SAND2011-2105P

Cloud Computing and Scientific Datasets

Craig Ulmer

Sandia National Laboratories, CA

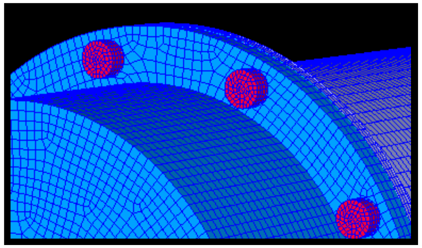
cdulmer@sandia.gov

March 29, 2011

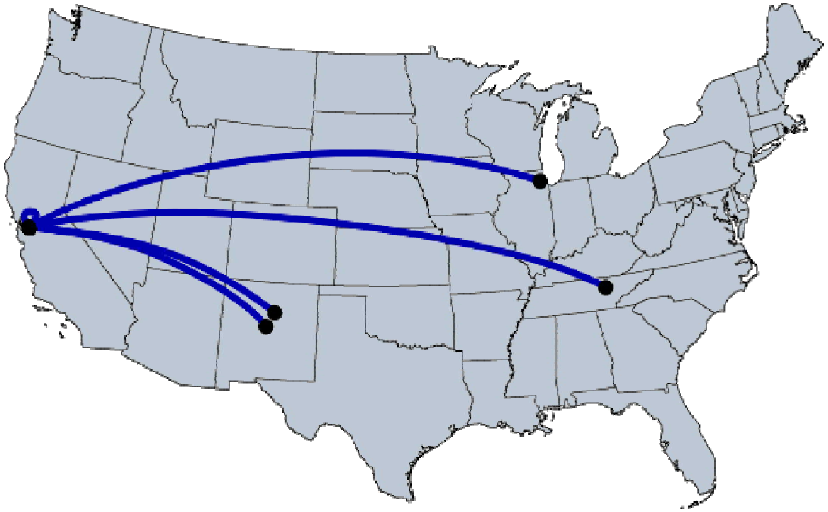
Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

High-Performance Computing in DOE

- HPC is essential to many aspects of DOE work
  - Scientific Computing: Compute-bound simulations
  - National Security: Graph algorithms, Data mining



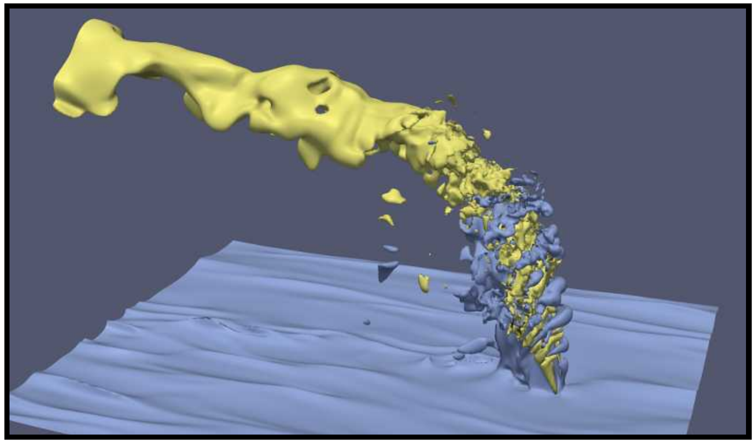
- Sandia/California Challenges
  - Local systems: Limited power (1MW), space, funding, staffing
  - Distance computing: Use external systems, 10Gb/s links



Sandia National Laboratories

Can We Leverage Cloud Computing?

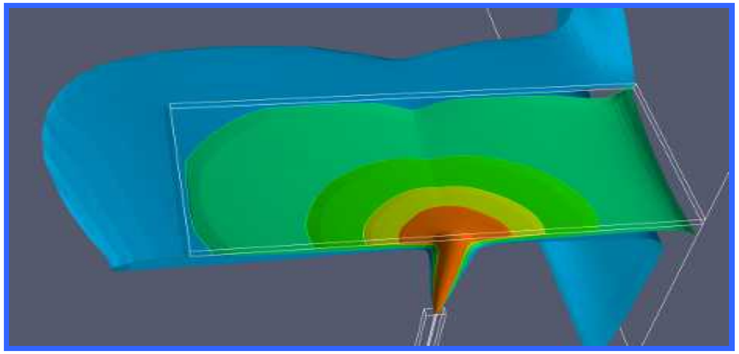
- ASC SICAIDA
  - Storage-Intensive Computing Architectures for In-situ Data Analysis
- Is cloud computing relevant to scientific computing?
  - Capability Computing: **No**
  - Capacity Computing: ***Possibly***
  - Post Processing: ***Yes***
- Motivating use case: S3D
  - Runs on ORNL Jaguar
  - 10-100TB Datasets
  - Provide collaborator access
  - Hadoop cluster

