Argonne NATIONAL LABORATORY

Future Trends in Nuclear Physics Computing, May 4, 2017

# Next-Generation Computing Challenges: HPC Meets Data


**Salman Habib**

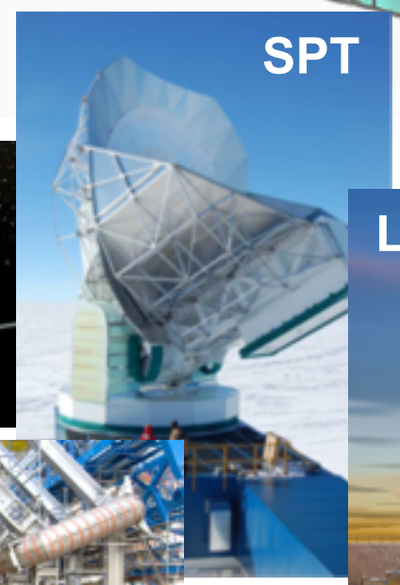
High Energy Physics Division
Mathematics & Computer Science Division
Argonne National Laboratory

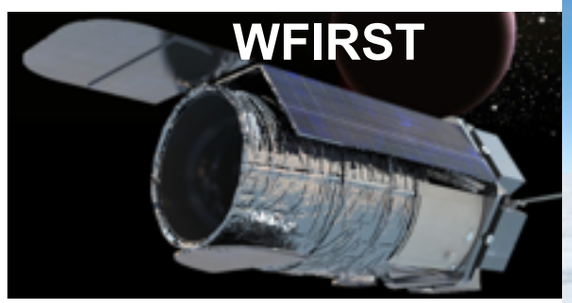
Computation Institute
Argonne National Laboratory
The University of Chicago

Kavli Institute for Cosmological Physics
The University of Chicago







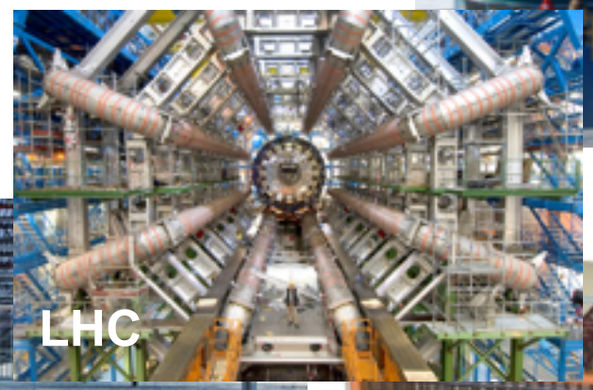





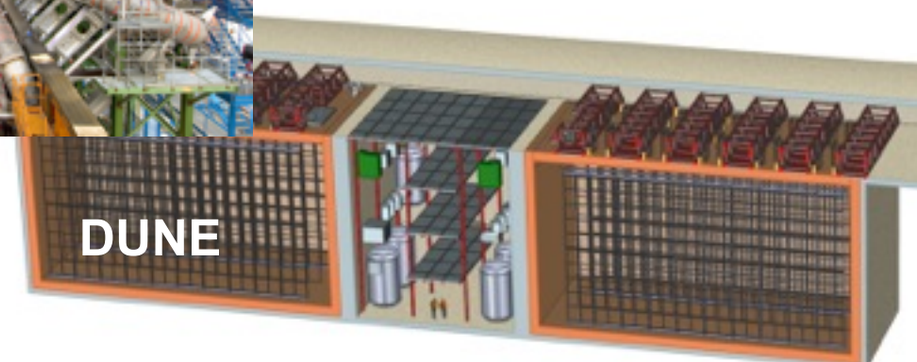
EXASCALE COMPUTING PROJECT

**HEP-CCE**

SciDAC
Scientific Discovery through Advanced Computing








NSF NASA

# Supercomputers: A Personal Historical Sample (~25 years)


TMC CM-200


Cray T3-D


Cray T3-E


Cray XT-4


IBM BG/P


TMC CM-5

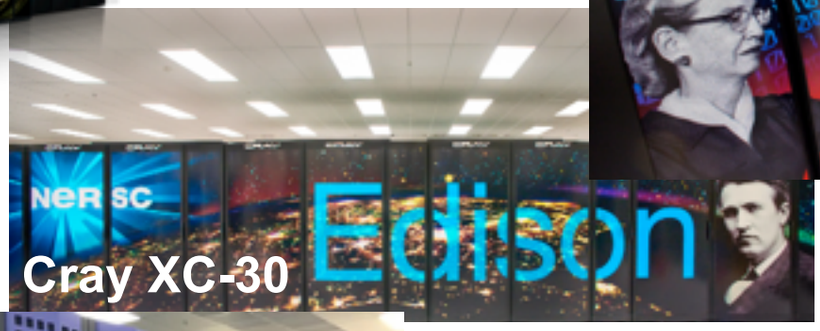

SGI O2000


IBM Roadrunner


IBM SP-2

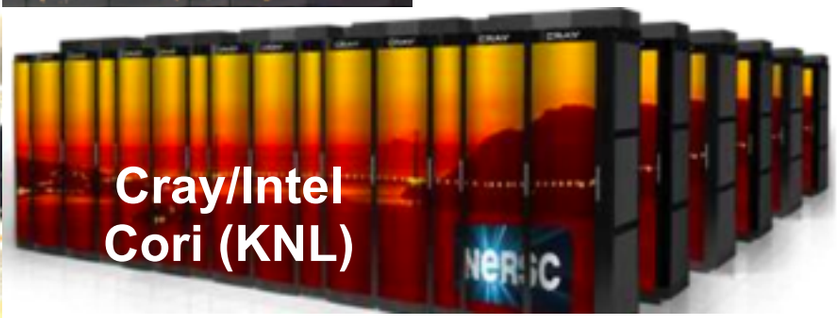

Compaq ASCI ‘Q’



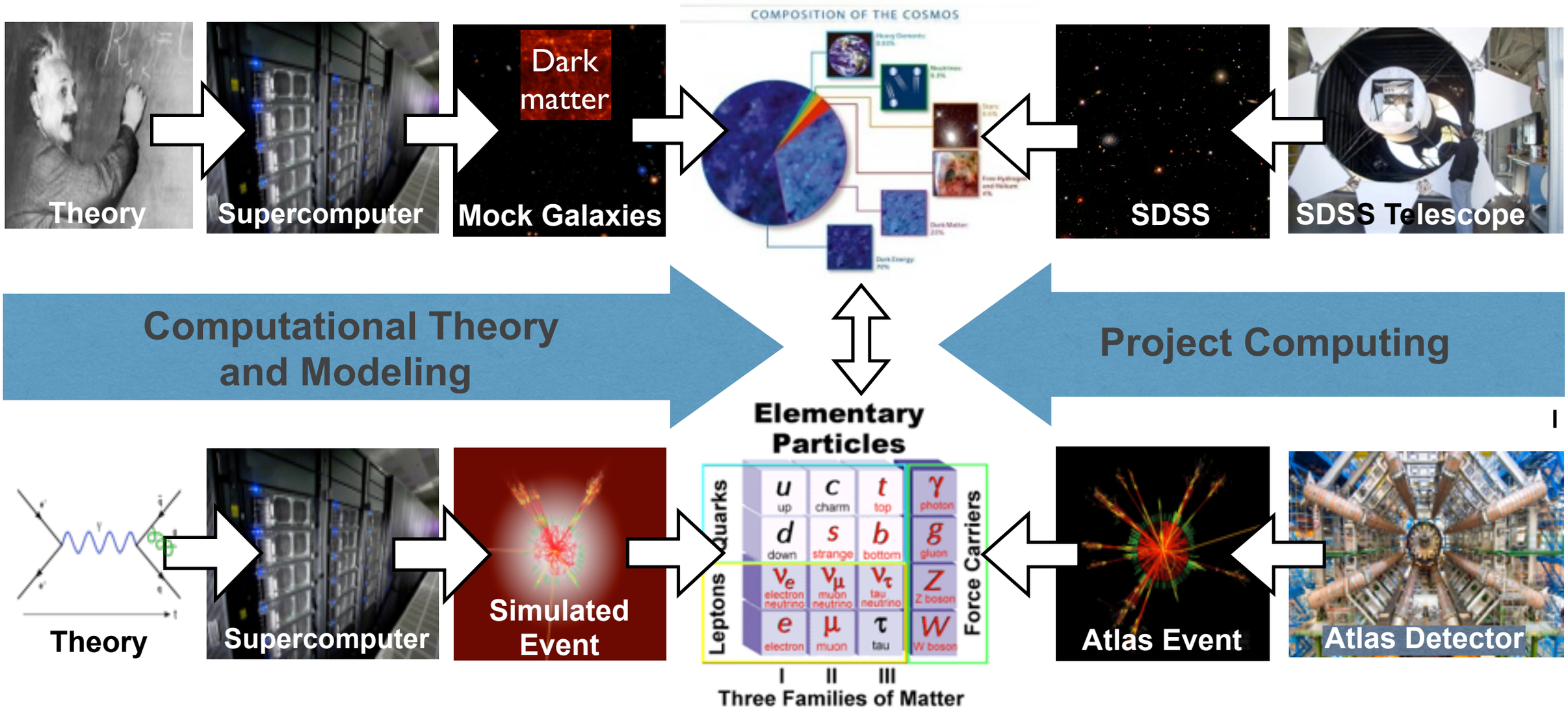



Cooley (CPU/GPU)



