The Computing Challenges Facing High Energy Physics

Rob Roser - Fermilab

Thank You

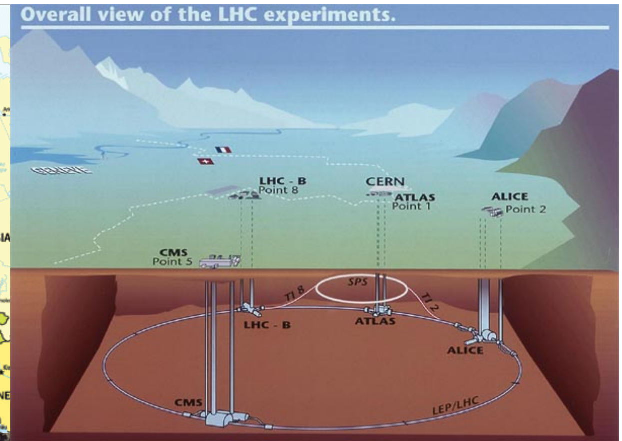
- Ollie Gutsche
- Lothar Bauerdick
- Panagiotis Spentzouris
- Harvey Newman

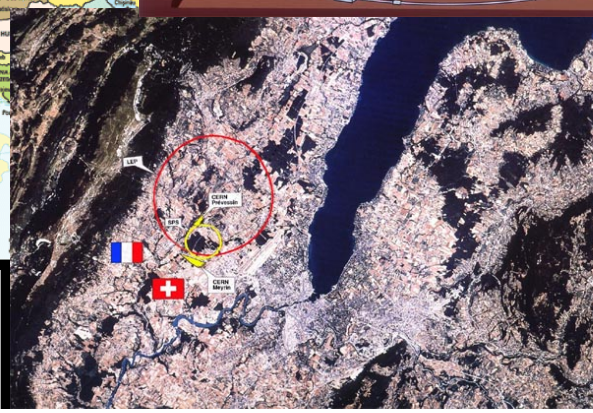
For help in preparing this talk!!!


Can we explain this!


The Large Hadron Collider
[LHC]


ICELAND
NORWAY
SWEDEN
FINLAND
RUSSIA
UNITED KINGDOM
IRELAND
GERMANY
POLAND
BELARUS
UKRAINE
CZECH REPUBLIC
AUSTRIA
FRANCE
SPAIN
PORTUGAL
ITALY
ALGERIA
MOROCCO
TUNISIA
Norwegian Sea
North Atlantic Ocean
North Sea
Mediterranean Sea

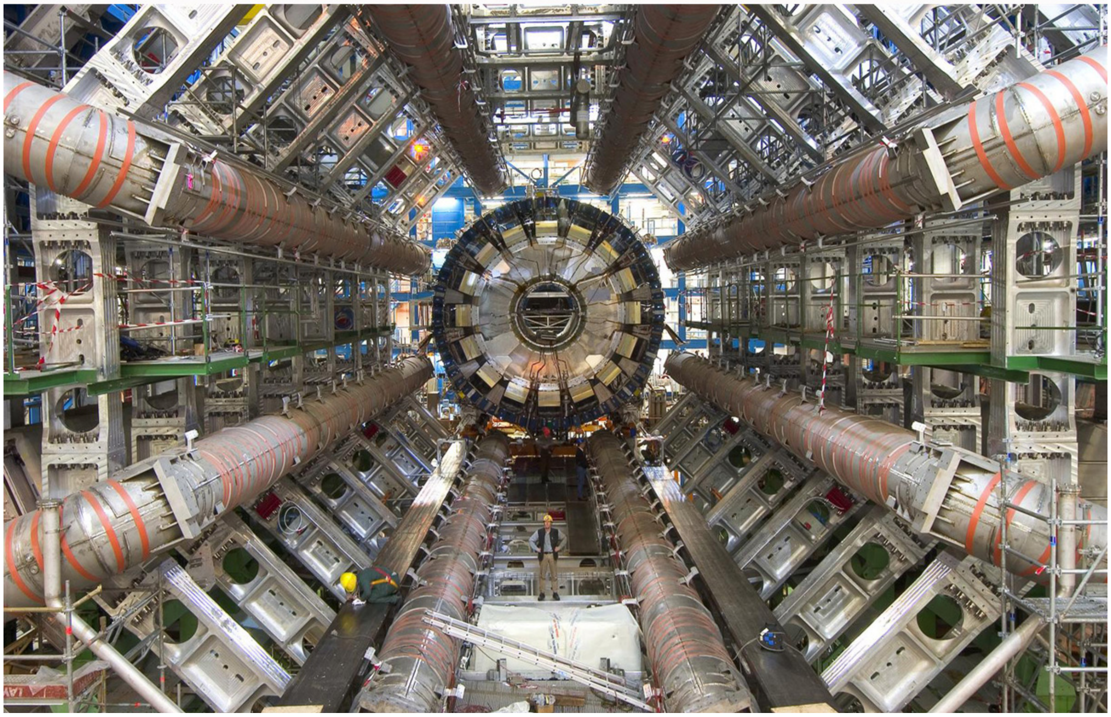

Overall view of the LHC experiments.
LHC - B
Point 8
CERN
ATLAS
Point 1
ALICE
Point 2
CMS
Point 5
SPS
TI 8
TI 2
LHC - B
ATLAS
ALICE
CMS
LEP/LHC



LHC

CMS

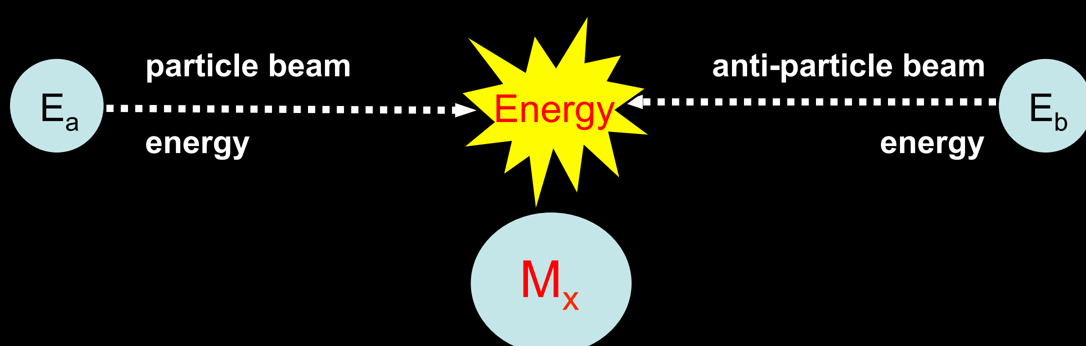
ATLAS



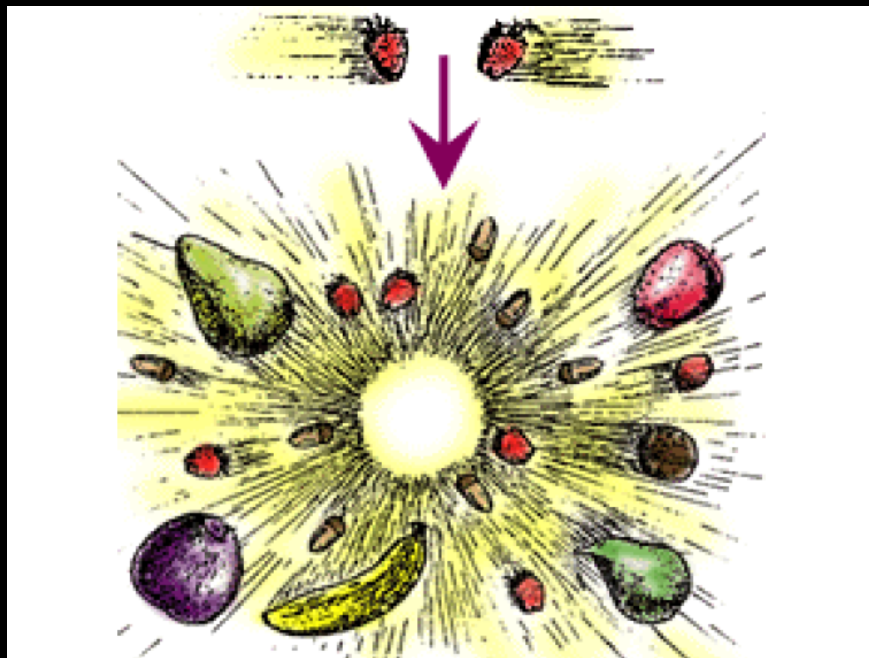
Making a Higgs

Thanks to Einstein we know that a high-energy collision of particle A and B can result in the creation of particle X

$$E = mc^2$$



As long as $E_a + E_b >> M_x c^2$



Can we make the worlds most delicious fruit?

