The S2I2-HEP Conceptualization Project and the High Luminosity LHC at CERN

Peter Elmer - Princeton University

1 May, 2017
2nd S2I2 HEP/CS Workshop

S2I2-HEP

The primary goal of the S2I2-HEP conceptualization project (`http://s2i2-hep.org`) is to produce a well-defined strategy for R&D for the software and computing models for use in high energy physics (HEP), in particular for the experiments collecting the very large data sets anticipated in the "High-Luminosity Large Hadron Collider" (HL-LHC) era of the 2020s.

Specifically the S2I2-HEP project will identify potential areas where U.S. university personnel can lead in key areas of software development to help realize the full potential of the HL-LHC program.

However HEP and the LHC are global projects, so no long-term planning exercise can exist in isolation, thus we are also pursuing a wider HEP community roadmap for software and computing in the 2020s.

LHC Grand Challenge Research Questions

The goal of HEP (and the LHC) is to understand the fundamental building blocks of nature, and their interactions. The potential of the LHC has been demonstrated in its first years with the discovery of the Higgs Boson.

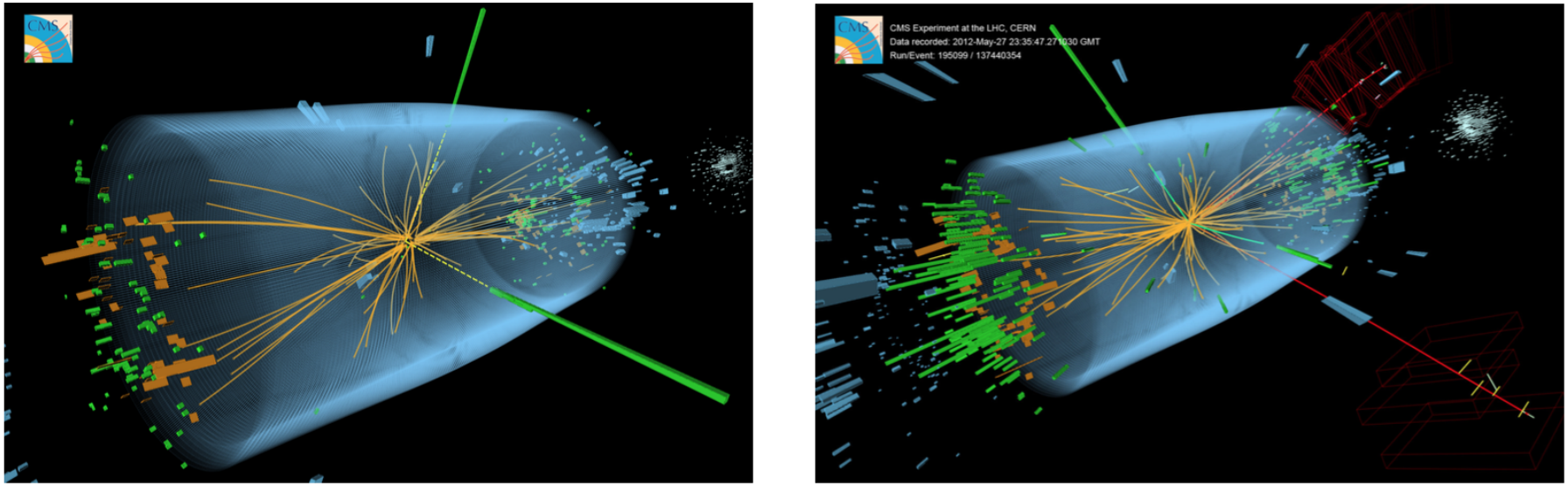
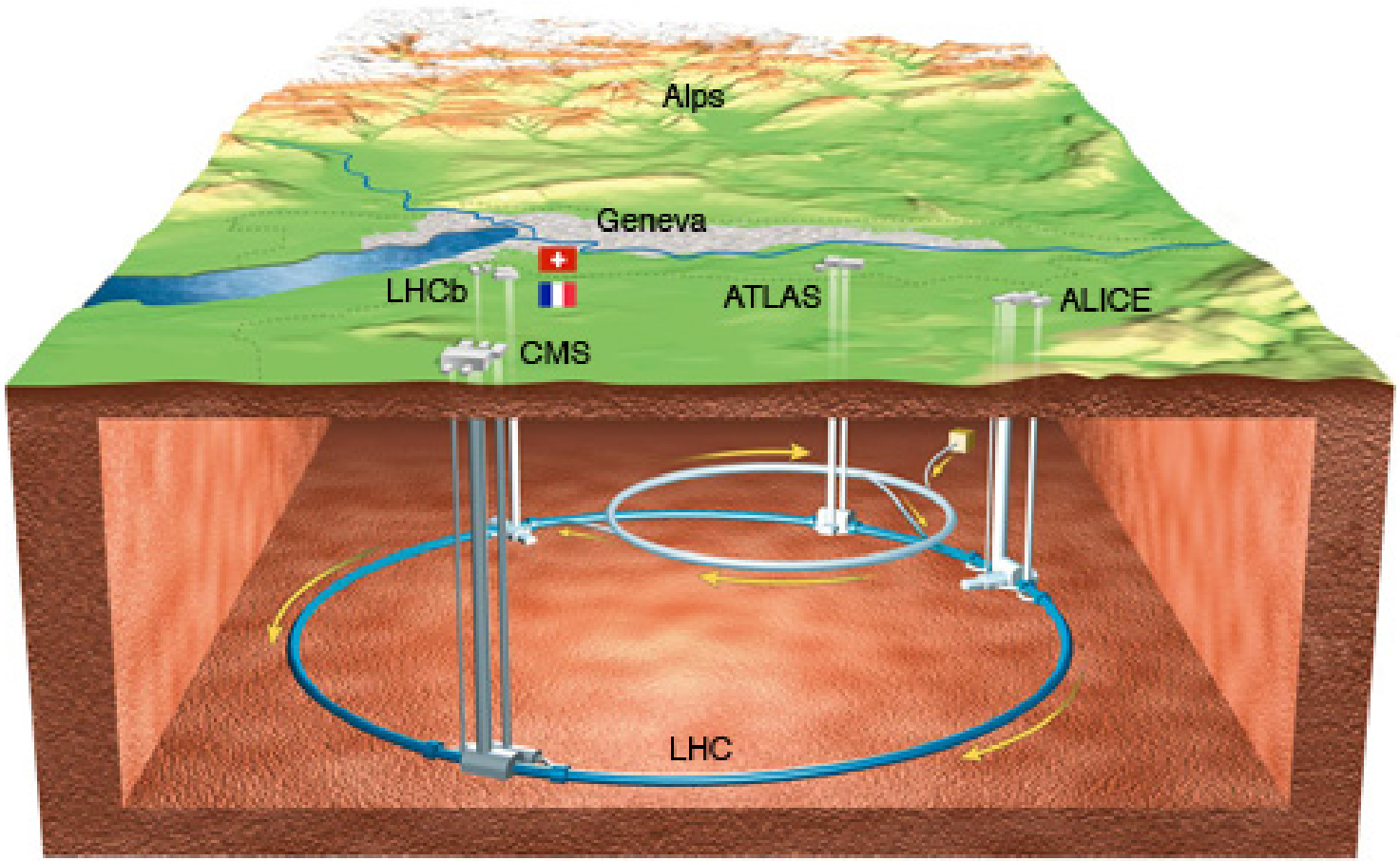


