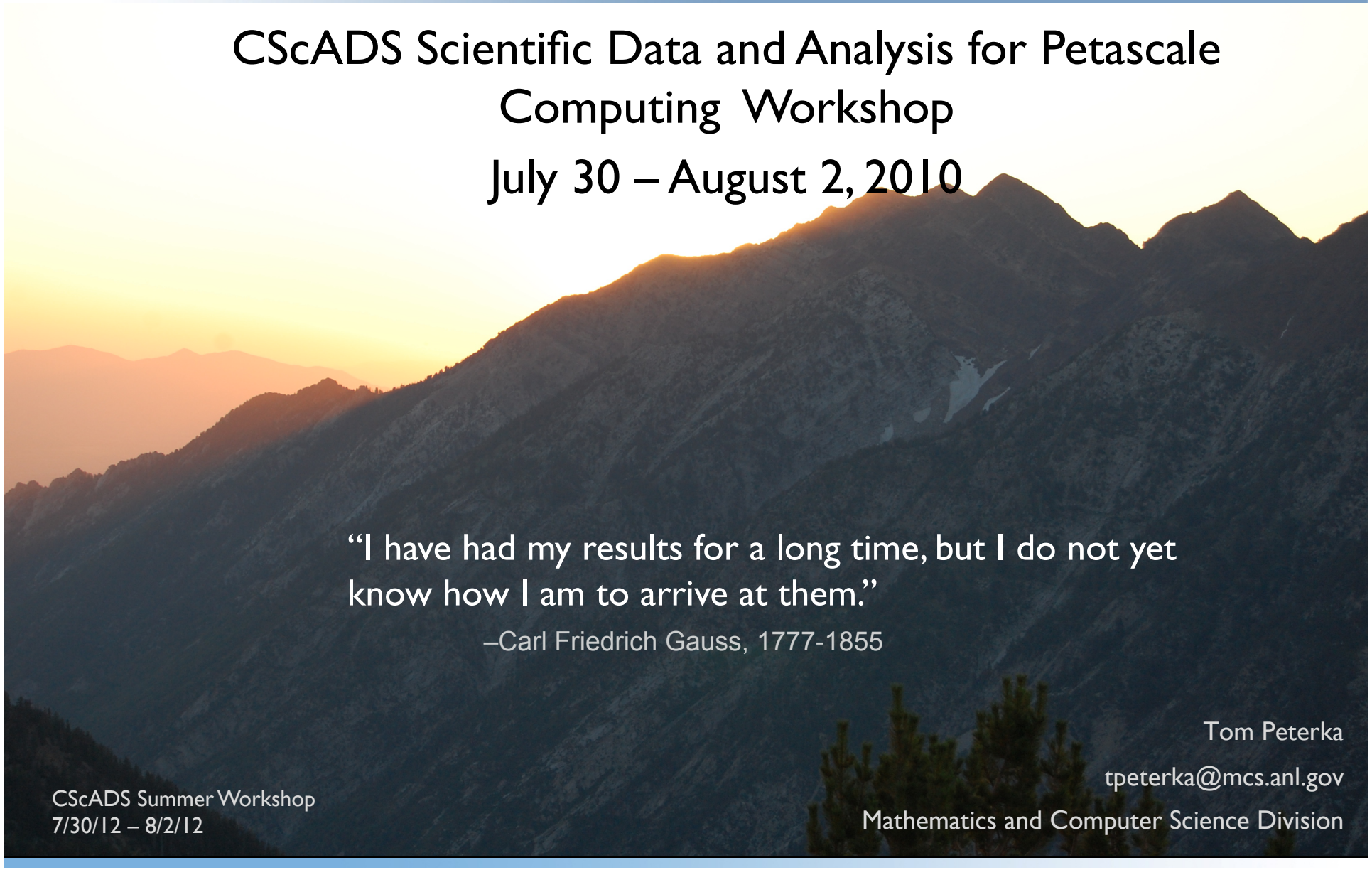
U.S. DEPARTMENT OF ENERGY

Argonne NATIONAL LABORATORY

CScADS Scientific Data and Analysis for Petascale Computing Workshop

July 30 – August 2, 2010



"I have had my results for a long time, but I do not yet know how I am to arrive at them."

–Carl Friedrich Gauss, 1777-1855

Tom Peterka

tpeterka@mcs.anl.gov

Mathematics and Computer Science Division

CScADS

Collaboration between Rice, ANL, UCB, UTK, UW-Madison

Mission is to catalyze software tools that enable applications to achieve scalability on DOE computing platforms

Research, infrastructure development, prototype software, application development, community outreach

6th year of Summer Workshop Series, 5th year of Data and Analytics

Other topics: autotuning, performance tools, petascale applications, libraries and algorithms

