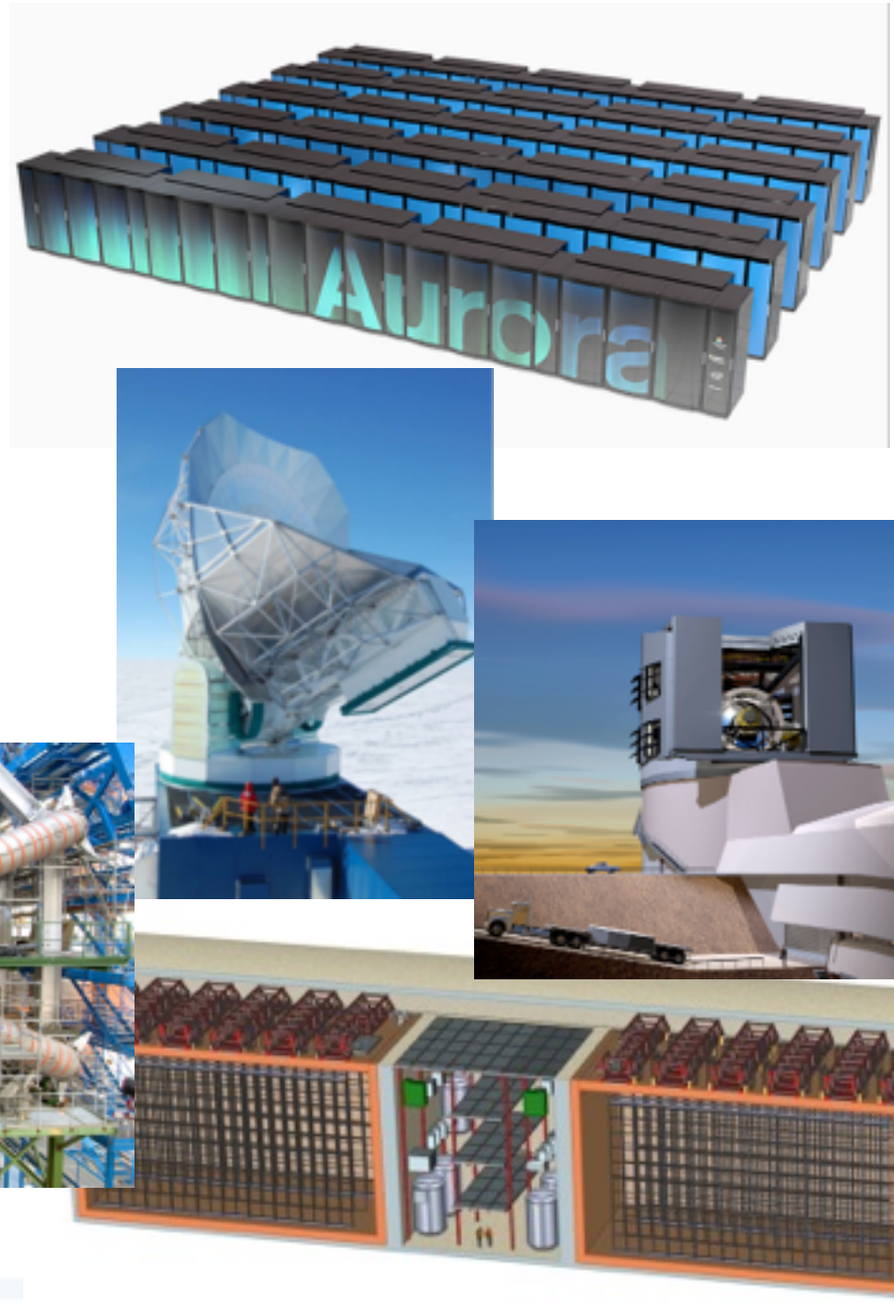
High Performance Computing for Scientific Data-Intensive Tasks: Friend or Foe?

Salman Habib
HEP and MCS Divisions
Argonne National Laboratory

Acknowledgments: Doug Benjamin, Lindsey Bleem, Franck Cappello, Taylor Childers, Katrin Heitmann, Tom LeCompte, Ravi Madduri, Dan Murphy-Olson, Adrian Pope, Tom Uram (Argonne)

Debbie Bard, Wahid Bhimji, Lisa Gerhardt, Peter Nugent (LBNL)



Will it Work?

- **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 is frustrating
  - HPC experts are not user-friendly
  - Machine downtime and crashes are common
  - Ability to ‘roll your own’ is limited

Running Jobs | Queued Jobs | Reservations

Total Queued Jobs: 172

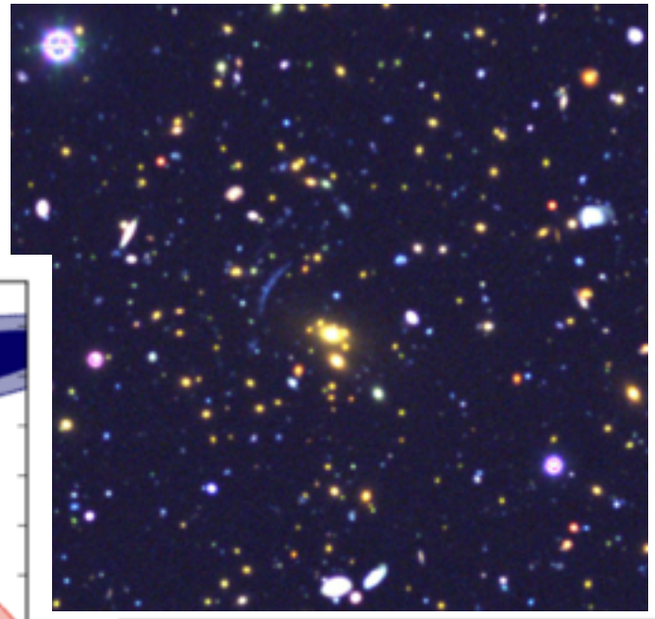
Job Id	Project	Score	Walltime	Queued Time	Queue	Nodes
307941	SkySurvey	8351.7	1d 00:00:00	5d 01:10:03	prod-capability	32768
307942	SkySurvey	8350.5	1d 00:00:00	5d 01:09:42	prod-capability	32768
309793	NucStructReact_2	7069.0	01:00:00	1d 19:13:34	prod-capability	32768
309794	NucStructReact_2	7065.1	01:00:00	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