Shamelessly stolen from Ian Shipsey

When Science meets popular culture

Higgs Boson

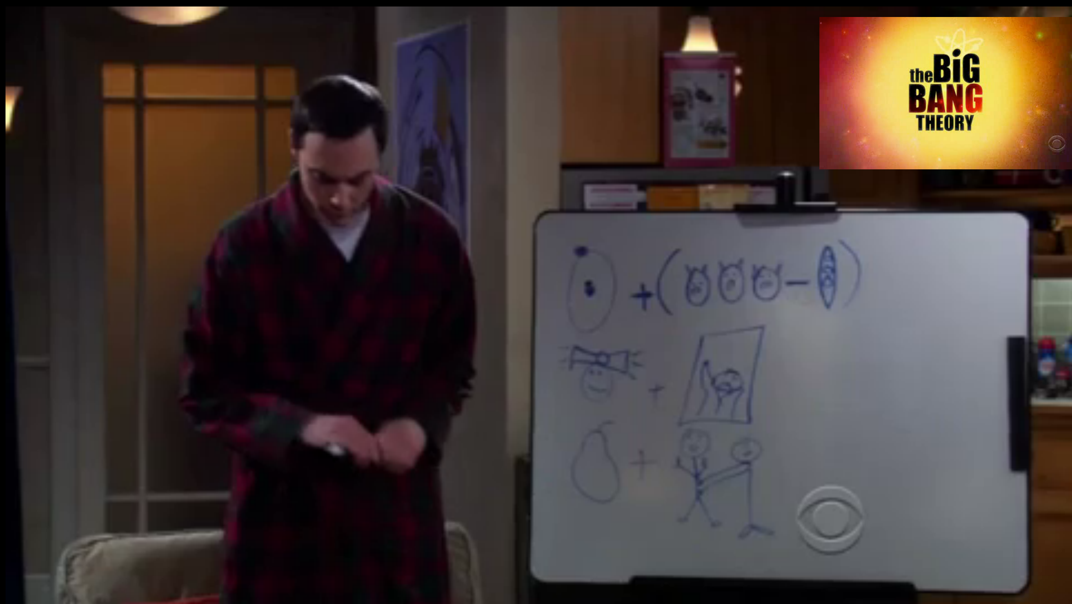
$E=MC^2$

The God Particle

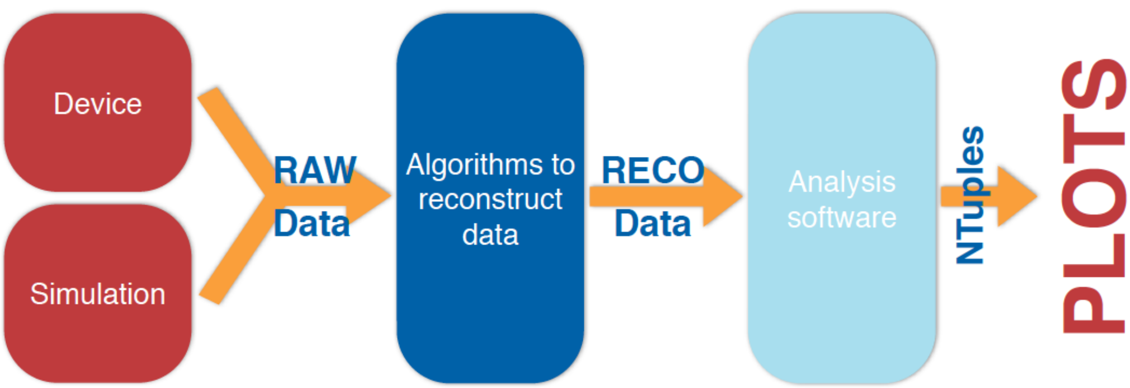
Nano technology

But you know you have arrived when….

The Big Bang Theory



How We Do our Work



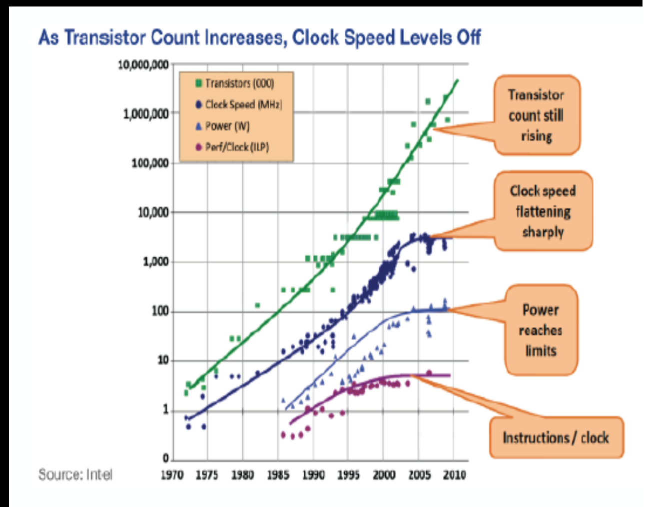
- **Software** is important for every step on the way to scientific results

Historical – Moore's Law

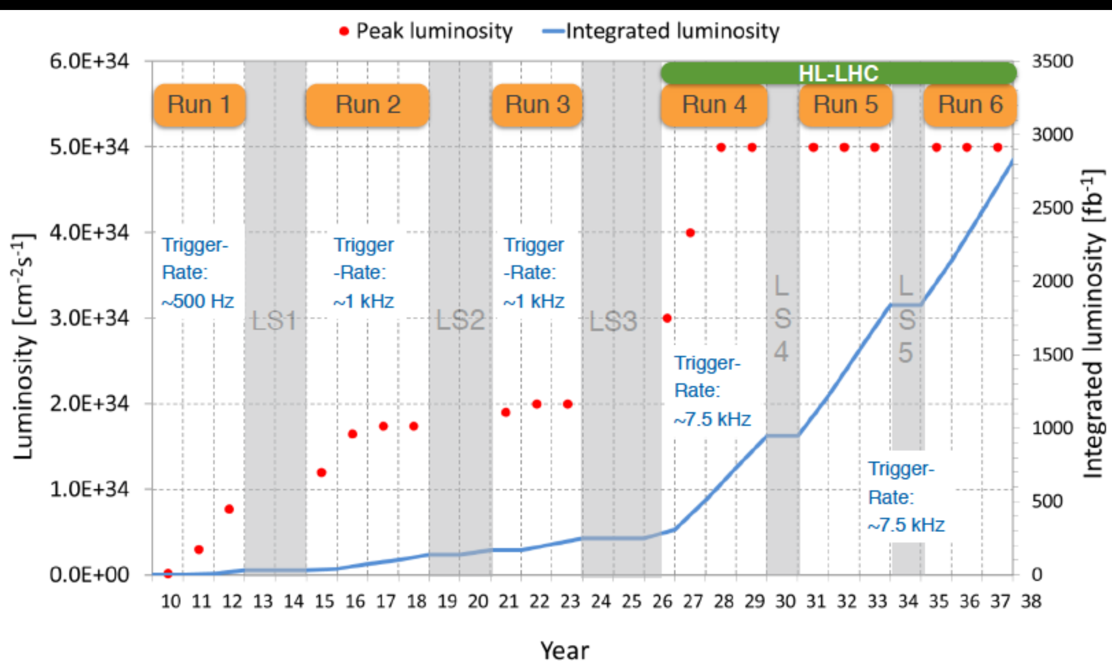
Traditionally, HEP software is optimized for a "simple" architecture

- X86 based Linux
- Machines
  - >1 CPU's with >1 Cores
- Shared memory
- Shared local Disk Space

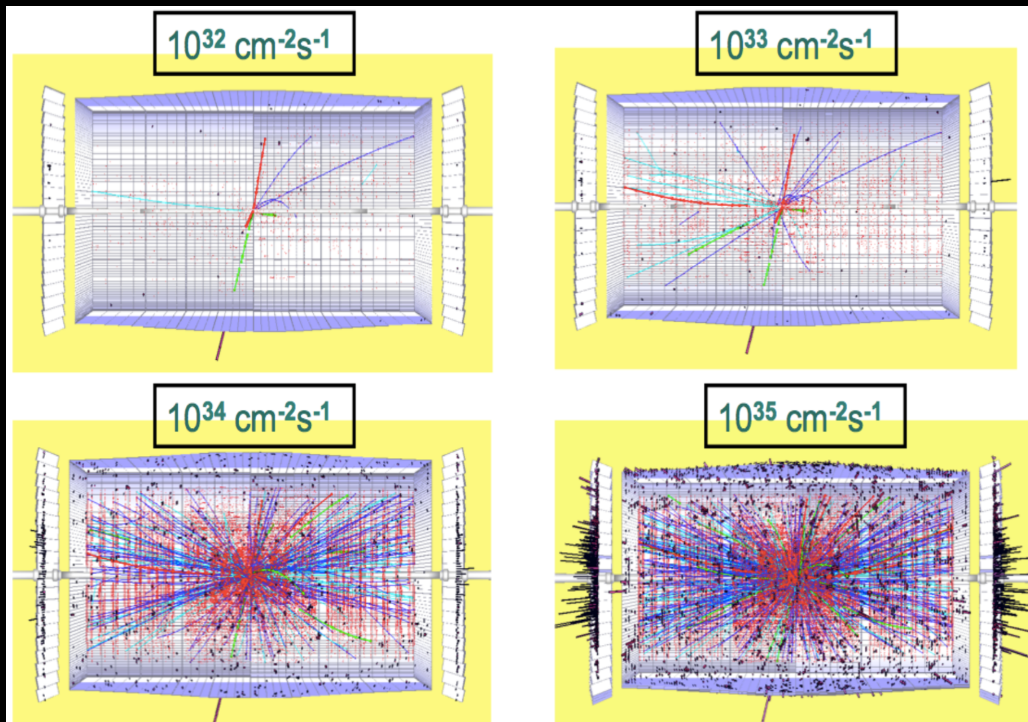
- A given application uses one core, memory and local disk space



LHC Schedule



First Challenge -- Event Complexity and Data Rates



The LHC: Spectacular Performance

Data Complexity: The Challenge of Pileup

~3.5 X 10^{15} pp Collisions
1M Higgs Bosons created in Run 1

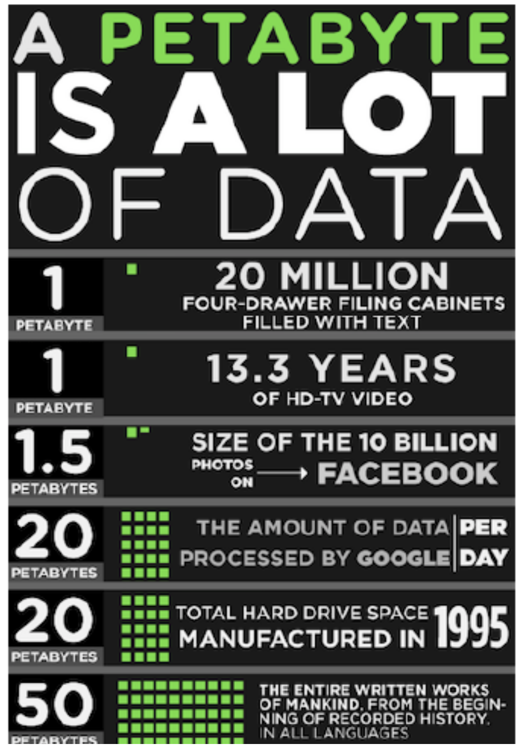
~50 Vertices, 14 Jets, 2 TeV

Run2 and Beyond will bring:
- Higher energy and intensity
- Greater science opportunity
- Greater data volume & complexity
- A new Realm of Challenges

