



---

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

---

# HEP ASCR Meeting

**Highlights from Snowmass**

**Computing & Data Challenges – focus LHC**

**CMS Projections and Challenges**

**CMS Tier 1**

Lothar A. T. Bauerdick

U.S. CMS Software and Computing Manager

Open Science Grid Executive Director

October 24, 2014

# Computing Study at Snowmass2013, as input for P5

- Snowmass 2013 **Community Study**, “Computing Frontier”
  - ♦ Detailed look at computing and software needs to advance experimental and theoretical HEP
    - computational requirements
    - facility and software infrastructure
    - research investments needed
    - training requirements
- Detailed report and high-level summaries as part of the Snowmass Report and a subsequent report to DOE from the “Topical Panel Meeting on Computing and Simulations in High Energy Physics”
- Used as input for the P5 recommendation on computing
  - ♦ “Strengthen the global cooperation...”



**SNOWMASS** CSS 2013  
**ON THE MISSISSIPPI**  
**JULY 29 – AUGUST 6, 2013**

ORGANIZED BY THE DIVISION OF PARTICLES AND FIELDS OF THE APS  
HOSTED BY THE UNIVERSITY OF MINNESOTA

**STUDY GROUPS**  
Energy Frontier  
Chip Brock (Michigan State),  
Michael Peskin (SLAC)  
Intensity Frontier  
JoAnne Hewett (SLAC),  
Harry Veerts (Argonne)  
Cosmic Frontier  
Jonathan Feng (University of California, Irvine),  
Steve Ritz (University of California, Santa Cruz)  
Frontier Capabilities  
William Barletta (MIT),  
Murdoch Gilchriese (LBNL)  
Instrumentation Frontier  
Marcel Demarteau (Argonne),  
Howard Nicholson (Mt. Holyoke),  
Ron Lipton (Fermilab)  
Computing Frontier  
Lothar Bauerdick (Fermilab),  
Steven Gottlieb (Indiana)  
Education and Outreach  
Marge Bardeen (Fermilab),  
Dan Cronin-Hennessy (Minnesota)  
Theory Panel  
Michael Dine (University of California, Santa Cruz)

**LOCAL ORGANIZING COMMITTEE**  
Marcela Carena (Fermilab and University of Chicago)  
Dan Cronin-Hennessy (Minnesota, Chair)  
Prisca Cushman (Minnesota)  
Lisa Everett (Wisconsin)  
Alec Habis (Minnesota, Duluth)  
Ken Heller (Minnesota)  
Jody Kaplan (Minnesota)  
Yuichi Kubota (Minnesota)  
Jeremy Mans (Minnesota)  
Bridget McCoy (Minnesota)  
Marvin Marshak (Minnesota)  
Jarek Nowak (Minnesota)  
Keith Olive (Minnesota)  
Gregory Pawloski (Minnesota)  
Ron Pelling (Minnesota)  
Marco Peloso (Minnesota)  
Yongzhong Qian (Minnesota)  
Roger Rusack (Minnesota)  
Wesley Smith (Wisconsin)

**DPF EXECUTIVE COMMITTEE**  
Chair: Jonathan Rosner (University of Chicago)  
Chair-Elect: Ian Shipsey (Purdue University)  
Vice Chair: Nicholas Hadley (University of Maryland, College Park)  
Past Chair: Pierre Ramond (University of Florida, Gainesville)  
Secretary/Treasurer: Howard Haber (University of California, Santa Cruz)  
Councillor: Marjorie Cooran (Rice University)  
Members at Large:  
• Jonathan Feng (University of California, Irvine)  
• Lynne Orr (University of Rochester)  
• Yuri Gershtein (Rutgers University)  
• Nikos Varelas (University of Illinois, Chicago)  
• Robert Bernstein (Fermilab)  
• Sally Seidel (University of New Mexico)

APS UNIVERSITY OF MINNESOTA NSF

WWW.SNOWMASS2013.ORG

POSTER DESIGN BY KATIE SCHALOW

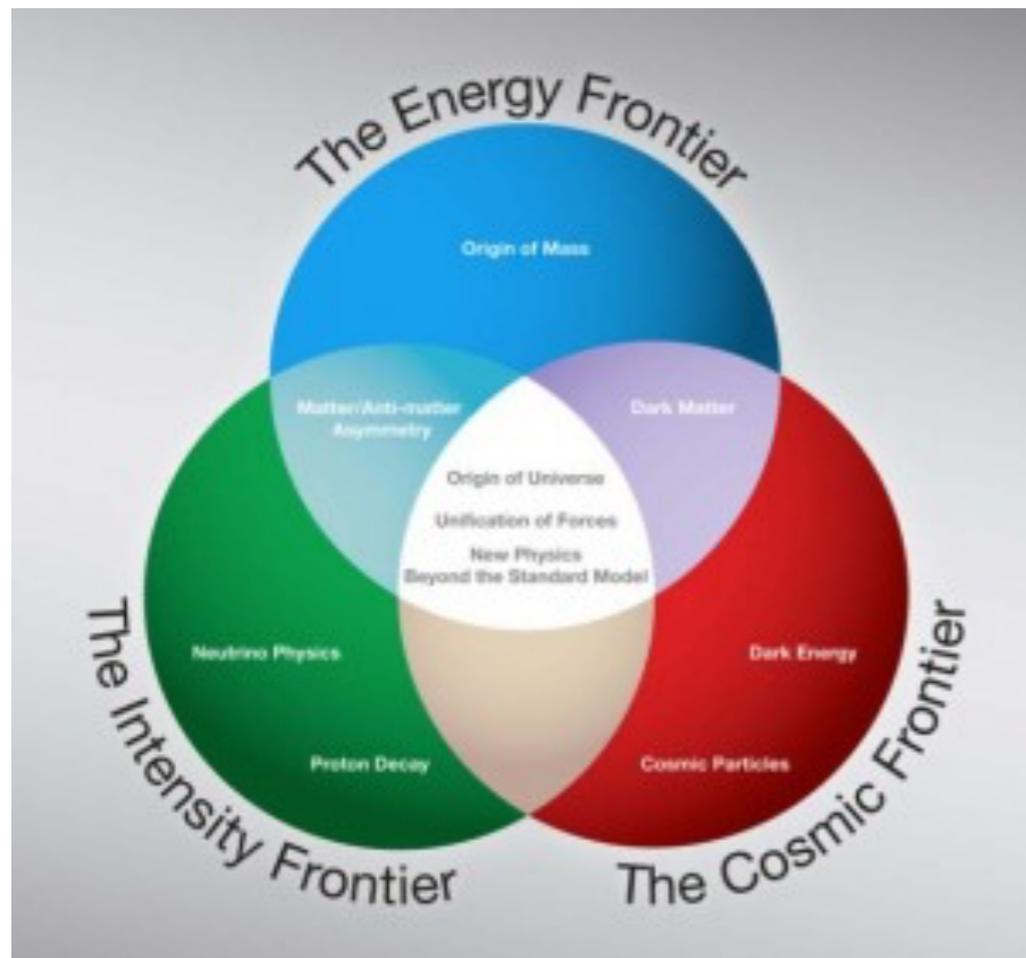


## Enabling R&D (Computing)

- The recent *Report from the Topical Panel Meeting on Computing and Simulations in High Energy Physics* articulated the challenges involved in meeting the increasing computational needs of the field and suggested steps to take full advantage of cost-effective computing solutions. The present practice is to handle much of the computing within individual projects. Rapidly evolving computer architectures and increasing data volumes require effective crosscutting solutions that are being developed in other science disciplines and in industry. Mechanisms are needed for the continued maintenance and development of major software frameworks and tools for particle physics and long-term data and software preservation, as well as investments to exploit next-generation hardware and computing models. Close collaboration of national laboratories and universities across the research areas will be needed to take advantage of industrial developments and to avoid duplication.

