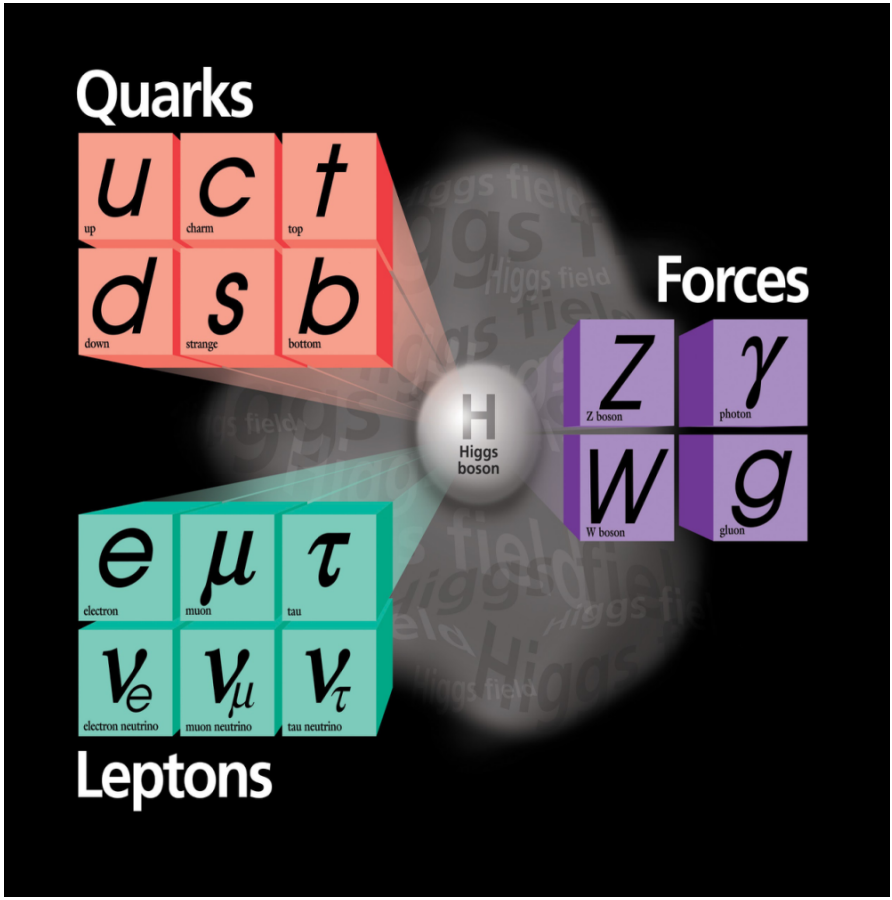
Managed by Fermi Research Alliance, LLC for the U.S. Department of Energy Office of Science

HEPCloud: efficient resource provisioning in the Science Cloud

Panagiotis Spentzouris
Fermilab
16/03/17

HEP: success through decades of data-driven science

The Standard Model of Particle Physics



A great scientific achievement

- Every particle physics experiment ever done fits with this model

But it is incomplete…

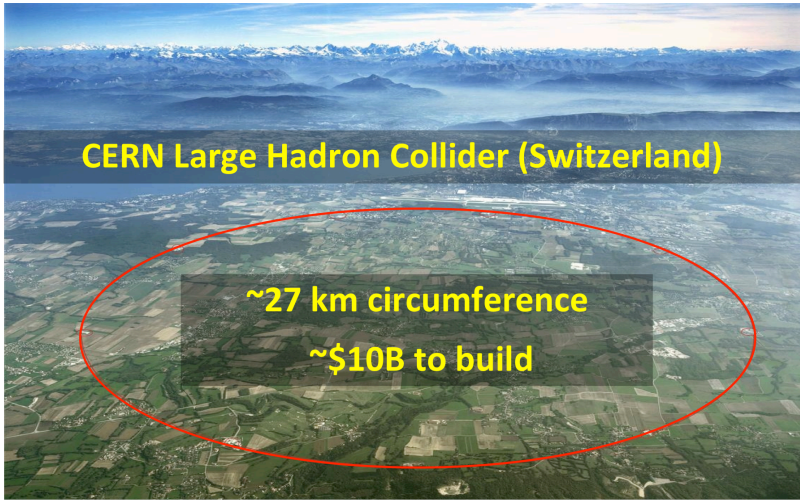
Why are there so many particles?
Why are their masses so different?
Why is the universe mostly matter?

- Is there a higher level theory providing natural explanation to these questions and unifying the different forces?

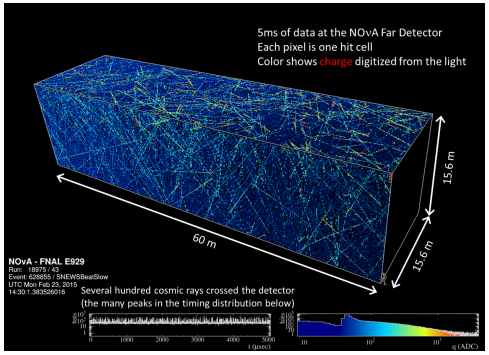
HEP Science Drivers

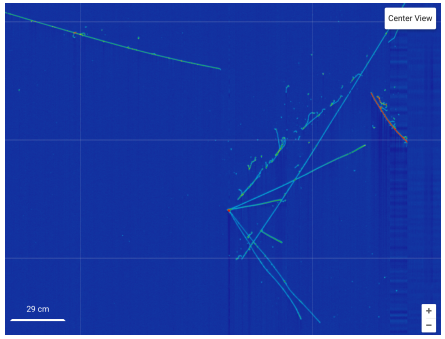
- Utilize high-energy particle beam collisions to discover
  - the origin of mass, the nature of dark matter, extra dimensions.
- Employ high-flux beams to explore
  - **neutrino interactions**, to answer questions about the origins of the universe, matter-antimatter asymmetry, force unification.
  - **rare processes**, to open a doorway to realms to ultra-high energies, close to the unification scale
- Commission surveys utilizing massive instruments to understand the nature of the contents of the universe
  - ordinary matter, dark mater and dark energy.

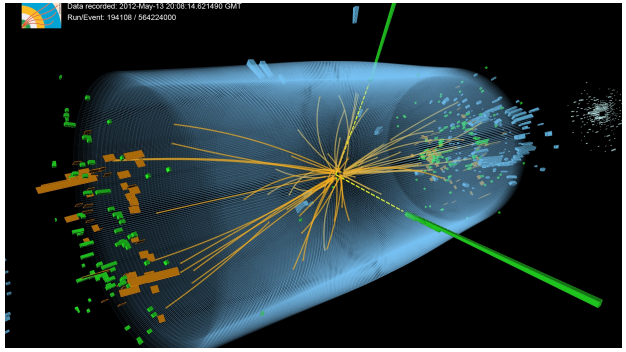
Massive Scientific Instruments generate Big Data



Courtesy S. Myers (IPAC 2012)







Multiple passes of data analysis by large, distributed collaborations

DUNE Collaboration…we have assembled the team

As of today:

960 collaborators from 163 institutions in 31 nations

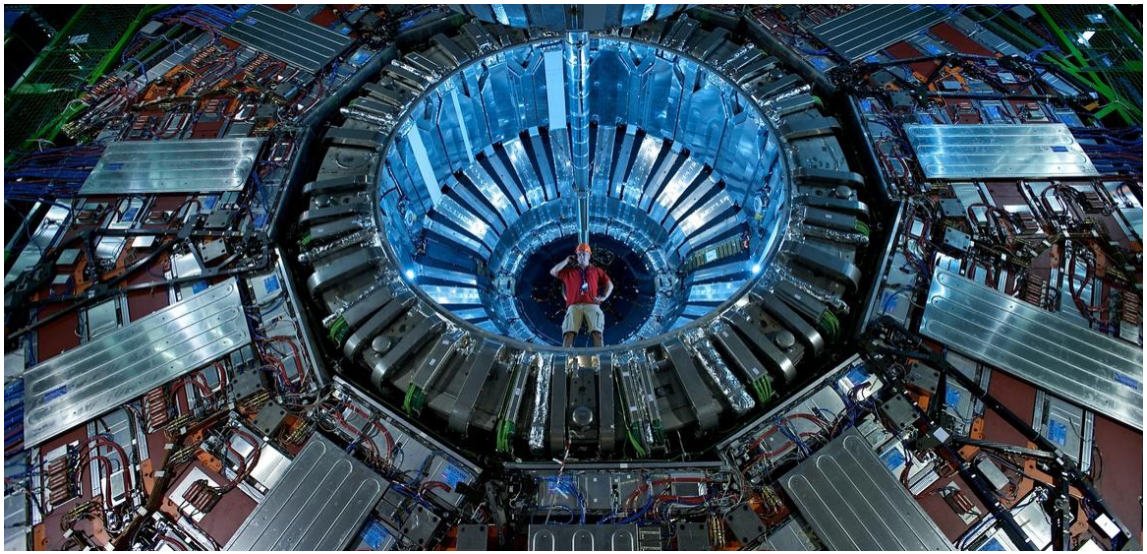
60 % non-US

Armenia, Brazil, Bulgaria, Canada, CERN, Chile, China, Colombia, Czech Republic, Finland, France, Greece, India, Iran, Italy, Japan, Madagascar, Mexico, Netherlands, Peru, Poland, Romania, Russia, South Korea, Spain, Sweden, Switzerland, Turkey, **UK**, Ukraine, USA

