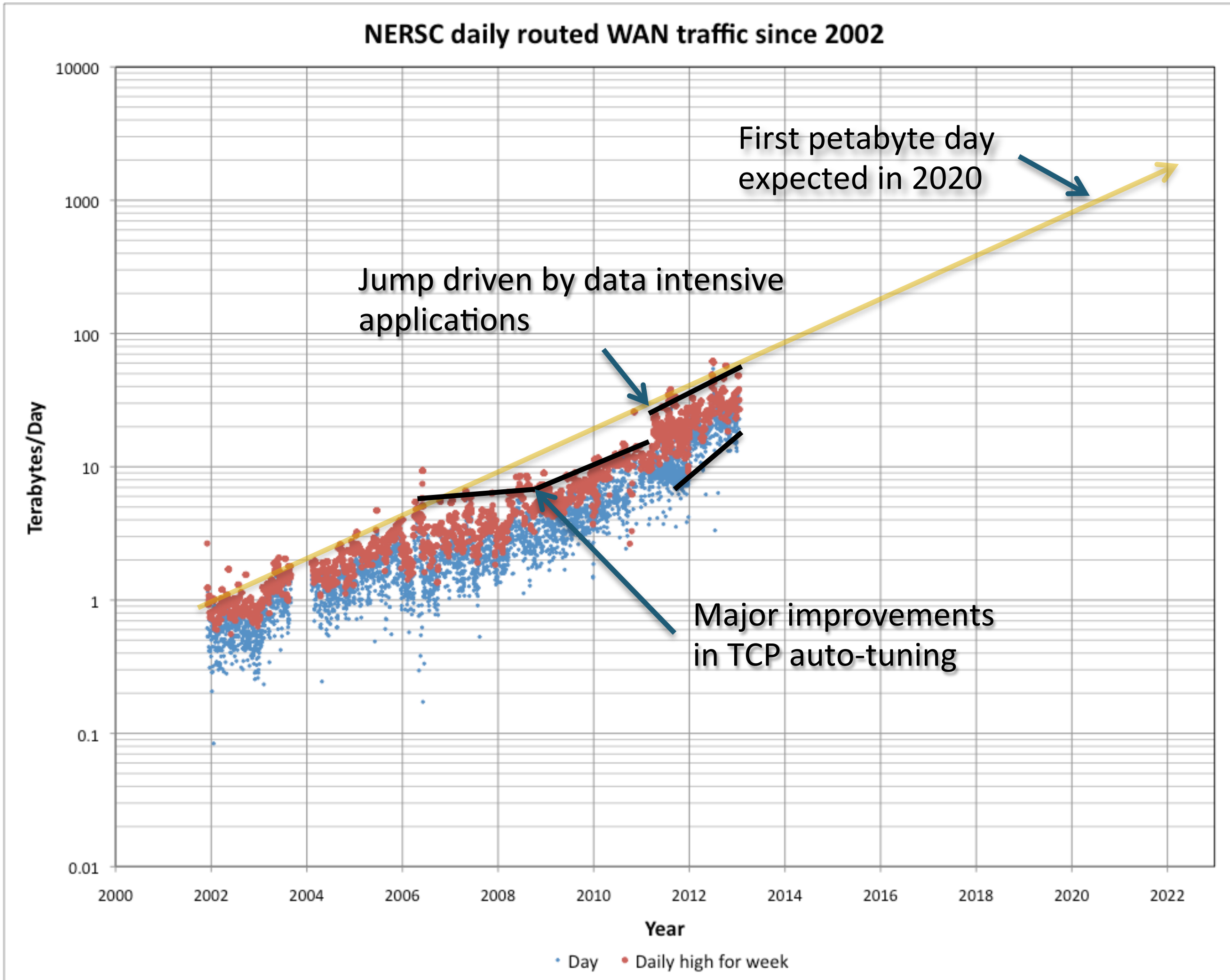
Extreme Data Science

Sudip Dosanjh, Shane Canon,
Jack DeSlippe, Kjiersten Fagnan, Richard Gerber,
Lisa Gerhardt, Jason Hick, Douglas Jacobsen,
David Skinner, and Nicholas J. Wright

Lawrence Berkeley National Laboratory

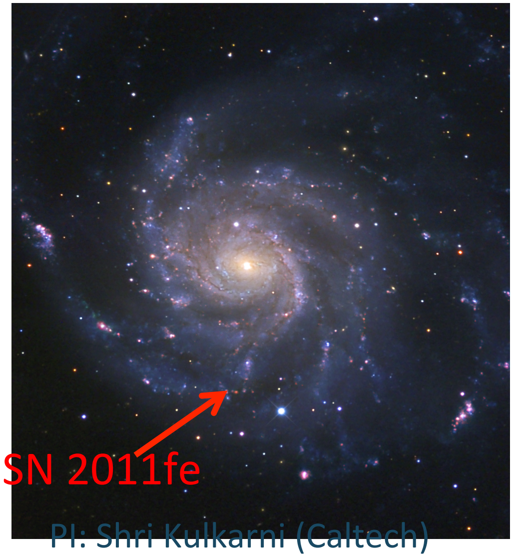
Exponentially increasing data traffic



Recent Scientific Breakthroughs Enabled by Extreme Data Science

NERSC

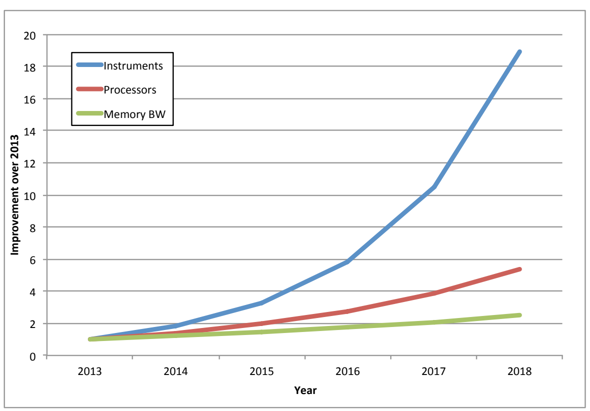
- Discovery of the Higgs Boson
- Measurement of the important "θ_{13}" neutrino parameter. One of Science Magazine's Top-Ten Breakthroughs of 2012.
  - Last and most elusive piece of a longstanding puzzle: why neutrinos appear to vanish as they travel
- The Palomar Transient Factory Discovered over 2000 supernovae in the last 5 years, including the youngest and closest Type Ia supernova in past 40 years
- Trillions of measurements by the Planck satellite led to the most detailed maps ever of cosmic microwave background
- Four of Science Magazines breakthroughs of the last decade were in Genomics
- Materials project has over 5000 users and was featured on the cover of Scientific American

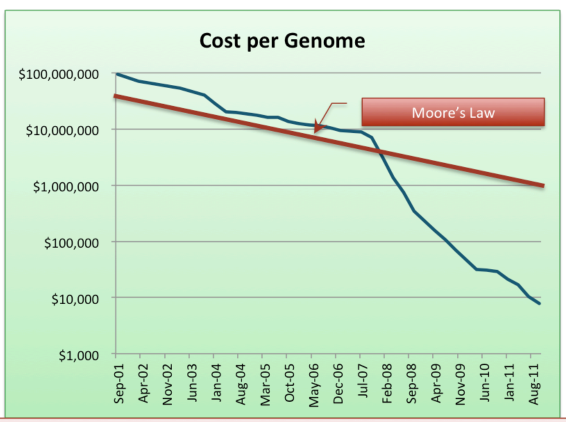


PI: Shri Kulkarni (Caltech)

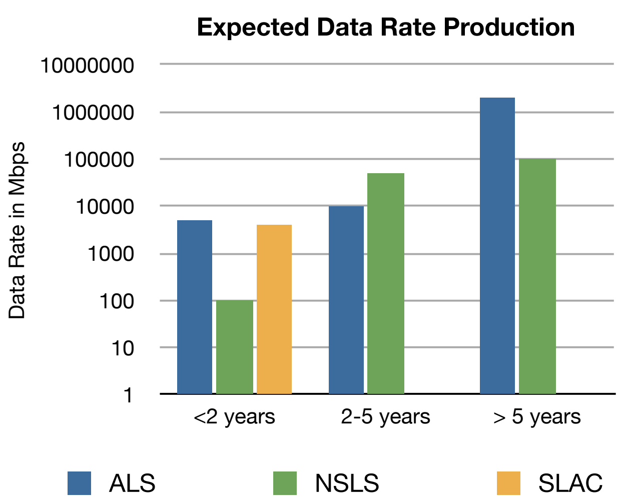
Data deluge will continue at DOE experimental facilities

