Argonne NATIONAL LABORATORY

APS DPF Meeting, August 4, 2017

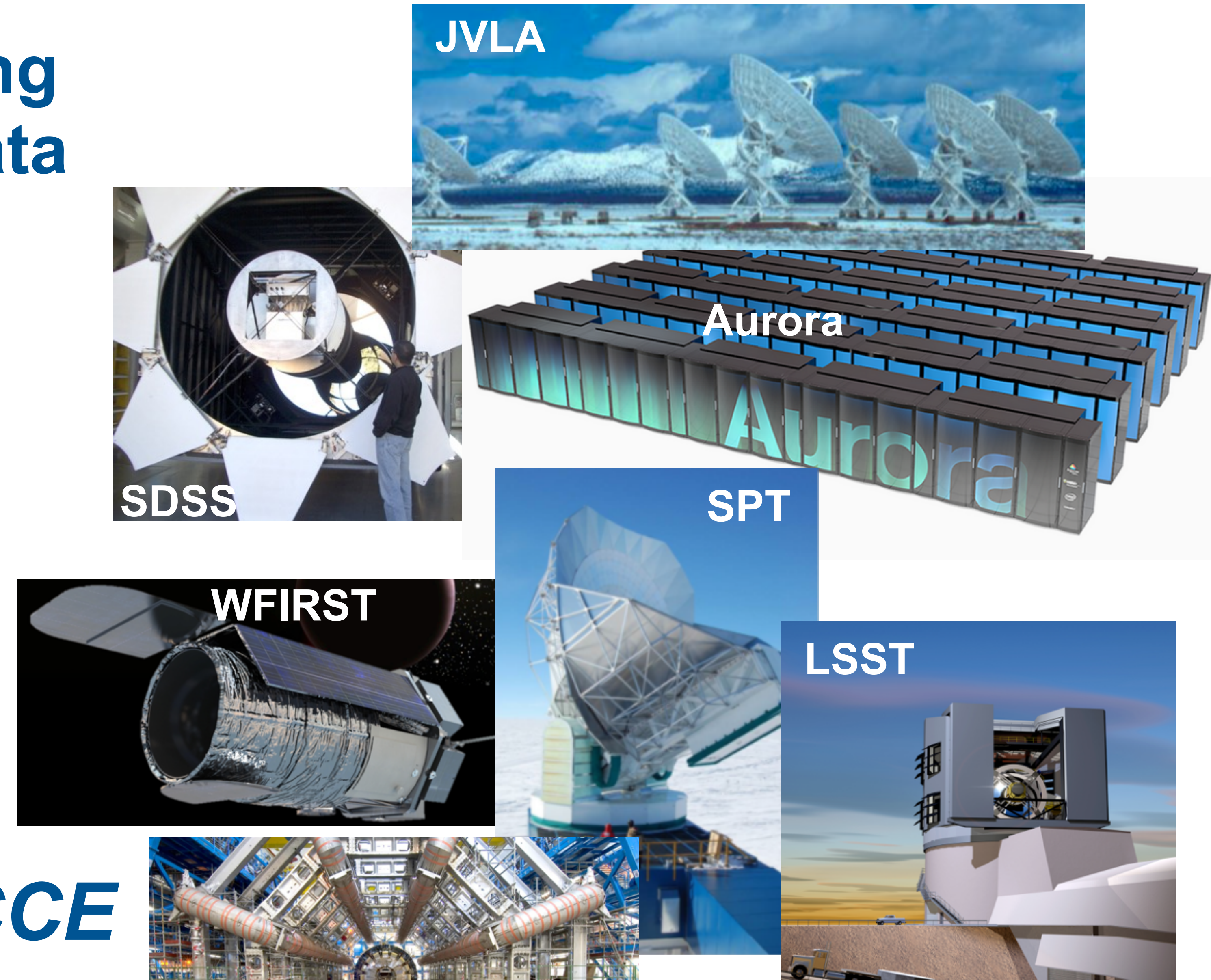
Next-Generation Computing Challenges: HPC Meets Data

Salman Habib

High Energy Physics Division
Mathematics & Computer Science Division
Argonne National Laboratory

Computation Institute
Argonne National Laboratory
The University of Chicago

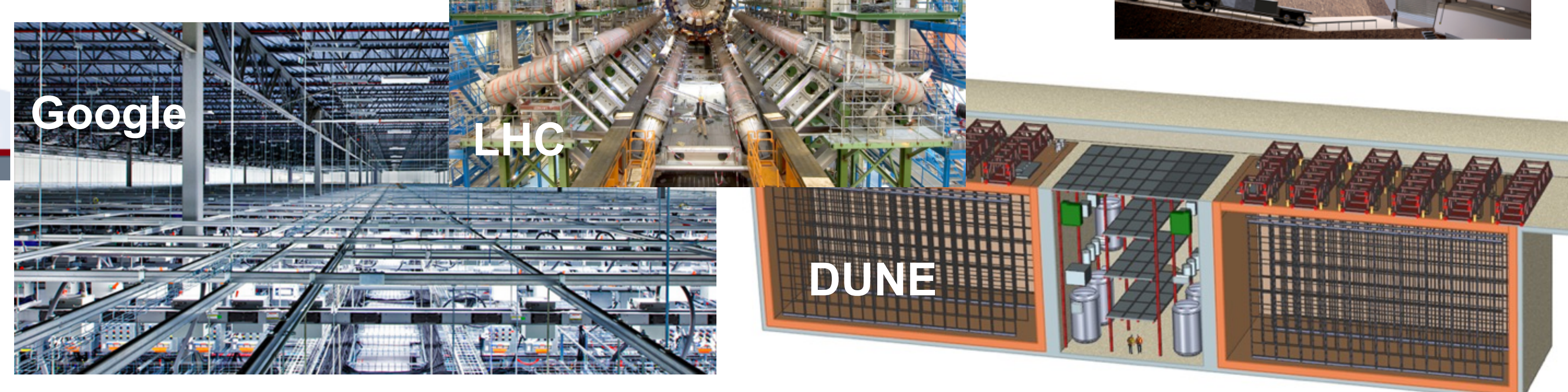
Kavli Institute for Cosmological Physics
The University of Chicago