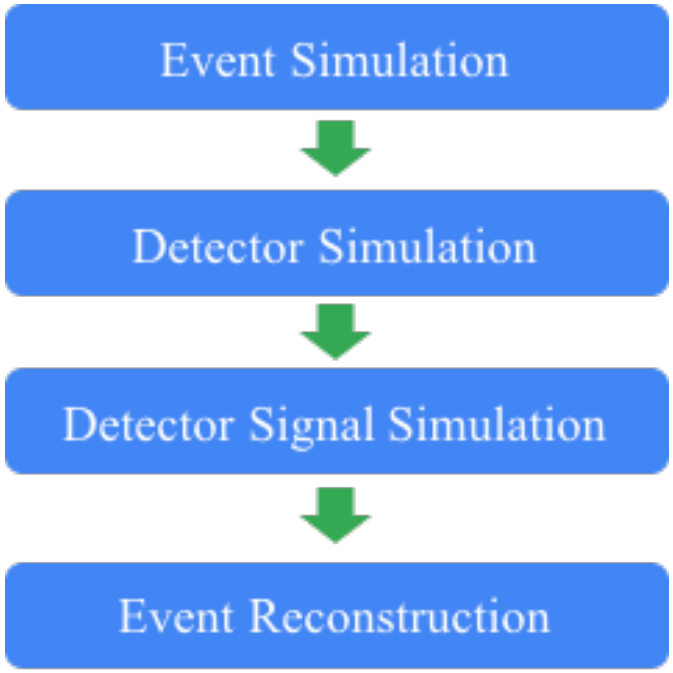
DUNE has broad international support

Compact Muon Solenoid (CMS) at the LHC

- **Protons collide** at the LHC
  ➡ **14 million times per second**
- **100 Megapixel** "camera" captures energy, position
  1000 times per second
  - Measurements during a collision constitute an "**event**"
- Huge number of events (6.3Ms live time in FY16), huge data sets to be analyzed by 1000's of collaborators

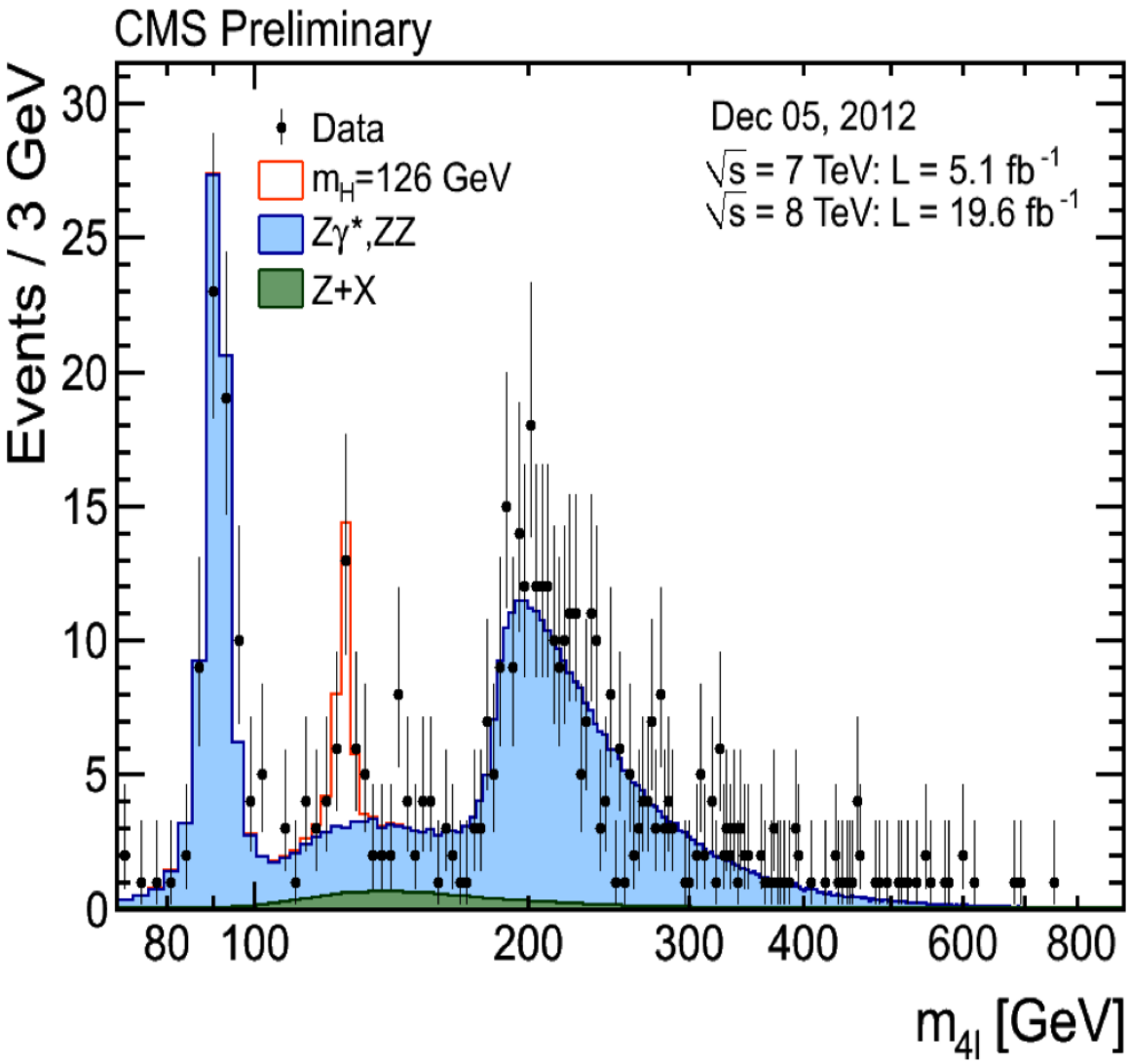


Extreme Simulation Needs



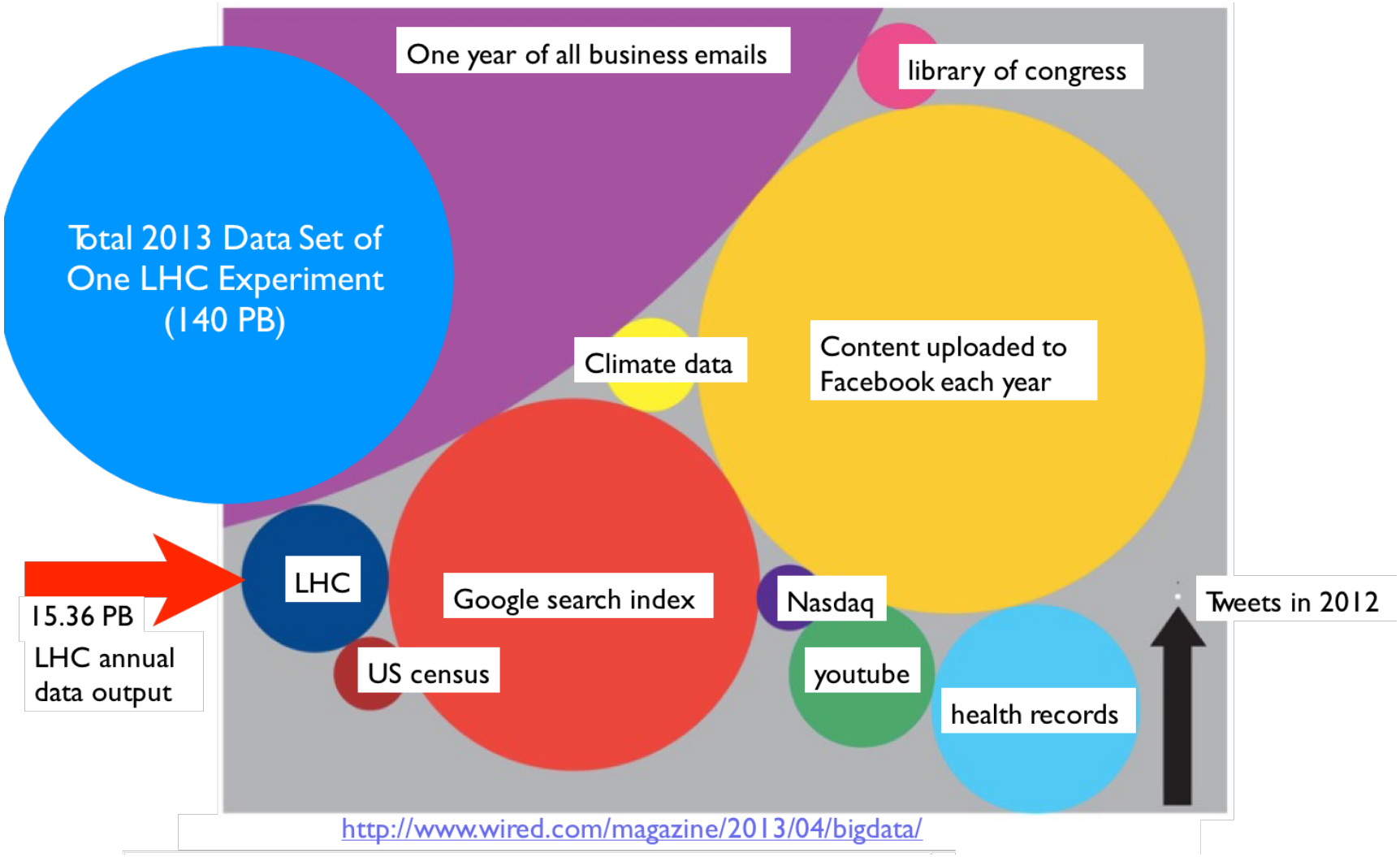
- Particle physics governed by quantum mechanics
  - Initial conditions not known precisely
  - Recorded particle collisions sample a large space of possibilities
- Detector effects, event selection
- We are using probabilistic techniques to sample this space in simulation
  - typical for CMS, **simulated sample x10 of detector sample**
- Analysis of both simulated and detector events necessary to extract physics results