Website: http://cscads.rice.edu

Workshop on Data Storage and Analysis for Scientific Discovery

website: http://cscads.rice.edu/workshops/summer-2012/data-analytics

-A little different than past years
  Emphasis on DATA as a first class citizen
  Storing, retrieving, moving, analyzing, visualizing
  Intersection of storage, analysis, computation of data-intensive applications
  Interplay between these; how each can assist the other to scale to peta- and exascale
-Open-ended research topics as well as solutions that can be deployed today
  Data models: what is a data model, how does its definition affect operations on data
  Analysis and visualization: When, where, how, techniques, infrastructures
  I/O systems: Present and future, scalability and performance, software infrastructure
  Applications: data-intensive, science drivers and needs
-Key players
  Applications data experts
  Analysis and visualization experts
  Systems software experts
-Objectives, what you want to get out of the workshop
  In a word, collaborations

Week at a Glance

-Monday
- Application introductions
- Data infrastructures, systems, libraries

-Tuesday
- Data models
- Data analysis and visualization infrastructures

-Wednesday
- Data analysis and visualization techniques
- Open time

-Thursday
- Panels
- Adjourn

Activities

- Hiking (trail maps available)
- Mountain biking
- Tram to the summit
- Peruvian chair and tunnel
- ZipRider
- Alpine Slide

- Restaurants

-Aerie, Atrium, El Chanate, Forklift, Steak Pit

-See the visitor guide for others

- Shops

-outdoor gear

-other stuff

Meals

- All are buffet style
- Breakfast
  - 7:00 - 8:30
  - Magpie A
- Lunch
  - 12:00 - 1:30
  - Magpie A except Monday White Pine ("C" level)
- Dinner
  - 6:00 - 8:30
  - Creekside Terrace (outside), Atrium in case of rain

Reminders

- Travel reimbursement forms
- Send me pdfs of slides so that I can post them

Panel 1 Suggested Topics: Emerging Architectures

- Many, multicores
- GPUs
- FPGAs
- Programming models
- Heterogeneity
- Locality
- Domain decomposition
- Portability
- Expanded memory / storage hierarchy
- Overall system design

Panel 2 Suggested Topics: Computer – Computational Science Interaction

One more volunteer needed

- Language barriers
- Amount and type of interaction
- Research vs. production
- Resources (funding, machine time)
- Publication of results
- Who's driving
- Integration of analysis and computation
    - In situ, coprocessing, intrusive, nonintrusive

Panel 3 Suggested Topics: Data Size and Complexity

- Define big data
- Impact on computation, analysis, storage
- Data models for big complex data
- Analysis techniques to alleviate size and complexity
    - Statistical summarization
    - Model fitting
    - Compression
    - Multiresolution
- Data-driven (unsupervised) vs. science-driven (supervised)

Supplemental Material: Discussion Questions from Previous Year

Data Models Discussion Questions

How will future exascale architectures affect data models? Eg., can data models be designed to:

- minimize data movement
- minimize memory size
- trade computation for space or communication?

What is the best balance of generality / specificity?

What is lacking in current data models or their discussion so far?

Data Analysis Infrastructures Discussion Questions

Will VTK continue to be a predominant analysis tool?

Can it scale to exascale, or is the code outdated?

Are all execution models (in situ, coprocessing, postprocessing equally useful to applications?

What level of involvement in analysis is most appropriate for applications?

What changes to current infrastructures will be needed for exascale analysis?

How can analysis infrastructures support some of the data models that were discussed and other new data models?

Computation in Science Codes Discussion Questions

What are your current highest priority problems in scaling your applications?

Describe your data input and output characteristics.

How do you see data storage and analysis being performed at exascale?

- Full storage of all data, or selective storage
- Determination of what to store
- Metadata storage

What degree of integration with storage and analysis tasks / experts do you see as necessary or useful for continued progress?

Analysis in Science Codes Discussion Questions

Describe where the line is between analysis directly as part of the simulation, in situ analysis performed on behalf of the simulation, and analysis separate from the simulation.

How can applications' native data models be exposed to analysis infrastructures and techniques?

Analysis and Visualization Techniques Discussion Questions

Describe the difference between visualization, analysis, visual analytics.

What techniques do you see as being most useful at exascale, besides the technique that you presented.

How will projected exascale architectures influence techniques?

Are some techniques more appropriate for certain execution models (in situ, postprocessing, coprocessing?)

What techniques do applications need most?

Can some techniques be used to reduce data storage more effectively than others?

Future Storage Systems & Software Discussion Questions

How will future storage architectures differ from today?

- Semantic information, object-oriented storage
- New layers in the memory / storage hierarchy

Active storage operations?

Write-behind / prefetching?

Will storage systems be built from commodity / enterprise / custom hardware?

What should applications anticipate when interacting with future systems?

- How much of the anticipated changes to storage will be visible to applications

Is exascale storage even feasible, or what needs to happen in order to succeed?

- Funding, awareness by applications and analysis, gaps in research

Closeout Discussion Questions

1. What level of involvement and integration with analysis and storage is appropriate or necessary for applications? What should applications anticipate when interacting with future systems, and how much of the anticipated changes to storage will be visible to applications?
2. What are the common components of data-intensive simulations, analysis, and storage? (data models, interfaces, object-oriented storage?)
3. How will future architectures affect these components?
4. Is exascale storage and analysis even feasible, or what needs to happen in order to succeed? Will storage systems be built from commodity / enterprise / custom hardware? Funding, awareness by applications and analysis, needed research?
5. Suggestions for improving this or other workshops?