and US CMS, funded through opportunities in DOE and NSF

- US CMS and ATLAS now have the outlines of and started to embark on a **work program for 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 has started last year;
  - ★ and a complementary DOE sponsored program for Fermilab and its university partners, building on the strength of the CompHEP, SciDAC, SCD, LDRD, and USCMS activities and capabilities

# Time Line for HL-LHC Computing R&D



# Areas That Must be Addressed (WLCG Strategy)

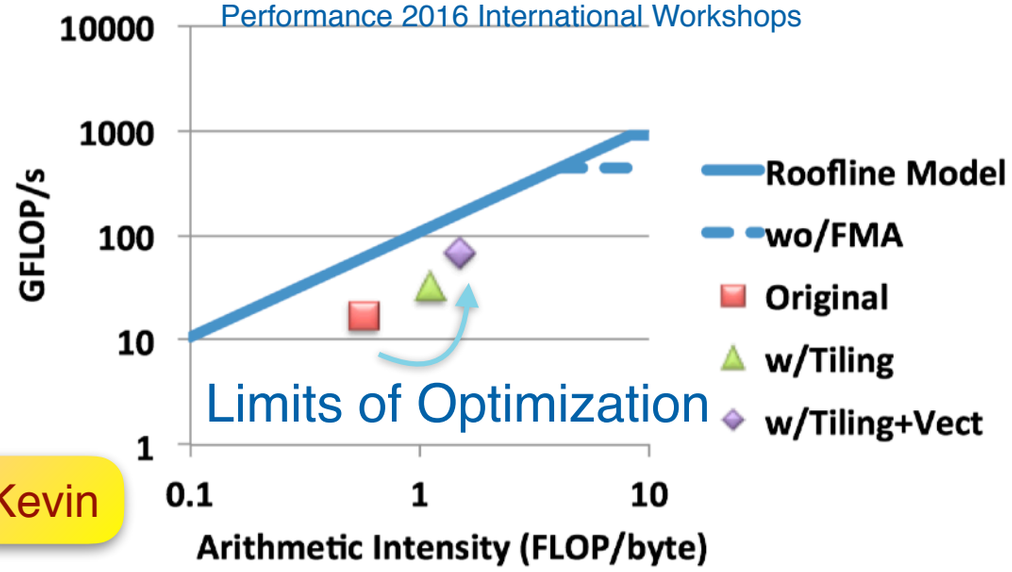
1. Modernizing Software
2. Improving Algorithms
3. Reducing Data Volumes
4. Managing Operations Costs
5. Optimizing Hardware Costs

# 1 Modernizing Software

- Today's LHC code performance is often far from what modern CPUs can deliver
  - Some is inherent to current algorithms:
    - typically nested loops over complex data structures, small matrices, making it hard to effectively use vector or other hardware units
    - complex data layout in memory, non-optimized I/O
  - Expect to gain only moderate performance factor (x2) by re-engineering the physics code