Example: Higgs discovery

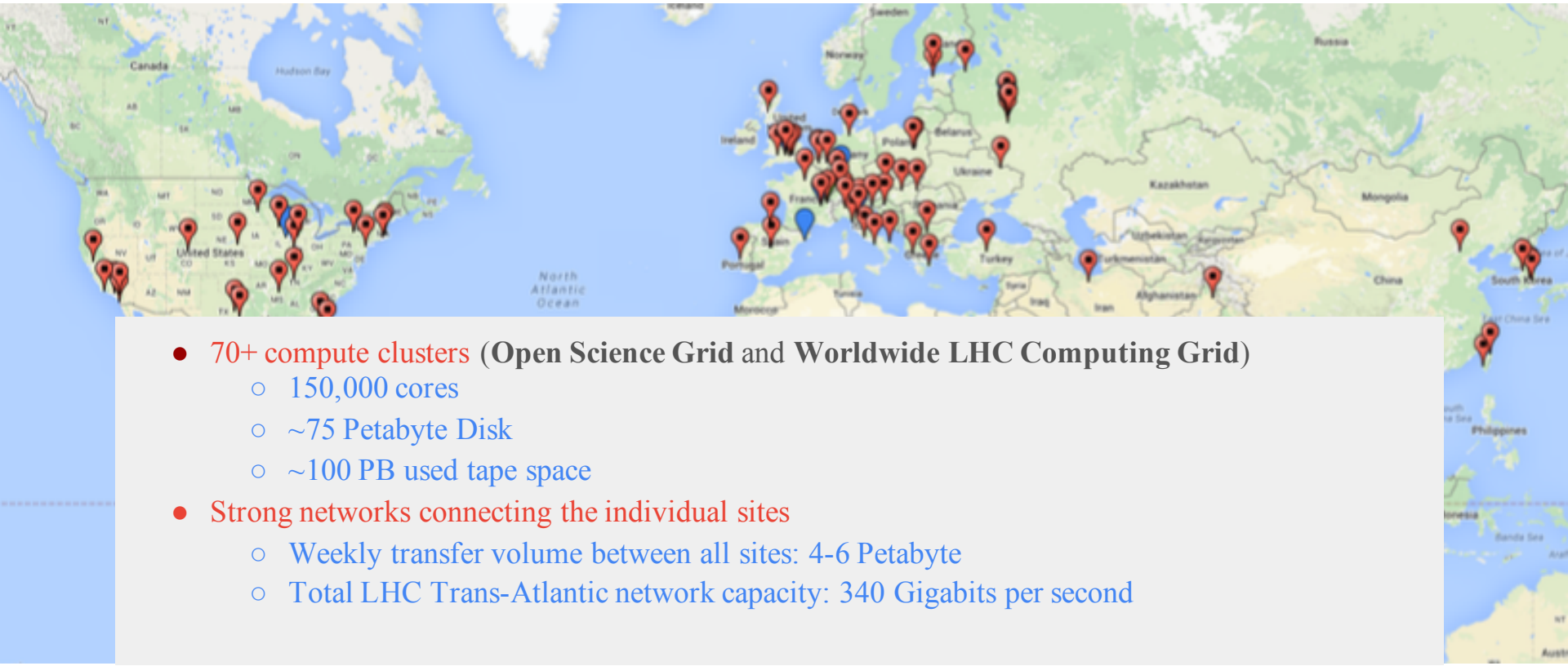


- Comparison with what we know (Simulations)
- Analyze data, compare, and look for deviations ➜ Needle in the Haystack
- Find additional ingredients that match the detector observation

The Big Data Frontier… from "*Wired*"



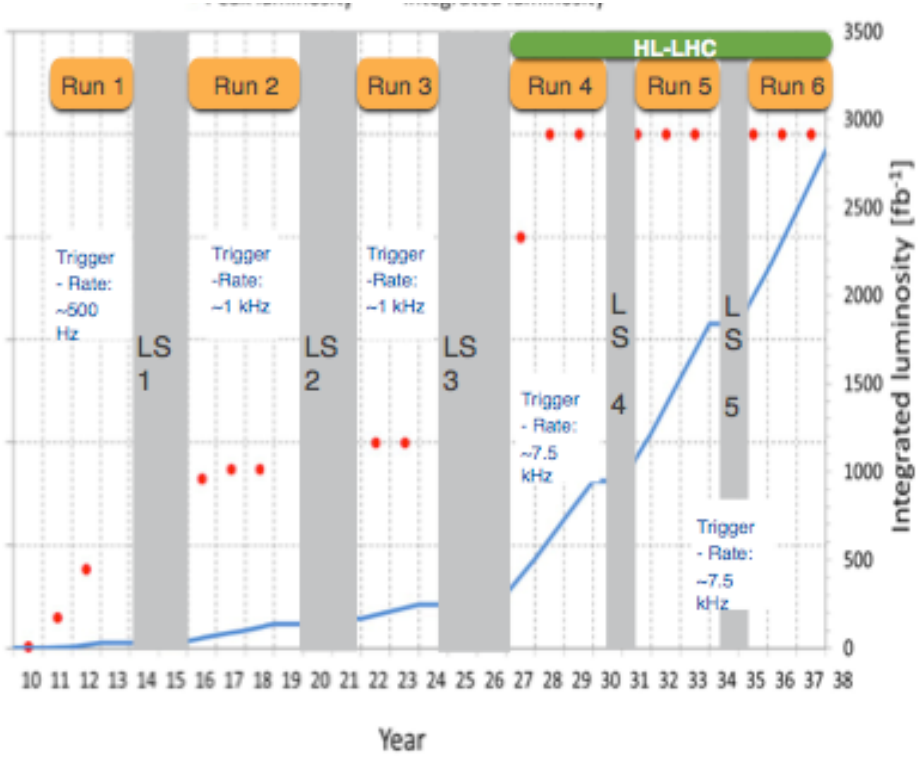
Processing the Data: global CMS (distributed) computing

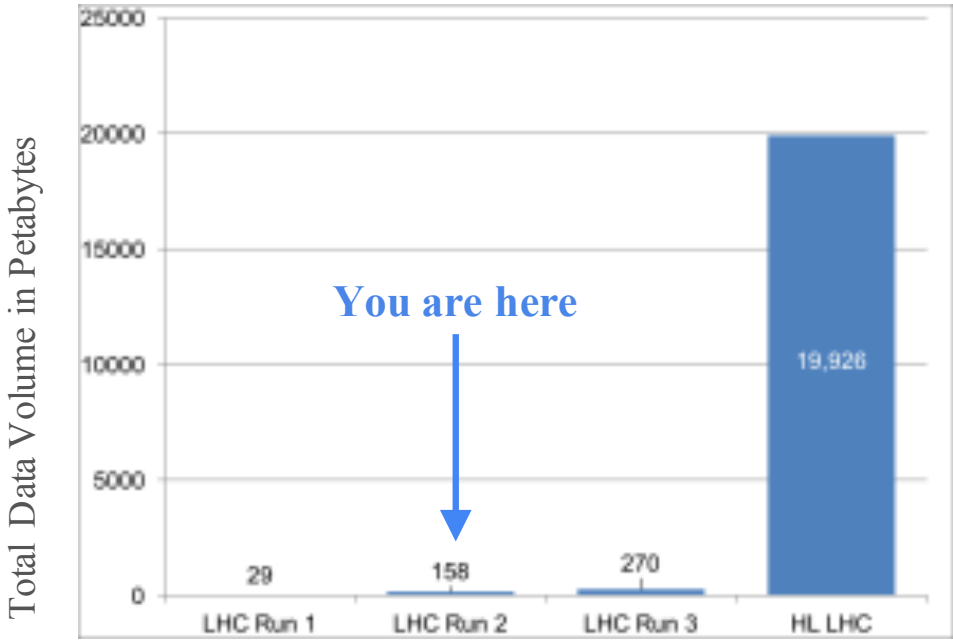


- 70+ compute clusters (**Open Science Grid** and **Worldwide LHC Computing Grid**)
  - 150,000 cores
  - ~75 Petabyte Disk
  - ~100 PB used tape space
- Strong networks connecting the individual sites
  - Weekly transfer volume between all sites: 4-6 Petabyte
  - Total LHC Trans-Atlantic network capacity: 340 Gigabits per second

