
5 Steps to Stress-Free, Large-Scale Data Management

Terence Critchlow

*Center for Applied Scientific Computing
Lawrence Livermore National Laboratory*

SIAM Computational Science and Engineering

February, 2007

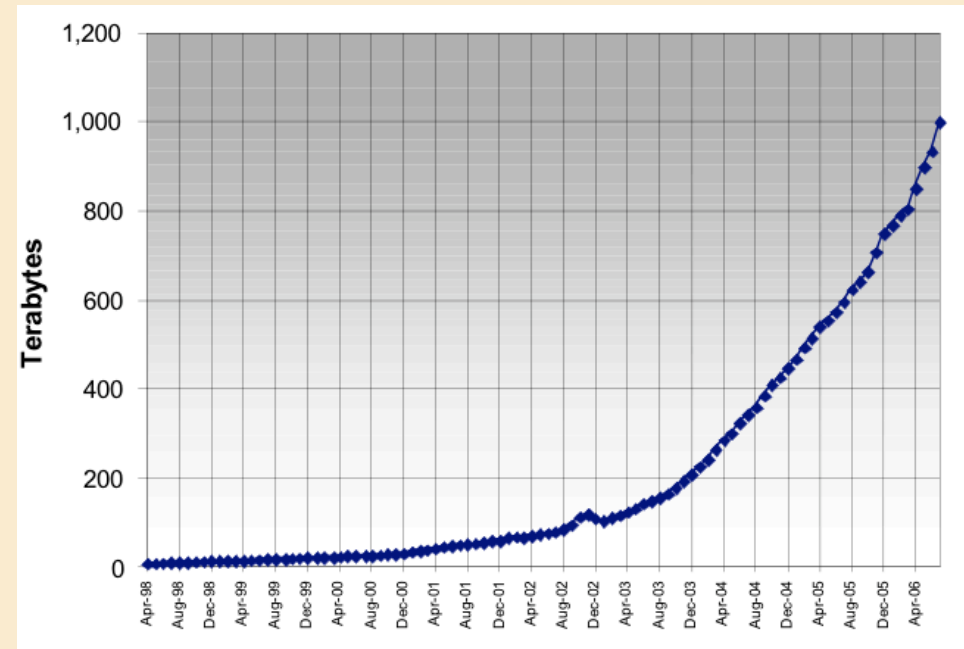
UCRL-PRES-228111



1) Stop worrying about Flops

build a balanced SC architecture

- Current machines are designed to provide high Flop performance on benchmarks
- Overall performance for real apps is the important metric
- Need to consider storage and data transfer costs and design systems that provide a holistic solution



ORNL HPSS storage size

Courtesy of Scott Klasky (ORNL)

BG/L

- 65,536 nodes, 131,072 CPUs, 512MB ram, 367TFlops/sec
- 1,204 I/O nodes, 806Tb disk