Figure 1.1: (a) Candidate for the decay Higgs $\rightarrow \gamma\gamma$, where the green lines are the two photons; and (b) Candidate for the decay Higgs $\rightarrow ZZ^*(ee\mu\mu)$, where here the green lines towards the center of the picture are the two electrons and the red lines in the center and at the upper right of the detector are the two muons.

LHC Grand Challenge Research Questions

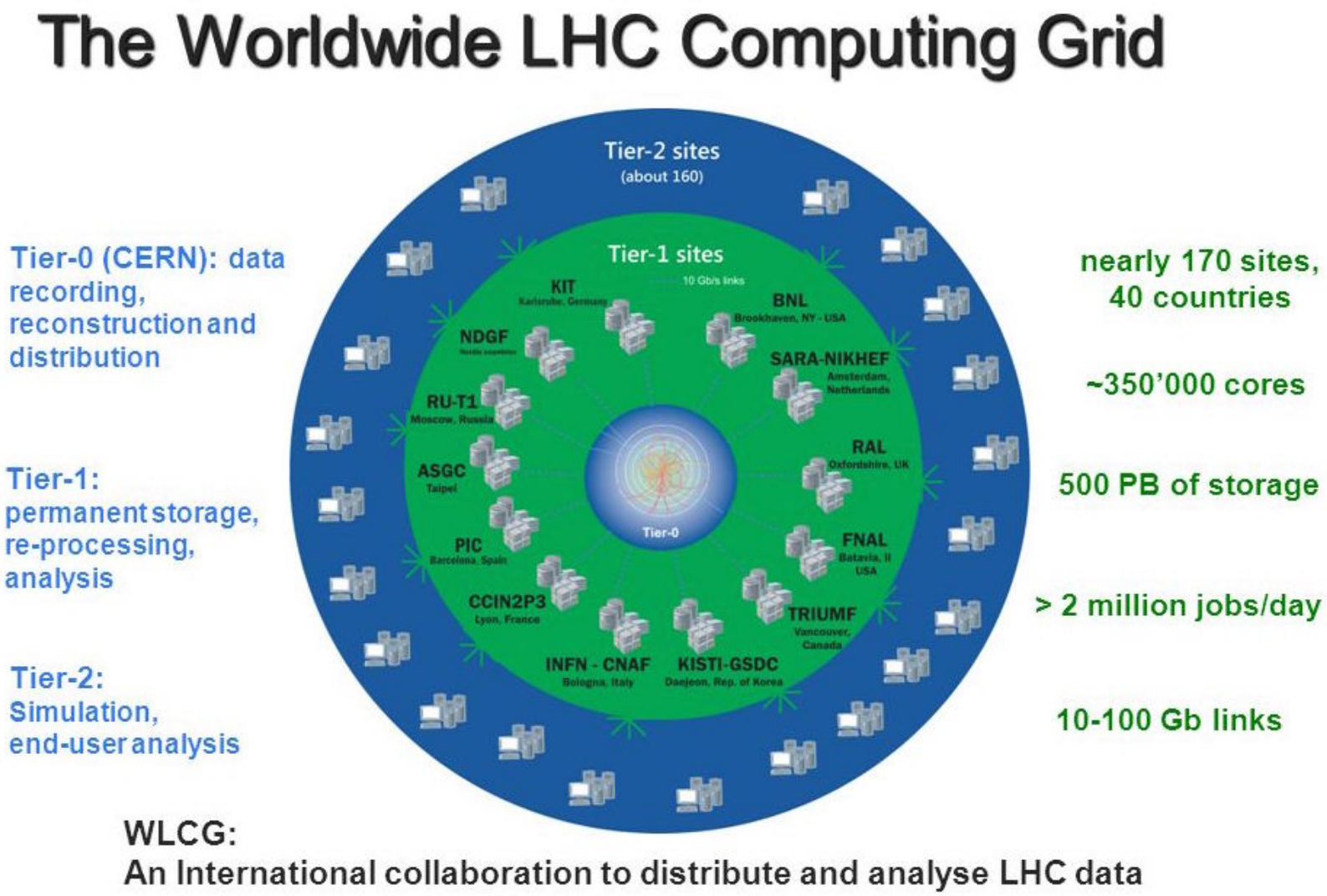
Many fundamental questions remain, however, including: Why does nature express the symmetries embodied in the SM, and not other equally elegant symmetries? Why are there (only) three generations of basic building blocks of matter? Why are the masses of these building blocks