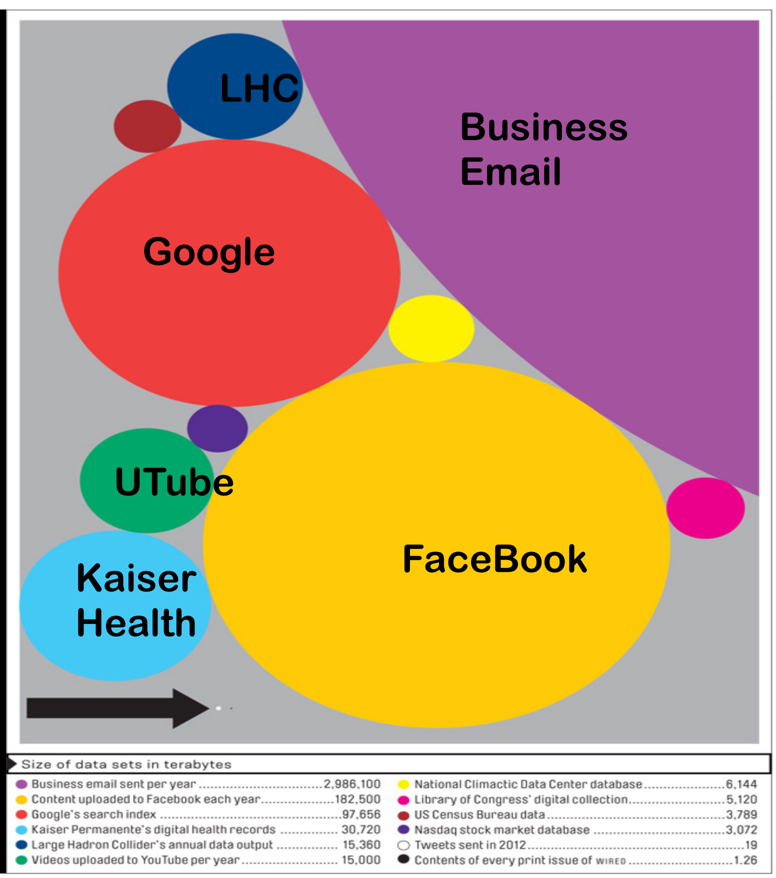
Average Pileup

Run 1 21
Run 2 42
Run 3 53
HL LHC 140-200

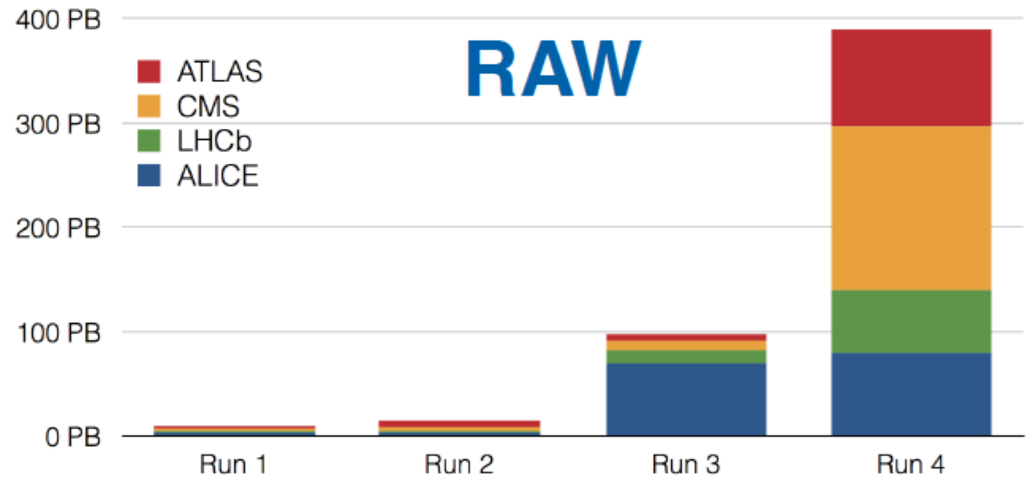
What is a Petabyte?



From Wired Magazine



LHC Expected Data Volumes



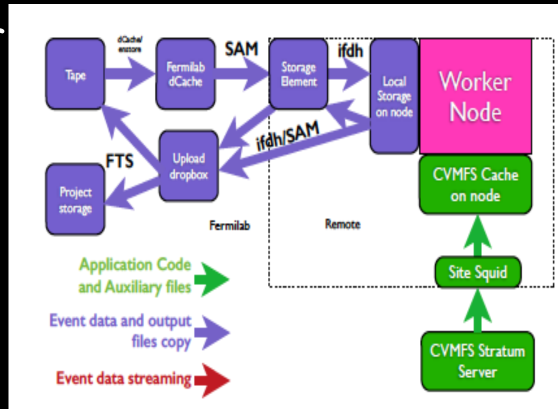
Shown is Raw Data

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

CERN Movie

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

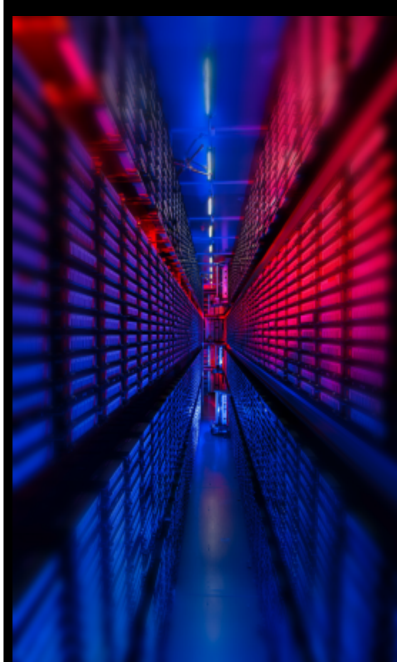


Sam – Fully integrated data distribution system – used by IF expt's

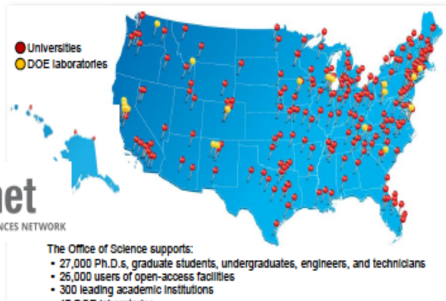
- All movement based on requests for datasets from jobs
- Interfaces to storage at sites – can do cache-to-cache copies

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



Strong Networks Crucial



ESnet ENERGY SCIENCES NETWORK

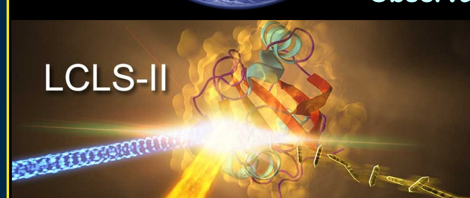
INTERNET2

Entering a New Era of Technical Challenges as we Move to Exascale Data and Computing

- The largest science datasets today, from LHC Run1, are 300 petabytes
  - Exabyte datasets are on the horizon, by the end of Run2 in 2018
  - These datasets are foreseen to grow by another 100X, to the ~50-100 Exabyte range, during the HL LHC era from 2025
- The reliance on high performance networks will thus continue to grow as many Exabytes of data are distributed, processed and analyzed at hundreds of sites around the world.
- As the needs of other fields continue to grow, HEP will face increasingly stiff competition for the use of large but limited network resources.

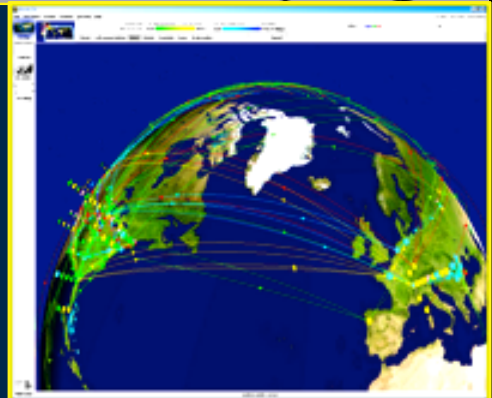
1 EB = 2 milligrams of DNA

Earth Observation



Entering a New Era of Technical Challenges as we Move to Exascale Data and Computing

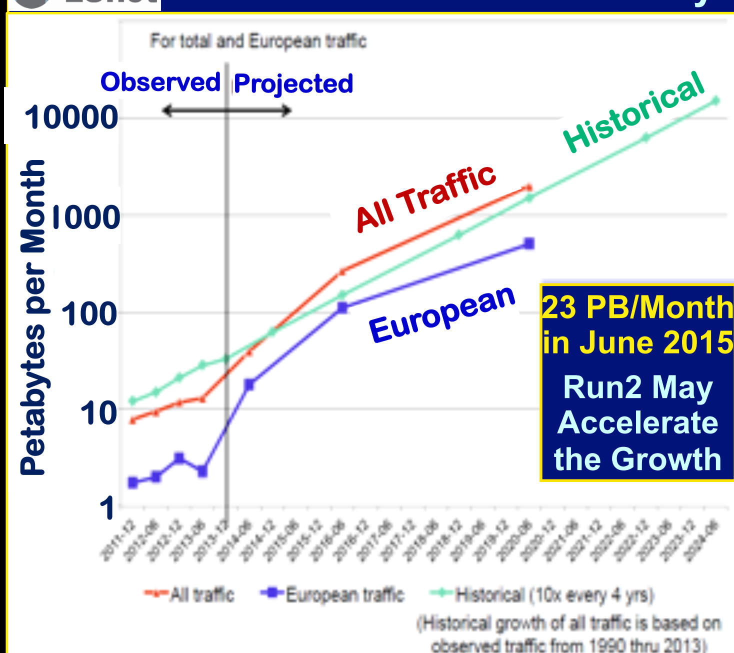
- Beyond network capacity and reliability alone, the keys to future success are next generation systems able to:
  - Respond agilely to peak and shifting workloads
  - Accommodate a more diverse set of computing systems from the Grid to the Cloud to HPC
  - Coordinate the use of globally distributed computing and storage, and networks that interlink them
    - In a manner compatible across fields sharing common networks
- The complexity of the data, and hence the needs for CPU power, will grow disproportionately: by a factor of several hundred during the same period



