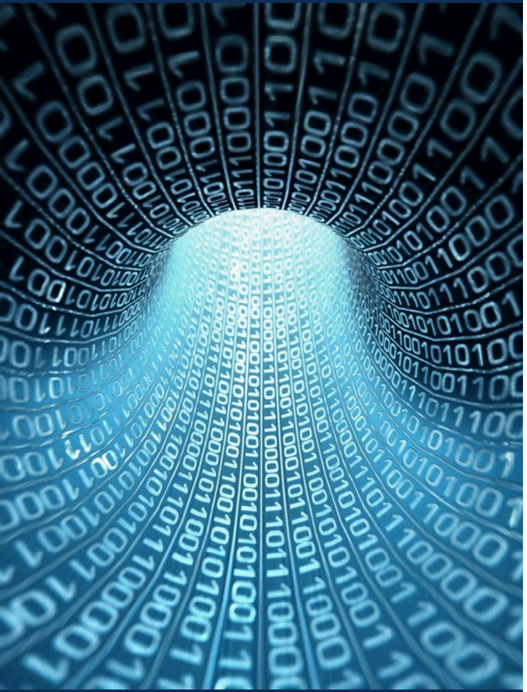
SAND2014-17570PE



Sandia National Laboratories

Exceptional service in the national interest

Data Analysis Priorities for a Successful Exascale Program

Janine C. Bennett

Sandia National Laboratories

DOE Data Council Meeting
Rockville, MD
September 16-17, 2014

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

Priorities

1. **Workflow management**
   - Increasingly complex workflows demand efficient, data-driven decision-making capabilities in-situ
   - Need to support conditional actions (e.g., I/O & analysis frequency)
     - Do only when “interesting science” is occurring
   - Both algorithmic research (to quickly identify “interesting”) and computer science infrastructure (to express workflow) are required

2. **Resilient workflows**
   - Different data challenges posed by different classes of errors
     - Soft errors can corrupt data by propagating through complex workflows
     - Hard errors (fail-stop node crashes) cause data loss
   - What errors should be handled transparently by the run-time?
   - What errors should be propagated to the user?

Priorities (continued)

3. **Programming models**
   - Changes in architectures are causing a shift in programming models
     - Many emerging asynchronous, many-task data flow models
     - Achieve both task & data parallelism
   - Need: unified programming model across all elements in workflow

4. **Data transformations and reduction techniques**
   - All raw data cannot be stored to disk
   - All analysis is not do-able in-situ (e.g. unanticipated events require iterative exploration)
   - We need techniques that reduce data while supporting/facilitating offline post-processing

Priorities (continued)

5. **Provenance and reproducibility of results**
   - How much and what data and meta-data must be stored?
   - What is the best way to capture provenance information in complex, dynamic workflows?