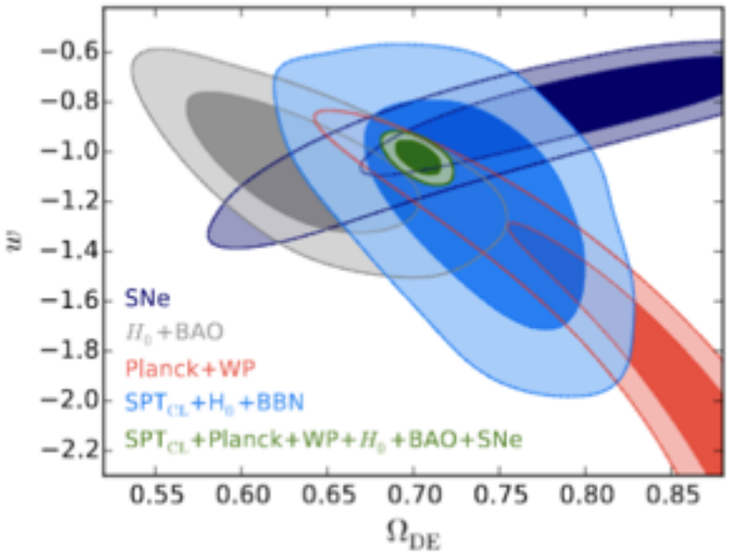
Computing Needs for Science

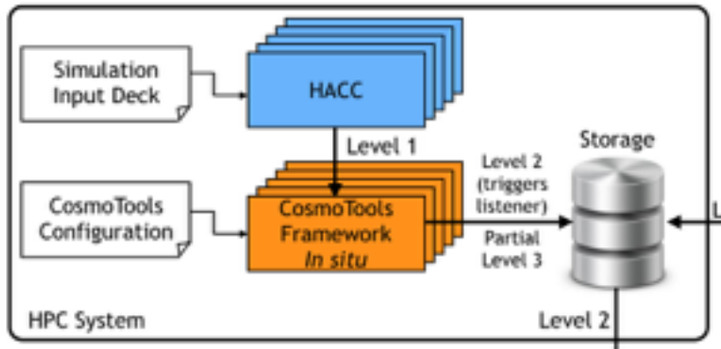
- **Many Communities Need Large-Scale Computational Resources**
  - **Light sources**
  - **Biology**
  - **Climate/Earth Sciences**
  - **High Energy Physics**
  - **Materials**
- **Message: Overall scientific computing use case is driven by large-scale data flow + volume**
- **Data-intensive applications will be ubiquitous, and will need performance, reliabiility, and usability**
- **Overall balance of compute + I/O + storage + networking will need to be thought through**

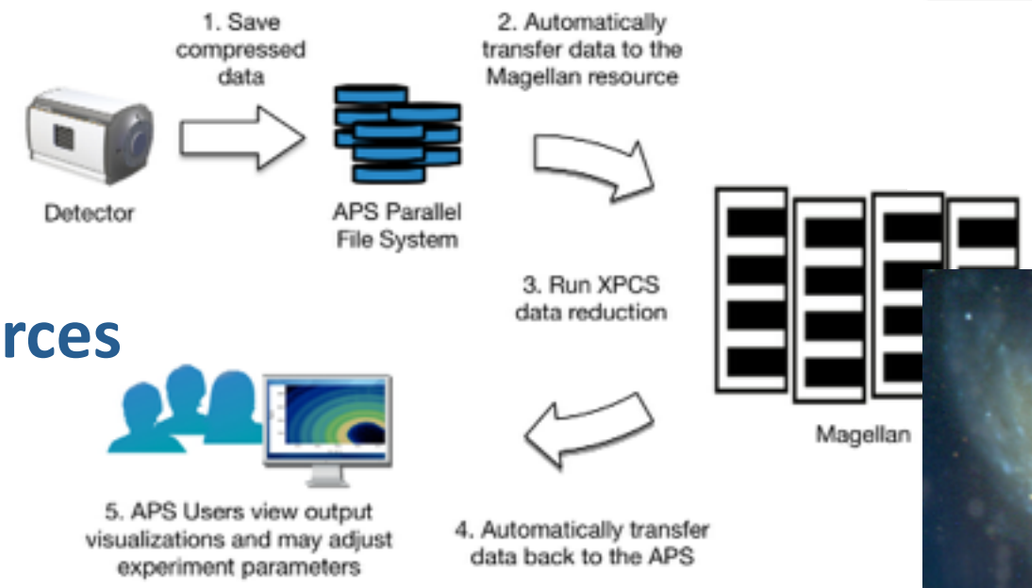
Huge Variety of Large-Scale Data-Intensive Tasks

- **Mining/Classification**
  - **Image Analysis**
- **Statistical Inverse Problems**
  - **Reconstruction**
- **Data Analysis/Management**
  - **Instrumental Pipelines**
- **Real-Time Analytics**
  - **Experiments and Data "In-Loop"**
- **Data Services**
  - **Fast queries on large datasets**
- **HPC Systems as Data Sources**
  - **In-Situ and Off-line analysis**



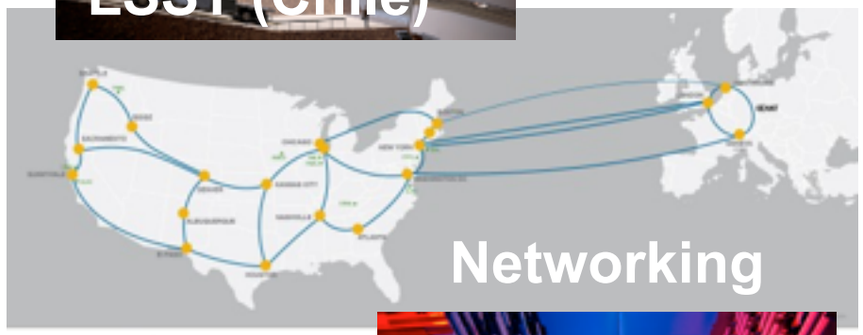


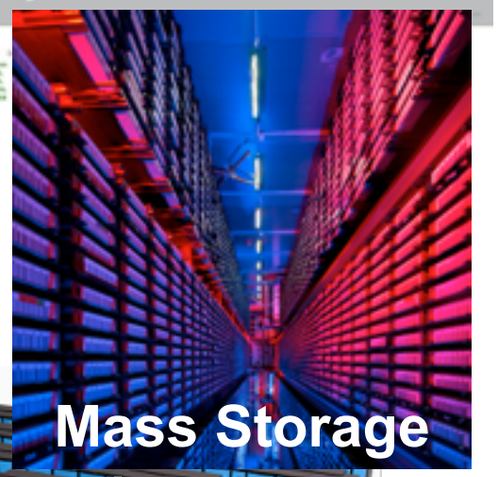




Scientific Data and Computing: 'Geography'

- **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'
- **Practical Response**
  - Make things better, but ***not unrealistically better***