# Computing Paradigm (Cosmic and Energy Frontiers)

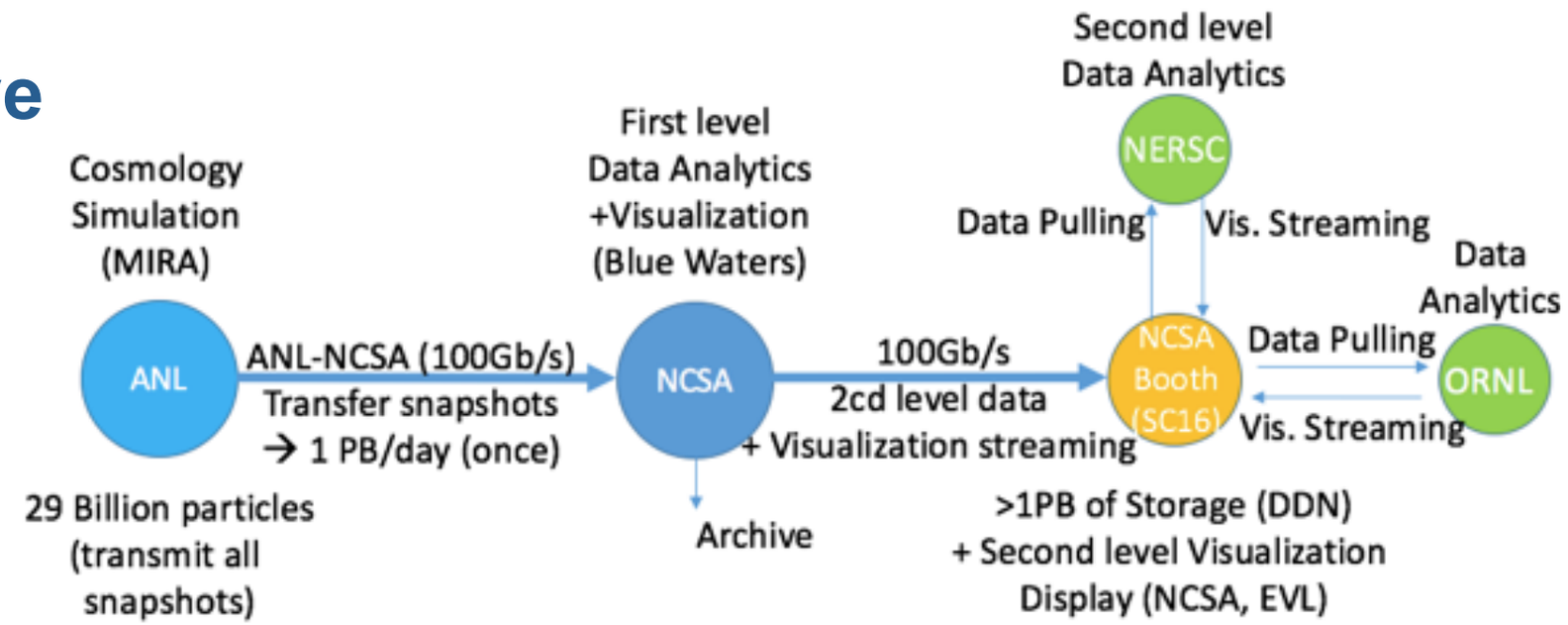
**Simulated Data:** 1) Large-scale simulation of the Universe, 2) Synthetic catalogs, 3) Statistical inference (cosmology); **Analysis:** Comparison with actual data

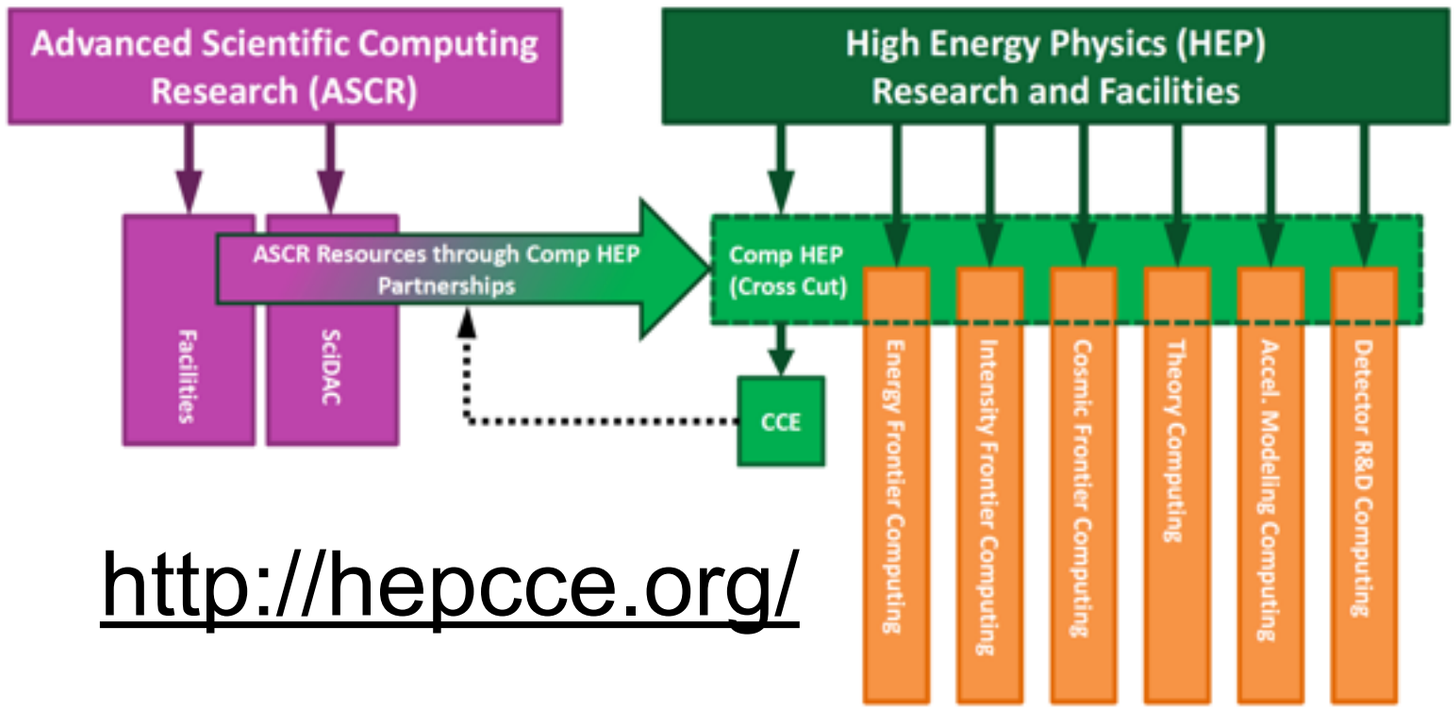



**Simulated Data:** 1) Event generation (lists of particles and momenta), 2) Simulation (interaction with detector), 3) Reconstruction (presence of particles inferred from detector response); **Analysis:** Comparison with actual data

# What this Talk Tries to Cover —

- **HPC meets Data-Intensive Computing**
  - **Cosmology context**
  - **HPC systems as data sources and sinks**
  - **Personal experience, provide reality check**
  - **DOE HEP response, joint work with ASCR — HEP-CCE**
- **Hope is to provide some general lessons that may possibly be useful to NP**
- **Suggestion: Explore possible HEP-NP connections via HEP-CCE**