2) Get the data where you need it

address data movement and integration issues

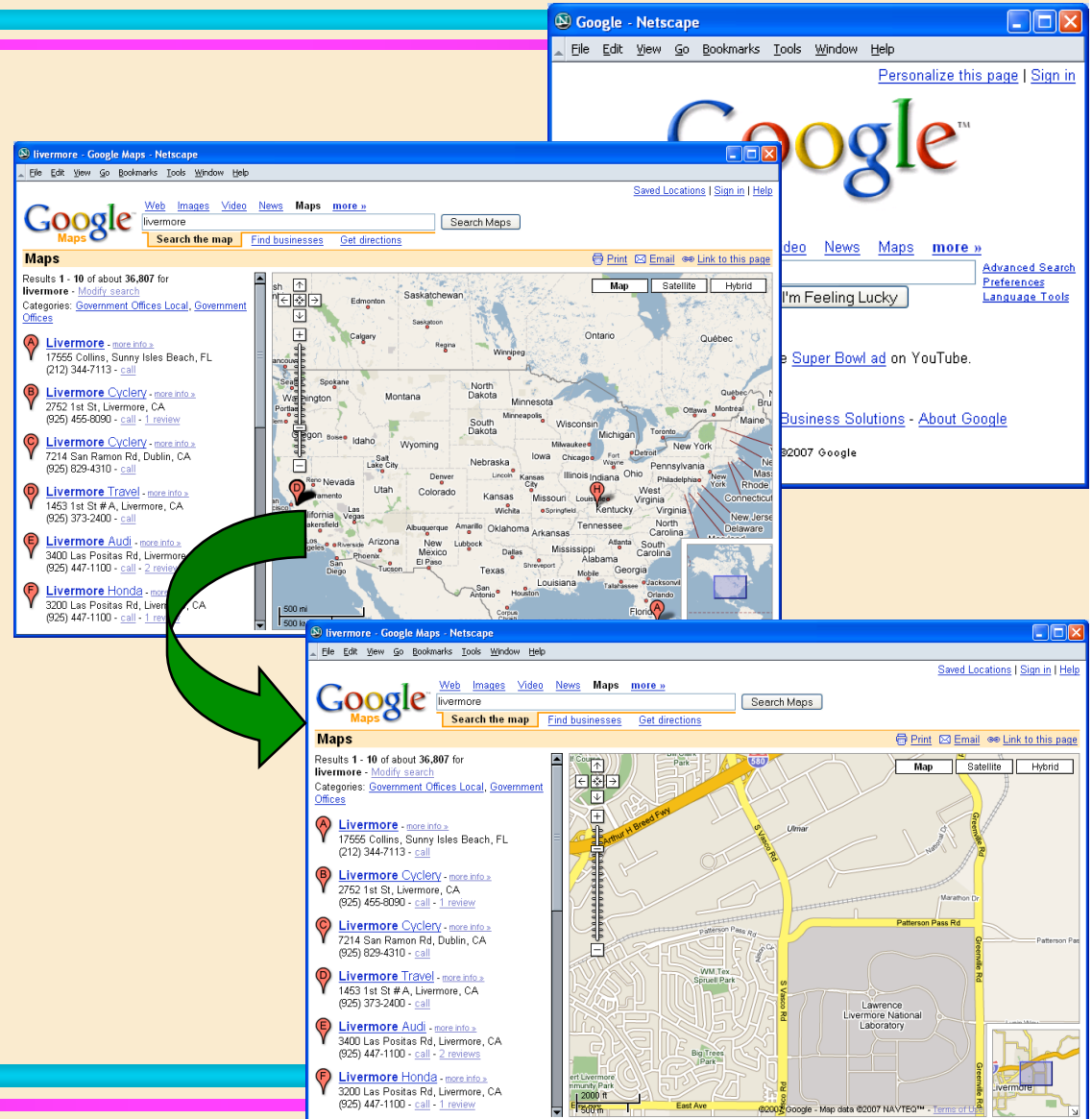
- It takes a significant amount of time to move data off of a SC
- Remote collaborations are hamstrung by data transfer capabilities
- Data integration still poses significant challenges
- Grid-like technology has promise but so far has not provided enhanced user capabilities



Need to pull together multi-modal information from distributed, heterogeneous sources

3) Figure out what data you care about *create indices and query multi-modal data*

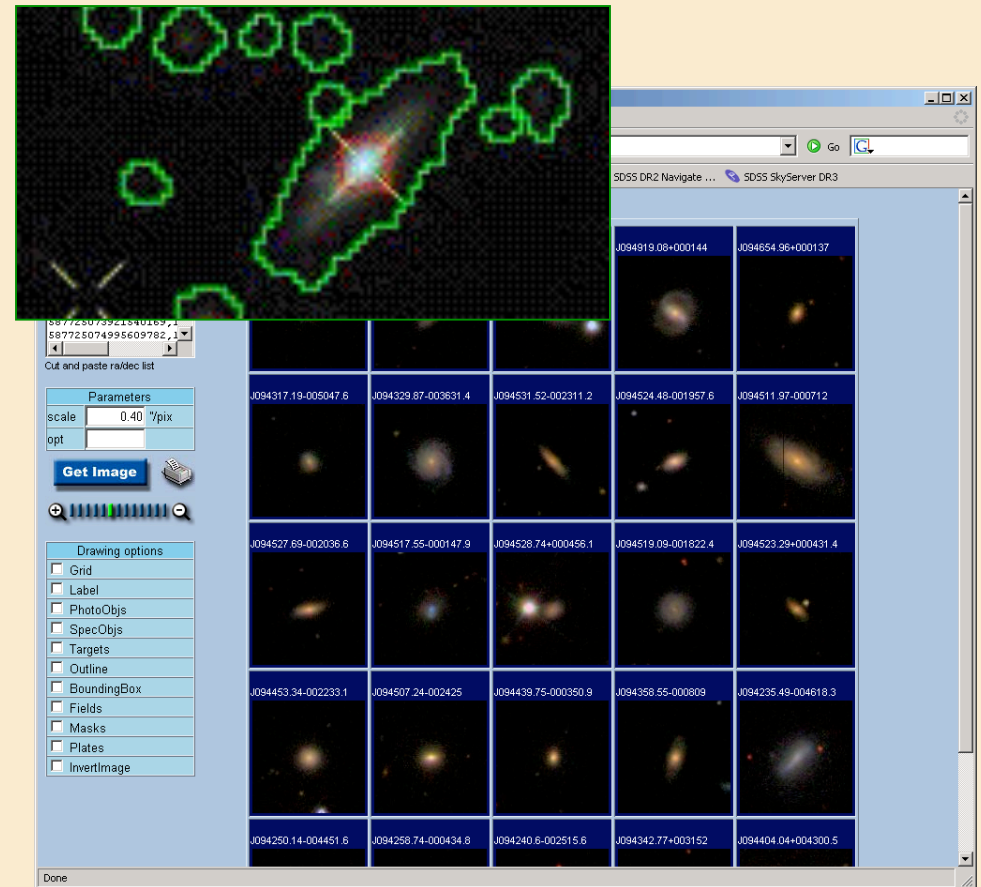
- Simulation and experimental data is being acquired faster than we can process it
- Need to perform quick filtering to eliminate data that we know we don't want to look at



