Meeting Big Data Challenges: infrastructure, tools, parallel workflows and algorithms

Dr. Richard Loft
Director, Technology Development Division
Computational and Information Systems Laboratory, NCAR
loft@ucar.edu

April 22, 2014
Outline

• Infrastructure
• Tools
• Parallel Workflows
• Algorithms
The NCAR-Wyoming Supercomputing Center...
High Performance Cyberinfrastructure...

1.5 Petaflops supercomputer (Yellowstone)

11 petabyte disk storage system

100 petabyte tape system
And 100 kilometers of interconnect wiring!
Yellowstone’s Data-Centric Architecture

Other Centers
- Data Transfer Gateway

People
- Science Gateways

Supercomputers
- scratch
- Project Space
- Data Collections

Long term storage
- HPSS

Data Analysis Visualization
Data Services

- Research Data Archive: http://rda.ucar.edu
- Earth System Grid https://www.earthsystemgrid.org
- Globus+ data cloud: dropbox on steroids
VAPOR Overview

- A 3D visualization environment
- Tailored towards our community
- A desktop solution designed to handle Big Data
- Support for multi-variate, time-varying data

http://www.vapor.ucar.edu/
VAPOR Wavelet Data Compression: Applied to new seadragon species

Data courtesy of Josephine Stiller and Dr. Greg Rouse, Scripps Institution of Oceanography
A sampling of scientific applications of VAPOR

- Hurricanes
- Solar Physics
- Wild Fires
- Astrophysics
- Thunder Storms
- Wind Turbines
NCAR Command Language (NCL)

“A scripting language for the analysis and visualization of geoscientific data”

http://www.ncl.ucar.edu/
NCL usage metrics

- 17,627 registered users in >160 countries
- 1,055 on email list
- 300 email postings a month
- 2,004 downloads a month
Lossy Compression – 5x with no loss of scientific information

• Tested several compression algorithms
• Chose highest compression ratio that passed statistical tests
• Result: a file 18% of its original size
Climate data “Pepsi test”: after 4:1 compression

After lossy compression...  Original...

ANN

recon fq2 (yrs 80-89)

Planet boundary layer height

mean = 589.20

meters

Min = 166.04 Max = 1299.78

LRC01 (yrs 80-89)

Planet boundary layer height

mean = 589.16

meters

Min = 165.83 Max = 1300.31
Reordering the data took 80% of the time in the last climate change assessment.

<table>
<thead>
<tr>
<th>Field 1</th>
<th>Header</th>
<th>Time 1</th>
<th>Time 2</th>
<th>...</th>
<th>Time m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field 2</td>
<td>Header</td>
<td>Time 1</td>
<td>Time 2</td>
<td>...</td>
<td>Time m</td>
</tr>
<tr>
<td>Field n</td>
<td>Header</td>
<td>Time 1</td>
<td>Time 2</td>
<td>...</td>
<td>Time m</td>
</tr>
</tbody>
</table>
Accelerating the manipulation of climate data

• 35x for 0.25° atmospheric data
  – 5 hours -> 8.5 minutes

• 7.9x for 0.10° ocean data
  – 14 hours -> 1.7 hours

Parallel computing:
Many computers working together
What if scientific data analysis were as fast as “googling”?
How can we estimate summer rainfall at places where there are no observations?
The keys to transforming atmospheric data analysis

- **Algorithmic acceleration**
  - Kriging -> LatticeKrig
  - 5000 obs = 30x speedup
  - 10000 obs = 75x speedup

- **Parallel speedup**
  - Near perfect scaling
  - 1000x on 1.5% of Yellowstone
Such analyses reveal underlying uncertainty in the data.
Thanks! Question?

Dr. Richard Loft
NCAR
loft@ucar.edu
April 22, 2014