http://hepcce.org/

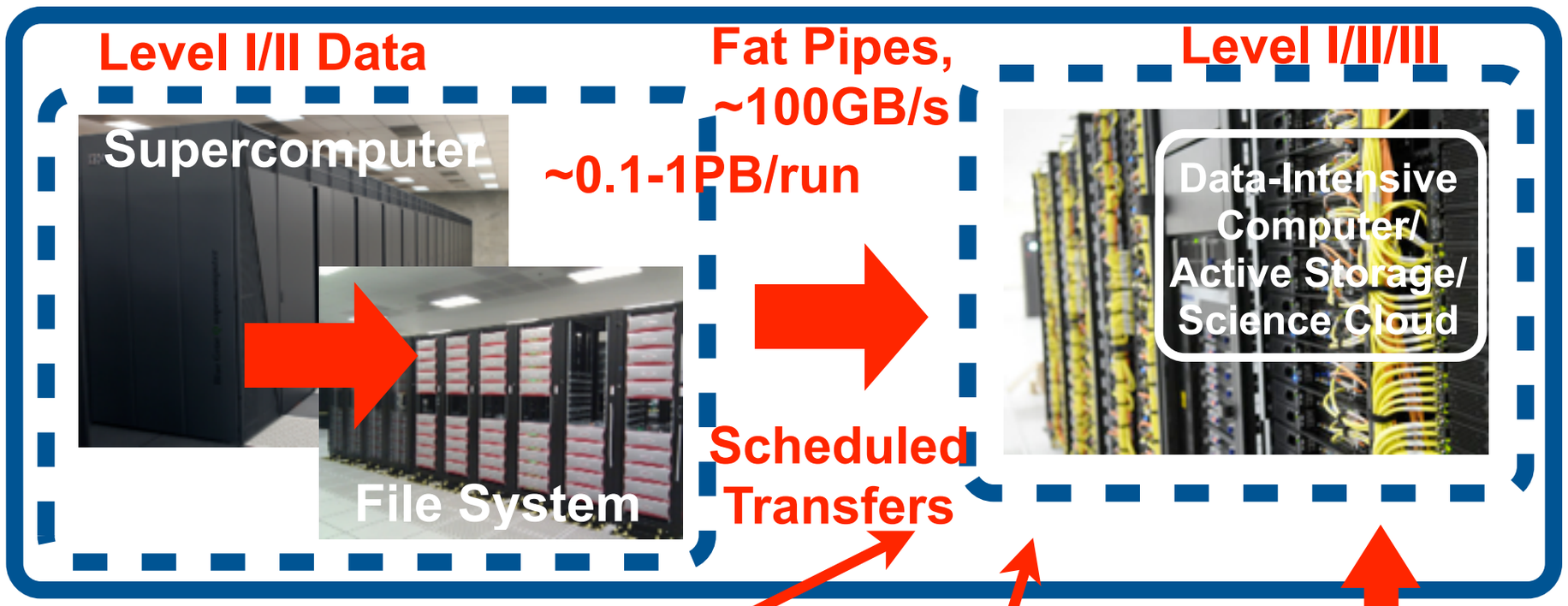
From Jim Siegrist's talk at ICHEP 2016

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

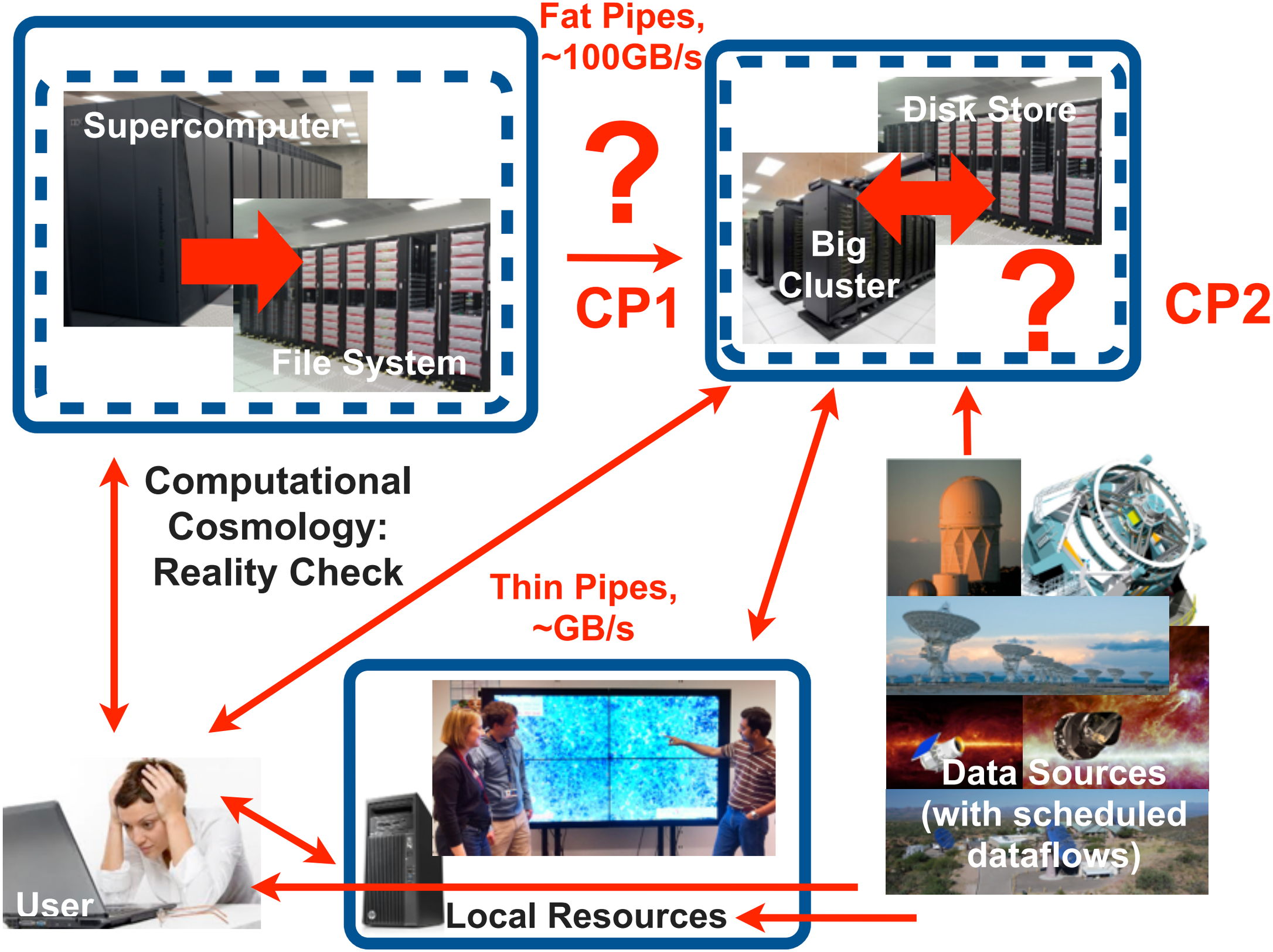
**Want more of this — ("Science Cloud"), but don't yet (really) have it (Data-Intensive Scalable Computing: DISC)**

# HPC + DISC Future: Desired Outcome



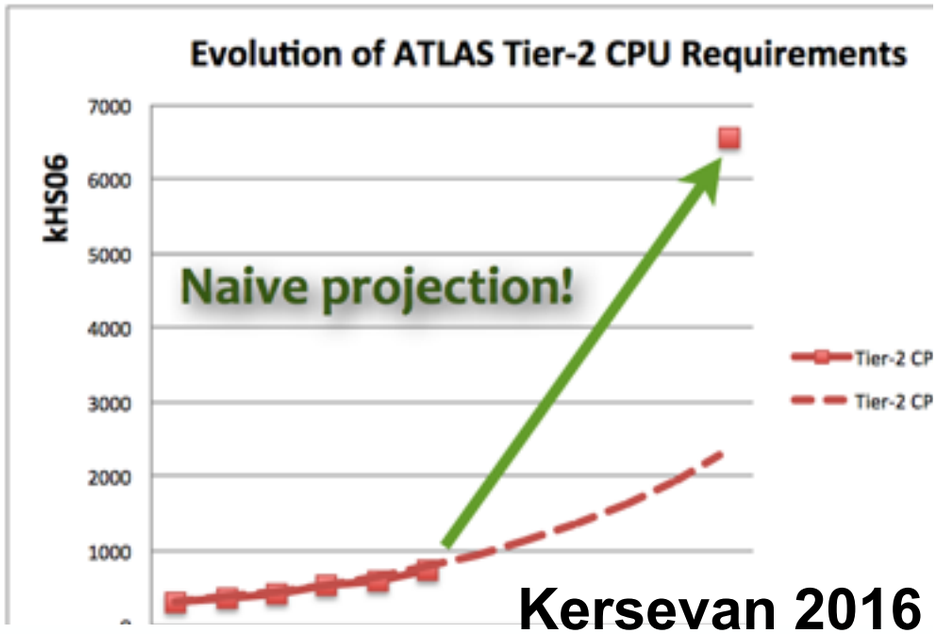

**Computational Cosmology: 'Ideal' Use Case**

# Reality — Very Much a Work in Progress

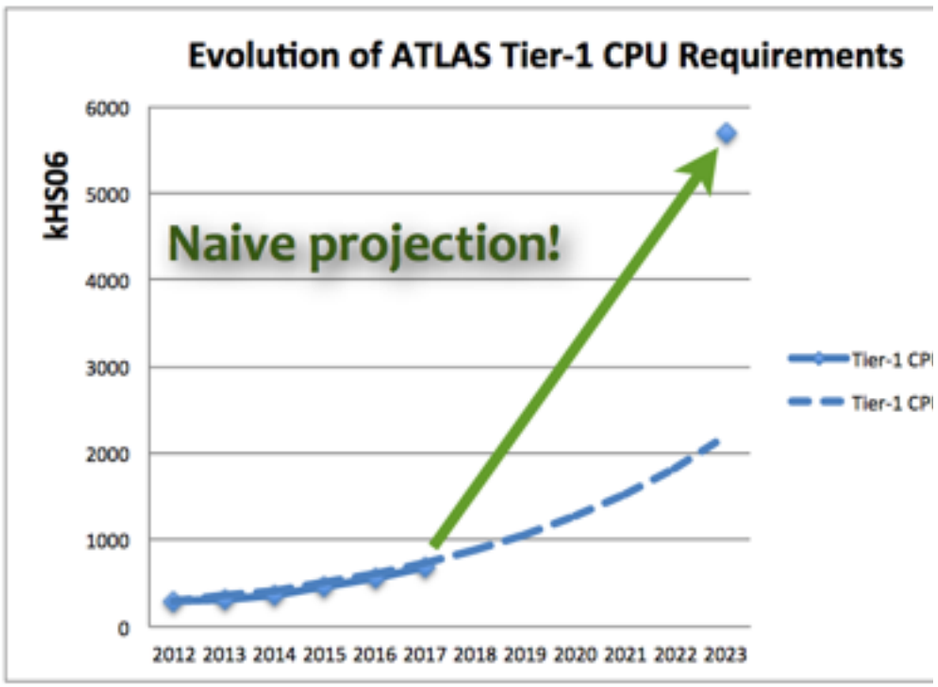


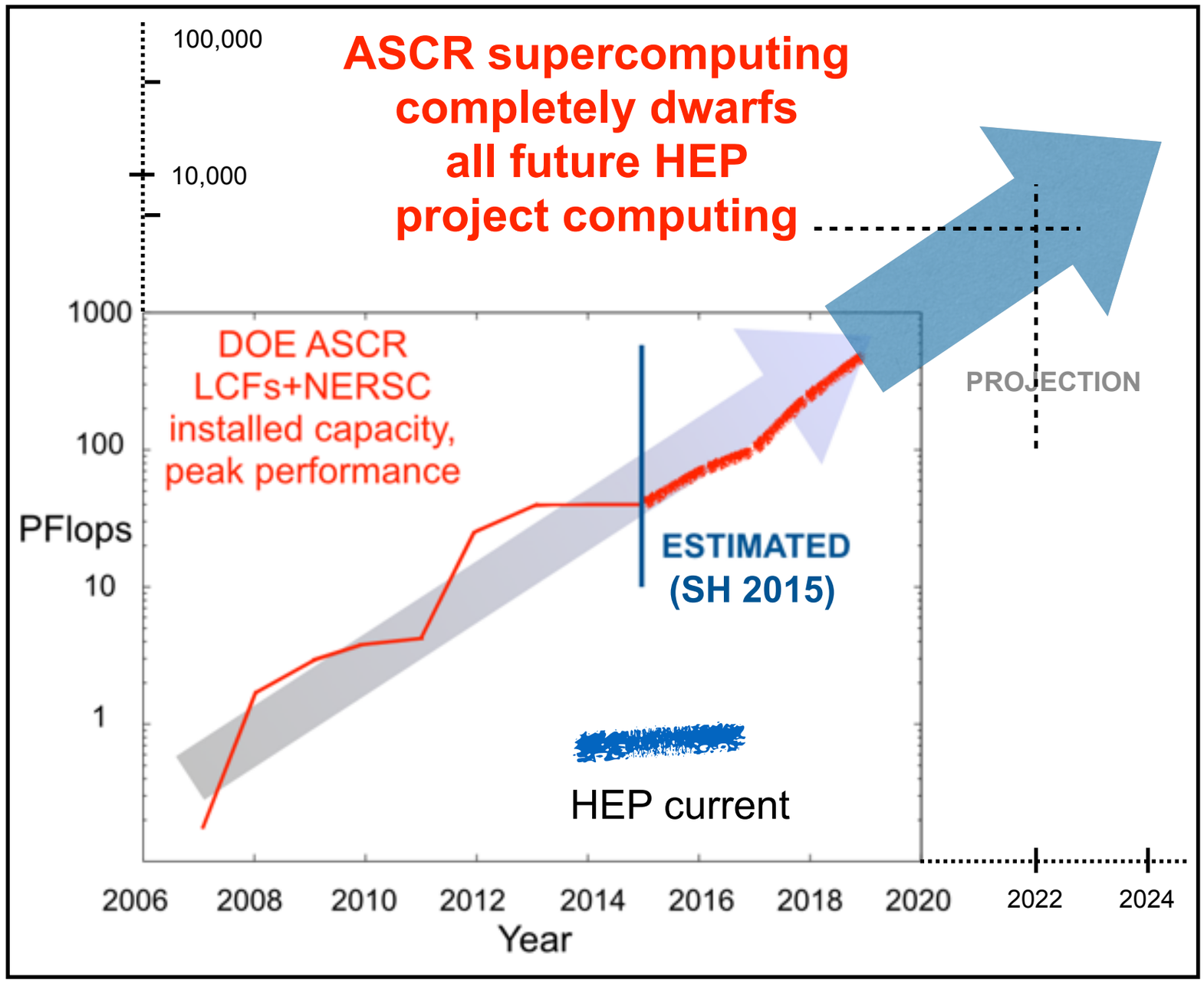
# HEP Computing Requirements for 'Energy Frontier'