- **Recommendation 29: Strengthen the global cooperation among laboratories and universities to address computing and scientific software needs, and provide efficient training in next-generation hardware and data-science software relevant to particle physics. Investigate models for the development and maintenance of major software within and across research areas, including long-term data and software preservation.**

# Cooperation and Collaboration in Computing



- Very concrete examples for cooperations in all directions
  - ♦ like the ESnet Extension to Europe EEX supporting LHC TA network
  - ♦ the HEP-FCE Forum for Computational Excellence
  - ♦ the HEP Software Collaboration with Europe, many more

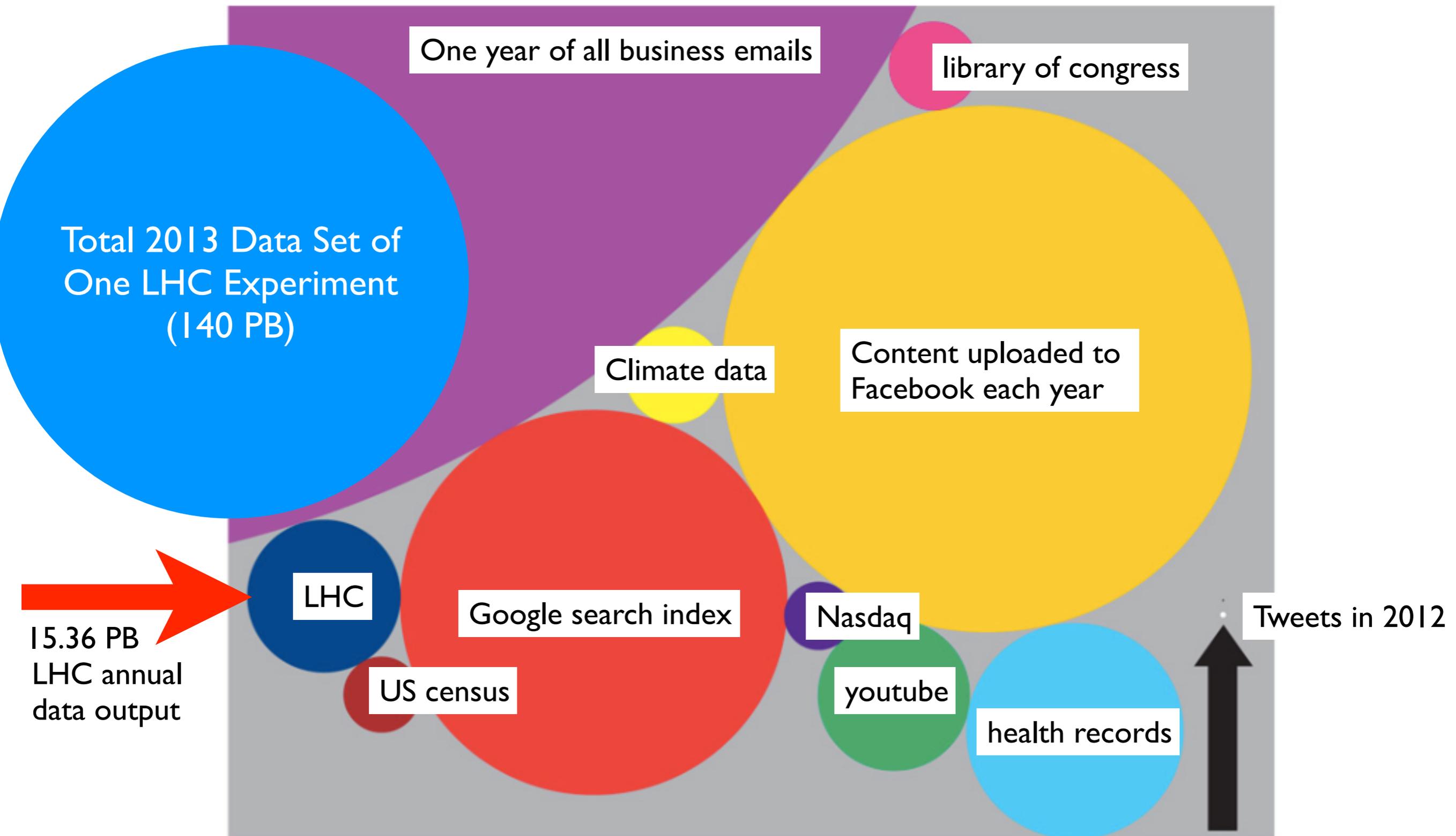
# Findings from the Computing Study: For the Energy Frontier and the Large Hadron Collider:

---

- Computing limitations already reduce the amount of physics data that can be analyzed
- The planned upgrades (HL-LHC) are expected to result in a
  - ✦ ten-fold increase in the number of events
  - ✦ and a ten-fold increase in event complexity.
  - ✦ LHC produces about 15 petabytes (PB) of raw data per year now, but in 2021 the rate may rise to 130 PB
- Efforts to increase code efficiency, parallelism, data processing
  - ✦ explore the potential of computational accelerators
  - ✦ advance from sequential to “big-data” type data analysis
- More than half of the computing cost is now for storage
  - ✦ in future it may be cost-effective to recalculate, rather than store
- Attention on data management and wide-area networking
  - ✦ assure network connectivity for distributed event analysis

# The Big Data Frontier (2013 Snowmass)

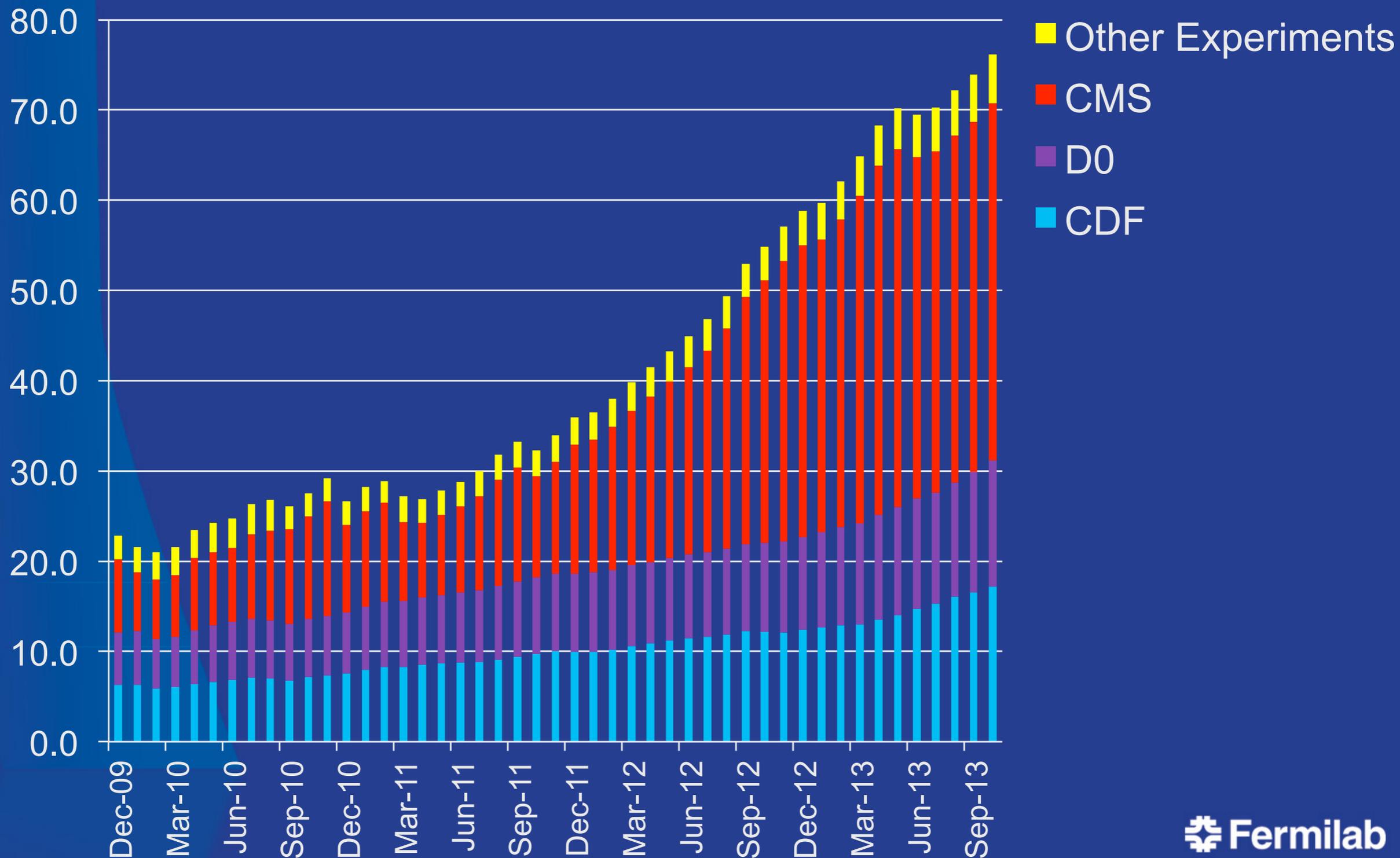
from Wired Magazine



<http://www.wired.com/magazine/2013/04/bigdata/>

# Long history of provisioning and operating storage for all Fermilab Users: approaching 100 Petabytes of data...

## Petabytes of Data on Tape



# Detailed modeling of expected compute and data needs

- Very large increase in online-storage (disk systems)
  - ✦ at Tier-1 (lab) and Tier-2 centers
- This is a main cost driver for LHC computing upgrades