- **The observational dataset for the Large Synoptic Survey Telescope will be ~100 PB**
- **The Daya Bay project will require simulations which will use over 128 PB of aggregate memory**
- **By 2017 ATLAS/CMS will have generated 190 PB**
- **Light Source Data Projections:**
  - 2009: 65 TB/yr
  - 2011: 312 TB/yr
  - 2013: 1.9 PB /yr
  - EB in 2021?
  - NGLS is expected to generate data at a terabit per second





Source: National Human Genome Research Institute



U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Unique data-centric resources will be needed

- Compute
- On-Package DRAM
- Capacity Memory
- On-node-Storage
- In-Rack Storage
- Interconnect
- Global Shared Disk
- Off-System Network

Compute Intensive Arch	Data Intensive Arch
Goal: ***Maximum computational density and local bandwidth*** *for given power/cost constraint.* Maximizes bandwidth density near compute	**Goal:** ***Maximum data capacity and global bandwidth*** *for given power/cost constraint.* Bring more storage capacity near compute (or conversely embed more compute into the storage). *Requires software and programming environment support for such a paradigm shift*

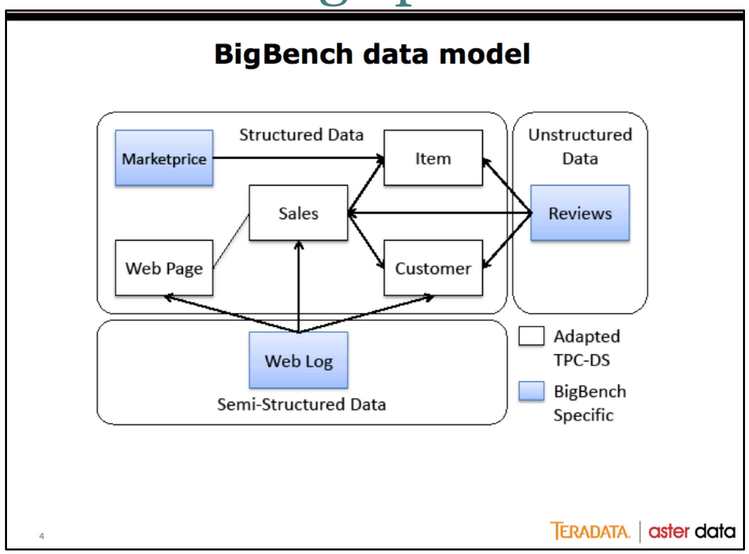
Direct from each node

Path Forward for Big Data and Extreme Computing

Chaitan Baru
Michael Norman
San Diego Supercomputer Center
UC San Diego

Application-level Benchmarking

- TPC-style: Schema + Workload
  - E.g.: BigBench: TPC+H with semistructured data and data mining, machine learning operations



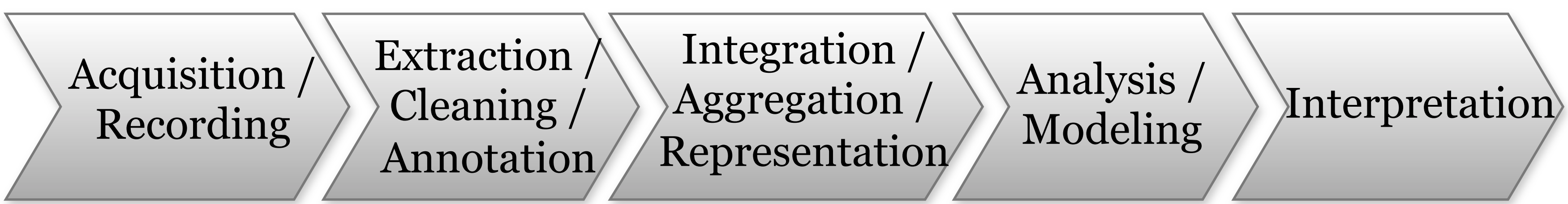
- Several other proposals under development:
  - HiBench, BigDecision, BigDataBench, Deep Analytics Pipeline
  - TPCx-HS: TPX Express – Hadoop Systems

SDSC UCSD SAN DIEGO SUPERCOMPUTER CENTER

CLDS Center for Large-scale Data Systems Research

Processing Pipelines: *Deep Analytics Pipeline*

- An end-to-end data processing pipline:
  - Data from multiple sources
  - Loose, flexible schema
  - Data requires structuring
  - ELT rather than ETL
- “User Modeling” is a prototypical application
  - Retail shoppers, Telecom subscribers, Healthcare patients, DataCenter HW and SW systems, Users in Ad-based Web
- Applications consist of
  - Pipelines of processing
  - Running models with data

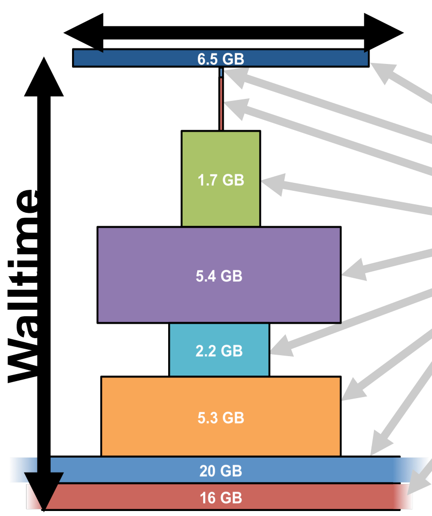


Processing Pipeline: Whole Genome Sequencing

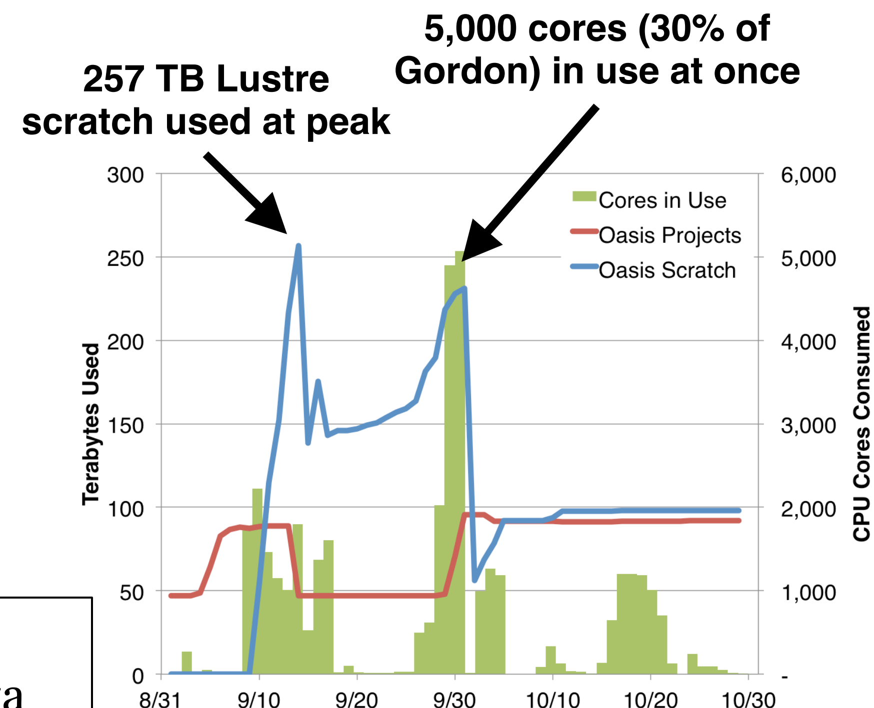
- By: Kristopher Standish*^, Tristan M. Carland*, Glenn K. Lockwood+^, Mahidhar Tatineni+^, Wayne Pfeiffer+^, Nicholas J. Schork*^

*Scripps Translational Science Institute, +San Diego Supercomputer Center, ^UC San Diego