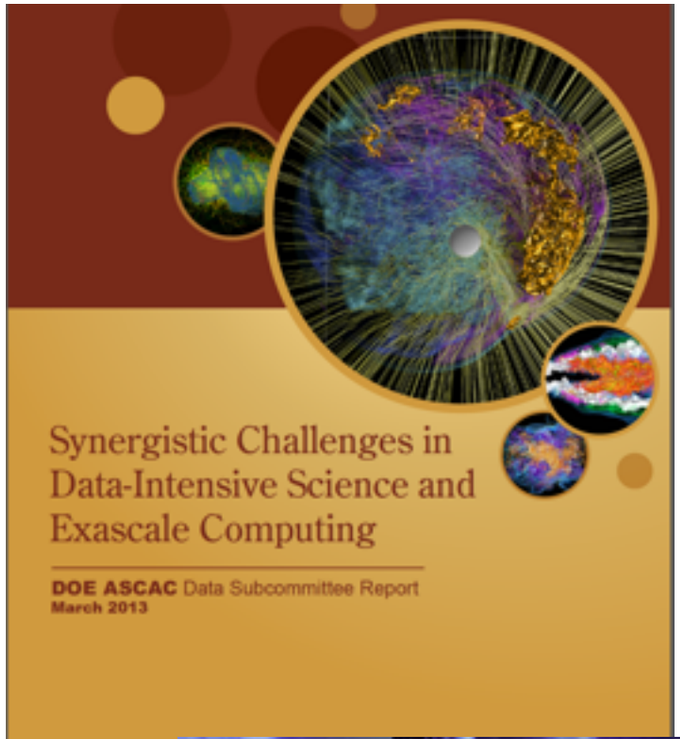


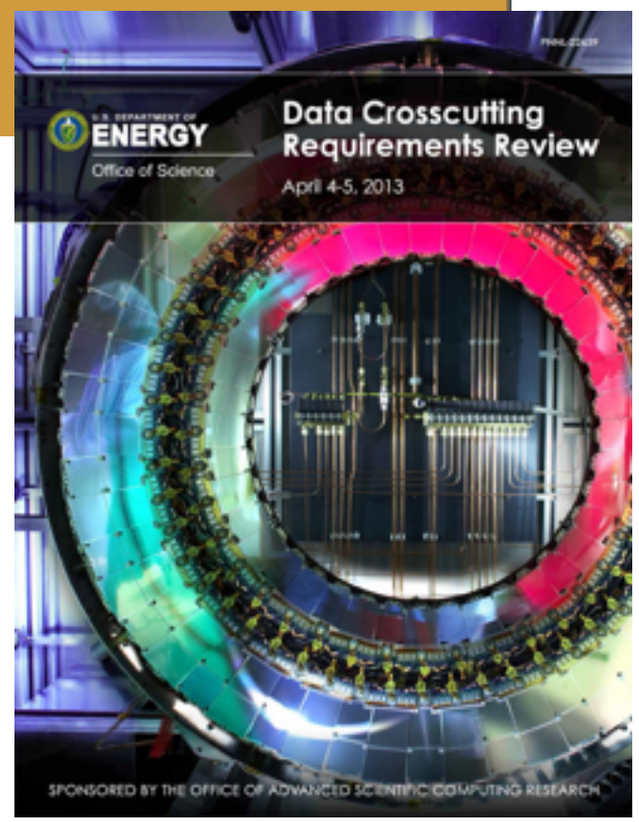


Different Flavors of Computing

- **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
  - 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 have it





Boundary Conditions

- **What's the Problem?**
  - Even if solutions can be designed ***in principle***, the resources needed to implement them are (usually) not available
  - This is because, ***despite all the evidence of its power***, computing 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) is an excellent option but is not issue-free
  - Storage and networking remain major problems

HEP Cosmic Frontier 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

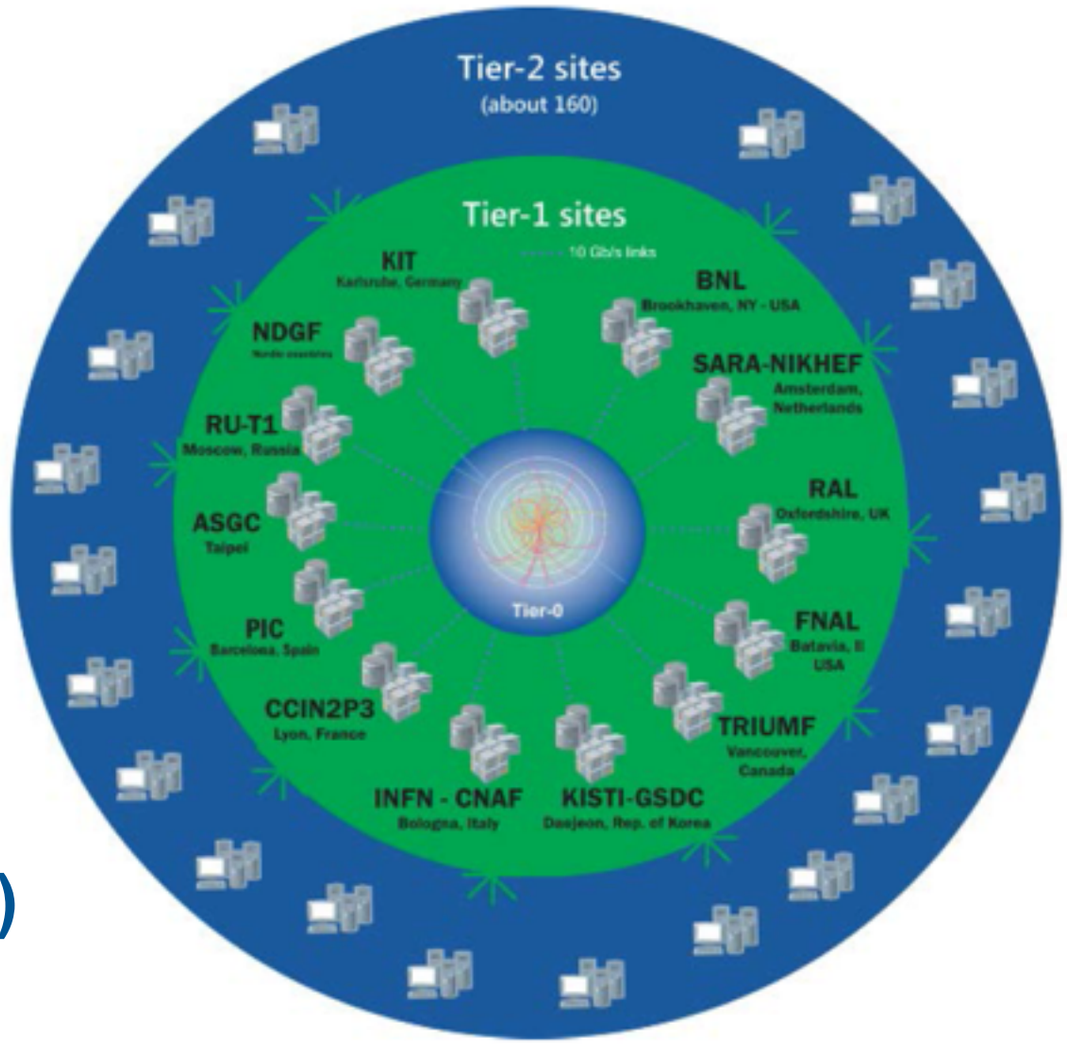
300 TFlops/10PB, 10kW in 2020 (Projection)

Case Example: High Energy Physics

- **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 can be complex and need to be run many times (individual campaigns can last for months)
  - Scale of data is variable — 10s of TB to 100s of PB/year
  - Multiple IO requirements
- **ASCR/HEP Exascale Requirements Review**
  - http://arxiv.org/abs/1603.09303, also http://hepcce.org/resources/reports/

