Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

The Future of Software and Computing for HEP

Pushing the Boundaries of the Possible

Elizabeth Sexton-Kennedy

ICHEP 2018, Coex Seoul

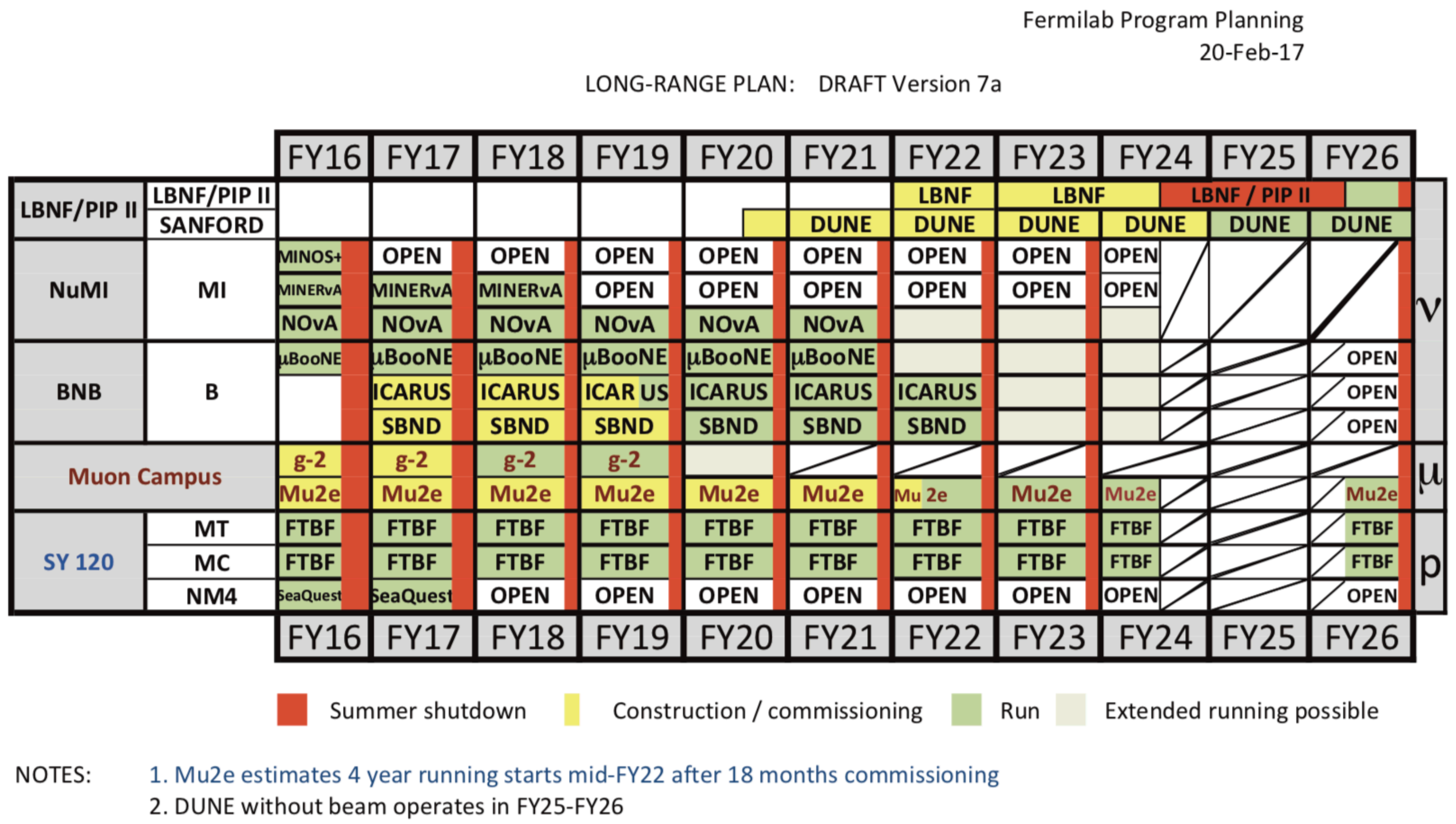
8 July 2018

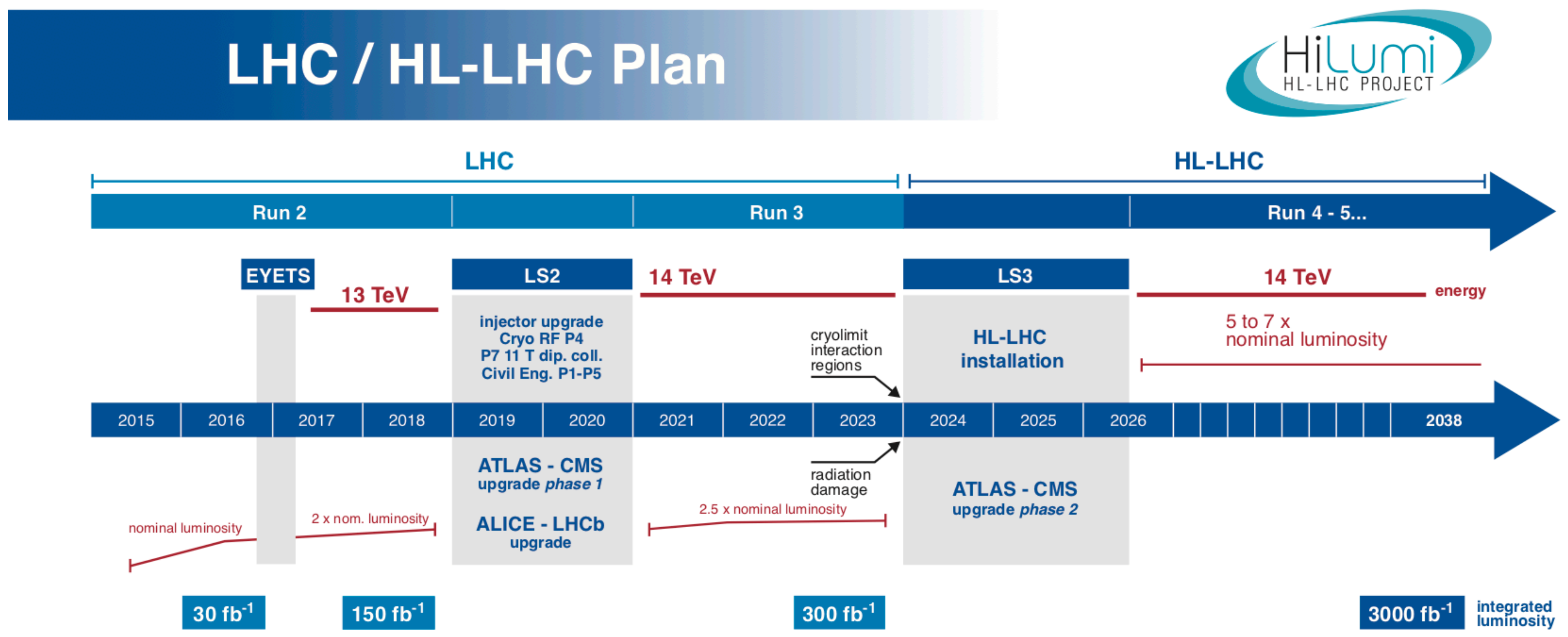
Outline

- Introduction
- A Data Centric Vision for the long-term future
- Community White Paper - Software and Computing tools
- The changing landscape of computing even quantum computing
- How much has been reflected in this conference… my observations
- Summary

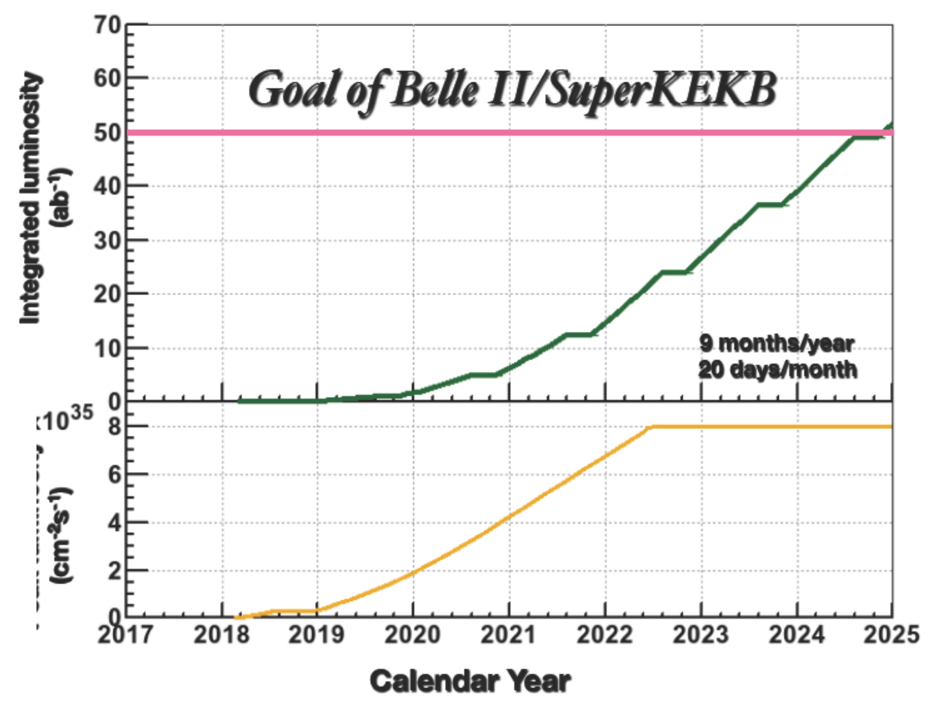
Introduction

- Both scientific computing tools and methods in HEP are changing.
- In the past it was possible to think about computing needs for a single experiment at a time. The number of participants and their growing requirements now make this impractical -> think community
- More sciences are becoming “Big Data” sciences.