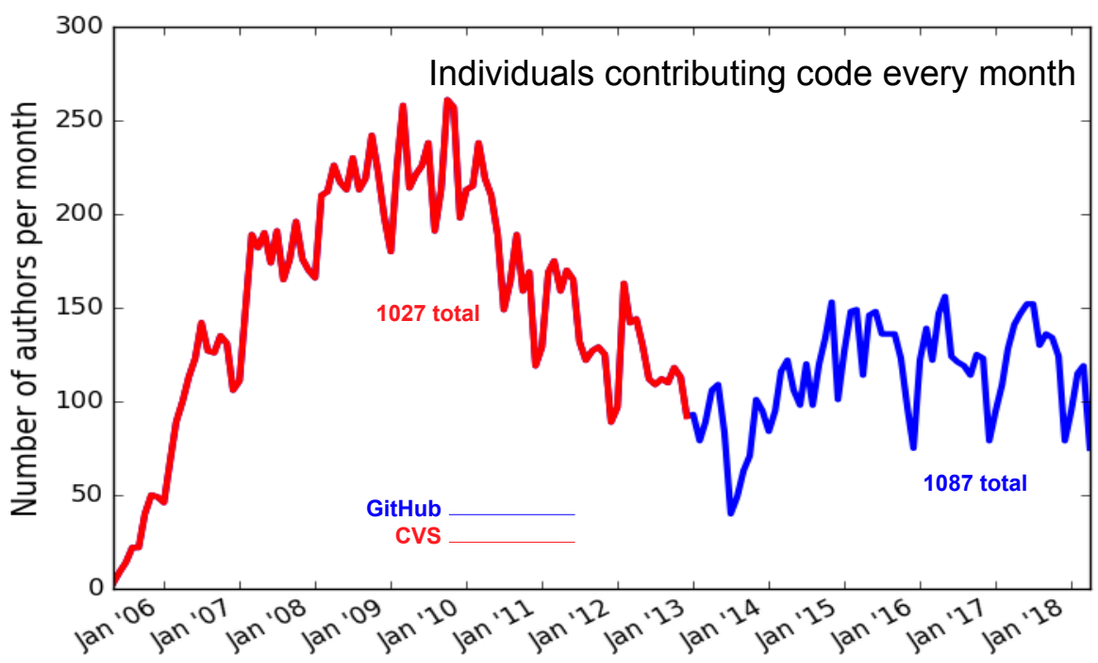
Allie Michalis Kevin

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




- CMS software is written by > a hundred of authors and domain experts
  - success in this area requires that the whole community develops a level of understanding of how to best write code for performance
- Support roles for USCMS Ops and HSF
  - automate physics validation of software across different hardware types and frequent changes
  - help with co-(re)design, best practices, codes

**CMS Software**



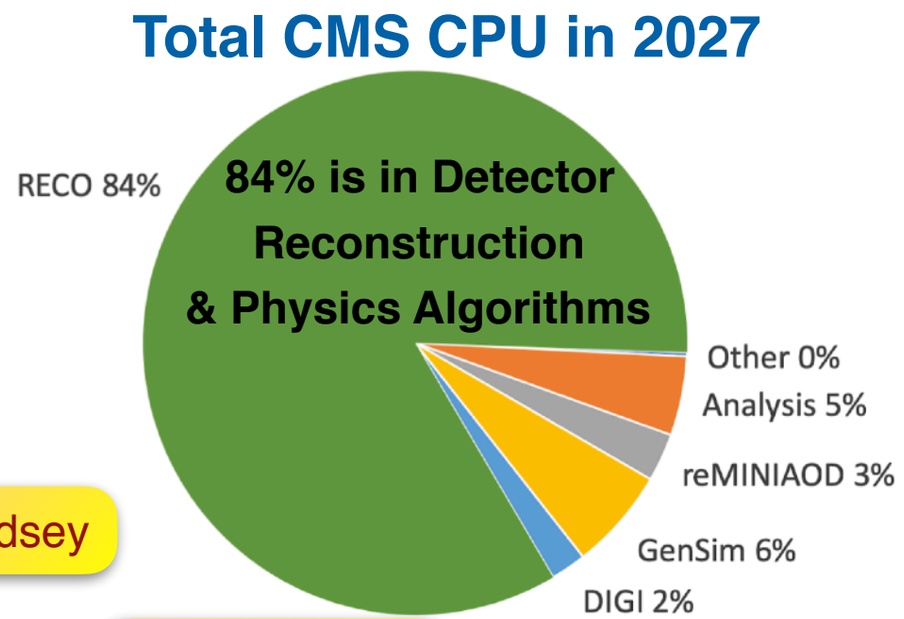



Fermilab

# 2 Improving Algorithms

- pile-up —> algorithms have to be improved to avoid **exponential computing time increases**
  - considerable improvement possible with some re-tuning, but new approaches are needed to have larger benefits
- New CMS **detector technologies**
  - very high-granularity calorimetry, tracking and timing require re-thinking of reco algorithms and particle flow

Lindsey




- Wider and deeper application of **Machine Learning/AI**
  - to **change the scaling behavior of algorithms** for disruptive improvements of triggers, pattern recognition, particle flow reco, to “inference-driven” event simul. and reconstruction

Jean-Roch

Javier

- Requires **expert effort PLUS engagement from the domain scientists**
  - and Fermilab has unique opportunities due to its advantageous in-house coupling of **computing/software and physics expertise** — Fermilab SCD and the LPC