- Project funding provided by Janssen R&D



1. **Map** (BWA)
2. **sam to bam** (SAMtools)
3. **Merge Lanes** (SAMtools)
4. **Sort** (SAMtools)
5. **Mark Duplicates** (Picard)
6. **Target Creator** (GATK)
7. **Indel Realigner** (GATK)
8. **Base Quality Score Recalibration** (GATK)
9. **Print Reads** (GATK)



36 core-years of computing;
438 full genomes; 50TB compressed data

What We Need

- Shared experimental infrastructure at scale for:
  - Systems R&D; software development and testing; and yes, education!
- Co-design, but also “co-education”!
  - Involve students: CS, science, computational science, data science
- A coordinated effort among science/CS—and also among agencies
- Reality: Ideas as well as funding may need to come from multiple sources

Thomas Lippert (Leader SP7: HPC Platform)
Boris Orth (SP 7 Project Manager)
Bernd Mohr (Task Leader T 7.2.4

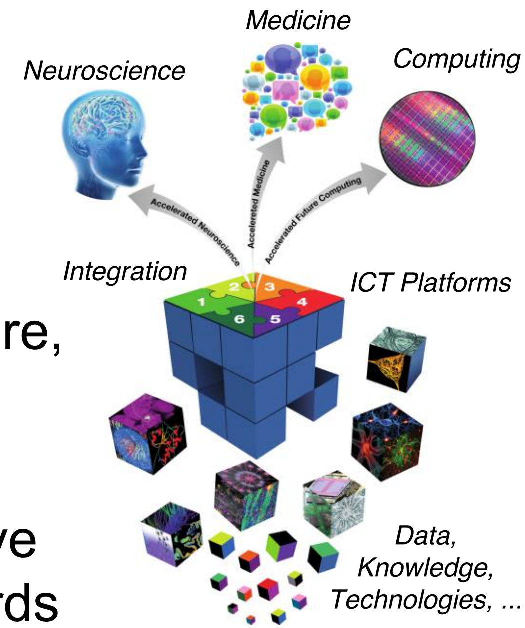
Human Brain Project

Basic Facts

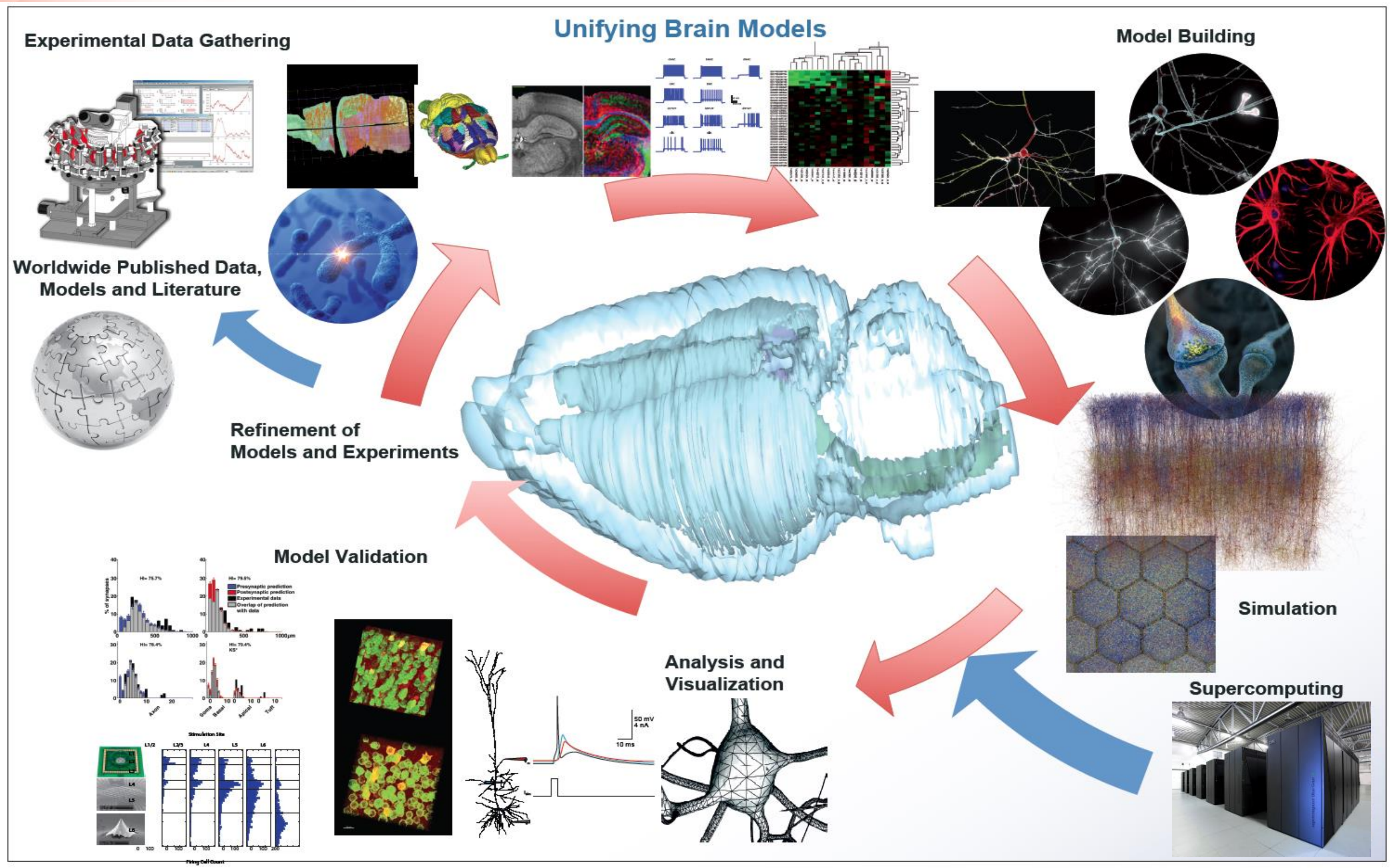
- European-led international large-scale project
- EU FET Flagship Programme
- 10 years duration (Oct 2013 →)
- EUR 1.1 billion total cost
- 12 subprojects (of which 2 led by Jülich)
- 80 partners / 23 countries
  - More via *Competitive Calls*
- Coordinated by EPFL (Henry Markram)
- www.humanbrainproject.eu

GOAL

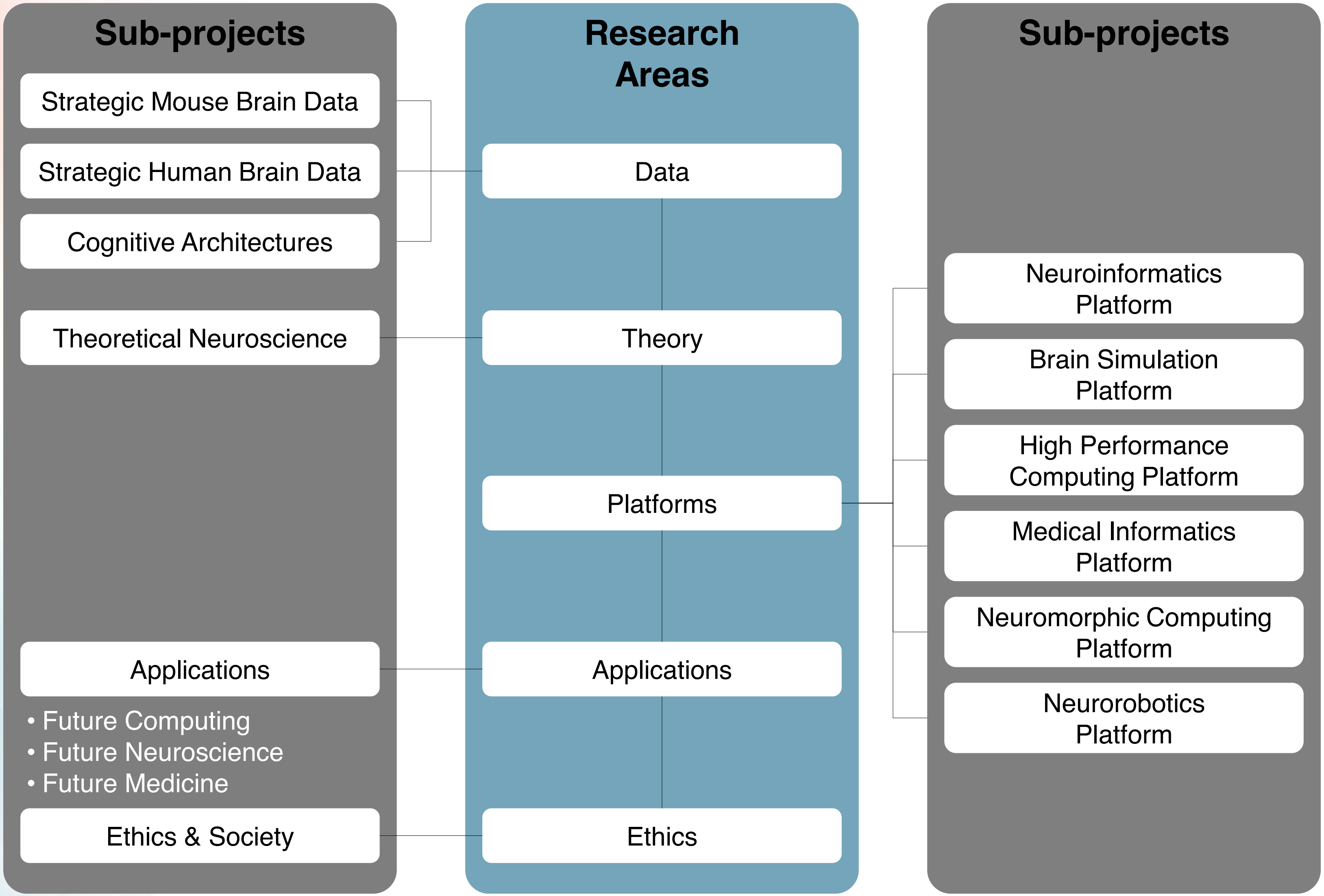
- Build an integrated ICT infrastructure, enabling
- A global collaborative effort towards understanding the human brain, and ultimately
- Emulate its computational capabilities



