High Energy Physics and the issues of Software Sustainability

Rob Roser, CIO
Fermi National Accelerator Laboratory
(roser@fnal.gov)

Can we explain this!

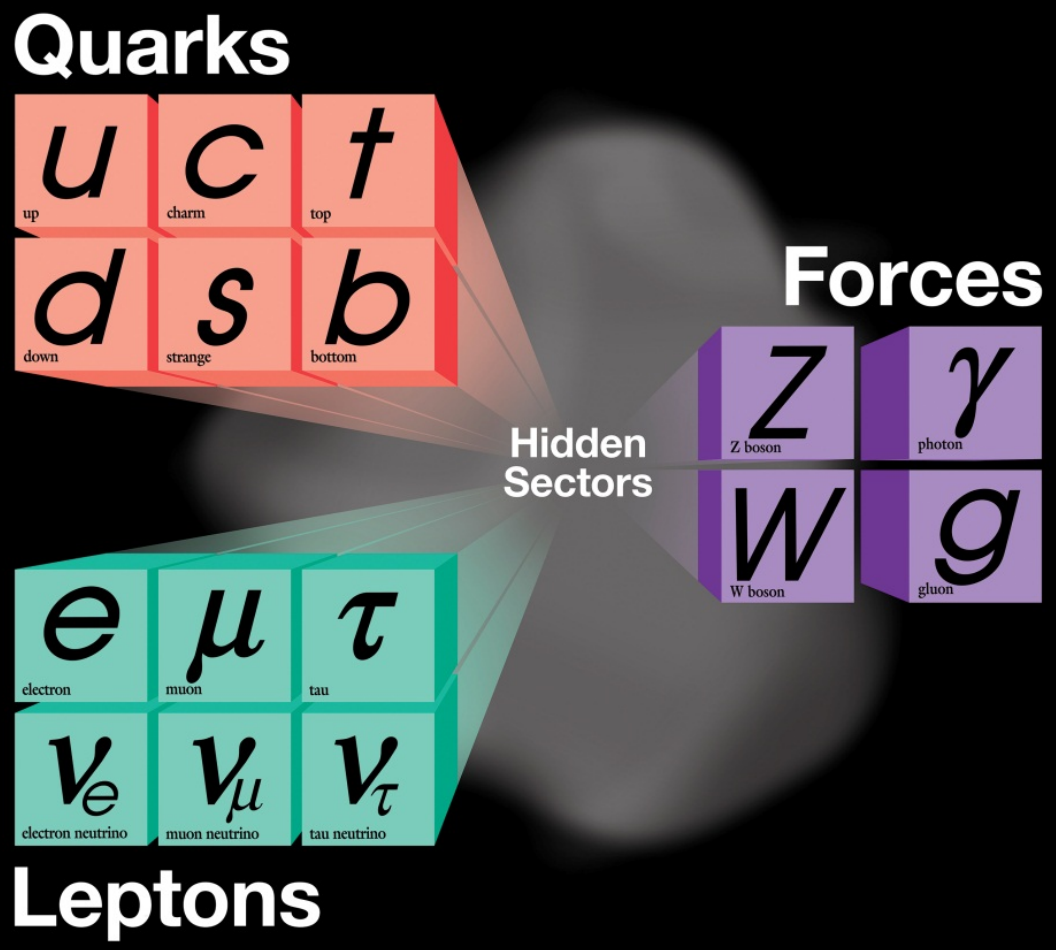
The Current State of Knowledge

The Standard Model

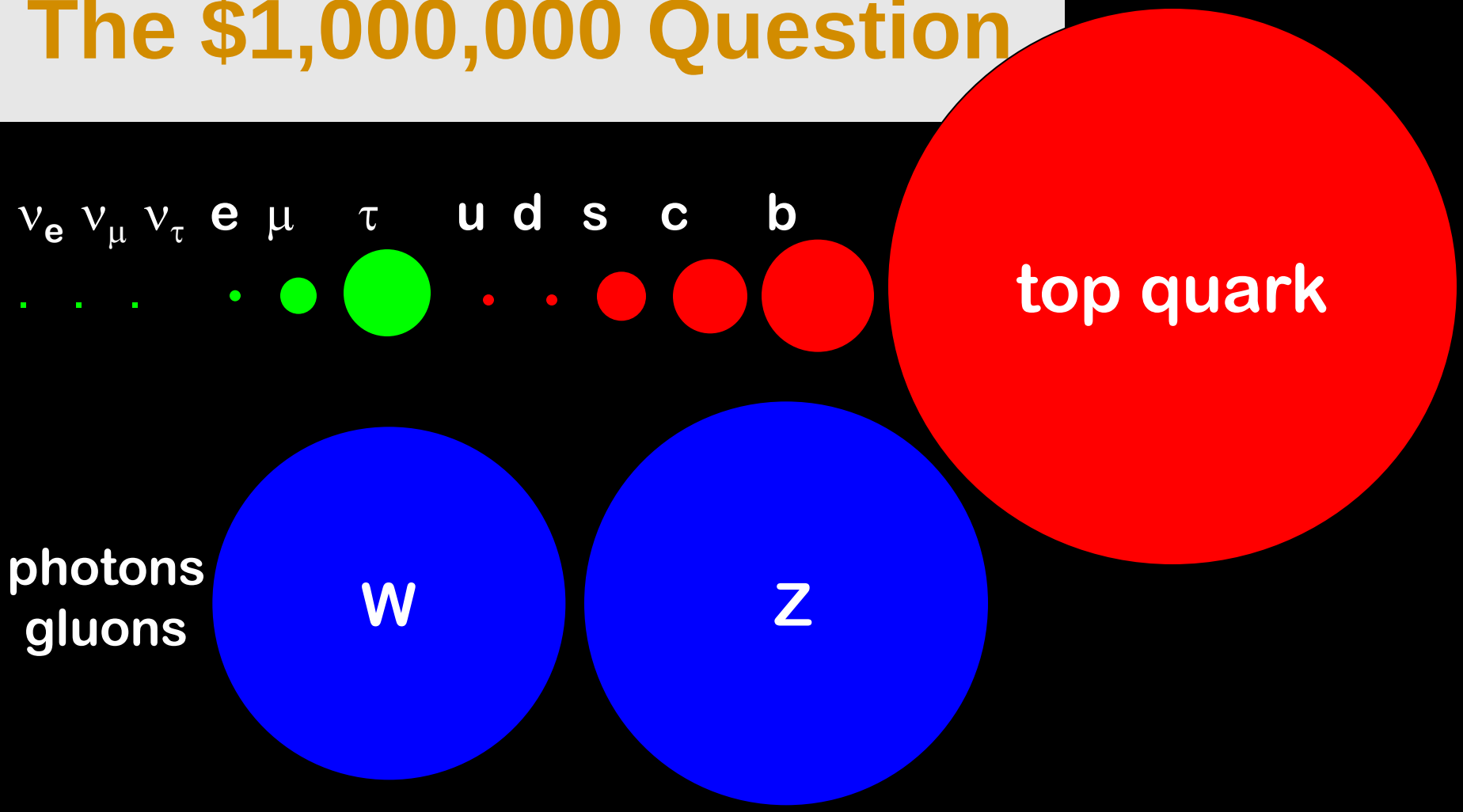
Greatest scientific achievement of 20th century