SuperKEKB luminosity projection

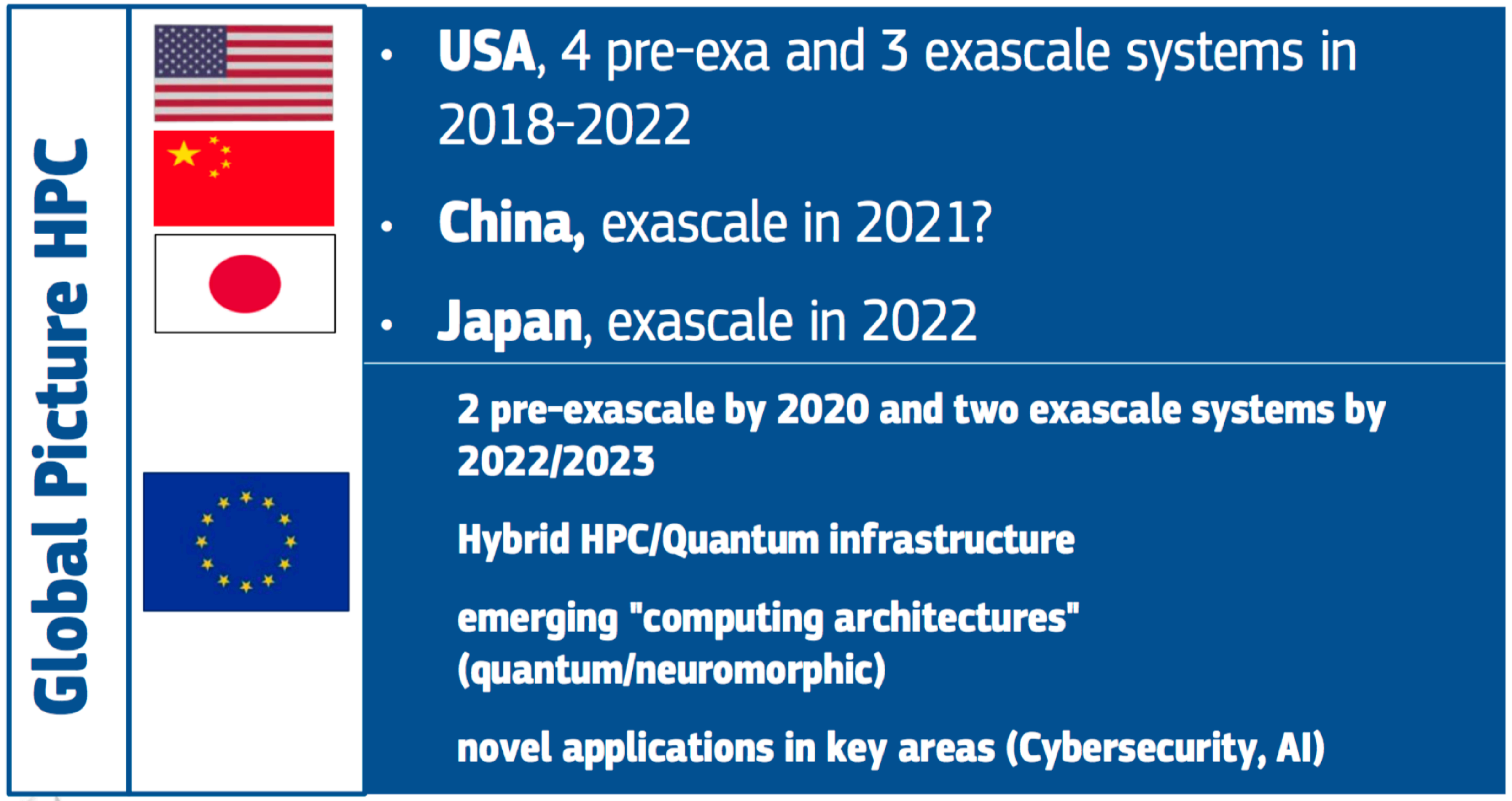


Fermilab

The Vision

International Big Data Science

- LHC, SKA, DUNE, LIGO, LSST are all data intensive sciences.
- While we know the computing challenges are equally large, others outside of HEP are planing to build exescale compute.

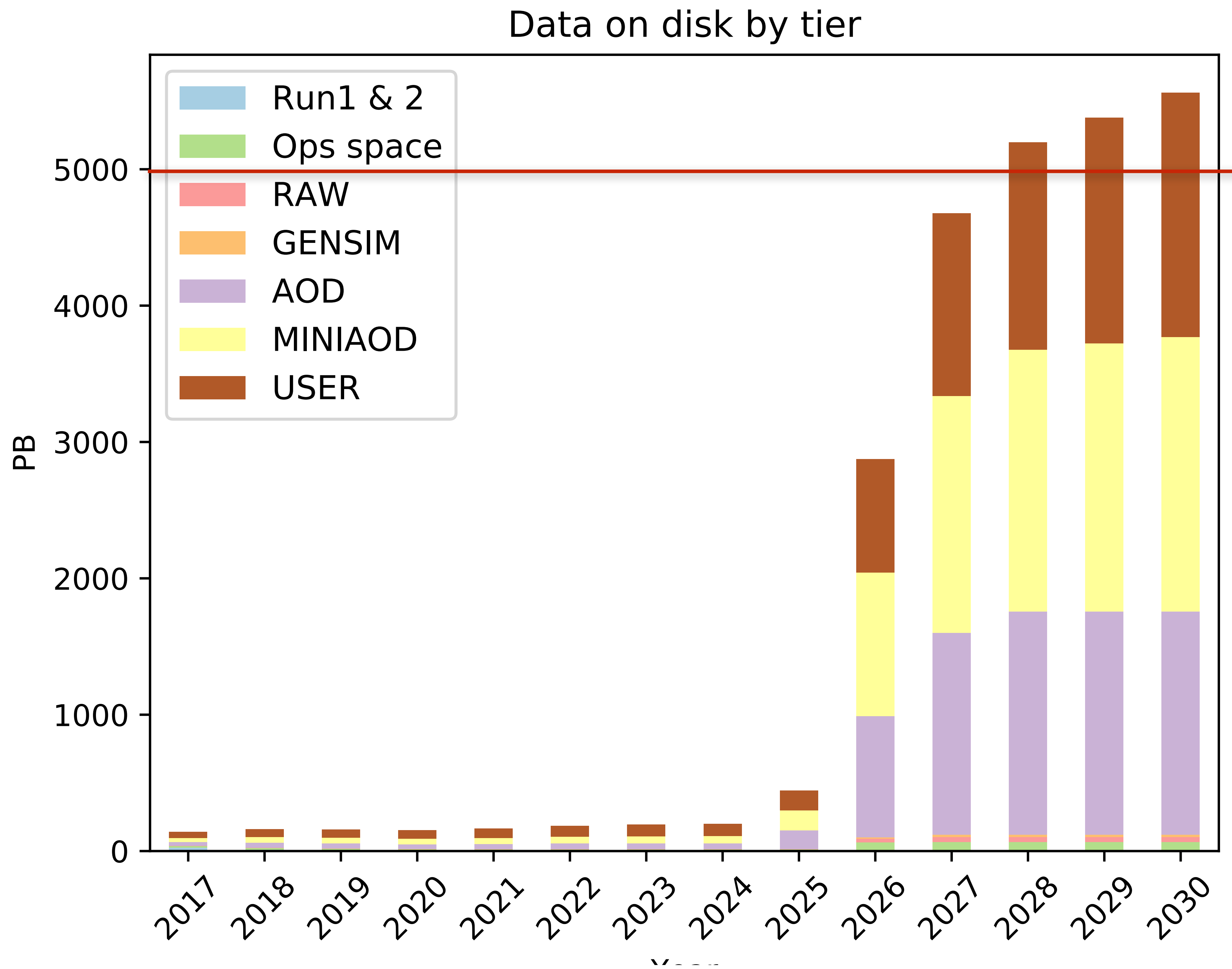


We will need to learn how to tap into this resource.

