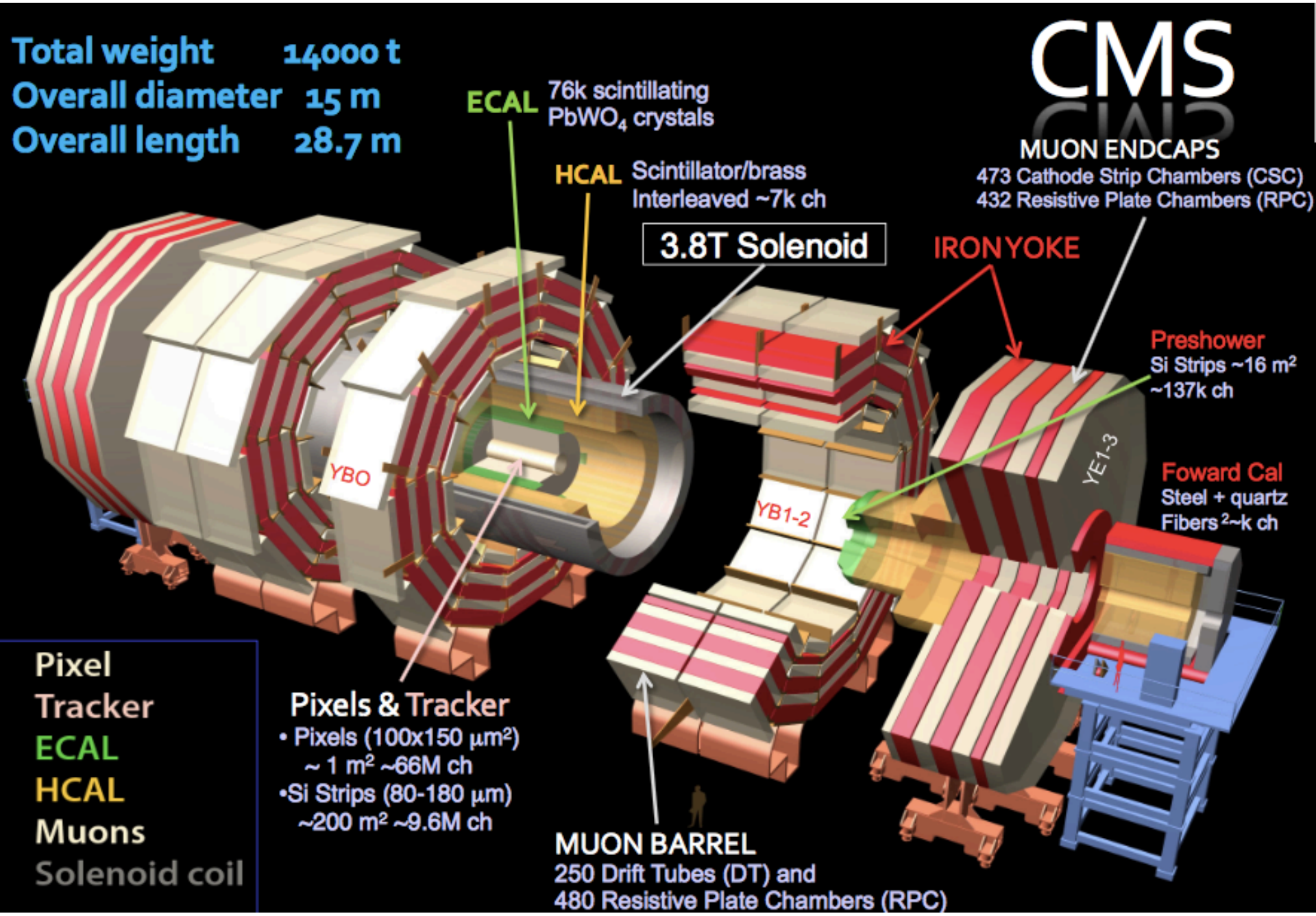
CMS software engineering and computing facilities