ECP EXASCALE COMPUTING PROJECT

HEP-CCE

SciDAC Scientific Discovery through Advanced Computing



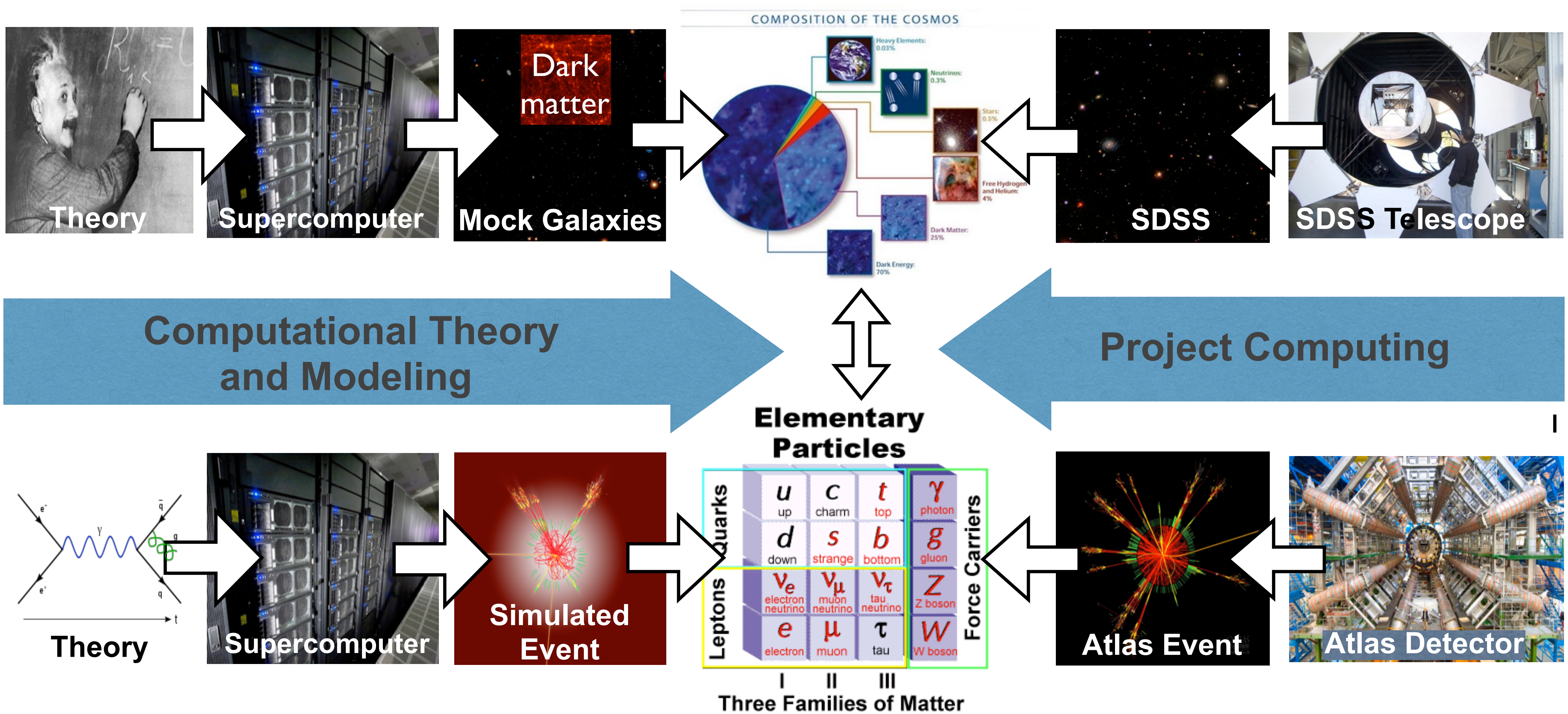
NSF NASA

High Energy Physics and Computing

- **Scales**
  - HEP science covers a number of scales (table-top to the most complex experiments in the world) and computing models (laptop to world-wide grid)
- **HEP Frontiers**
  - Energy Frontier (large experiments at colliders, O(1000) researchers/expt)
  - Intensity Frontier (small/medium/large, O(10-1000) researchers/expt)
  - Cosmic Frontier (small/medium/large scale, O(10-1000) researchers/expt)
- **Data**
  - Most experimental data requires fine-grained, ‘event’ style analysis
  - Data pipelines are complex and need to be run many times (individual campaigns can last for months)
  - Scale of data — 10s of TB to 100s of PB/year **(Exabyte already)**
  - Multiple IO requirements
- **ASCR/HEP Exascale Requirements Review (good place for details)**
  - http://arxiv.org/abs/1603.09303, also http://hepcce.org/resources/reports/

Computing Paradigm (Cosmic and Energy Frontiers)

Simulated Data: 1) Large-scale simulation of the Universe, 2) Synthetic catalogs, 3) Statistical inference (cosmology); **Analysis:** Comparison with actual data



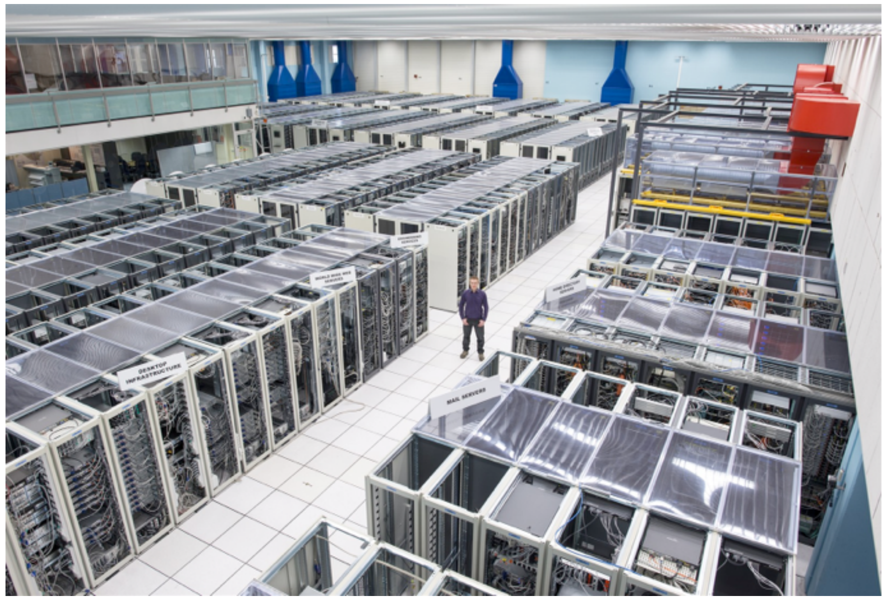
Simulated Data: 1) Event generation (lists of particles and momenta), 2) Simulation (interaction with detector), 3) Reconstruction (presence of particles inferred from detector response); **Analysis:** Comparison with actual data

Different Flavors of Computing

- **High Performance Computing ('PDEs')**
  - Parallel systems with a fast network
  - Designed to run tightly coupled jobs
  - "High performance" parallel file system
  - Batch processing
- **Data-Intensive Computing ('Interactive Analytics')**
  - Parallel systems with balanced I/O
  - Designed for data analytics
  - System level storage model
  - Fast Interactive processing

Want more of this — ("Science Cloud"), but don't yet (really) have it (Data-Intensive Scalable Computing: DISC)

- **High Throughput Computing ('Events'/'Workflows')**
  - Distributed systems with "slow" networks
  - Designed to run loosely coupled jobs
  - System level/Distributed data model
  - Batch processing



