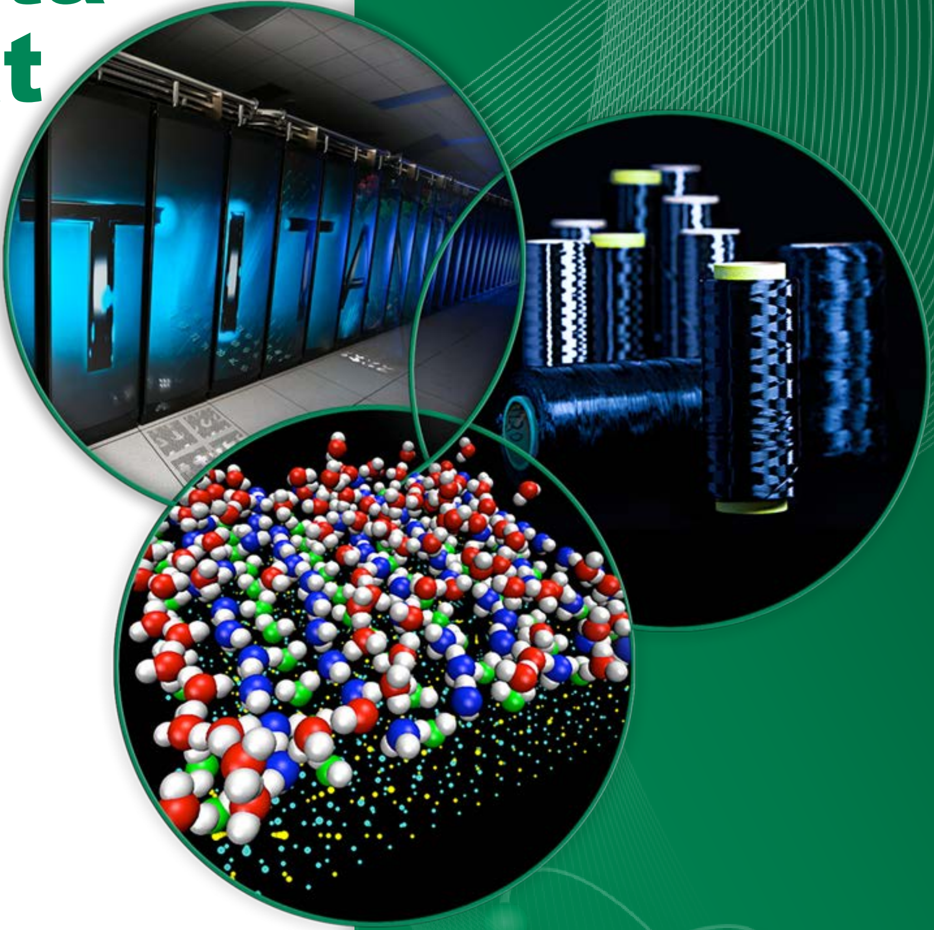
# The Future of Data and Computing at ORNL



Barney Maccabe

April 29, 2013

# DOE is a world leader in HPC

## *- leaders in architecting, acquiring and deploying computers used as instruments of discovery*

- A record of leadership in computing performance
- Thousands of users from government labs, universities, and industry
- Forefront computing facilities
  - Used to solve mission problems in nuclear defense, science, and engineering
  - Enabling prize-winning science (e.g., Nobel), engineering solutions, and thousands of publications

TOP500
November 2012

Titan at ORNL
(#1, 17+ PF)

Sequoia at LLNL
(#2, 16+ PF)

Mira at ANL
(#4, 8+ PF)

Cielo at LANL/SNL
( #18, 1+PF)

Hopper at LBNL
(#19, 1+PF)