But this won't be enough as the program evolves…

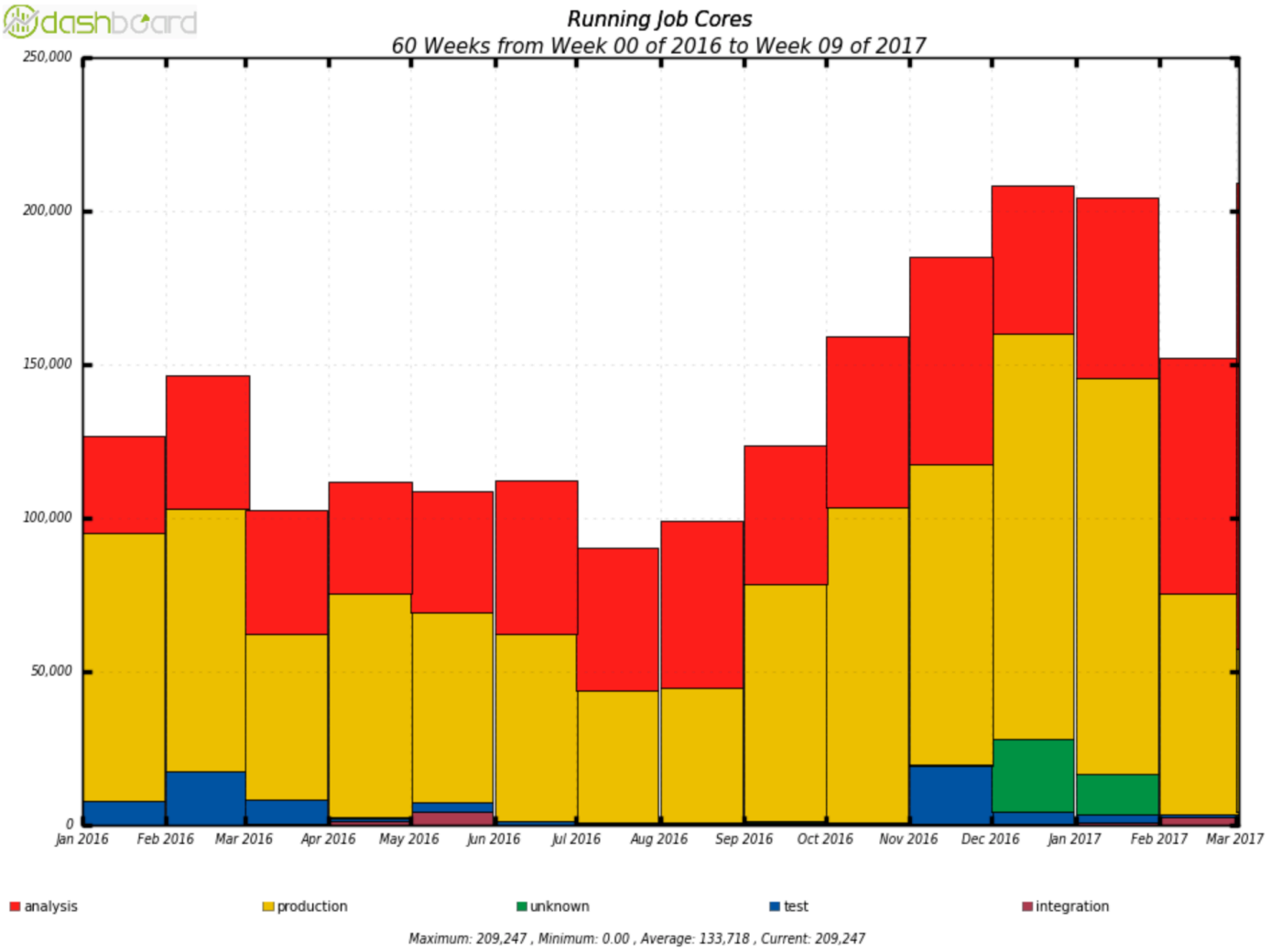
Better Accelerator and Detector performance ➡ increase discovery potential





But, computing will need to reach **10-100x current capacity**

Furthermore, resource utilization not at steady state



A new approach necessary

- Scale of required computing, combined with departure from Moore's law, calls for a paradigm change in procuring and managing computing resources
  - Similar trends for HEP programs outside LHC
- Utilization efficiency and cost-effectiveness is essential
  - Deploying resources when needed (elasticity)
    - ☞ **high frequency provisioning** instead of the traditional "annual procurement cycle"
  - Enables "goal oriented" resource allocation
  - Requires expanding pool of potential resources beyond "owned, on the floor, program specific" resources and developing tools to effectively manage them

The HEPCloud concept

HEPCloud is envisioned as a **portal** to an ecosystem of **diverse computing resources**, commercial or academic, to:

- provide "complete solutions" to users, with agreed-upon levels of service
- route to **local or remote** resources based on policy, workflow requirements, cost, and efficiency of accessing various resources (compute, storage, networking)
- manage user allocation of "owned" resources and supercomputing facilities, and credits to commercial clouds
- provide cost effective and efficient **"elastic" resource deployment**, by utilizing sophisticated **decision engine** and middleware for **automation**.

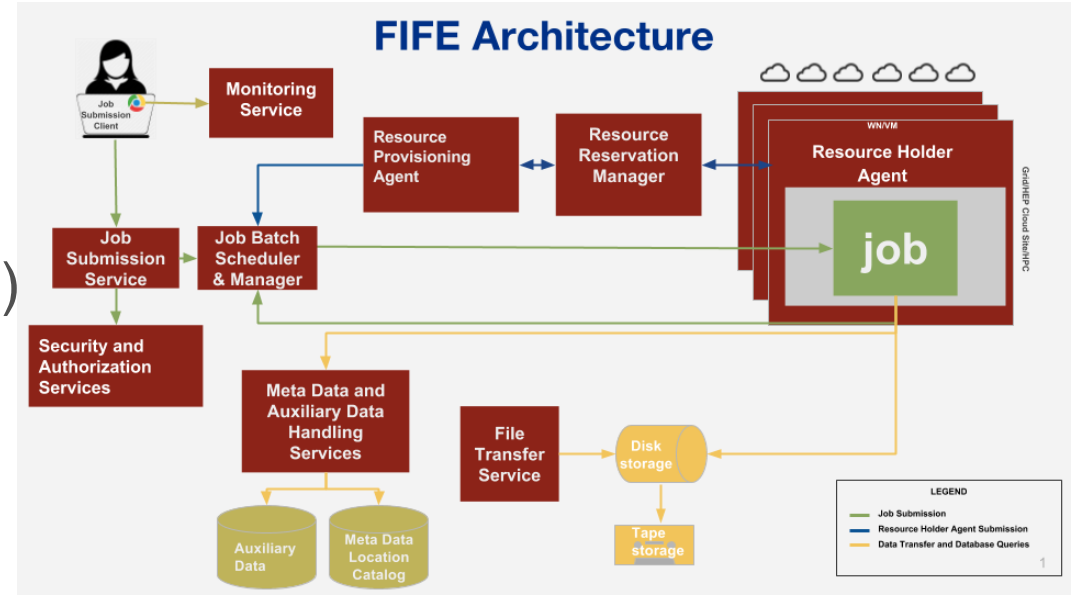
➢ Enabling high-frequency and goal oriented provisioning