Every particle physics experiment ever done fits with this model

But it is incomplete…



The $1,000,000 Question



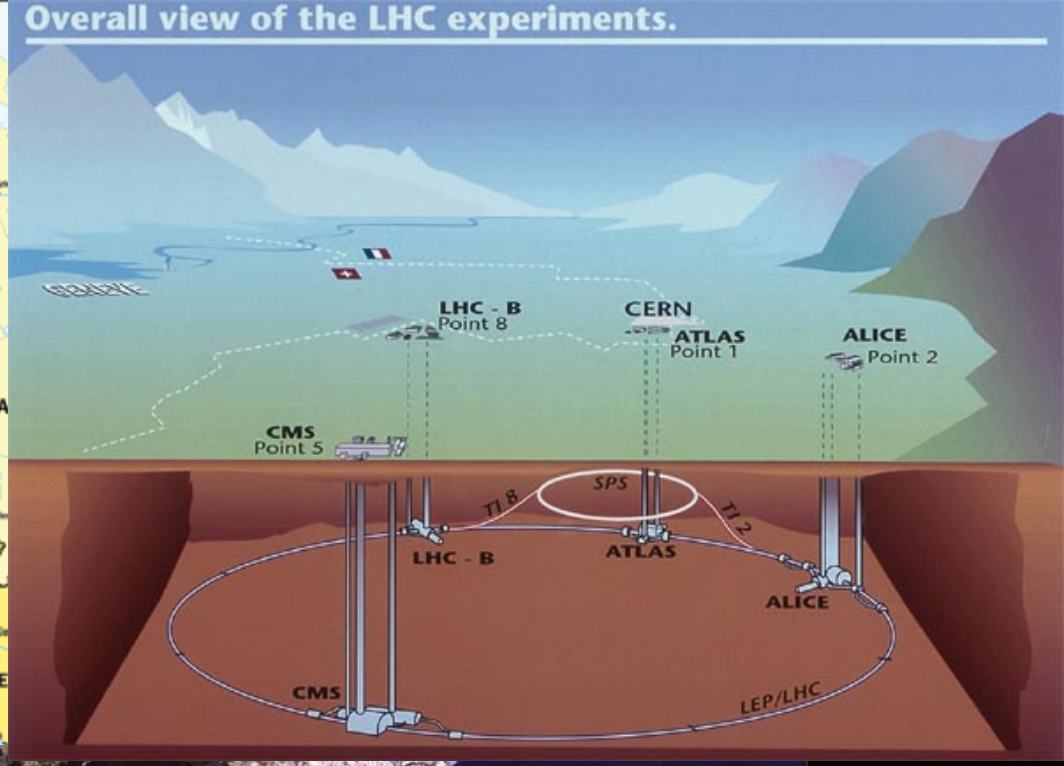
Why is the top quark so heavy?

Why are there so many particles?

Where does mass come from?

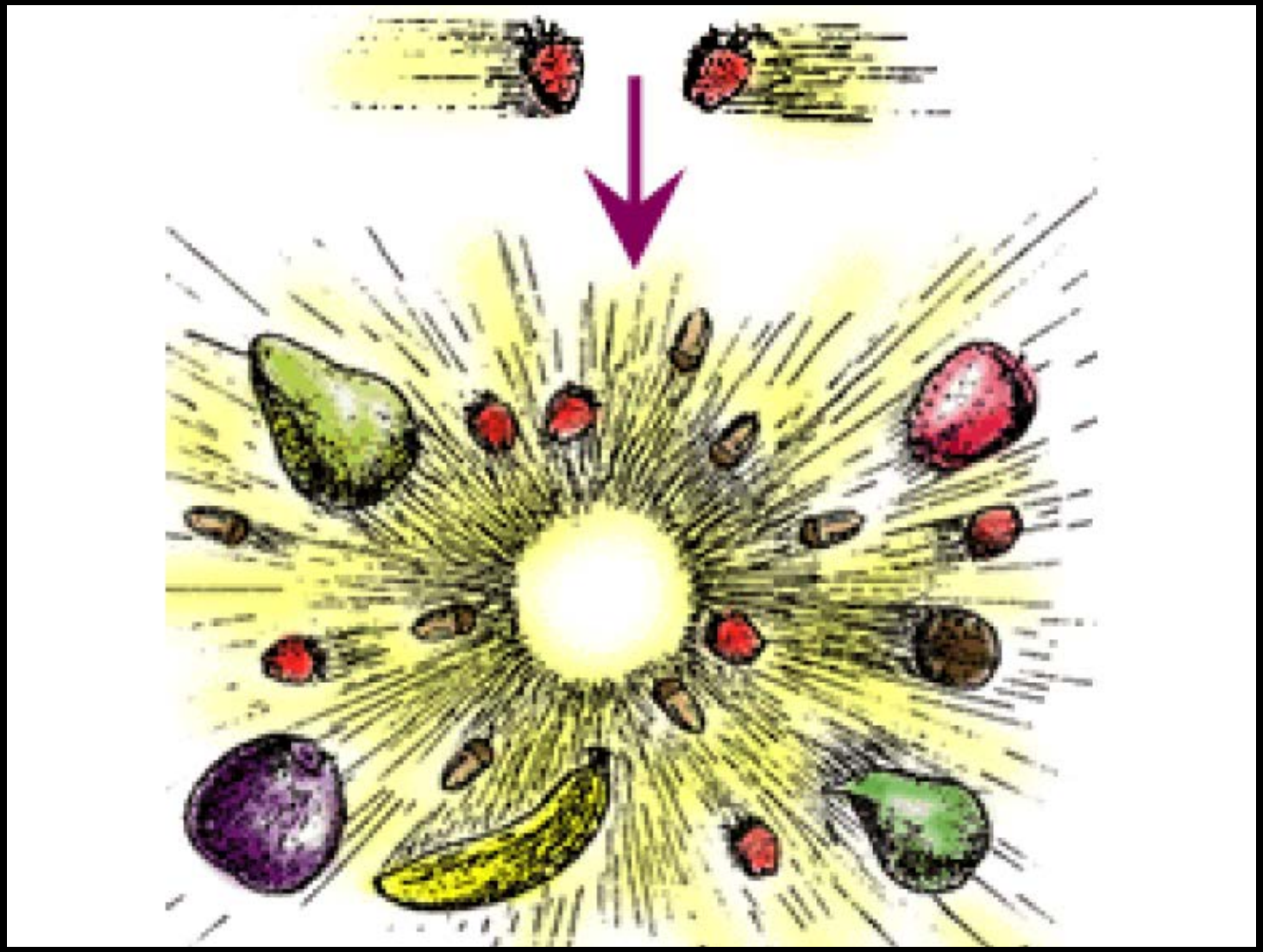
The Large Hadron Collider [LHC]

ICELAND
NORWAY
SWEDEN
FINLAND
RUSSIA
ESTONIA
LATVIA
LITHUANIA
BELARUS
DENMARK
UNITED
KINGDOM
IRELAND
NETH.
BEL.
GERMANY
POLAND
UKRAINE
CZECH REPUBLIC
SLOVAKIA
AUSTRIA
FRANCE
SWITZERLAND
ITALY
CROATIA
PORTUGAL
SPAIN
MONACO
ANDORRA
MOROCCO
ALGERIA
TUNISIA
MALTA
Norwegian Sea
Greenland Sea
North Atlantic Ocean
North Sea
Mediterranean Sea
Celtic Sea
Bay of Biscay
Barents Sea