Kevin Nhan

- To sustain such efforts & exploit opportunities of the LHC Physics Center, Fermilab should get into a position where it can make effective **connections** between **CMS domain experts** and **DOE computing experts**
  - e.g. connect the experiment to ECP and co-design projects etc.
- Decisive “Chicago Area” advantage could be in a close(r) tie b/w **FNAL & ANL**

Fermilab

# 3 Reducing Data Volumes and 4 Managing Operations Costs

- A **key cost driver** is the amount of **storage** required
  - focus on 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
  - —> CMS is ahead of the game, and last year has successfully introduced a nanoAOD format, already for Run2

Brian

Nick Joosep

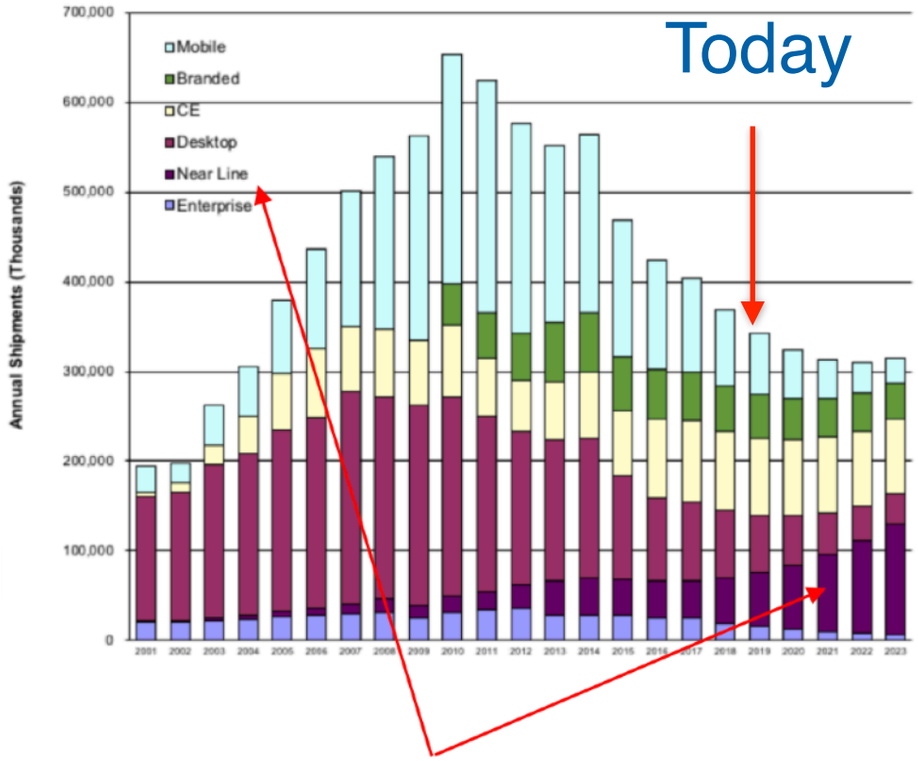
- **Storage consolidation** to **optimize operations cost**
  - The idea of a **data-lake** where few large centers manage the long-term data, while needs for processing are managed through streaming, caching, and related tools, allows the cost of managing / operating large storage systems to be minimized, reduces complexity
  - 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
  - move common data management tools out of the experiments into a common layer
    - ★ allows optimization of performance and data volumes, easier operations, and common solutions

Brian

- Fermilab needs to prepare taking on this central role, on behalf of the experiment, focusing on **providing data services**, and **brokering CPU services**, from wherever they are “cheapest”

# 5 Optimizing Hardware Costs

Example from: HEPIX Tech Watch




- Storage cost can be reduced by more **actively using "cold storage"**.
  - highly organized access to tape or "cheap" low-performant disk could remedy the need to keep a lot of data on high-performance disk
- optimization storage vs compute, **optimizing the granularity of data** that is moved — dataset level vs event level
- **Moving away from random access** to data
  - Modern systems like in Joosep's and Nick's talks show the power of this approach
- Judicious use of **virtual data**: re-create samples rather than store
  - This could save significant cost, but requires the experiment workflows to be highly organized and planned, and CMS is working towards those goals (helped by framework)
- **Data Analysis Facilities** could be provided as a centralized and optimized service, also allowing caching and collating data transformation requests
  - we are developing the concepts of a centralized analysis service, and Nhan's example of a "inference as a service" shows possible architectures how to include HPC facilities