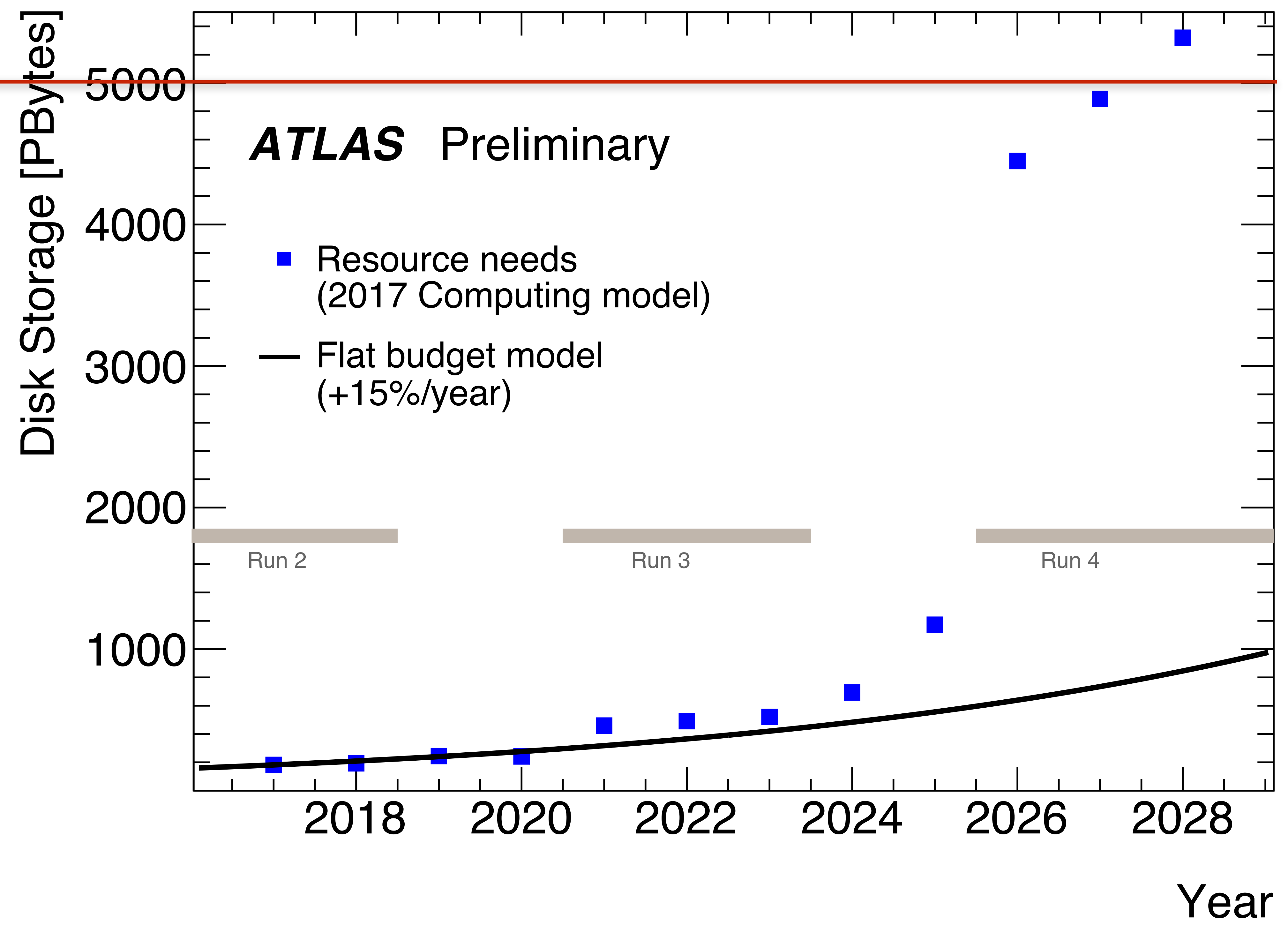
- International science requirers international data movement and storage.
- Most likely our community will have to build exe-scale data to match the exe-scale compute along with our partners in other communities.
- Going forward the LHC will not be alone in using this infrastructure.
- In fact Bell2 and DUNE have already started using it.
- For a subset of these collaborations I will have one slide each on their data needs.

HL-LHC Current Data Predictions

- These plots were created at the request of our funding agencies and represent what the needs would be extrapolating from current practice.



5 Exabytes
of Data on Disk



DUNE Data Needs

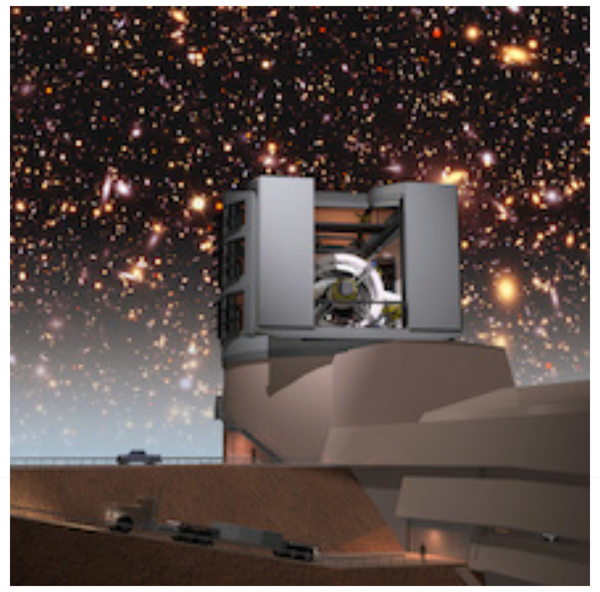
- Full Stream Data* for DUNE is impossibly large, order 150EB/year
- Much of the detector research will go into reducing that to reasonable levels
- suppression of 39Ar decay, cold electronics noise, space charge effects, argon purities all play a role
- above means that most challenging data needs for DUNE are during it's prototyping phase - now untill 2020
- Needs proposed at review: low/high = 4/59 PB, most probable 16PB

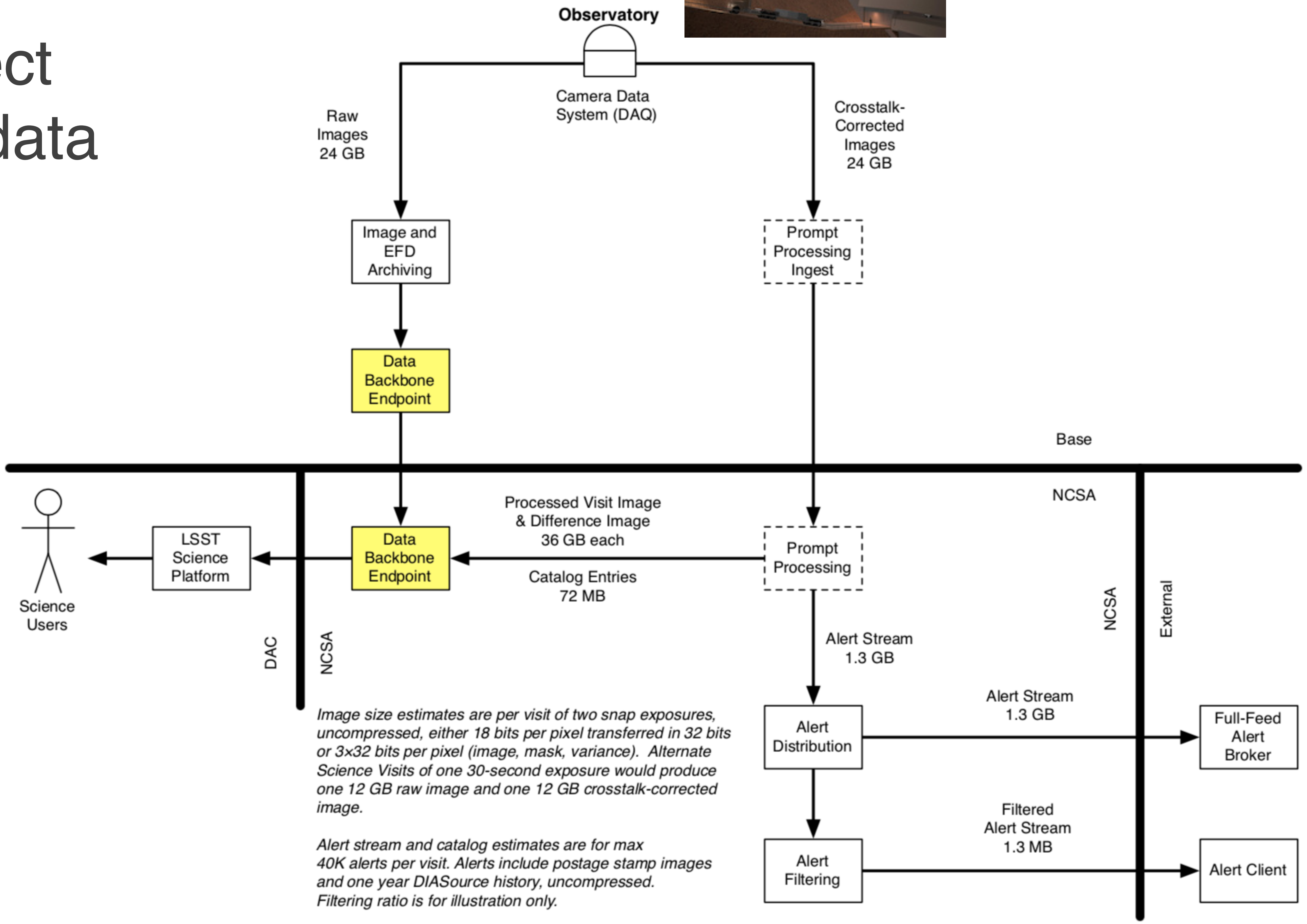
Year	*CPU (10^6 Hr)*	*Storage (TB)*	*Tape (PB) low/high*
FY18	*9.25*	*703*	*0.8/5.9*
FY19	*28.6*	*1938*	*3/49.8*
FY20	*12.5*	*237*	*0.04/3.4*