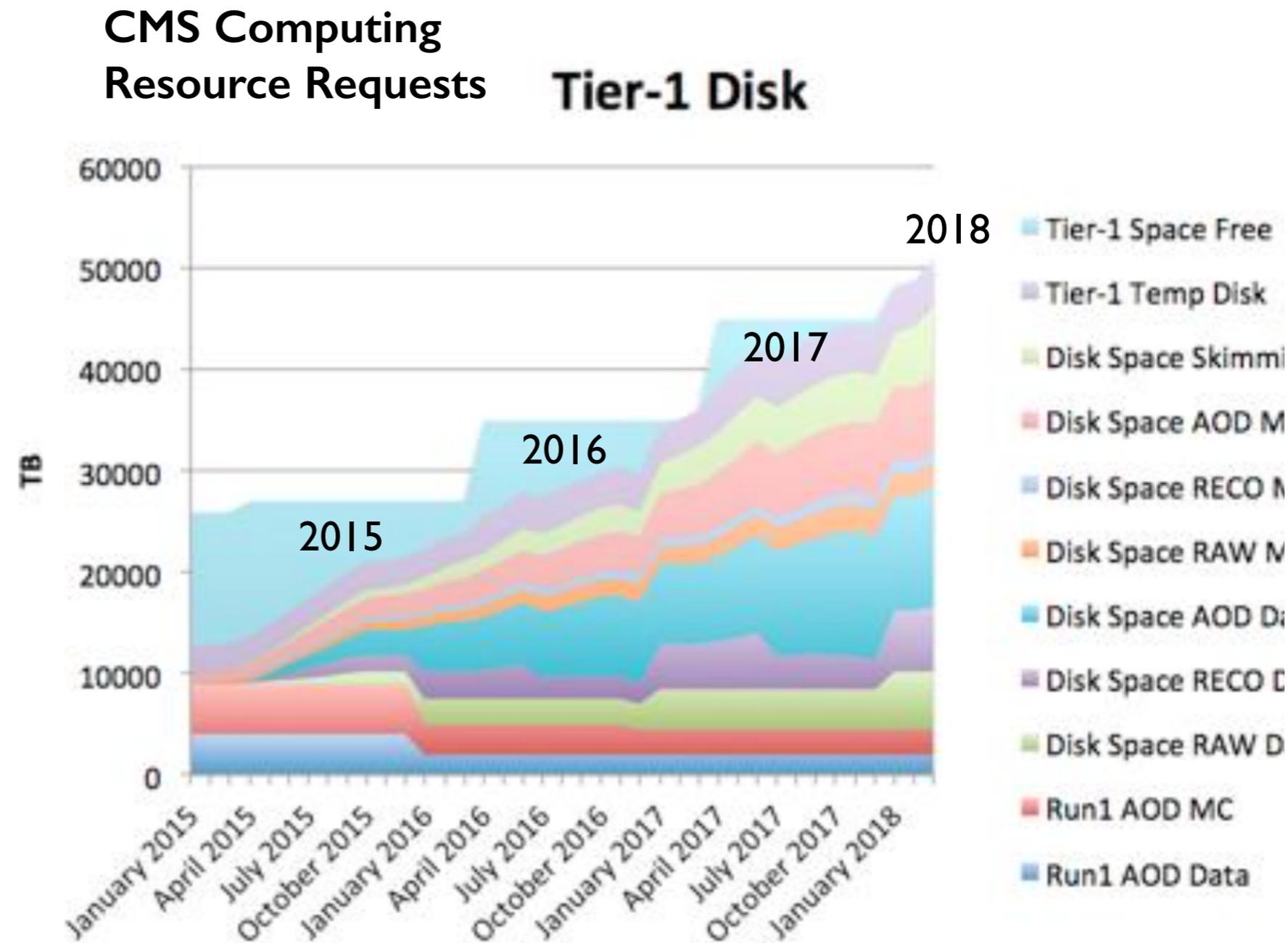
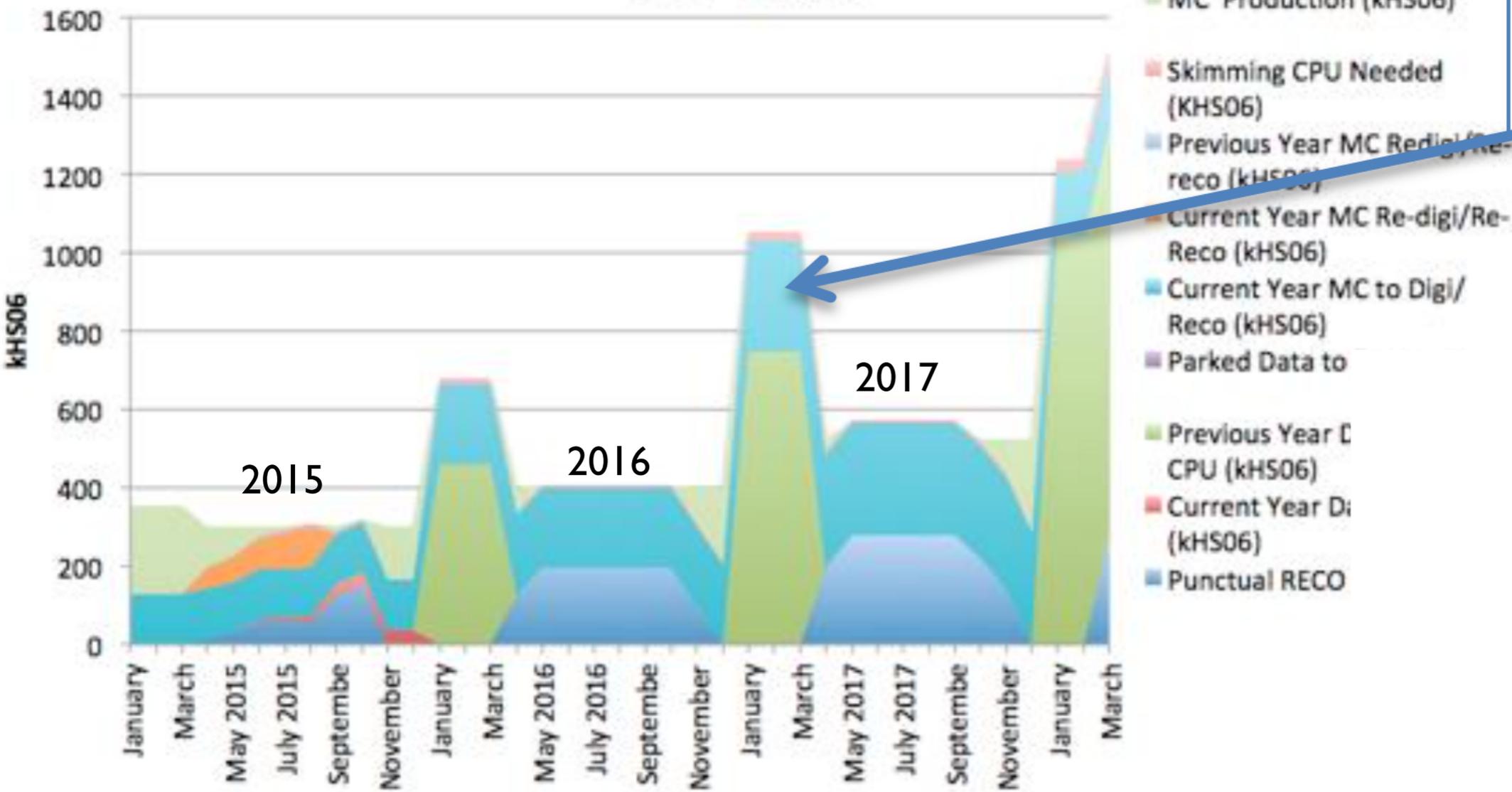


Figure 2: Tier-1 disk usage by data type.

# Compute Needs: Huge “spikes” above a large “baseline”

- LHC experiments are looking for ways of fulfilling **peak demands** on the time scale of 2016-2018