HEPCloud pilot project

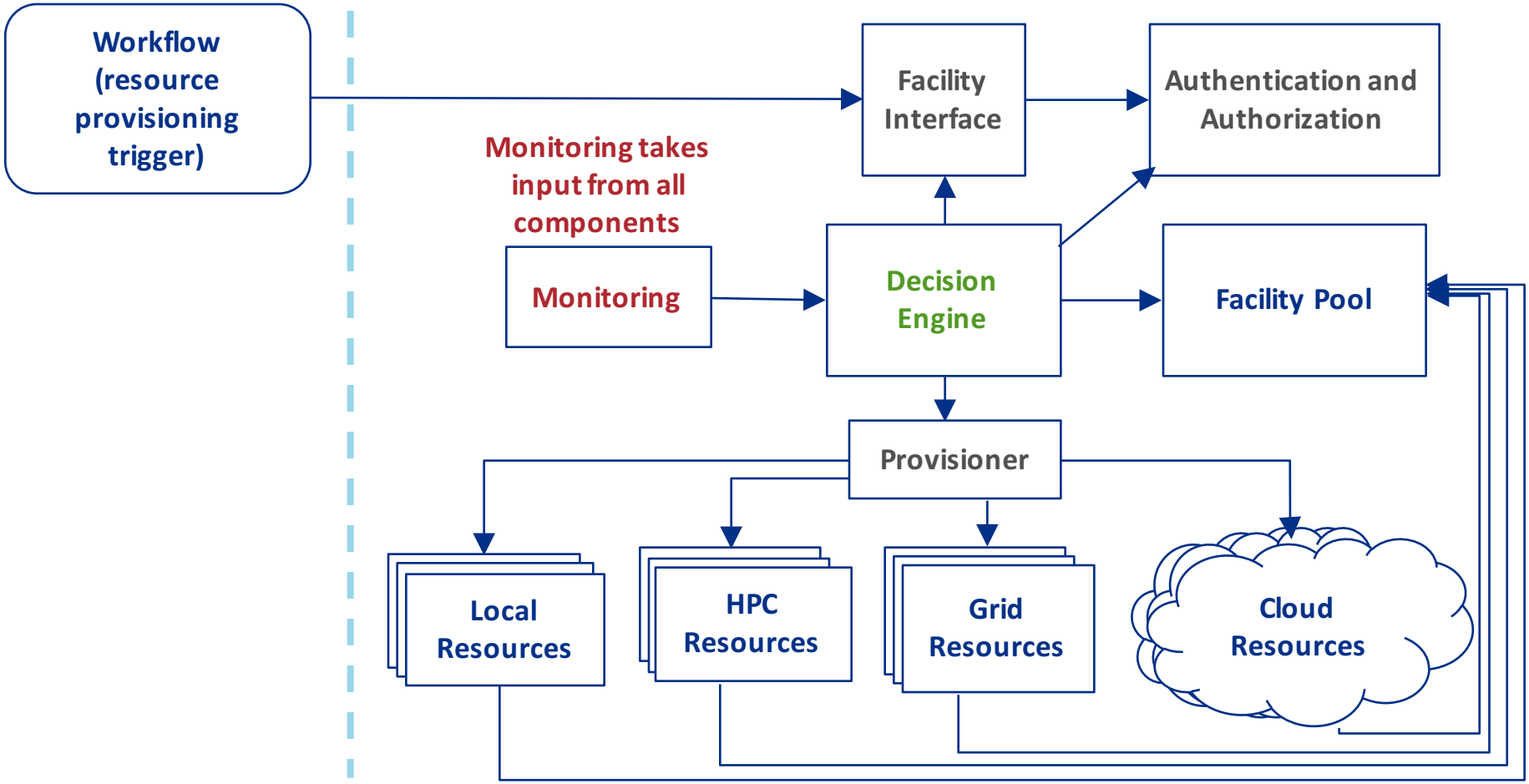
- Pilot project to demonstrate feasibility, explore potential
  - Establish partnerships with industry and academic institutions
- Develop architecture and components
- Mature enough policies, cost model, monitoring and decision engine to
  - Demonstrate scaling and sustainability with commercial clouds
  - Initiate work with HPC facilities to understand constraints and requirements for further developing policies and tools necessary to enable access
    - Identify use cases that are feasible within the constraints of allocation, security and access policy of HPC facilities.

HEPCloud pilot project

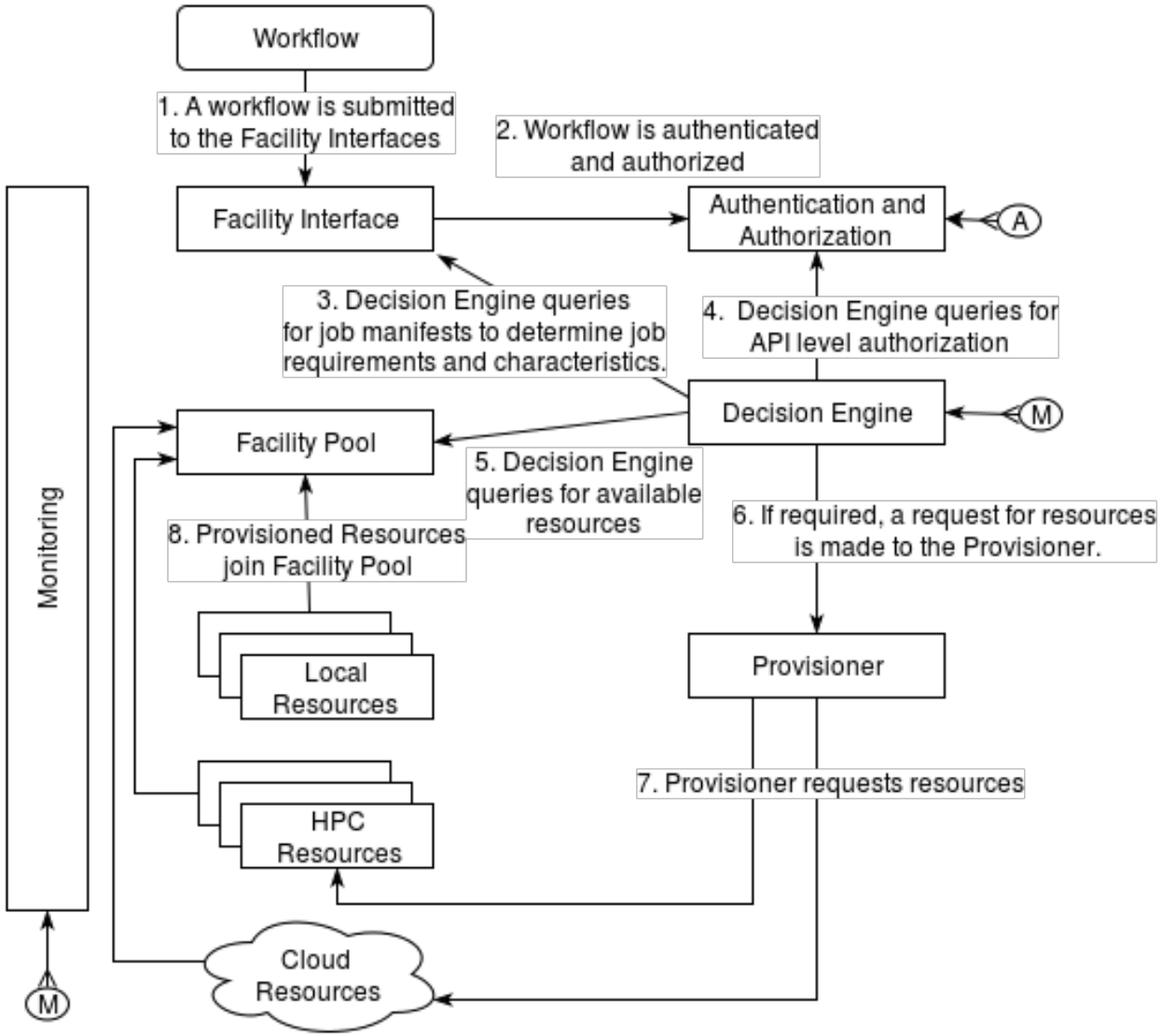
- Utilize Open Science Grid (OSG) stack for provisioning
- Fermilab developed and co-developed tools for workflow management, data management (and their integration), monitoring
- Partnerships:
  - ESnet
  - BNL (ATLAS, Facility)
  - ANL/CCE (edge services)
  - OSG (provisioning)
  - FNAL (CMS, muons, neutrinos, Facility)



HEPCloud architecture

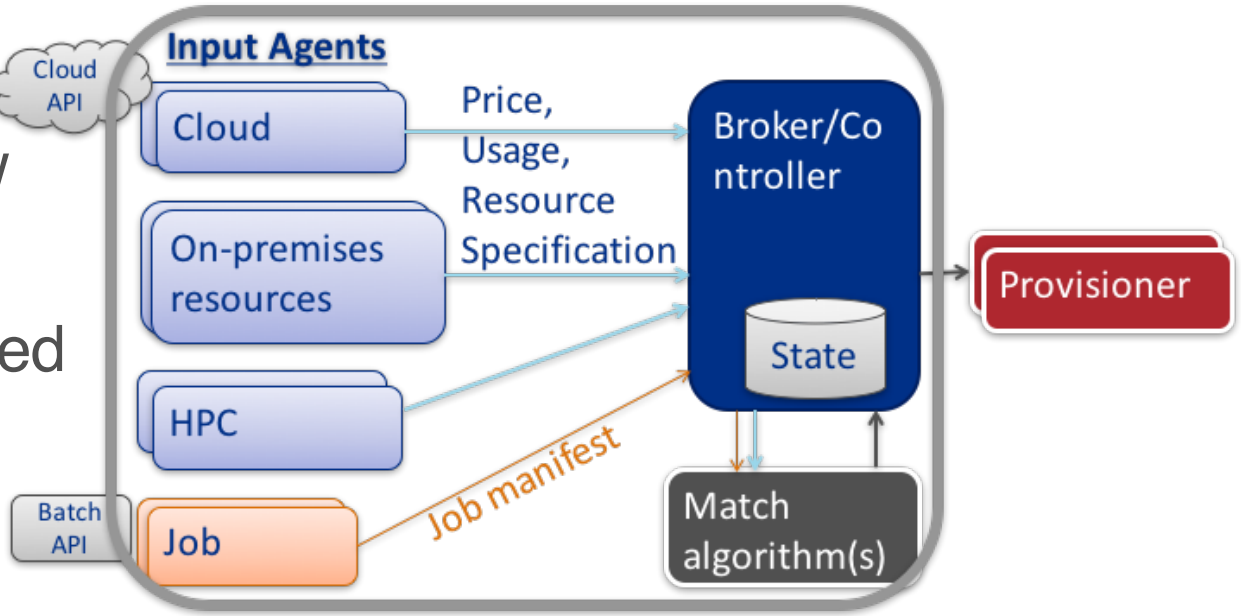


And for the CS types…



The Decision engine

- High Frequency “trading” of resources
- Assign “value” to resources
  - Owned, allocated, commercial credits
  - Data movement/storage, speed/availability
  - …
- Aware of policies, workflow requirements, …
- ➢ Enables automated goal oriented acquisition