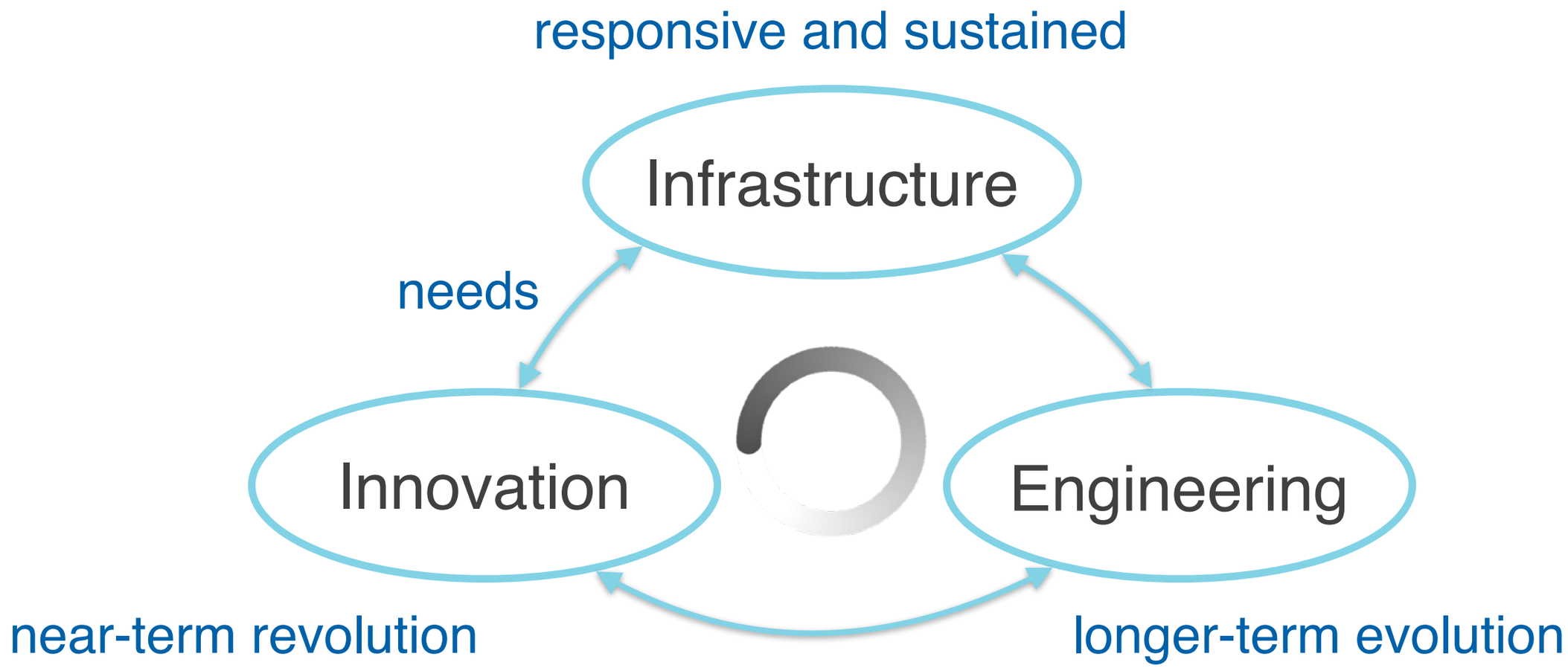
# Resource Limitations → Need to Invest in R&D & Innovation

- CMS has always been more resource limited, compared to others
  - ★ Given our demographics, CMS will always have to do more with less resources
  - ★ CMS makes more aggressive choices to reduce resource needs
  - → Invest in computing R&D to **optimize**, and in social engineering to **compromise**
- Compromises in computing approaches, without compromising Physics
  - **chose generators** that need less resources (only a few % of 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!
  - Huge possible future savings in **optimizing analysis process**: declarative approach and optimized data access, bring rich open-source data science environment to bear
- Data Volumes and Data Discipline
  - past successes in social-engineering to convince CMS to fully embark on advantages of a **smaller (compromise) data format** —> wide adoption of nanoAOD
  - **Pile-up simulation** using pre-mixing, reduces I/O load & CPU time needed significantly
- Early investments into multi-threading framework & scheduling to accelerators
  - CMS is executing all workflows multi-threaded, reducing the memory needs per core
  - enables to run on smaller core architectures (e.g. KNL and other many-core processors)