HEP Computing Paradigms

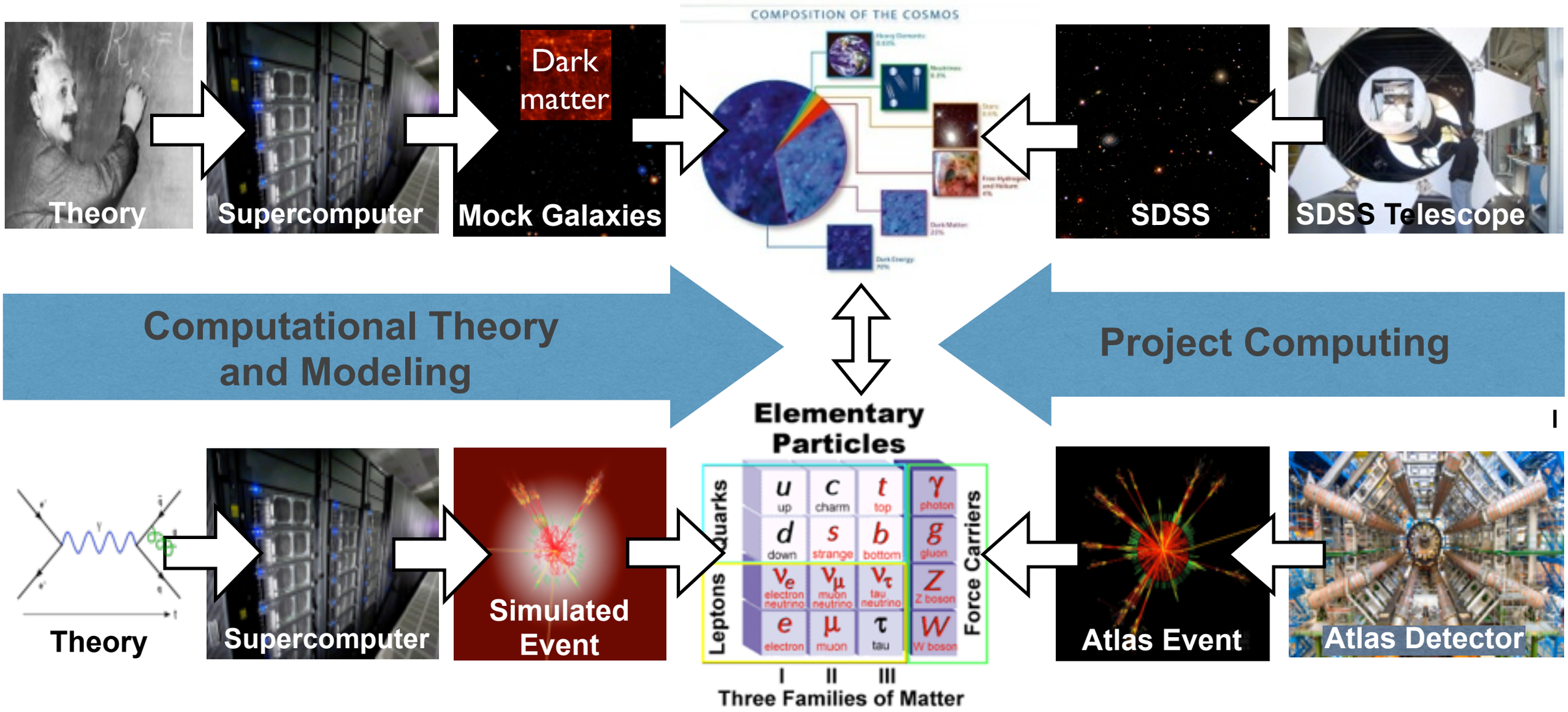
- **High Throughput Computing (HTC)**
  - Major exploitation of Grid resources. Co-evolution of HEP experimental software and the Grid is reaching a potential breaking point (not enough resources to handle demand). New hardware/software exploits needed.
- **High Performance Computing (HPC)**
  - Classic use of HPC resources by theorists. New ideas for simulating experimental events include event services and dedicated front-ends for job packaging.
- **Data-Intensive Scalable Computing (DISC)**
  - Analysis of datasets generated from simulations and co-analysis of simulation and observational data without HTC lead times. Desire for true interactive large-scale computing (‘power cloud’).



Large Hadron Collider (LHC) worldwide computing infrastructure

HEP 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



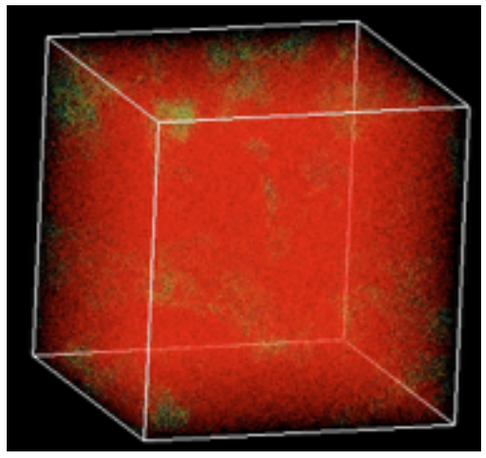
Simulated Data: 1) Event generation (lists of particles and momenta), 2) Simulation (interaction with detector), 3) Reconstruction (presence of particles inferred from detector response); **Analysis:** Comparison with actual data

HPC

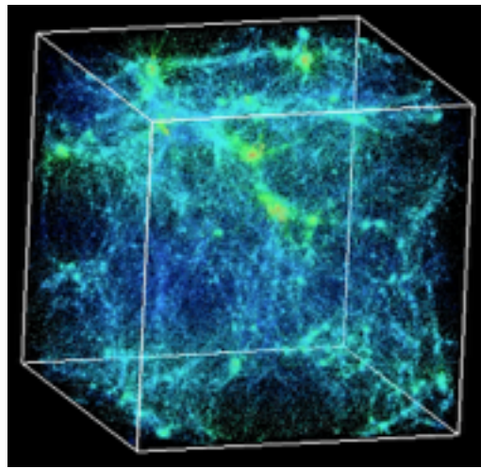
DISC

Analytics/Workflow Complexity Example

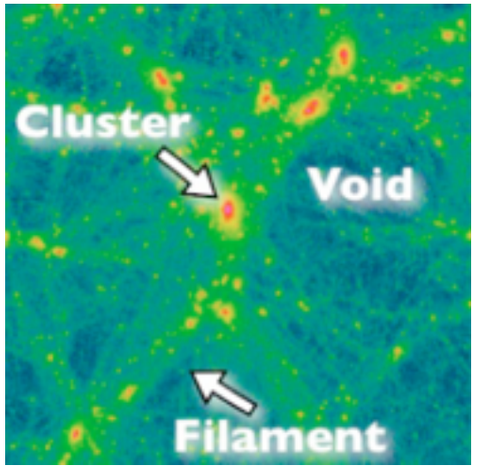
Gaussian Random Field Initial Conditions



High-Resolution N-Body Code (HACC)



Multiple Outputs Halo/Sub-Halo Identification



Halo Merger Trees

Semi-Analytic Modeling Code (Galacticus)

