BLUE WATERS

SUSTAINED PETASCALE COMPUTING

July 8, 2014

Data @ Scale Working Group – Report

Shaowen Wang

CyberGIS Center for Advanced Digital and Spatial Studies
CyberInfrastructure and Geospatial Information Laboratory
Department of Geography and Geographic Information Science
Department of Computer Science
Department of Urban and Regional Planning
Graduate School of Library and Information Science
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

NCSA Blue Waters Symposium for Petascale Science and Beyond, May 12 – 15, 2014, Champaign, IL, USA

NCSA NSF GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION CRAY

Team

- Jason Alt
- Gregory Bauer
- Michelle Butler
- Kalyana Chadalavada
- Mark Klein
- Bill Kramer
- Shaowen Wang

Data @ Scale

Compute @ Scale

Science @ Scale

Image source: http://blog.allstream.com/six-keys-to-successful-data-centre-convergence/

Scientific Domains

- Astrophysics
- Computational and data sciences
- Computer science
- Climate
- Earthquake
- Geospatial sciences
- Hydrology
- Meteorology
- Molecular dynamics
- Social sciences

Science Drivers for Innovative Data @ Scale Solutions

- Wide range of domains
- Wide range of input/output patterns
  - One file per process, single shared file (100K+ files)
- Wide range of file sizes
  - 1K to TB+
- Wide range of software and tools
  - MPI-IO, NetCDF, HDF, BoxLib, etc.
- Wide range of data transfers
- Wide range of analytics
  - Pre- and post-processing
  - *In situ*
  - Visualization
- Wide range of workflows

Data @ Scale Requirements

- Management
  - Metadata, not just data
  - Explosive growth
- Validation and verification
  - Fault tolerance
- Data movement
- Analysis
- Visualization
- Workflow
- Software and tools
  - Reusable
- Etc.

Every

- **Application**
- **Use case**
- **File size**
- **Format**
- **Etc.**

Performance!

Discussion – General Questions

- What are the major challenges of data handling for your applications?
- What new architecture, software, and tools will likely improve your data @ scale practices?
- What should NSF/UIUC/NCSA be doing to help your sciences achieve desirable data handling?

Questions – Data Movement

- How easy and practical is it to move your data sets today?
- Is it sufficiently fast and simple?
- Are today's software and tools adequate for your data movement needs?
  - If not, what are your recommendations for addressing the inadequacies?

Questions – Data Sharing

- What are your requirements for sharing your data within your community? How about publicly?
- What obstacles do you face that complicate your data sharing?
- How could today's software and tools be improved to advance data sharing capabilities?
- What is missing from today's capabilities?

Questions – Software and Tools

- What are major limitations of current software and tools for your data handling?
- How do these limitations affect your sciences?
- Do you have any suggestions for eliminating these limitations?
- Do you need any software and tools for data handling that are important to your sciences but currently missing?

Recommendations

Addressing Full Life Cycles of Data @ Scale

- Avoid data movement needed for analysis and visualization
- Support data access beyond allocations to maximize scientific analysis and impact
- Enable analytics @ where data are located
  - Dedicated resources for analysis

Data Archival & Sharing

- Provide data repository with efficient access
- Easy and secure data sharing
  - Minimal impact on computational work

Algorithms, Software, and Tools

- Common libraries and utilities for data manipulation @ scale
- Using machine learning to extract data important to sciences out of large generated datasets
- Data compression for efficient storage and transfer
- Software-as-a-service for data analytics @ scale

Education and Workforce Development

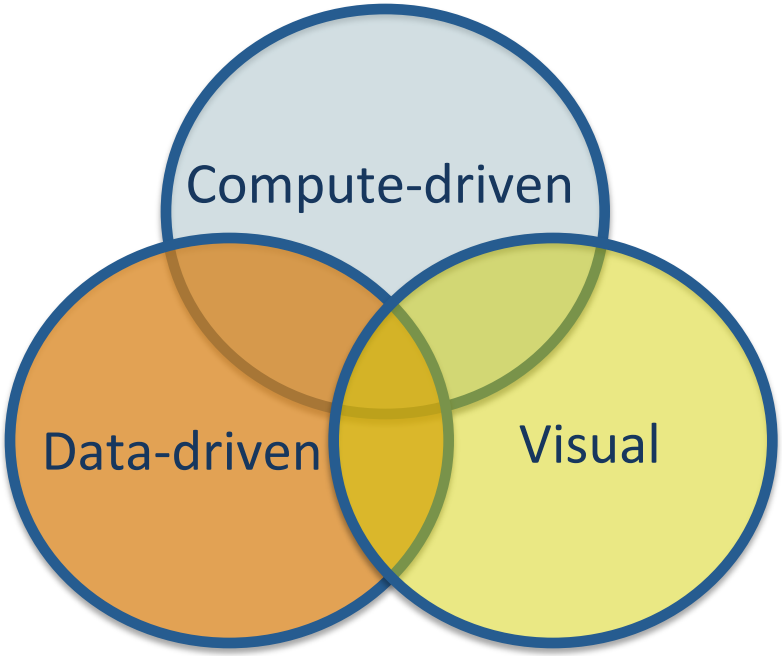
- Improve education of application scientists on capabilities for the state-of-art data management, analysis, and visualization
- Fault tolerance built into applications

Big Challenges and Opportunities

How to fuse petascale (or beyond) data from multiple geographically distributed sites to generate new scientific data products?

How to perform interactive data analytics @ scale for steering simulations?

CyberGIS Workflow



Massive ice melt in Antarctica 'appears unstoppable,' NASA says

What's the pathway forward?

Data @ Scale

Compute @ Scale

Science @ Scale

Image source: http://blog.allstream.com/six-keys-to-successful-data-centre-convergence/

Comments & Questions?

- Thanks!