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

LEP
CERN Prévessin
SPS
CERN Meyrin

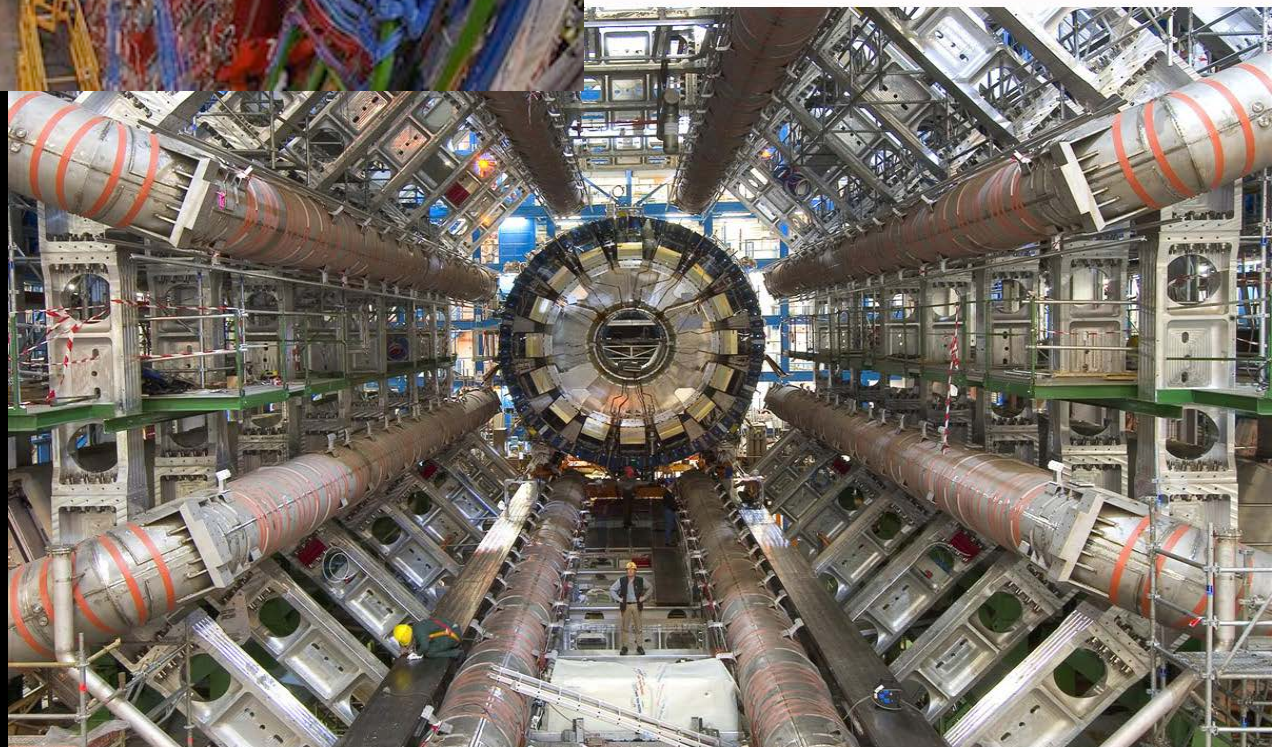
LHC



Can we make the worlds most delicious fruit?

Shamelessly stolen from Ian Shipsey

CMS and ATLAS



When Science meets popular culture

Higgs Boson

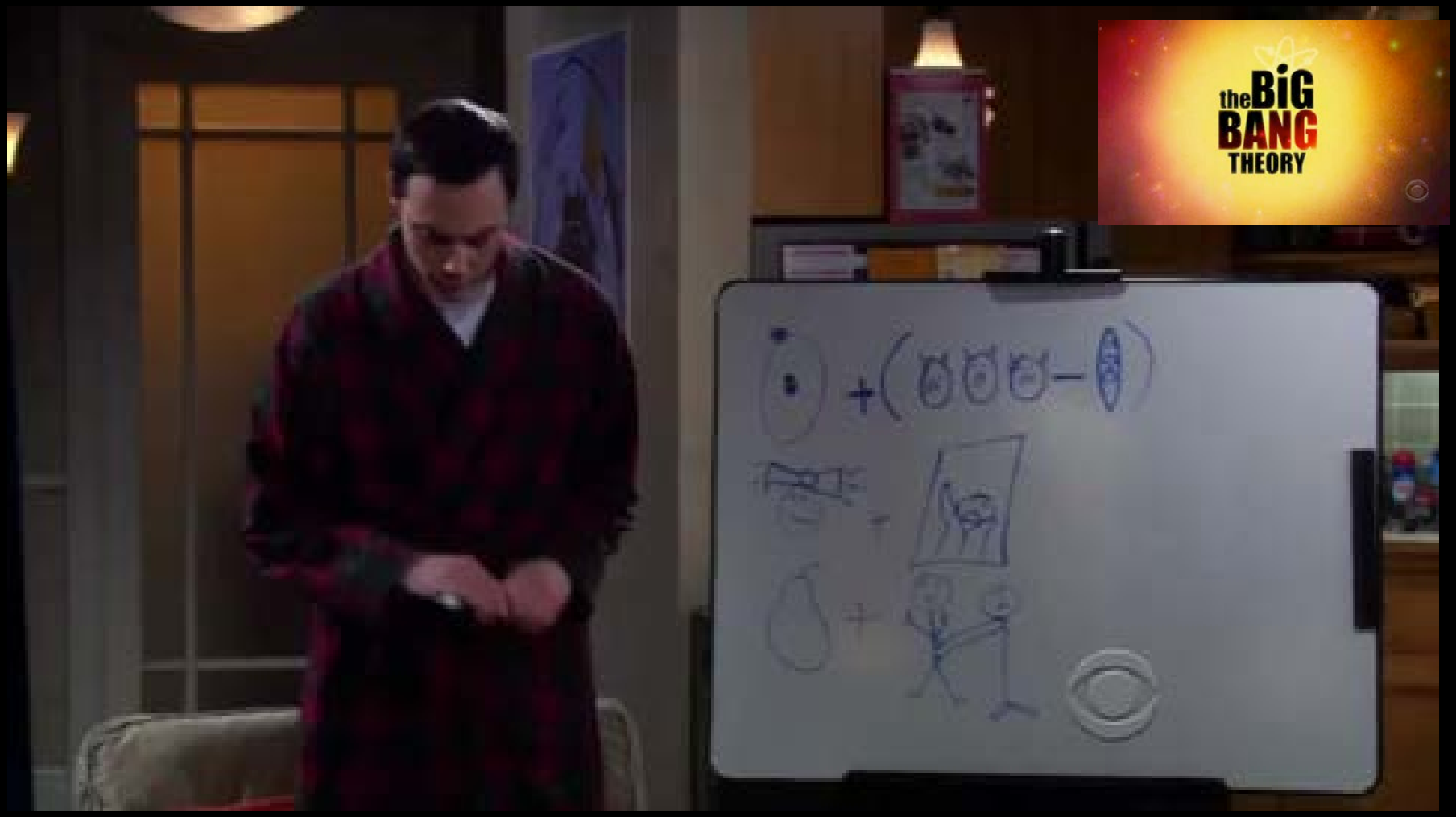
$E=MC^2$

The God Particle

Nano technology

But you know you have arrived when….