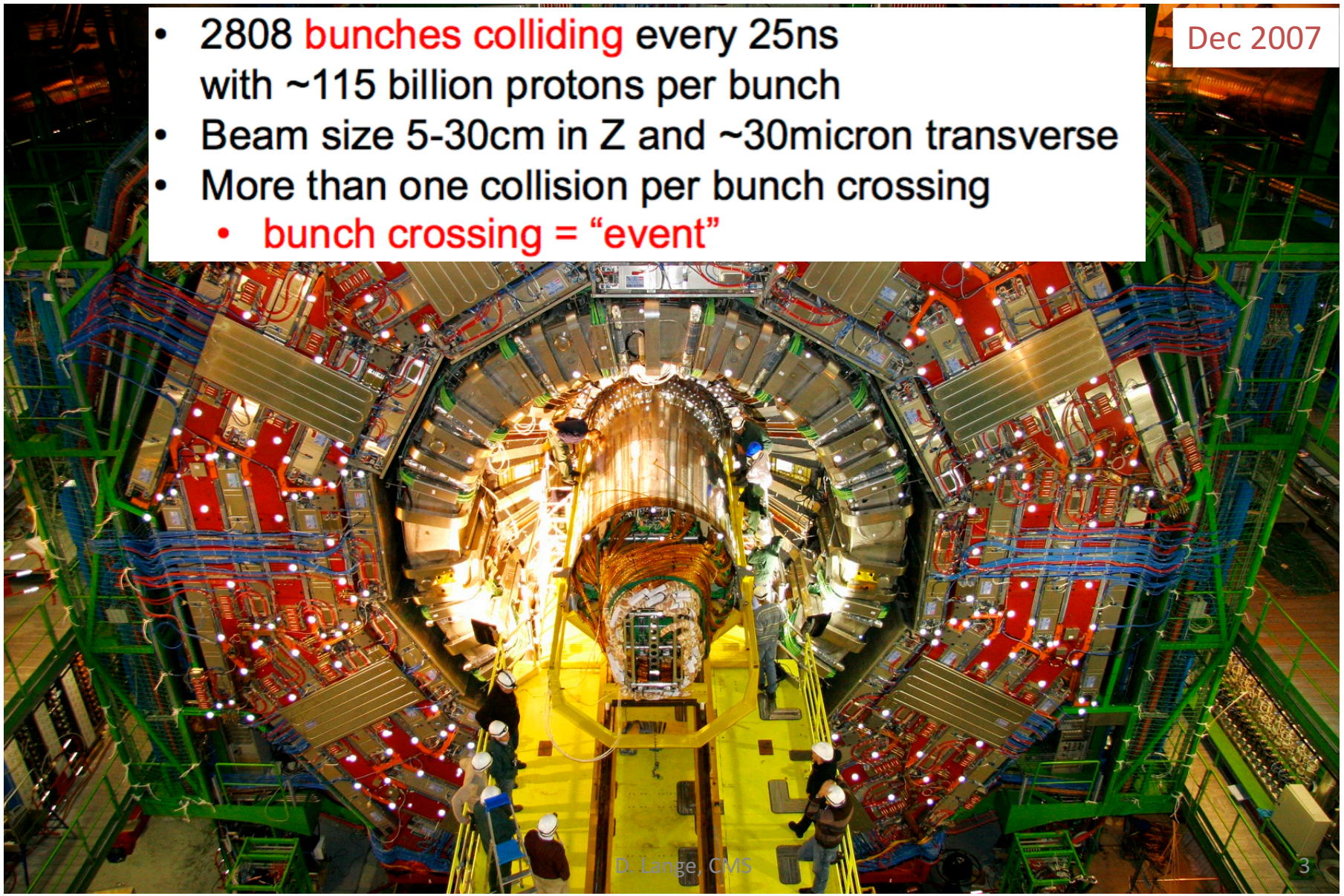
David Lange

December 13, 2016

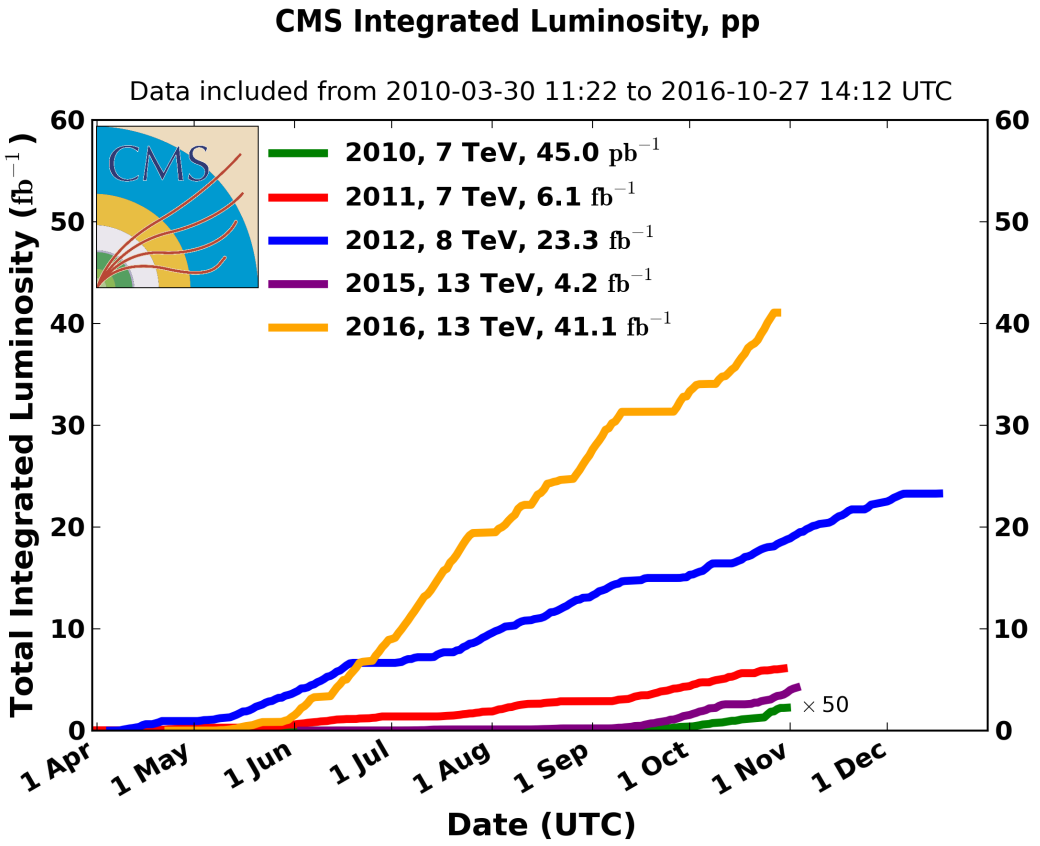
PRINCETON UNIVERSITY


Total weight 14000 t
Overall diameter 15 m
Overall length 28.7 m
CMS
ECAL 76k scintillating PbWO4 crystals
HCAL Scintillator/brass Interleaved ~7k ch
3.8T Solenoid
IRON YOKE
MUON ENDCAPS
473 Cathode Strip Chambers (CSC)
432 Resistive Plate Chambers (RPC)
Preshower
Si Strips ~16 m²
~137k ch
Foward Cal
Steel + quartz Fibers ²~k ch
YBO
YB1-2
YE1-3
Pixel
Tracker
ECAL
HCAL
Muons
Solenoid coil
Pixels & Tracker
• Pixels (100x150 μm²) ~ 1 m² ~66M ch
•Si Strips (80-180 μm) ~200 m² ~9.6M ch
MUON BARREL
250 Drift Tubes (DT) and
480 Resistive Plate Chambers (RPC)

A slice of the CMS experiment



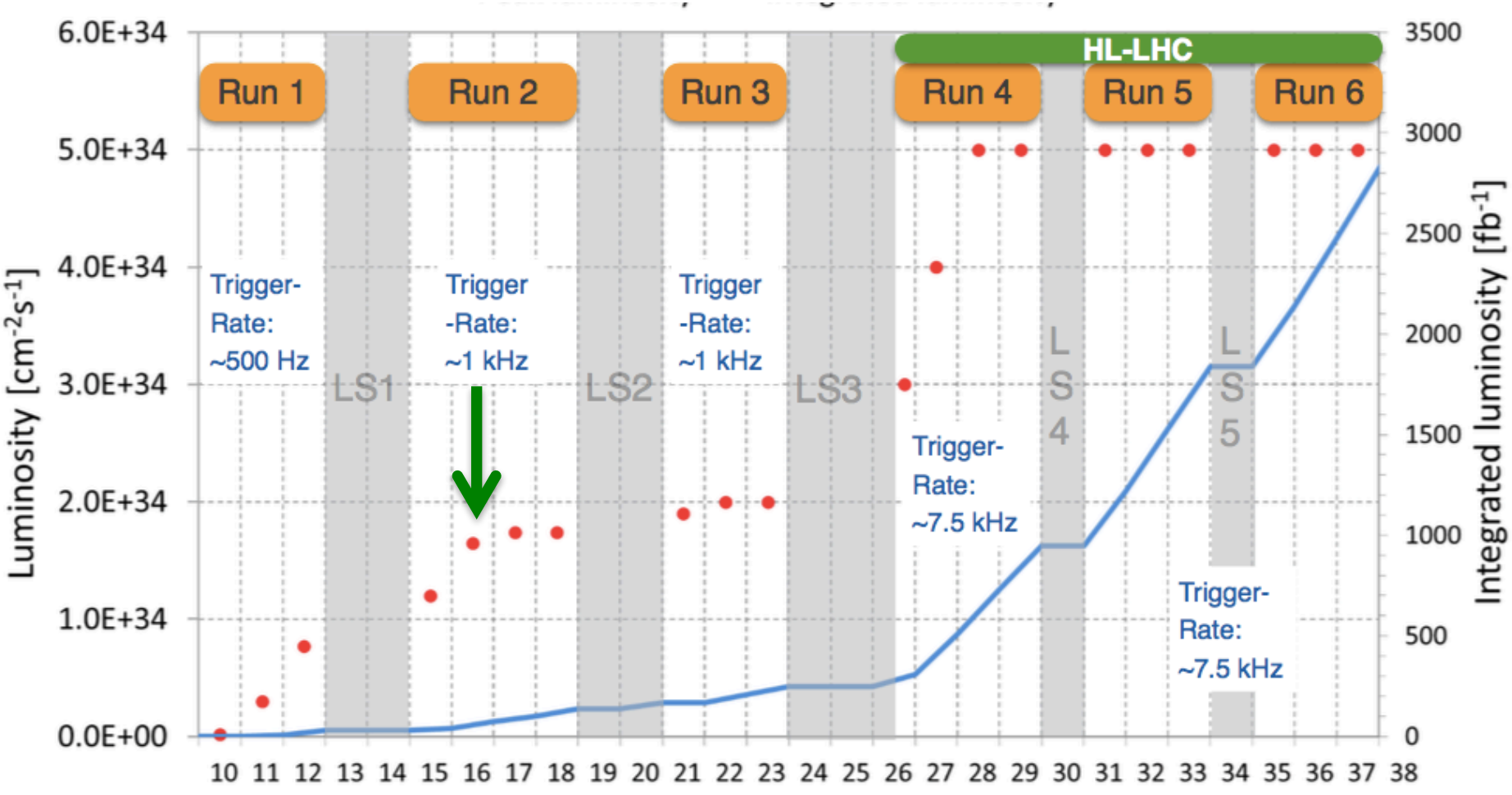
2016 has been an exciting year



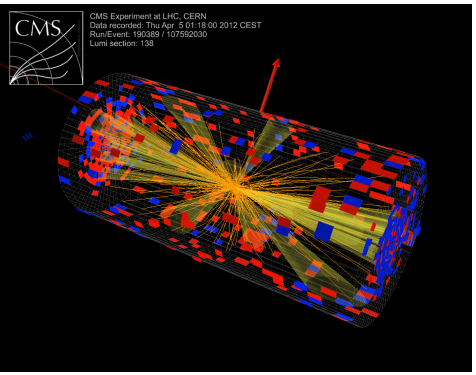
- More data than ever.
- More data than expected

Production computing and analyzers have to be more efficient than before

The long-range LHC schedule:
20 more years of exciting physics

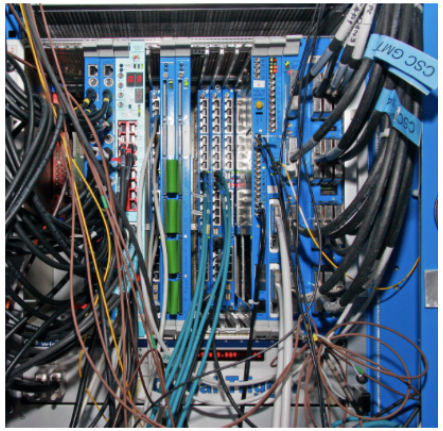


CMS data pipeline



- CMS Detector:

40 TB/sec



- First stage trigger (FPGAs):

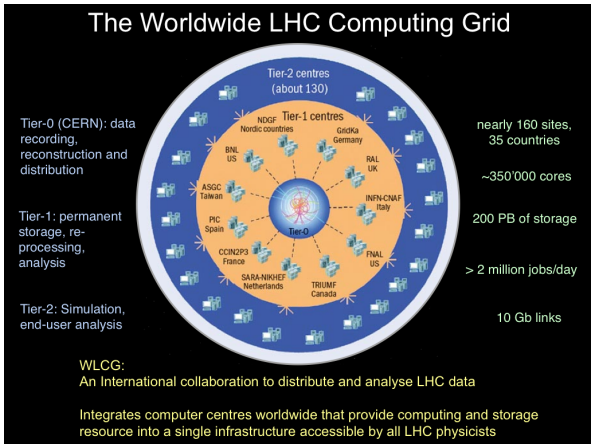
100 GB/s

- Second stage trigger (CPUs):