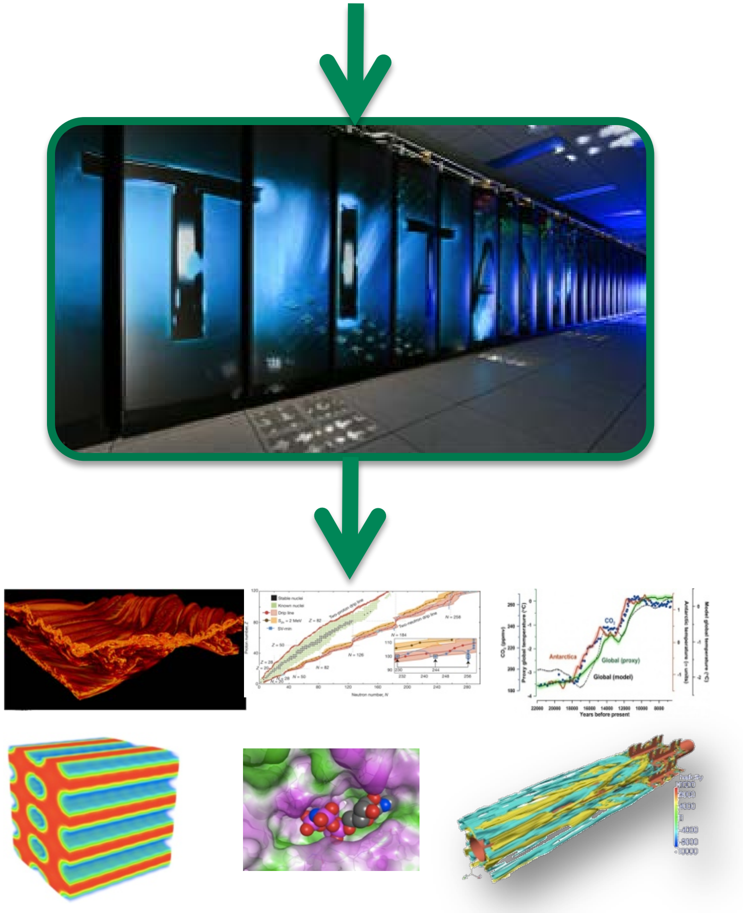
# Scientific Computing as an Instrument

$$\left(-\nabla^2+V\right)\Psi=E\Psi$$

$$\Psi=-\left(-\nabla^2-E\right)^{-1}V\Psi$$

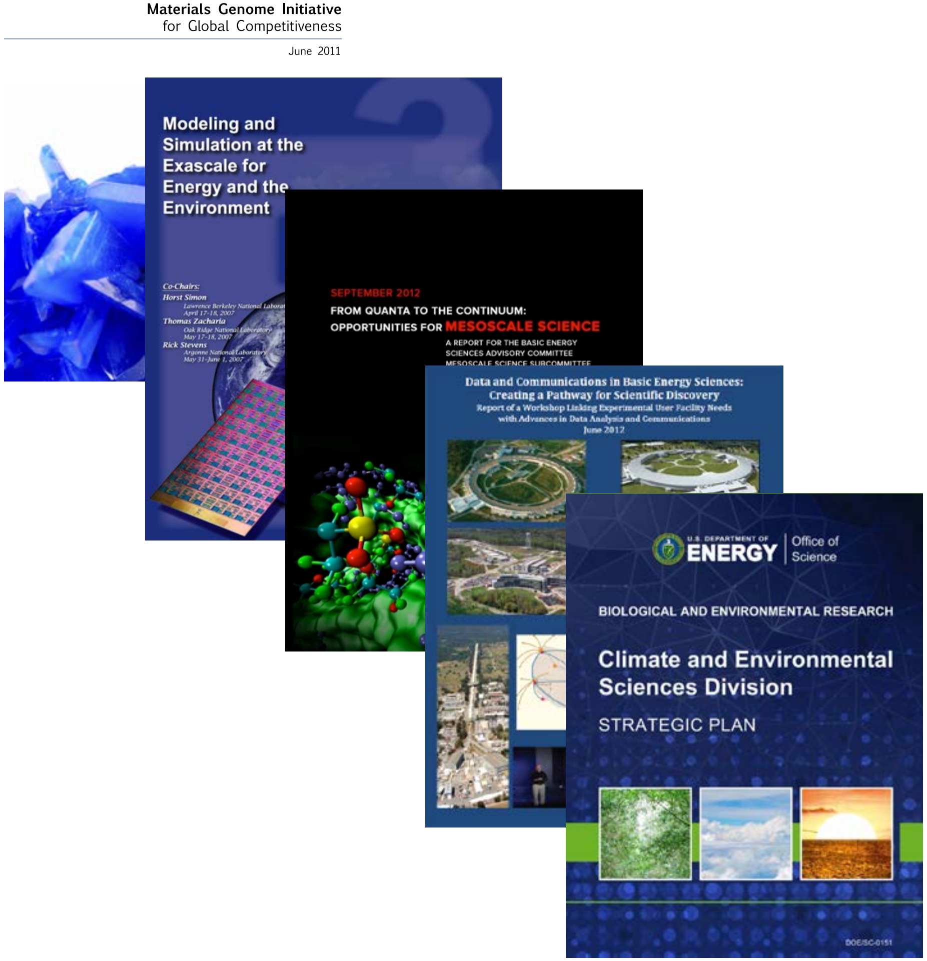
$$=-G*\left(V\Psi\right)$$

$$\left(G*f\right)(r)=\int ds\,\frac{e^{k|r\;\;s|}}{4\pi|r-s|}f(s)\ \text{ in 3D };k^2=-E$$



- Over the past few decades we have developed world-class facilities for scientific simulation
- These "simulation instruments" are now integral to theory based scientific discovery
- Titan will support a broad range of theoretic studies in Astrophysics, Climate Science, Materials Science, Nuclear Physics, Fusion, etc.