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




Kersevan 2016



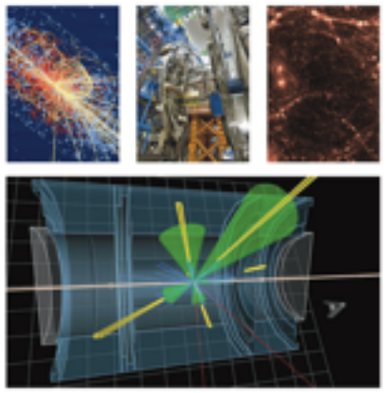
# Many White Papers and Reports —

http://hepcce.org/files/2016/11/DOE-ExascaleReport-HEP-Final.pdf

HEP

HIGH ENERGY PHYSICS

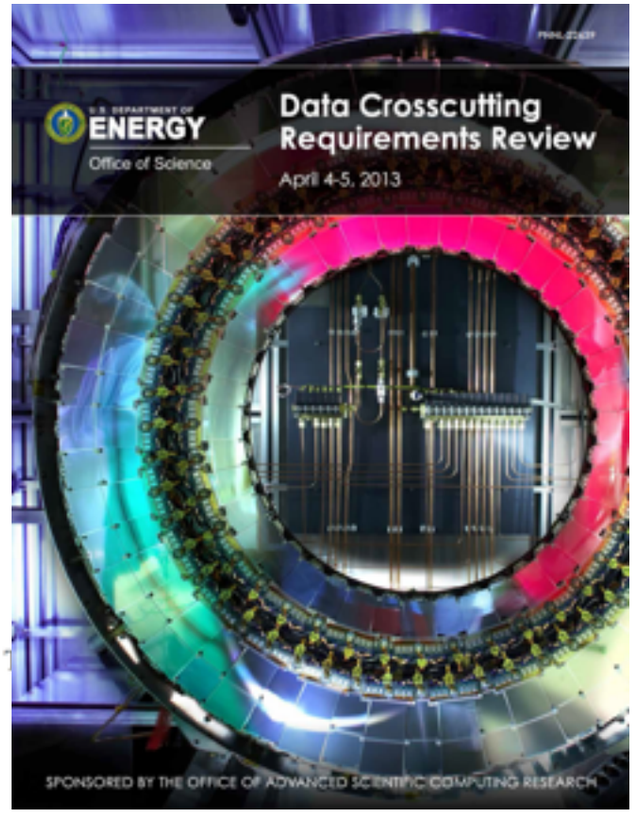
EXASCALE REQUIREMENTS REVIEW



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] [illegible]
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)

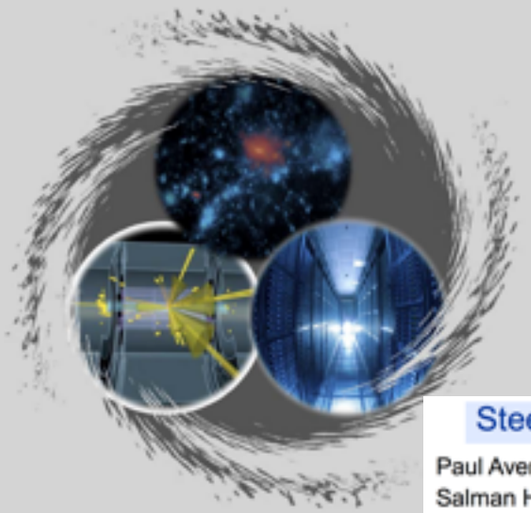
**Applications Software Leads:** Tom LeCompte[1], Zach Marshall[3]
**Software Libraries and Tools Leads:** Anders Borgland[4], Brett Viren[5]
**Systems Lead:** Peter Nugent[3]

**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**

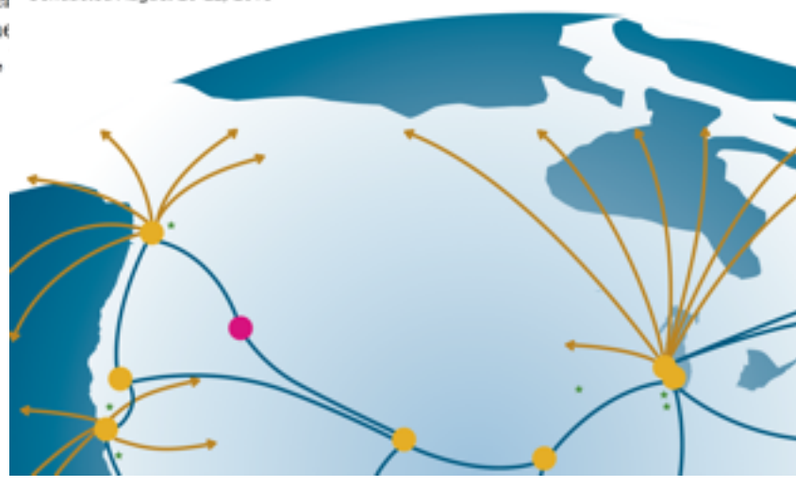
**L. A. T. Bauerdick, S. Gottlieb**, G. Bell, K. Bloom, T. Blum, D. Brown, M. Butler, E. Cormier, P. Elmer, M. Ernst, I. Fisk, G. Fuller, R. Gerber, S. Habib, M. Hildreth, S. Hoeche, C. Joshi, A. Mezzacappa, R. Mount, R. Pordes, B. Rebel, L. Reina, M. C. Sanchez, J. Shank, A. Szalay, R. Van de Water,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

## Chapter 9: Computing

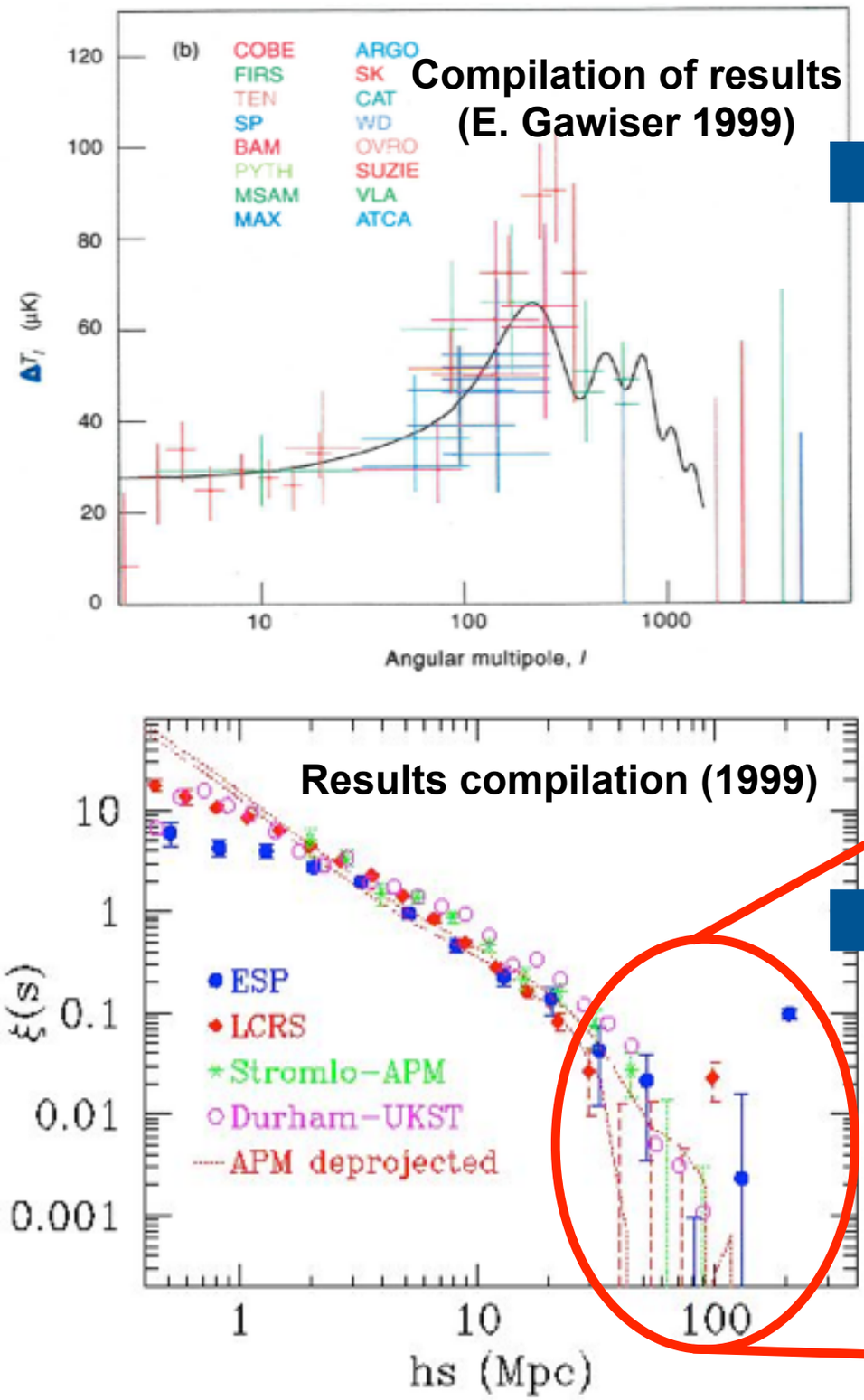
**High Energy Physics and Nuclear Physics Network Requirements**