The Big Bang Theory

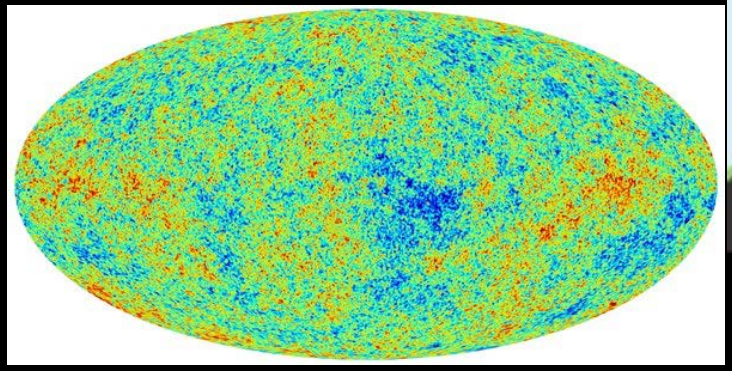


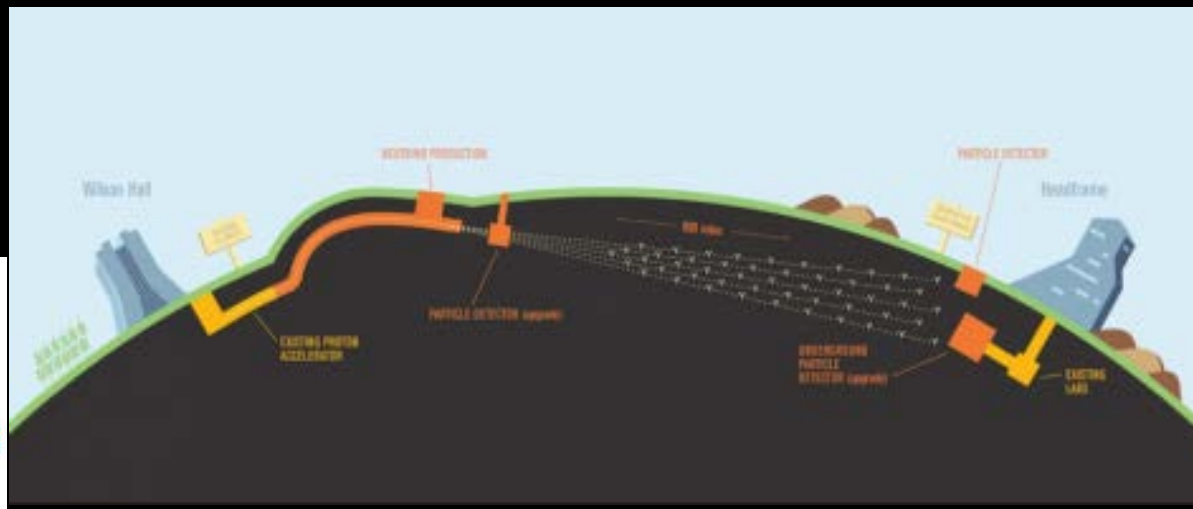
Fermilab Overview

- Only National Laboratory solely dedicated to HEP research
- More than 4,200 scientists worldwide use Fermilab and its particle accelerators, detectors and computers to perform their research.

Fermilab Moving Forward

A Vibrant Program





Major Software "Products" at Fermilab

Scientific toolkits
- ROOT
- OSG gLExec / authorization
- CVMFS
- Geant 4
- Pythia
- GENIE
- Synergia
- MARS

Production software frameworks
- Art
- Artdaq
- LArSoft
- CMSSW MT FWK

R&D - advanced technologies
- ***SciDAC and Computational HEP***
  - Synergia - Accelerator modeling on GPGPU, MIC
  - Geant4 - reengineering on GPGPUs
  - ART – Computational Cosmology on HPC platforms
  - LQCD - algorithms on GPGPU/MIC
- ***Programs & projects***
  - CMSSW multi-threaded framework
  - IF trigger on multi-core architectures
  - LSST Dark Energy analysis toolkit
  - CosmoSIS – MCMC framework
  - Big data explorations

Large-scale computing
- Job & data management
  - SAM (+IDFH+FTS)
  - Jobsub
  - WMAgent
  - Glide-in WMS
  - FIFEmon
  - Aux File Cache
- Database & collaborative
  - FroNTier
  - IFBeam
  - ECL (control room logbook)
  - Gratia

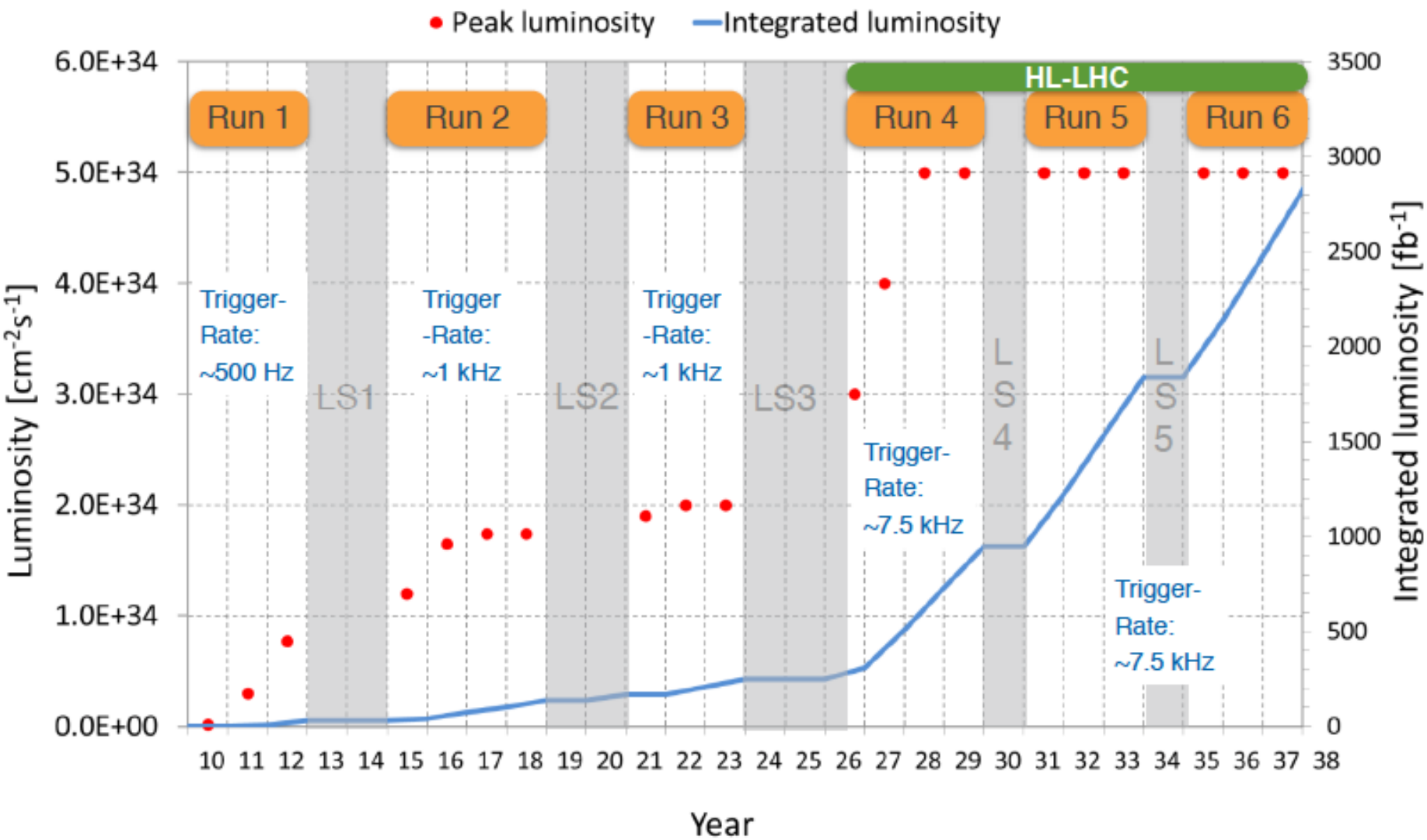
R&D – LDRD
- GPUs for network packet analysis
- DAQ & the internet of things
- Automata Processors for next generation pattern recognition

Challenges Facing HEP Moving Forward

- Big Data – ability to process, store, move and plot
- Emerging Technologies – offer new opportunities but also new difficulties -- paradigm shift as we move to more complex parallelism computing architectures
- Computing is more specialized – need experts that don't exist right now in HEP
- Tight budgets – HEP can provision resources for average – need ways to deal with peak loads (commercial cloud, HPC, University clusters, etc)
- History – huge and successful codebase that needs to be adapted as computational technologies change
- A customer base that is used to Amazon.com, Ebay, etc – sets expectations both in terms of useability as well as agility – how do we live up to those expectations?