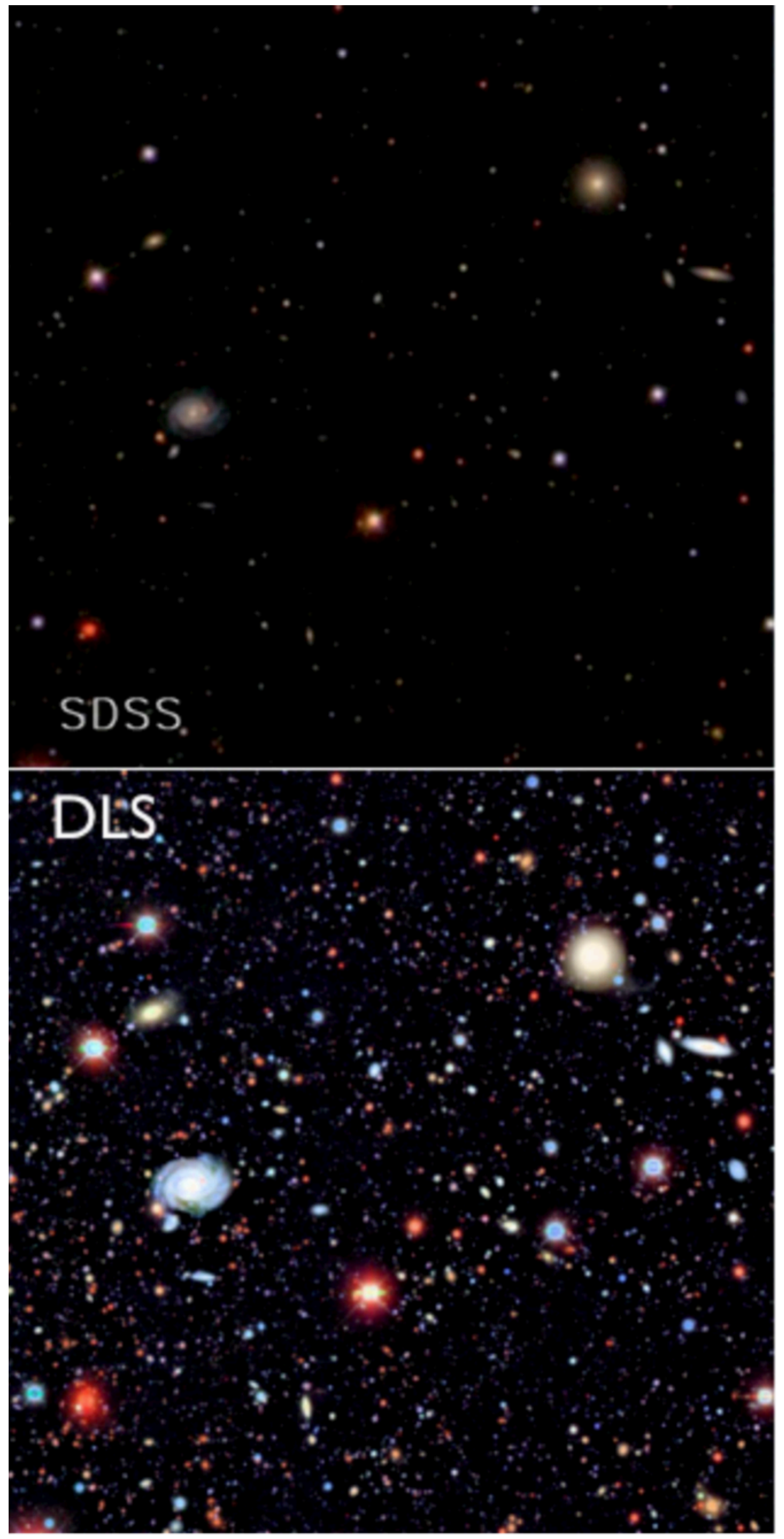
* multiply the frontend data taking rates by the number of channels