Integration Strategy



HBP Research Areas and Subprojects



Key technical aspects of future HPC platform

Vision of **Interactive Supercomputing**:
data-intensive interactive simulations, analysis and visualization

- Efficient data management
  - Significantly increased memory capacity to keep data within system
- Tightly integrated visualization
  - Rendering close to data, scalable image compositing
- Dynamic resource management
  - Dynamic relocation of resources within session and dynamic resizing of session resources
  - Co-scheduling of heterogeneous resources

The Need for Resilience Research in Workflows of Big Compute and Big Data Scientific Applications

Franck Cappello ANL&UIUC and Tom Peterka, ANL

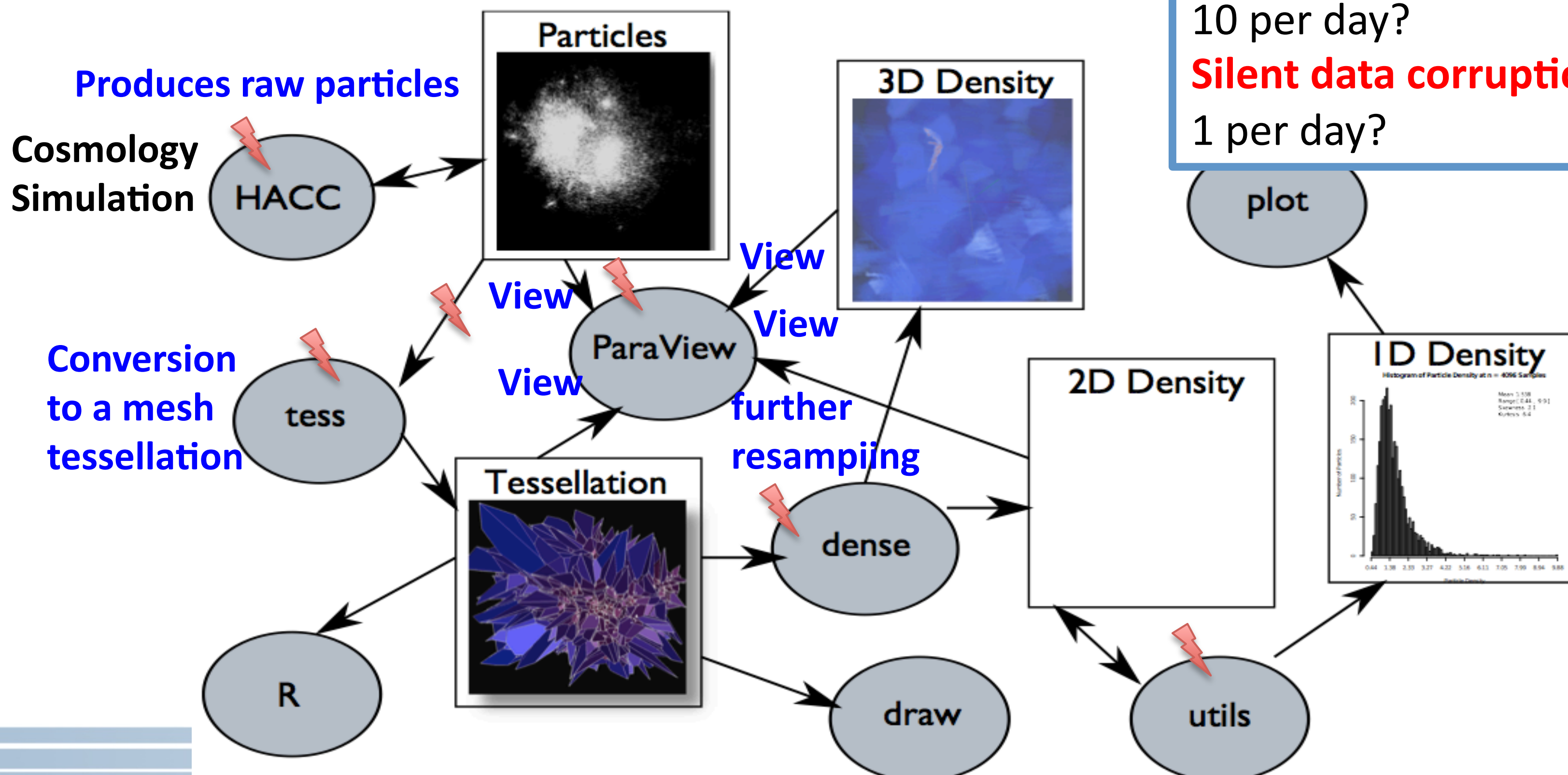
In situ BigCompute + BigData: A new Class of Executions

-increasing need of coupling simulations with Data analytics

(generated data too large to fit on storage for off-line analytics)

-different types of analytics: physics, visualization

Key problems@Exascale:
Fail stop errors, process crashes
10 per day?
Silent data corruptions
1 per day?

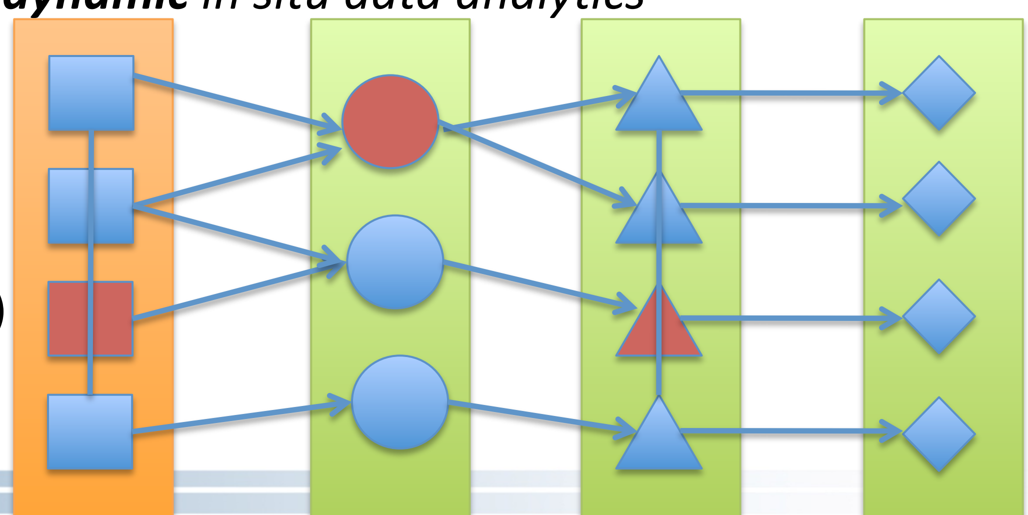


What is the problem?