LHC Schedule

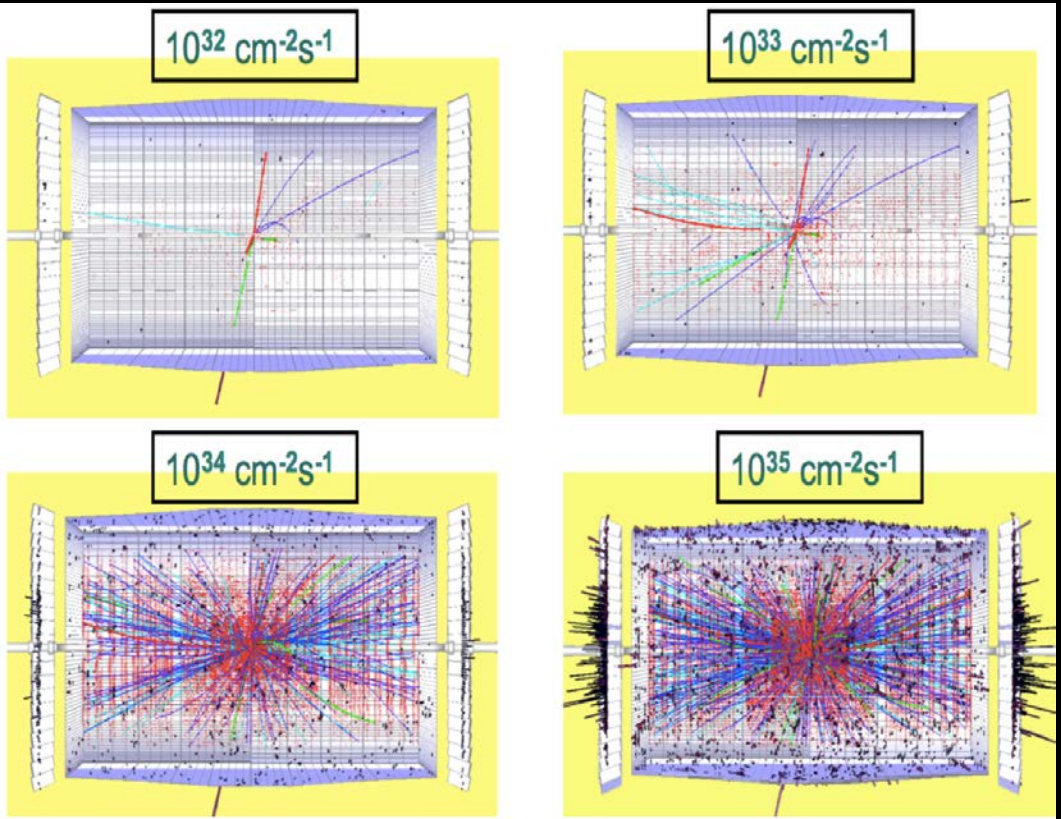
Plot Spans 28 Years!!



Software and Experimental Complexity

- Code base for a given experiment is significant – millions of lines of code written by hundreds of people over time.

- Software Complexity increases as experiments push the envelope of what is possible. Algorithms must keep pace.



Technology Changes

- Computing models must be flexible – the will change as the technology evolves.
  - Constant tension between compute, storage, and networking and where to emphasize ones investment
- The road to the Exascale
  - National Super Computing Initiative will be very important to HEP
  - 5% of an exascale machine is 50 Petaflops!!!
  - Commercial clouds starting to become players in scientific computing
  - HEP needs to be part of this initiative
- Bulk of HEP software is currently not set up for multi threading or a highly parallelized environment.

HEP must modernize its software stack

Data/Knowledge Preservation

- Never undertaken before by HEP – until recently with the Tevatron and BABAR experiments
- Curation of Data – relatively straight forward – just requires time, $$$ and desire
- However a digital library of data w/o the knowledge of how to read and understand the data is useless
- Thus we tend to refer to data preservation as “knowledge preservation” – archiving not only the data – but the tools to understand it
- Knowledge preservation places significant demands on software sustainability
  - Compilers
  - Documentation of code
  - Test jobs
  - Etc.

Data/Knowledge Preservation

- Knowledge Preservation Means we have to preserve
  - the reconstruction software
  - Monte Carlo model of detector and ability to run a “modern” event generator and then simulate it
  - Analysis software and definition of variables
  - Host of data base information
    - A way to validate (making a well understood physics measurement before searching for something new
  - Curation would be easier if experiments thought about it during the software design process – thus far – that has not been the case
  - Tools to help with this curation are highly desierable.

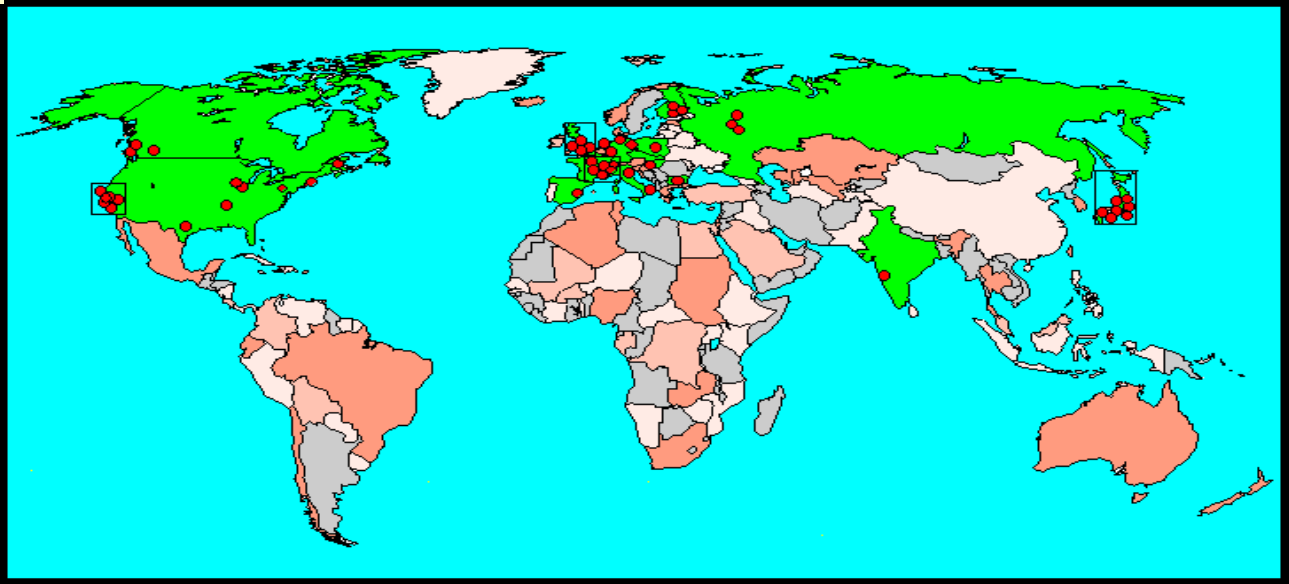
The Role of Simulation

- Its ALL about Simulation
- Spend much more computing time on simulation than we do with the instrument data
- Everything we do involves simulation
  - Investigating whether a proposed experiment has any scientific merit
  - Designing the detector
  - Designing the analysis
  - Understanding what the “signal” looks like
  - Understanding what other physics mimics that signal “background”
  - Calculating how many signal events we should see?
  - ….

What is Geant4?

- Geant4 is a tool kit that tracks particles through matter, breaking the particle motion into small segments, applying appropriate physical processes and probabilities at each segment.
- These processes can destroy old particles, modify state or create new ones
- Processes include atomic processes like ionization and excitation, decay processes, photonic transitions, secondary emission, etc..
- The wide coverage of physical processes comes from mixture of theory-driven, parameterization, and empirical formulae.
- Software package contains millions of lines of code.