Galaxy Catalog

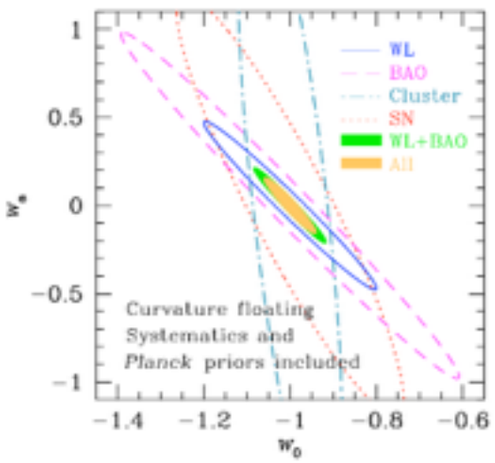
Realistic Image Catalog

Atmosphere and Instrument Modeling

Data Management Pipeline

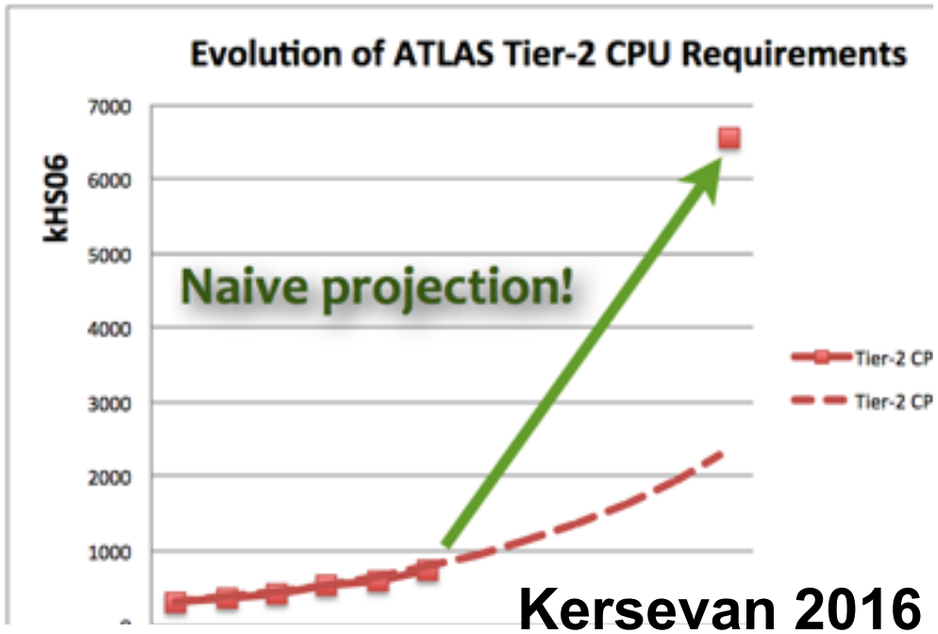
Data Analysis Pipeline

Scientific Inference Framework

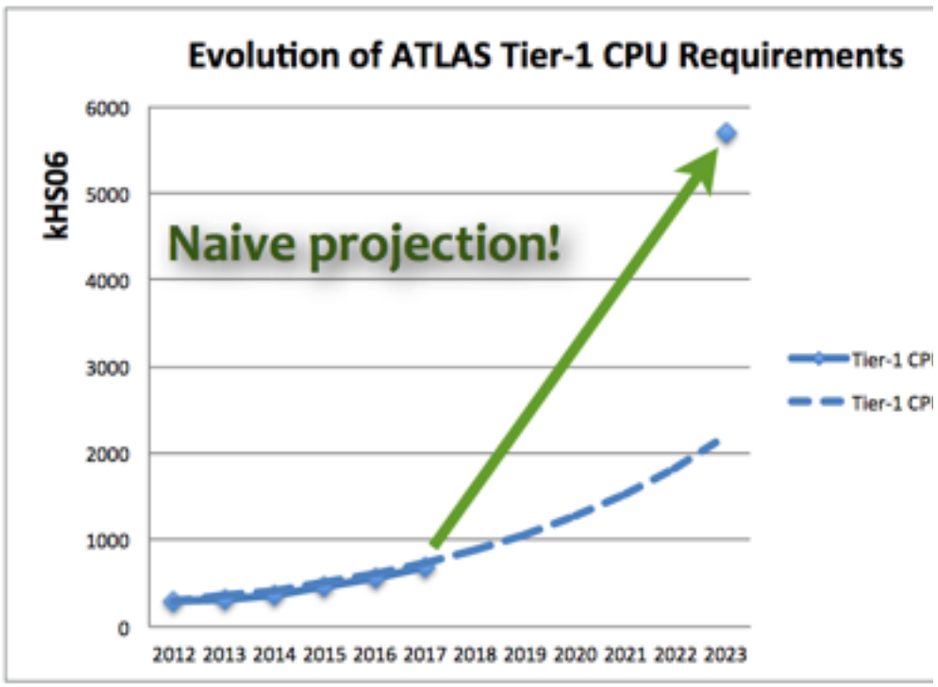


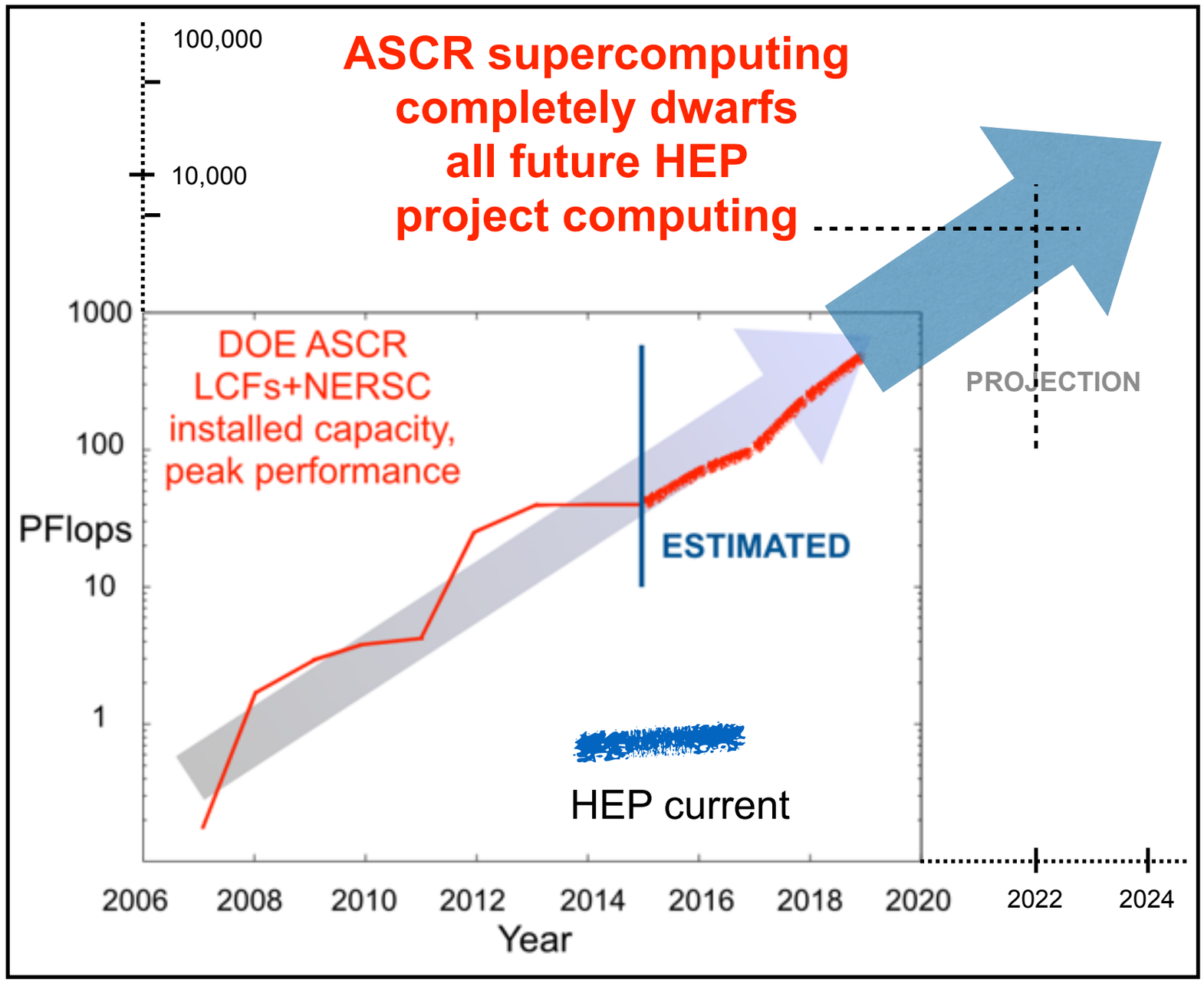
HEP Computing Requirements for Energy Frontier