ESnet Science projection to 2024 Compared to historical traffic

E. Dart
W. Johnston

ESnet Total traffic handled in Petabytes per Month



23 PB/Month in June 2015
Run2 May Accelerate the Growth

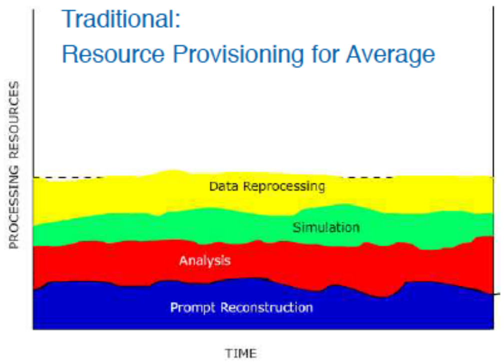
Projected Traffic Reaches 1 Exabyte Per Month. by ~2020
10 EB/Mo. by ~2024

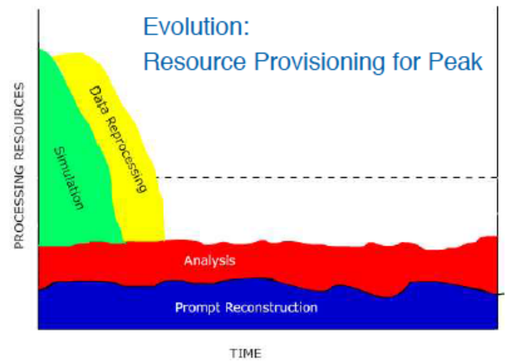
Rate of increase follows or exceeds Historical trend of 10X per 4 Years

HEP traffic will compete with BES, BER and ASCR

Processing – Evolving the GRID

- Experiments don’t need all the resources all the time
- Conference schedules, accelerator schedules, holidays all influence demand
- Resource Needs vary with time – and provisioning needs to adapt





Provisioning for Peak Demands

The dream of short turn-around times for workflows

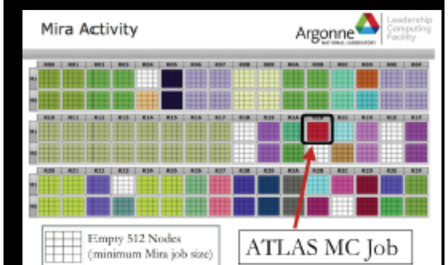
- Short latencies in particular in analysis workflows are important for science efficiency
- Using resources from a larger pool when they are needed, should also result in more cost-effective solutions

- Separating the processing and storage services allows them to scale independently
- e.g. ATLAS and CMS are looking at ways to double available resources for periods of time
  - Using Amazon services

Provisioning for peak requires that we use pooled resources
➡ Clouds or a large HPC Center!

High Performance Computing in HEP

- HTC: High Throughput Computing
  - Independent, sequential jobs that can be individually scheduled on many different computing resources across multiple administrative boundaries(*)
- HPC: High Performance Computing
  - Tightly coupled parallel jobs, must execute within a particular site with low-latency interconnects(*)
- Long history in HEP in using HPC installations
  - Lattice QCD and Accelerator Modeling exploit the low latency interconnects successfully for a long time
- Community effort: enable traditional HEP framework applications to run on HPC installations
  - Example: Mira at Argonne (PowerPC, ~49k nodes each 16 cores, almost 800k cores)
  - Generating Atlas LHC Events with Algren



Exascale Ecosystems
for Next-Generation Data Intensive

- The opportunity for HEP (CMS example):
  - CPU needs will grow 65 to 200X by HL LHC
  - Dedicated CPU that can be afforded will be an order of magnitude less; even after code improvements on the present trajectory
- DOE ASCR/HEP Exascale Workshop:
  - Identified key opportunities for harnessing the special capabilities of ECFs
  - Exposed the favorable outlook and issues for HEP to take this key step + meet the needs
  - Highlighted the *Network Dimension*
- Important added benefits to HEP + ASCR, the facilities, programs and the nation
  - Shaping the future architecture and operational modes of ECFs
  - Folding LCFs into a global ecosystem for data intensive science
  - Developing a "modern coding workforce"
  - Enabling many fields to "think out of the box"

Leadership

A favorable HEP platform:
- LHC experiments are gearing their S&C operations for more flexible use of diverse resources: Grid, Cloud, HPC

Vision: Next Gen Integrated Systems for Exascale Science: Synergy ➡ a Major Opportunity

Exploit the Synergy among:

1. Global operations data and workflow management systems developed by HEP programs, being geared to work with *increasingly diverse and elastic resources* to respond to peak demands
   - Enabled by distributed operations and security infrastructures
   - Riding on high capacity (but mostly still-passive) networks

2. Deeply programmable, agile software-defined networks (SDN) Emerging as multi-domain network "operating systems"
   - \+ New network paradigms focusing on content: from CDN to NDN
3. Machine Learning, modeling and simulation, and game theory methods Extract key variables; optimize; move to real-time self-optimizing workflows

✸ The Watershed: A new ecosystem with LCFs as focal points in the global workflow; meeting otherwise daunting CPU needs

Key Developments from HPC Facility Side Enabling the Vision: LCF Architecture

❐ Developing appropriate system architectur in hardware + software that meet the needs

- ✸ Edge clusters with petabyte caches
  - ✸ Input + output pools: ~10 to 100 Pbytes
- ✸ A handful of proxies at the edge
  - ✸ To manage and focus security efforts
- ✸ Identifying + matching HEP units of work to specific sub-facilities adapted to the task
- ✸ Extending Science DMZ concepts
  - ✸ Enabling 100G to Tbps SDNs with Edge/W Coordination + DTN Autoconfiguration
- ✸ Site-Network End-to-End Orchestration
  - ✸ Efficient, smooth petabyte flows

Leadership

Next Gen Science DMZ

➡ Dynamic agile systems that learn to adapt to peaking workloads

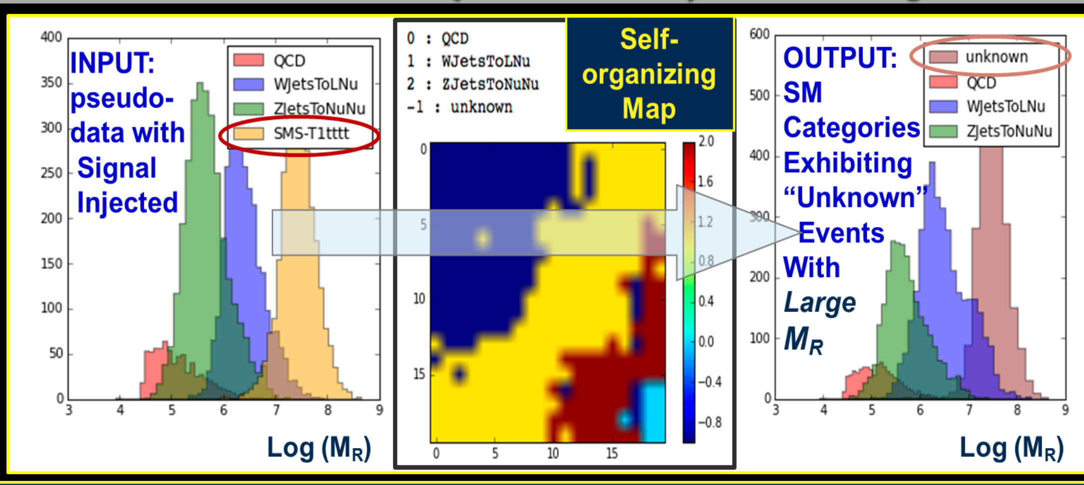
Data Intensive Exascale Facilities for Science Deeper Implications

- Bringing these facilities into the ecosystem of globally distributed
  information and knowledge sources and sinks
  - The hallmark of science, research and everyday life this century
- Will open new avenues of thought and new modes of the
  pursuit of knowledge in the most data intensive fields
  - By responding to petascale inquiries
    on human time scales, irrespective of location
  - Bringing our major networks, once again, into sharp focus
- This will broaden the function and architecture of ECFs
  and ultimately shape them in future generations
  - While also shaping the leading edge of
    "modern computing and networking"
- And place the US science community in a new position of leadership
  Being the first to cross this conceptual threshold

BACKUPS

Machine Learning: Exploring New Methods

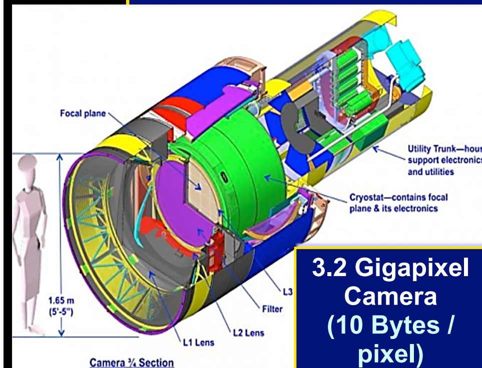
Aim to extend CMS' (and HEP's) Discovery Reach



Targets: Analysis - Identification/discovery of *unknown BSM signals;*
Optimization of LHC workflow and distributed system operations

- Synergy with previous Computing Model work on optimization of global grid and network systems using Self-organizing Neural Nets in MONARC

LSST + SKA Data Movement
Upcoming *Real-time* Challenges for Astronomy



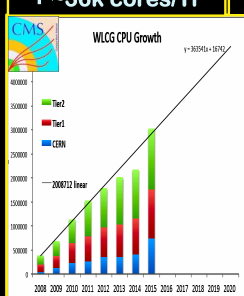
3.2 Gigapixel Camera (10 Bytes / pixel)