1 GB/s

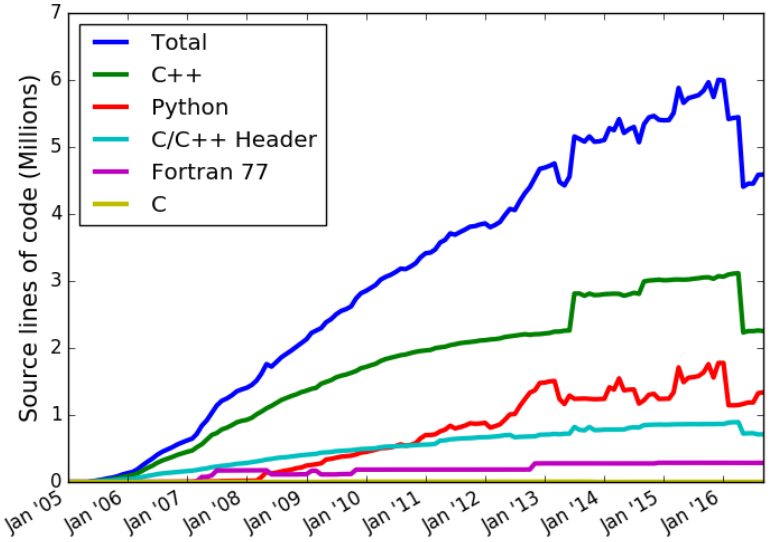
- Analysis data (Distributed):

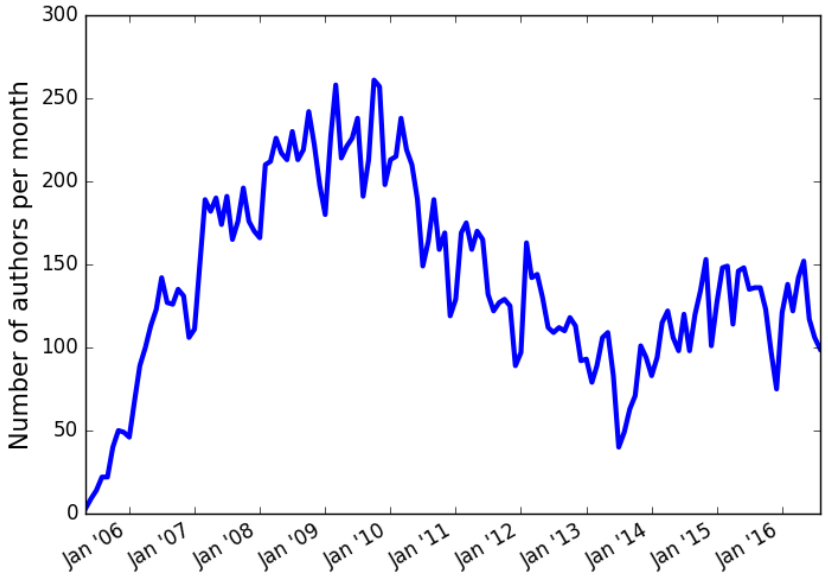
0.4 GB/s



CMS has a global effort to develop its computing and software

Going from raw data from the detector to analysis formatted and high-quality data requires complex and reliable reconstruction, simulation, calibration, monitoring software algorithms





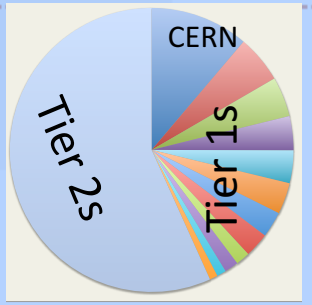
100-200 people contribute to CMS software each month!

Worldwide LHC Computing grid (WLCG)

Global data movement: 15 GB/s

1.5 PB/week recorded
2-3 GB/s from CERN

170 sites, ~8000 users
nearly 40 countries



Resource distribution

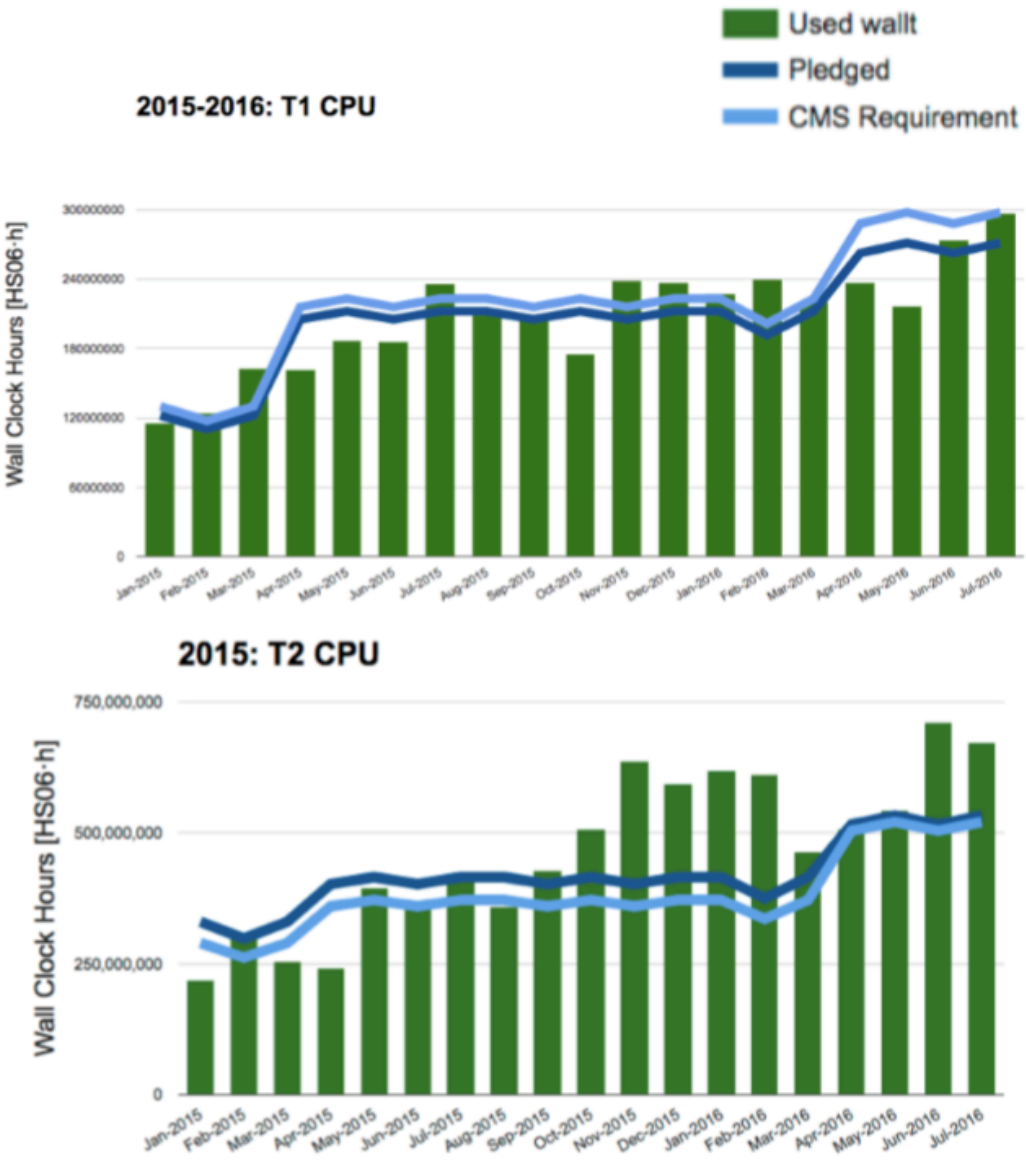
2 M jobs / day

250 000 CPU days/day

200PB Storage

Recent GRID usage for CMS compared with expectations

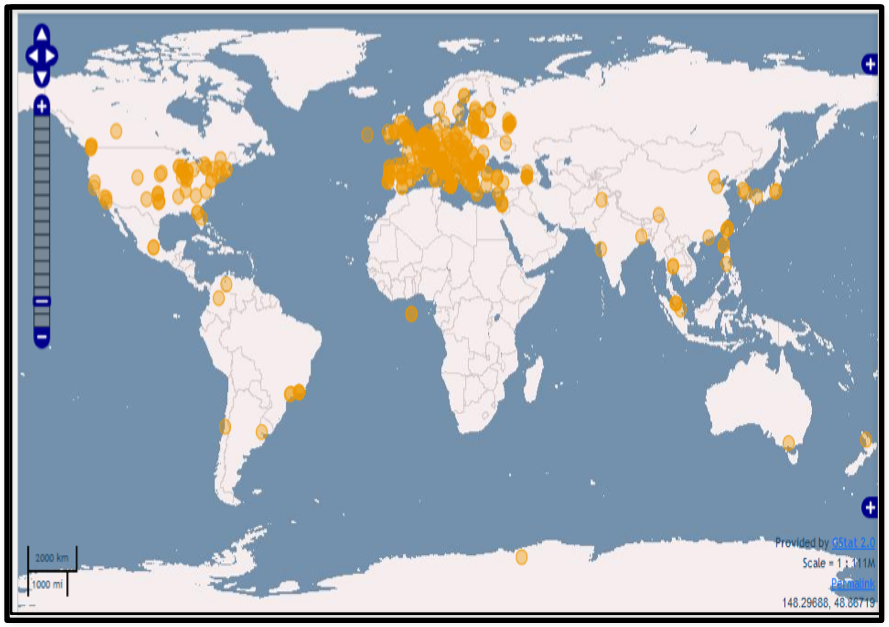
- GRID resources are dedicated for CMS (even if part of shared facilities).
- Part of our job is to keep them busy and to use them efficiently
- Grid computing facilities continue to evolve to increase their flexibility:
  - Workflows handling
  - Reducing time needed to produce analysis datasets



