DTI Image Processing Pipeline and Cloud Computing Environment

Kyle Chard
Computation Institute
University of Chicago and Argonne National Laboratory
Introduction

• DTI image analysis requires the use of many tools
  – QC, Registration, ROI Marking, Fiber Tracking, ..
• Constructing analyses is challenging
  – Data & tool discovery, selection, orchestration, ..
• We have made huge strides in terms of data
  – Data formats, repositories, protocols, metadata, CDEs
• We now need infrastructure to reduce the barriers that exist between data providers, tool developers, researchers, and clinicians
  – Big Science. Small Labs
    o We have exceptional infrastructure for the 1%, what about the 99%?
Common Approach to Analysis

- Modify
- Install
- (Re)Run Script

FSL
Camino

BMC Neuroscience
Google

www.ci.anl.gov
www.ci.uchicago.edu

DTI Pipelines and Cloud Infrastructure
How can we improve?

- We need a platform where users can easily construct and execute analyses
  - Using best of bread tools and pipelines
  - Abstracting low level infrastructure and platform heterogeneity
  - Supporting automation and parallelism
  - Supporting experimentation

=> Make existing tools and common analyses mundane building blocks
• Ultimate Goal: Investigate the feasibility of using DTI in clinical practice

• Automatic calculation of DTI metrics (FA, MD) from 48 automatically generated ROIs
  – Using existing tools to create a reusable analysis workflow that can be easily repeated
  – Investigate the ability to scale analyses over large datasets

• Explore the reproducibility over a group of 20 subjects with 4 scans spread over 2 sessions
DTI Processing Pipeline (1)

BVEC & BVAL

1. ECC DTI (FSL)

2. BET DTI (FSL)

3. BET T1 (FSL)

4. Linear Registration DTI / T1 (FSL FLIRT)

5. DTI Fitting (FSL/Camino)

6. DTI Fitting (FSL/Camino)

7. Linear Registration T1/Template (FSL FLIRT)

7. Non-linear Registration T1/Template (FSL FNIRT)

8. Calculate ROI Mean FA/MD (AFNI 3dmaskave)

9. Transform FA/MD to MNI space (FSL Applywarp)

Atlas Mask

Template
DTI Processing Pipeline (2)

1. ECC DTI (FSL)
2. BET DTI (FSL)
3. DTI Fitting (FSL/Camino)

Linear Registration (FSL FLIRT)
Non-Linear Registration (FSL FNIRT)

FA image → FA in MNI space
MD image → MD in MNI space

FA in MNI space → 6. Calculate ROI Mean (3dmaskave)

MD in MNI space → Apply Warp coefficient

Atlas Mask

FA Template

DTI BVEC & BVAL
Approaches for Implementing Pipelines

**Scripts**
- Bash scripts written to execute tools on a single computer
- Time consuming, error prone, hard to transfer knowledge
- Little support for parallelization

**XNAT Pipeline Engine**
- Defined by code (XML + scripts)
- Overhead to include tools, develop interfaces and create pipelines
- Difficult to change tools/pipelines
- Some support for parallelization

**Globus Genomics**
- SaaS for genomics
- Graphical interface for creation and execution
- Supports on-demand provisioning based on pricing policies
- Tools installed dynamically when required

---

[Image of XNAT Pipeline Engine]

[Image of Globus Genomics]
DTI Pipeline Platform

Globus Endpoints

Galaxy & Manager
- Globus Transfer
- Galaxy
- Condor

Shared File System
Dynamic Scheduler

Dynamic Worker Pool

www.ci.anl.gov
www.ci.uchicago.edu
DTI Pipelines in the Cloud

DTI Pipelines and Cloud Infrastructure
Cloud Computing

- Leverages economies of scale to facilitate utility models
  - Pay only for resources used
  - 1 * 100 hours == 100 * 1 hour
- On-demand and elastic access to “unlimited” capacity
  - Addresses fluctuating requirements
- Web access to data through defined interfaces
- Platform as a Service
  - No management of hardware or low level tools
Challenges Moving to the Cloud

- **Resource Selection**: Comparing price, capabilities, performance, instance types (EBS, Instance store), tool performance

- **Tool Selection and Management**: Finding tools, installing, configuring and using them in different environments

- **Analysis/Resource Management**: Developing structured and repeatable analyses with different tools.