Geant4 International Collaboration



J.W.Goethe Universität

Forum for Computational Excellence

- Cross cutting HEP initiative to better leverage skills in our different research area's and provide a single group to engage with outside agencies
- Sponsored a series of reports to understand the computing "landscape" in HEP today – www.fce.org
- Led workshop with ASCR on exascale requirements – want to leverage HPC for both computational intensie and data intensive use cases – report to be released soon
- Building a library of mini-apps to describe many of our key software kernels, workflows etc
- Other initiatives…

$$$ Money $$$

- We need to talk about this
- HEP does not explicity fund software maintenance
- Is that saying that Sustainability is not important?
- People who take on this important task find little in the way of rewards in terms of careers/career path

- How do we change this?

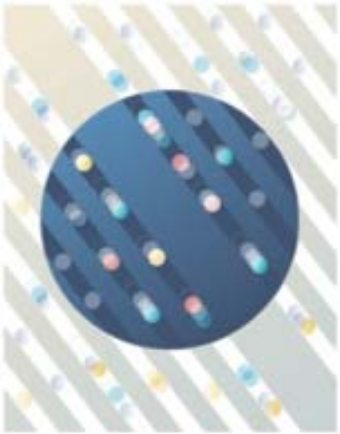
Take-Away Messages

- HEP can not accomplish its mission without sophisticated computational Tools
- Software sustainability is more important now than ever before
- HEP is not tackling this important problem in any uniform way
- Are there tools that can be developed to help?

"That's all Folks!"