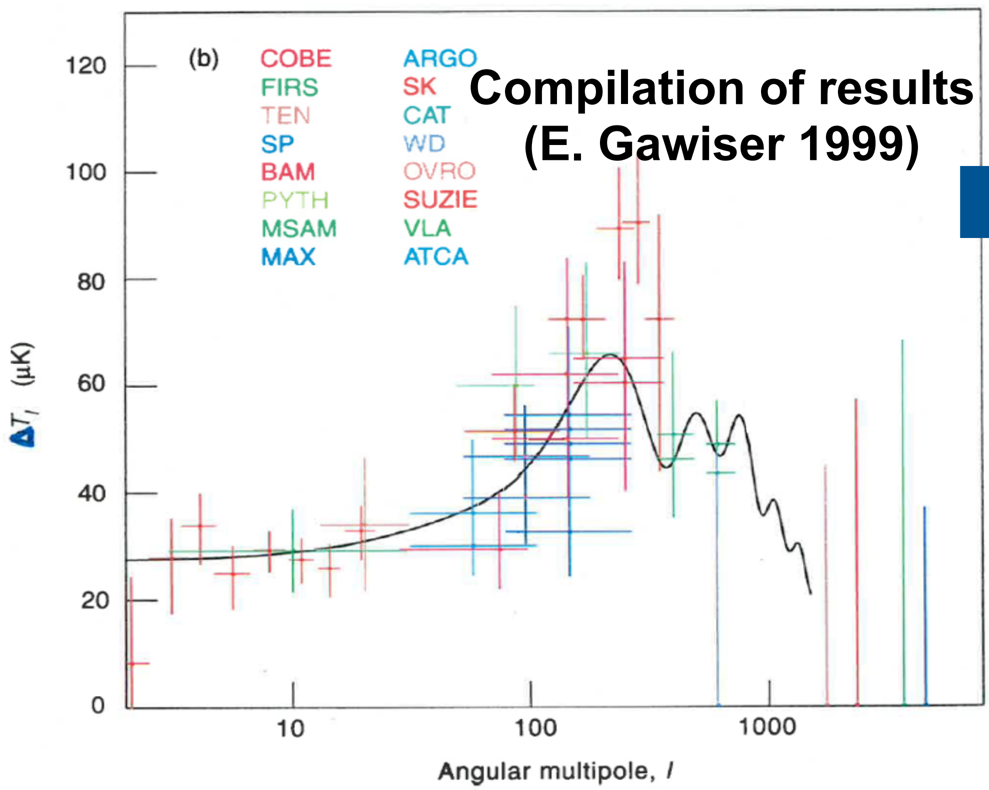
Timing Example: LSST and Computing

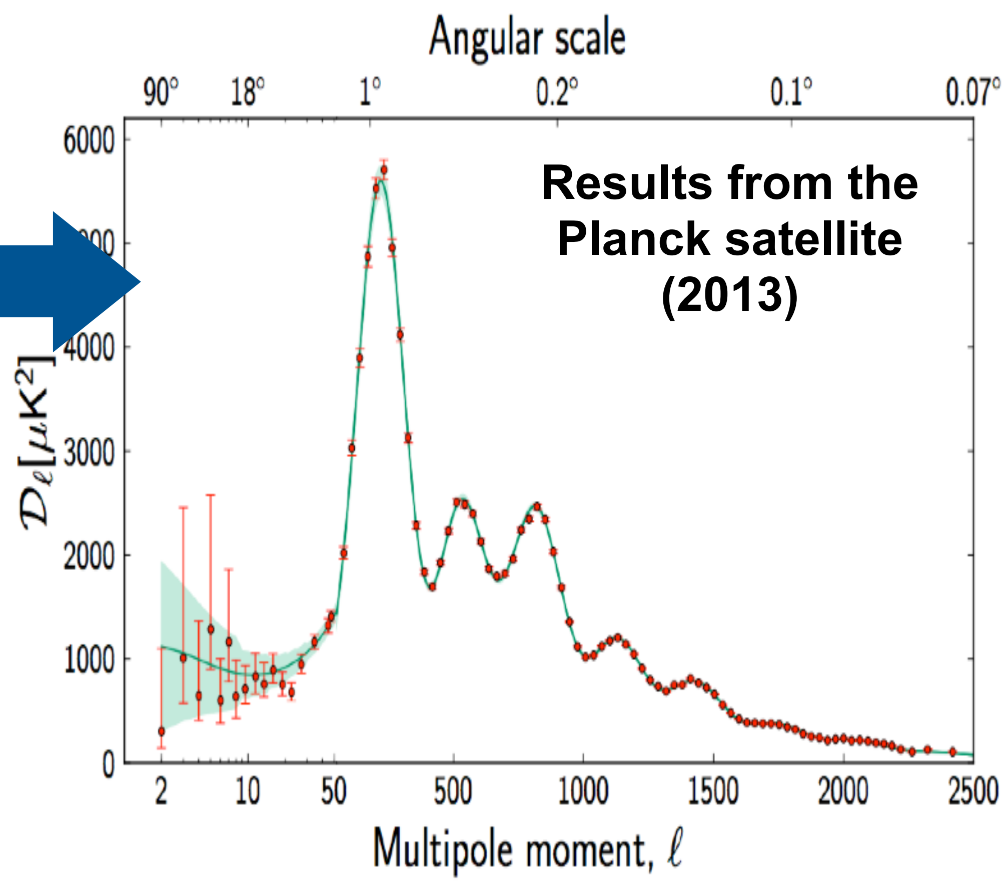
- **LSST computing (pipeline + analysis)**
  - Estimates of initial computing needs are unclear, ranging from 150-350 TFlops/year
  - Initial storage needs are ~PB, growing linearly
  - Based on this, we would want (at least) the #1 machine in the Top 500 in 2006
  - In 2022 there may be O(1000-10,000) such machines in the US alone!
  - Storage requirement is already ‘trivial’, LSST is NOT ‘Big Data’
- **So what’s the problem?**
  - Analyses will be complex (and there will be multiple reprocessing steps)
  - These tasks will expand to fill available computational space
  - Programming models may be very different from those in use today

IBM BG/L, Top 500 #1 in 2006

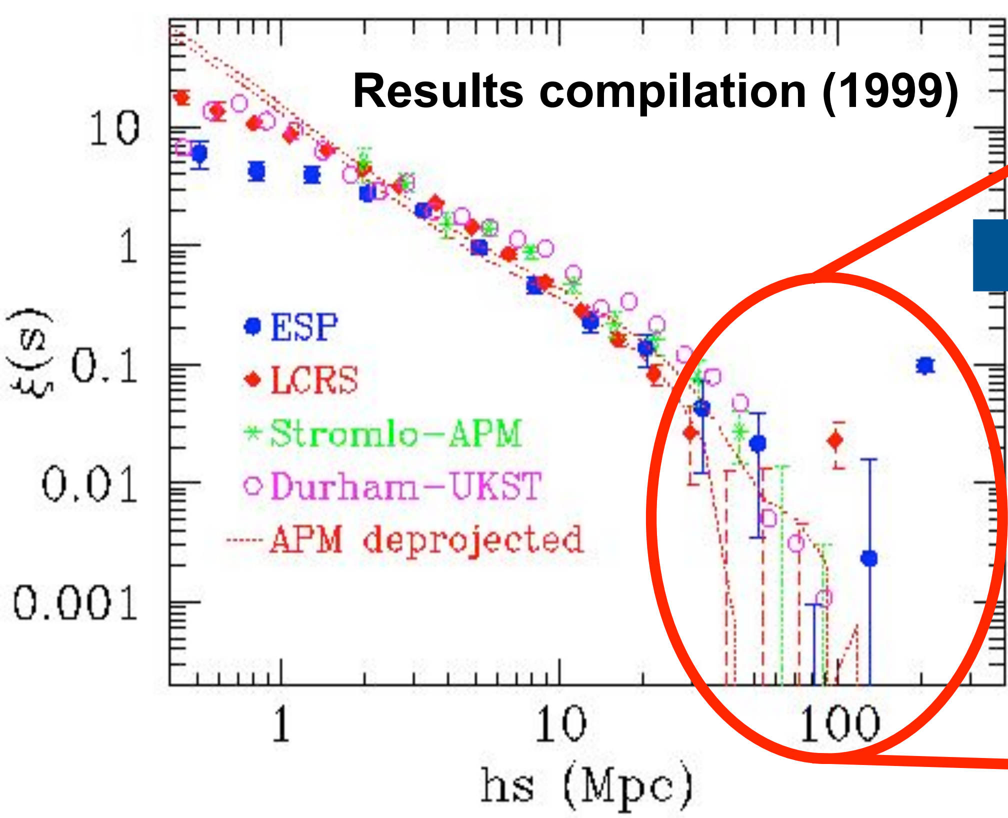
300 TFlops/10PB,
10kW in 2020
(Projection)

