CIS 602-01: Computational Reproducibility

Cloud Computing and Scalability

Dr. David Koop
Numerical Reproducibility Results

<table>
<thead>
<tr>
<th>Description of the simulation (no. of processors)</th>
<th>Minimal value of the sheet thickness change (%)</th>
<th>Maximal value of the sheet thickness change (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>−29.21</td>
<td>+9.54</td>
</tr>
<tr>
<td>2</td>
<td>−29.34</td>
<td>+9.14</td>
</tr>
<tr>
<td>3</td>
<td>−29.04</td>
<td>+9.02</td>
</tr>
<tr>
<td>4 (1st attempt)</td>
<td>−29.04</td>
<td>+8.98</td>
</tr>
<tr>
<td>4 (2nd attempt)</td>
<td>−29.09</td>
<td>+8.96</td>
</tr>
</tbody>
</table>

Max = 9.54

Max = 8.96

[Diethelm, 2012]
Scalar Product Example

• Computation of scalar product

\[ s = \sum_{i=1}^{n} x_i y_i \]

• Now assume we have a 10^6-dimension vector

• Parallelize! Divide among p processors the n summands; l-th processor computes the following partial sum:

\[ s_l = \sum_{i=(l-1)n/p+1}^{ln/p} x_i y_i \]

• Then, compute the final sum:

\[ s = \sum_{i=1}^{p} s_l. \]
Fixed number of processors

• Processors complete in different orders...
• Compute the sum as the solutions come in
• Example: \( n = 8, \ p = 4 \)

\[
\begin{align*}
x_1 &= 10^{12}, \ x_2 = 0, \ x_3 = 10^{-8}, \ x_4 = 0, \ x_5 = -10^{12}, \ x_6 = 0, \ x_7 = 10^{-8}, \ x_8 = 0, \text{ and } y_j &= 1 (j = 1, 2, \ldots, 8) \\
s_1 &= 10^{12}, \ s_2 = s_4 = 10^{-8}, \ s_3 = -10^{12}
\end{align*}
\]

• If we have order \( s_1, s_3, s_2, s_4 \)

\[
\begin{align*}
z_1 &= s_1 + s_3 = 10^{12} - 10^{12} = 0 \\
z_2 &= z_1 + s_2 = 0 + 10^{-8} = 10^{-8} \\
s &= z_2 + s_4 = 10^{-8} + 10^{-8} = 2 \cdot 10^{-8}
\end{align*}
\]

• Order \( s_1, s_2, s_3, s_4 \) gives

\[
\begin{align*}
s' &= ((10^{12} + 10^{-8}) - 10^{12}) + 10^{-8} \\
&= (10^{12} - 10^{12}) + 10^{-8} = 10^{-8}
\end{align*}
\]
Varied number of processors

• What is we have only two processors?

\[ s_1 = 10^{12} + 10^{-8} \]
\[ s_2 = -10^{12} + 10^{-8} \]
\[ s^* = s^*_1 + s^*_2 = 10^{12} - 10^{12} = 0 \]

• A third solution!

• Solutions to these issues?
  - Wait for all sums to complete so the final sum can be done deterministically
  - Use virtual processors that is fixed no matter what the environment
Assignment 2

- Keep your project on Github, keep images on Docker Hub
- Put a link to your Docker Hub images in your Github README.md
- Questions?
  - Running inside of Docker versus on your local machine
  - MacOS X stdout: start VisTrails with VisTrails.command
Projects

• Report
  - Reproducibility:
    • Summary of paper's main contributions
    • Reproducibility overview (what is available?, does it work?)
    • Experiments: what experiments are being reproduced and what is the ground truth
    • Results: what was reproducible, what wasn't, why
  - Research:
    • If research on reproducibility, like a standard research paper
    • If adding reproducibility to your own research, introduce ideas, but then focus on what reproducibility requirements exist, what was implemented, and how reproducibility was evaluated
Projects

• Presentation (8-10 minutes):
  - Reproducibility:
    • **Short** summary of the paper
    • Summary about what is available, what works, etc.
    • Best: show reproducibility in action
    • Good: show side-by-side results
  - Research (focus on **reproducibility** aspects!):
    • **Short** motivation and introduction
    • Reproducibility requirements
    • Discuss Results
    • Best: show tool/reproducibility in action
Cloud Computing

- Cloud = Virtualization + Resources + Services
- Any code, any data
- Scalable storage and compute for everyone
- Services for processing big data, various data models
- Services for managing VMs/images/containers
- Secure, reliable, available
- "Download it to my laptop" is insufficient

[B. Howe, 2009]
Amazon EC2 Scaling

3000 CPU's for one firm’s risk management application

300 CPU's on weekends

[Deepak Singh, Amazon.com]
The Value of Data

- Who pays to maintain data?
- How do we measure important data? Use?
- Should the use of data incur charges?