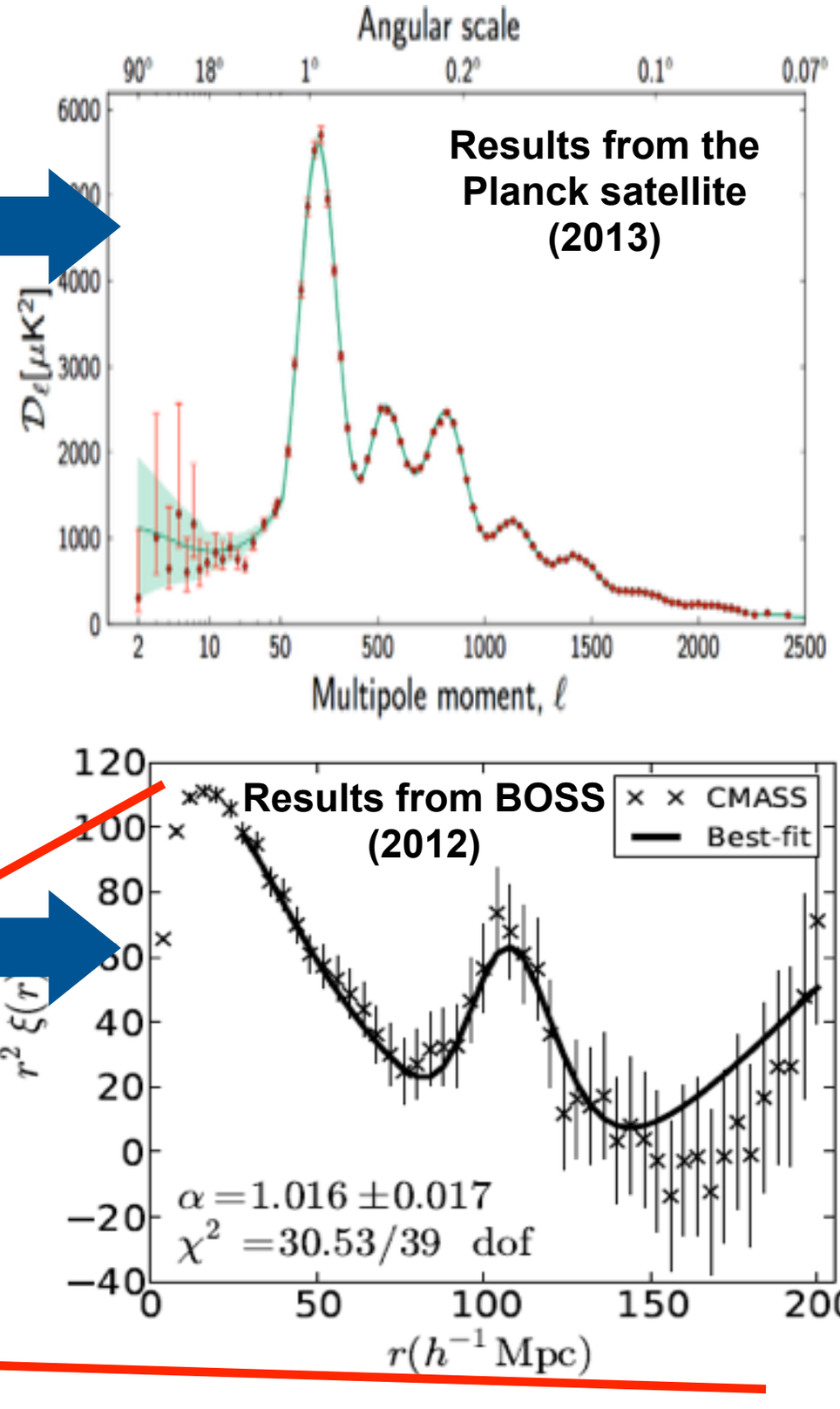
HEP and NP Network Requirements Review
Final Report

*Conducted August 20-22, 2013*



# Back to the Universe: Science Drivers

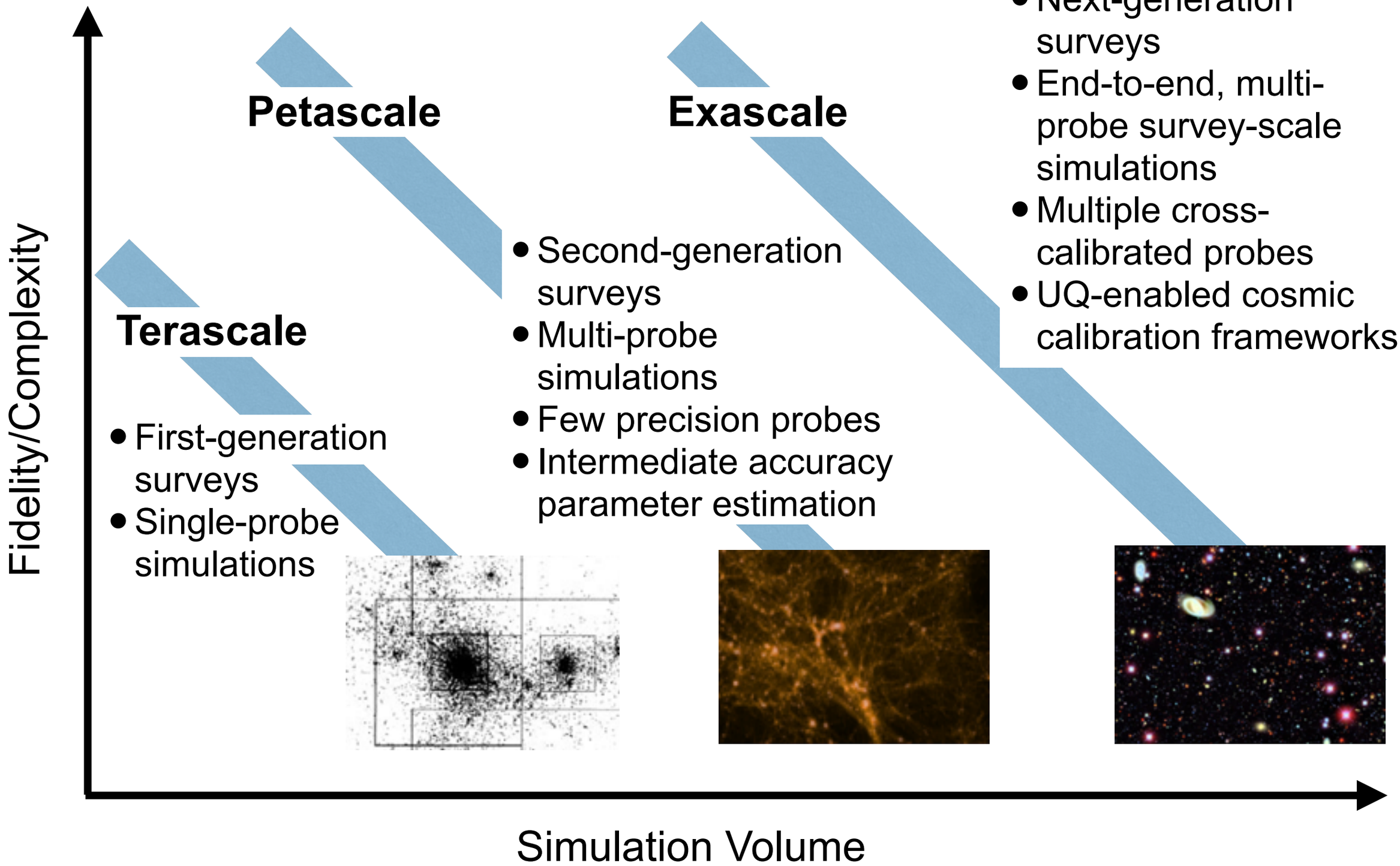



- Massive increase in sensitivity of cosmic microwave background (CMB) observations
- Cross-correlation with galaxy surveys
- New era of CMB modeling/simulations