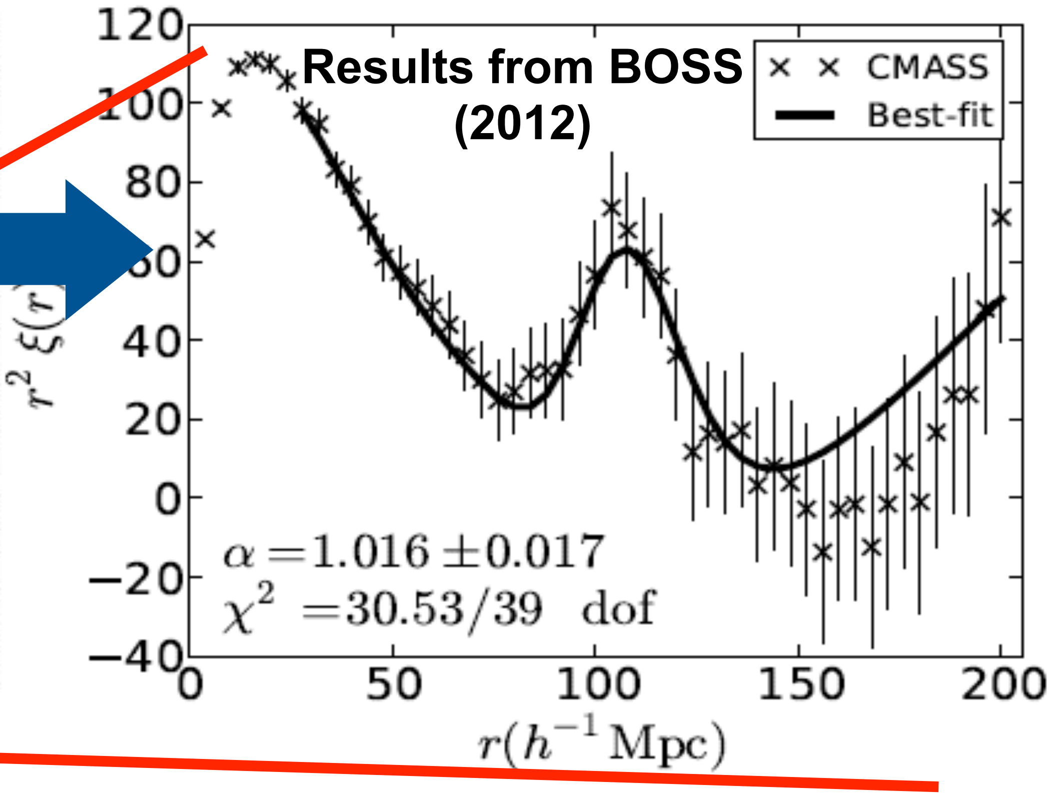
Computing Science Drivers: Cosmology





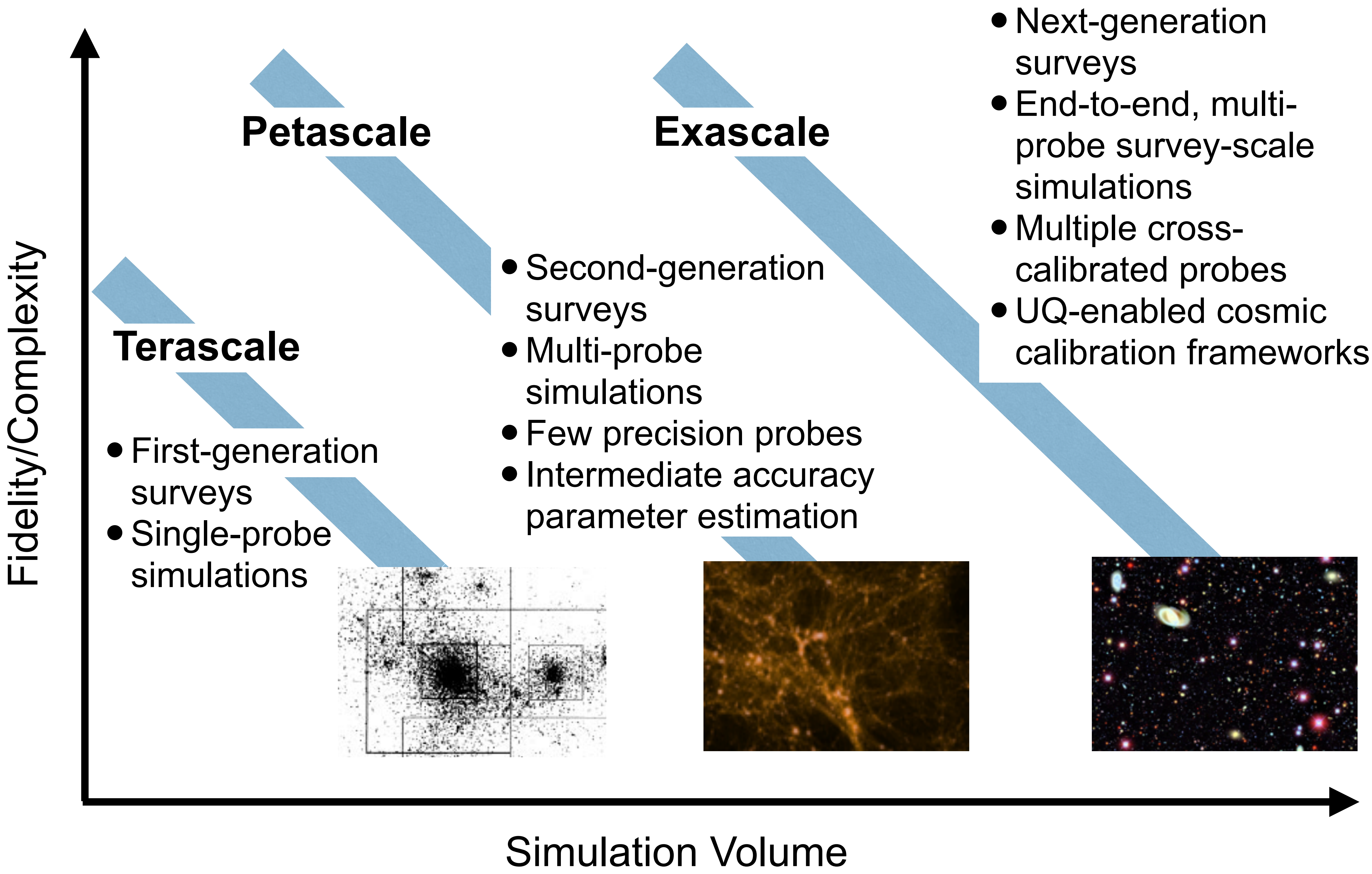
- Massive increase in sensitivity of cosmic microwave background (CMB) observations
- Cross-correlation with galaxy surveys
- New era of CMB modeling/simulations





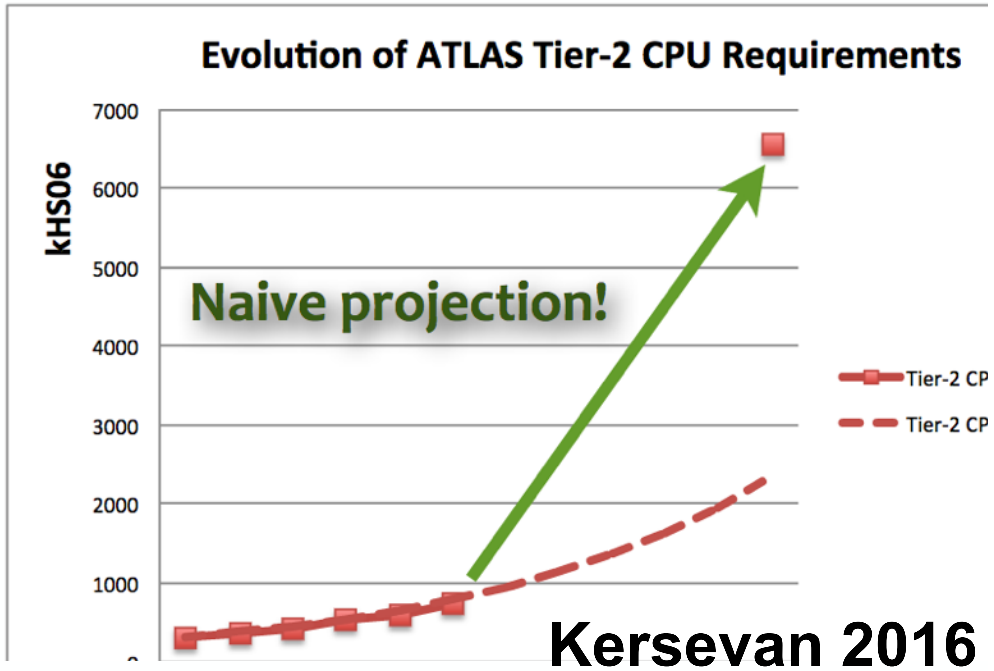
- Massive increase in volume of galaxy surveys
- Next-generation galaxy clustering simulations
- Multi-physics codes needed to meet accuracy requirements