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



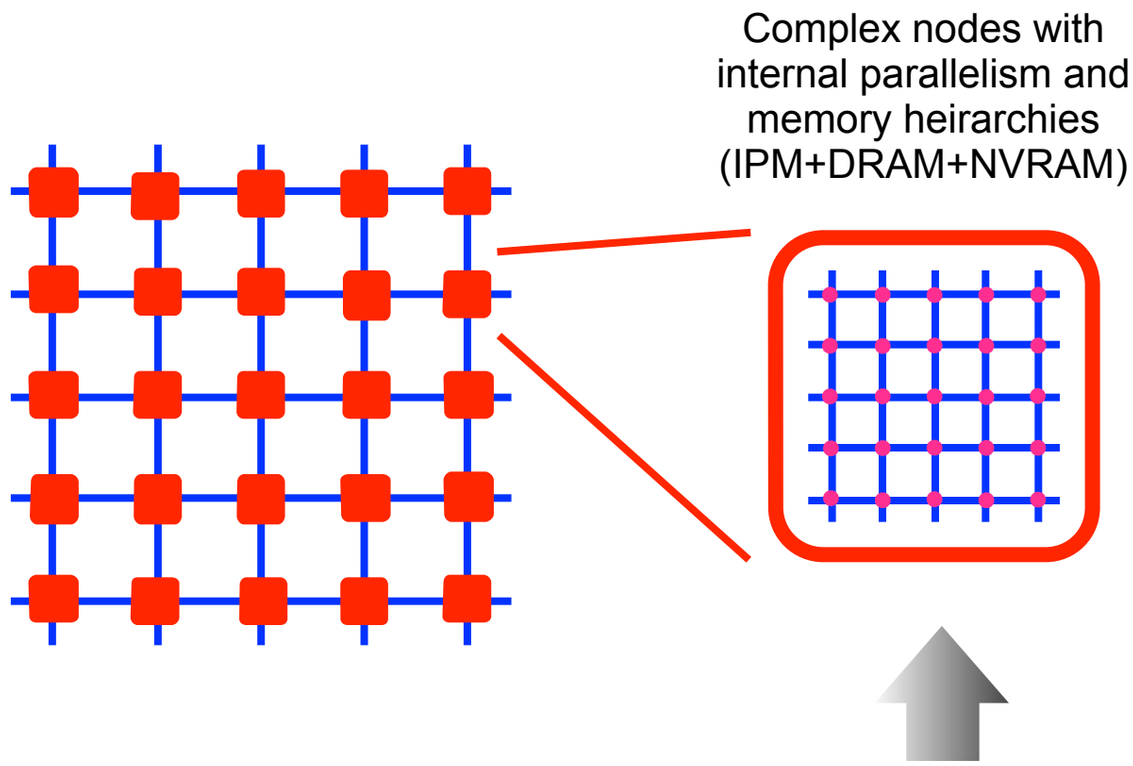
Kersevan 2016





HPC-based DISC: Likely Exploits

- **Most use cases likely to be DISC/HTC**
  - Note HPC systems can easily handle these in the very near future
  - Possibly fall into two classes — **1)** many runs of a simple, not highly optimizable code, **2)** smaller, but still sizable number of runs of a potentially optimizable code
- **'Single node' application span**
  - Nodes are big enough: >100GB RAM + NVRAM (total memory ~PB with I/O BW at ~TB/s)
  - Key parallelism exploit at the node level
- **Exceptions**
  - Large-scale spatio-temporal statistics (will need system level parallelism — essentially an HPC application)



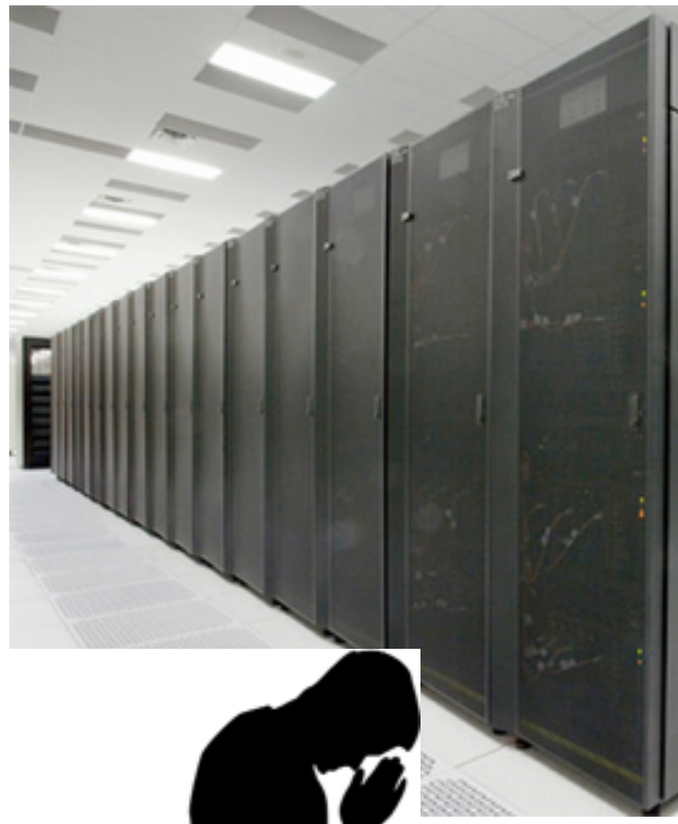
Focus on node-level parallelism: Quasi-independent tasks run on individual nodes; intermittent communication across nodes