SKA

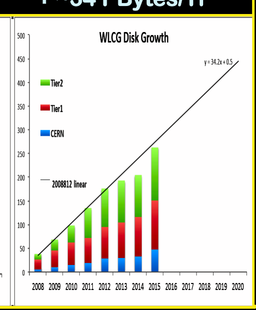
- ❒ **Planned Networks: Dedicated 100G for image data, Second 100G for other traffic, and 40G for a diverse path**
- ❒ **Lossless compressed Image size = 2.7GB (~5 images transferred in parallel over a 100 Gbps link)**
  - ❒ **Custom transfer protocols for images (UDP Based)**
- ❒ **Real-time Challenge: delivery in seconds to catch cosmic "events"**
- ❒ **+ SKA in Future: 3000 Antennae covering > 1 Million km2;** ***15,000 Terabits/sec to the correlators➡ 1.5 Exabytes/yr Stored***

CMS Offline Computing Requirements
HL LHC versus Run2 and Run1 [*]

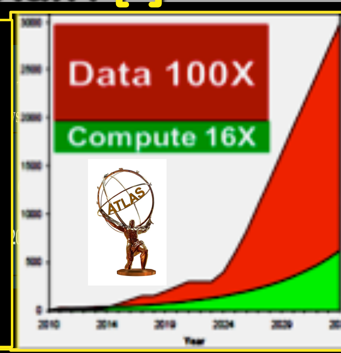
+ ~36k cores/Yr



+ ~34 PBytes/Yr



- Ratios in Computing and Storage for Run 2/Run1 are ≈ 2X.
- Hence HL-LHC to Run1 CPU: 130X to 400X



CPU Requirements Projections
- Projected CPU Needs: HL LHC/Run2 = 65 to 200X
- Anticipated increase in CPU resources at fixed cost/year: 8X
- Anticipated code efficiency improvements: 2X
- Projected shortfall at HL LHC 4X to 12X

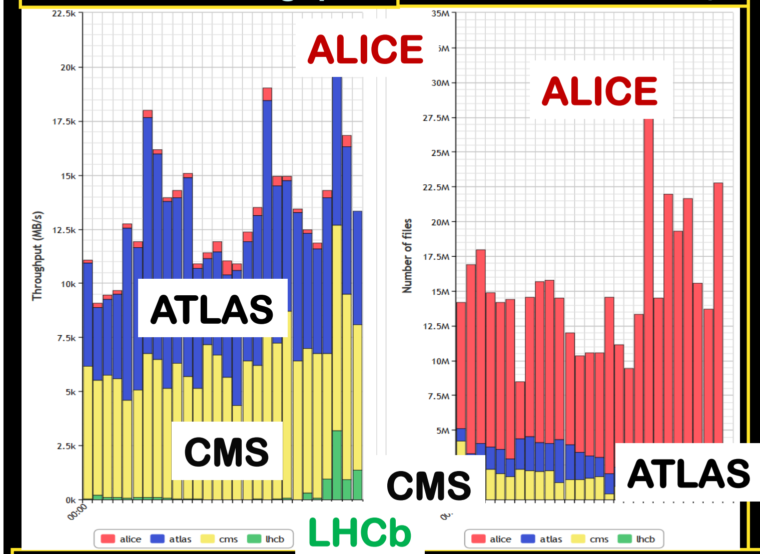
Storage Requirements Projections
- Projected Events: HL LHC / Run2 = 5 to 7.5X
- Event Size: HL LHC / Run2 = 4 to 6X
- Anticipated growth in Storage HL-LHC / Run2: 20-45X
- Projected shortfall at HL LHC 3X or More

[*] CMS Phase2 Technical Proposal: https://cds.cern.ch/record/202088

Complex Workflow: the Flow Patterns Have Increased in Scale and Complexity, even at the start of LHC Run2

WLCG: 170 Centers in 40 Countries. 2 Million Jobs Per Day

Transfer Throughput | Transfers Done/Day



13 GBytes/s Typical

25 GBytes/s Peak Transfer Rates

Complex Workflow

- Multi-TByte Dataset Transfers
- Transfers of Millions of Output Files Daily
- Access to Tens of Millions of Object Collections/Day
- >100k of remote connections (e.g. AAA) simultaneously

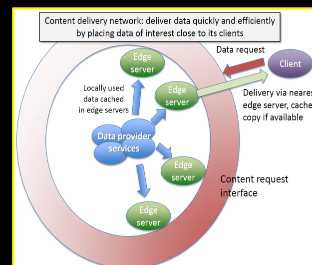
WLCG Dashboard Snapshot May-June. Patterns Vary by Experiment

Convergence and Collaboration: Tackling the Big Issues for LHC Run2

- Short-term Milestones: Enabling more efficient, manageable workflow by integrating advanced networking into CMS and ATLAS' mainstream software and data systems, along with CPU and Storage
  - Developing network awareness, resource management + path control though SDN in CMS (PhEDEx) and ATLAS (PanDA):
- Collaborating in Developing Key Technologies
  - Production use of Terabyte to Petabyte Transfers with State of the Art High Throughput [CHOPIN]
  - Dynamic Circuits for guaranteed bandwidth to Tier2s and 3s in the US and Across the Atlantic: Integrated in LHCONE [ANSE]
  - Software Defined Networking: [OliMPS; OpenDaylight Controller]
  - Named Data Networking: A possible Future Internet paradigm

LCF-Edge Data Intensive Systems (LEDIS)
Operational Model

- In the context of a new HEP – LCF – ESnet partnership for Joint system and architecture development
- Data brought to LCF edge ~petabyte chunks: ~2 hrs at 1 Tbps
  - Far enough in advance: chunks ready and waiting in a buffer pool
- Using secure systems at the site perimeter: Security Efforts (human and AI) can be focused on a limited number of entities (proxies)
  - Keeping manpower + risk at acceptable levels
- Multiple chunks for different stages of the workflow
  - Each chunk's provenance + attributes identified
  - Examples: Input/Output Data size, memory, CPU to IO ratio; delivery deadline, authorization level
- Enables matching to appropriate HPC subsystems, to meet the needs while operating at high efficiency
- Conceptual Extension: Caching in the Network, or at nearby HEP Lab Sites; as in "Data Intensive" CDNs (NDNs)
  - Adapting to the future Internet architecture that may emerge



Follow on to a CDN Model
T. Wenaus BNL/ATLAS

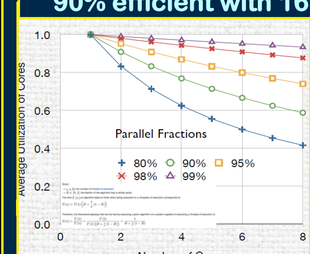
Key Developments on the HEP Side
Enabling the Vision: Coherent Parallel Architectures

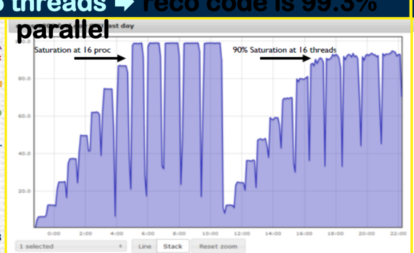
- We need to recast HEP's code and frameworks for the highly parallel, energy efficient architectures (GPU, Knights Landing, etc.) of modern HPC systems

  Significant progress in specific HEP areas exists
- CMS threaded memory-efficient concurrent framework for multicore CPUs

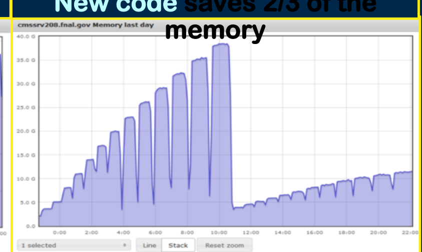
CMS Multithreaded Reconstruction Framework
E. Sexton-Kennedy at CHEP2014

90% efficient with 16 threads ➔ reco code is 99.3% parallel





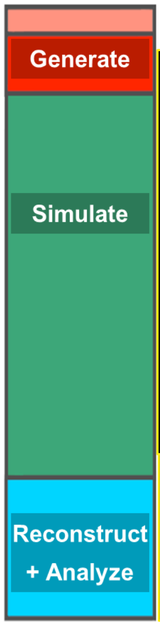
New code saves 2/3 of the memory



HEP Collider HPC Use, Prospects and Wishes

Tom Lecompte (Argonne) at the Exascale Workshop

❖ Computing to reach the Science Goals: Argonne LCF Use



To MIRA

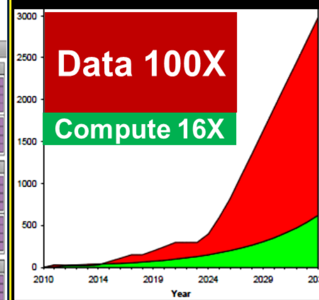
- Mira: 65M Hrs Compare the Grid: 1B Hours
- 2 FTEs
- Equal to the 7th largest "country" in CPU power in ATLAS in 2015
- Focus: Generators

- Simulation next
- Enabling "extra dimensions" in HEP Analysis

ALPGEN on MIRA

- 256k/768k Cores
- Code Improved 23X: 1 core went from 1/15 to 1.5X a Grid core
- 6-8X the ATLAS Grid CPU when running

Issues for the HL LHC



Meeting the CPU and data handlng needs

Adapting HEP codes + SW frameworks

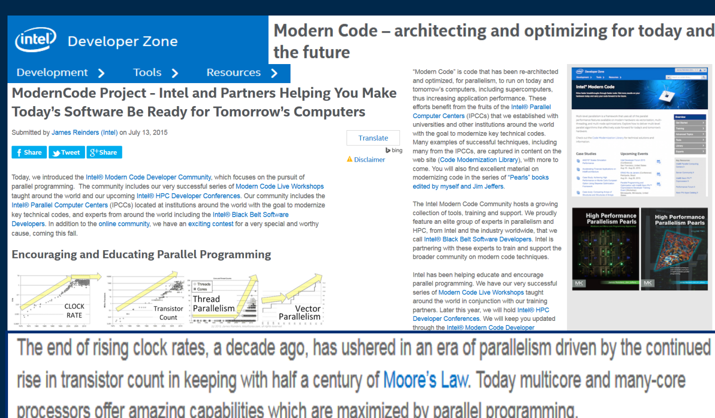
Beyond application software alone: A New Class of System