so different from each other, both within a generation and between generations? What is the dark matter which pervades the Universe? Why is matter so dominant over antimatter in the Universe? Does space-time have additional symmetries or extend beyond the 3+1 dimensions of which we know? What mechanism stabilizes the Higgs mass from large quantum corrections at high energy? Are neutrinos their own anti-particles? Can gravity and quantum mechanics be described in a consistent theoretical framework?

Large Hadron Collider (LHC) and Experiments



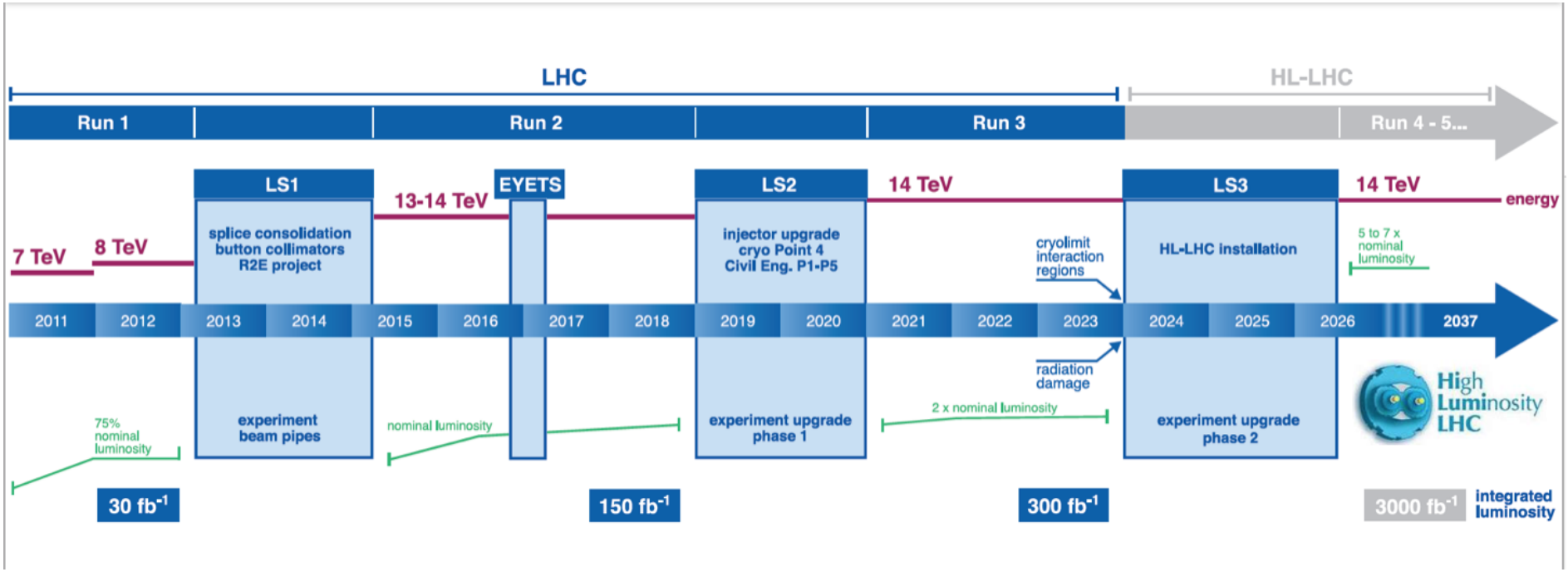
Two very large experiments (Atlas, CMS) with 3500+ people, and two large experiments (Alice, LHCb) with 500+ people