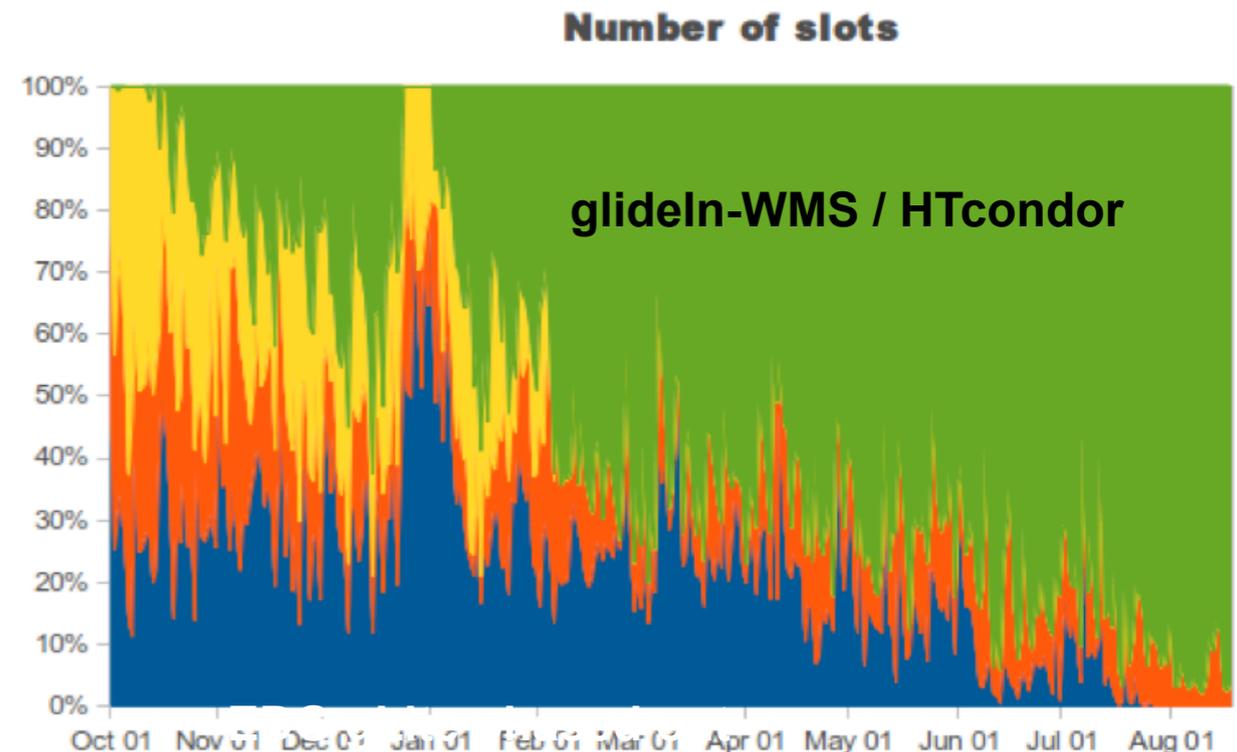
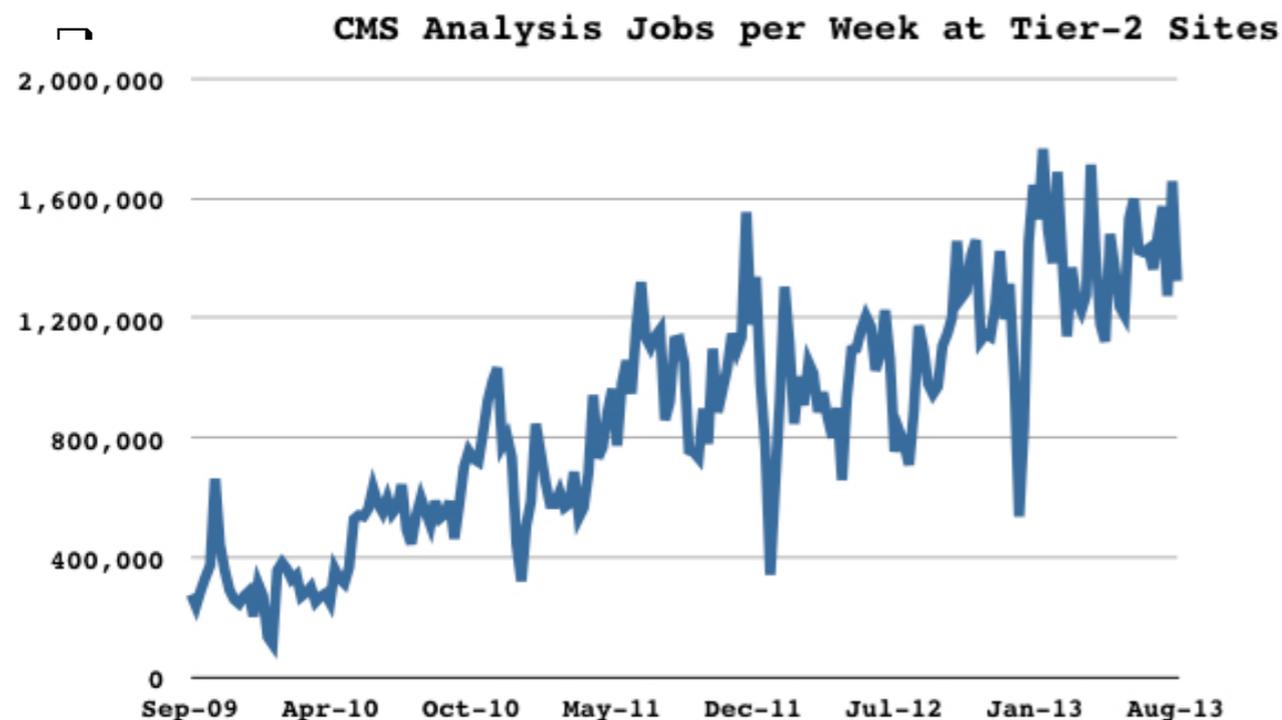
**CMS Computing Resource Requests Tier-1 CPU**



Requires on-demand additions to LHC computing i/s!

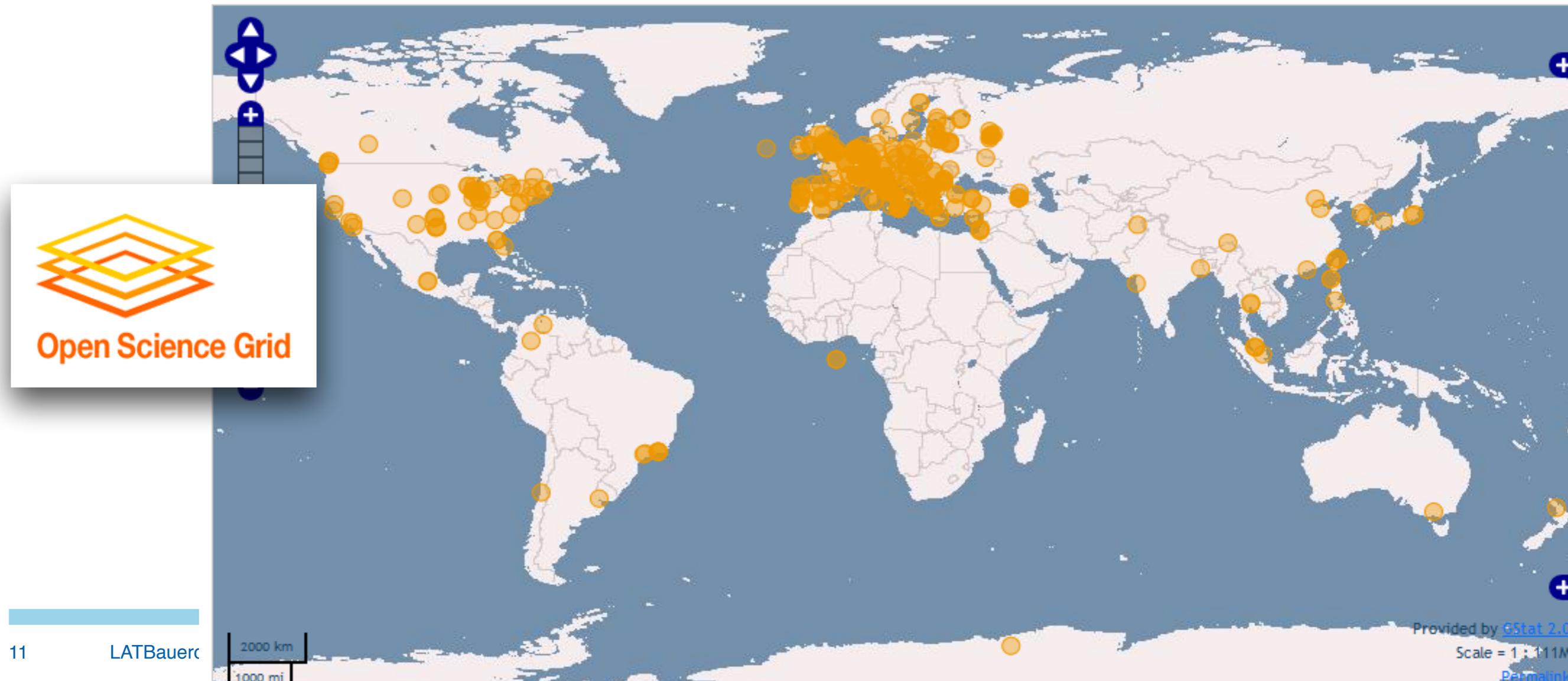
# Data Analysis: Distributed High-Throughput Computing

- Hundreds of individual researcher cause huge increase in data analysis
  - ♦ University computing centers (Tier-2s) run  $\gg$  1M jobs/week
- CMS moved to HTcondor-based “overlay job manager” (green, below)
  - ♦ homogeneous, stable execution environment, phase out old Grid software
- allows use of cloud resources and other dynamically provisioned resources
  - ♦ CMS HLT farm and CERN general purpose “Agile” infrastructure
    - Use of HLT farm increases computing capacity by a  $\sim$ factor of 2!
  - ♦ provisioned as IaaS through amazon EC2-compatible OpenStack interface
  - ♦ through HTCondor, support for security infrastructure