Cosmology: Simulation Frontiers

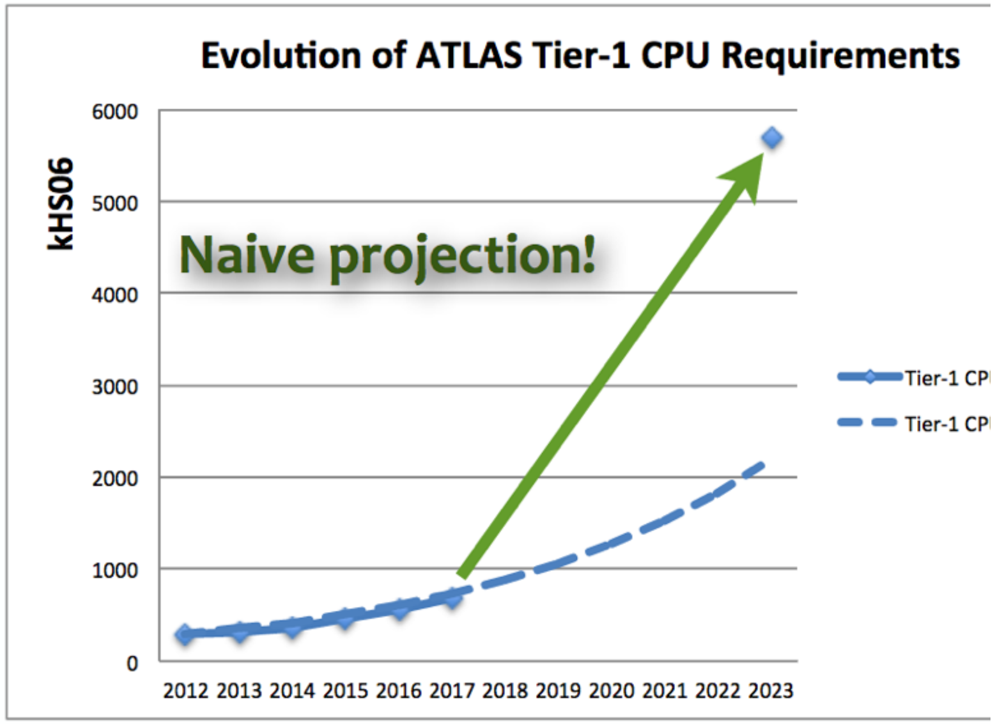


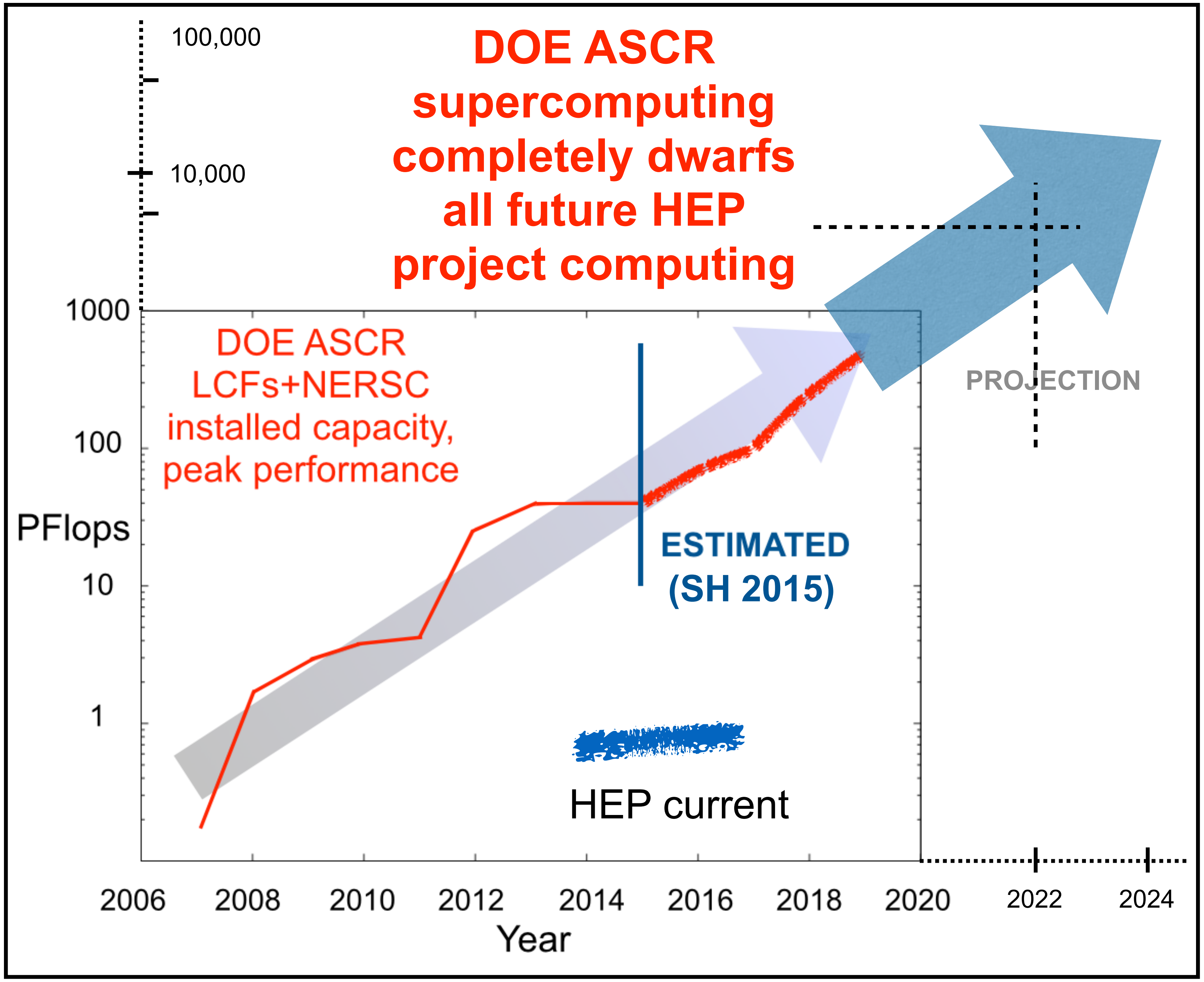
Computing Requirements: Energy Frontier

- **HEP Requirements in computing/storage will scale up by ~50X over 5-10 years**
  - Flat funding scenario fails — must look for alternatives!