HEPCloud pilot use cases: cloud

NoVA Processing

Processing the 2014/2015 dataset
16 4-day "campaigns" over one year
Demonstrates stability, availability, cost-effectiveness
Received Amazon Web Services (AWS) academic grant

CMS Monte Carlo Simulation

Generation (and detector simulation, digitization, reconstruction) of simulated events in time for Moriond16 conference
56000 compute cores, steady-state
Demonstrates scalability
Received AWS academic grant

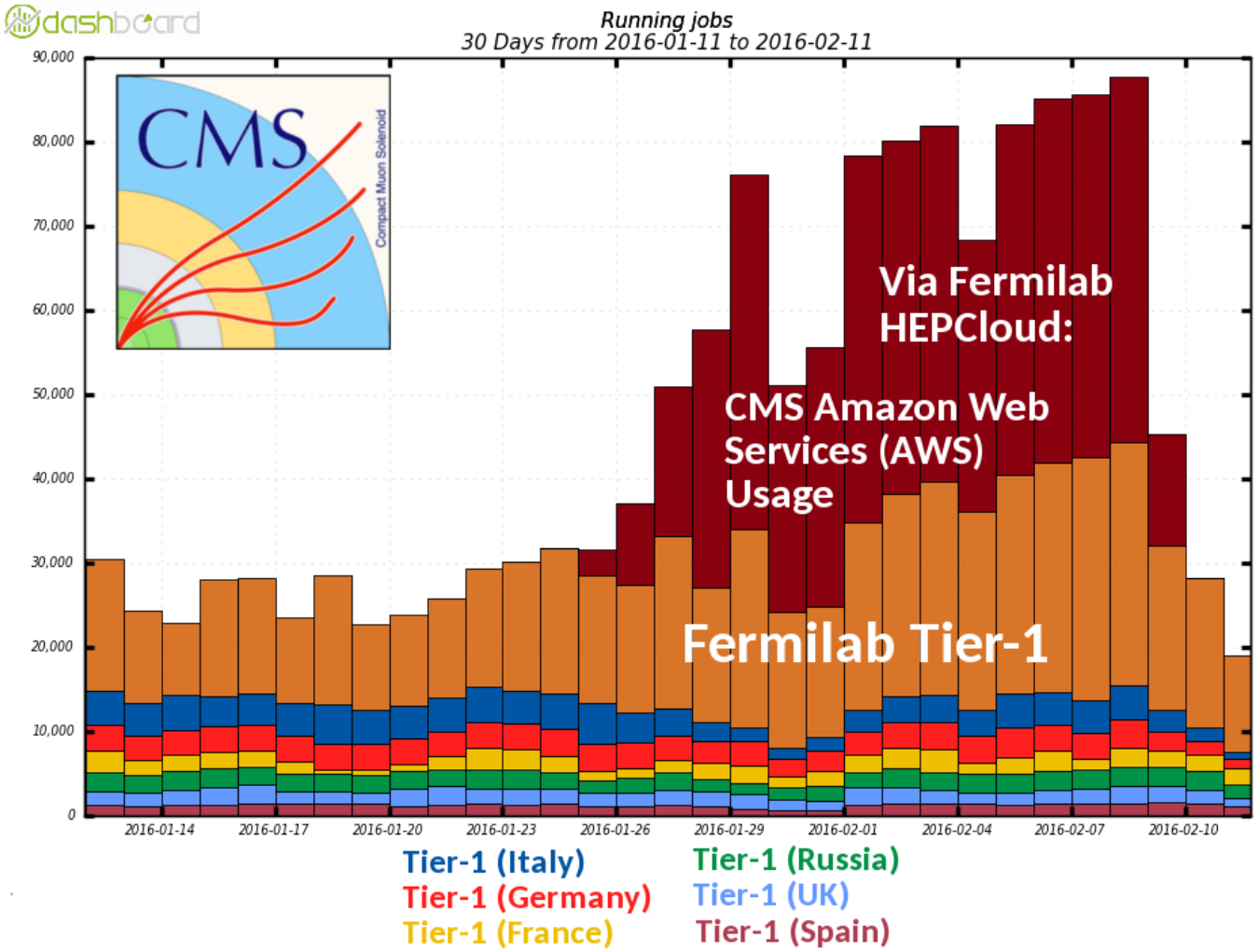
CMS Monte Carlo Simulation

Generation (and detector simulation, digitization, reconstruction) of simulated events in time for Moriond17 conference
160000 compute cores during Supercomputing 2016 conference (~48 h)
Demonstrates scalability, capability
Received Google Cloud Platform grant

mu2e Processing

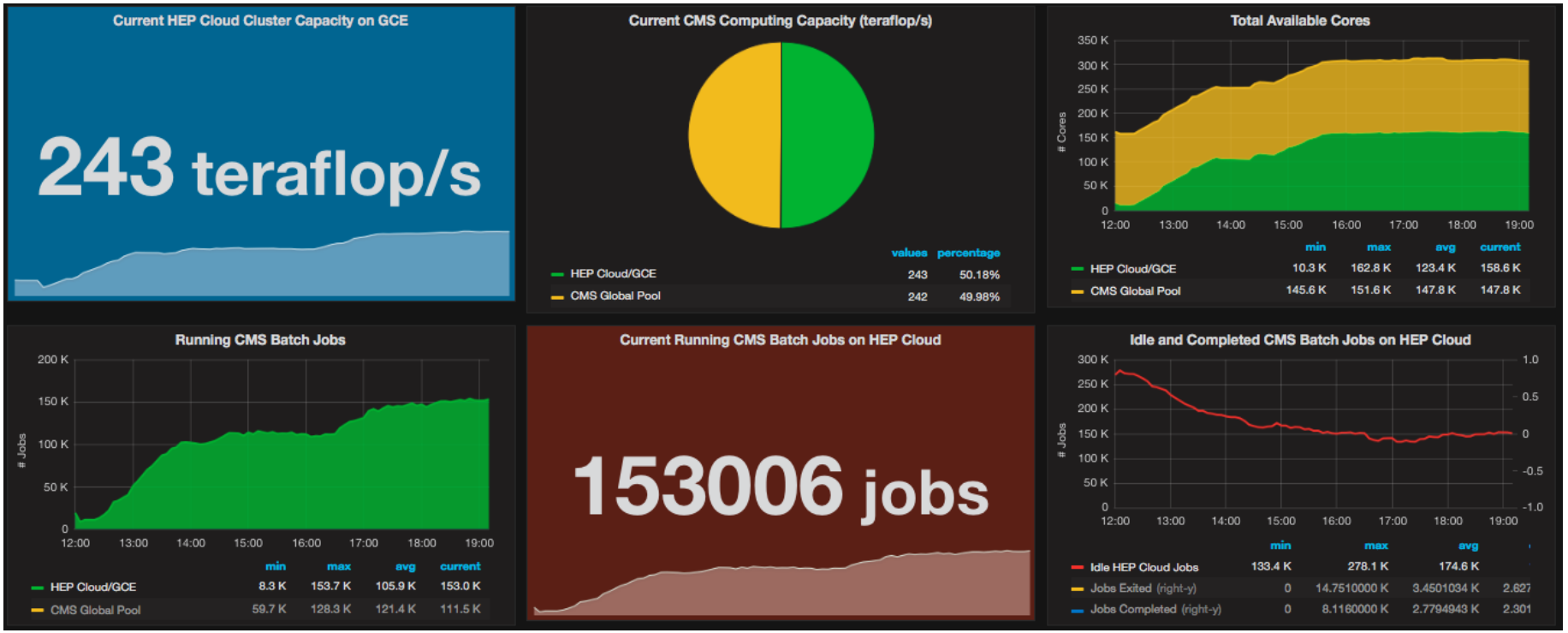
Simulating cosmic ray veto detector and beam particle backgrounds
3M integrated core-hours
Demonstrates rapid on-boarding
Received Google Cloud Platform grant

