Summary of the SDM Breakout Session

Feb 23, 2011

Scope

- Workflow systems,
- Metadata generation, capture and evaluation
  - Standards, provenance, semantics
- Data models, representation, formats
- Data movement/transfers
- Data fusion, data integration
- Support for Analysis, Informatics, Information visualization
- Data reuse, archiving, persistence, (integrity: annotation)
- V&V, UQ, data quality (tools and statistical measures of quality)
- Acceleration technology (indexing, reordering, transformations)
- Publication, data discovery (as oppose to analysis)

- Focus on OASCR needs, as opposed to industry or science domains

Motivation

- Enable new and better science and improve scientific productivity
- Computer architectures will be changing, ASCR needs to be ready for the change to minimize disruption to science
  - Data management will be the key issue to enable useful application of the machine (power consumption is directly related to data management)
  - In-memory workflow changes data management requirements from current paradigm
  - Reducing memory transfers will require accommodation of tools such as staging, workflow, etc.

- With complex systems, we need to understand what has happened and be able to address issues around reproducibility of analysis on the exascale?
  - Bitwise reproducibility is likely not possible.
  - What is an acceptable system to compare results? Statistical comparisons.

Current state of the art

- Workflow technology: kepler, SWIFT, Taverna
- Schedulers, system software
- Data staging: ADIOS
- Data movement tools / IO

- Coherent, persistent data sets are assumed by the current set of tools

Issues with current technology (Research Opportunities)

- Coherence will become a problem in exascale
- Must support in-memory data sharing among different tools, data objects management in memory
- Indexing system for facilities and resources must be available on nodes for resource discovery and allocations
- Data transformations and analytics on-node / in-memory are required to prepare for output, data reduction
- Current data models do not provide the required abstractions
  - IO based interface may not be sufficient
  - Ease of use is important
  - Need new data models for supporting in-memory, out of memory, and streaming data access
- Will require new scheduling and provisioning techniques to support co-scheduling of I/O and compute and storage
  - The scheduler will need to manage workflows not just jobs

Cross-cutting and co-design opportunities

- System / runtime
- Programming models
  - Deep memory hierarchy requires new design of many tools such as data management and workflow
  - New definition of data consistency
  - Direct control of memory hierarchy for some users
  - Runtime system environment awareness of data/ objects used by user programs
  - Leveraging compiler and library research work
- Need better levels of abstraction

Metrics

- Time to quality of solution:
- Scientific productivity
  - relevant publications
- Reduction in power required to solve the same problem

- What can be enabled by the new tools:
- Adoption: Number of users/codes successfully using the machines/tools/technologies

Take away messages

- This is not your father's data management, data will become an integral part of computational science
  - Data management is not just about files anymore
  - Many cross-cutting issues to be addressed
- We will need to adjust to a more complex world
  - Need new data models and approaches to effectively manage data and workflows in this environment

Cast of Characters

- Terence Critchlow
- Dean Williams
- John Wu
- Manish Parashar
- Keith Jackson
- Jay Lofstead
- Norbert Podhorszki
- Doron Rotem
- Suren Byna
- Jinghau Ge
- Kerstin Kleese Van Dam
- Matthew Wolf
- Karsten Schwan
- Quincey Koziol
- Rob Latahn
- Terry Ligocki
- James Ahrens