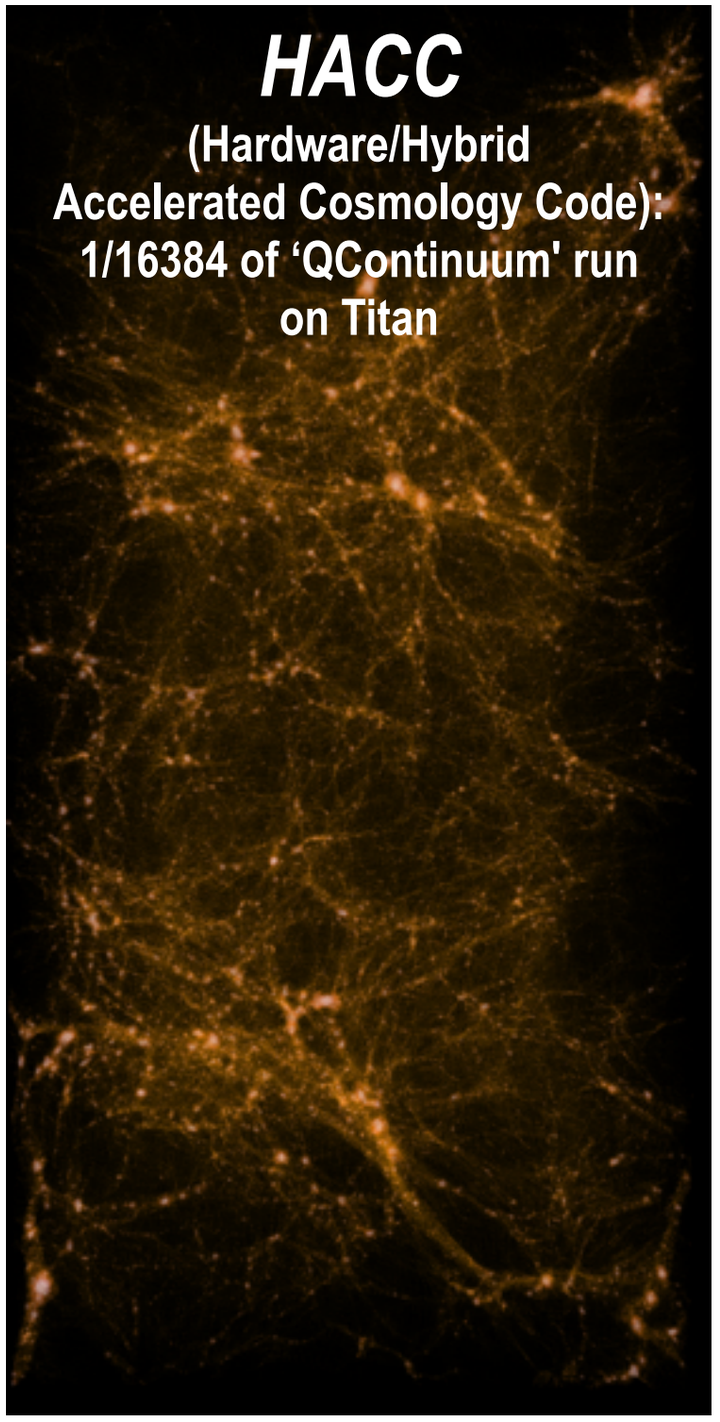
- Massive increase in volume of galaxy surveys
- Next-generation galaxy clustering simulations
- Multi-physics codes needed to meet accuracy requirements

# Precision Cosmology: Simulation Frontiers



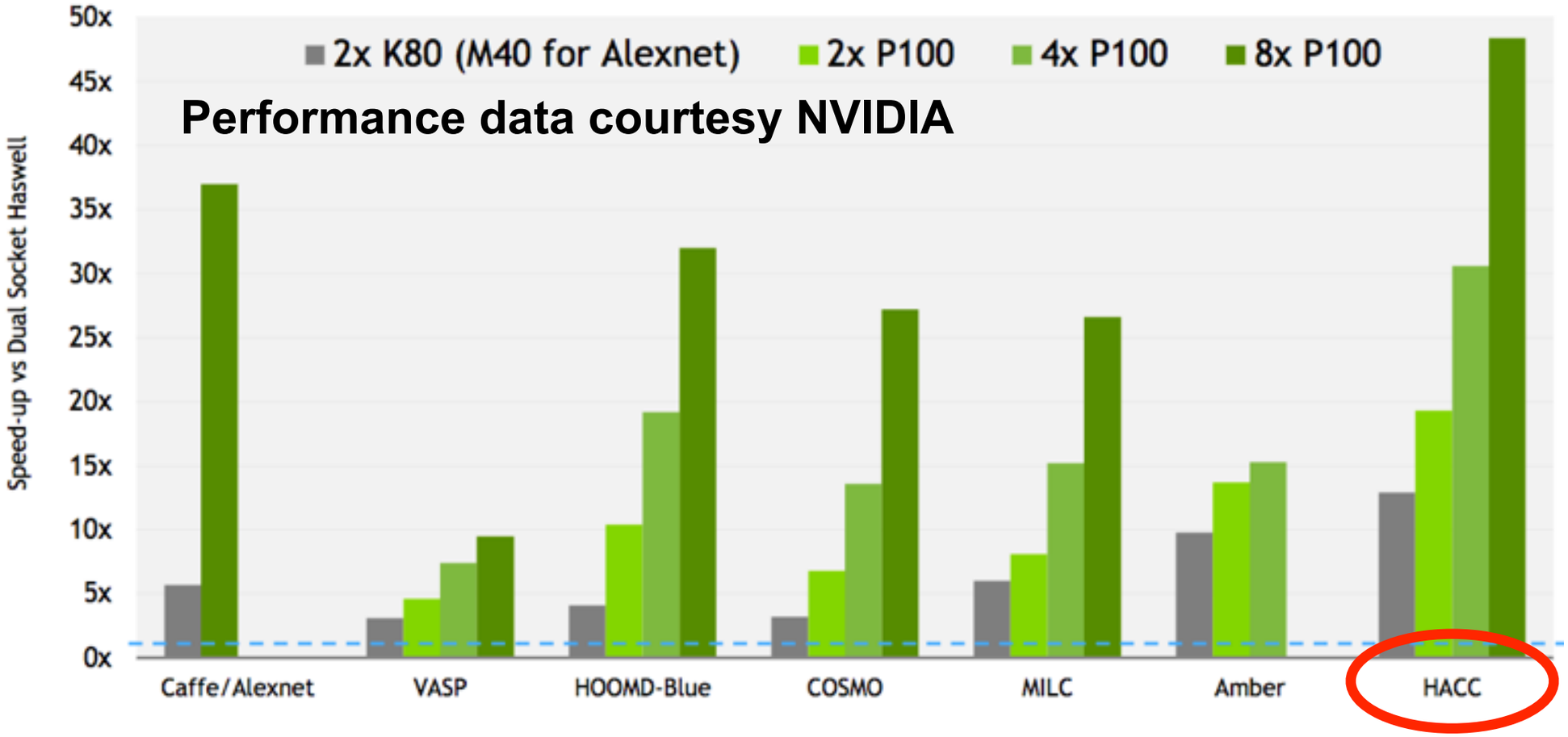
# Cosmology Simulation: HACC Framework (PIC+PP+Hydro)

- **High-Order Spectral Particle-Mesh:** Short-range forces tuned by spectral filters, high-accuracy polynomial fits, custom 3D FFT
- **Particle Overloading:** Particle replication at ‘node’ edges reduces communication, eases “soft portability” design
- **Performance Focus:** Aim for high absolute performance on all platforms, C++/MPI + ‘X’ programming model, first production science code to cross 10PFlops sustained
- **Task-Based Load Balancing:** Transfer of work packages using overloading concept
- **Flexible Chaining Mesh:** Optimizes tree/P3M methods
- **Optimized Force Kernels:** Very high compute intensities, use of mixed precision
- **Adaptive Time-Stepping:** Sub-cycling of short-range time-steps, adaptive time-stepping at the individual particle level
- **Custom Parallel I/O:** Topology-aware parallel I/O with lossless compression (GenericIO)
- **CCRK-SPH Hydro:** New hydrodynamics capability underway
- **Analysis:** CosmoTools library (in situ/co-scheduled/offline)




*Habib et al., New Astron. 42, 49 (2016)*

# HACC on Pascal and KNL




## $512^3$: 64 cores, 4 nodes of BG/Q, 1 node of KNL

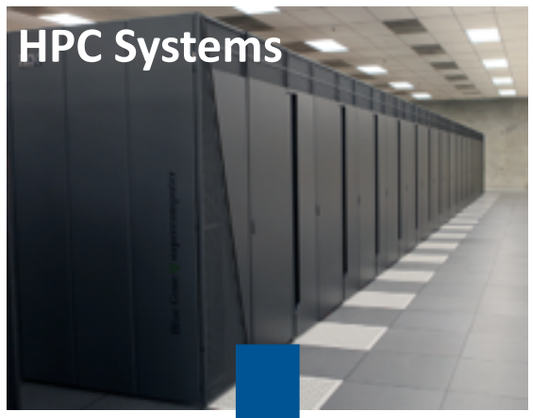
| Cores | RPN | OMP | TH | BG/Q Time, s | KNL B0, cache mode Time, s | KNL B0, flat mode Time, s | Ratio |
|---|---|---|---|---|---|---|---|
| 64 | 16 | 4 | 64 | 4542 | 678.7571 | 678.2269 | 6.69 |
| 64 | 16 | 8 | 128 | 2823 | 606.1815 | 609.2007 | 4.66 |
| 64 | 16 | 16 | 256 | 2556 | 587.2716 | 587.4443 | 4.35 |
| 64 | 32 | 2 | 64 | 4747 | 620.7261 | 621.2356 | 7.65 |
| 64 | 32 | 4 | 128 | 2824 | 536.1650 | 534.9907 | 5.27 |
| 64 | 32 | 8 | 256 | **2503** | **503.0927** | 501.8637 | 4.98 |
| 64 | 64 | 4 | 256 | 2539 | 510.3745 | 506.7107 | 4.98 |

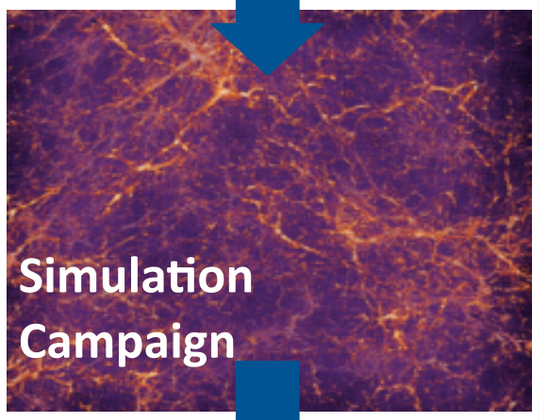
# Exascale Cosmology: 'Big Data' Meets Supercomputing