WLCG Distributed Computing System

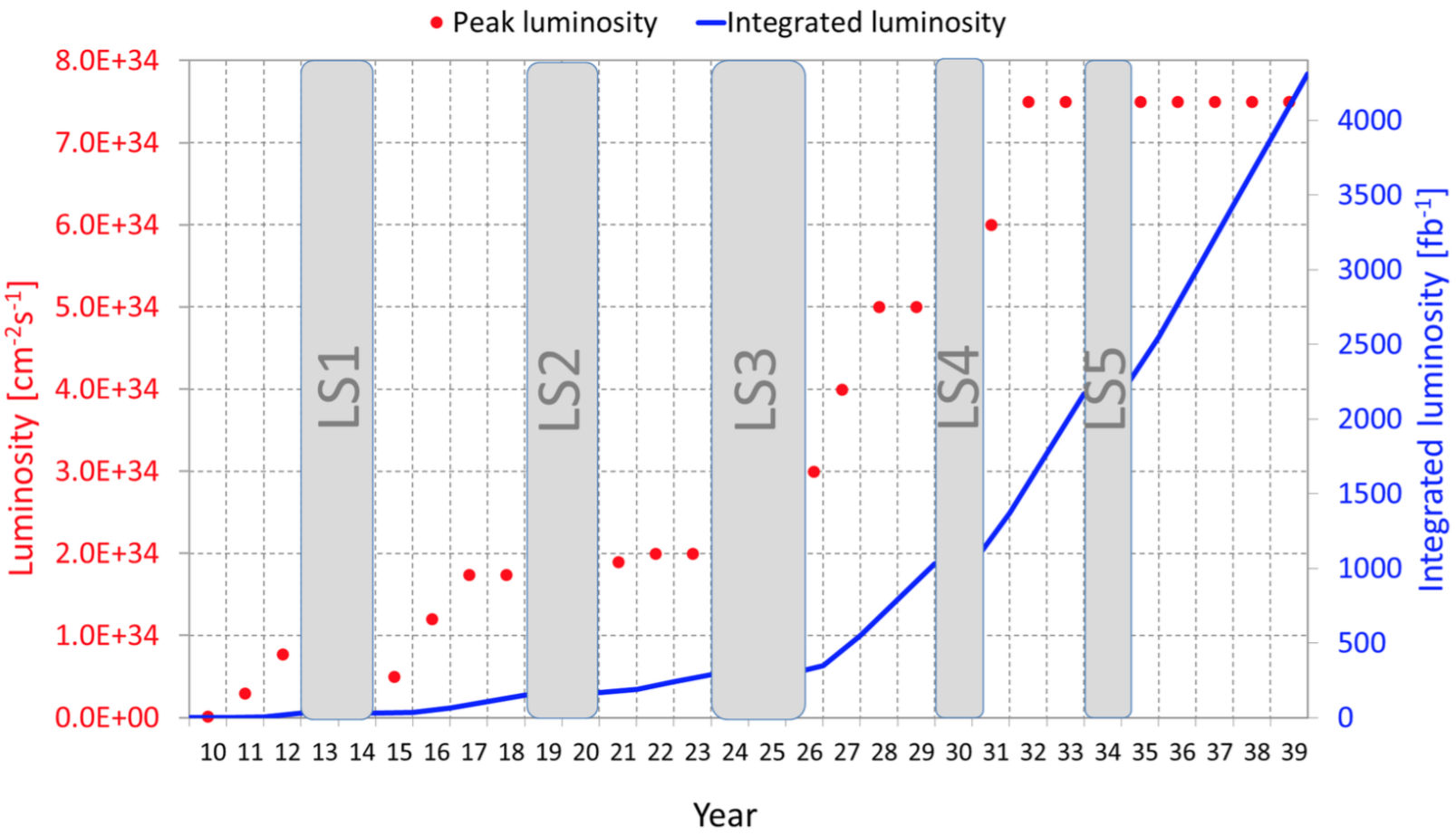


Slide from 2015

Plans for upgrading the LHC and Experiment Detectors



Plans for upgrading the LHC and Experiment Detectors

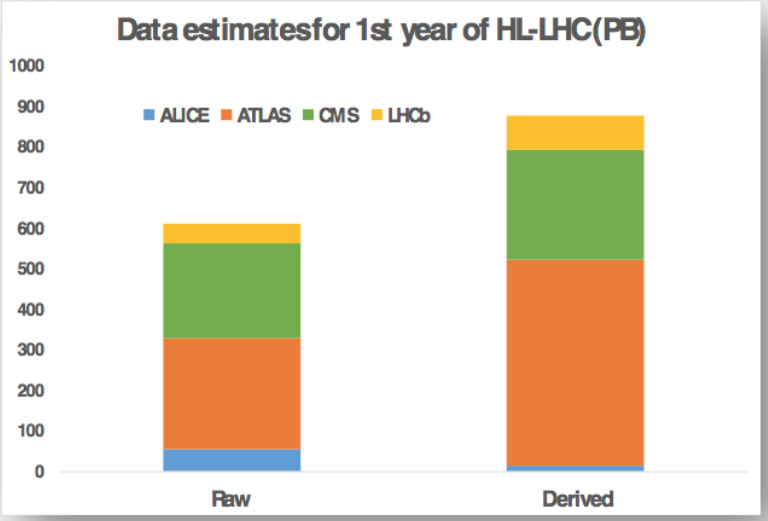


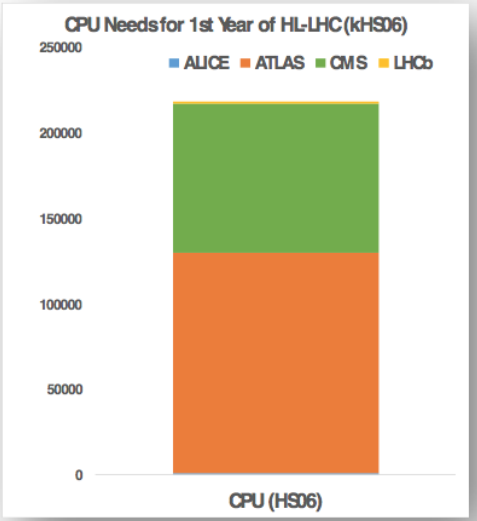
A Software "Upgrade" for HL-LHC and 2020s HEP?

Looking forward to the next 10 years, we see a number of challenges for HEP software and computing:

- **Performance/cost:** Estimates of computing needs run faster than Moore's Law by factors of o(10)
- **Technology/Market evolution:** the return of heterogeneity; technology change will also make it challenging to exploit Moore's Law without software evolution.
- **Scale:** The HL-LHC will integrate 100 times the current data, with significantly increased data (pileup) and detector complexity.