AWS CMS use case: simulation campaign for Moriond 2016



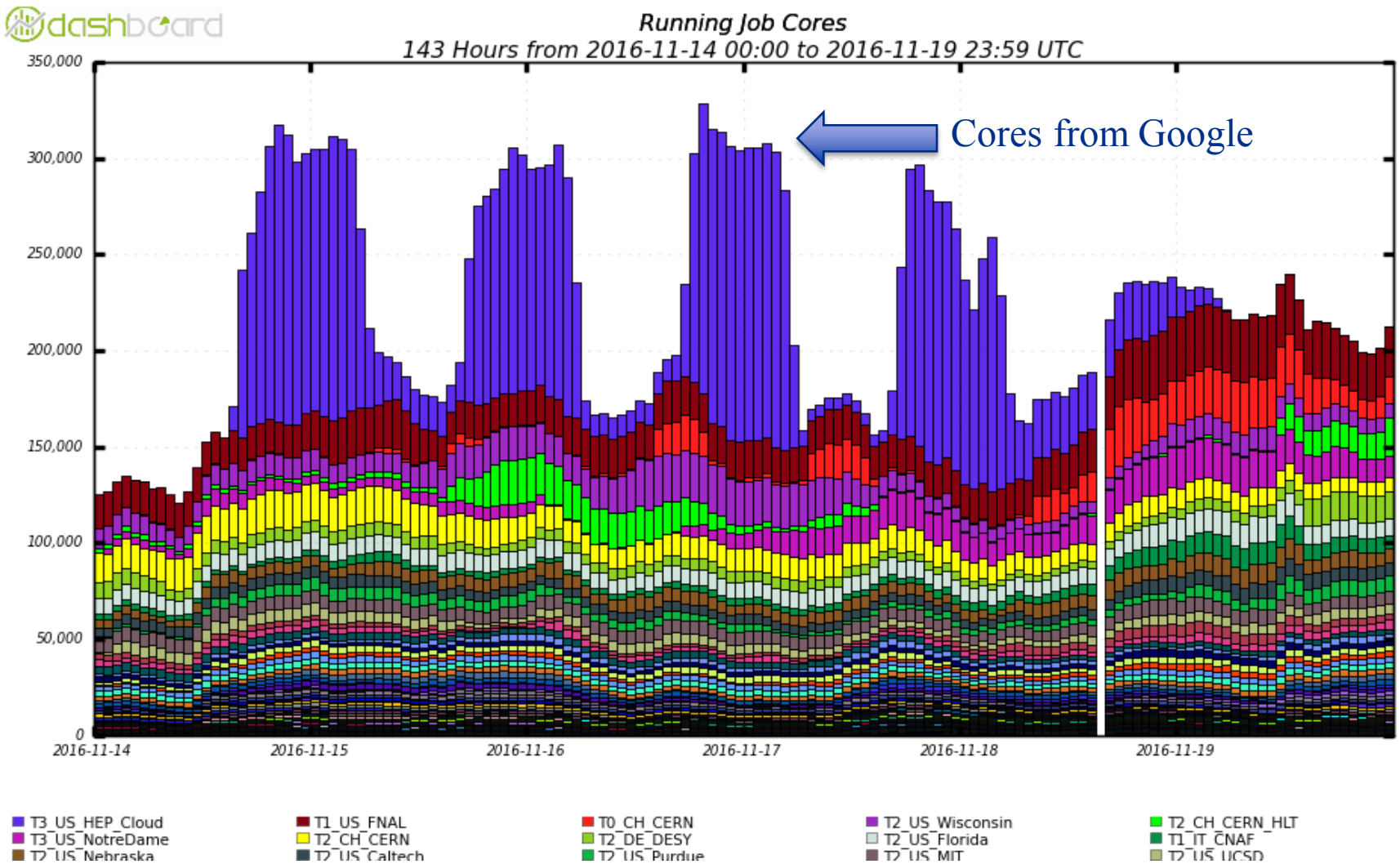
60k slots for ~2weeks, **25% of global CMS capacity,** sustained

Google CMS use case



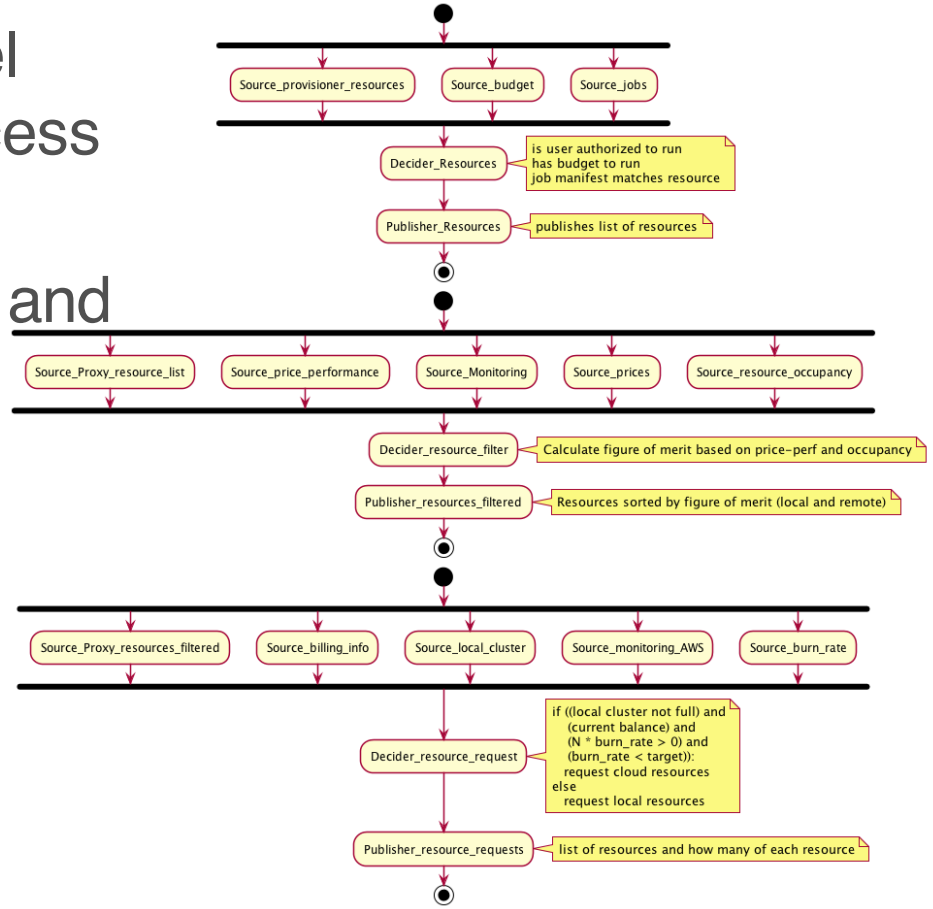
205 M physics events generated, yielding 81.8 TB of data, in 48 hrs, during SC2016

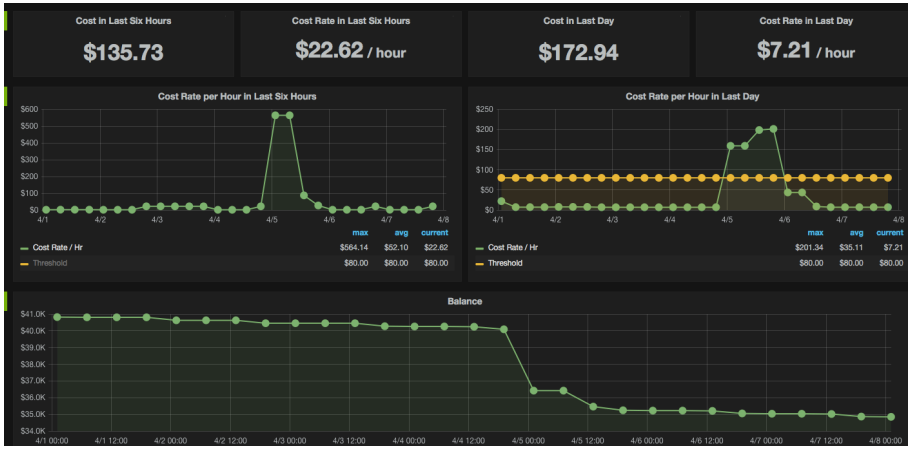
Doubling the size of global CMS capacity