# Data Infrastructure and Scientific Discovery



Materials Genome Initiative
for Global Competitiveness
June 2011

Modeling and Simulation at the Exascale for Energy and the Environment

Co-Chairs:
Horst Simon
Lawrence Berkeley National Laborat
April 17-18, 2007
Thomas Zacharia
Oak Ridge National Laboratory
May 17-18, 2007
Rick Stevens
Argonne National Laboratory
May 31-June 1, 2007

SEPTEMBER 2012
FROM QUANTA TO THE CONTINUUM:
OPPORTUNITIES FOR MESOSCALE SCIENCE
A REPORT FOR THE BASIC ENERGY
SCIENCES ADVISORY COMMITTEE
MESOSCALE SCIENCE SUBCOMMITTEE

Data and Communications in Basic Energy Sciences:
Creating a Pathway for Scientific Discovery
Report of a Workshop Linking Experimental User Facility Needs
with Advances in Data Analysis and Communications
June 2012

U.S. DEPARTMENT OF ENERGY | Office of Science
BIOLOGICAL AND ENVIRONMENTAL RESEARCH
Climate and Environmental Sciences Division
STRATEGIC PLAN
DOE/SC-0151

- To date, our primary focus in has been simulation
- However, the rate of scientific progress is increasingly dependent on the ability to efficiently capture, integrate, analyze, and steward large volumes of diverse data
- Increasing data volume, variety, and velocity require a new environment for scientific discovery
- Major facilities and research programs across the Office of Science must develop adequate infrastructure to support this new environment

# Data Intensive Computing as Integrative Infrastructure

ARM
CLIMATE RESEARCH FACILITY

SPRUCE

CLM
COMMUNITY LAND MODEL

## Observation

- Observation of natural phenomena

## Experiment

- Experimental methods and quantification

## Theory

- Simulation of complex phenomena

**Integration of Complex Data From Observation, Experiment and Theory**

- **Access to data from different modalities**
- **Differing temporal and spatial scales**
- **Differing data representations**

# Office of Science Response To Data Intensive Challenges

- The ASCR (Advanced Scientific Computing Research) program is gearing up to address these needs

  *SC 2013 Budget request*

  - Computer science research

    "FY 2013 supports new research efforts to address the challenges of data-intensive science with focus on full data lifecycle management and analysis for the massive data from DOE scientific user facilities."

  - Computational Partnerships

    "New research efforts will engage partners across the Office of Science to address the data-intensive science challenges at the science application level."

  - Next Generation Networking

    "FY 2013 supports new research efforts to address the data-intensive science challenges facing Request scientific communities using unique DOE facilities and engaging in large-scale collaborations."

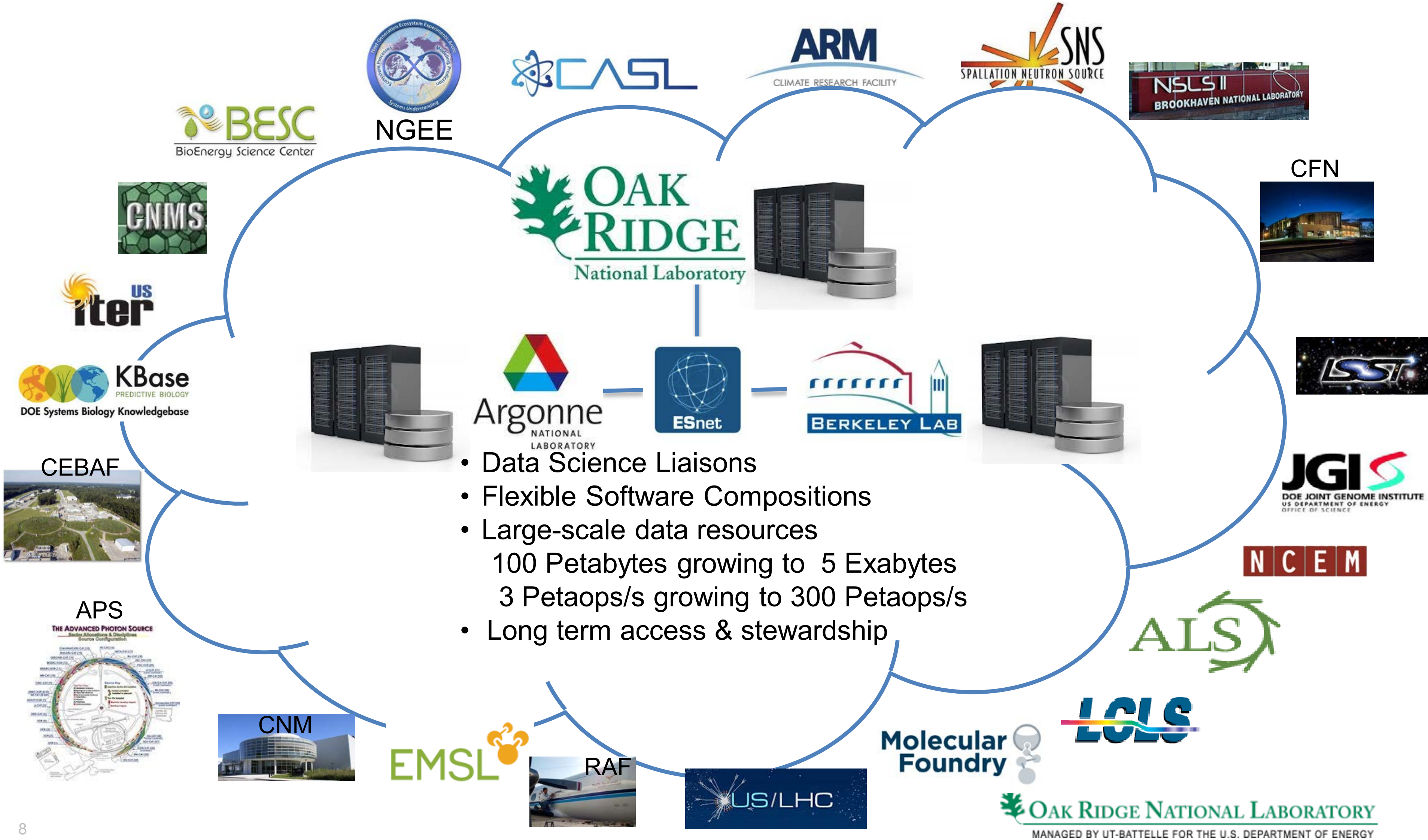
- **What is missing?**
  - **Facilities to support the data challenges facing the Office of Science**