An excellent very promising start. A lot of work remains

Key Issue for HEP for this Vision

Establish the Workforce: New Traditions

- ❖ Establishing the expertise, and then the tradition of using the code architectures and algorithms well-adapted to HPC systems today
  - ❖ More generally, *To the computers of tomorrow* (viz. the Intel *"Modern Code Project"* launched on July 13)



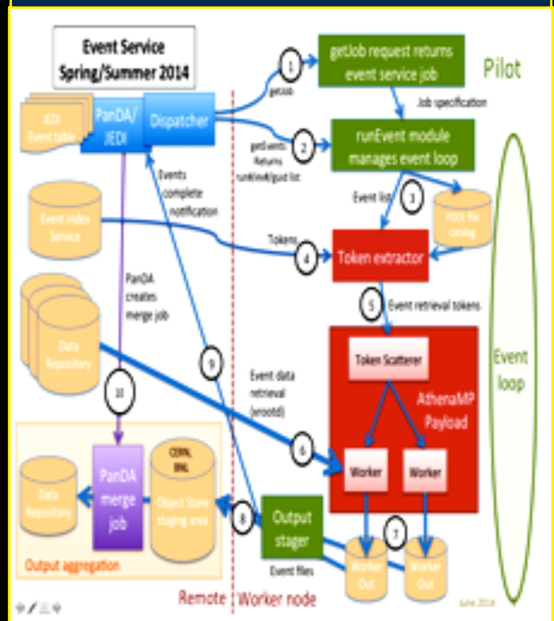
It is the Computing Per Joule Imperative That shapes our Future

- ❖ Developing and building adaptive global systems that effectively co-schedule: CPU, memory, storage, local + wide area networks
- ❖ Developing the Workforce: With *new* system + network expertise

Key Developments from the HEP Side
Enabling the Vision: ATLAS Granular Event Service

- ATLAS PanDA/Jedi Event Service:
  ➡Serve *event chunks,*
  - Smallest possible work units
  - Object store loaded on output,
  - ➡ *Granular, flexible workflow*
- Scaling Issues:
- Named Data Networking

ATLAS PanDA/JEDI Event Service



Data Intensive ECF Systems
and HEP: A Sense of Scale Circa 2016-28

- The current exercises on MIRA with generators (1 million threads) run at "6-8X the speed of the ATLAS Grid" [Lecompte]
- If HEP production codes circa 2019 all run efficiently on Aurora (17X Mira) they would occupy just a few percent of the CPU capacity
  - ➡ Perhaps ~10% including inefficiencies
- HL LHC projected needs circa 2028 could fully occupy Aurora, but by then systems ~30X Aurora or Summit are foreseen
  - Example: NERSC Upgrade Circa 2028 of "5-10 Exaflops" [S. Dosanjh]
  - HEP would thus remain at several percent of this one ECF; perhaps ~10% with inefficiencies; and a significant co-developer
- Also consider one strategic aim of such systems: To enable new and "larger" methods of exploration and discovery
  - Energy Frontier examples: Probing high dimensionality model spaces; Deeper, scaled up simulations; new intensive reconstruction methods

Conclusion: a compatible partner
and a source contributing solutions

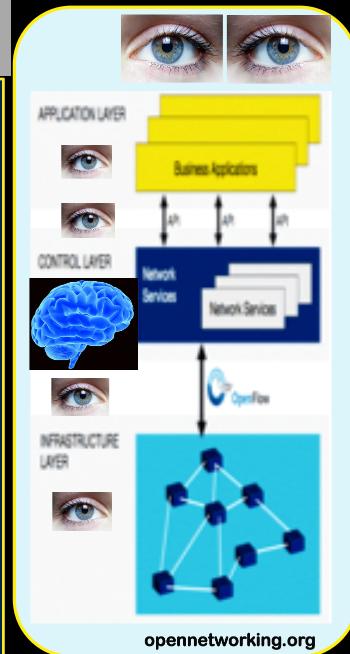
SDN in SDN-NGenIA and SENSE
Ideas Building on Caltech/Esnet/FNAL Experience

Vision: Distributed computing environments where resources can be deployed easily and flexibly to meet the demands of data-intensive science, giving transparent access to an integrated system of enormous computing power

SDN is a natural pathway to this vision: separating the functions that control the flow of network traffic, from the switching infrastructure that forwards the traffic itself through open deeply programmable "controllers".

With many benefits:

- ❒ Replacing stovepiped vendor HW/SW solutions by open platform-independent software services
- ❒ Imagining new methods and architectures
- ❒ Virtualizing services and networks: lowering cost and energy, with greater simplicity



A system with built in intelligence

Requires excellent monitoring at all levels

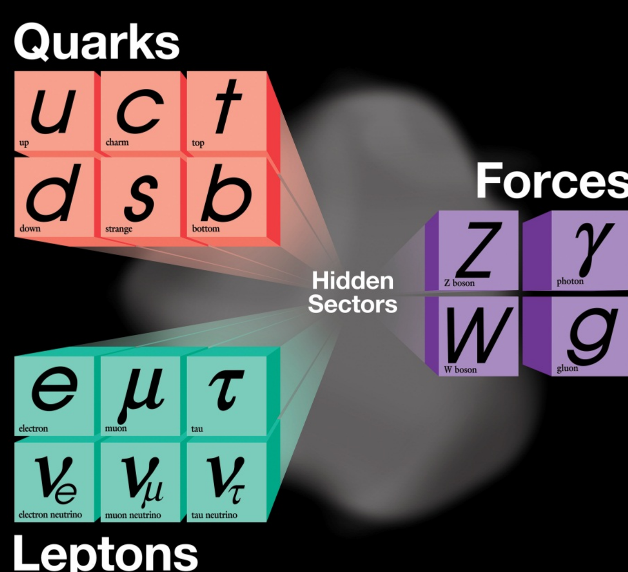
The Current State of Knowledge

The Standard Model

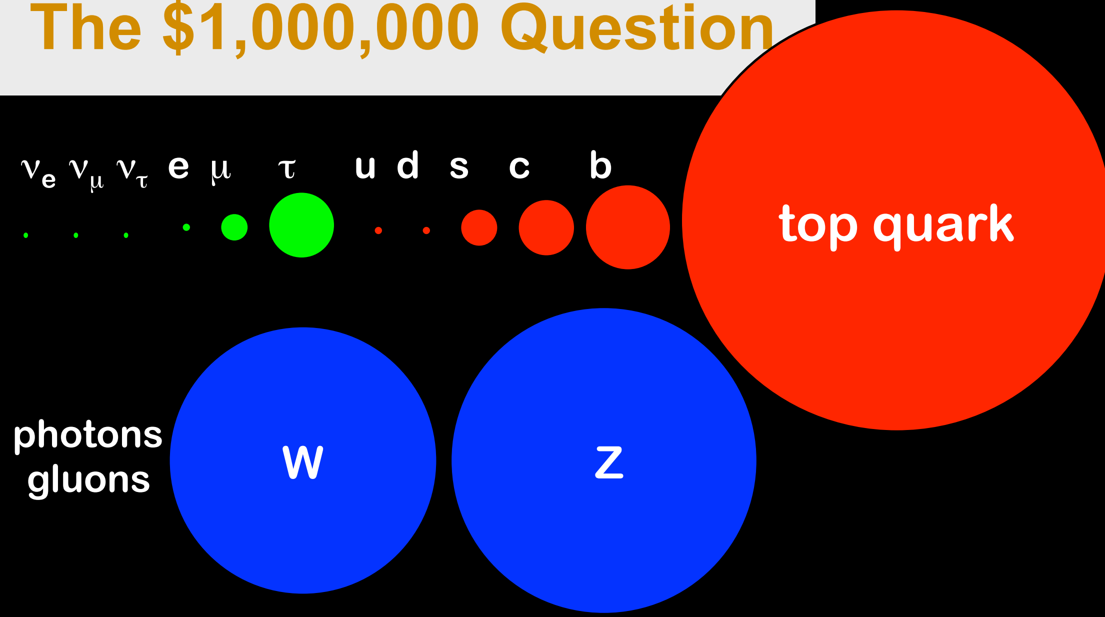
Greatest scientific achievement of 20th century

Every particle physics experiment ever done fits with this model

But it is incomplete…



The $1,000,000 Question



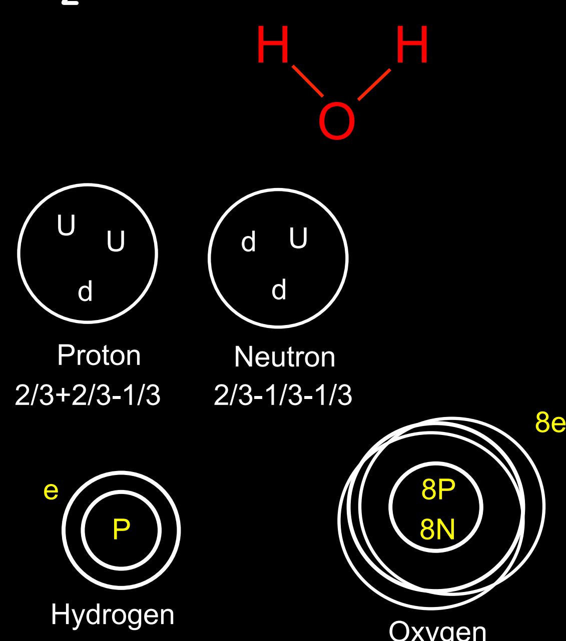
Why is the top quark so heavy?

Why are there so many particles?

Where does mass come from?

An example – Water!