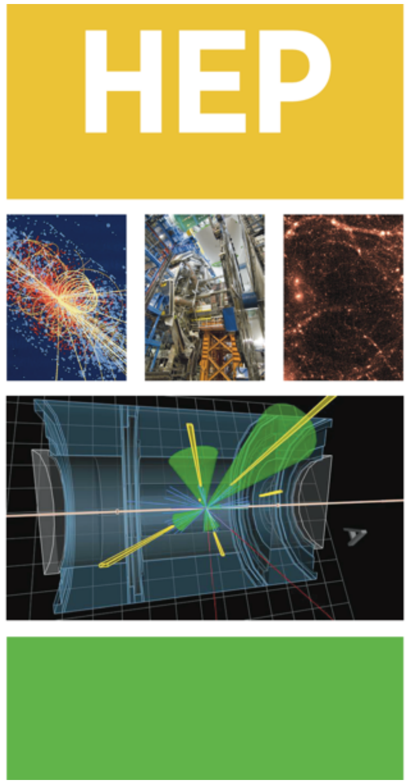


Kersevan 2016





What to Do? Many White Papers and Reports —



HIGH ENERGY PHYSICS

EXASCALE REQUIREMENTS REVIEW

An Office of Science review sponsored jointly by Advanced Scientific Computing Research and High Energy Physics

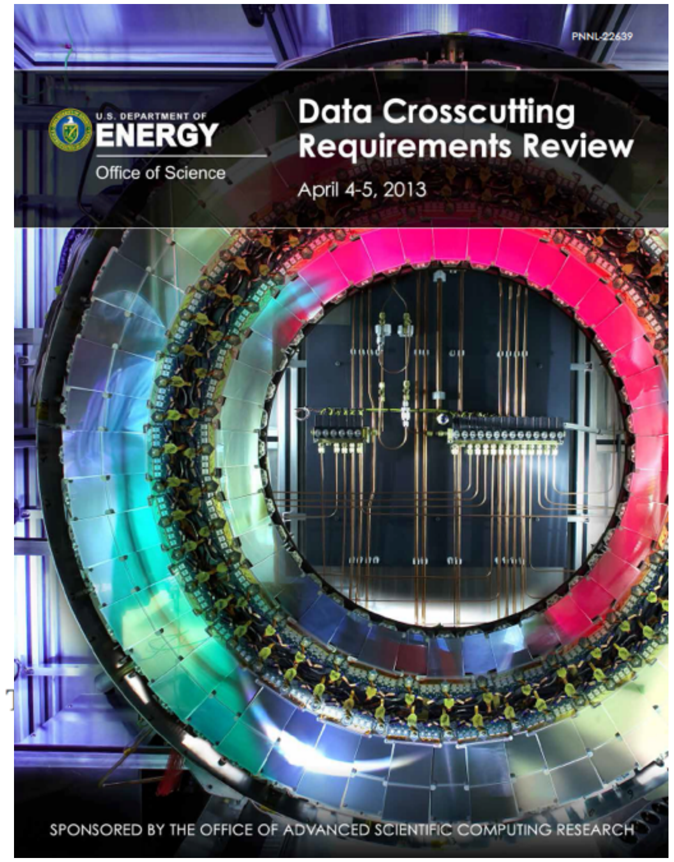
Lead Authors

HEP

Salman Habib[1] and Robert Roser[2]

ASCR

Richard Gerber,[3] Katie Antypas,[3] Katherine Riley,[1] T
and Tjerk Straatsma[4]



HIGH ENERGY PHYSICS FORUM FOR COMPUTATIONAL EXCELLENCE: WORKING GROUP REPORTS

I. APPLICATIONS SOFTWARE
II. SOFTWARE LIBRARIES AND TOOLS
III. SYSTEMS

Lead Editors: Salman Habib[1] and Robert Roser[2] (HEP-FCE Co-Directors)

Applications Software Leads: Tom LeCompte[1], Zach Marshall[3]
Software Libraries and Tools Leads: Anders Borgland[4], Brett Viren[5]
Systems Lead: Peter Nugent[3]

Applications Software Team:
Makoto Asai[4], Lothar Bauerdick[2], Hal Finkel[1], Steve Gottlieb[6], Stefan Hoeche[4], Tom LeCompte[1], Zach Marshall[3], Paul Sheldon[7], Jean-Luc Vay[3]

Software Libraries and Tools Team:
Anders Borgland[4], Peter Elmer[8], Michael Kirby[2], Simon Patton[3], Maxim Potekhin[3], Brett Viren[3], Brian Yanny[2]

Systems Team:
Paolo Calafiura[3], Eli Dart[3], Oliver Gutsche[2], Taku Izubuchi[5], Adam Lyon[2], Peter Nugent[3], Don Petravick[9]

Report from the Topical Panel Meeting on Computing and Simulations in High Energy Physics

Sponsored by the U.S. Department of Energy,
Office of Science, High Energy Physics
December 9-11, 2013 Rockville Hilton Hotel, Rockvil

Steering Committee

Paul Avery (co-Chair)	U Florida
Salman Habib (co-Chair)	Argonne
Amber Boehnlein	SLAC
Robert Roser	Fermilab
Stephen Sharpe	U Washington
Heidi Schellman	Northwestern
Craig Tull	LBNL
Torre Wenaus	BNL

Planning the Future of U.S. Particle Physics

Report of the 2013 Community Summer Study

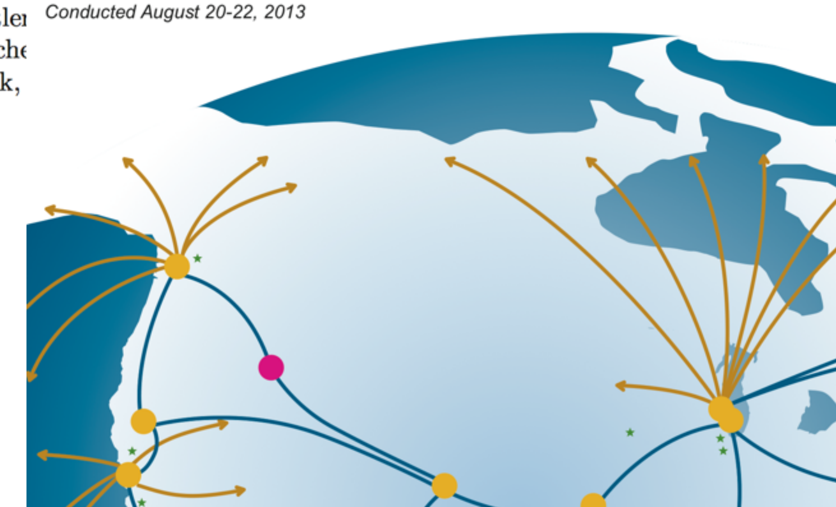
L. A. T. Bauerdick, S. Gottlieb, G. Bell, K. Bloom, T. Blum, D. Brown, M. Butle
E. Cormier, P. Elmer, M. Ernst, I. Fisk, G. Fuller, R. Gerber, S. Habib, M. Hildreth, S. Hoeche
C. Joshi, A. Mezzacappa, R. Mount, R. Pordes, B. Rebel, L. Reina, M. C. Sanchez, J. Shank,
A. Szalay, R. Van de Water, M. Wobisch, S. Wolbers

Chapter 9: Computing

High Energy Physics and Nuclear Physics Network Requirements

HEP and NP Network Requirements Review
Final Report

Conducted August 20-22, 2013



Are Supercomputers a Universal Solution?