# A “Virtual Data Facility” for Office of Science

- The Virtual Data Facility will provide these capabilities
  - Rich data analysis environment - tightly coupling compute and data storage
  - Coupling simulation, experiment, and observation data
    - Model improvement, validation, steering, site selection, etc.
  - A flexible compute and data environment that can be tailored to specific needs
  - Cataloging and long-term stewardship of scientific datasets
  - Long-term (many year) allocations for major facilities and projects
- The Virtual Data Facility will be built as an extension of the three compute facilities at Argonne, Berkeley, and Oak Ridge
  - Core ASCR facilities (OLCF, ALCF, and NERSC) have already been designed to meet the needs of computational science (simulation) workloads
  - These facilities will be extended to meet the data science needs
    - Leverages facilities investments and staff expertise at the OLCF, ALCF, NERSC, and ESnet
- The Virtual Data Facility will  create a new environment for scientific discovery for the office of science (ASCR, BER, BES, NP, FES, HEP, NE, etc.)

# The Virtual Data Facility

- Integrated Data Science Infrastructure across ASCR Facilities in support of major projects and facilities



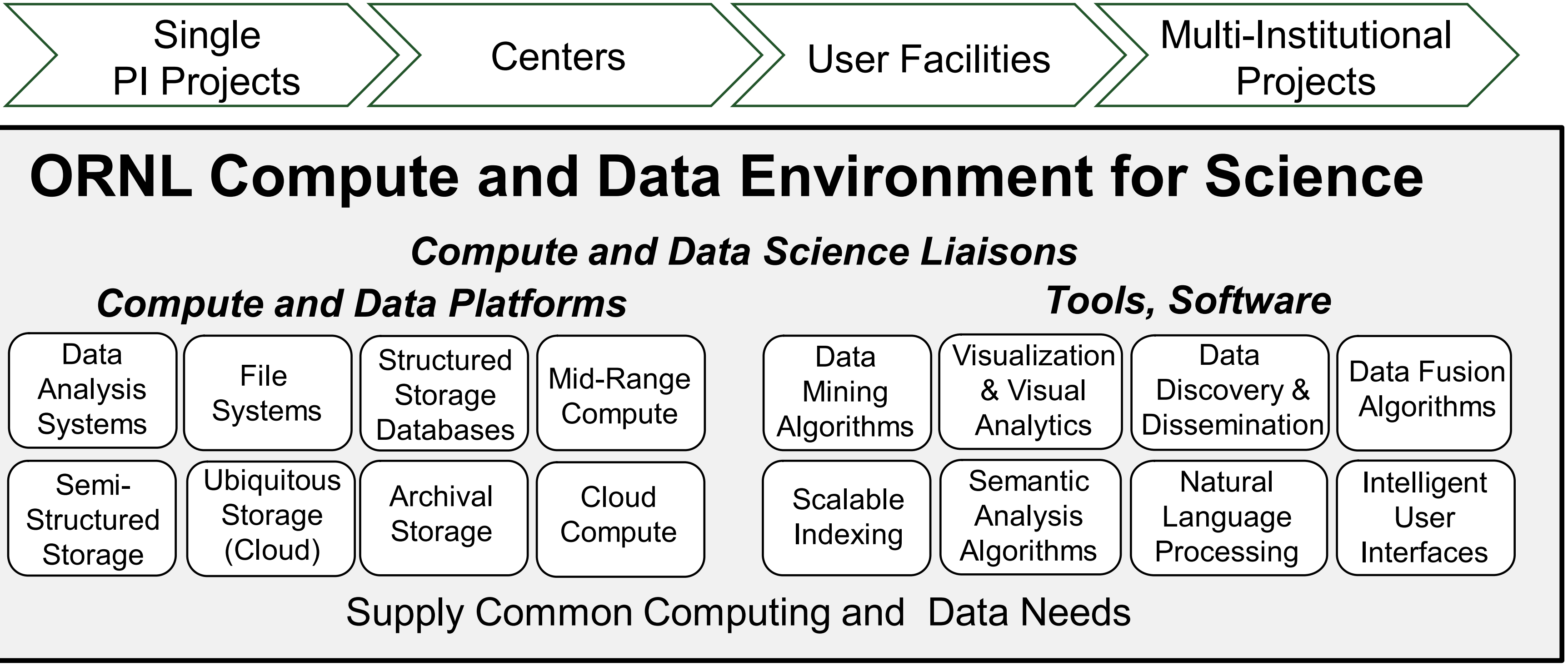
- Data Science Liaisons
- Flexible Software Compositions
- Large-scale data resources
  100 Petabytes growing to 5 Exabytes
  3 Petaops/s growing to 300 Petaops/s
- Long term access & stewardship