# Distributed Computing and Sharing of Resources

- LHC was the first in HEP to adopt a **distributed computing model**
  - ♦ Before LHC most of the computing and storage capacity was at a central experiment data center
  - ♦ now most HEP experiments have adopted distributed computing
  - ♦ Open Science Grid helps with opportunistic sharing of resources
    - > 100M CPU hours on the OSG were “opportunistic”, ~10%, increasing



# Energy Frontier Computational Strategies

---

- High Throughput Computing
  - ✦ process as many events as possible in the shortest period of time
  - ✦ tasks (data analysis, reconstruction, simulation) split into parallel jobs
    - no/loose inter-process communications, it's an “embarrassingly parallel” problem
  - ✦ distributed computing resources, presented to users/work flow engines through “overlays” as a coherent job execution environment
    - pilot-job based Overlay Job Managers through glideinWMS, Panda, HTcondor
- Work flow aspects are very important for the quality of the result
  - ✦ e.g. significant data management, book keeping, provenance tracking
- Very significant data component
  - ✦ managing and extracting science from tens of PetaByte active data
    - observational and simulations data ~ same size
    - expect to grow to ~.3 ExaByte of active data on OSG during this decade
  - ✦ across a set of data centers for local and remote access to data
- The biggest computational challenges are **volume** and **complexity** of the data to be processed

# Computational Challenges for the LHC

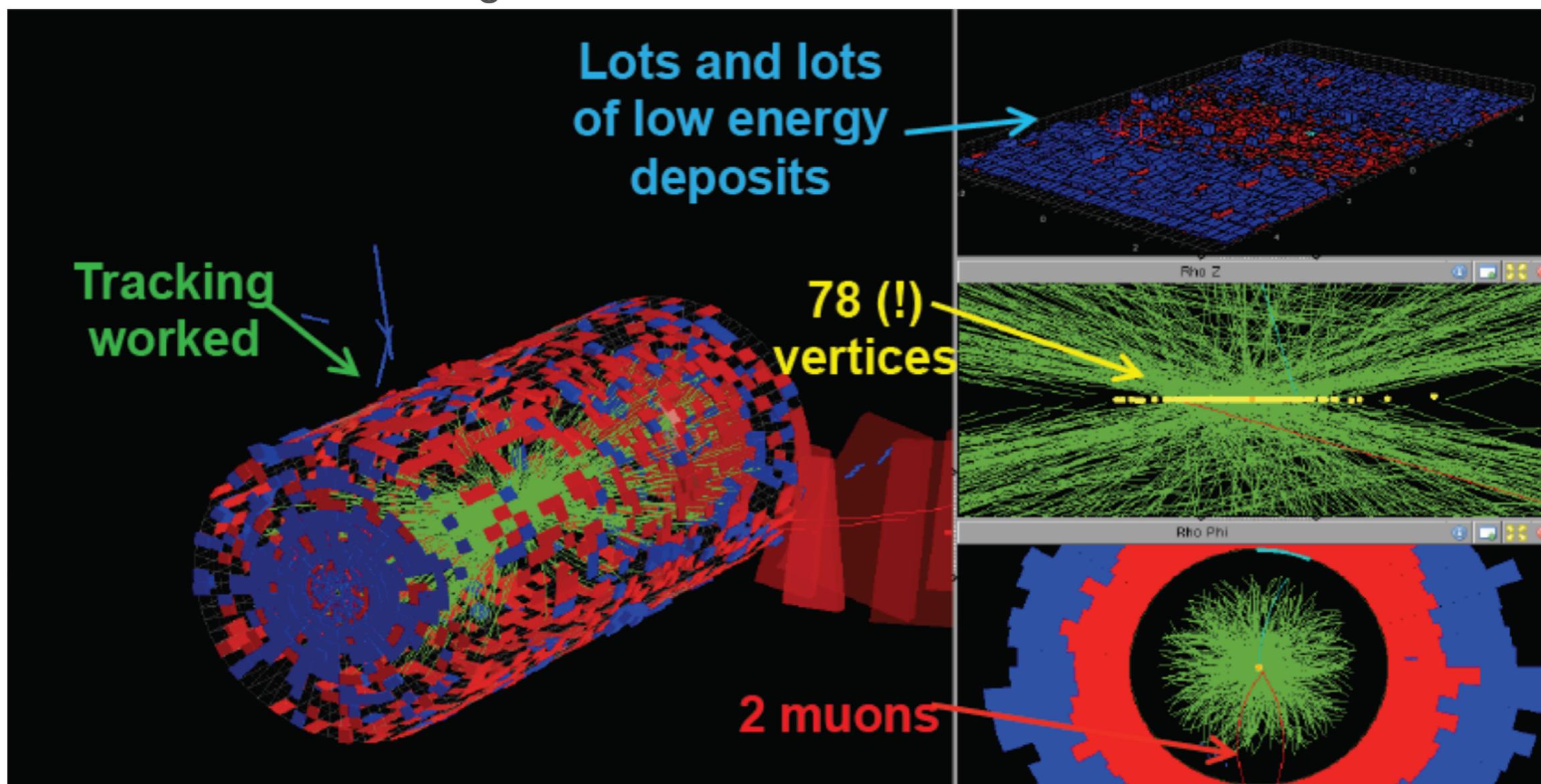
---

- Challenging resource needs require efficient and **flexible use of all resources**
  - ♦ We're proactively looking into ways of tapping into new kinds of resources
  - ♦ both **Distributed High-Throughput Computing** (Grids, Clouds)
  - ♦ and **High-Performance Computing** (some successes with NERSC, LCF, XSEDE)
  - ♦ **Sharing** and **opportunistic use** help address resource needs, from all tiers of computing, and now including community or commercial clouds etc
- To stay on the Moore's law curve, need to proactively make full/better use of **advanced architectures**: multi-threading, GPU environments, low-energy CPUs
  - ♦ With the need for more parallelization the **complexity of software and systems** continues to **increase**: frameworks, workload management, physics code
  - ♦ Important needs for **developing and maintaining expertise** across offline, computing, POGs etc, including re-engineering of frameworks, libraries and physics codes, adapting key software tools
- Unless corrective action is taken we could be **frozen out of cost effective computing solutions** on a time scale of 10 years.
  - ♦ There is a large code base to re-engineer
  - ♦ We currently do not have enough people trained to do it

# Increasing Complexity of Events at EF

## Exponentially Increasing Resource Needs

- each recorded interaction consists of a hypothesis-dependent complex hierarchy of data structures
  - ♦ 2d-hits vs 3d hits vs track elements vs particle hypothesis etc
  - ♦ raw and reconstructed signals etc



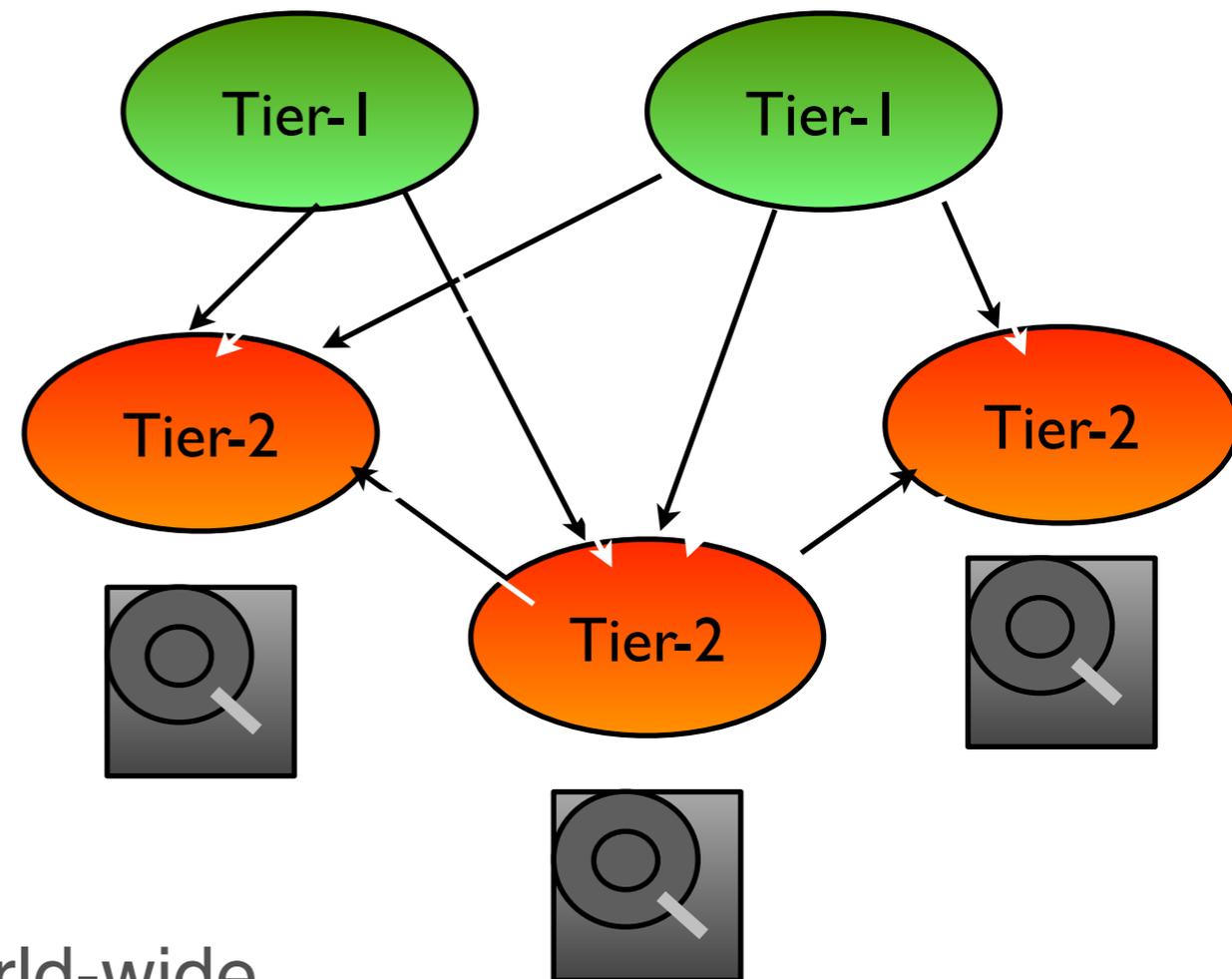
# LHC Tier-1/2 centers make huge data sets accessible for analysis