Main themes:

1. **Locality, locality, locality, —**
2. **Threading/Vectorization**
3. **I/O**

“Data Meets HPC” — Basic Requirements

- **Software Stack:** Ability to run arbitrarily complex software stacks (***software management***)
- **Resilience:** Ability to handle failures of job streams (***resilience***)
- **Resource Flexibility:** Ability to run complex workflows with changing computational ‘width’ (***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 (***integration***)
- **Automated Workloads:** Ability to run large-scale automated production workflows (***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***)



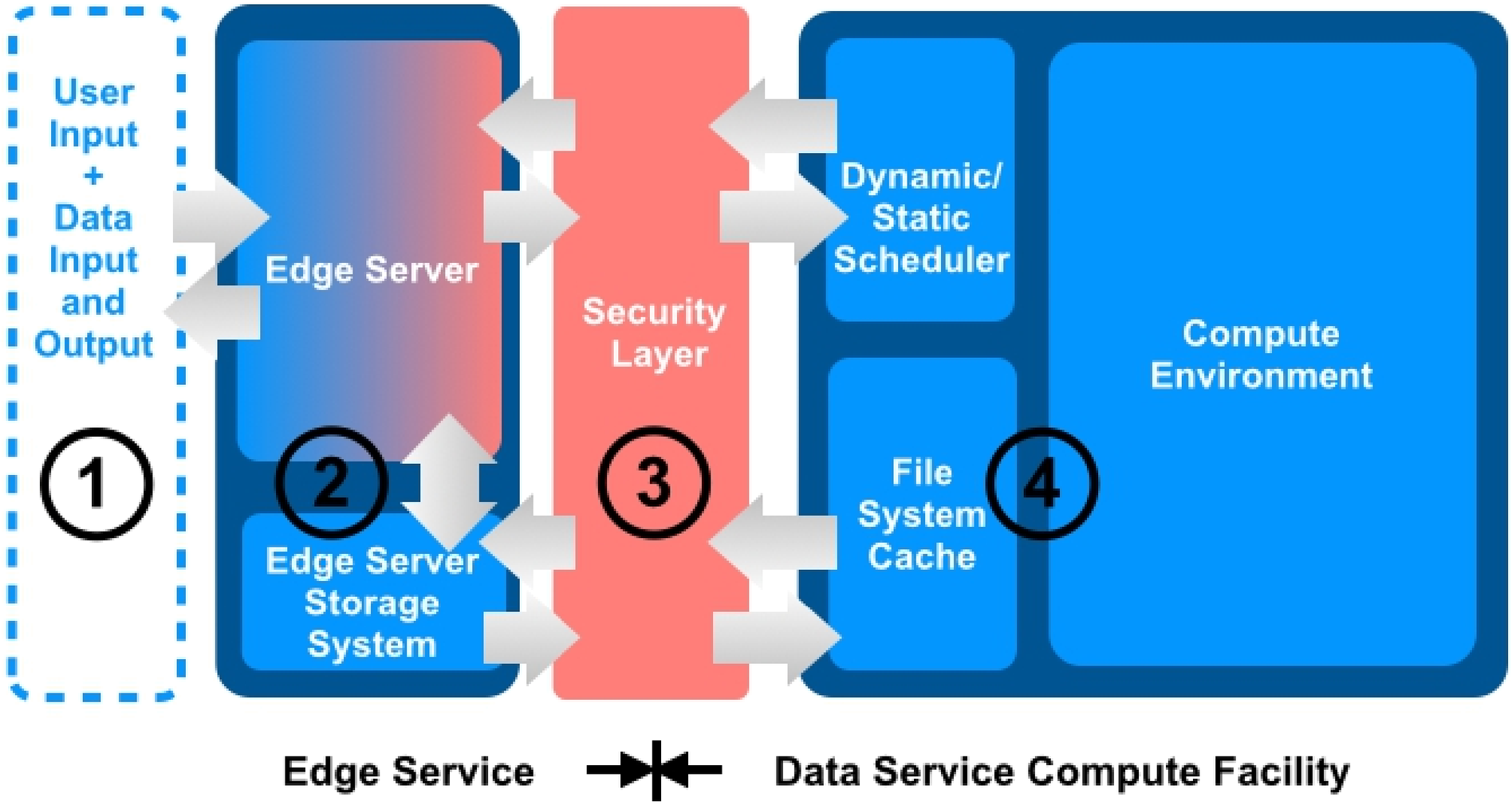
HPC Systems in HEP World: Nuts and Bolts

- **HEP vs. HPC Practice**
  - HEP community used to ‘owned’ resources
  - HPC systems belong to someone else — no root access!
  - HPC systems have higher levels of security requirements
- **Data Transfers**
  - Large data transfers on HPC systems via dedicated data transfer nodes, unlike the LHC Grid, where transfers are to worker nodes
  - HPC I/O not optimized for fine-grained file I/O
- **Compute Architecture**
  - Node-level architecture supports compute-heavy applications that can potentially scale up; most HEP applications are not compute-intensive and scalability is not needed (event level analysis, 1-10MB of data/event)


Titan at Oak Ridge


CERN data center

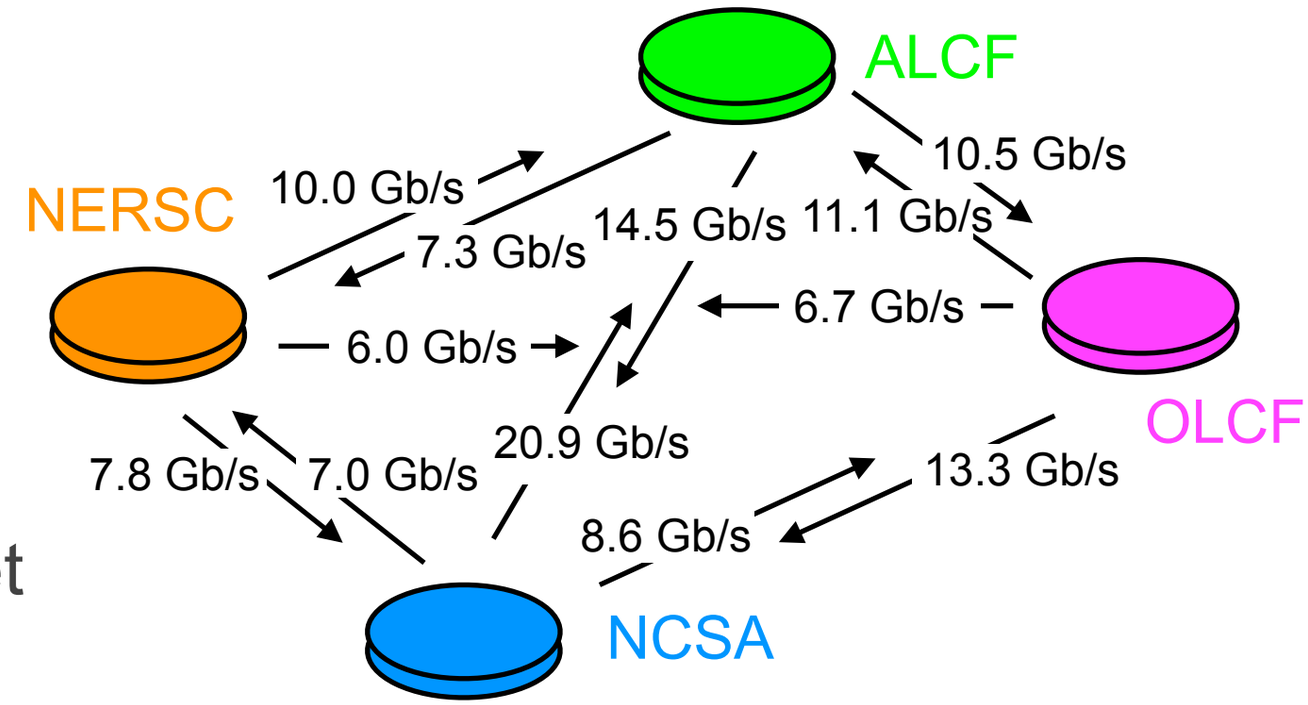
Connecting to HPC Systems: Edge Services