The scale of data Management in HEP

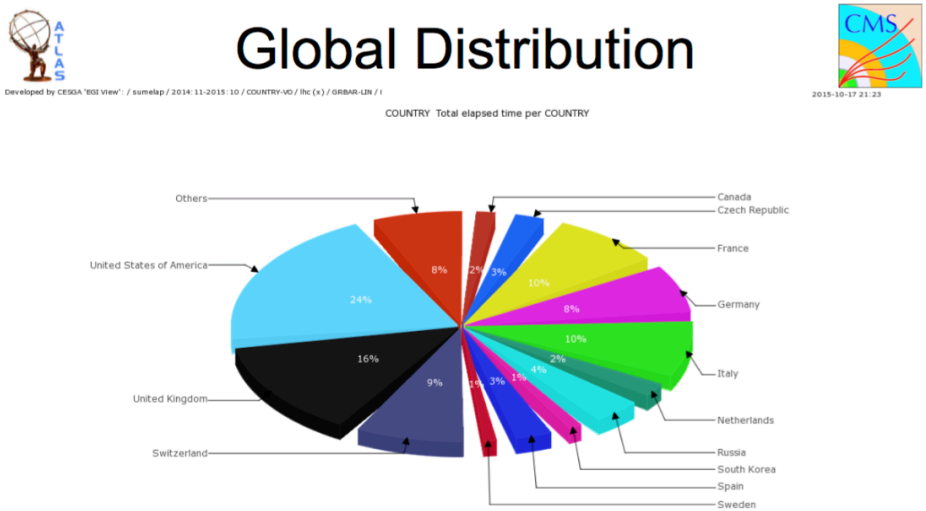
There are close to 200 sites in WLCG

- 246 PB of disk
- 267 PB of tape



WLCG has 140PB of unique data and 280PB under management

- More than 1B files
- Average file size 0.2GB to 2.5GB

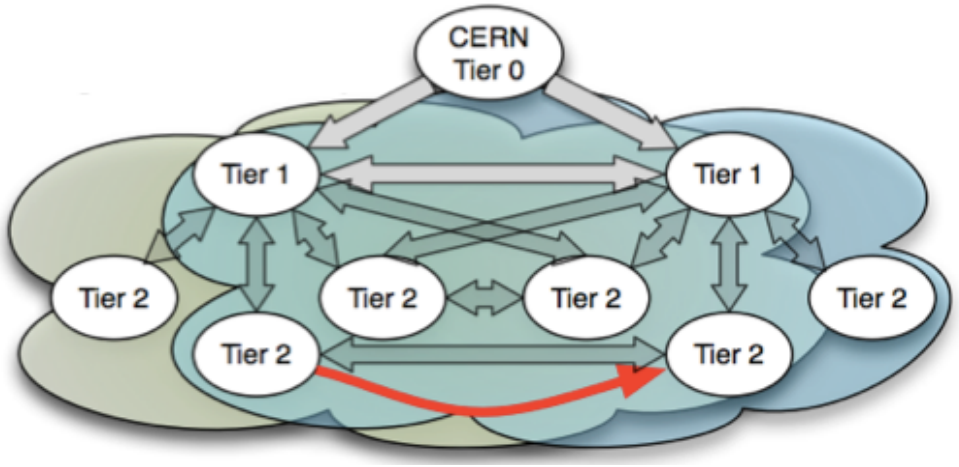


Largest national contribution is only 24% of total resources.

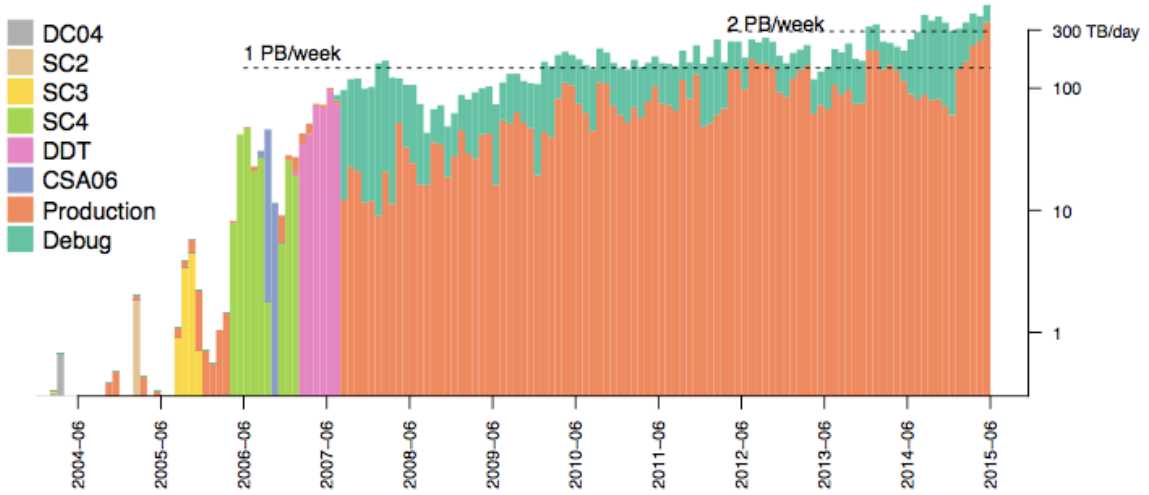
We rely heavily on good network performance

Our challenge:

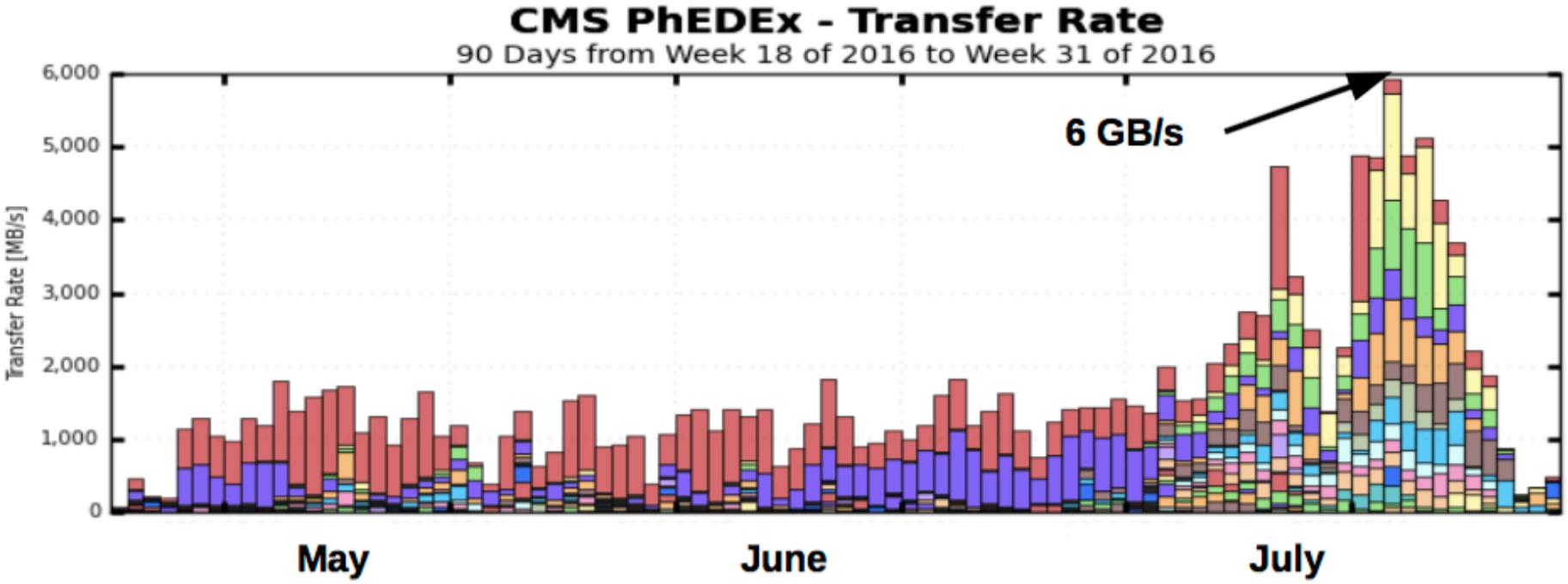
- How to deliver samples to 150k processor cores as directed by the experiment centrally and thousands of scientists



CMS transfers more than 2 PB per week



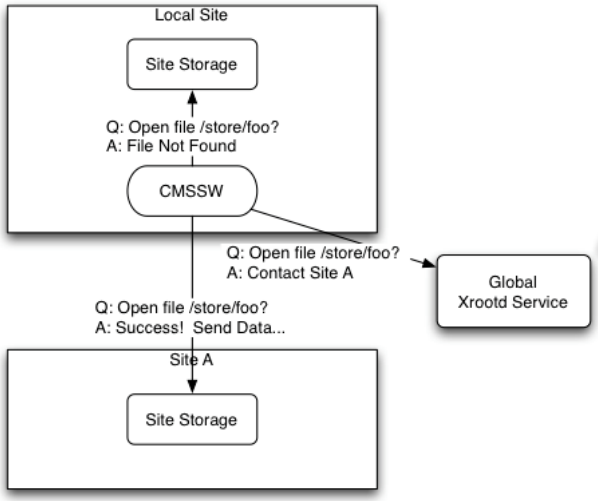
We even see peaks up to 6PB just from transfers out of CERN in 2016