- **Dealing with supercomputers is painful!**
  - HPC programming is tedious (MPI, OpenMP, CUDA, OpenCL, —)
  - Batch processing ruins interactivity
  - File systems corrupt/eat your data
  - Software suite for HPC work is very limited
  - Analyzing large datasets on HPC systems is painful
  - HPC experts are not user-friendly
  - Downtime and mysterious crashes are common
  - Ability to ‘roll your own’ is limited

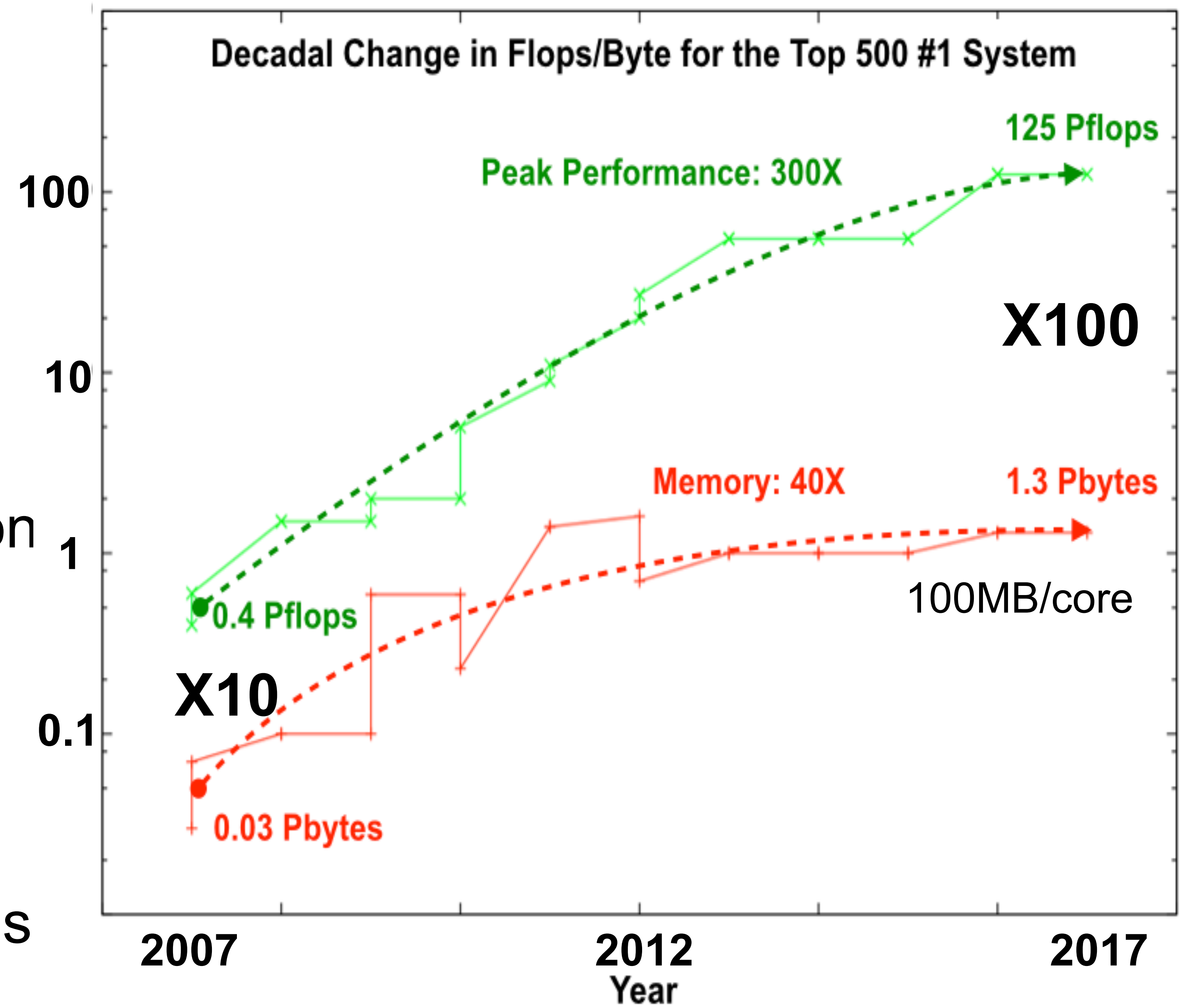
Running Jobs | Queued Jobs | Reservations

Total Queued Jobs: 172

Job Id	Project	Score	Walltime	Queued Time	Queue	Nodes
307941	SkySurvey	8351.7	1d 00:00:00	5d 01:10:03	prod-capability	32768
307942	SkySurvey	8350.5	1d 00:00:00	5d 01:09:42	prod-capability	32768
309793	NucStructReact_2	7069.0	01:00:00	1d 19:13:34	prod-capability	32768
309794	NucStructReact_2	7065.1	01:00:00	1d 19:12:28	prod-capability	32768
309795	NucStructReact_2	7056.8	01:00:00	1d 19:10:04	prod-capability	32768
309271	LatticeQCD_2	6121.1	03:00:00	3d 03:40:34	prod-capability	12288
309314	LatticeQCD_2	5036.1	04:50:00	2d 22:51:59	prod-capability	12288
309315	LatticeQCD_2	5034.8	03:00:00	2d 22:51:38	prod-capability	12288
309316	LatticeQCD_2	5034.0	04:50:00	2d 22:51:24	prod-capability	12288
309317	LatticeQCD_2	5033.0	03:00:00	2d 22:51:08	prod-capability	12288
309318	LatticeQCD_2	5032.6	04:50:00	2d 22:51:01	prod-capability	12288

Where is Computing Headed?

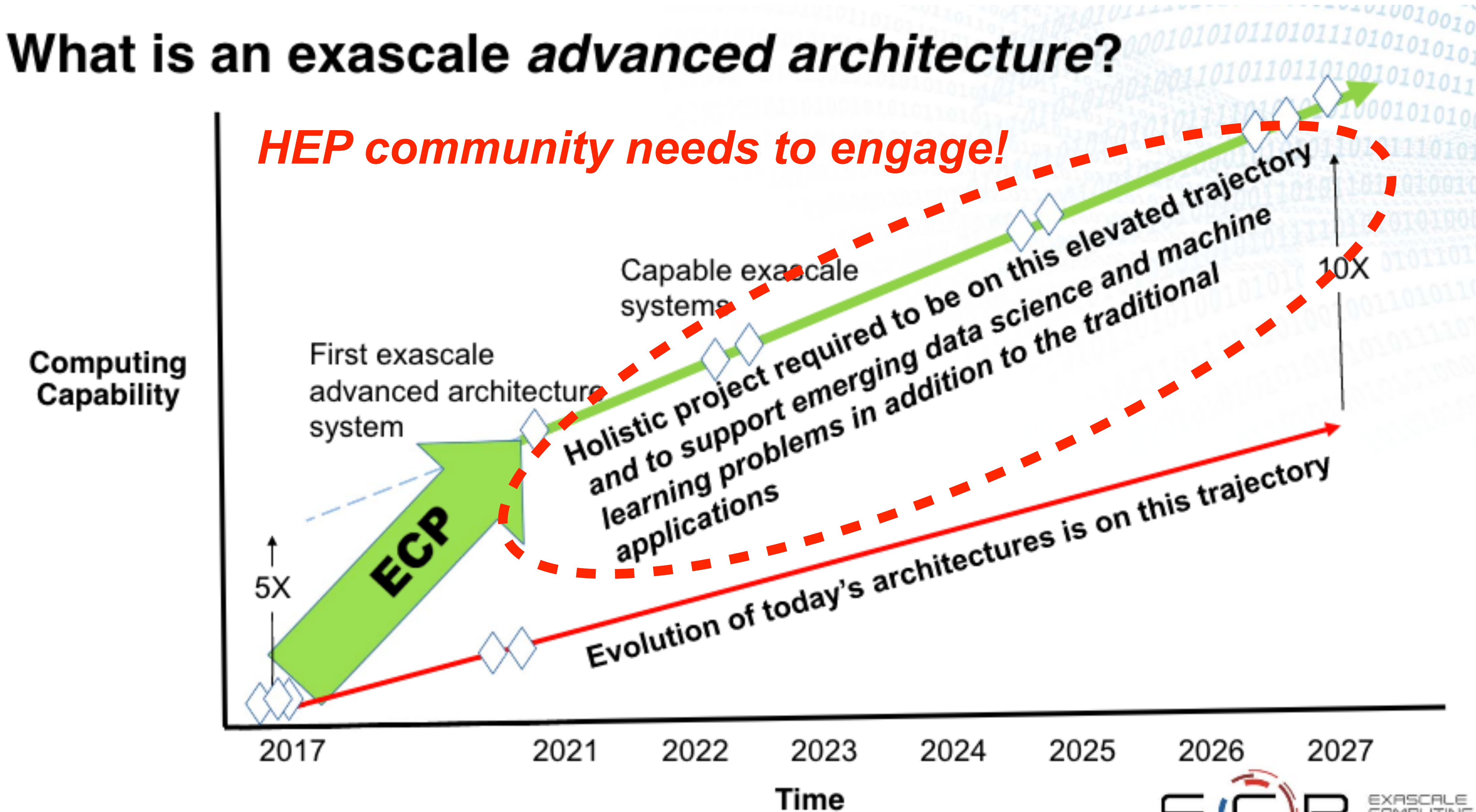
- **Evolution of HPC Systems**
  - Optimized for raw Flops
  - Poor Memory to Flops ratio
  - Poor Comm/IO to Flops ratio
  - Insufficient storage
  - Multiple technology ‘swim lanes’
  - Rapid node architecture evolution
  - Major lag in software development
- **Mitigation Strategies**
  - Rethink computer architecture and design for science use cases
  - Storage caches with direct connectivity to compute nodes
  - Faster/fatter data pipes to compute platforms
  - Software strategies for portability