Oak Ridge National Laboratory
Argonne National Laboratory
ESnet
Berkeley Lab
BESC
BioEnergy Science Center
NGEE
CASL
ARM
Climate Research Facility
SNS
Spallation Neutron Source
NSLS II
Brookhaven National Laboratory
CFN
CNMS
US iter
KBase
Predictive Biology
DOE Systems Biology Knowledgebase
CEBAF
APS
The Advanced Photon Source
CNM
EMSL
RAF
US/LHC
Molecular Foundry
LCLS
ALS
NCEM
JGI
DOE Joint Genome Institute
US Department of Energy Office of Science
LSST

# Data Science and Compute Strategy at ORNL

- Consolidate resources and develop a flexible and elastic infrastructure to satisfy needs of existing and future computing and data projects
- Establish a Compute & Data Environment for Science (CADES)
- Provide competitive advantage in R&D and for national data centers (DSF)

Compute and Data Environment for Science targets full range of needs



Single PI Projects

Centers

User Facilities

Multi-Institutional Projects

**ORNL Compute and Data Environment for Science**

***Compute and Data Science Liaisons***

***Compute and Data Platforms***

Data Analysis Systems

File Systems

Structured Storage Databases

Mid-Range Compute

Semi-Structured Storage

Ubiquitous Storage (Cloud)

Archival Storage

Cloud Compute

***Tools, Software***

Data Mining Algorithms

Visualization & Visual Analytics

Data Discovery & Dissemination

Data Fusion Algorithms

Scalable Indexing

Semantic Analysis Algorithms

Natural Language Processing

Intelligent User Interfaces

Supply Common Computing and Data Needs

# CADES Infrastructure

## People

**Matrix staff with expertise in all areas of compute and data science**

- Domain-specific compute and data science
- Algorithms, data analysis, and visualization
- Compute and data system architecture
- Operations

## Flexible and Elastic Resources

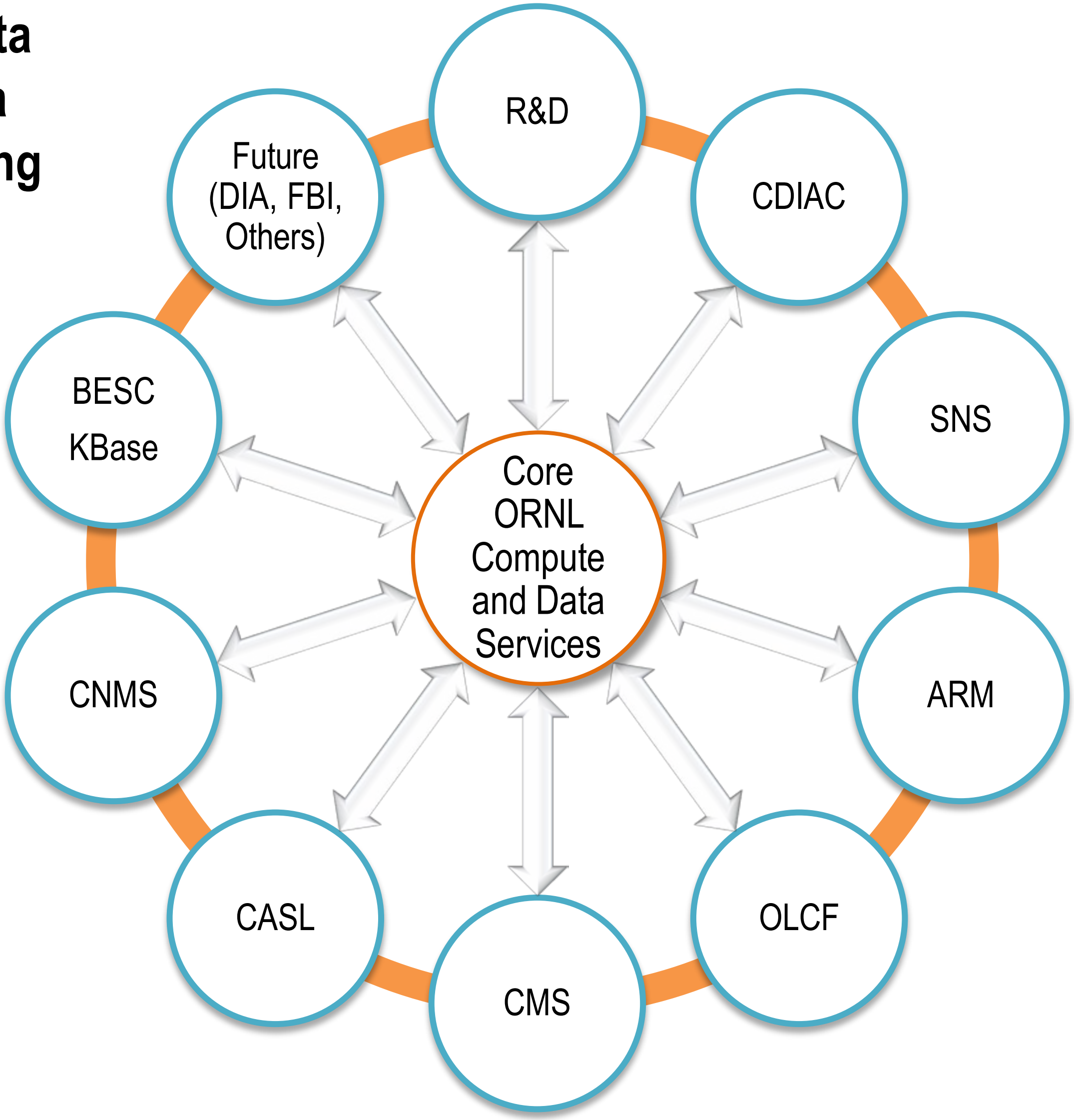
- Flexible to meet requirements of a broad set of initiatives
  - Performance, scalability, manageability, security
  - Compute: From enterprise servers and clusters to specialized systems
  - Storage: From network attached storage and parallel file systems to archive
  - Software: From commercial or community to custom packages
- Elastic to meet on-demand compute and storage requirements for data-intensive workloads
  - Projects no longer constrained by a fixed system resource