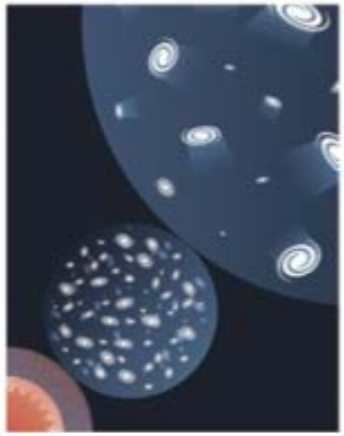
Fermilab is a Research Lab

Higgs boson



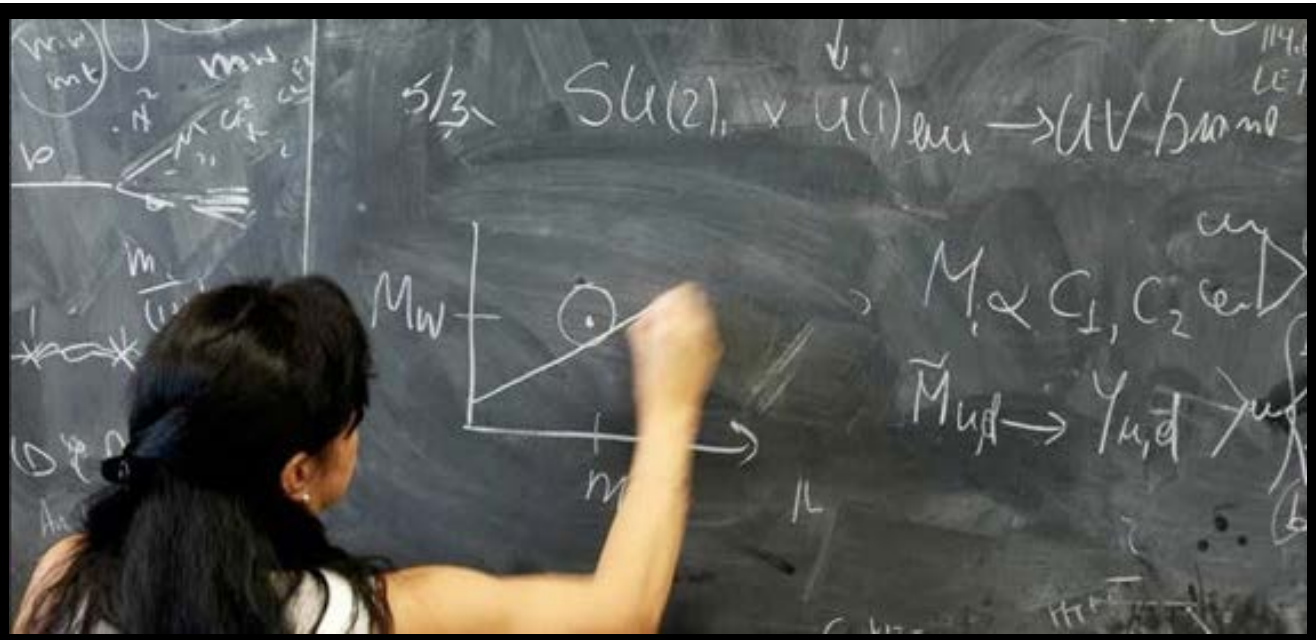
Neutrino mass

Dark matter



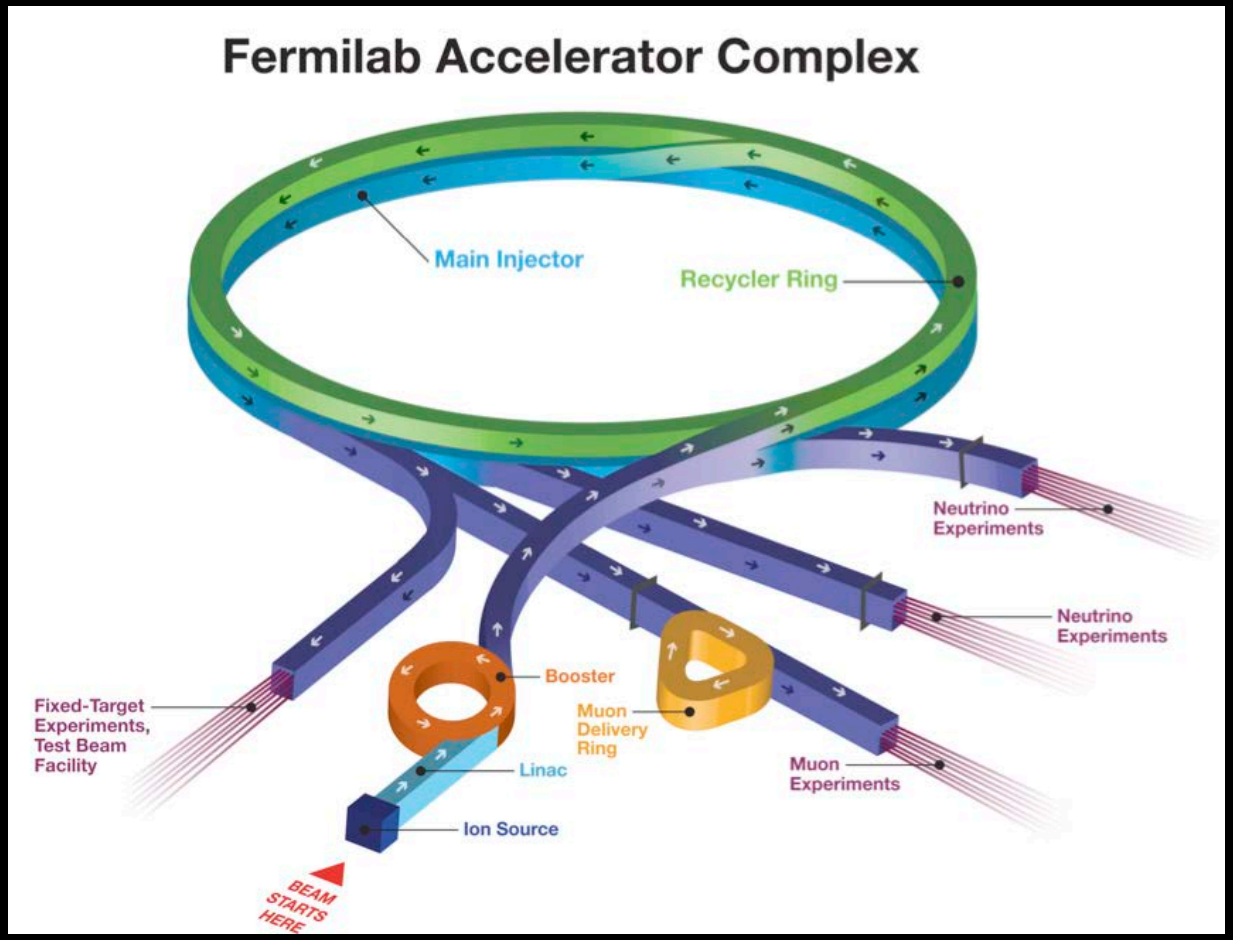
Cosmic acceleration

Explore the unknown



Fermilab is an Operations Lab

- Accelerator
- Detector
- Computing



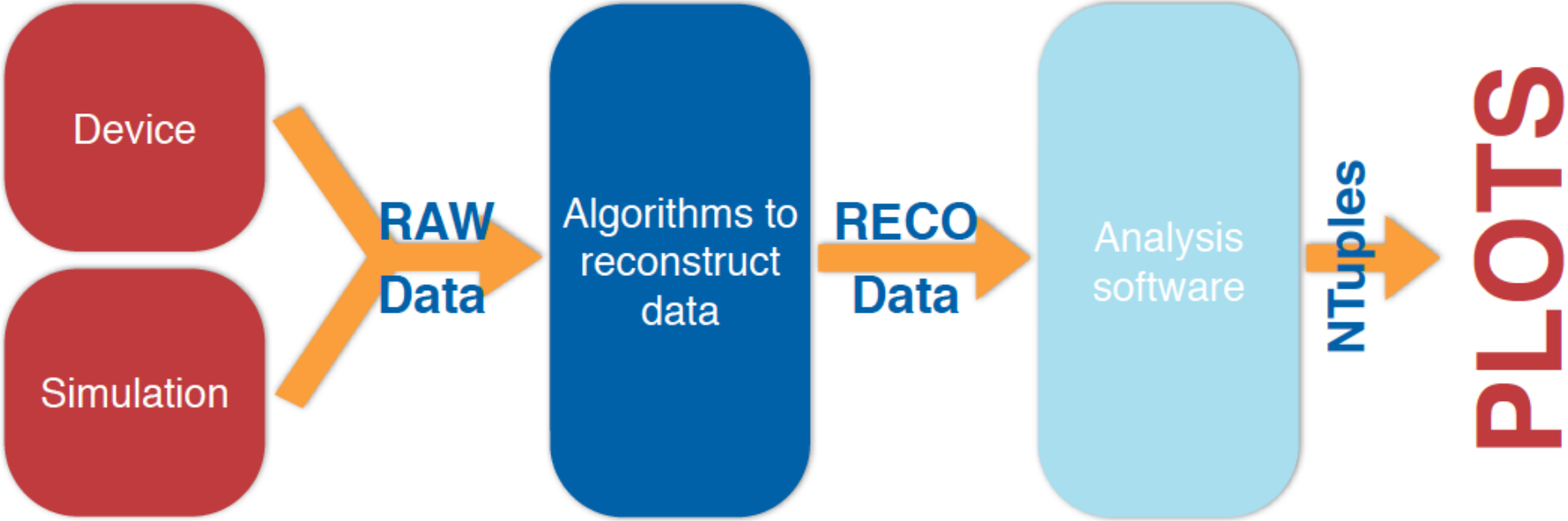
Why Does One Do High Energy Physics?

- The Experiments are large -- many operating experiments EACH have ~2000 physicists.
- Author list might take more pages in a journal than the actual article
- Physicists have to travel great distances to get to their experiments
- Startup times for experiments are large; already working on experiments and accelerators that won't run till 2025 or beyond
- The apparatus is far from being table top in size
- A single experimenter does not have total control over his or her environment

One Person's Answer – My Answer

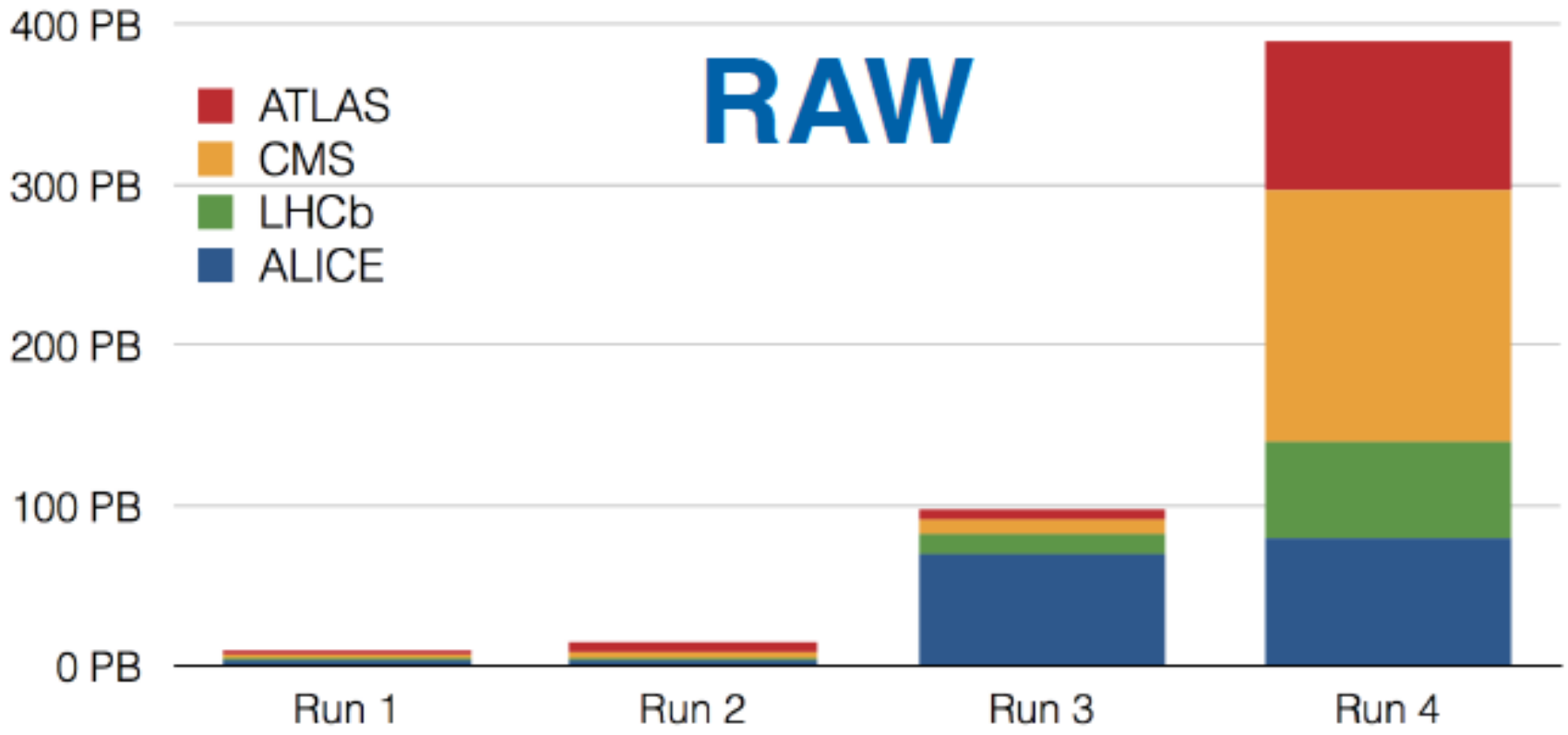
- We want to understand some of the most fundamental questions in nature
- We want to understand how the universe was created and the laws that govern it
- We, as physicists have an insatiable curiosity for the world we live in and how it works
- We enjoy working with a large group of smart people toward a common goal
- The variety of physics topics that these new experiments can address means life is never boring
- BECAUSE ITS FUN