Edge service design must consider a number of factors; security, resource flexibility, interaction with HPC schedulers, external databases, requirements of the user community — several specific examples are in production use. ***Key point*** — nothing from a user's job message is ever executed on a command line, only applications registered in the edge service database can be run

Large-Scale Data Movement

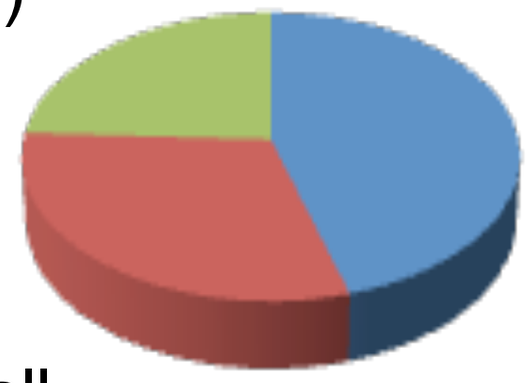
- **Offline Data Flows:** Cosmological simulation data flows already require ~PB/week capability, next-generation streaming data will require similar bandwidth
- **ESnet Project:** Aim to achieve a production capability of 1 PB/week (FS to FS, also HPSS to HPSS) across major compute sites
- **Status:** Very close but not there yet (600+ TB/week); 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, FNAL, SDSC, —)



Petascale DTN project, courtesy Eli Dart, HEP-CCE/ESnet support

Energy Frontier Status

- **HEP Payloads on HPC/Next-Gen Architectures**
  - X86 clusters are fine
  - Xeon Phi (KNL) looking good (Geant4, etc.)
  - IBM BG/Q systems also ok
  - GPUs problematic (too different from X86)
- **Data Transfers**
  - ASCR facilities prefer a single solution for all users, petascale data transfer project ongoing, using Globus
- **Software Management**
  - Containerization work with multiple projects (including Cosmic and Intensity Frontiers); uses NERSC's Shifter technology — work ongoing with CMS and ATLAS teams
- **I/O on HPC Systems**
  - Burst buffers have led to factor of 2 improvements in HEP I/O tests, more possible

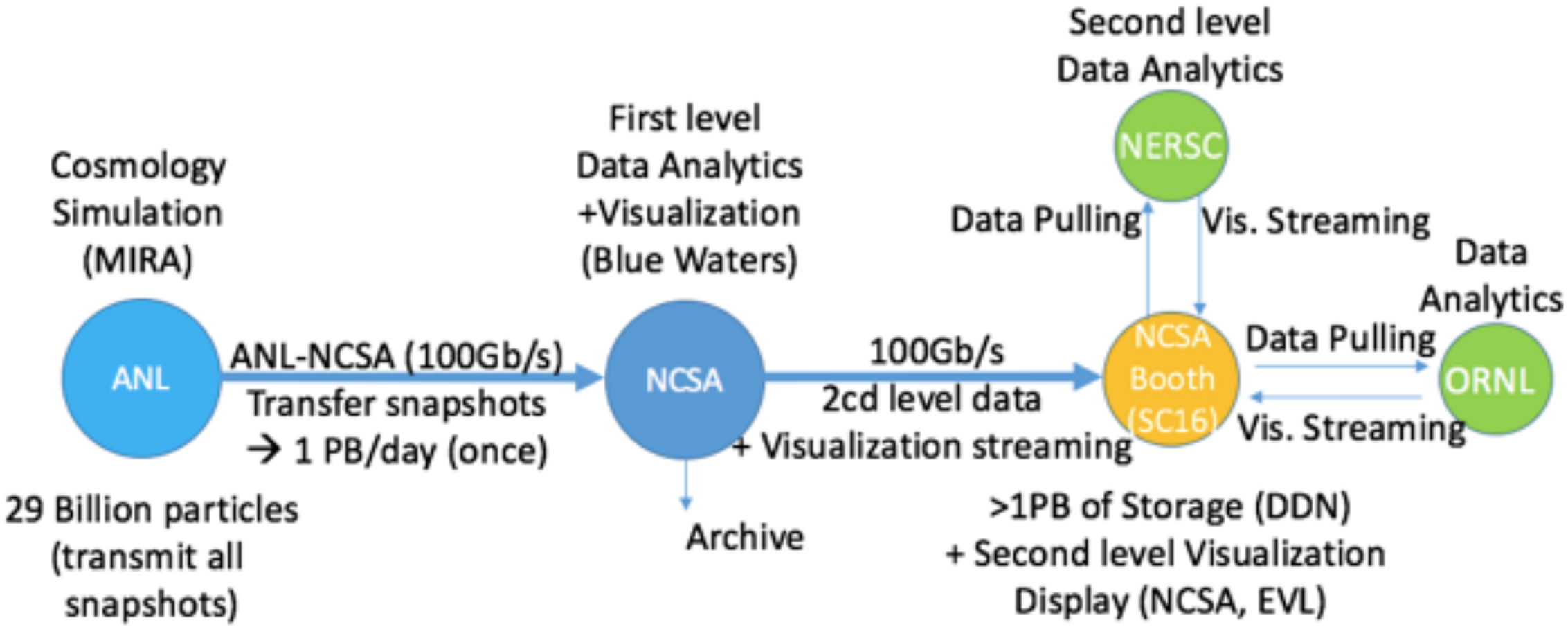


HPC systems already providing ~150M core-hours/year, roughly equivalent to 15% of the ATLAS global Grid resources

See Childers/ Gerhardt at ICHEP2016

Cosmic Lab of the Future Demo (SC16)

- **SC16 SciNet Demo**
  - HPC system — Mira or Theta at Argonne
  - DISC system — Blue Waters at NCSA
  - Data Center — SC16 booth, NERSC, ORNL systems
  - PDACS (Portal for Data Analysis services for Cosmological Simulations) as analysis engine



Summary

- **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/