LSST Data Needs

- LSST will collect 50PB/year of data

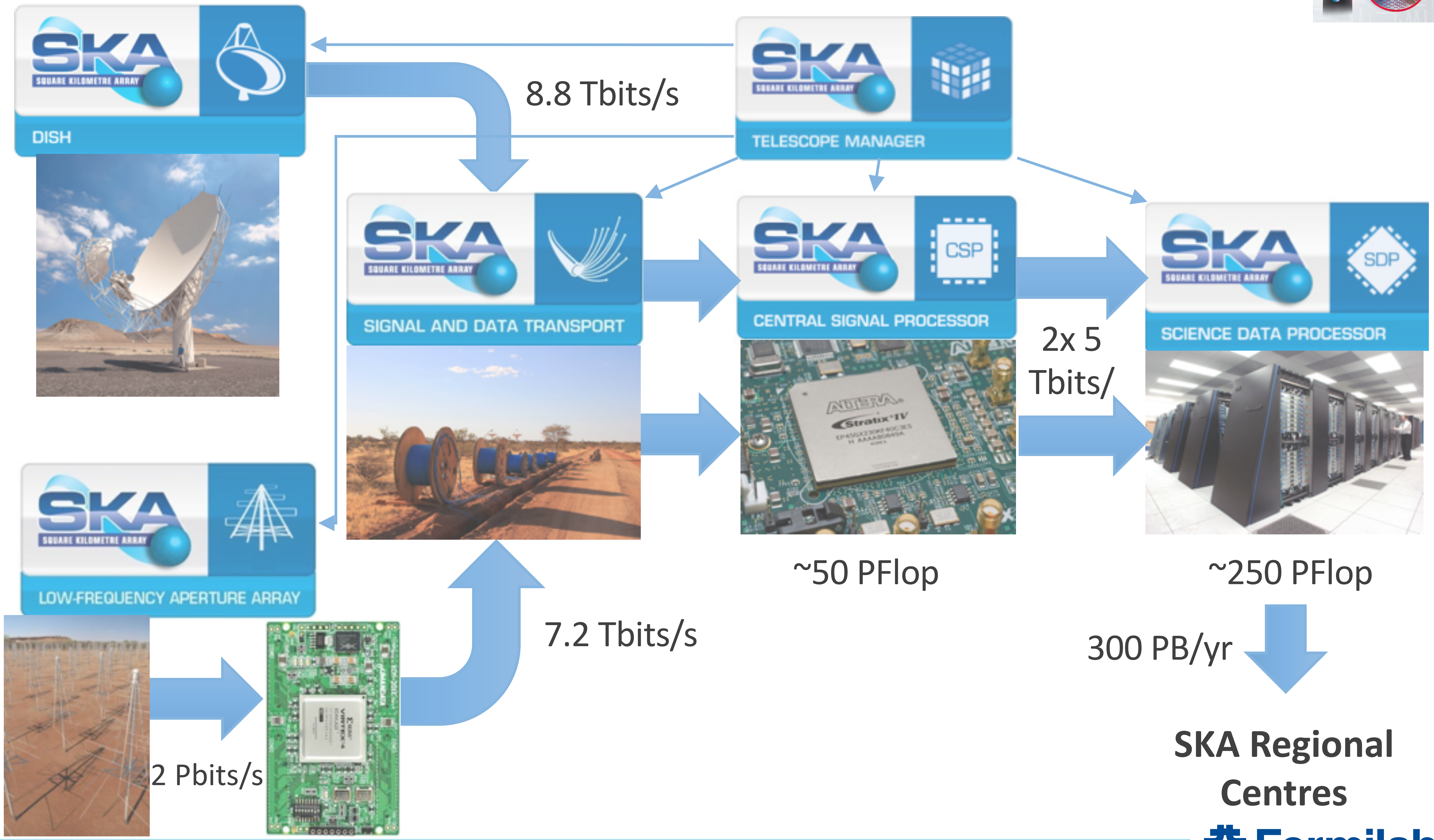






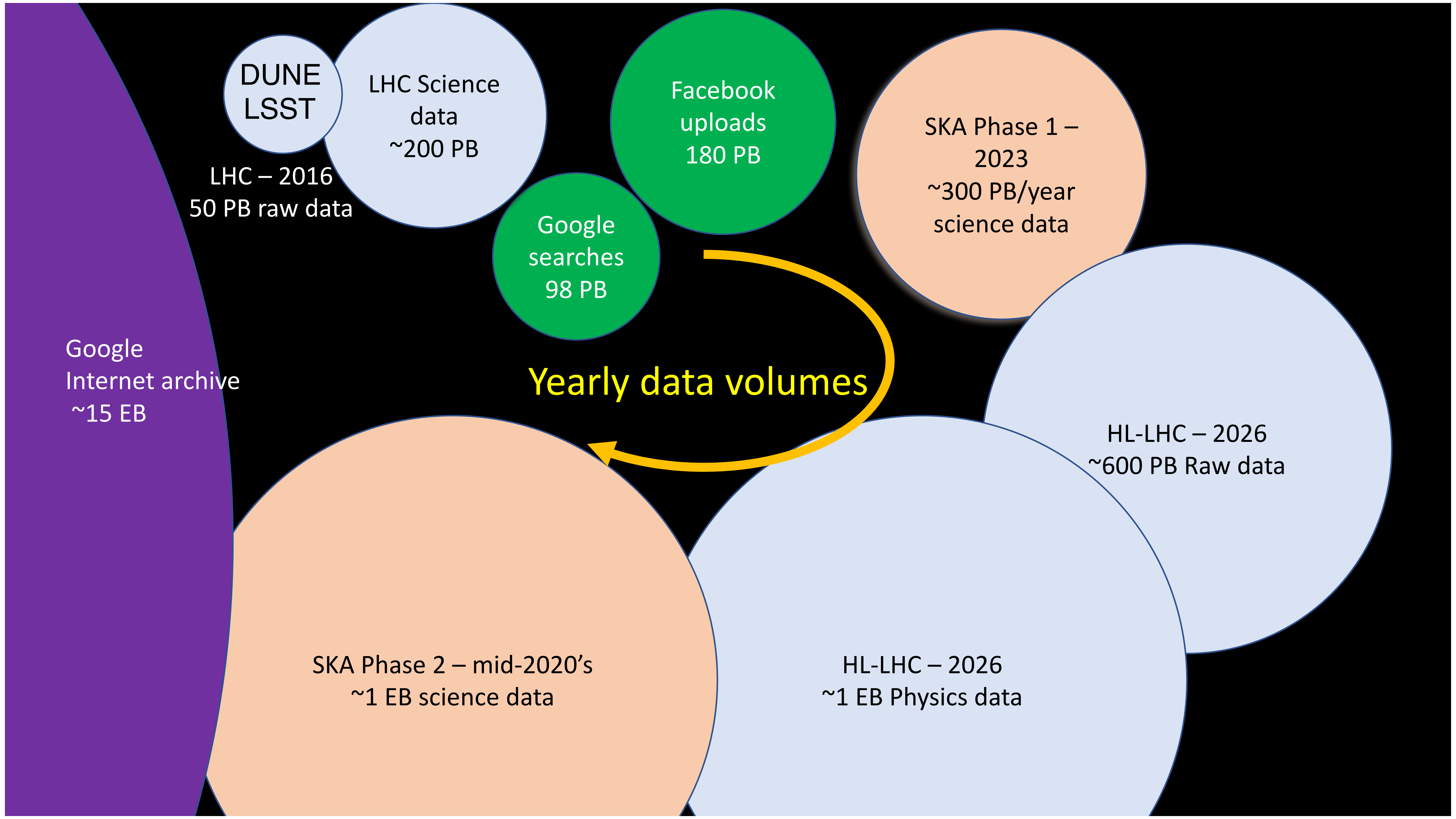
SKA Data Challenge

- SKA is a software telescope
- Very flexible and potentially easy to reconfigure
- Major software and computing challenge
- Bottom line: will collect 300PB/ year



Yearly International Data Needs

- We do this today with a world wide computing grid. It will need to grow.
- Reliable and performant networking is key to our federated data model.
- Usage of this infrastructure will have to expand to support other HEP domains as well.



Fermilab

Overheard: What is being said in the halls…

“Funding agencies will not buy computing for just HEP anymore”

“We can no longer afford to continue with business as usual.”

“We have reached the end of Denard/Moore’s law scaling and what homogeneous resources like the WLCG can deliver.”

“HL-LHC salvation will come from software improvements, not from hardware”

“The experimental physics community needs to take a page from the lattice gauge community…

The R&D Roadmap

Community White Paper [1]

A Roadmap for HEP Software and Computing R&D for the 2020s

- Inspired by the P5 process and guided by its goals
- The Global Community White Paper provides a roadmap to extend commonality to a broader set of software.
  - 70 page document
  - 13 topical sections summarising R&D in a variety of technical areas for HEP Software and Computing
  - Almost all major domains of HEP Software and Computing are covered
  - 1 section on Training and Careers
  - 310 authors (signers) from 124 HEP-related institutions

[1] https://arxiv.org/pdf/1712.06982.pdf

Contents

1 Introduction 2

2 Software and Computing Challenges 5

3 Programme of Work 11
- 3.1 Physics Generators 11
- 3.2 Detector Simulation 15
- 3.3 Software Trigger and Event Reconstruction 23
- 3.4 Data Analysis and Interpretation 27
- 3.5 Machine Learning 31
- 3.6 Data Organisation, Management and Access 36
- 3.7 Facilities and Distributed Computing 41
- 3.8 Data-Flow Processing Framework 44
- 3.9 Conditions Data 47
- 3.10 Visualisation 50
- 3.11 Software Development, Deployment, Validation and Verification 53
- 3.12 Data and Software Preservation 57
- 3.13 Security 60

4 Training and Careers 65
- 4.1 Training Challenges 65
- 4.2 Possible Directions for Training 66
- 4.3 Career Support and Recognition 68

5 Conclusions 68

Appendix A List of Workshops 71

Appendix B Glossary 73

References 79

Fermilab

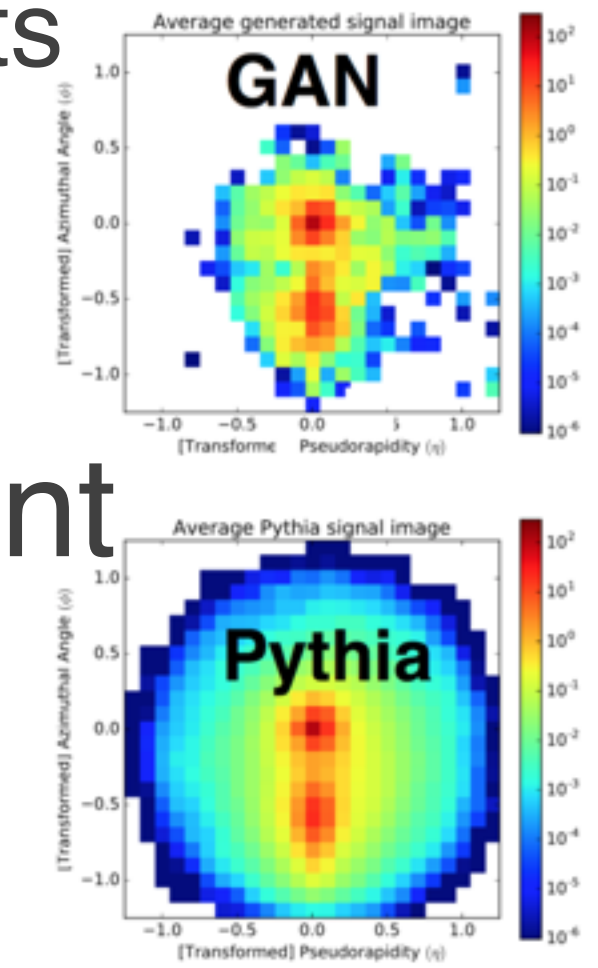
CWP Overlap with ICHEP18: Simulation

- Simulating our detectors consumes huge resources today
  - Remains a vital area for HL-LHC and intensity frontier experiments in particular
- Main R&D topics
  - Improved physics models for higher precision at higher energies (HL-LHC and then FCC) Geant4 Detector Simulations for Future HEP Experiments
  - Adapting to new computing architectures
    - Can a vectorised transport engine actually work in a realistic prototype
    - (GeantV early releases)? How painful would evolution be (re-integration into Geant4)?
  - Faster simulation - develop a common toolkit for tuning and validation of fast simulation
    - How can we best use Machine Learning profitably here? from processes to entire events

Fast calorimeter simulation in LHCb

New approaches using machine learning for fast shower simulation in ATLAS

  - Geometry modelling
    - Easier modeling of complex detectors, targeting new computing architectures
- CWP brought a more consistent view and work-plan among the different projects