- **Sustainability:** Most of the current software, which defines our capabilities, was designed 15-20 years ago: there are many software sustainability challenges.

Estimates of Resource Needs for HL-LHC (WLCG)





Data:
- Raw 2016: 50 PB → 2027: 600 PB
- Derived (1 copy): 2016: 80 PB → 2027: 900 PB

CPU:
- x60 from 2016

Technology at ~20%/year will bring x6-10 in 10-11 years

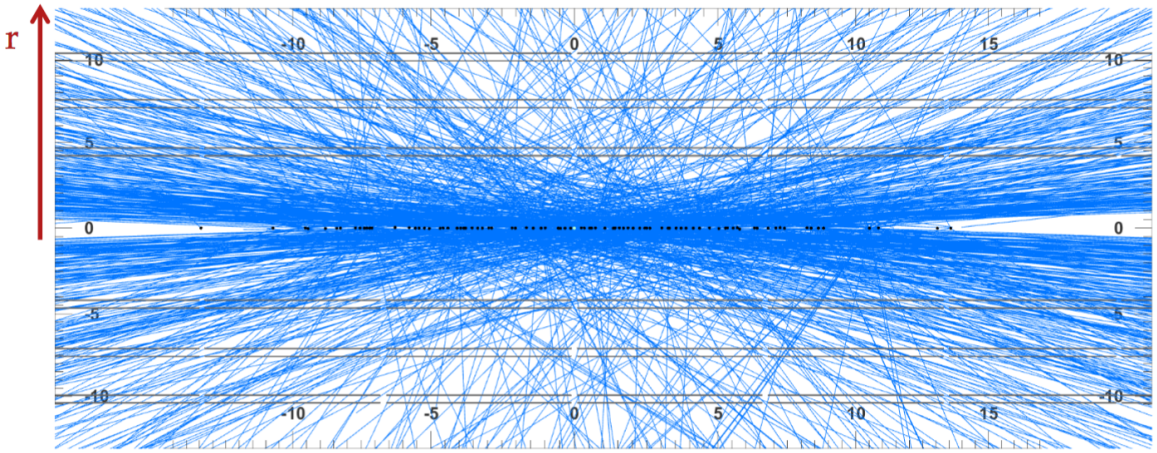
- Simple model based on today's computing models, but with expected HL-LHC operating parameters (pile-up, trigger rates, etc.)
- At least x10 above what is realistic to expect from technology with reasonably constant cost

(Slide from WLCG Workshop Intro, Ian Bird, 8 Oct, 2016)