[B. Howe, 2009]
Big Data

• The days of FTP are over
  - It takes days to transfer 1TB over the Internet, and it isn’t likely to succeed.
  - Copying a petabyte is operationally impossible

• The only solution: Push the computation to the data, rather than push the data to the computation
  - Upload your code rather than download the data

[B. Howe, 2009]
In silico research in the era of cloud computing

J. T. Dudley and A. J. Butte
Reproducibility in the Cloud

- Whole System Snapshot Exchange (WSSE)
- May be huge (GBs or TBs), but do this in the cloud not between desktops
- Layers:
  - Data
  - System
  - Service
Data Layer

In cloud computing, computation and data are 'virtualized', meaning that software and data are not tied to physical computing resources, such as a specific server with hard drives plugged into it. Instead, cloud-computing infrastructures are comprised of large and often geographically disparate clusters of computing hardware that are made to appear as a single, homogeneous computational environment. The virtualization of data in the cloud makes it possible to move or copy 'snapshots' of large data sets from point to point within the cloud at high transfer rates, without the need to associate particular machines or storage drives at either the source or destination of the data transfer.

Concerns have been voiced that scientific computing in the cloud could make results less reproducible. One concern is that cloud computing will be a computing 'black box' that obfuscates details needed to accurately interpret the results of computational analyses. Although this is an important concern, we argue that cloud computing could actually enhance reproducibility.

Figure 1

Layers of reproducible computing in the cloud. (a) Data layer. Generators of large scientific data sets can publish their data to the cloud (1) and substantial updates to these data sets can exist in parallel without loss or modification of the previous data set (2). Primary investigators can clone entire data sets within the cloud (3) and apply custom scripts or software computations (4) to derive published results (5). An independent investigator can obtain digital replicates of the original primary data set, software and published results within the cloud to replicate a published analysis and compare it with published results (6). (b) System layer. Investigators can set up and conduct scientific computations using cloud-based virtual machine images that contain all of the software, configurations and scripts necessary to execute the analysis. The customized machine image can be copied and shared with other investigators within the cloud for replicate analyses. (c) Service layer. Instead of replacing an old version of a scientific computing service with a new version of the software, a service that is virtualized in the cloud can be easily replicated and the old version made available alongside the new. Requests made by applications through the external service interface could incorporate a version parameter. This could enable published results that used older versions of the service to be evaluated for reproducibility.
Data Layer

- Publish data and update it with versioning
- Clone entire datasets
- Apply custom scripts to derive results in the cloud
- Others can obtain replicates of the original dataset
- Enable greater data availability and preservation
will be exchanged exclusively using cloud computing, defined here as "computing in which dynamically scalable and virtualized resources are provided as a service over the Internet".

In cloud computing, computation and data are 'virtualized', meaning that software and data are not tied to physical computing resources, such as a specific server with hard drives plugged into it. Instead, cloud-computing infrastructures are comprised of large and often geographically disparate clusters of computing hardware that are made to appear as a single, homogeneous computational environment.

The virtualization of data in the cloud makes it possible to move or copy 'snapshots' of large data sets from point to point within the cloud at high transfer rates, without the need to associate particular machines or storage drives at either the source or destination of the data transfer. Some cloud providers offer technologies that enable the creation of large databases that can persist as omnipresent resources that can be accessed by any computation in the cloud.

Concerns have been voiced that scientific computing in the cloud could make results less reproducible. One concern is that cloud computing will be a computing 'black box' that obfuscates details needed to accurately interpret the results of computational analyses. Although this is an important concern, we argue that cloud computing could actually facilitate reproducibility.

**Figure 1**

Layers of reproducible computing in the cloud. (a) Data layer. Generators of large scientific data sets can publish their data to the cloud (1) and substantial updates to these data sets can exist in parallel without loss or modification of the previous data set (2). Primary investigators can clone entire data sets within the cloud (3) and apply custom scripts or software computations (4) to derive published results (5). An independent investigator can obtain digital replicates of the original primary data set, software and published results within the cloud to replicate a published analysis and compare it with published results (6).

(b) System layer. Investigators can set up and conduct scientific computations using cloud-based virtual machine images that contain all of the software, configurations and scripts necessary to execute the analysis. The customized machine image can be copied and shared with other investigators within the cloud for replicate analyses.

(c) Service layer. Instead of replacing an old version of a scientific computing service with a new version of the software, a service that is virtualized in the cloud can be easily replicated and the old version made available alongside the new. Requests made by applications through the external service interface could incorporate a version parameter. This could enable published results that used older versions of the service to be evaluated for reproducibility.
System Layer

• Set up computations in VM images that contain all software, configurations, and scripts
• Images can be copied and shared easily
will be exchanged exclusively using cloud computing, defined here as “computing in which dynamically scalable and virtualized resources are provided as a service over the Internet.”

In cloud computing, computation and data are ‘virtualized,’ meaning that software and data are not tied to physical computing resources, such as a specific server with hard drives plugged into it. Instead, cloud-computing infrastructures are comprised of large and often geographically dispersed clusters of computing hardware that are made to appear as a single, homogeneous computational environment. The virtualization of data in the cloud makes it possible to move or copy ‘snapshots’ of large data sets from point to point within the cloud at high transfer rates, without the need to associate particular machines or storage drives at either the source or destination of the data transfer.