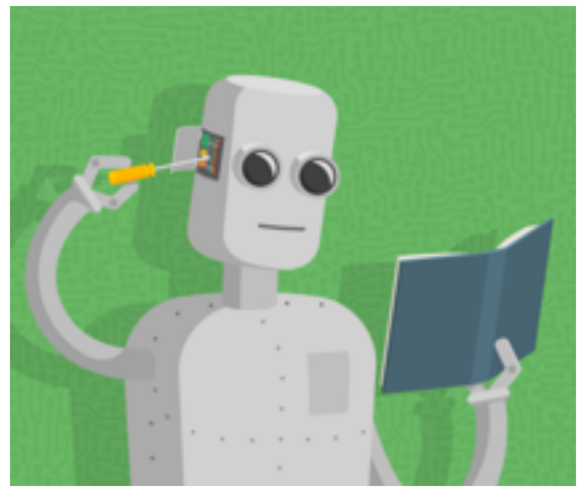


CWP Overlap with ICHEP18: SW Trigger & Reconstruction

- Moving to software triggers is already a key part of the program for LHCb and ALICE in Run 3
  - ‘Real time analysis’ increases signal rates and can make computing more efficient (storage and CPU)
- Main R&D topics
  - Controlling charged particle tracking resource consumption and maintaining performance
    - Do current algorithms’ physics output hold up at pile-up of 200 (or 1000)
    - Can tracking maintain low pT sensitivity within budget?
  - Detector design itself has a big impact (e.g., timing detectors, track triggers, layout)
    - Improved use of new computing architectures: multi-threaded and vectorised CPU code, GPGPUs, FPGAs
  - Robust validation techniques when information will be discarded
    - Using modern continuous integration, multiple architectures with reasonable turnaround times
  - Reconstruction toolkits adapted to experiment specificities: ACTS, TrickTrack, Matriplex

CWP Overlap with ICHEP18: Machine Learning

- Neural networks and Boosted Decision Trees have been used in HEP for a long time. e.g., particle identification algorithms
- The field has been significantly enhanced by new techniques (DNNs), enhanced training methods, and community-supported (Python) packages
  - Very good at dealing with noisy data and huge parameter spaces
  - A lot of interest from our community in these new techniques, in multiple fields
- Main R&D topics
  - Speeding up computationally intensive pieces of our workflows (fast simulation, tracking)
  - Enhancing physics reach with better classification than our current techniques
  - Improving data compression by learning and retaining only salient features
  - Anomaly detection for detector and computing operations

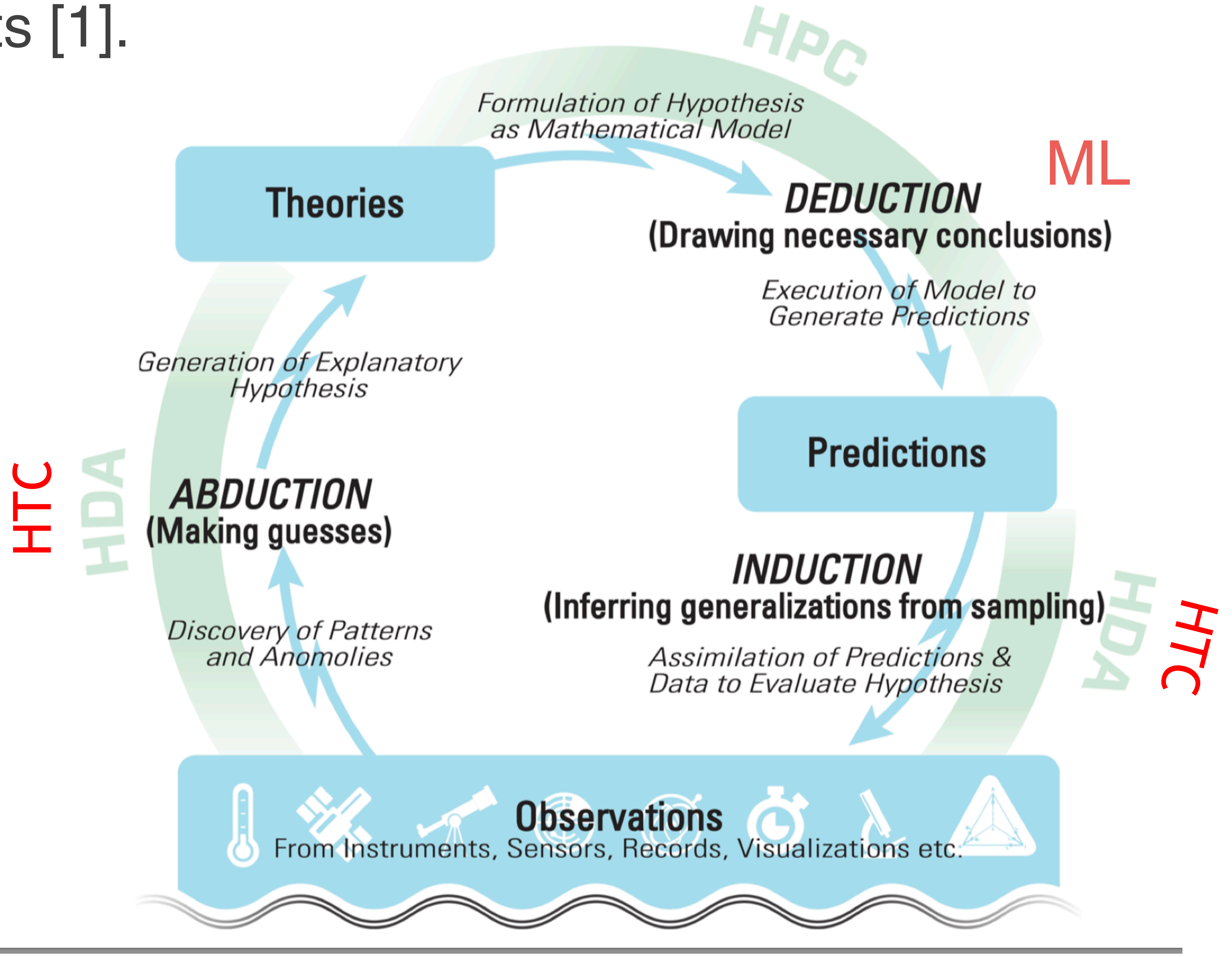


Reports from 4 experiments and the community challenge

BIG DATA AND EXTREME-SCALE COMPUTING: PATHWAYS TO CONVERGENCE

- HEP should be a major player in reconciling the split between traditional HPC and HTC ecosystems, discussed by an international group of HPC experts [1].

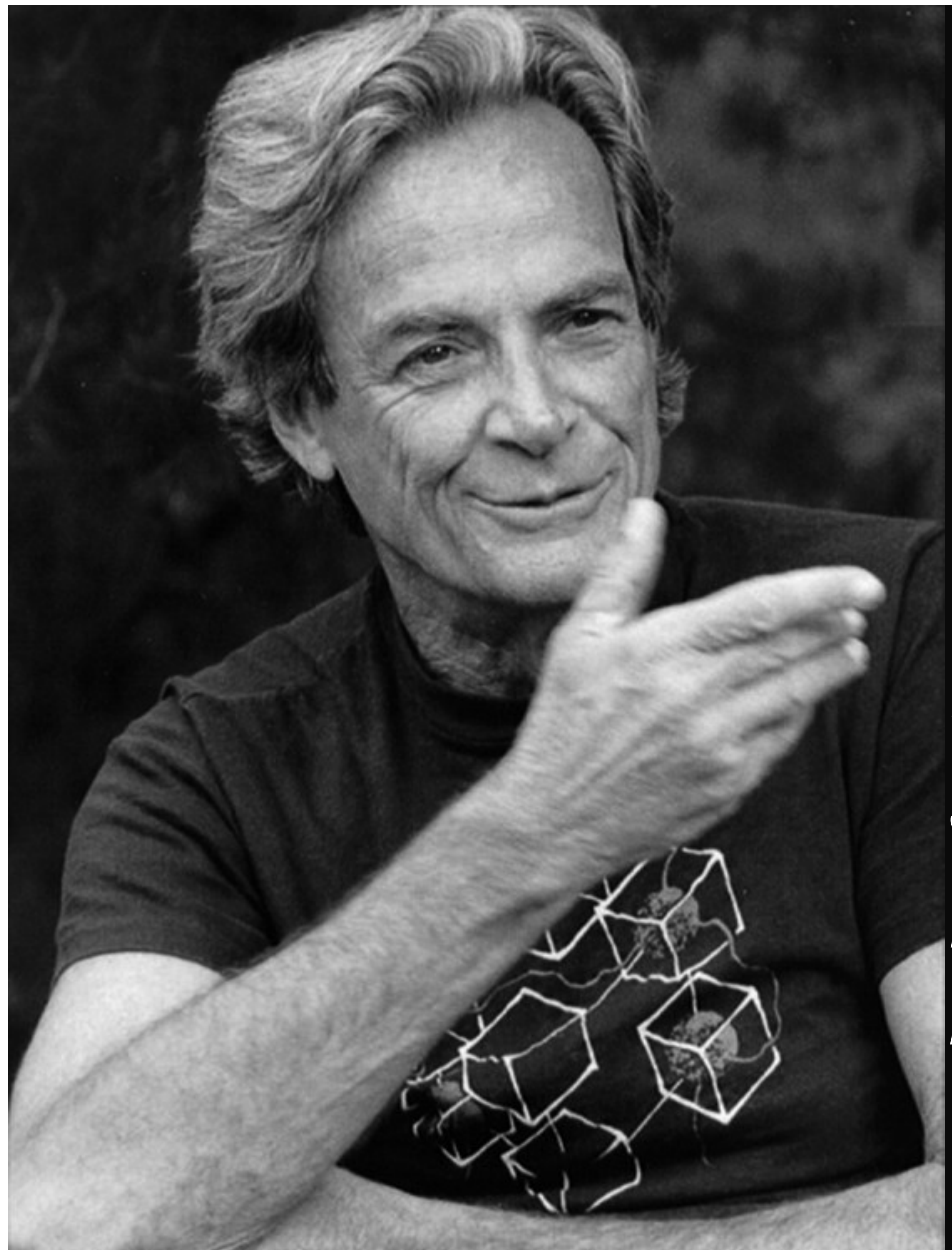
"Combining HPC and HTC applications and methods in large-scale workflows that orchestrate simulations or incorporate them into the stages of large-scale analysis pipelines for data generated by simulations, experiments, or observations"



[1] http://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/whitepapers/bdec2017pathways.pdf

A Quantum Take on Quantum Computing

Feynman was one of the originators of the idea…



Trying to find a computer simulation of physics seems to me to be an excellent program to follow out

. . .

the real use of it would be with quantum mechanics

. . .