Fail stop errors + SDCs

- The **execution** is a multi-stage pipeline, **workflows** (graph)
- **Producers** and **consumers** components
- Communications as **streams** (Unidirectional) BW components
- **Bidirectional** (burst) **communications** inside components
- **Heterogeneous** parallel **applications** (some tightly coupled, some loosely, different nature, different #processes, etc.)
- Performance → implement **communication BW components in memory**
- Potentially **Heterogeneous Hardware/software**
- **Different user recovery needs** depending on where/when the fault happened
- ***Static versus dynamic*** *in situ data analytics*

Mix of tightly
Coupled and loosely
Coupled stages
(**simulation in orange**)



Multiple failure
Scenarios:
-simulation fails
-2cd stage fails
-Multiple stages fail
-corruption in
Simulation
-corruption in final
stage

What are the main technical issues?

→ How users **express** their resilience needs/expectations?

→ How do we **handle fail stop errors?**

→ Checkpoint? How to capture the state of a gigantic workflow? Can we?

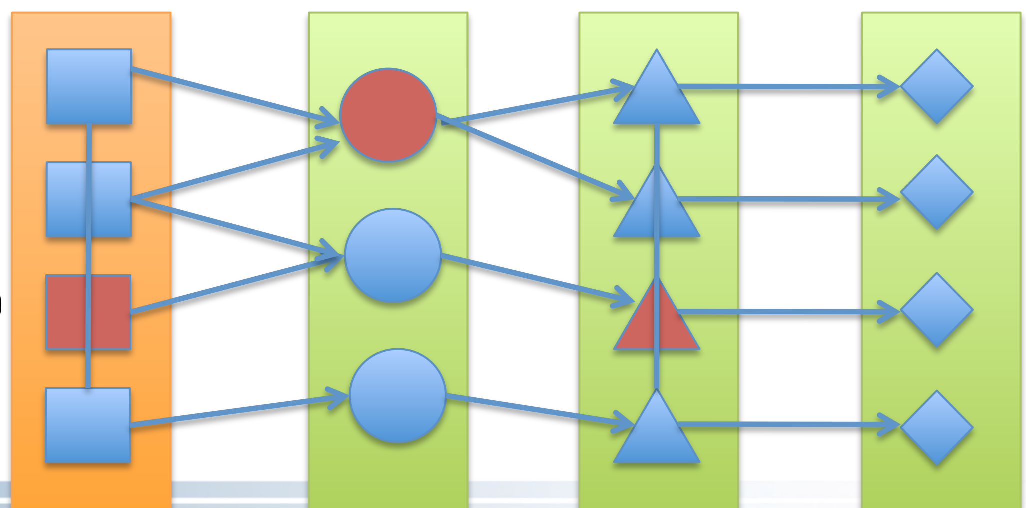
→ Restart?, from where: beginning?, simulation checkpoint? Workflow state?

→ How do we **prepare for SDCs?**

→ Don't care?, try to detect as much as possible?, depends on the components?, on the location of the component in the graph?

→ Do we use replication in the data analytics modules?, ABFT for data analytics ? Approximate computing? More robust hardware?

Mix of tightly
Coupled and loosely
Coupled stages
(**simulation in orange**)



Multiple failure
Scenarios:
-simulation fails
-2cd stage fails
-Multiple stages fail
-corruption in
Simulation
-corruption in final
stage

Why is this different?

- != Large scale parallel execution (bidirectional communication, homogeneous)
- != Workflows on GRID (loosely coupled, intermediate storage on disk, security)
- != Coupled Applications (CESM, etc: Interaction symmetry, global checkpoint)

At least 4 new resilience problems/dimensions for the BDEC roadmap:

1) **Understand** the effects of SDC on the workflow results.
   - Depending on the data product, the combination of resolution and location in the workflow may make some data products more sensitive to SDCs than others.
2) **Establish clear response modes** with respect to failure modes + user needs
   - Depending on the failure type (FS+SDC) and on where it happens in the workflow,

     static versus dynamic in situ analytics
   - Is speculative execution of a module during the recovery of another of interest?
3) **Design** workflow components & coupling methods
   - Maximize performance AND at the same time maximize failure containment
4) **Architect** the right fault tolerance approach for each component and for the workflow as a whole → more than a problem of orchestration: optimization

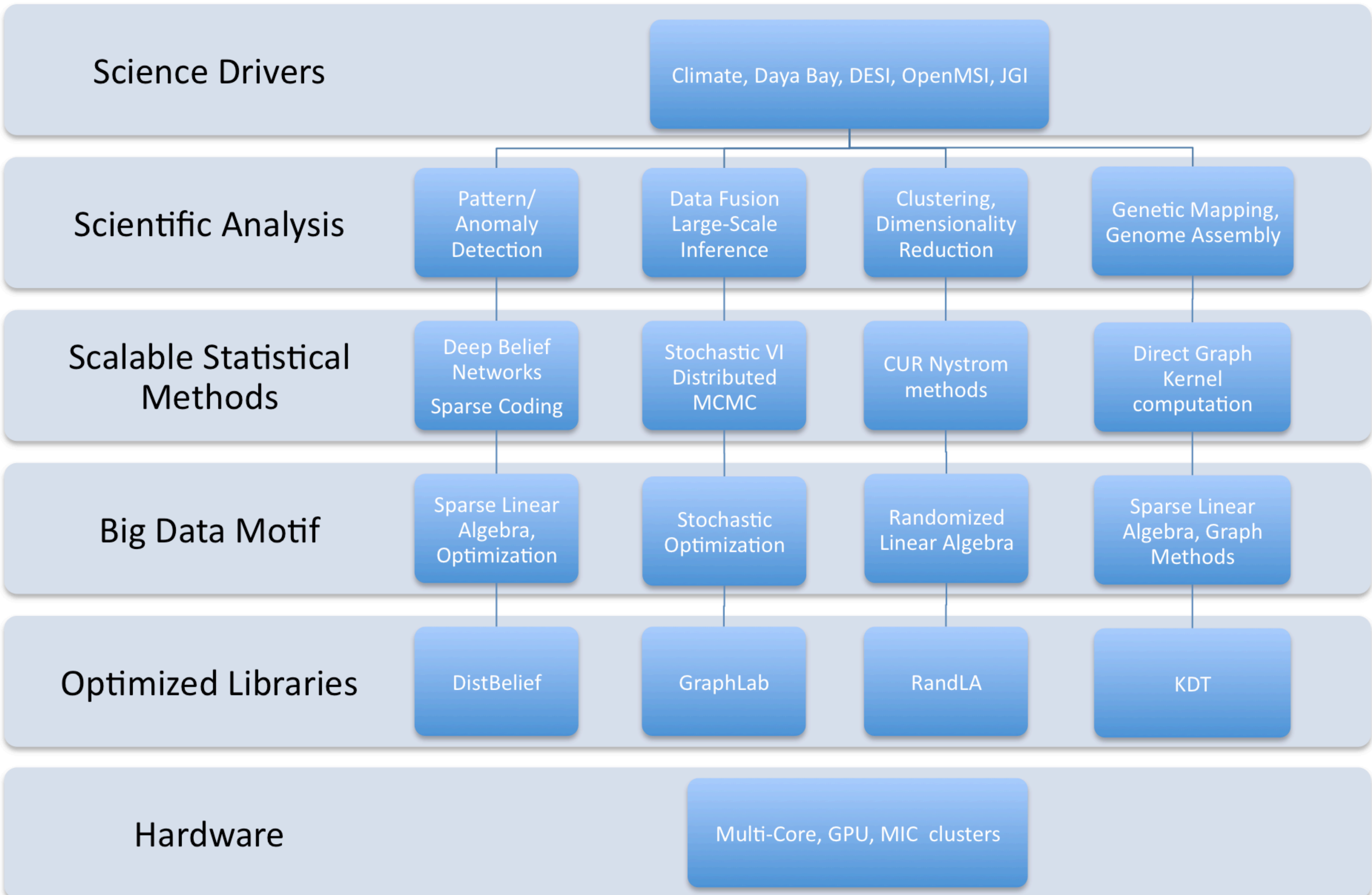
Holistic View of Composable Data Analysis: Insights From Software Frameworks for Extreme Scale Computing