Matti

# LHC Community Energizes the Lab's Expertises and Facilities

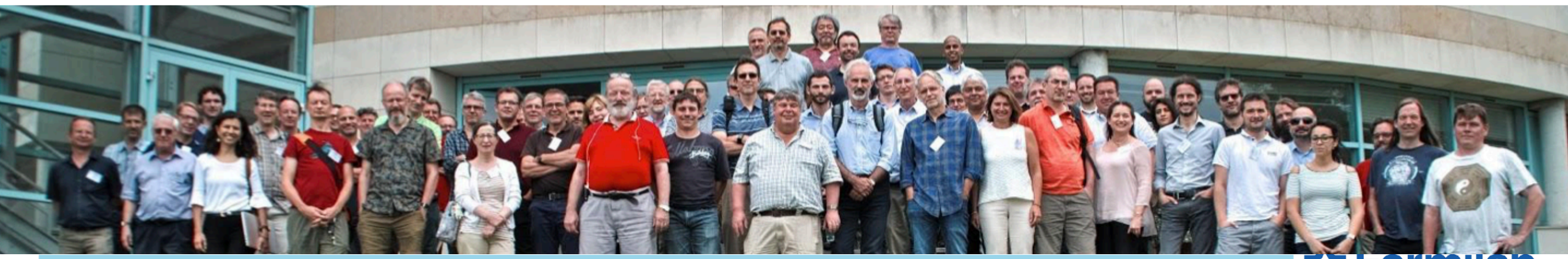
- LHC needs & requirements + community engagement is a powerful motor for innovation in computing



# "A Coordinated Ecosystem for HL-LHC Computing R&D"

- Nov 2017 meeting w/ DOE and NSF program managers at CUA
  - "Multiple R&D efforts must be coordinated to achieve coherence and alignment between a multitude of stakeholders and effort providers, US and international. Strong DOE/NSF partnerships will be required. A joint blueprint activity will be critical to building this coordination."

- HL-LHC Computing R&D Eco-system becomes to be effective
  - NSF Software Institute for the LHC, IRIS-HEP, is funded and active
  - DOE CompHEP, SciDAC, CCE, LDRDs, are supporting important efforts
  - US CMS and US ATLAS are starting a HL-LHC computing postdocs program



# Focus Areas for HL-LHC Computing R&D, as agreed at CUA

## *Focus Areas for HL-LHC R&D*

**Data Analysis Systems**
**Reconstruction and Trigger Algorithms**
**Applications of Machine Learning**
**Data Organization, Management and Access**
**Simulation**
**Storage infrastructure and Facilities**
**Data Transfer and networking infrastructure**
**Workflow and Resource management**
**Event Processing Frameworks**
**Data and Software Preservation**
**Physics Generators**
**Visualization**
**Software Development, Deployment and Validation/Verification**
**…**

**$S^2I^2$ Focus Areas**

*(highest-priority areas for initial S2I2 investment)*

Evolved by Blueprint Activity

12

# Focus Areas for HL-LHC Computing R&D, as agreed at CUA

## *Focus Areas for HL-LHC R&D*

| | |
|---|---|
| C1 | Data Analysis Systems |
| C2 | Reconstruction and Trigger Algorithms |
| C3 | Applications of Machine Learning |
| C4 | Data Organization, Management and Access |
| C5 | Simulation |
| C6 | Storage infrastructure and Facilities |
| C7 | Data Transfer and networking infrastructure |
| C8 | Workflow and Resource management |
| C9 | Event Processing Frameworks |
| C10 | Data and Software Preservation |
| C11 | Physics Generators |
| C12 | Visualization |
| C13 | Software Development, Deployment and Validation/Verification |