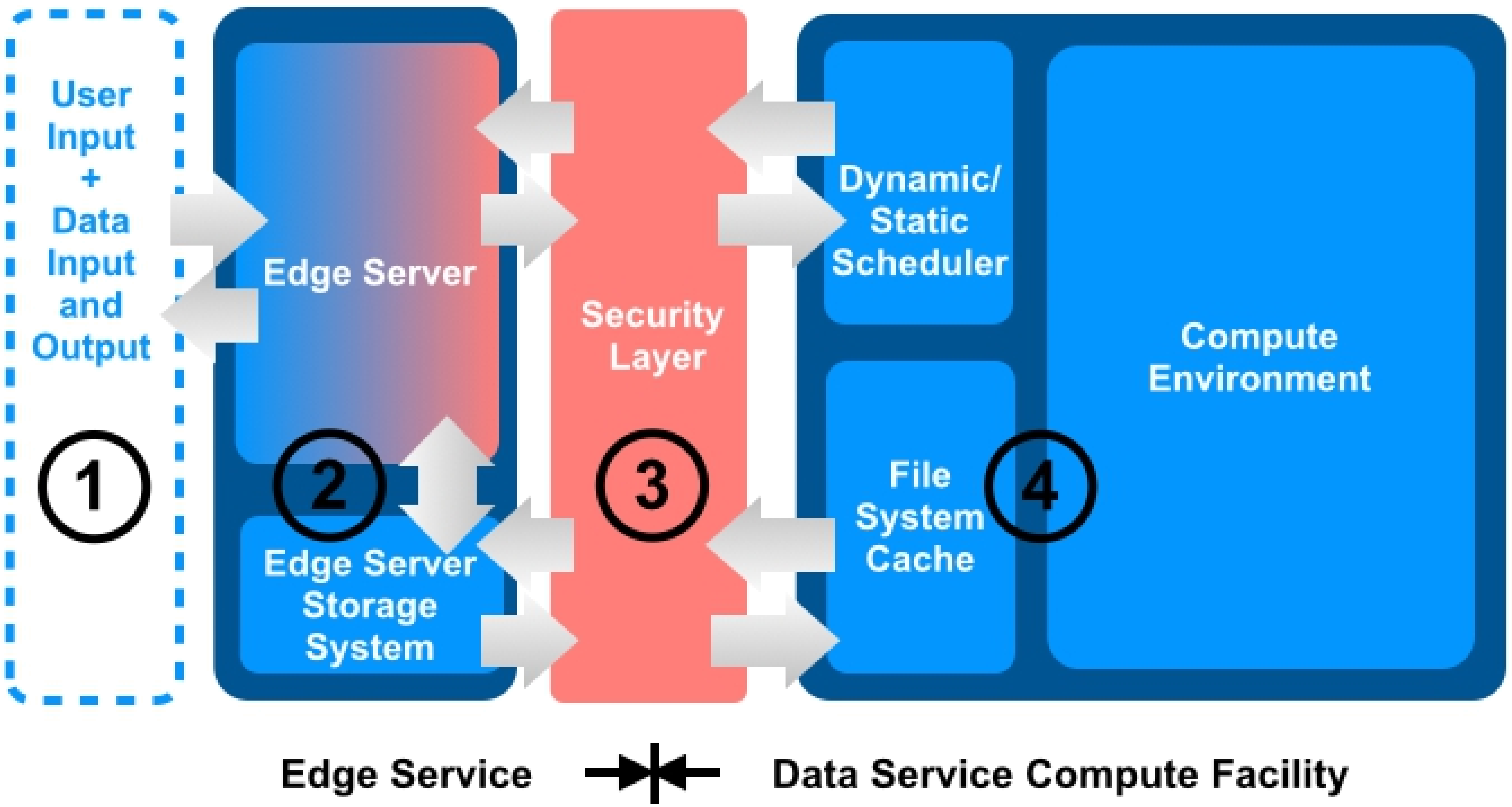
Example of current supercomputer evolution: driven by a number of imperatives — economic and technological — leading to specialized nodal architectures (end of the ‘PC’ model)

Exascale Computing Project

Major DOE SC and NNSA joint project to arrive at a scientifically usable architecture for exascale computing in the early 2020's — *largest science project within DOE*



Connectivity Example: Edge Services



Edge service design must consider a number of factors; security, resource flexibility, interaction with HPC schedulers, external databases, requirements of the user community — modern supercomputers are once again ‘strategic’ resources, not a ‘pile of PCs’!

Boundary Conditions

- **What's the Problem?**
  - Even if solutions can be designed ***in principle***, the resources needed to implement them are (usually) not available
  - ***Despite all the evidence of its power***, computing still does not get high enough priority compared to building "things"
  - In part this is due to the success of computing — progress in this area is usually much faster than in others, so one can assume that ***computing will just happen (Moore's Law)*** — to what extent is this still true?
- **Large-Scale Computing Available to Scientists**
  - Lots of supercomputing (HPC) available and more on the way
  - Not enough data-intensive scalable computing (DISC) available to users, hopefully this will change over time
  - Publicly funded HTC/Grid computing resources cannot keep pace with demand
  - Commercial space (Cloud) may be a viable option but is not issue-free
  - Storage, networking, and curation are major problems (***sustainability***)

“Data Meets HPC” — Basic Requirements

- **Software Stack:** Ability to run arbitrarily complex software stacks on HPC systems (***software management***)
- **Resilience:** Ability to handle failures of job streams, still rudimentary on HPC systems (***resilience***)
- **Resource Flexibility:** Ability to run complex workflows with changing computational ‘width’, possible but very clunky (***elasticity***)
- **Wide-Area Data Awareness:** Ability to seamlessly move computing to the data (and vice versa where possible); access to remote databases and data consistency via well-designed and secure edge services (***integration***)
- **Automated Workloads:** Ability to run large-scale coordinated automated production workflows including large-scale data motion (***global workflow management***)
- **End-to-End Simulation-Based Analyses:** Ability to run analysis workflows on simulations using a combination of in situ and offline/co-scheduling approaches (***hybrid applications***)

Summary

- **Is HPC the solution we have been waiting for?**
  - Not quite, but —
  - It might be a solution we can live with (provided software upgrades are doable and straitjacketing is acceptable)
  - It might be a (partial) solution we will *have* to live with (power, funding)
- **Compute/data model evolution**
  - What happens when compute is free but data motion and storage are both expensive?
  - Investment in appropriate networking infrastructure and storage
  - Major refactoring of software, especially where the computational payload meets the compute platform
- **Will require nontraditional cross-office agreements**
  - Individual experiments too fine-grained, need a higher-level arrangement
  - Will require changes in ASCR's computing vision ("superfacility" variants)
  - ASCR is not a "support science" office, prepare for the bleeding edge!