- Tier-2s vary from 10s of TB at the smallest site to several PB of disk at the larger sites
  - ♦ centers employing diverse set of disk-based storage solutions
  - ♦ In 2014 there are  $\gg$  100PB of disk storage in use at LHC Tier-2 centers
- The ability to procure and operate large disk-based storage resources at a large number of computing centers has been critical to the success
  - ♦ LHC Tier-2s are very heavily utilized
  - ♦ challenging IO applications
- DataManagement Complexity



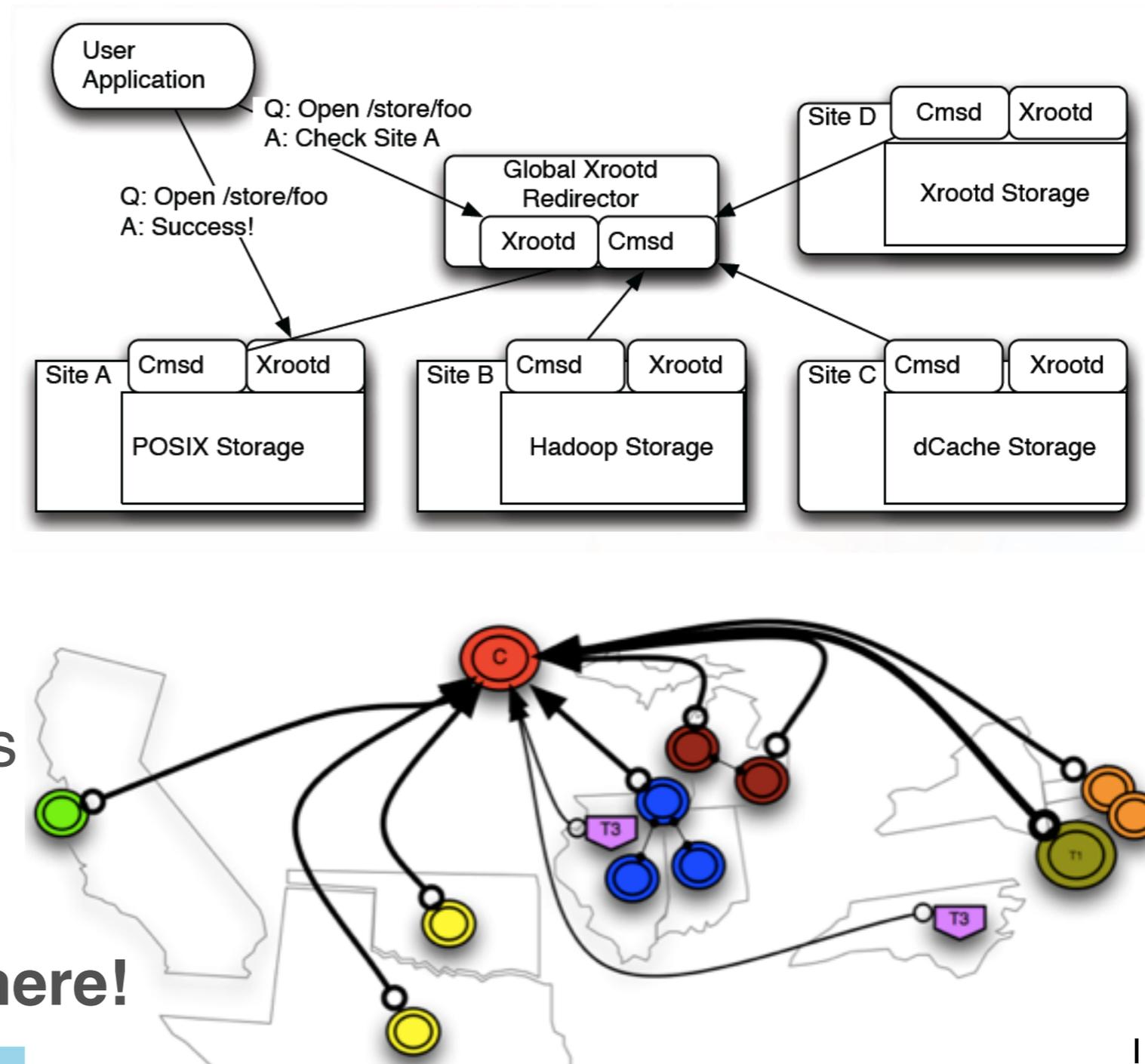
# Federating the Data Centers

- LHC experiments **federate** their distributed storage systems
  - ✦ Create a common global namespace across all storage sites, accessible from anywhere, with local file catalogs and “redirection”
  - ✦ Make easy to use, homogeneous access to data, based on **xrootd** system
  - ✦ Empower sites to add/modify/operate their systems locally, federating it to the overall experiment’s globally distributed system
- Many uses and advantages
  - ✦ Failover for load balance or in case of problems with local storage systems
  - ✦ Gain access to more CPUs using WAN direct read access
    - brokering to Tier 2s with partial datasets
    - opportunistic resources w/o local storage
  - ✦ Use as caching mechanism at sites to reduce local data management tasks
    - Eliminate cataloging, consistency checking, deletion services
- System is now being commissioned world-wide



# Seamless Data Access through the Wide Area Network

- Optimized IO allows over-the-WAN direct access (“streaming”)
  - ✦ Sending data directly to applications over the WAN (using xrootd)
- Small impact on WAN traffic compared to total HEP use
  - ✦ transfers vs direct access
  - ✦ caching feasible
  - ✦ traffic being closely monitored
- Huge impact on managing the data and the storage: local data management, bottom-up, empowering sites
- Huge impact on Science:
  - ✦ **Any data, anytime, anywhere!**



# The Future?

- Data federation is a good first step for data ubiquity etc
- in future we need a robust system following industry trends:  
data intensive content delivery networks for scientific data
  - ♦ delivering to super computers, clouds, clusters, people etc



# Content Delivery Networks: Compare to Netflix (I.Fisk)

- HEP problem harder than Netflix?
  - ✦ Netflix delivers streaming video content to > 20M subscribers
  - ✦ Routinely quoted as the single largest user of bandwidth in the US
    - More than 30% of the traffic
- HEP has a different working point:
  - < # clients,
  - < distribution,
  - > bandwidth per client
- However, **much** larger data set
  - ✦ HEP can't make many multiple static copies
  - ✦ need different strategies instead:
    - make **dynamic** replicas and clean up when no longer useful
    - access data directly over the wide area networks