Nature isn't classical . . . and if you want to make a simulation of Nature, you'd better make it quantum mechanical, and by golly it's a wonderful problem, because it doesn't look so easy.

—1981

- Almost 40 years later, we still don't think it is easy. However research in this field is accelerating.
  - –Electron-Phonon Systems on a Universal Quantum Computer
  - •arXiv:1802.07347

MAGIS100 - Matter-Wave Atomic Gradiometer Interferometric Sensor

ICHEP2018 SEOUL

XXXIX INTERNATIONAL CONFERENCE ON high Energy PHYSICS

ICHEP 2018 Observations

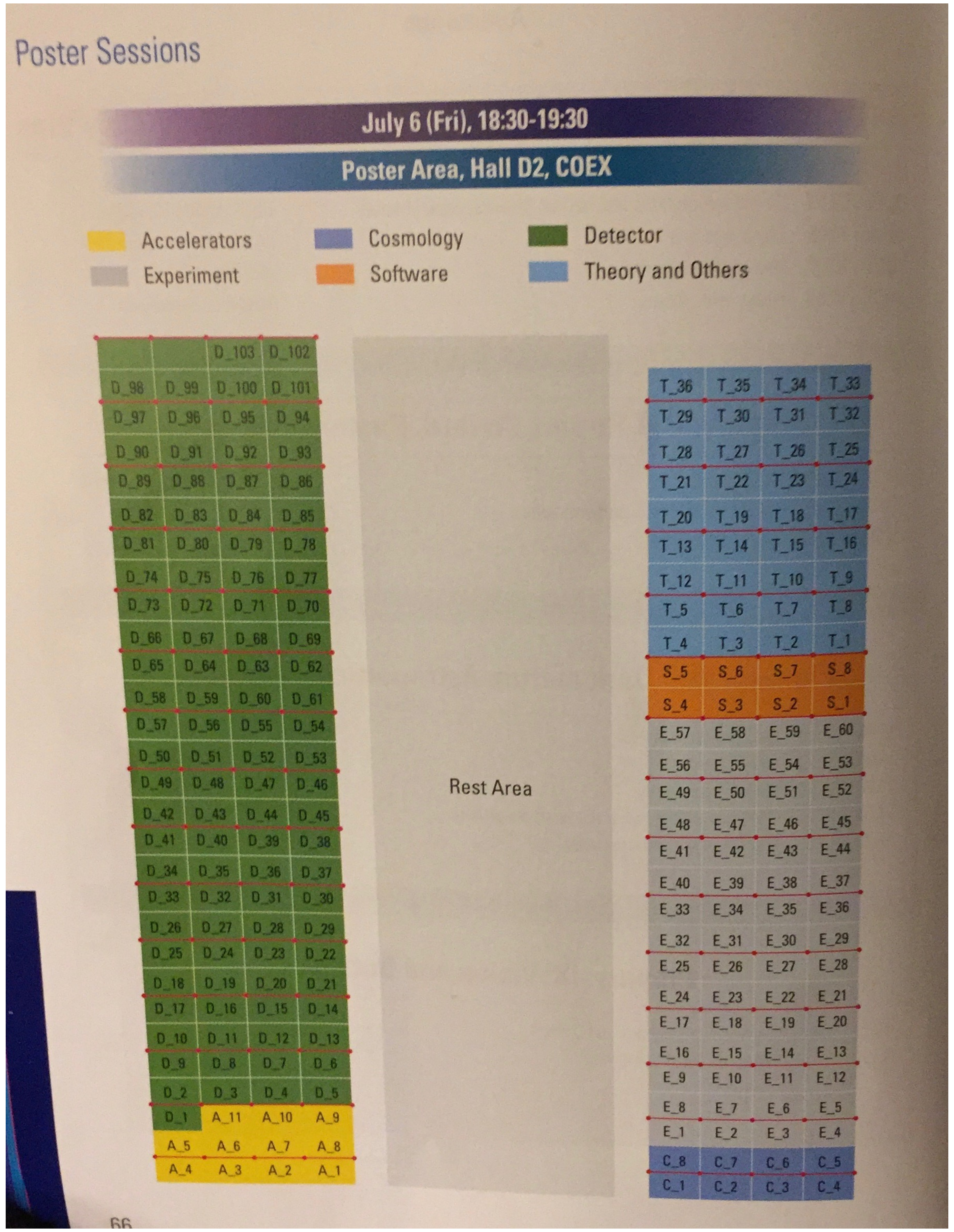
Instrumentation

- Large scientific achievements in the past decades have been enabled by large advances in instrumentation.

- Large silicon detectors and cameras with high granularity are driving us to large computing and data challenges.

- Large costs of these projects require an international scope.

Parallels and Posters



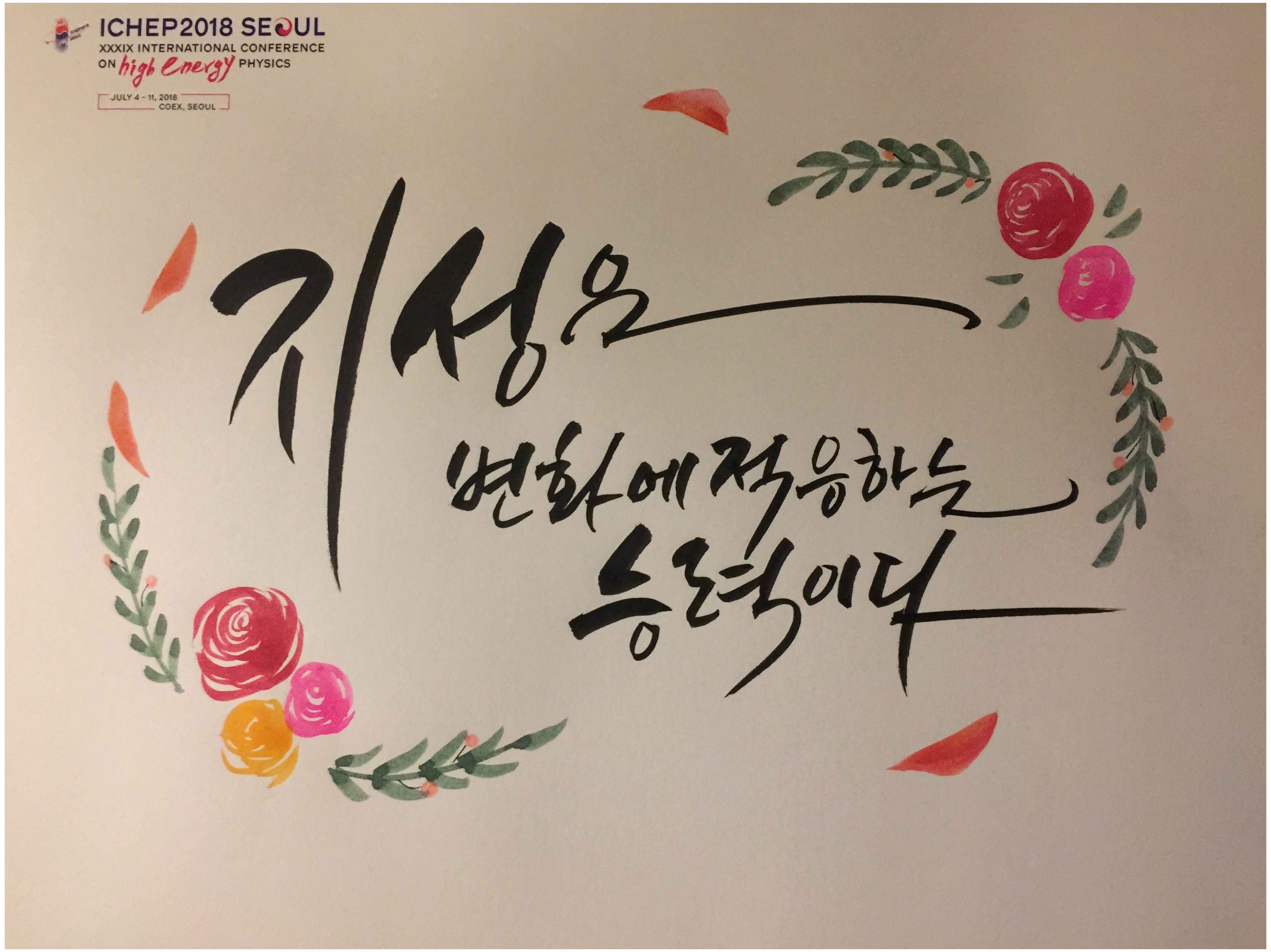
- High interest in detector
- John Kogut would say, “This interest should spill over into computing and software, take a page from LQCD”
- Think of S&C as another device necessary to extract the science.
- Indeed some talks submitted to detector were moved over to computing.
  - Lively conversation on algorithms and methods for jet, muon, and tau object reconstruction
- The large attendance at the Machine Learning parallels is a good sign.