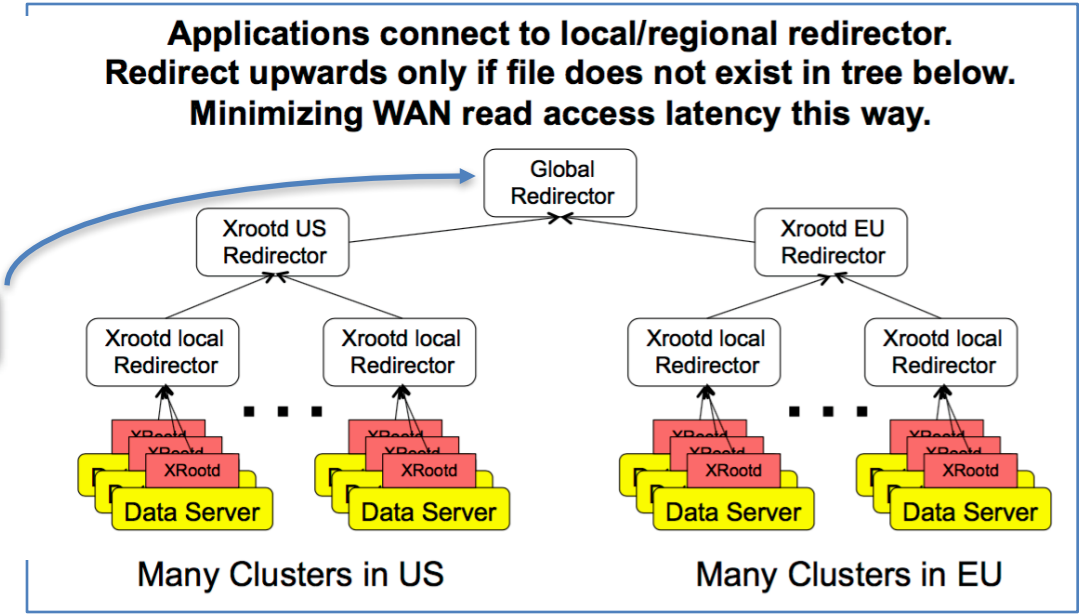


- Networks are very important to CMS GRID computing
  - Data taking (RAW data distribution)
  - Processing (Analysis data distribution)
  - Analysis (Remote access of data for users)

GRID CPU and storage no longer need to be located at the same place

- "AAA" data federation has the goal to provide CMS users access to "Any data, Any time, Anywhere"

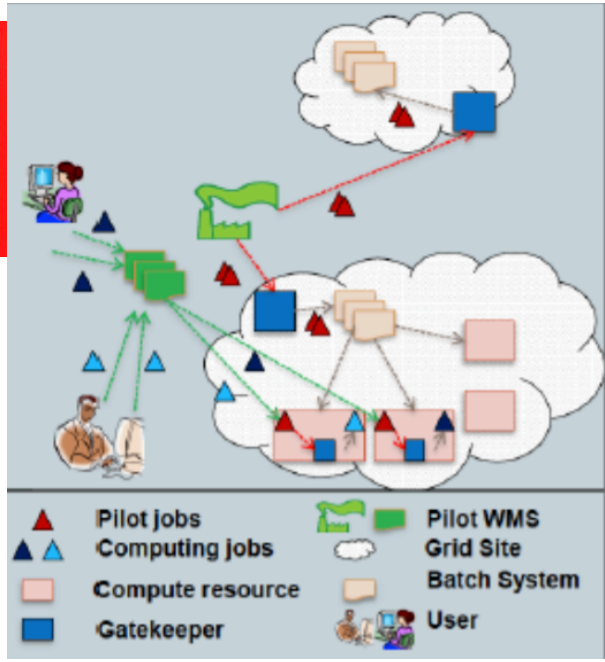


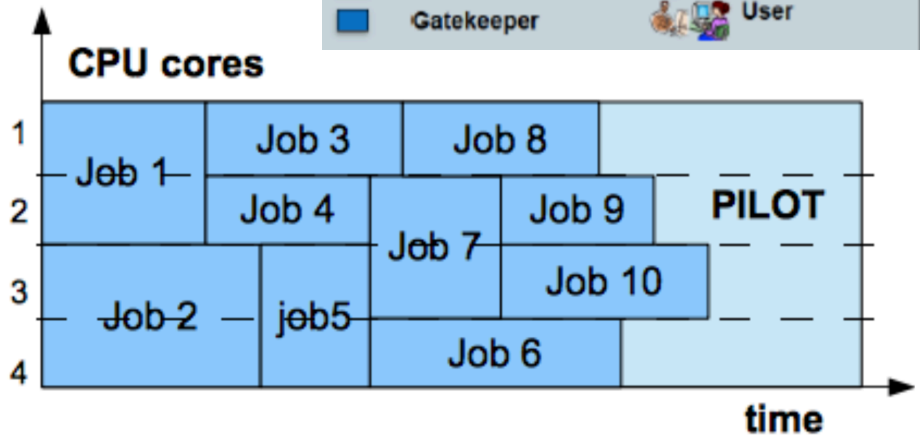


- What does this mean?
  - CMS applications can read data efficiently over wide-area networks
  - Job location no longer tied to data location: Relaxes constraints on locations of datasets and workflows

Additional GRID challenge: How to efficiently schedule dynamic and varied resources

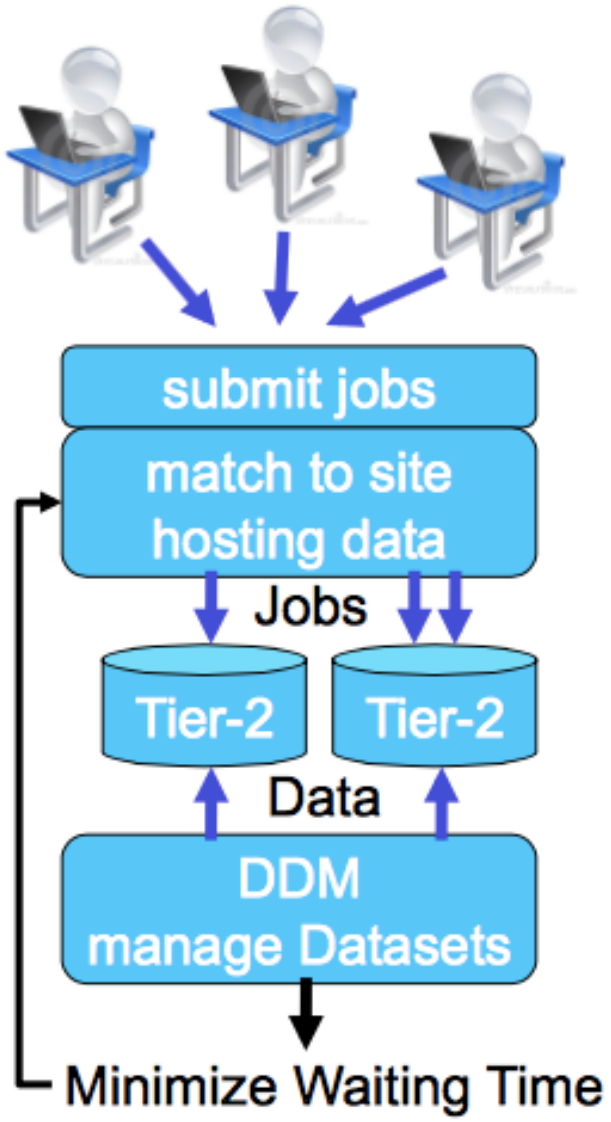
- "Pilots" are key concept for hiding the complexity of batch systems from users
  - Pilot jobs run on sites where they match to resources and pull user jobs from the CMS Global pool
  - Multicore pilots deployed across the GRID for CMS add extra degree of complexity for efficient job scheduling





Dynamic data management

- Traditionally: Humans make decisions about what data is available where
- Current implementation: Automated system uses available disk to replicate popular data (based on recent accesses)
- Next generation system:
  - Data replication based on data analytics metrics
  - Improved metrics to judge how many replicas are needed



Automation is key for efficient computing on the GRID

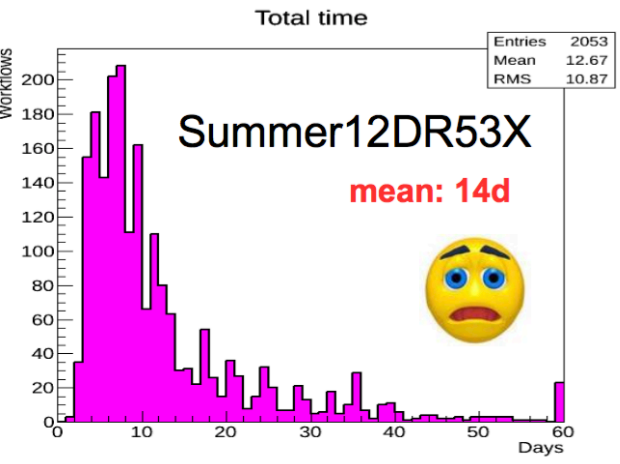
Challenges:

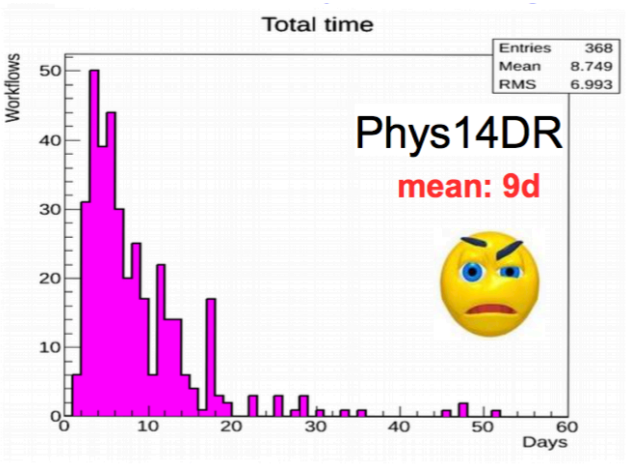
- Large number of user requests for data samples (each unique in some way)
- Many requests are urgent ("Can you finish it yesterday?")
- Demand, availability, accessibility variations across sites (each unique in some way)

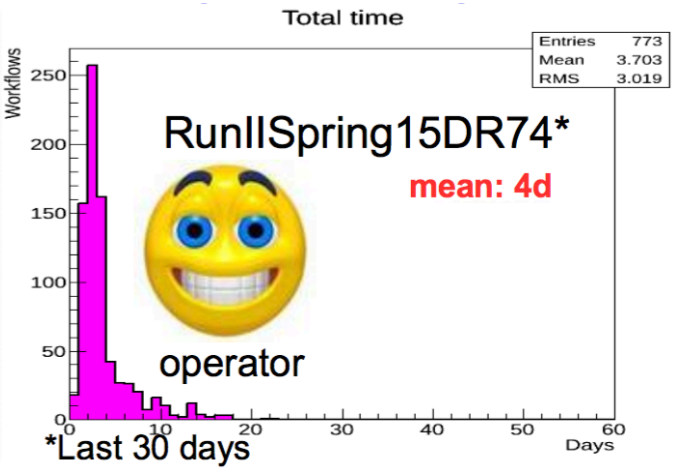
Solutions:

- Automate request handling based on request data availability, site status, etc..
- Simplify and automate job recovery procedures

Increased automation continues to benefit CMS users!



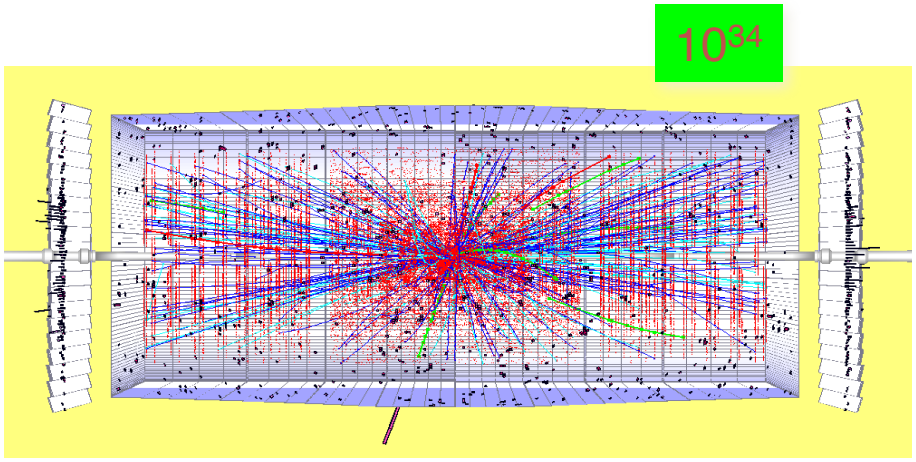


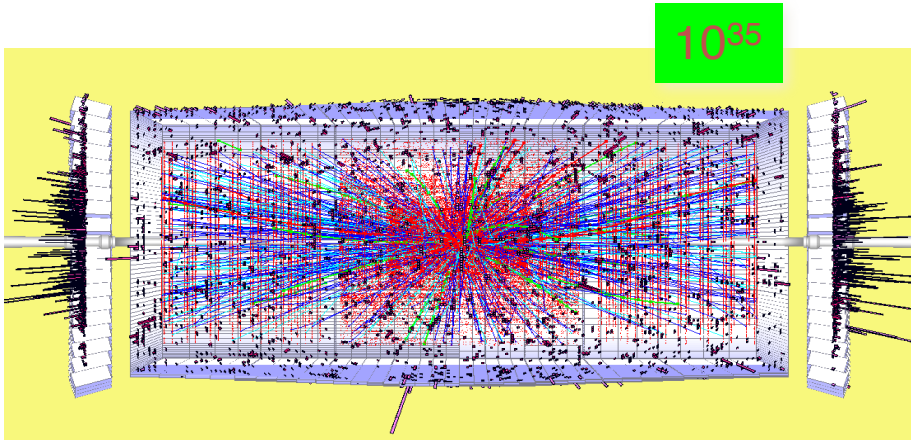


Our future: Major upgrades planned

Preparing for CMS at the start of Phase 2 (HL-LHC):

- The CMS detector configuration is still to be determined
- Even higher output rate of trigger (potentially 10kHz)
- Even higher luminosity and pileup (140+ interactions/crossing)

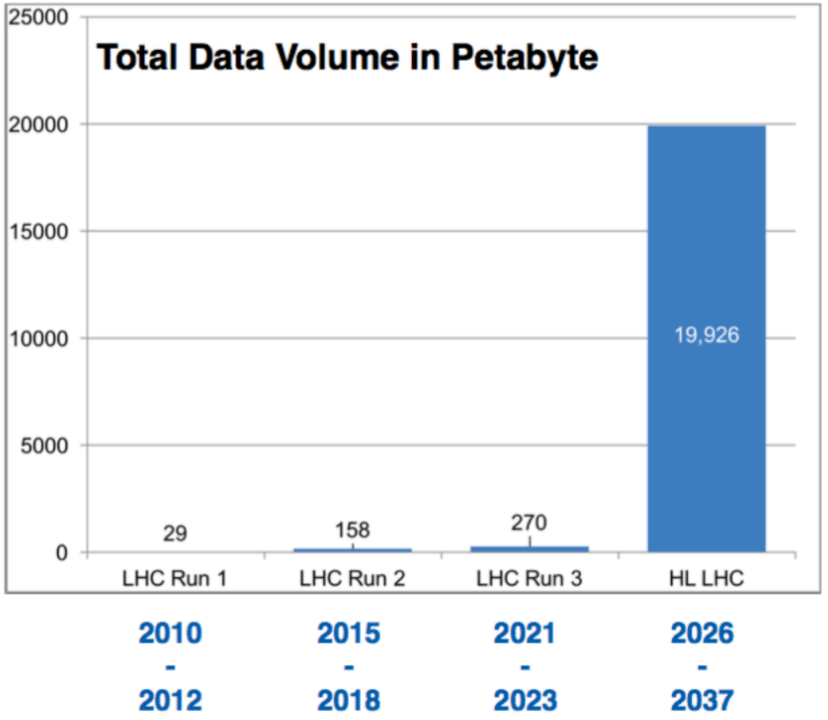




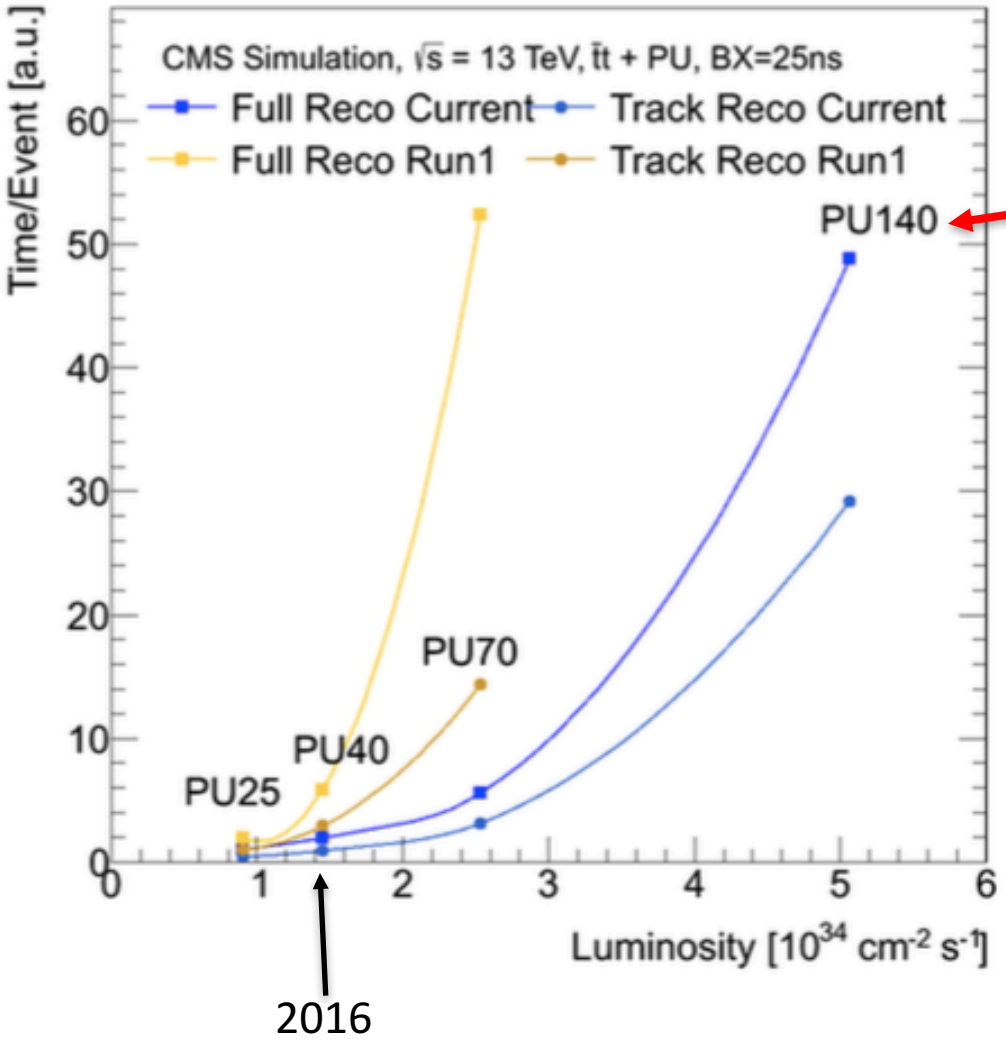
HL-LHC presents increased challenges for Triggering, Tracking and Calorimetry, in particular for low to medium P_T objects