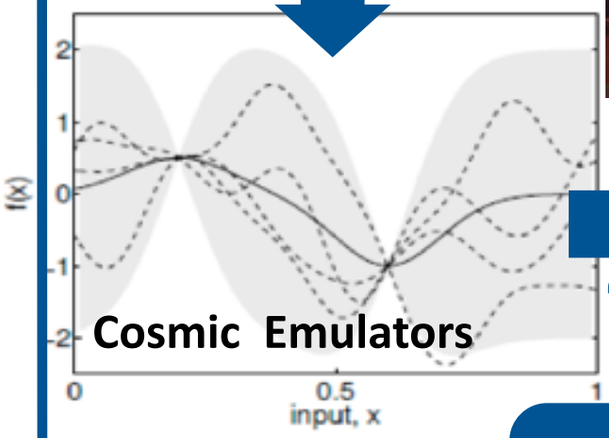
**Supercomputer simulation campaigns**

**Statistics + machine learning + optimization methods**

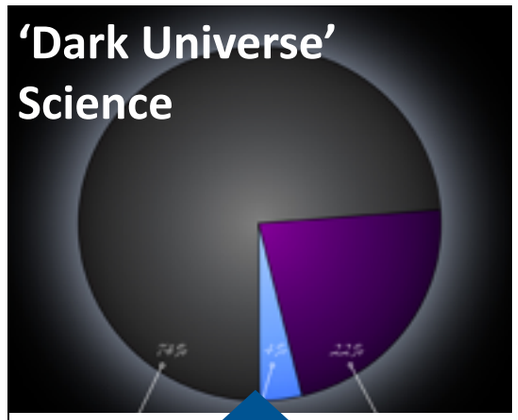
**Emulator based on Gaussian process interpolation in high-dimensional spaces**



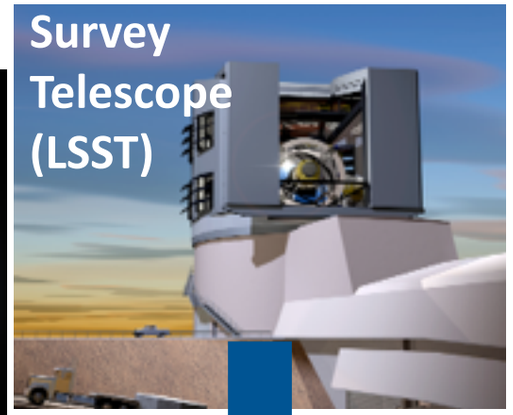



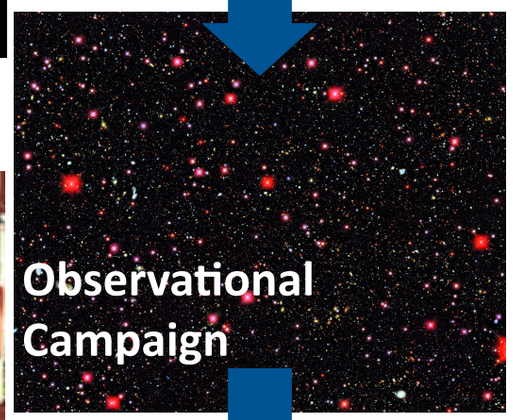

## Cosmic Calibration

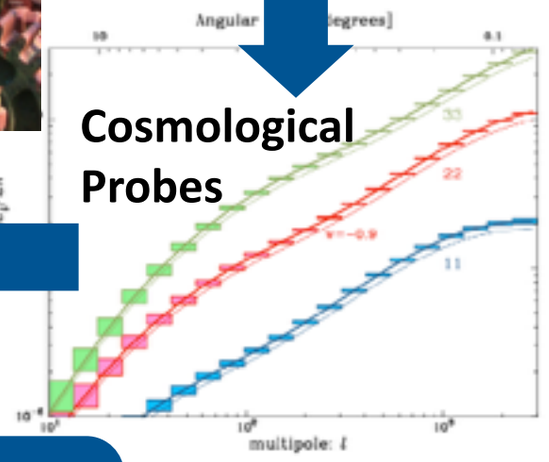




'Precision Oracle'




**Science with Surveys: HPC meets Big(ish) Data**

**Mapping the sky with multiple survey instruments**

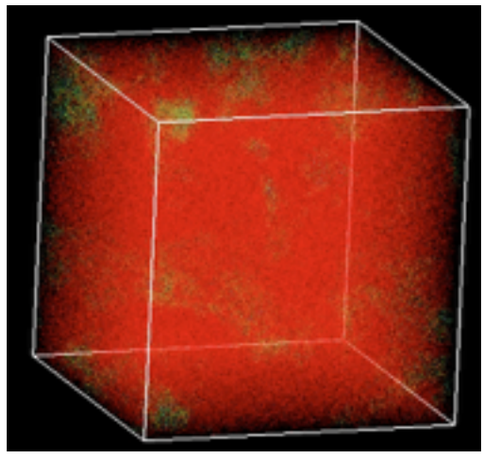
**Extraction of summary statistics from survey sky maps**

**Observations: Statistical error bars very small, systematics dominate**

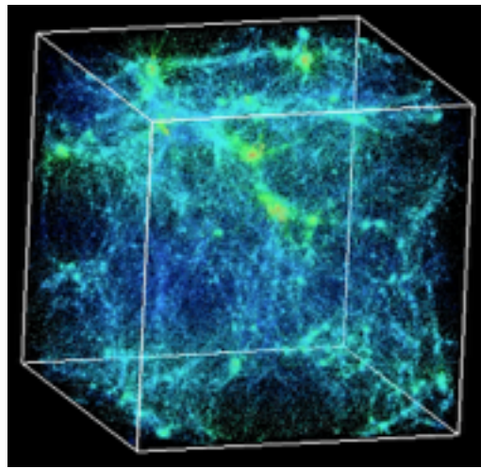
**Heitmann et al. 2006, Habib et al. 2007, Higdon et al. 2010, etc. etc.**

# Exascale Analytics/Workflow Complexity

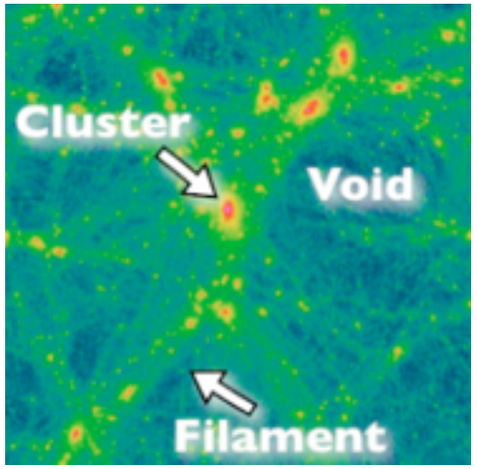
Early Universe Initial Conditions



High-Resolution Cosmological Simulation



Multiple Outputs Halo/Sub-Halo Identification




Halo Merger Trees

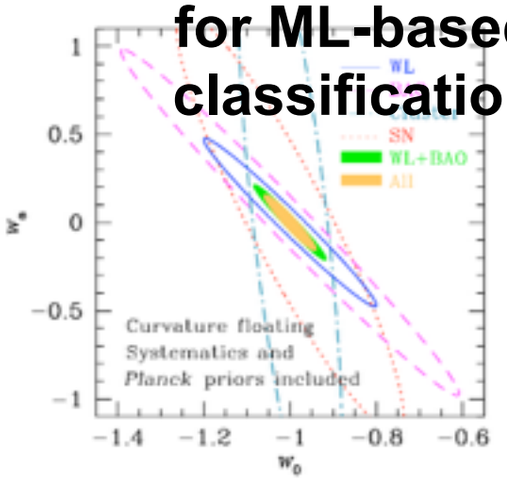
Semi-Analytic or Empirical Galaxy Modeling