# Providing Expertise in Compute & Data For Mission

**CADES is like a hub -- it shares data infrastructure and Compute & Data Science capabilities with and among many projects**

**Projects with core competencies in Compute and Data Science can share these expertise via the HUB**

**CADES is delivering to initiatives today and provides the necessary capabilities required by the SC Data Science Facility in the future**

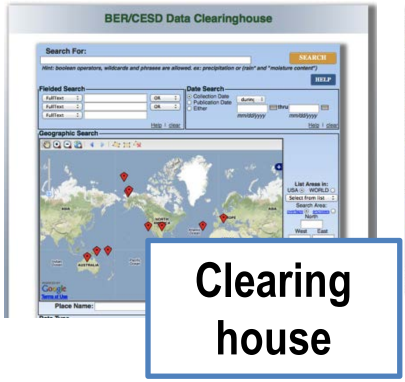

R&D
CDIAC
SNS
ARM
OLCF
CMS
CASL
CNMS
BESC
KBase
Future (DIA, FBI, Others)
Core ORNL Compute and Data Services

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

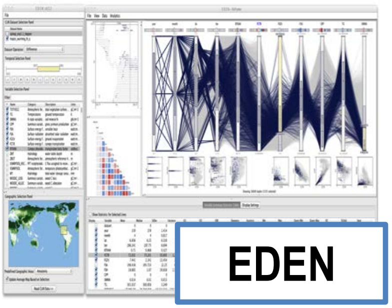
# CADES – Delivering to Initiatives Today

## Scalable tools and data infrastructure for climate science

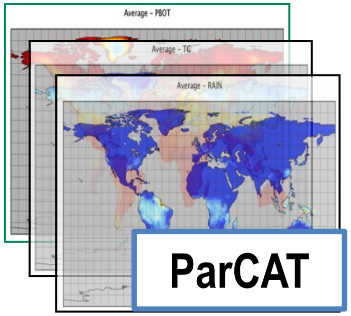
- Data discovery and dissemination tools
- Scalable data analytics and data fusion



Clearing house



EDEN



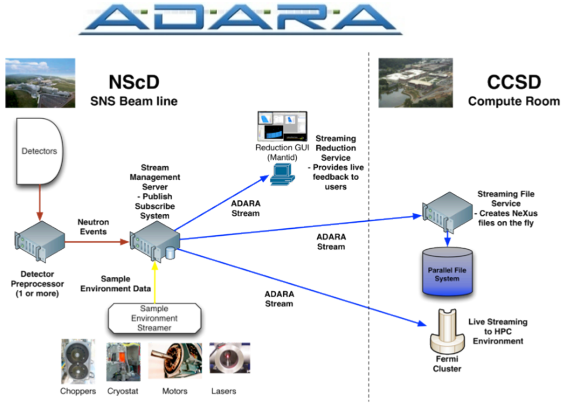
ParCAT

ESGF

Obs4MIPs

## An integrated compute and data environment for neutron science

- Provide live feedback from experiment through streaming data capture and reduction
- Integrating simulation (MD/DFT) directly into the neutron scattering data analysis chain



ADARA

NScD
SNS Beam line

CCSD
Compute Room

Detectors

Detector Preprocessor (1 or more)