Decision Engine development

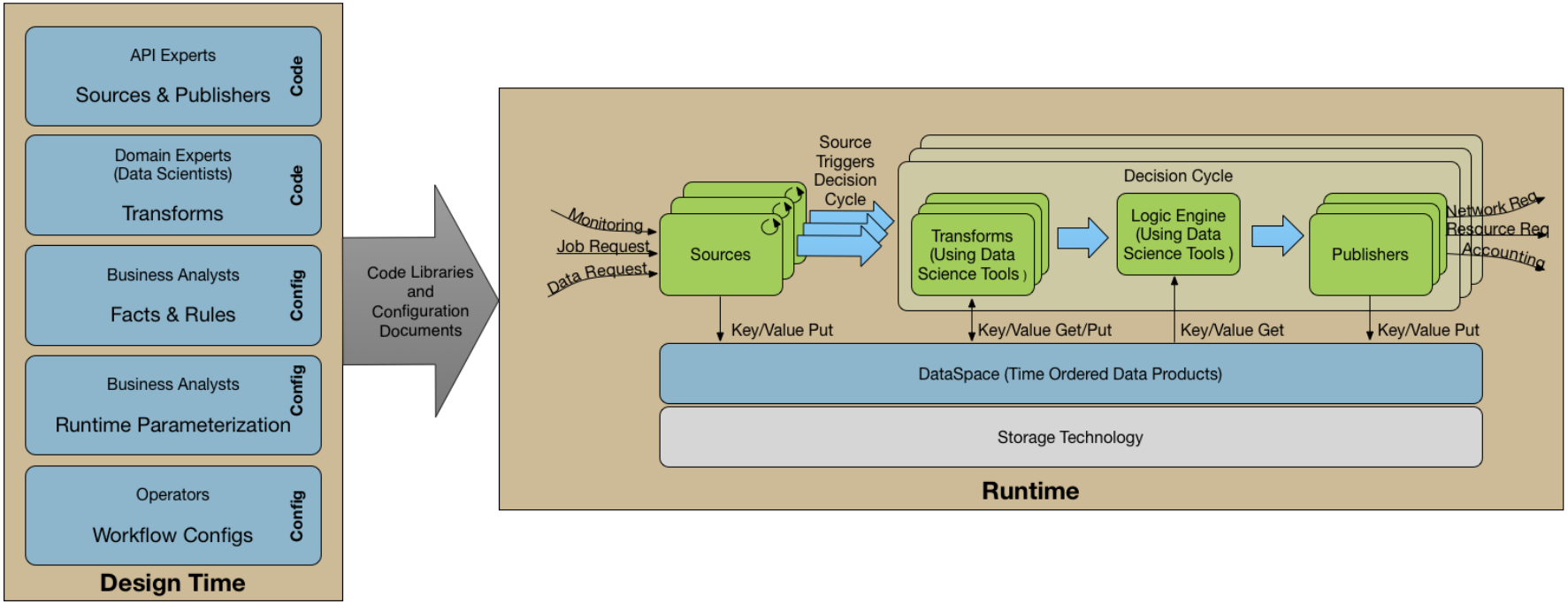
- Prototype focus on Cost Model logic for commercial cloud access
- Current focus: complete architecture, incorporate HPC and corresponding policy modules





Proto-prototype Decision Engine implementation

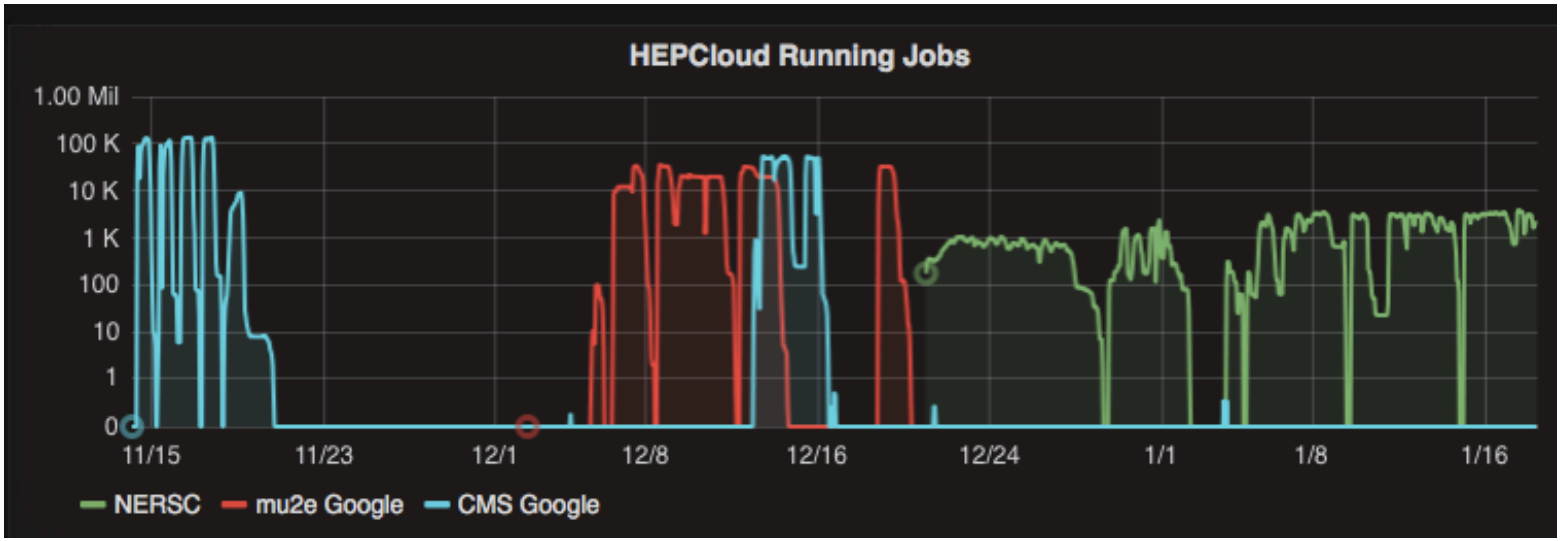
Decision Engine Architecture Summary



HEPCloud pilot project on HPC

- **Early steps**: adapt HTC workflows to HPC facilities
  - Neutrino experiment production on Cori @ NERSC
  - Physics Generators on Mira @ ALCF: multi-parameter tuning of event generators using collider data
  - **CMS production on Edison, Cori @ NERSC**: provisioned resources and executed various Generator-Detector Simulation-Reconstruction workflows
- Current status:
  - run corresponding workflows through the HEPCloud production team @ NERSC
    - Identify and implement necessary additional security controls and processes
  - Investigate “edge service” solutions, security policies and controls with ALCF

Starting production on NERSC



- Use cases have been defined
- Ramping up processing
  - Identifying/resolving issues with data movement efficiencies, performance, …

HEPCloud and HPC

- An intresting possibility, for the exascale era, is to consider HEPCloud as a common layer for ”offering” exascale facility services to data-driven science domains
  - and, in the other direction, though HEPCloud services (e.g. storage/data management) integrate other resources (especiaally data centers) to the exascale facilities
- The HEPCloud concepts are not HEP specific, other data driven sciences have similar needs (as we learned at the exascale crosscut review)
  - Common layer for accessing facility (authentication, policy, tools…); software deployment; workflow and data management integration;…

➢ So, HEPCloud could evolve as a common layer across different science domains (DOECloud?) –and, possibly, serve as an attraction point for further developing cross-cut cloud strategy and infrastructure

Summary

- Successful demonstration of concept with commercial cloud and NERSC use cases
  - Plan to transition aspects of HEPCloud to production end of 2018
- Continuing R&D for Decision engine, policy development, integration and development of software infrastructure…
  - **Partnerships with other domain sciences and ASCR welcome!**
- Next stage: focus on incorporating HPC resources
  - Build on work to provision cycles on Edison, Cori at NERSC
- Extend HPC use cases to non-pleasingly parallel problems
  - Deep learning, …
- Complete interfaces with commercial cloud providers
  - Done: Google Cloud Platform, Amazon Web Services; Next: Microsoft Azure, …
- Long term: Explore possibility of **partnerships** in integrating concept to exascale facilities software infrastructure

Backups

HEPCloud and exascale

- Big Data on exascale: more diverse, larger data sets; larger, distributed data set ensembles; many stages of analysis and several “passes” of data processing; need to get data in and out of the exascale facility
  - Different from the current LCF “Big Data” paradigm (data flow networks for analysis of large, structured data sets generated by massive simulations)
  - Assuming utilization of exascale for processing instrument (observational) data and simulated data for modeling response of instruments
- Workflow and data management, tracking provenance, …; desirable to present exascale facilities to users using common interface, tools, access, means to effectively utilize allocations on different facilities based on load and status
  - Functionality essential even if LCFs remain focused on running massive simulations, since these will have to be coupled to observational data outside the LCF
- The HEPCloud concept could provide this functionality

Research/Partnership Areas

- Developing policies and interfaces for LCF
- Data management
- Debugging and Monitoring
- Workflow end-to-end execution automation, including resource management
- Workflow ensemble management
- Automated workflow lifecycle
- Failure Management
- Integrating Advanced Networking to Decision Engine and Monitoring
- Exploring in-situ analysis approaches, especially for filtering simulated data
  - Becomes desirable if exascale machines used extensively for simulation campaigns

Accessing NERSC and LCFs

- FNAL security team (on behalf of HEPCloud) interacted with NERSC (through our liaison Lisa Gerhardt) and ALCF security team (contact Piotr Zbiegiel)
- Analyzed access models for each facility and mapped to HEPCloud access model and controls
- Mapping was successful for production teams, after modifications of HEPCloud controls (sensitive country)
  - We are operating at NERSC under the pilot project. If we move to operations, we could further advance by better integration of security teams (monitoring information, advancing common controls, …)
  - Haven't closed the loop with ALCF. MFA requirement an additional change needed on our end for automated production workflow management (not an issue/reason for not proceeding)


Overall view of the LHC experiments.
LHC - B
Point 8
CERN
ATLAS
Point 1
ALICE
Point 2
CMS
Point 5
SPS
TI 8
TI 2
LHC - B
ATLAS
ALICE
CMS
LEP/LHC
Battery Spencer
Alcatraz Island
Fort Point
Golden Gate
Baker Beach
SEA CLIFF
OUTER RICHMOND
Golden Gate Park
16th Avenue Tiled
OUTER SUNSET
BERNAL HEIGHTS
GLEN PARK
Radius 2.6
Units
Click to place a circle, right click to remove
Google
Map data ©2017 Google