Fermilab

Summary and Conclusions

- The data and compute challenges of the next decade are large, even daunting.
- In order to satisfy the scientific needs of our community, we will need to build unprecedented scientific facilities and capabilities
- The scientific harvest that is arriving with this new era of big data science, and exascale computing is extremely compelling.

Special thanks to Ian Bird, Michel Jouvin, Rob Gardner, Ken Herner, Marcelle Soares-Santos, and others who contributed to the slide content

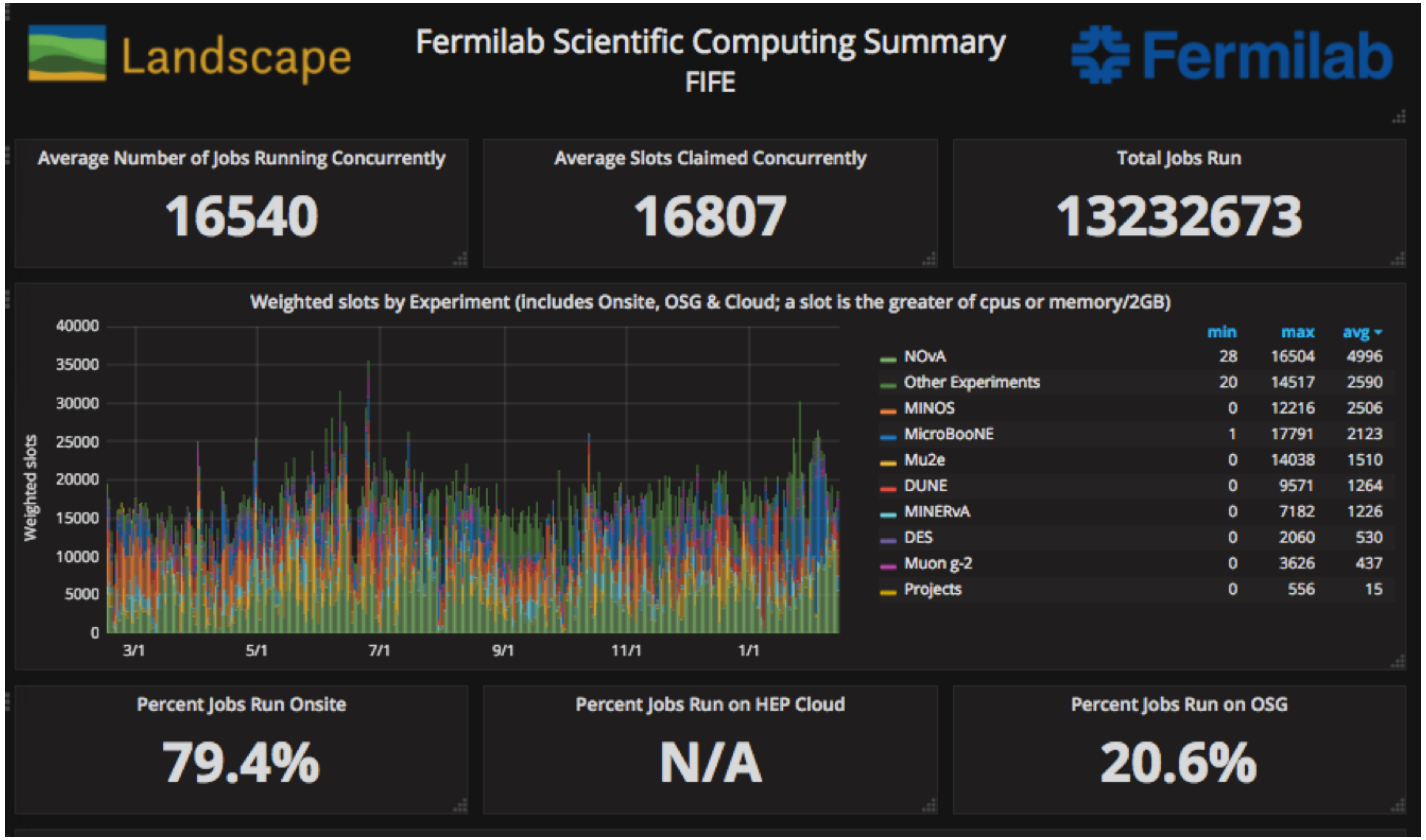


Intelligence is the ability to adapt to change. - Stephen Hawking

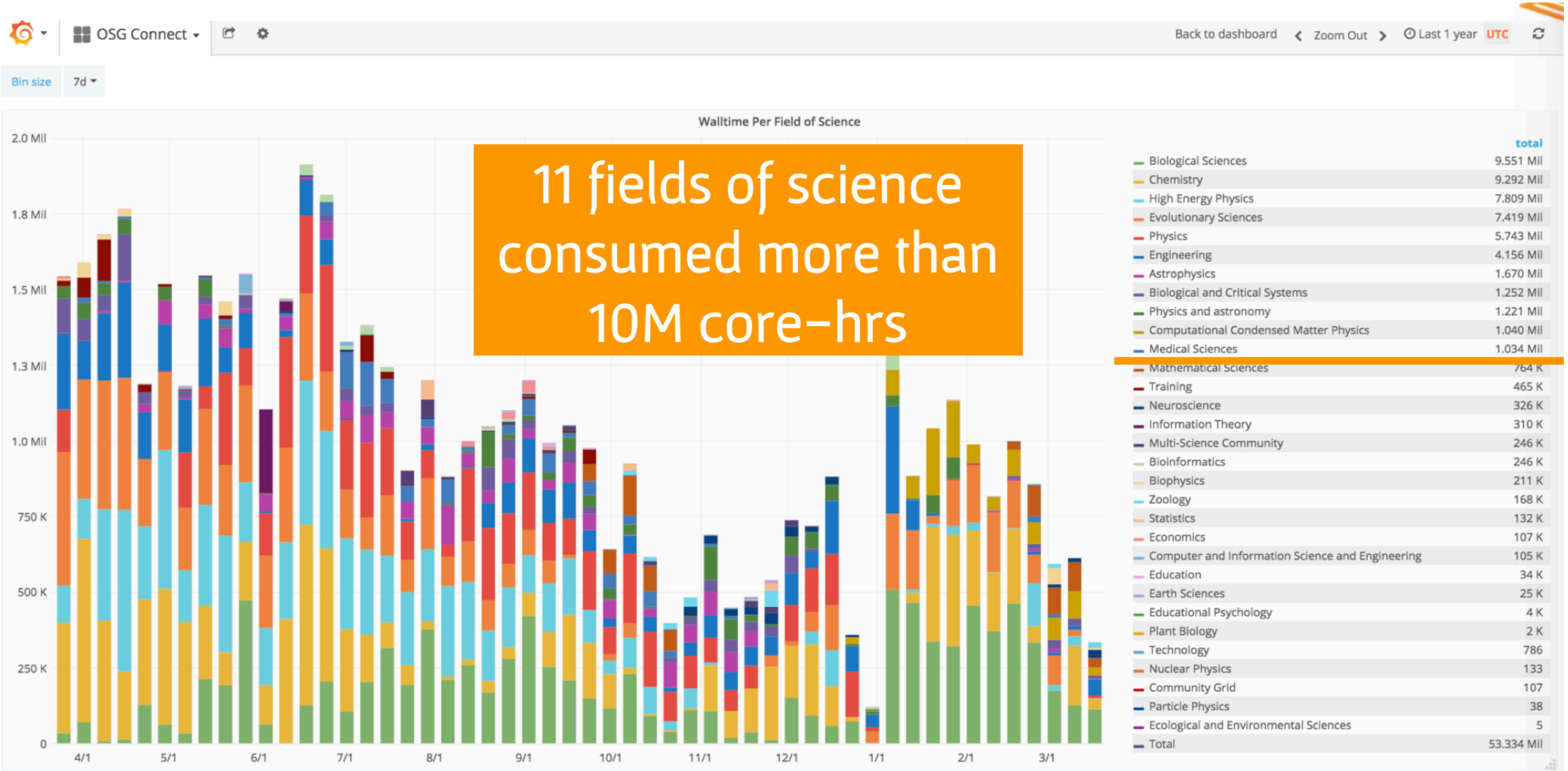
Backup

Sharing Computing Infrastructures

- Most science needs are spiky, a large number of users keeps facility utilization high.
- The mean value theorem works in computing as well.



The Long Tail of Science and the OSG



- Allowing opportunistic use of our large facilities is powerfully enabling