Neutron Events

Stream Management Server
- Publish Subscribe System

Sample Environment Data

Sample Environment Streamer

Choppers
Cryostat
Motors
Lasers

Reduction GUI (Mantid)

Streaming Reduction Service
- Provides live feedback to users

ADARA Stream

ADARA Stream

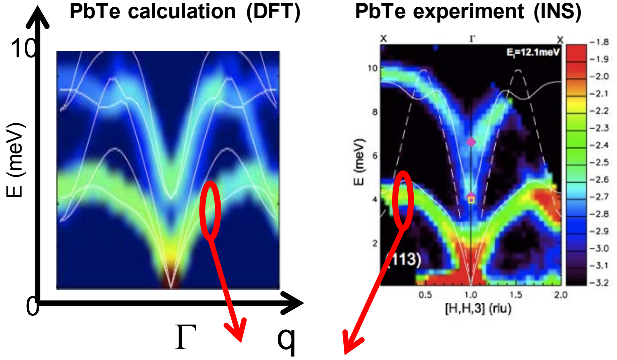
ADARA Stream

Streaming File Service
- Creates NeXus files on the fly

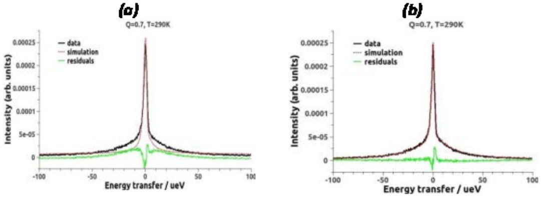
Parallel File System

Fermi Cluster

Live Streaming to HPC Environment



Example: , *ab-initio* MD simulations for ferroelectrics/thermoelectrics. Focus on *width* of dispersions



Fit between experiment (black) and simulated (red) dynamics structure factors for water-hydrogen.

# Case Study: Neutron Scattering
## Enabling near real-time feedback from experiment

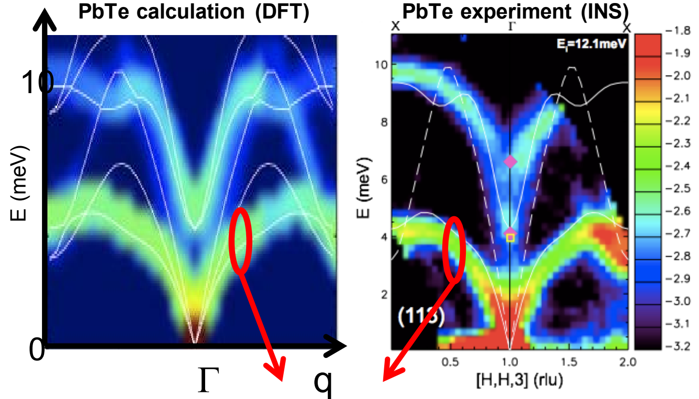
| Challenge | Response | Status |
|---|---|---|
| Realizing the full potential of SNS requires near real-time feedback to users and integration of experiment and simulation/modeling | • Accelerating Data Acquisition, Reduction and Analysis (ADARA)<br>– Leverages ORNL'S neutron scattering and computational expertise<br>– Stream data to computational resources and provide live feedback from experiment in real-time<br>– High performance data backplane for reduction, analysis, and coupling with simulation; forms basis for future work to integrate experiment and simulation | • ADARA is being field tested on the SNS HYSPEC beam line<br>–Complete field testing on HYSPEC beam line<br>–Continue development and deployment of ADARA on subsequent SNS beam lines |

# Neutron Science: Integrating Simulation In the Data Analysis Infrastructure

- The CAMM will integrate materials modeling/simulation (MD/DFT) directly into the chain for neutron scattering data analysis, **offline** and **online** (in near real time)

- Developing workflows for refinement, integration of MD codes, **neutron scattering corrections** ..

- The CAMM is working with ORNL's Materials Science and Technology Division to study coarse grained MD simulations of polymers PEO-AA (CNMS), *ab-initio* MD simulations for ferroelectrics/thermoelectrics



Example: *ab-initio* MD simulations for ferroelectrics/thermoelectrics. Focus on *width* of dispersions