How We Do our Work



- **Software** is important for every step on the way to scientific results

LHC Expected Data Volumes

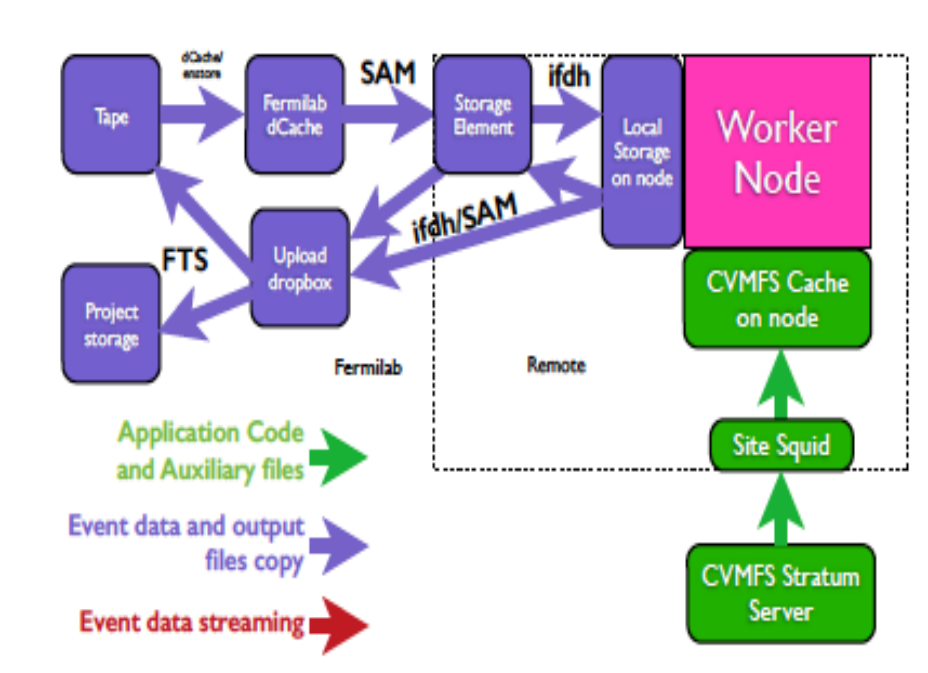


Shown is Raw Data

- Derived Data – (reconstructed+Simulation) ~8X raw!
- LHC Run 4 Starts the Exabyte era (2025)
- How do we deal with that?

Trend – Dynamic Data Placement

- Subscription Based Transfer Systems
- PhEDEx (CMS) and Rucio (ATLAS)
- LHC Run 1 – data movement was a mostly manual operation
- LHC Run 2 – dynamic Data management
  - Popularity is tracked per data set
  - Replica count across sites is increased or decreased according to popularity



Sam – Fully integrated data distribution system – used by IF expt's

- All movement based on requests for datasets from jobs
- Interfaces to storage at sites – can do cache-to-cache copies

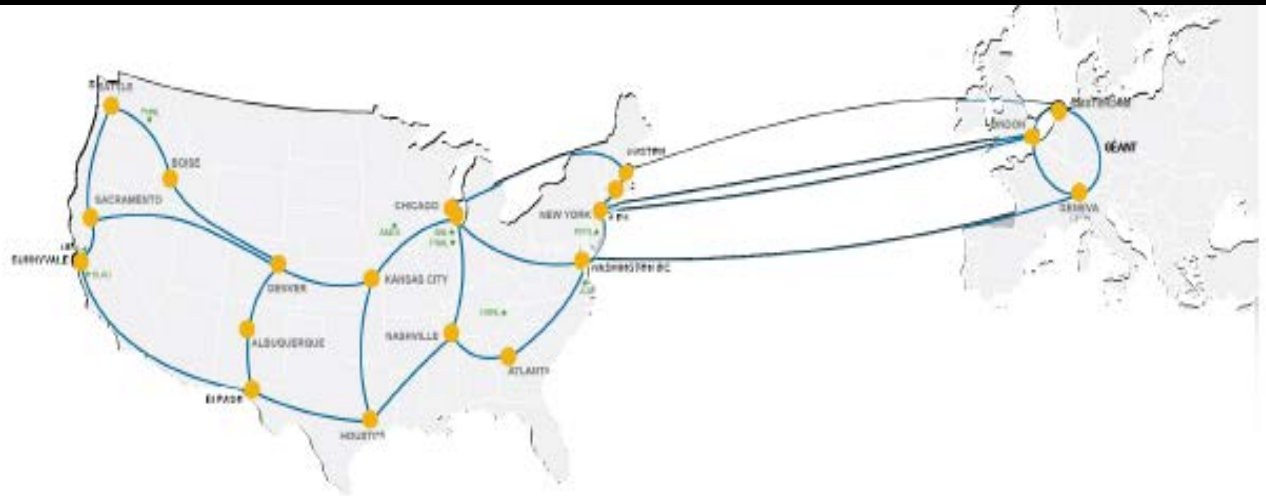
Active Archival Facility

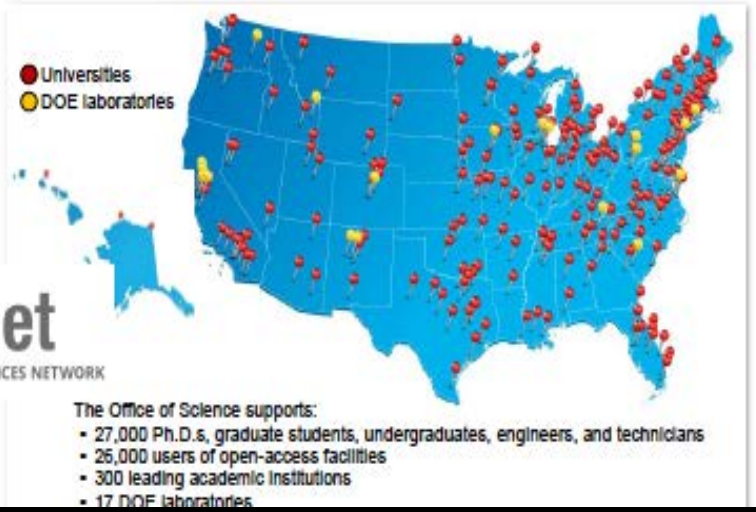
- HEP has the tools and experience for the **distributed exabyte scale**
  - We are "best of class" in the field of scientific data management
- We are working with and for the whole science community
  - To bring our expertise to everyone's science
  - To enable everyone to manage, distribute and access their data, globally
- Example: Fermilab's Active Archival Facility (AAF)
  - Provide services to other science activities to preserve integrity and availability of important and irreplaceable scientific data
  - Projects:
    - Genomic research community is archiving datasets at Fermilab's AAF and providing access through Fermilab services to ~300 researchers all over the world
    - University of Nebraska and University of Wisconsin are setting up archival efforts with Fermilab's AAF

What Makes Geant4 Unique?

- Geant 4 is distinguished from other Monte Carlo Particle Transport codes by
  - The comprehensive suite of physics processes and particle types
  - The complexity of geometrical descriptions leads to realistic representations.
  - A collaborative open source model leveraging international expertise
  - Enables the user to select physics processes/models and choice of GUI/visualization/persistency/histogramming technologies

Strong Networks Crucial





ESnet
ENERGY SCIENCES NETWORK

INTERNET2