- H_2O

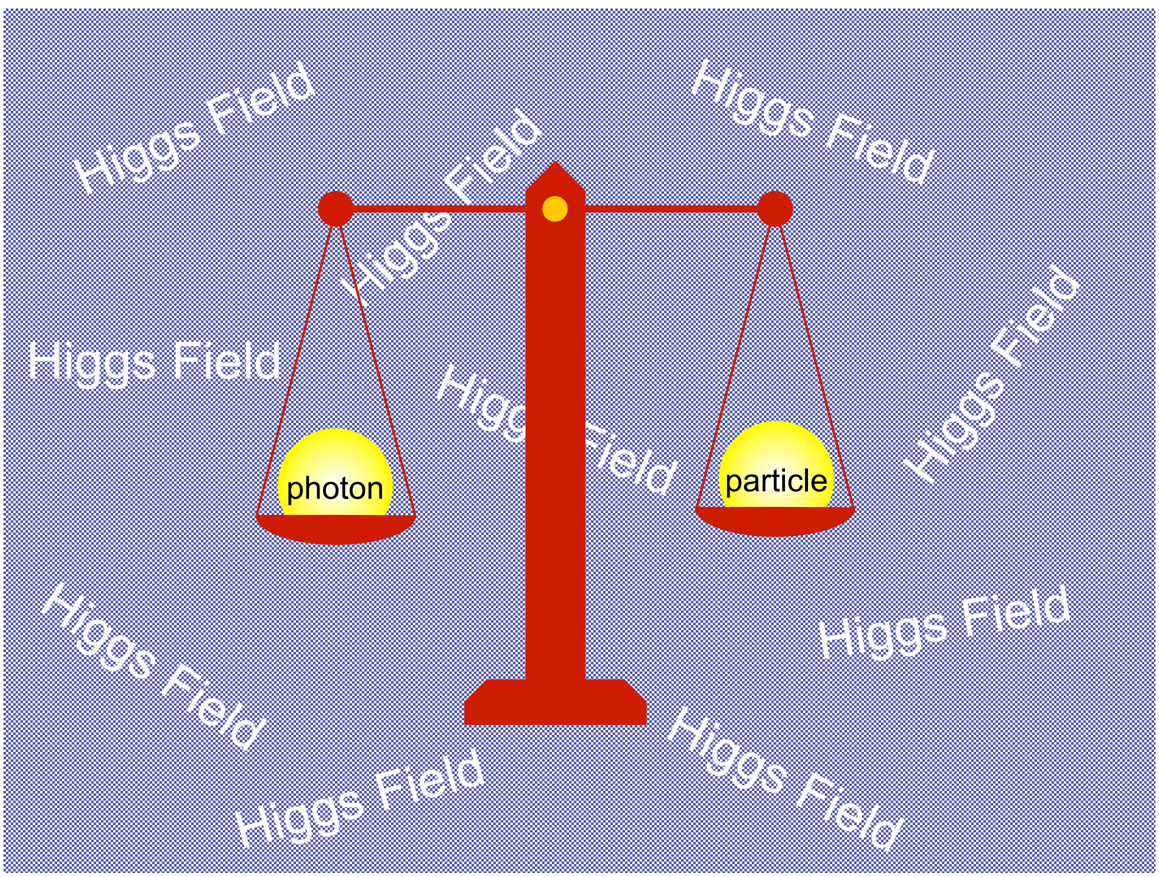


H_2O ➡ 28 Up Quarks
26 down quarks
10 electrons

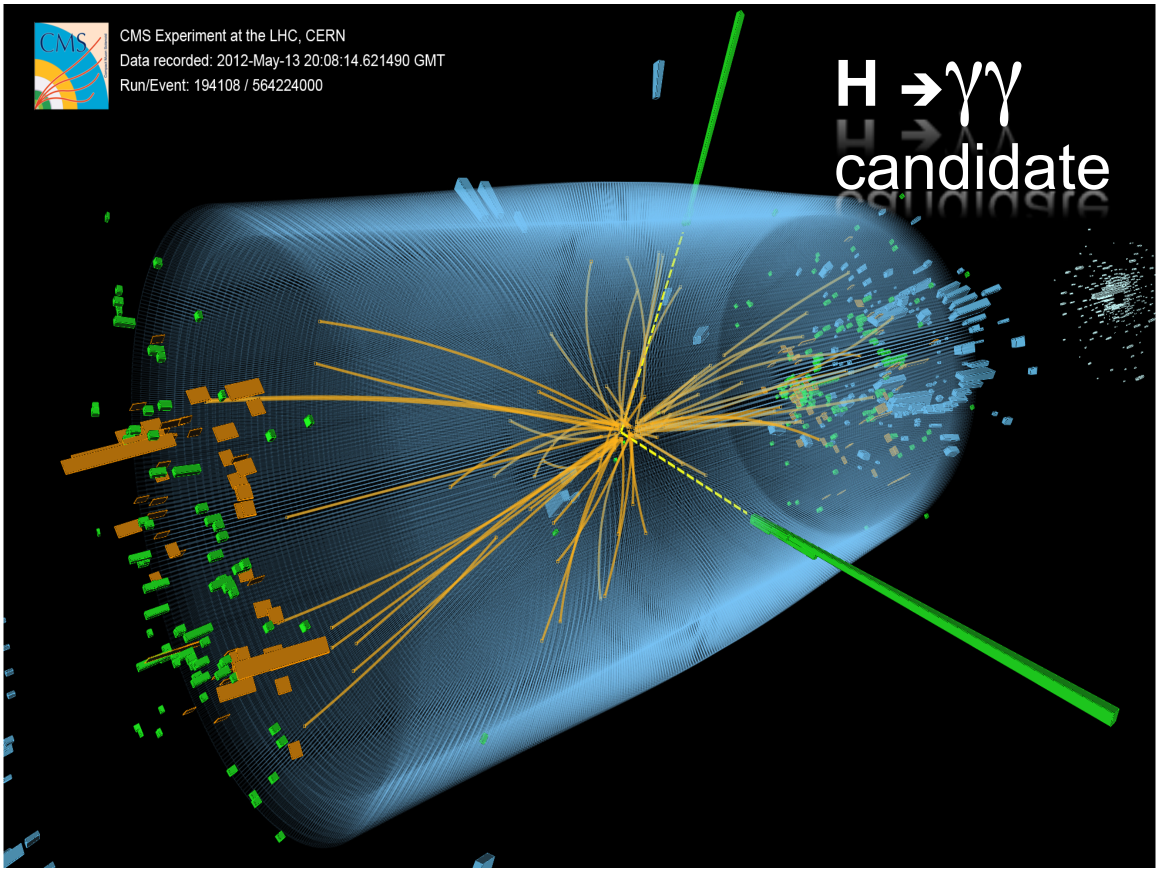
Enter the Higgs Mechanism

Popularity = Mass

Analogy by Prof. David Miller
University College of London



Consequence of the Higgs Mass



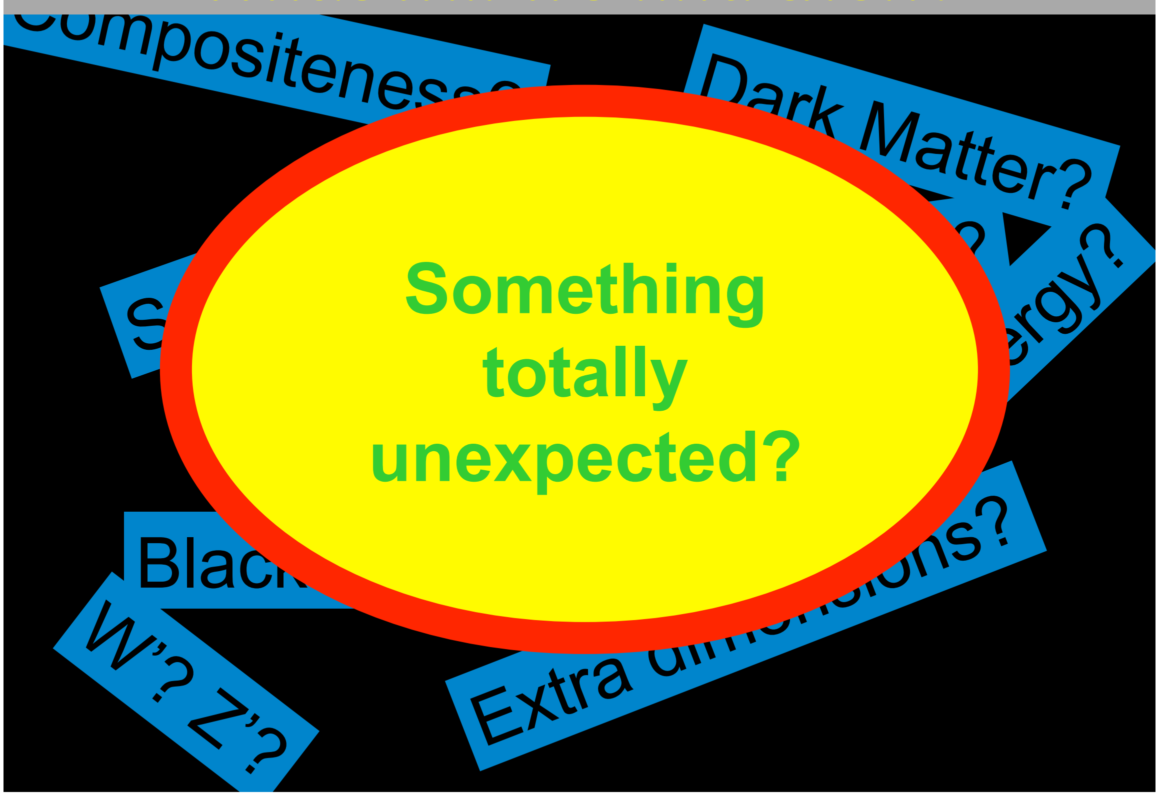
The Role of Simulation

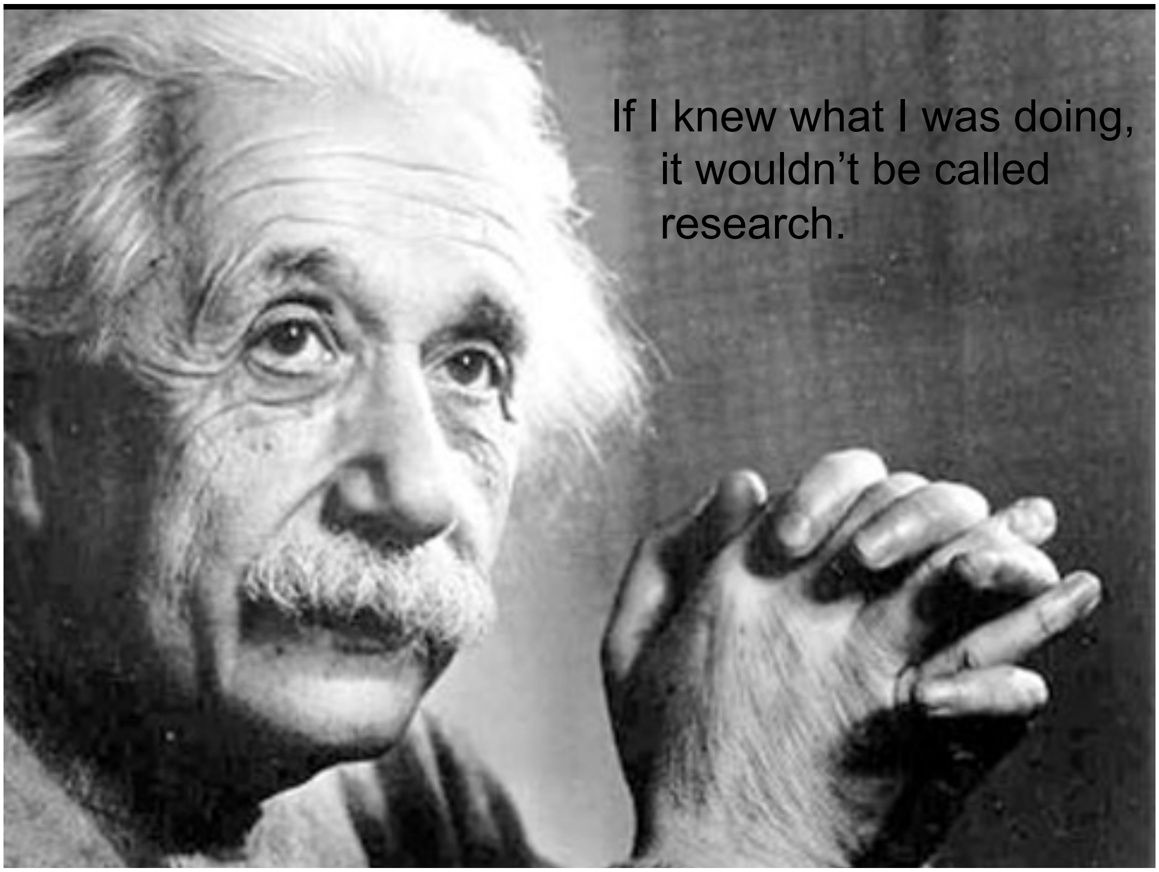
- Its ALL about Simulation
- Spend much more computing time on simulation than we do with the instrument data
- Everything we do involves simulation
  - Investigating whether a proposed experiment has any scientific merit
  - Designing the detector
  - Designing the analysis
  - Understanding what the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

A Snapshot today…

- Currently, CMS and ATLAS has each produce between 7-10 billion events over its lifetime.
- These events are produced WORLD-WIDE via the GRID and then cataloged and stored centrally
- An average size of a simulated event is ~500kb
- Full simulation of one event on one core today ~2-5 min depending on complexity of physics process
- Demand for simulation has peaks as the collaboration prepares for certain prestigious conferences

What will we find then?





Closing Remarks

- Finding the Higgs is not the end of the story – but the start of a new chapter
- Simulation is the cornerstone of particle physics – our ability to simulate well gives us the confidence to make discoveries with relatively relatively few events
- I believe the next decade will be very exciting as we uncover more of the mysteries of the world we live in

Why Does One Do High Energy Physics?

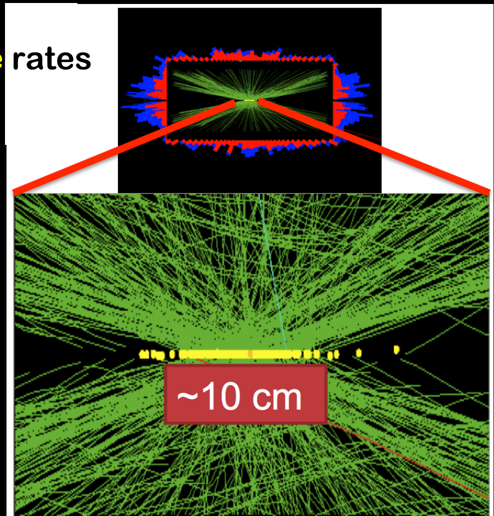
- The Experiments are large -- many operating experiments EACH have ~2000 physicists.
- Author list might take more pages in a journal than the actual article
- Physicists have to travel great distances to get to their experiments