Some cloud providers offer technologies that enable the creation of large databases that can persist as omnipresent resources that can be accessed by any computation in the cloud.

Concerns have been voiced that scientific computing in the cloud could make results less reproducible. One concern is that cloud computing will be a computing ‘black box’ that obfuscates details needed to accurately interpret the results of computational analyses. Although this is an important concern, we argue that cloud computing could actually be a tool for improving reproducibility.

**Figure 1**

Layers of reproducible computing in the cloud. (a) Data layer. Generators of large scientific data sets can publish their data to the cloud (1) and substantial updates to these data sets can exist in parallel without loss or modification of the previous data set (2). Primary investigators can clone entire data sets within the cloud (3) and apply custom scripts or software computations (4) to derive published results (5). An independent investigator can obtain digital replicates of the original primary data set, software and published results within the cloud to replicate a published analysis and compare it with published results (6). (b) System layer. Investigators can set up and conduct scientific computations using cloud-based virtual machine images that contain all of the software, configurations and scripts necessary to execute the analysis. The customized machine image can be copied and shared with other investigators within the cloud for replicate analyses. (c) Service layer. Instead of replacing an old version of a scientific computing service with a new version of the software, a service that is virtualized in the cloud can be easily replicated and the old version made available alongside the new. Requests made by applications through the external service interface could incorporate a version parameter. This could enable published results that used older versions of the service to be evaluated for reproducibility.
Service Layer

- Service is virtualized in the cloud
- Keep all versions of services
- Incorporate version parameter to replicate past results
- Create a cloud-based scientific computing commons
Concerns

• Black box = less reproducible…or more transparent?
Data Sharing Features

• Traditional Challenges:
  - Large data sets difficult to share over standard internet connections; can require substantial technical resources to obtain and store.
  - Public data sets change frequently. Difficult to archive and share entire data repositories used for analyses.

• Cloud Solutions:
  - Large data sets can be stored as ‘omnipresent’ resources in the cloud. Easily copied and accessed directly from any point in the cloud.
  - ‘Snapshots’ of large public data sets can be rapidly copied, archived and referenced.
Software and Applications

• Traditional Challenges:
  - Reproducibility of results often requires replication of the precise software environment (that is, operating system, software and configuration settings) under which the original analysis was conducted. Specific versions often required for reproducibility.
  - Analyses typically conducted by several types of software or scripts executed in a precise sequence across one or several systems as part of an analysis pipeline. Only the individual programs or scripts are usually provided with published results. Substantial technical resources typically required.
  - Standard software packages cannot serve all the needs of a scientific domain. Investigators develop nonstandard software and computational pipelines to facilitate computational analysis exceeding the capabilities of common tools.
Software and Applications

- Cloud-computing solutions:
  - Computer systems are virtualized in the cloud, allowing them to be replicated wholesale without concern for the underlying hardware. Snapshots of a fully configured system or group of systems used in analysis can be rapidly archived as digital machine images. System machine images can be copied and shared with others in the cloud, allowing reconstitution of the precise system configuration used for the original analysis.
  - System images can be preconfigured with common and customized software and tools in a standardized fashion to facilitate common tasks in a scientific domain (e.g., assembly of genome sequences from DNA sequencer data). Preconfigured images can be shared as public resources to promote reproducibility and follow-up studies.
System and Technical Features

• Traditional Challenges:
  - Substantial computational resources might be required to replicate an analysis. Original computational analyses requiring several hundred processors to complete becoming more common. Reproducibility limited to those with requisite computational resources.
  - Substantial technical support often required to reproduce a computational analysis and to replicate the software and system configuration required by the analysis. Prevents reproducibility by nontechnical investigators lacking substantial IT support.
System and Technical

- Cloud-computing solutions:
  - Cloud-based computational resources can be scaled up in a dynamic fashion to provide necessary computational resources. Investigators can create large computational clusters on demand and disperse upon analysis completion.
  - Complete digital representations of a computational pipeline can be shared as machine images along with deployment scripts that can be executed by nontechnical users to reconstitute a complete computational pipeline.
Access and Preservation

• Traditional Challenges:
  - Grant-funded software and data repositories often disappear from the public domain after funding is discontinued or the maintainers abandon the project. Leads to loss of access by dependent users and loss of public investment into the resource.

• Cloud-computing solutions:
  - Software, code and data from grant-funded projects can be archived and provided as publicly accessible resources in the cloud. Economies of scale in the cloud allow for active preservation of grant-funded resources for many years past funding for nominal cost.
  - Cloud-computing providers already show a willingness to host public scientific data sets at no cost.
Discussion
Review

• Code Availability, Collaboration, and Version Control
• Data Availability, Citation, and Curation
• Virtual Machines and Containers
• Scientific Workflows
• Provenance
• Tools
• Numerical Reproducibility and Scalability
• Cultural, Ethical, and Legal Challenges
Importance of Reproducibility
Better Understanding of how to do Reproducible Work
What's Next
Final Project

- Tuesday, Dec. 20, 3-6pm in Dion 101
- 8-10 minutes
- Look forward to seeing your work and reading your reports