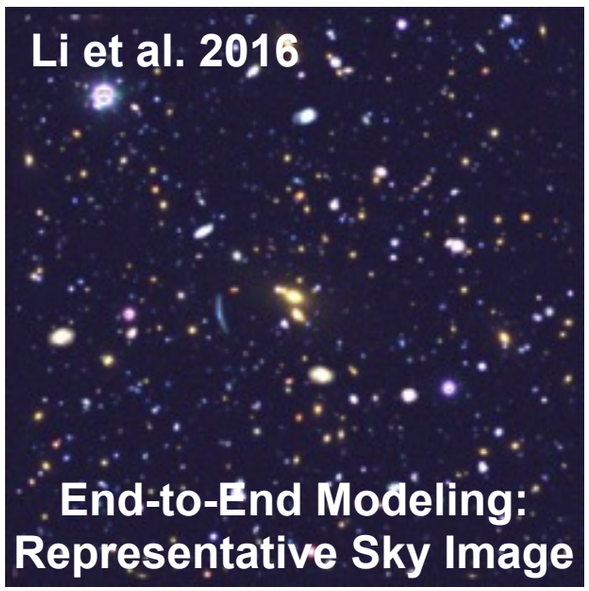
Galaxy Catalog

Realistic Image Catalog

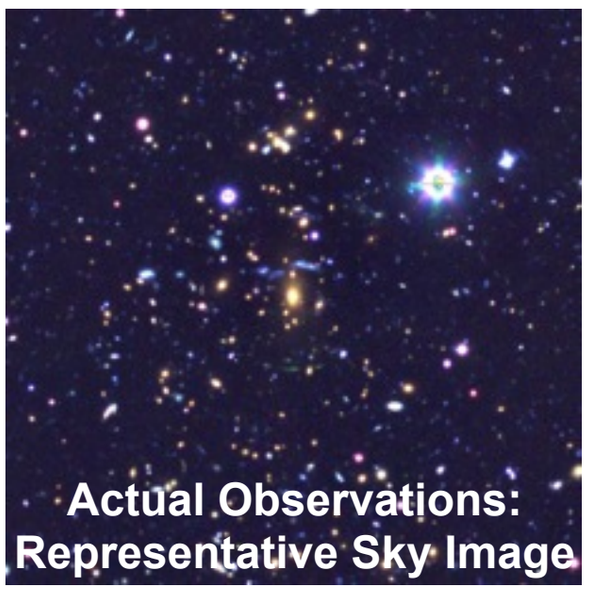
Atmosphere and Instrument Modeling

Millions of images for ML-based classification




End-to-End Modeling: Representative Sky Image



Actual Observations: Representative Sky Image

Data Management Pipeline

Data Analysis Pipeline

Scientific Inference Framework

# HPC and Data Science — A Difficult Marriage?

- **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

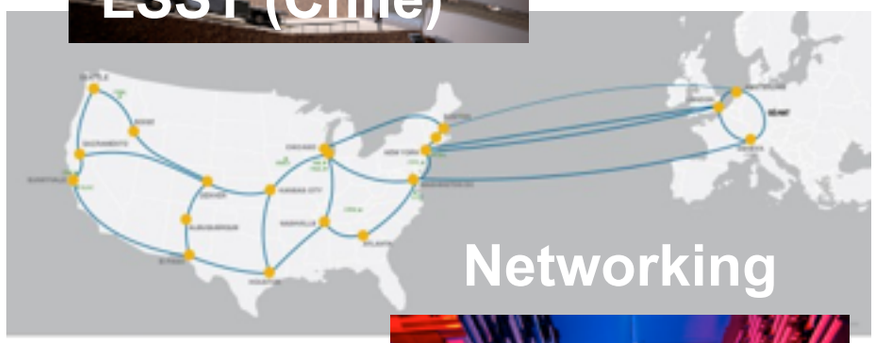
**Total Queued Jobs:** 172

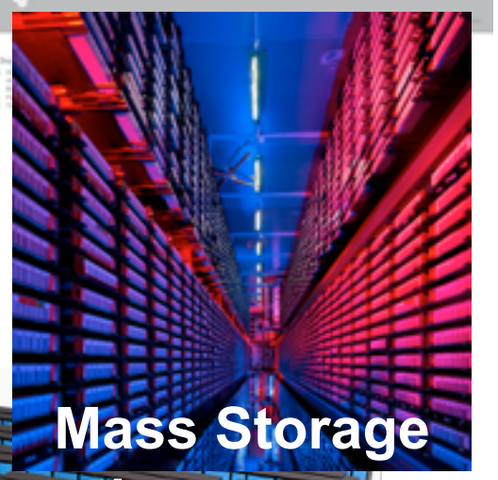
| Job Id | Project | Score | Walltime | Queued Time | Queue | Nodes |
|---|---|---|---|---|---|---|
| 307941 | SkySurvey | 8351.7 | 1d 00:00:00 | 5d 01:10:03 | prod-capability | 32768 |
| 307942 | SkySurvey | 8350.5 | 1d 00:00:00 | 5d 01:09:42 | prod-capability | 32768 |
| 309793 | NucStructReact_2 | 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 |

# Scientific Data and Computing: 'Geography'

- **Optimal Large-Scale Efficiency**
  - Desire data and computing in the same place, but — for a number of reasons — often not ***realistic***
- **Optimal Usability**
  - Mix of small/medium/large-scale computing, data, and network resources, but often not ***affordable***
- **Real-World Issues**
  - Distributed ownership of data, computing, and networking creates ***policy barriers***
  - ***Lack of shared priorities*** across owners
  - Multiple use case ***collisions***: hard to optimize at the system level
  - Funding ***politics*** creates and (sometimes) stabilizes nonoptimal 'solutions' (top-down does ***not*** work)
  - Noodling around with data is ***not science***
- **Practical Response**
  - Make things better, but ***not unrealistically better***






# 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*** — 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

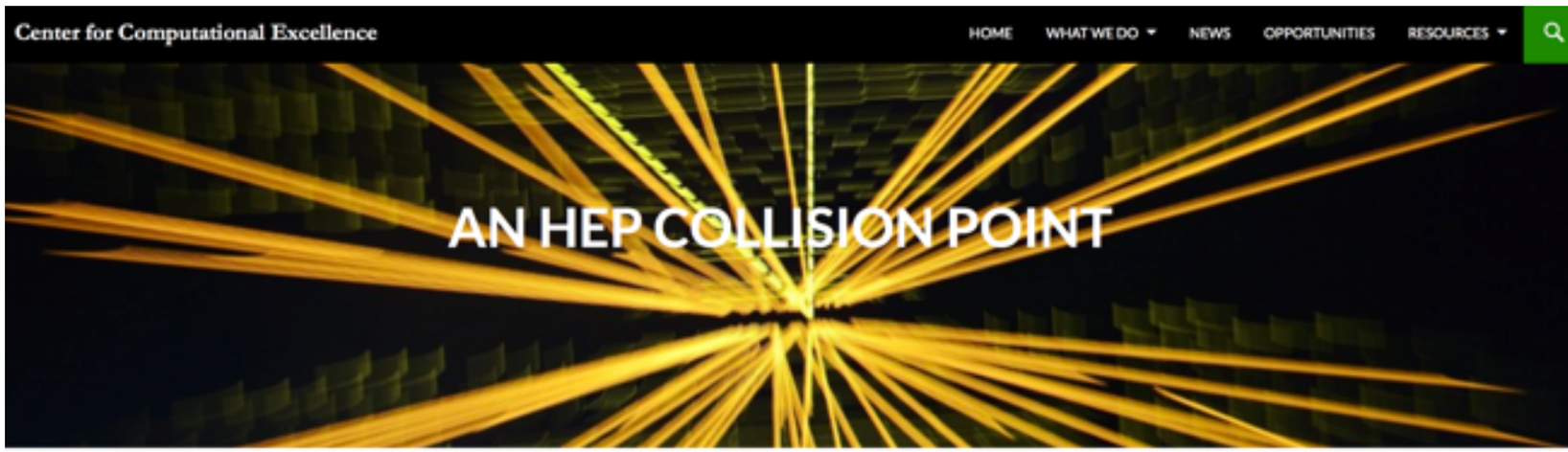
- **Software Stack:** Ability to run arbitrarily complex software stacks on HPC systems (***software management***)
- **Resilience:** Ability to handle failures of job streams, still rudimentary on HPC systems (***resilience***)
- **Resource Flexibility:** Ability to run complex workflows with changing 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***)

# HEP-CCE

- **HPC systems ARE useful for data-intensive tasks:** Current estimates are that up to 70% of HEP computing can be done on HPC platforms
- **Will HPC systems deliver on this promise?:** This is largely a policy issue, not primarily determined by technical bottlenecks
- **Is the HEP case unique?:** The HEP community is very “data-aware” as compared to some others; the number of competing efforts is not large
- **What about other fields?:** There is likely to be an “effort barrier” — the use case must be at large-enough scale to make a supercomputing-based attack worthwhile; cloud or local resources will remain attractive options for many applications

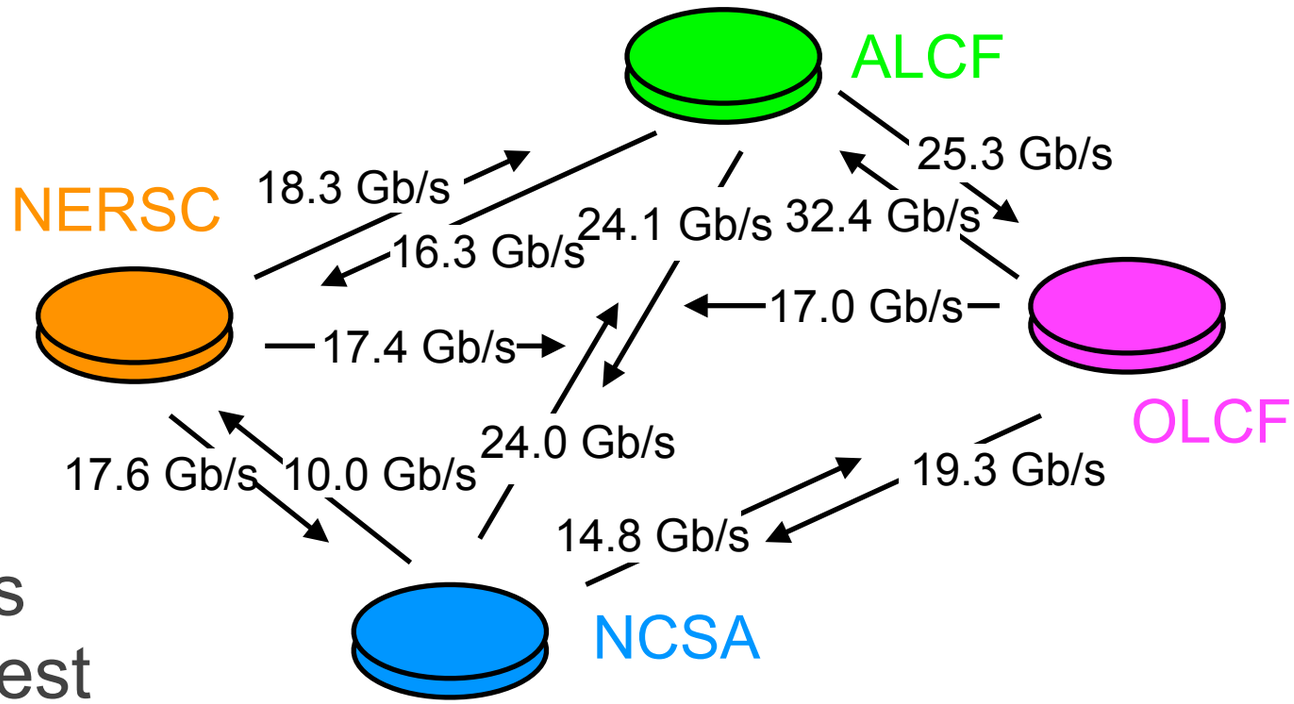
Making the exascale environment work for HEP through interaction with ASCR — HEP-CCE

http://hepcce.org/




# “Production” Example: Large-Scale Data Movement

- **Offline Data Flows:** Cosmological simulation data flows already require ~PB/week capability, next-generation streaming data will require similar bandwidth
- **ESnet Project:** Aim to achieve a production capability of 1 PB/week (FS to FS, also HPSS to HPSS) across major compute sites
- **Status:** Success achieved! numbers from a simulation dataset “transfer test package” (4 TB)
- **Future:** Automate entire process within the data workflow including retrieval from archival storage (HPSS); add more compute/data hubs (BNL underway, just solved Globus-dCache handshake problem)




**Petascale DTN project, courtesy Eli Dart, HEP-CCE/ESnet supported joint project**

globus online

ESnet
ENERGY SCIENCES NETWORK

**HEP-CCE**

# Summary

- **Is HPC the solution you have been waiting for?**
  - Not quite, but —
  - It might be a solution you can live with (provided software upgrades are doable and straitjacketing is acceptable)
  - It might be a solution you will *have* to live with (power, money)
- **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
- **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!
- **Natural synergy with HEP in many places**
  - Use this to leverage available software/experience/capabilities
  - Use HEP-CCE, HSF, other points of interaction such as ECP and SciDAC