…

**$S^2I^2$ Focus Areas**

*(highest-priority areas for initial S2I2 investment)*

Evolved by Blueprint Activity

12

# Work Breakdown for HL-LHC Computing R&D

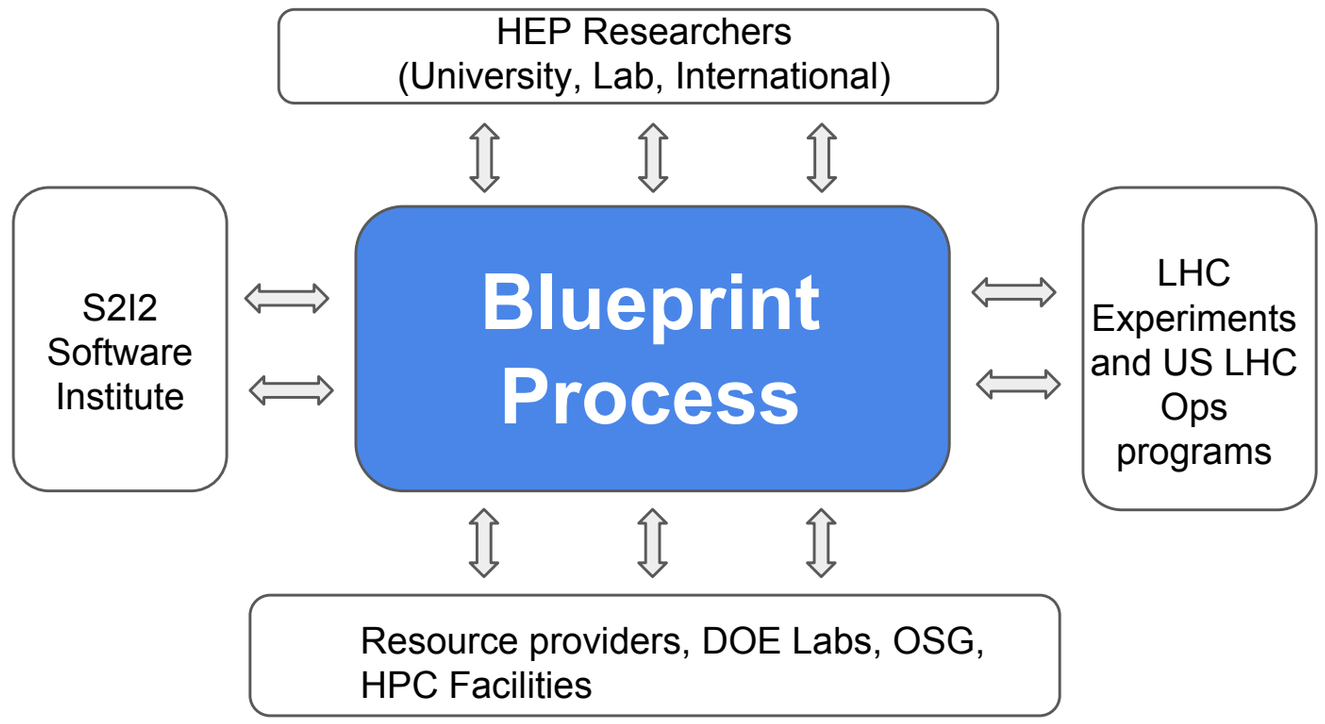
| U.S. CMS HL-LHC deliverable | Computing Challenge | Computing Area of US | Presentation | Technologies and Paradigms | Connections, Stakeholders | Sources of Support |
|---|---|---|---|---|---|---|
| **Tracker** | pattern reco, pile-up, 4D tracking and vertexing | C2, C3 | Allie, Nhan, Matti, Jean-Roch | FPGAs, GPUs, optimized data structures | ATLAS, SciDAC RAPIDS, ORNL, IRIS-HEP,NSF | CompHEP, SciDAC (FNAL + ORNL), LDRD, HSF, NESAP |
| **High-Granularity Calorimeter** | fine granularity clustering, pile-up, complex geometries, particle flow | C2, C3, C5 | Lindsey, Nhan, Matti, Jean-Roch, Kevin | GNN, VecCore, kokkos, RAJA | IRIS-HEP, ECP | LDRD, ECP |
| **Trigger** | event rates, pile-up, track trigger | C2, C3, C6 | Michalis, Javier, Matti, Jean-Roch | FPGAs, HLS4ML, Microsoft Azur Brainwave | ATLAS, DUNE, Accelerator Controls | LDRD |
| **Data Analysis** | DOMA, event throughput, optimizing algorithms and innov. approaches, usability, interactivity, data analysis facility | C1, C6, C7, C8, C9, C10, C11 | Nick, Joosep, Kevin, Nhan, Brian, Matti | Data Science eco-system, SPARK, Fermi-Striped, uproot, awkward arrays, aws | ATLAS, IRIS-HEP, NOvA, DUNE, ECP | CompHEP, ECP, IRIS-HEP, HSF, Intel, CERN Openlab |