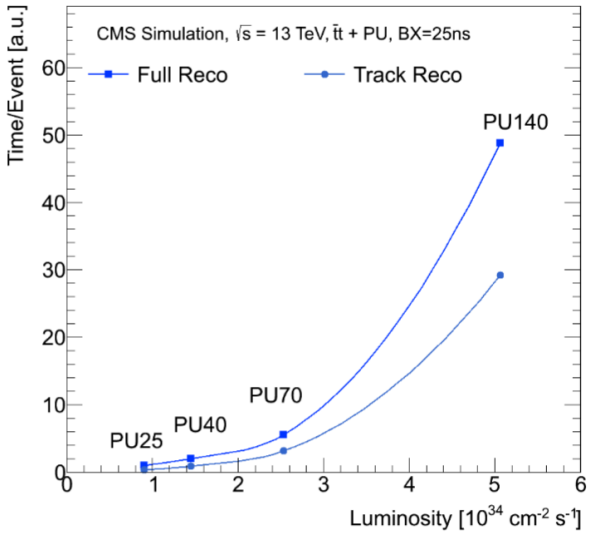
CPU and bulk processing capacity

The vast majority of the CPU needs at the HL-LHC come from pattern recognition algorithms ("reconstruction" for HEP). This probably a factor of 10 short, assuming we get and can exploit 20%/year Moore's Law gains.

The Challenge of the HL-LHC

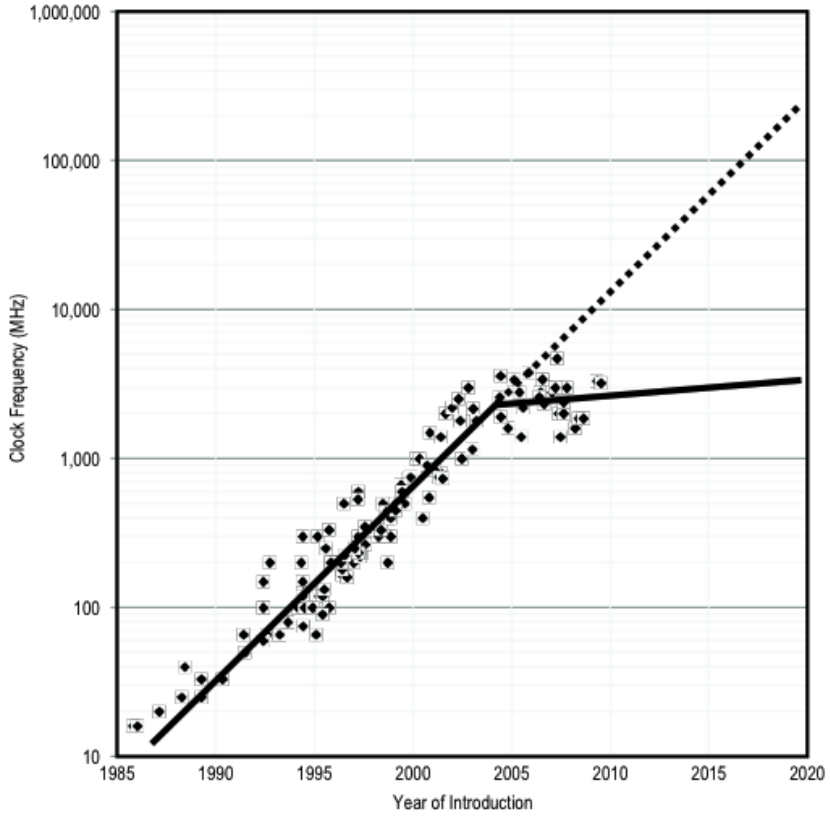


Simulation of pile-up = 140
at CMS in r-z plane



- HL-LHC poses serious challenges to reconstruction, and tracking will be hit the hardest
- Combinatorics from track finding will be a nightmare for efficiency, fake rate, and time

Processor evolution and software impact



Clock Frequency vs Time

- Single core performance has stalled, leading to multi/manycore and specialization
- To even realize Moore's Law gains, we are pushed towards parallelization of algorithms and design for performance.
- The software designs and implementations themselves need to evolve, not just be recompiled

Back to heterogeneous systems?

Building the worldwide distributed LHC computing grid was largely made possible by the convergence on Linux on (commodity) Intel x86 processors around the year 2000. Building the WLCG at this scale in the heterogeneous workstation era would have been quite difficult. For better or for worse, heterogeneity is returning:

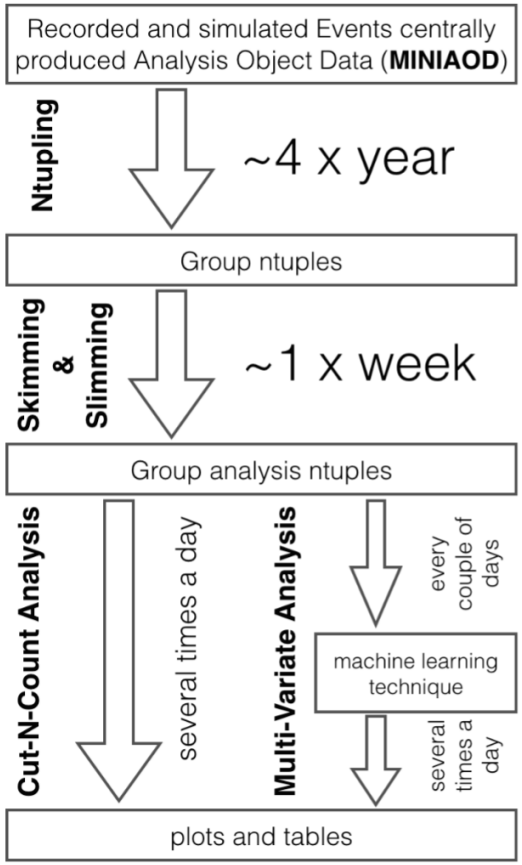
- Diversity of computing processor architectures (general purpose cores vs specialized processors)
- Owned vs commercial/cloud providers
- Some pressure to use systems traditionally designed for other types of applications (e.g. HPC/supercomputer as opposed to HTC/high-throughput systems)
- Possible further commoditizing market pressures (e.g. mobile)

Data Access, Organization and Management

Similarly managing and accessing a factor of 10 increase in data volume/year is a challenge.

This can also be exacerbated by any evolution required to use processing capacity (Grid, Cloud, HPC center). The evolution of tape, networks and heterogeneous computing will likely a more dynamic model. Optimizations and trade-offs within that will be critical.

Analysis Data Reduction - “Last Mile”



- Interactivity is the key to successful analysis: “Search for the needle in the haystack”
  - Select events, calculate new properties, train neutral nets, etc.
- Current Analysis Workflow
  - Touches only a subset of the total data volume, but subset varies from analysis to analysis
  - Complicated multi-step workflow because dataset is too large for interactive analysis
    - Slimming & Skimming, analysis dependent
    - Calculation of new quantities
    - Rerun framework code (b-tagging with non-default parameters, etc.)
    - Recipes on top of centrally produced samples to correct problems/mistakes
  - Can take weeks using GRID resources and local batch systems
  - Not all time spent is actual CPU, a lot of time is bookkeeping, resubmission of failed jobs, etc.
- Input:
  - Centrally produced output of reconstruction software, reduced content optimized for analysis
- Ntupling:
  - Convert into format suited for interactive analysis (still too big for interactive analysis)
- Skimming & Slimming:
  - Reduce number of events and information content
- Question: Will this scale for HL-LHC?

Why Software? Software is *the* Cyberinfrastructure

Computer hardware is a consumable.
Software is what we keep, and invest in, over time.