Anshu Dubey, W. Bethel, Prabhat, J. Shalf, A. Shoshani, B. Van Straalen

Scientific Process Closed Loop

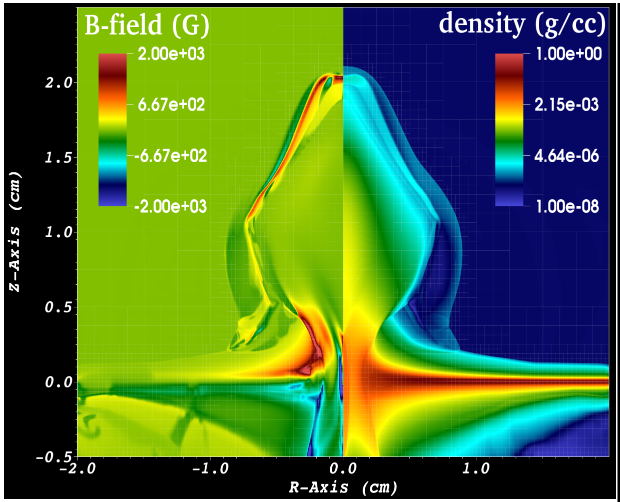
- There is a hypothesis
  - Experiments, observations and/or simulations are designed around the hypothesis.
    - Often complex **multi-stage** data analysis involved
    - Analysis might lead to a new hypothesis
    - Process is repeated
- Data analysis and curation has become comparable or even bigger exascale challenge than simulations
- Workflows for big data and extreme computing share many characteristics
  - Many stages in the computations, different algorithms for each stage
    - Diverse and often conflicting demands from system resources
    - Interoperability is a challenge

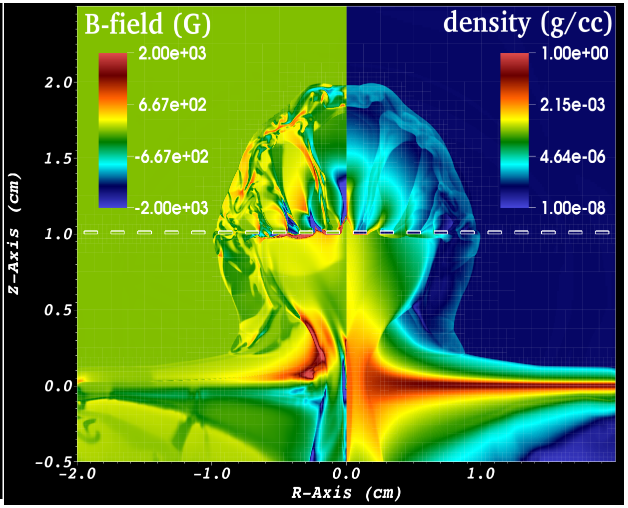
Big Data Analytics Stack

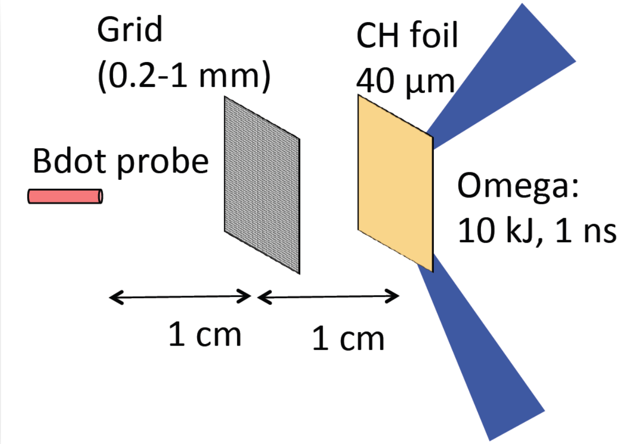


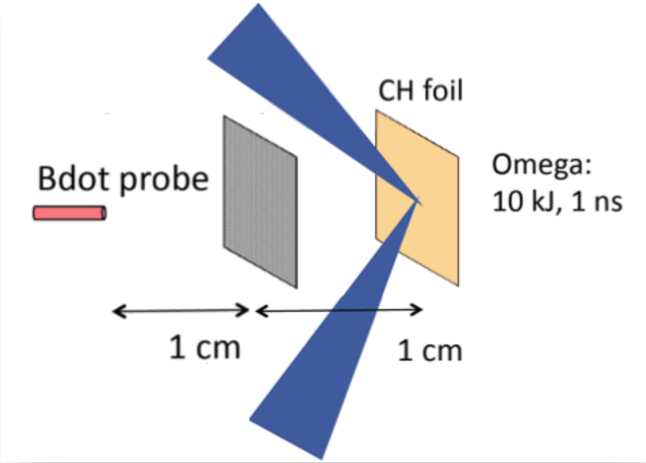
Experimental Validation and Design Through Simulations

- Data plays the role of intermediary
  - Stream of data from experiments and simulations
- Gregori et al. (2012) demonstrated in the laboratory the generation of magnetic fields by asymmetric shocks – a widely invoked mechanism for the creation of seed fields in the universe
- Higher magnetic Reynolds number needed in the experiments for the next step
  - Increased laser energy
- Use FLASH Simulations of two configurations to design experiments









Images from The Flash Center for Computational Science

Publications: http://www.sciencedirect.com/science/article/pii/S157418181200095X
http://www.sciencedirect.com/science/article/pii/S1574181812001280
http://www.sciencedirect.com/science/article/pii/S157418181200119X

BERKELEY LAB
LAWRENCE BERKELEY NATIONAL LABORATORY

Insights from Petascale Computations

- Takes a combination of robust software design, hard-nosed trade-offs and careful orchestration
- Software Design:
  - Separating algorithmic concerns from infrastructure
  - Reusable components
  - Well designed, extensible interfaces
  - Framework for composability
- Trade-offs:
  - Also consider sub-optimal solutions for components
    - Algorithms and implementations
    - Example of a simulation campaign: http://hpc.sagepub.com/content/27/3/360
- Orchestration:
  - Take a holistic view of the solution
  - Leverage heterogeneity and
  - Expose optimization possibilities during design

From Simulations to Numerical Laboratories

Alex Szalay (JHU)

- HPC is an instrument in its own right
  - *Largest simulations approach/exceed petabytes*
- Need public access to the best and latest
- Also need ensembles of simulations for UQ
- Creates new challenges
  - *How to access the data?*
  - *What is the data lifecycle?*
  - *What are the analysis patterns?*
  - *What architectures can support these?*
- On Exascale everything will be a Big Data problem

Usage Scenarios for Big Simulations

- Huge variations in data lifecycle
  - *On-the fly analysis* *(immediate, do not keep)*
  - *Private reuse* *(short/mid term)*
  - ***Public** reuse* *(mid term)*
  - ***Public** service portal* *(mid/long term)*
  - *Archival and curation* *(long term)*
- Very different from supercomputer usage patterns
- Not every data set is equally important!
- Important data sets are naturally emerging
- Opportunity to build network of data resources

Numerical Laboratories

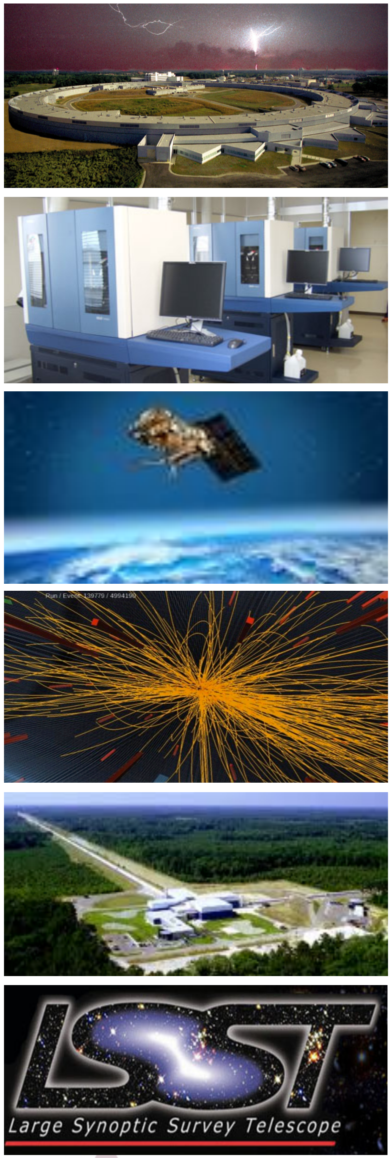
- Similarities between Turbulence/CFD, N-body, ocean circulation and materials ccience
- Differences as well in the underlying data structures
  - *Particle clouds / Regular mesh / Irregular mesh*
- Innovative access patterns appearing
  - *Immersive virtual sensors/Lagrangian tracking*
  - *User-space parallel operators, mini workflows on GPUs*
  - *Posterior feature tagging and localized resimulations*
  - *Machine learning on HPC data*
  - *Joins with user derived subsets, even across snapshots*
  - *Data driven simulations/feedback loop/active control of sims*