The scale of computing increases dramatically

- Doing “more of the same” does not work: Either major budget increases or significant improvements in the way we operate CMS computing are need



Similar increases expected in CPU needs for HL-LHC



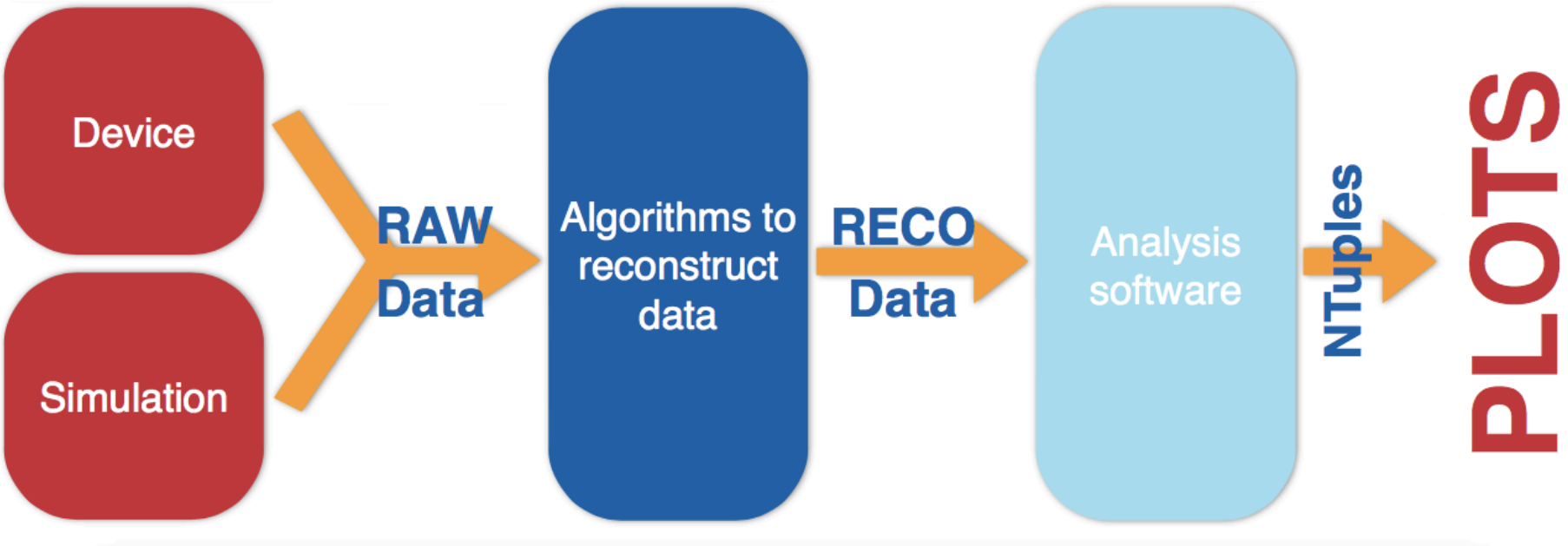
- More CPU needed to reconstruct as event complexity increases
- While we have made large improvements with time in the software performance, this will remain an important area of research for us

Potential ways R&D can close the gap? (Only example thoughts)

1. Improve software, calibrations, verification, validation techniques so that we process raw events only once
   - Close relationship between online and GRID computing?
2. New approaches to clustering, pattern recognition and filtering techniques
   - Can current machine learning techniques (and toolkits from industry) radically change the time required to identify physics objects from our raw data?
3. Smaller analysis data formats?
   - Most CMS searches have adopted a 10x smaller data format for Run 2. How do we expand this?
4. Faster detector simulation
   - Detailed simulation with dramatically faster tools (e.g., GeantV) or use parameterized simulation techniques?
5. Faster data analysis techniques

Computing community at CERN and beyond thinking about how to organize these R&D efforts to be ready for HL-LHC

Many analysis workflows are unique to HEP (partly to adapt to today's GRID computing)

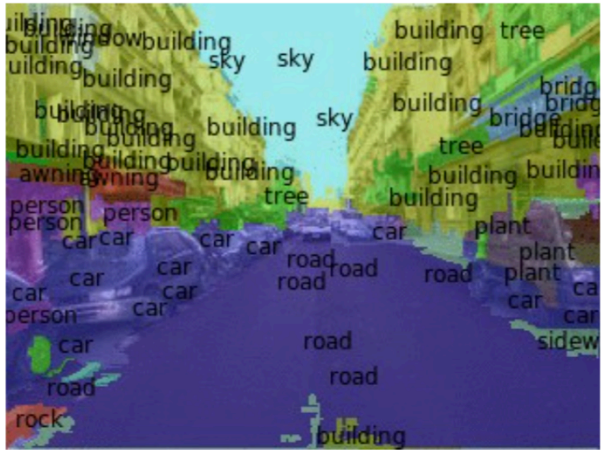


- It often takes weeks of processing (organized by users) for their final "ntuple" step.
- Big data techniques and computing infrastructure changes can bring a big improvement
  - Can data reduction centers dramatically improve the time it takes to complete computations for analysis?

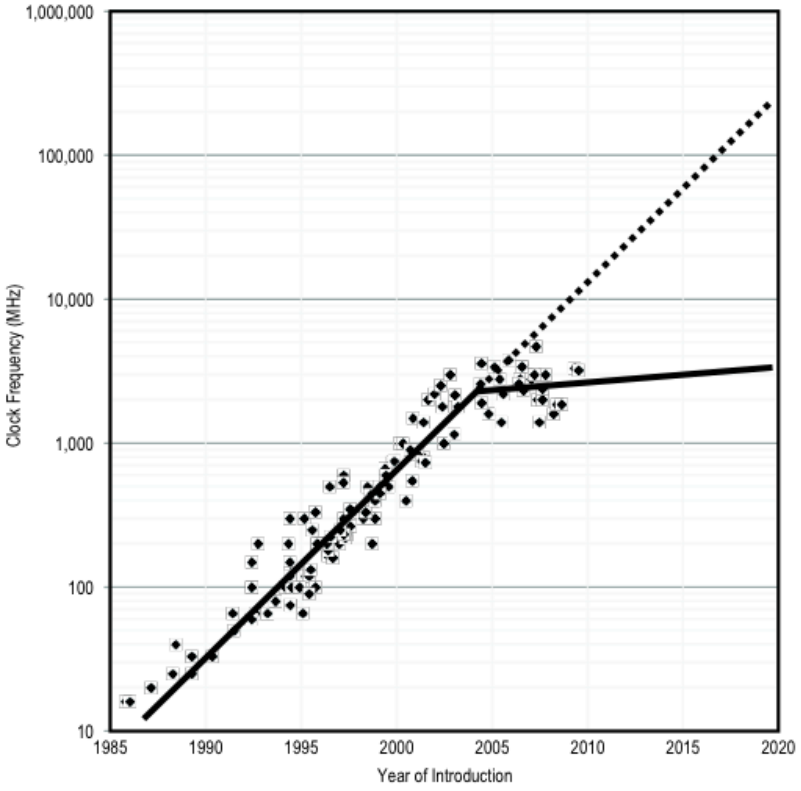
New approaches to analysis, clustering, pattern recognition and filtering techniques

- Pattern recognition is one of our big CPU consumers

- Can deep learning approaches identify "physics objects" in our data just as they do physical objects in scenes?