HEP Software Ecosystem

indico

TMVA

ROOT
Data Analysis Framework

IgProf

FroNTier

FairRoot

Roo Fit

XRootD

FastJet

INSPIRE HEP

RooStats

EvtGen

THE GAUDI PROJECT

CernVM File system

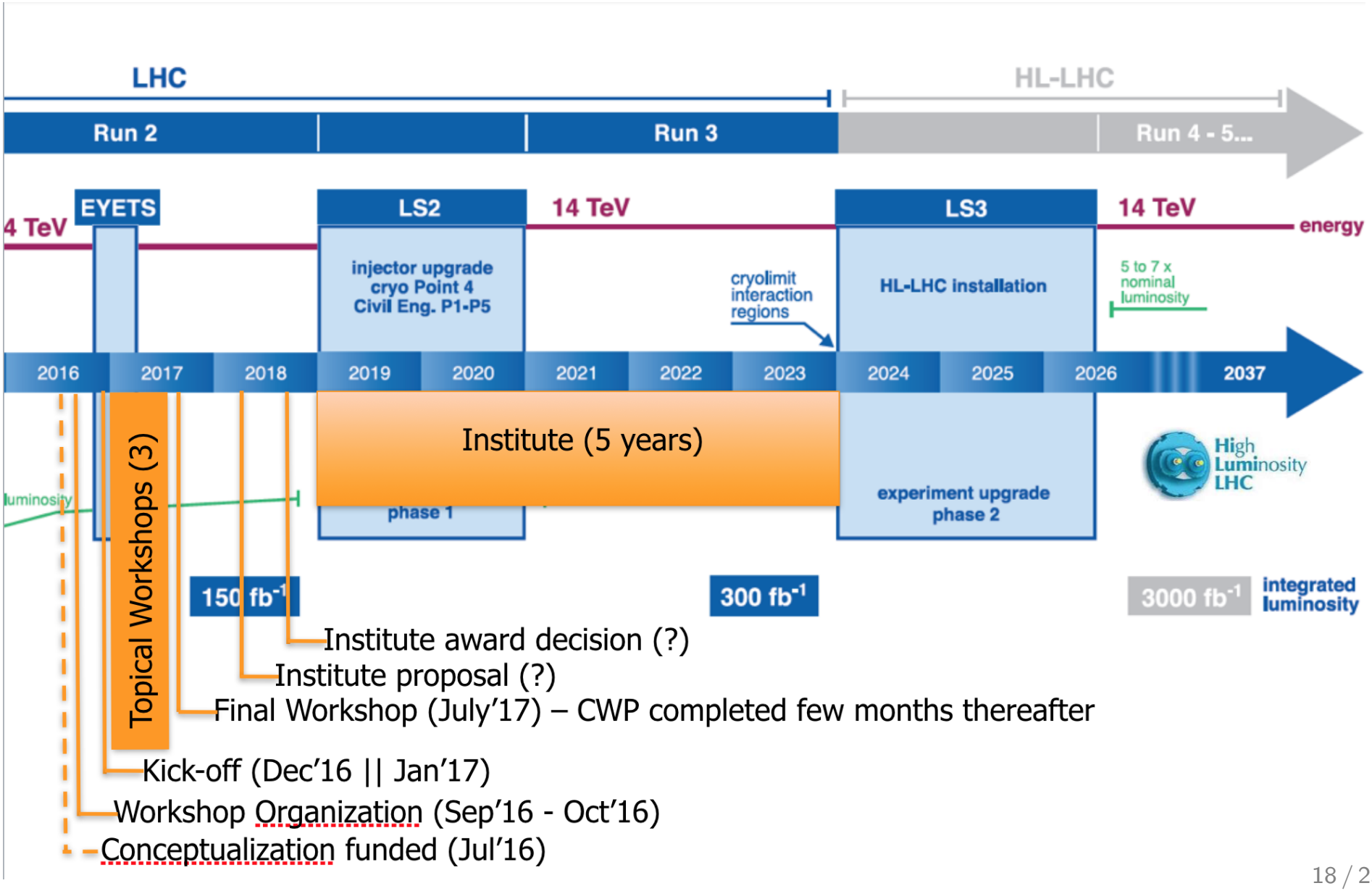
Geant 4

Examples, definitely incomplete!

dCache.org

Plus 15-20M Source Lines of Code (SLOC) of "experiment specific" codes, as well as dependencies on non-HEP scientific software.

S2I2-HEP (Success-oriented) timeline



Defining Longer-term Strategy

- HL-LHC computing requires a major 'software upgrade' and an eventual S2I2 institute for HEP would be a major player in that task
- Planning for such an "upgrade" cannot be done for the US (Universities) in isolation
- Thus we are intiating a larger community process to produce a Community White Paper (CWP) with an overall consensus strategy and roadmap for software and computing in HEP
  - Initiated as WLCG charge to the LHC experiments and HSF as a step towards the LHC experiment TDRs in advance of HL-LHC
  - The scope should not be restricted only to HL-LHC
  - Some early software components could be built, tested and used by experiments in LHC Run3
- Organised by the HEP Software Foundation (HSF) [next slide]
- Paper to be delivered by Summer 2017
- The S2I2-HEP Strategic Plan will be derived from this global plan

HEP Software Foundation (HSF)

The HSF (http://hepsoftwarefoundation.org) was created in early 2015 as a means for organizing our community to address the software challenges of future projects such as the HL-HLC. The HSF has the following objectives:

- Catalyze new common projects
- Promote commonality and collaboration in new developments to make the most of limited resources
- Provide a framework for attracting effort and support to S&C common projects (new resources!)
- Provide a structure to set priorities and goals for the work

Community White Paper (CWP)

- The CWP will identify and prioritise the software research and development investments required:
  - to achieve improvements in software efficiency, scalability and performance and to make use of the advances in CPU, storage and network technologies
  - to enable new approaches to computing and software that could radically extend the physics reach of the detectors
  - to ensure the long term sustainability of the software through the lifetime of the HL-LHC
- The HSF is engaging the HEP community to produce the CWP via a "community process"
  - Initiated as an HL-LHC planning process
  - Aiming for a broader participation (LHC, neutrino program, Belle II, linear collider so far)

Goals for this workshop

In the next talk Mark Neubauer will review the discussions we had in the 1st S2I2 HEP/CS workshop at NCSA/UIUC. In this workshop we would like to focus more on possible research collaborations in specific areas:

- Science Practices & Policies, Sociology and Community Issues
- Machine Learning and Algorithms
- Software Life Cycle / Software Engineering
- Data Management, Access and Organisation / Data Streaming
- Data Intensive Analysis Tools, Visualization
- Scalable Platforms
- Software/Data/Workflow Preservation & Reproducibility
- Training, Education, Professional Development, Advancement

Goals for this workshop

Although the general goal is to identify topics for possible HEP/CS research collaborations, we would like to call out some specific points in that direction:

- Are there places where the HEP language to describe a problem or system doesn't match how a CS person would describe the same problem?
- Which things does the HEP community want to do, does not know how, and believes that Computer Scientists may be able to figure out?
- How do these problems map to CS research questions?
- Are there CS workshops/meetings on these questions which HEP people should consider attending?