Architectual Challenges

- How to build a system good for the analysis?
- Need to define razor sharp tradeoffs
  - *Cannot build a system that is everything for everybody*
  - *BDEC system is different from supercomputer*
- Need high bandwidth to data
  - *Computations/visualizations must be on top of the data*
  - *For subsetting also need fast random access*
- Lessons from the database world:
  - *It is hard to schedule complex I/O patterns*
  - *For subsets we must use indexing, cache resilient storage*
  - *Complex architecture => use a declarative language, the users should tell **what** to do but not how to do it*
- Big Data in simulations more structured than commercial

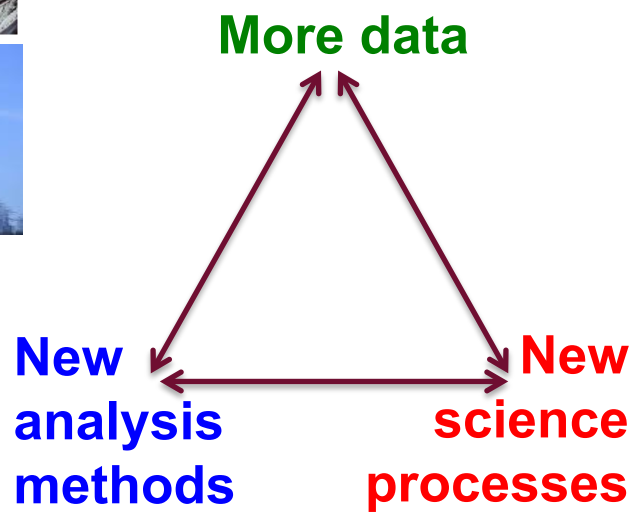
Extreme-scale computing for new instrument science

Ian Foster, Argonne National Laboratory and University of Chicago



New sensors with high data rates
High-performance simulations
Multi-modal data
Databases and knowledge bases
Scientific literature

Contact:
foster@anl.gov
compinst.org
globus.org
ianfoster.org
@ianfoster



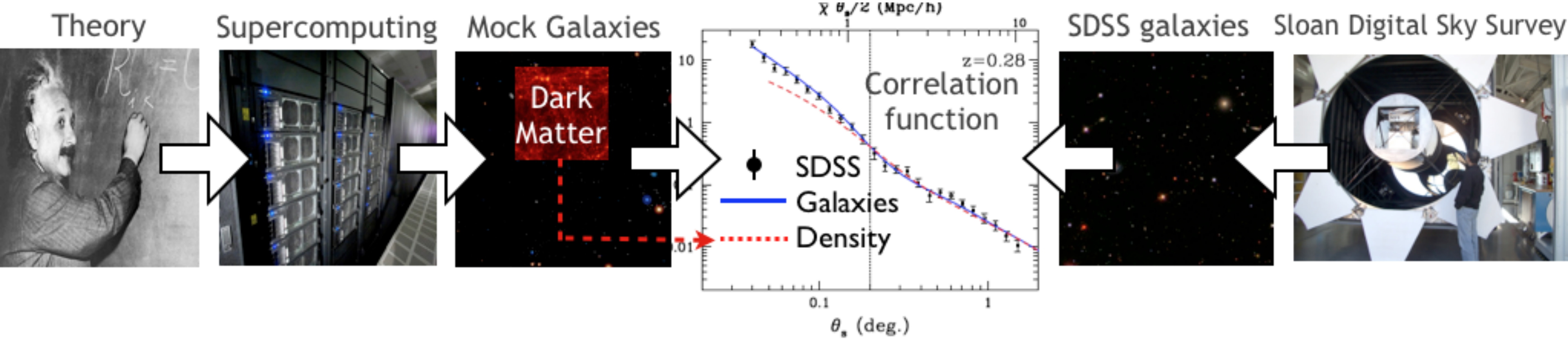
Automated feature detection
Flag interesting events
Real-time data integration
Classification, clustering, etc.

Online quality control
Integrate observation, simulation
Knowledge-based feedback
Knowledge-based control

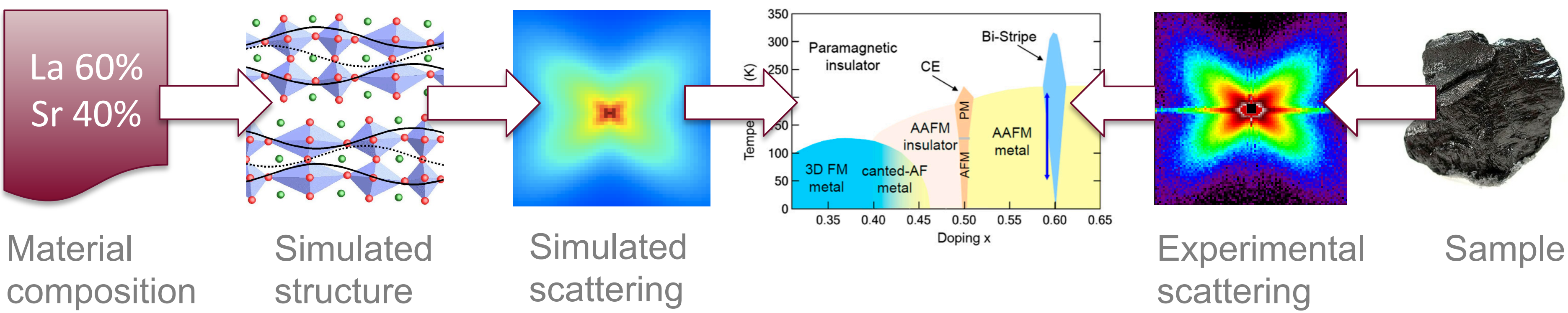
Based in part on discussions within DOE “Accelerating Scientific Knowledge Discovery” group: Deborah Agarwal, Amber Boehnlein, Ian Foster, Barbara Jennings, Scott Klasky, Kerstin Kleese-Van Dam, Ruth Pordes, David Skinner

Discovery Engines for Big Data: New knowledge by coupling observation and simulation

Cosmology: The study of the universe as a dynamical system



Materials science: Diffuse scattering to understand disordered structures

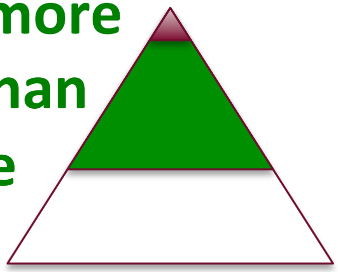


Discovery engine = Advanced instruments + large knowledge bases + extreme-scale computing + collaborative groups

Images from Salman Habib et al. (HEP, MCS, etc.) and Ray Osborn et al. (MSD, APS, etc.)