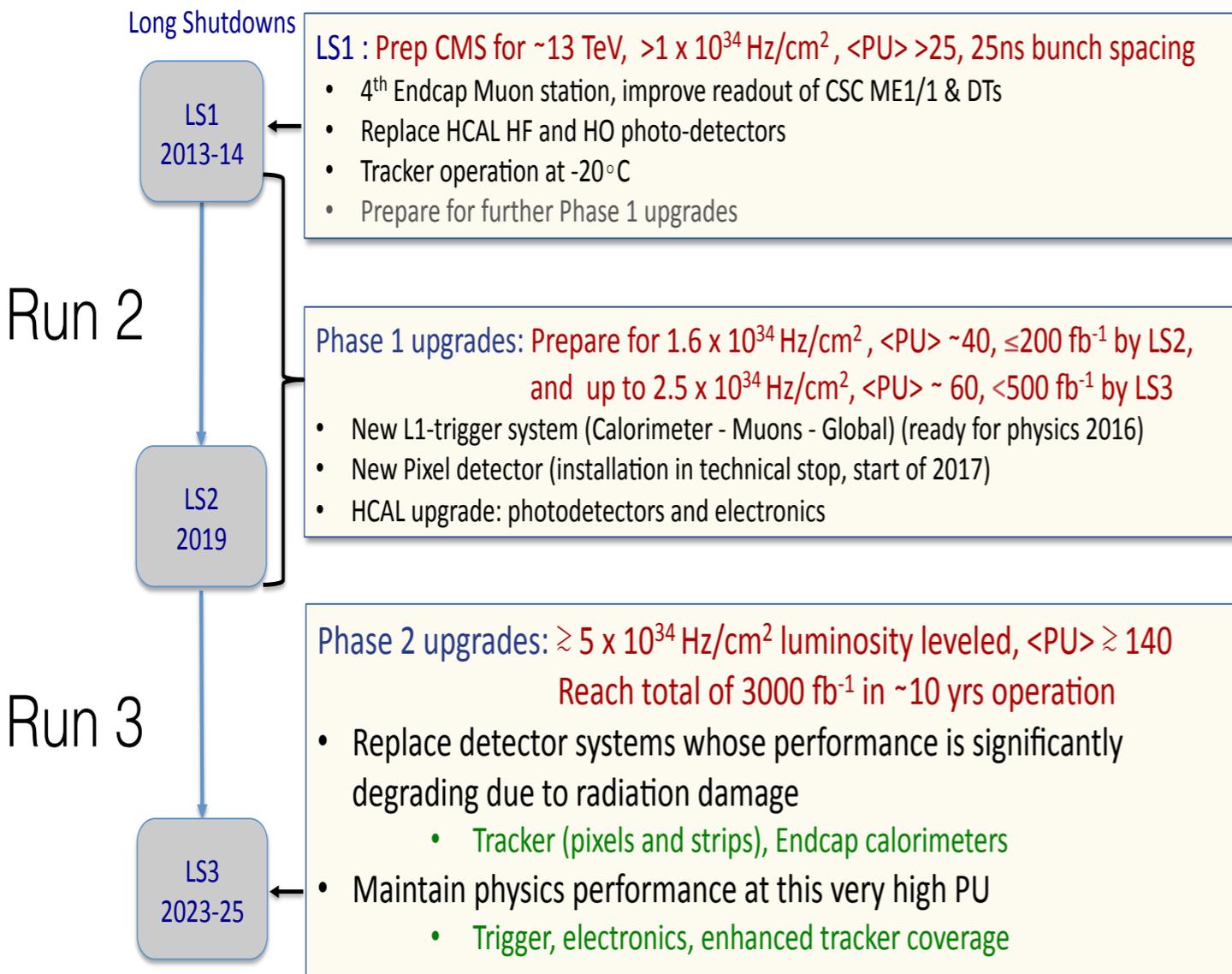
	NETFLIX	LHC Computing
Bandwidth per client	1.5Mbit/sec	1MByte/sec
Clients	1M*	100k cores
Serving	1.5Tbits	0.8Tbits
Total Data Distributed	12TB	20PB
Annual Budget	>\$4B	< \$.04B

Similar Problems:  
Not all files  
are equally  
accessed



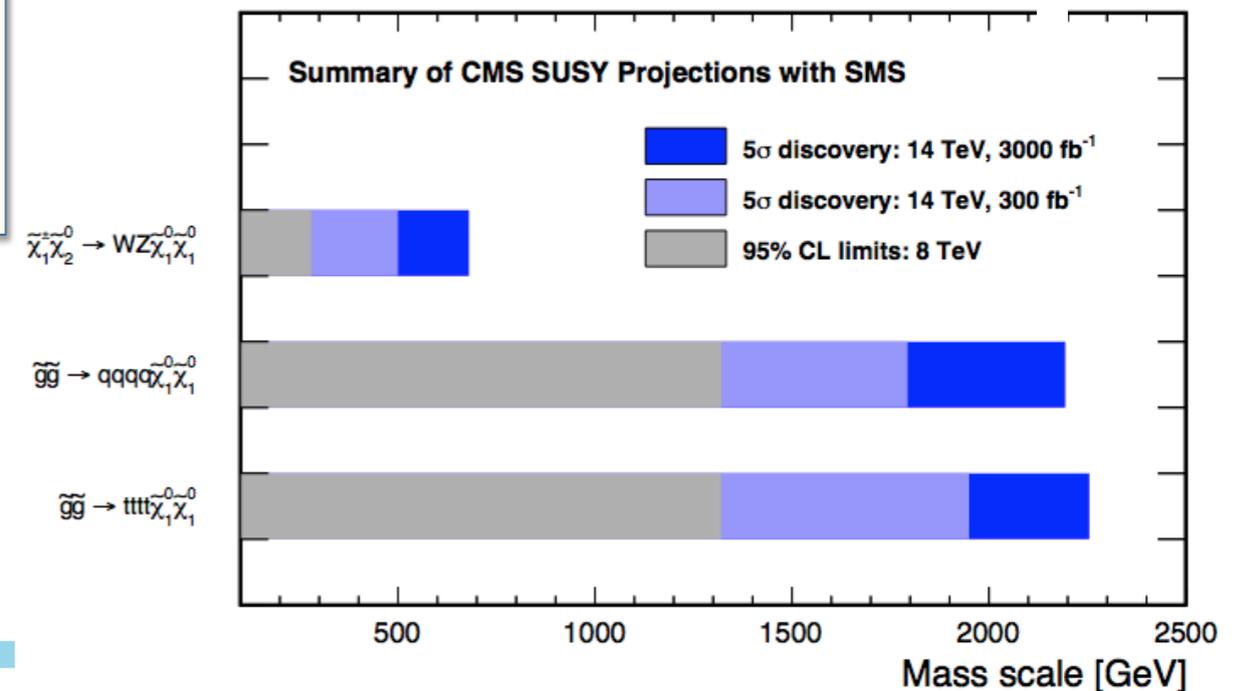
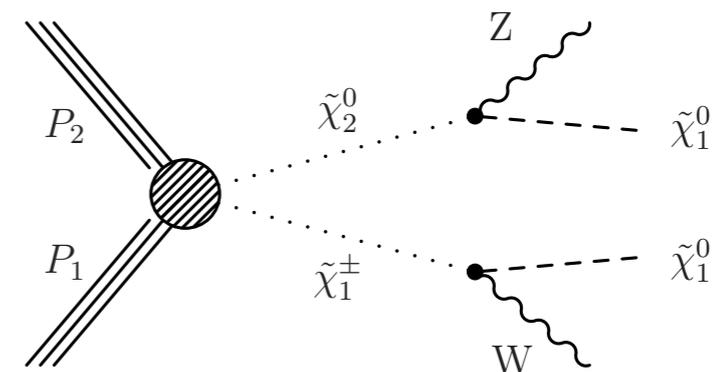
e.g. Forward  
Physics ;-)