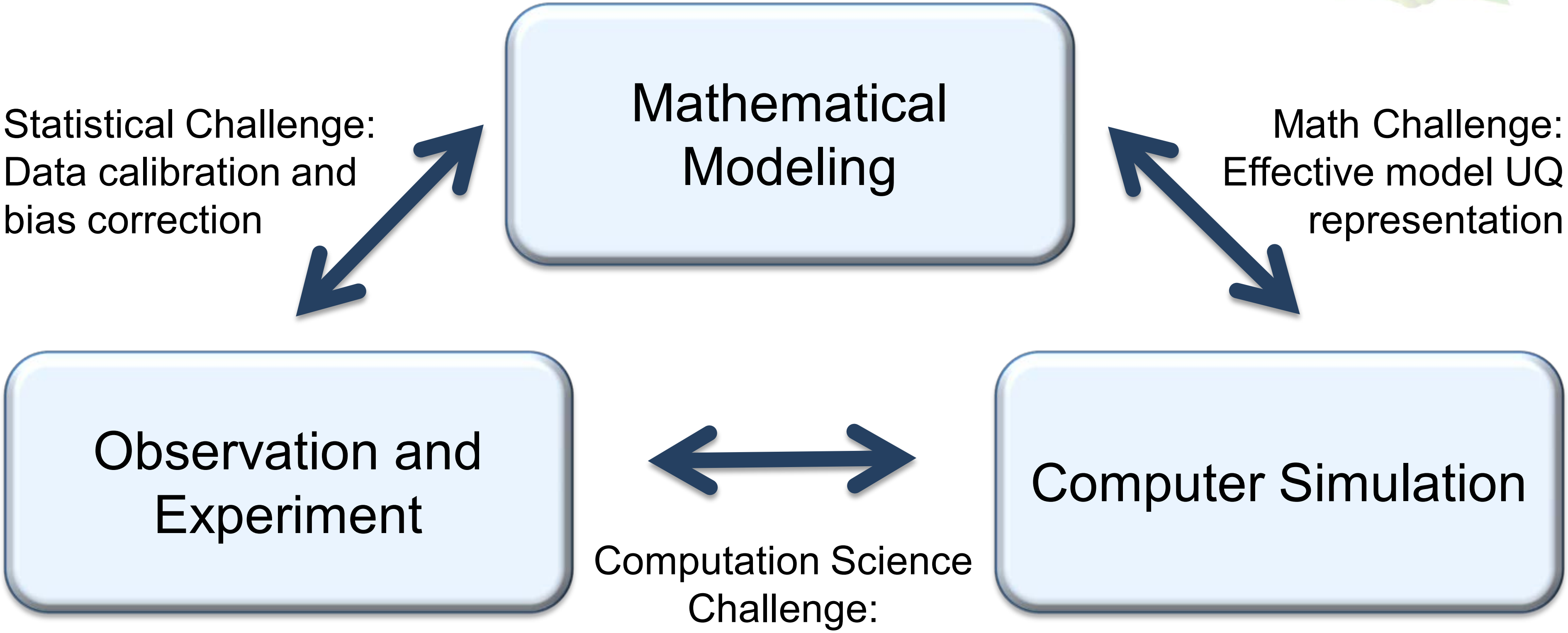
**The Center for Accelerating Materials Modeling (CAMM)**

- *Partnership between ORNL's Neutron Sciences, Physical Sciences and Computing and Computational Sciences Directorates*
- *ORNL SEED money and DOE funds provided to study force field refinement from quasi-elastic and inelastic neutron scattering data*
- *CAMM formed in response to BES proposal call for* Predictive Theory and Modeling

APPLIED & COMPUTATIONAL MATHEMATICS PMG Predictive Methods Group

# Case Study: Neutron Imaging

## Developing mathematical foundations for UQ-V\


Mathematical Modeling

Statistical Challenge: Data calibration and bias correction

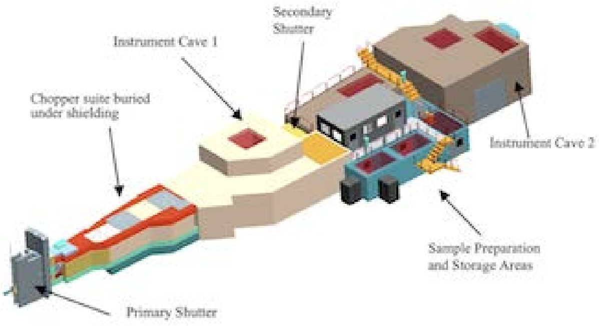
Math Challenge: Effective model UQ representation

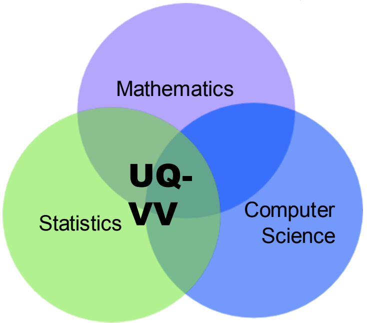
Observation and Experiment

Computer Simulation

Computation Science Challenge: Architecturally effective embedded UQ-VV

VENUS Neutron Imaging


Secondary Shutter
Instrument Cave 1
Chopper suite buried under shielding
Instrument Cave 2
Sample Preparation and Storage Areas
Primary Shutter



ORNL Leadership Computing

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

# Neutron Science: An Integrative Data Infrastructure

Neutron Data

SMS

Translation

Automated Reduction

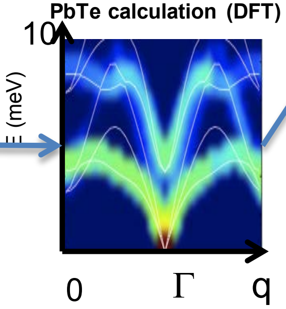
Data Catalog

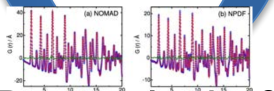
High-performance Scalable Storage

Access, Citation & Data reuse

osti.gov

Simulation





Re-analysis & Derived data Products

# Questions?