4) Gain insight from your data

perform multi-modal data analysis

- Identifying most interesting information requires complex analysis
- Need automated and semi-automated tools to identify interesting information quickly
- Data comes in a wide variety of formats (simulations, text, images, sensors, GIS) and analysis techniques need to be useable on all of them



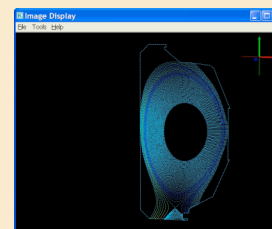
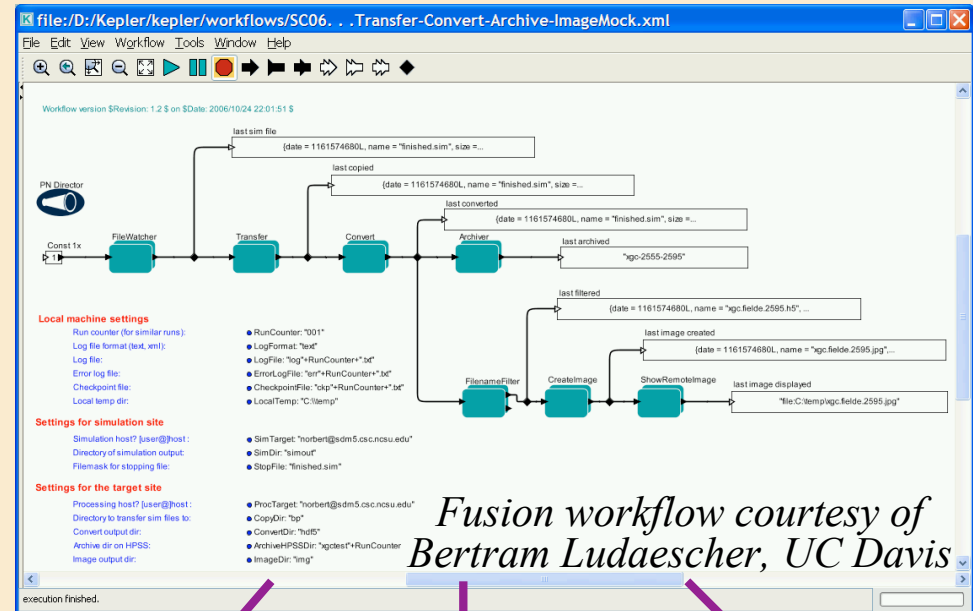
Sloan Sky Survey Toolkit
Screen Shot from Jordan Raddick

5) Automate record keeping

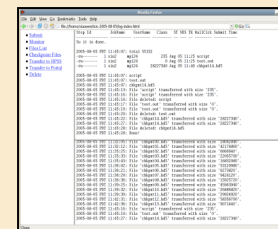
use workflows, provenance, and metadata

- Simulations and experiments generate tremendous amounts of data
- Need a way to automatically keep track of this information in a way that makes it retrievable

- Where did a data set come from
- What are the values of key parameters within it
- How does this compare to other data sets



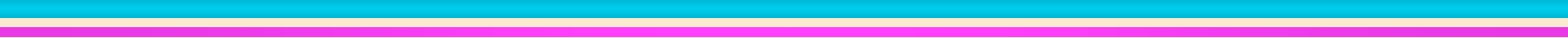
Simulation output



Log files



Provenance information



This work was performed under the auspices of the U.S.
Department of Energy by University of California Lawrence
Livermore National Laboratory under contract No. W-7405-
ENG-48.