Discovery engines and extreme-scale computing

Reach many more researchers than extreme-scale simulation



Urgent research agenda

Knowledge management and fusion

Rapid knowledge-based response

Human-centered science processes

Challenges for exascale technologies

Reliable, secure, high-speed system integration beyond the machine room

On-demand scheduling to match with human decision taking timelines

New computational problems that stress computer architectures in new ways

Supporting Big Data @ NAS

Piyush Mehrotra
L. Harper Pryor
NASA Advanced Supercomputing (NAS) Division), NASA Ames
{piyush.mehrotra,laura.h.pryor}@nasa.gov

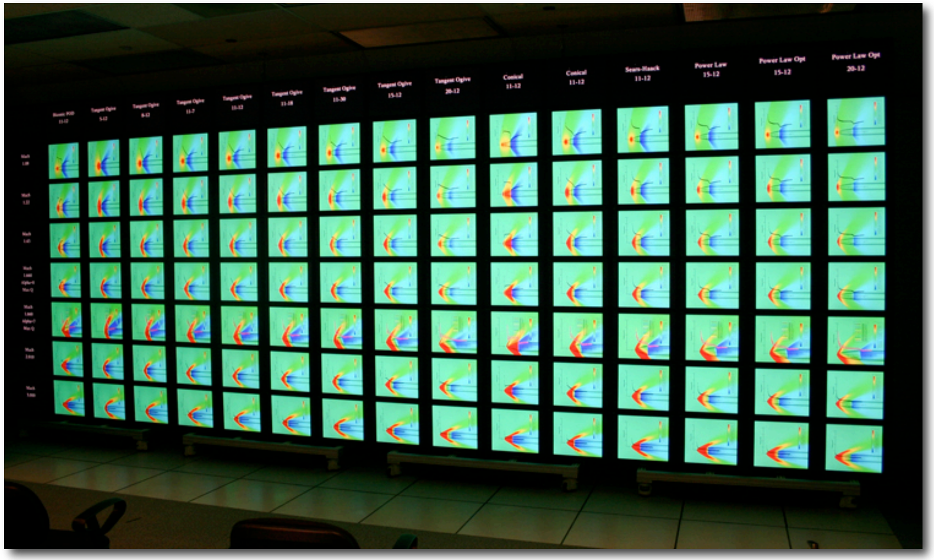
- NASA has enormous collections of observational and model data
- Observational Data:
  - Estimate 100+ active satellites producing 50PBs per year
    - Solar Dynamics Observation (SDO) satellite produces 1 GB per minute => > 1/2 PB/ year ; ~ 3PB in its 5 year life cycle
  - NASA Earth Science operates 12 DAACs (archive centers); National Space Science Data Center
- Model Data:
  - NAS has 20+ PB storage; 115 PBs archive storage & archiving 1+ PB per month
    - MITgcm 35K core run produced 1.44 PB in its 5 day run; full run will produce 9-18 PB; adding bio-geo-chemistry will increase data 100-fold

Fun Fact: The term "Big Data" was first used by Michael Cox & David Ellsworth of NAS in a paper: "*Visualizing flow around an airframe*" Visualization 97, Phoenix AZ.

- Biggest data set considered 7.5GB; high-end analysis machines had less than 1GB memory

Advanced Visualization: hyperwall-2 and CV

- Supercomputer-scale visualization system to handle massive size of simulation results and increasing complexity of data analysis needs
  - 8x16 LCD tiled panel display (23 ft x 10 ft)
  - 245 million pixels
  - Interconnected to NAS supercomputer via IB
- Two primary modes
  - Single large high-definition image
  - Sets of related images (e.g., a parameter space of simulation results)
- *<u>Traditional Post-processing:</u>* Direct read/write access to Pleiades filesystems eliminates need for copying large datasets
- *<u>Concurrent Visualization:</u>* Runtime data streaming increases temporal fidelity at much lower storage costs:
  - ECCO: images every integration time step as opposed to every 860+ time steps originally



NASA Earth Exchange (NEX)

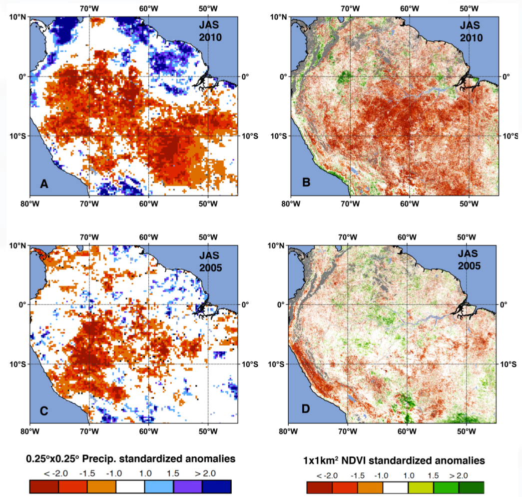
Collaborative Computing for Earth Science

VISION

To engage and enable the Earth science community in addressing global environmental challenges

GOAL

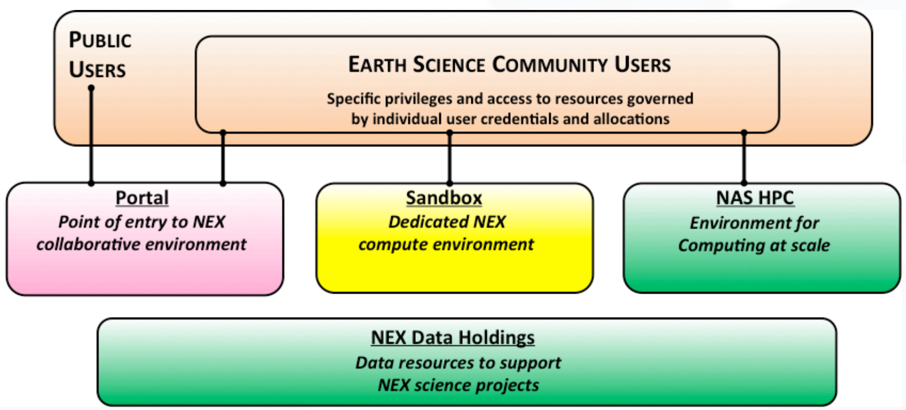
To improve efficiency and expand the scope of NASA Earth science technology, research and applications programs



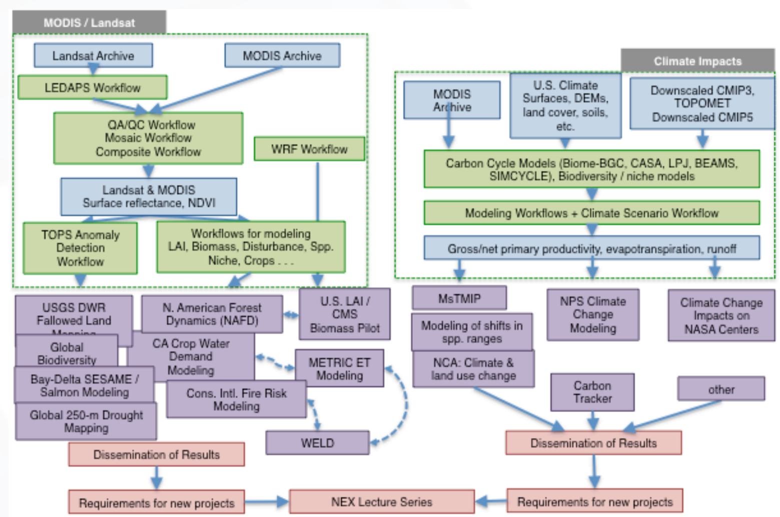
A tale of two droughts/Amazon 2005 & 2010

Samantha et al., GRL, 2010
Xu et al., GRL, 2011

Faster (24 months vs. 3 months), consistent (same analytical methods, quality flags) and reproducible;



NEX: Three-tier environment



Representative workflow; tools currently being investigated VisTrails & ParaView

Big Data Effort @ NAS

- Current infrastructure => Big compute:
  - Pleiades #16 on Top500, undergoing augmentation to 3.5 PF; Endeavour – SGI UV nodes 2TB & 4TB; 20+ PB storage; 115 PB of archive storage
- **Big Data Focus:** Develop and implement a roadmap for an infrastructure to support analysis & analytics
  - Conducted survey of projects dealing with big data (available soon)
  - Currently conducting prototype experiments
- Challenges (extracted from survey):
  - Data management – storage/access/transport
  - Data discovery - Indexing/archiving, metadata – requires semantic reasoning
  - Tools/models/algorithms: development & discovery
  - Data Analysis/Analytics infrastructure
    - Most NASA data is structured, gridded, geospatial
    - Shared memory systems with large I/O pipes; data preferably co-located with compute
    - Visualization support
  - Workflow to tie all components together
  - Collaboration environments
    - Dissemination and sharing of results/tools/models/algorithms