# LHC Upgrades and High-Luminosity LHC



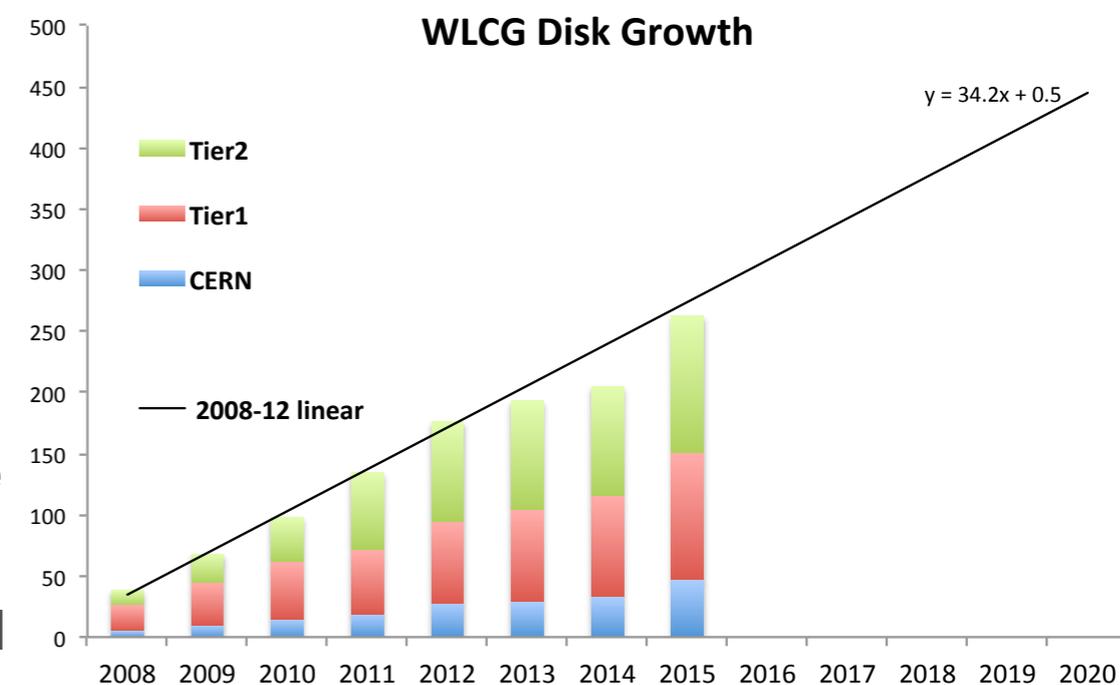
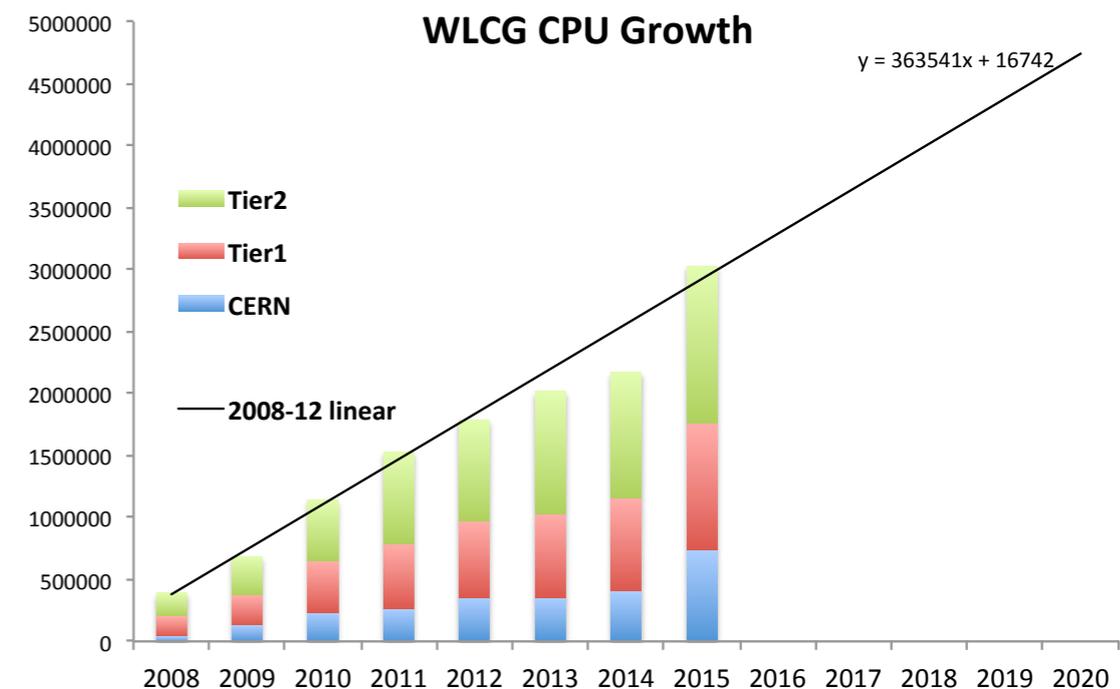
- 3000 fb<sup>-1</sup> at  $\sqrt{s}=14$  TeV
- Precision studies of Higgs and any other new particles yet to be found

- New physics reach example
- SUSY EWK gauginos



# The High-Luminosity LHC will bring new Computing Challenges to the Energy Frontier

- The HL-LHC program will likely have a 10-fold increase in trigger rate and data complexity
  - ✦ computing costs will constrain data rate, and thus the possible choices on triggers and analyses
- LHC raw data: ~15 PB now; ~130 PB in 2021
  - ✦ the total dataset sizes could be up to **10x larger**, once we include processing steps and simulated data
  - ✦ data management must become much more efficient
- With flat budgets (optimistic) we might gain a factor of 4-5 in capacity over the coming decade
  - in the past LHC computing world-wide added ~25k processor cores and ~34 PB of disk, each year
- In future need to make better use of resources as the technology evolves
  - we're still not optimized, and don't use the full capacity
  - Adapting to new processor architectures becomes more challenging, requires specialized (and valuable) expertise
- Storage is cost driver, disks get cheaper only slowly
  - ✦ not all data needs to be on disk — \$10 puts 1M additional events on tape (CMS) — advanced data caching, workflows, data access etc: Big Data technologies!



# R&D and Improvements Needed

---

- In software, CMS has focused on new architectures
  - ✦ low cost and low power ARM processors, high performance GPU and co-processor systems
  - ✦ need to deal with much lower memory/core
- Substantial development effort to improve ability to run across many cores
  - ✦ thread safe code and libraries, algorithm re-engineering
  - ✦ matching the multi-core architecture, but running against Amdahl
- New resources and more techniques in resource provisioning
  - ✦ developing access to opportunistic computing and migrating to cloud provisioning tools
- More efficient use of storage
  - ✦ move to SSD, dynamic data placement, more reliance on data served remotely and content delivery networks
  - ✦ subsequently an even larger reliance on networks (see M.Ernst)
- Reduce sample of data that needs to be re-processed and re-analyzed
  - ✦ use of Big Data tools and techniques to reduce size of samples needed for analysis

# Need for Training and Career Paths

---

- Encourage and support training, as a continuing activity
  - ✦ Use certification to document expertise and encourage learning new skills
  - ✦ Use mentors to spread scientific software development standards
  - ✦ Involve computing professionals in training of scientific domain experts
  - ✦ Use online media to share training
  - ✦ Use workbooks and wikis as evolving, interactive software documentation
- We need to provide young scientists with opportunities to learn computing and software skills that are **marketable** for non-academic jobs
- We need training and **career paths** (including tenure stream) for researchers who work at the forefront of computation techniques and science is critical

# Another Problem: Data Longevity

---

- What to do with 100s of PB of data over 10s of years?
  - ♦ Irreplaceable resource, should be preserved, some how, for the future
    - in the past, e.g. the entire data for a LEP experiment fits on a desktop hard drive
    - still, just 1 of 4 LEP experiments actually demonstrated ability to reanalyze data
- DPHEP Working Group
  - ♦ Convened by International Committee on Future Accelerators (ICFA)
  - ♦ 2 reports, concluded: “an urgent and vigorous action is needed to ensure data preservation in HEP”, “A clear and internationally coherent policy should be defined and implemented”
  - ♦ Tier model: from documentation, outreach level to reconstructed, full data
- Effort at Fermilab to address preservation of Tevatron data
- Emerging LHC Data Preservation and Access plans
  - ♦ outreach efforts using “Tier 2” data already
  - ♦ serious amount of work needed for Tiers 3 and 4
- DASPOS project to link to Biology, Astrophysics, Digital Curation

# Summary

---

- LHC computing at the current scale is hardly limited by technology
  - ✦ but it is already limited by what we can afford or are willing to spend!
  - ✦ experiments make choices and set priorities about which type of events can be collected, what data analyses can be followed through, based on how much computing and storage resources can be afforded
- ==> Computing and ability to process data is often a limiting factor for the physics program
  - ✦ fundamentally the data set size is given by trigger rate, i.e. physics
  - ✦ computing resource needs ~follow, modulo optimizations of work flows
- LHC data set size is continuing to grow, driven by the even rates
  - ✦ expect to go someplace in the neighborhood of 10-20 kHz sustained rate
  - ✦ only known alternative: store just pre-processed/synthesized data
    - Would be a big change of mentality and in the scientific process
- More storage and processing capabilities make for better physics
  - ✦ while the existing data tiering techniques and the successful federated/distributed data storage/access model predict good scalability

# Outlook

---

- Success of the ambitious LHC program and upgrades requires significant advances in computing and software
  - ✦ New technologies and approaches have been and will be transformative
  - ✦ Distributed computing, networks, parallelization, virtualization, GPUs etc
- Industry has caught up to us, and in cases has surpassed us
  - ✦ Should take advantage of progress in industry and advances in other sciences
  - ✦ There's much to leverage! Big Data is good news, Clouds are good news...
- Investing in community planning, solve common problems in partnerships
  - ✦ Investments in common development are important, and require support
- Collaboration and partnerships are essential
  - ✦ Distributed computing requires partnerships, between sites, science communities, with computer scientists, between agencies etc
  - ✦ The FCE is starting with a software focus, and a computing institute would help !
- We have amazing challenges, and progress in computing immediately translates into advances in the science domain
  - ✦ Many opportunities to partner with ASCR experts and facilities