- R&D presentations today cover a large subset of the computing areas of interest to U.S. CMS, and highlight the connections between projects and stakeholders

# Joint DOE/NSF Blueprint Process Binding it Together

- Drive the evolution of R&D efforts to address the software & computing challenges of the HL LHC, co-sponsored by:
  - **US LHC Ops program**
  - **S2I2**
  - **OSG**
  - **CCE**

- Involving the DOE facilities, and key personnel at both DOE labs and US Universities.
- Long term sustained set of workshops to drive coherence across projects and experiments.



# The Fermilab LPC—SCD Connection is a Strategic Advantage

- For CMS, almost everything "Computing" touches Fermilab and the US
- 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**
- Fermilab has a unique connection of SCD computing experts and LPC domain experts and collaborators
  - A major asset to "solve" the HL-LHC computing challenge
- Now need a period of Innovation and R&D, embedded with, informed by, and with support from computer professionals and computer scientists

- Facilities
  - ★ Storage: disk and tape, high-throughput computing
- Software
  - ★ Core framework ➜ core expertise
  - ★ Physics algorithms ➜ domain expertise
  - ★ 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

# Coordination and Collaboration

- CCE: closer communication and coordination with LHC
  - Recent Ops Program review recommendations (jointly to US CMS and ATLAS)
  - Develop an HL-LHC S&C R&D strategic plan […] with specific milestones for deliverables[…].
    - Carry out a set of open workshops in coordination with US [CMS and ATLAS], HEP-CCE, IRIS-HEP, Open Science Grid (OSG)-LHC, and WLCG.
    - Coordinate with the DOE and NSF a plan to sustain such an R&D activity for the next 3-5 years

- ECP: possible involvement of ECP in LHC applications
  - Co-design Center or Energy Applications project or something similar?
  - Co-design targets crosscutting algorithmic methods that capture the most common patterns of computation and communication, in ECP applications
    — can LHC be targeted for one of these?

# Closing Remarks

- CMS aims to for HL-LHC Computing to succeed **within current funding levels**, but we need an **infusion of R&D** now
  - Physics choices will be made to contain cost, and CMS has a record of being able to do that

- 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
  - Storage is the cost driver, our data storage systems cannot be done “opportunistically”, Fermilab has to be the central data hub, while CPU will come from several sources
  - The CPU challenge is about physics algorithms, how they run at high pile-up, on various machine architectures
  - 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
  - US CMS would like to partner more closely, sustained, and coherently coordinated with Fermilab and CCE, with other OHEP computing initiatives, including in ASCR and ECP