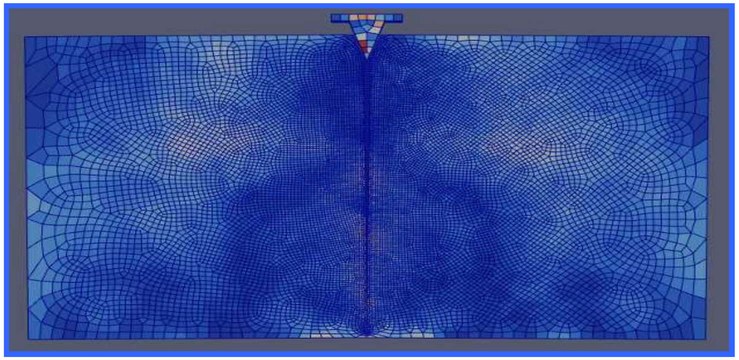


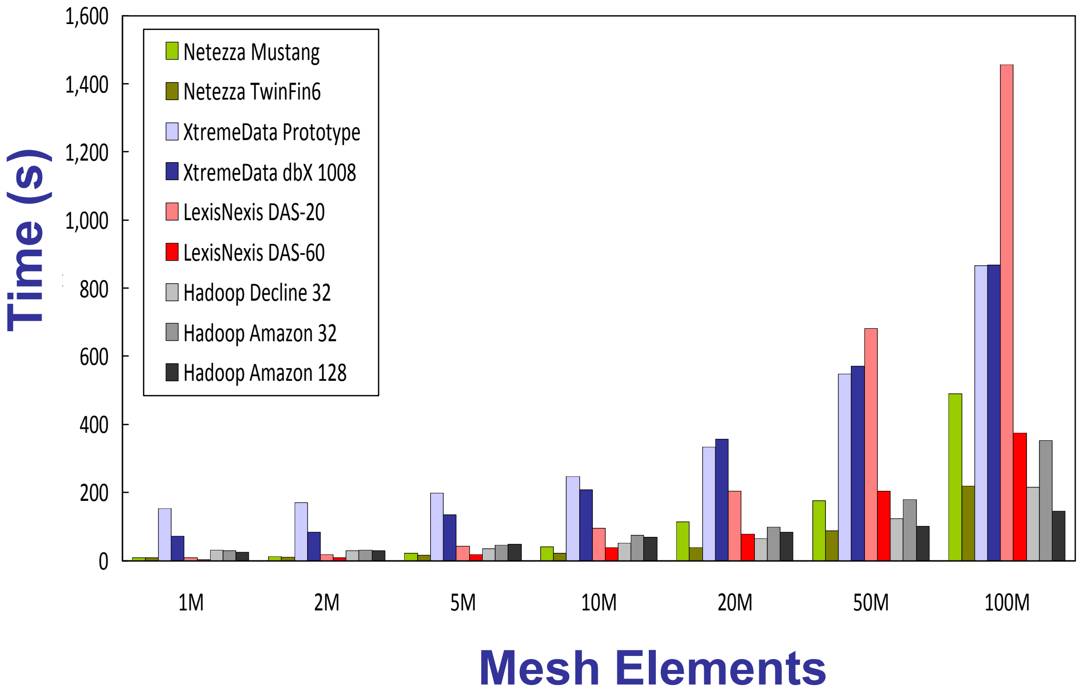
Sandia National Laboratories

Evaluating Different Platforms

- Ported **mesh analysis** algorithms to multiple platforms
  - Traditional SQL Parallel Database: Netezza, XtremeData
  - “NoSQL” Platforms: LexisNexis DAS, Hadoop (Local + Amazon)

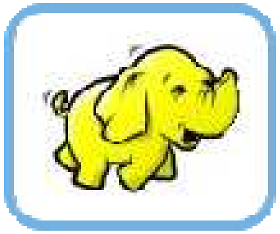






Current Status

- MapReduce is good and bad for scientific data analysis
- Evaluated many cloud technologies
  - Frameworks: Hadoop (MapReduce, Streaming, Pig), Sector/Sphere
  - Stores: Cassandra, GlusterFS, Direct HDFS, MongoDB
- Ongoing interests
  - Improving Hadoop w/ better resources (10GigE/IB, SSDs)
  - Scheduling non-Hadoop jobs on cluster via Hadoop Streaming
  - Working directly w/ parallel data stores
  - Data ingestion from other HPC resources
  - Security issues of multi-tenant systems



Sandia National Laboratories