- Startup times for experiments are large; already working on experiments 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

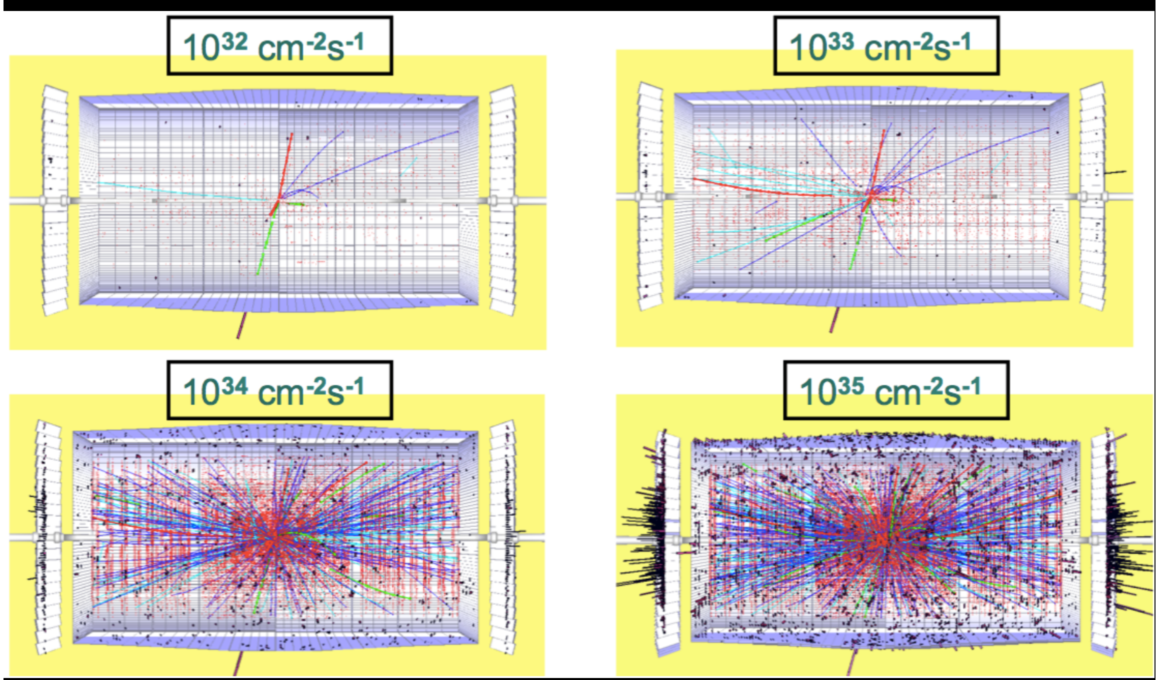
Unprecedented Computing Challenges: Trigger and Offline Reconstruction

- **Hadron colliders:**
  enormous cross section, unattainable rates
  - Event Data Complexity and Data Rate
  - Trigger needs to be very selective
  - Additional complications due to pile-up
-
  - 40 MHz →100kHz to HLT
    →O(100) Hz to offline
- **Physics Challenges:**
  - EW physics and Higgs
    - Soft Leptons 20-30 GeV
    - Several Hz rate
  - Natural SUSY
    - Jets, Leptons
    - Moderate Missing ET
  - QCD Background
    - Jet of 200 GeV – rate 1kHz
    - Jet fluctuations – lepton BG



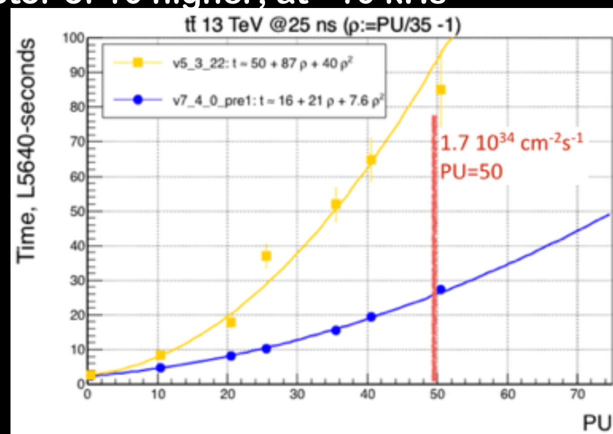
Shown: 78 collisions in one bunch crossing
Expect up to 200 in HL-LHC

Event Data Complexity and Data Rate



Software Improvements to Deal with Complexity

- Combinatorics of pattern recognition cause exponential explosion: for High-Luminosity LHC need a factor of 50…
  - extrapolate current event reconstruction performance to HL LHC
  - at pile-up of 140-200 factor 5 CPU time needed
  - also rate expected a factor of 10 higher, at ~10 kHs



- There is no way we can buy ourselves out of this problem with computing purchases
- We will have to revamp the computing model and improve the software

Industry to the Rescue?

- Processor and Architecture Trends
  - Moore's law gives us 64 times more transistors, but we're not using them
  - highly vector, heterogeneous, multi-core
- Network Trends
  - just finished 10->100Gbit transition, will there be another one
- Storage Trends
  - move to SSDs or NVRAM which could also change memory system architectures.
- Commercial Computing Trends
  - Cloud computing is becoming cheaper, will be able to handle our peak needs?