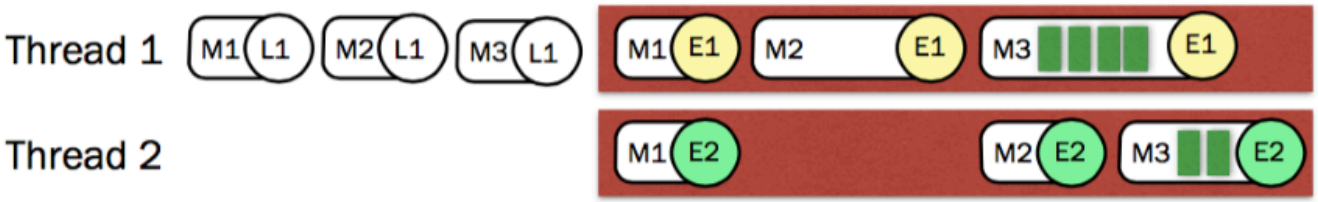
Computing architecture evolution: the introduction of multi/many core processors has pushed us away from embarrassingly parallel



- Trends we look to adapt to (or to take advantage of)
  - Multi-core / many-core systems
  - Low power / mobile
  - Importance of memory access
- The first step of our work was to develop a multi-threaded framework

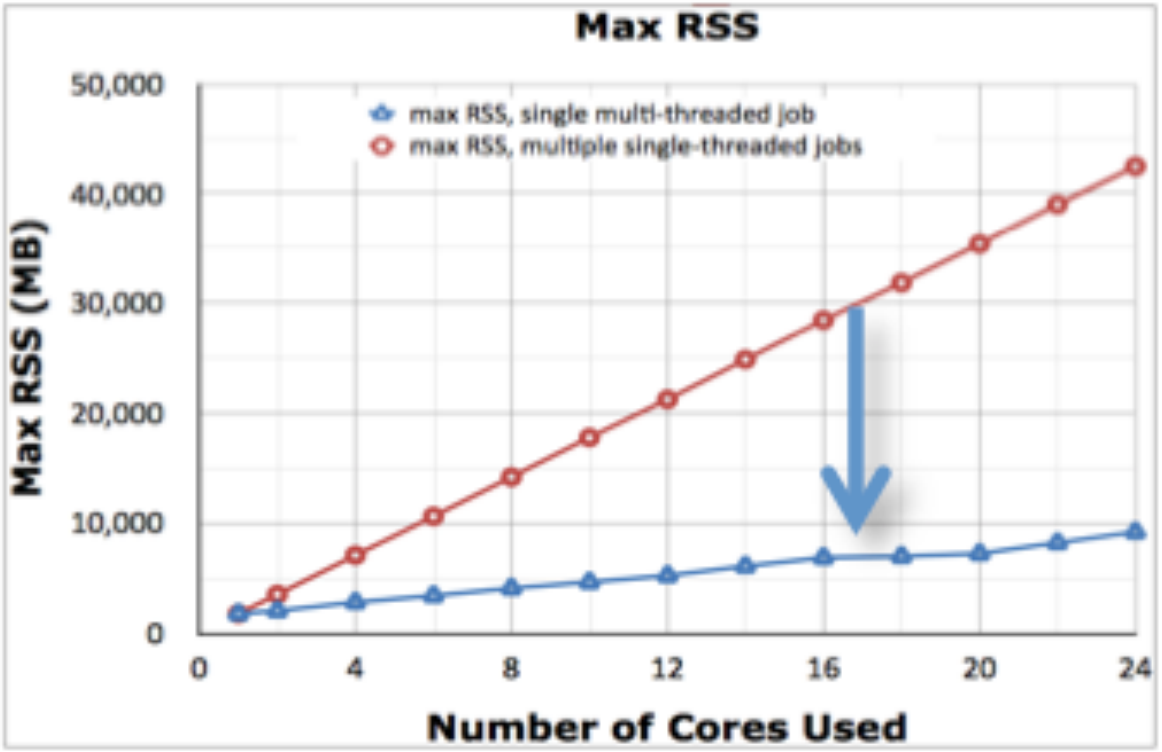
We have developed a multithreaded framework

- Advantage: save memory by sharing between threads
- current state: run each event in own thread



- future: run parts of events in different threads ➜ higher optimization results with even less memory usage

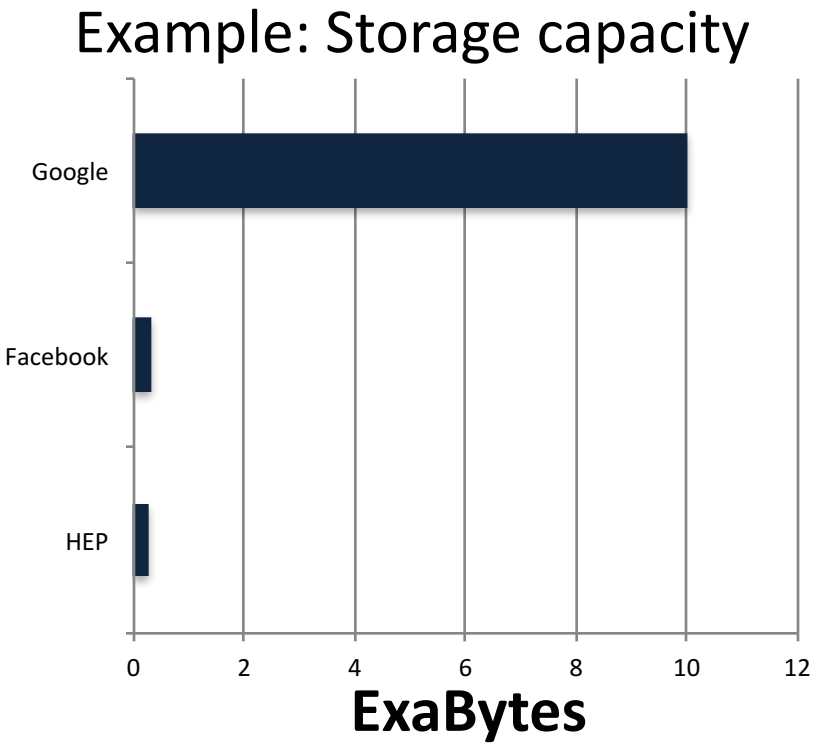
Use of this framework in production has been a big success in 2015



- Big challenge remains:
  - How do we upgrade all algorithms to be thread friendly
  - How do we teach developers to program in this new regime?

Science Clouds? HEP data manage is large, Cloud providers are dramatically bigger

- Idea: Commercial centers can provide semi-infinite CPU resources when you need them
  - You have (and pay for) resources (only) when you need them)

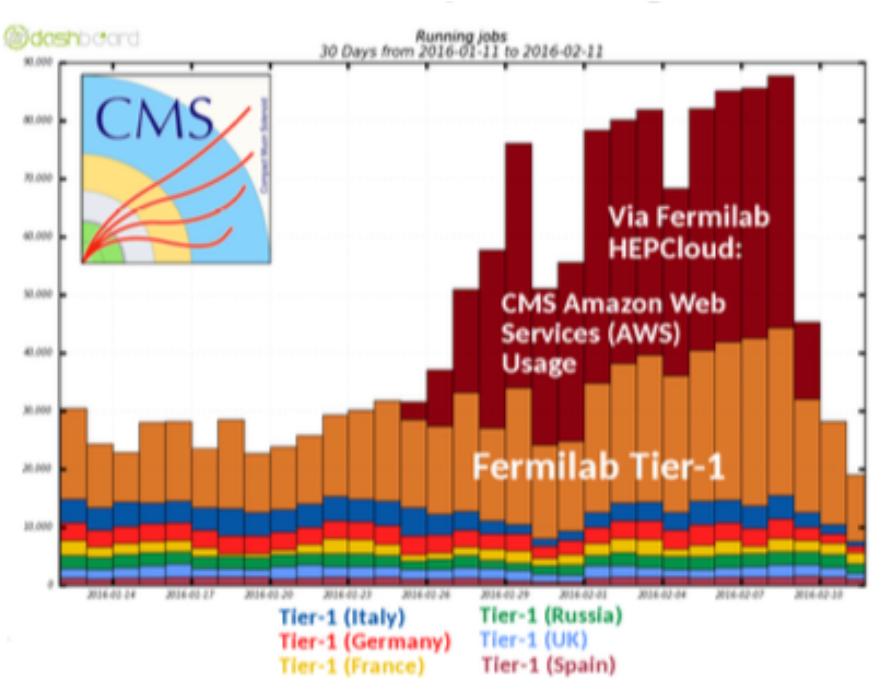


The scale of HEP data processing is big, but small compared to industry provides. We are looking at how to take advantage

We see a movement towards science clouds

- Recent work confirms costs and constraints on our workflows
  - Example: Spot market costs on AWS within reach of GRID facility
  - Push towards further infrastructure virtualization
  - I/O intensive workflows need to be monitored
  - GRID facilities continue to house data (for now!)

Example: FNAL facility augmented by Amazon web services CPUs



Clouds look to be the future of our computing resources.
We have a long ways to go to fully adopt this model

Outlook

- The success and continued evolution of GRID computing is essential for CMS

- Challenge for CMS computing for the next ten years: Deliver on analysis needs in CMS through run 2 and run 3 while preparing for HL-LHC
  - Be faster and more efficient without giving up physics quality (Example: D. Bonacorsi presentation today)
  - Keep up with technology evolution: The only way to be cost effective into the future (Example D. Abdurachmanov presentation today)
  - This means a software and computing ”upgrade”. Work is starting with HEP community to make that happen (P. Elmer presentation today)