- 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**

# Current US CMS HL-LHC R&D Efforts

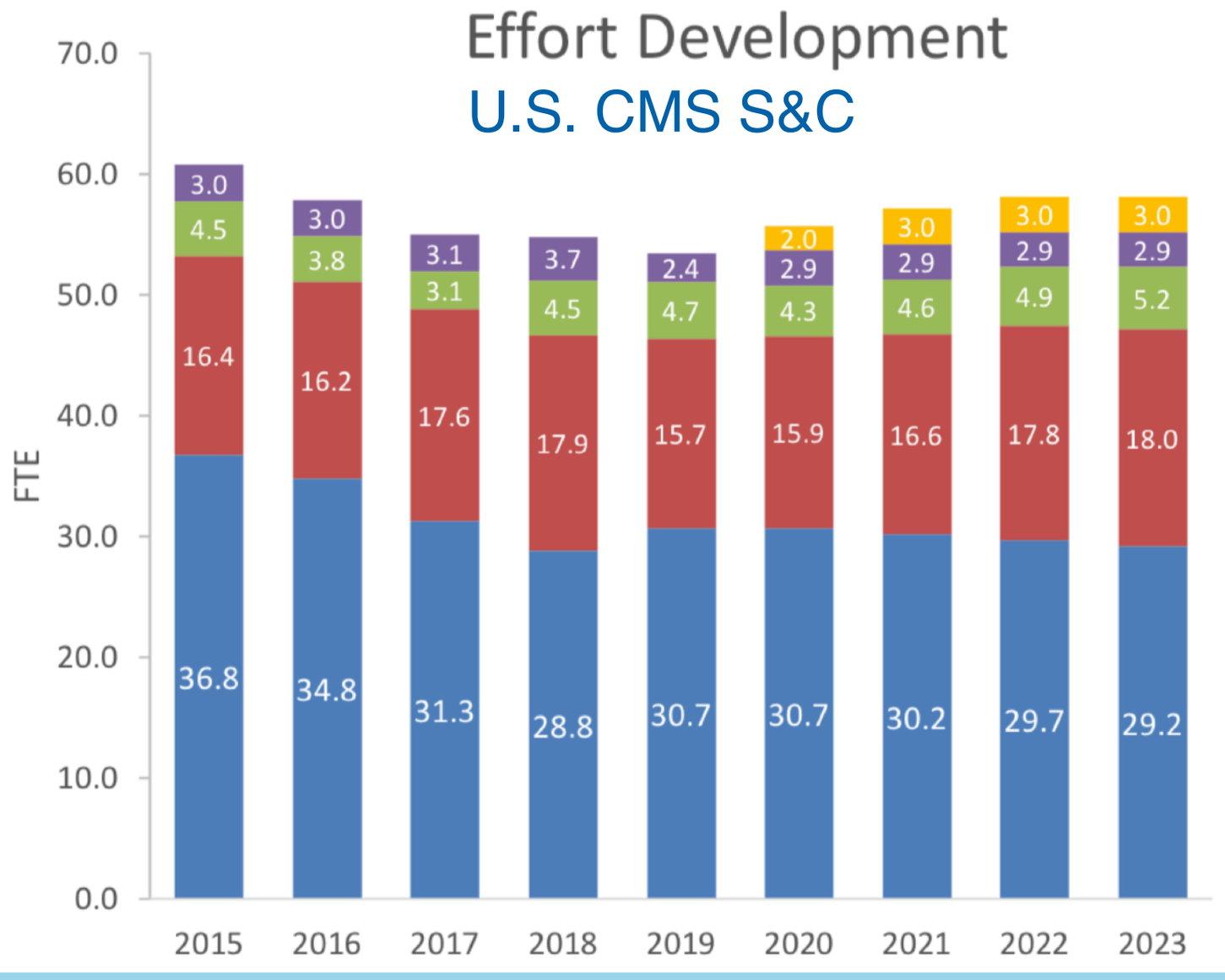
| WBS area | Task | DOE | NSF | Grand Total |
|---|---|---|---|---|
| Computing Infrastructure and Services | | 0.8 | 1.0 | 1.8 |
| | HEPCloud R&D | 0.8 | | 0.8 |
| | DOMA R&D: xrootd towards data lakes | | 0.5 | 0.5 |
| | Workflow Management R&D | | 0.5 | 0.5 |
| Software and Support | | 0.8 | 2.1 | 2.9 |
| | Analysis Facility R&D | | 0.3 | 0.3 |
| | Machine Learning R&D | | 0.7 | 0.7 |
| | ROOT7 EvE Visualization R&D | | 0.4 | 0.4 |
| | DD4HEP Geometry Description R&D | 0.8 | | 0.8 |
| | Tracking on Advanced Architectures R&D | | 0.7 | 0.7 |
| Grand Total | | 1.6 | 3.1 | 4.7 |

- In 2019, US CMS has staffing levels of 4.7 FTE for R&D, about half or them are available for R&D towards HL-LHC

# Ramping up HL-LHC R&D effort

- ★ current funding levels allow for adding a Postdoc-level R&D program,
  - ramping to 6 postdocs (half position funded by Ops),
- ★ plus ~2 FTE additional software engineering support (“Development”)
  - by gaining efficiencies in operations

**Needed** HL-LHC computing effort level estimated to **15 FTE in 2023** (current plan: <10 FTE)




Joint Postdoc Program



Fermilab

# Bringing non-HEP Resource to Bear (from PAC July 2018)

- 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