- **Data transfer**: Moving large amounts of data in/out of Cloud environment reliably and efficiently

- **Scale and Parallelism**: Scaling analyses by efficiently parallelizing across elastic infrastructure

- **Security**: Data and computation security - HIPAA?
## Amazon EC2 Pricing

<table>
<thead>
<tr>
<th>System Specifications</th>
<th>Pricing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CPU Units</td>
</tr>
<tr>
<td>m1.large</td>
<td>4</td>
</tr>
<tr>
<td>m1.xlarge</td>
<td>8</td>
</tr>
<tr>
<td>m3.xlarge</td>
<td>13</td>
</tr>
<tr>
<td>m3.2xlarge</td>
<td>26</td>
</tr>
<tr>
<td>m2.xlarge</td>
<td>6.5</td>
</tr>
<tr>
<td>m2.2xlarge</td>
<td>13</td>
</tr>
<tr>
<td>m2.4xlarge</td>
<td>26</td>
</tr>
</tbody>
</table>
Spot Pricing Volatility

![Graph showing pricing volatility for Linux/UNIX instances over a 1-week period with different instance types and zones. The graph displays the price changes for different regions such as us-east-1a, us-east-1b, us-east-1c, us-east-1d, and us-east-1e.](image-url)
Instance Performance and Pricing

- **EBS**
- **Instance Store**
- **Spot (Low)**
- **Spot (High)**
- **On-Demand**

<table>
<thead>
<tr>
<th>Instance Type</th>
<th>Time (Minutes)</th>
<th>Cost per Subject ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>m1.large</td>
<td>100</td>
<td>0.3</td>
</tr>
<tr>
<td>m1.xlarge</td>
<td>80</td>
<td>0.3</td>
</tr>
<tr>
<td>m3.xlarge</td>
<td>60</td>
<td>0.4</td>
</tr>
<tr>
<td>m3.2xlarge</td>
<td>40</td>
<td>0.5</td>
</tr>
<tr>
<td>m2.xlarge</td>
<td>60</td>
<td>0.6</td>
</tr>
<tr>
<td>m2.2xlarge</td>
<td>80</td>
<td>0.7</td>
</tr>
<tr>
<td>m2.4xlarge</td>
<td>100</td>
<td>0.8</td>
</tr>
</tbody>
</table>

www.ci.anl.gov
www.ci.uchicago.edu
Pricing - Multiple Analyses Per Node

---

- **On-Demand**
- **Spot (Low)**
- **Spot (High)**

<table>
<thead>
<tr>
<th>Node</th>
<th>Cost per Subject ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>m1.large</td>
<td>0.1</td>
</tr>
<tr>
<td>m1.xlarge</td>
<td>0.2</td>
</tr>
<tr>
<td>m3.xlarge</td>
<td>0.3</td>
</tr>
<tr>
<td>m3.2xlarge</td>
<td>0.4</td>
</tr>
<tr>
<td>m2.xlarge</td>
<td>0.5</td>
</tr>
<tr>
<td>m2.2xlarge</td>
<td>0.6</td>
</tr>
<tr>
<td>m2.4xlarge</td>
<td>0.7</td>
</tr>
</tbody>
</table>

---

www.ci.anl.gov
www.ci.uchicago.edu

DTI Pipelines and Cloud Infrastructure

Argonne

THE UNIVERSITY OF CHICAGO

www.ci.anl.gov
www.ci.uchicago.edu
Elastic Startup Cost

New Worker

- ROI Calculation
- Tensor Fitting
- ECC & Registration
- Contextualize
- Spot Price
- Queue

Existing Worker
Data Transfer with Globus Online

- Reliable file transfer, sharing, syncing.
  - Easy “fire and forget” file transfers
  - Automatic fault recovery
  - High performance
  - Across multiple security domains
- In place sharing of files with users and groups
- No IT required.
  - Software as a Service (SaaS)
    - No client software installation
    - New features automatically available
Transfer Comparison

The graph illustrates the comparison of different transfer rates for various file sizes. The x-axis represents the file size in bytes, while the y-axis shows the rate in mbps.

- **Globus Transfer**: The blue line with square markers shows the highest transfer rate, which increases with file size.
- **FTP**: The red line with diamond markers indicates a lower transfer rate compared to Globus, also increasing with file size.
- **HTTP**: The green line with triangle markers shows the lowest transfer rate among the three, with a steady increase.

The graph highlights the efficiency differences between these transfer methods, with Globus Transfer being the most efficient for large file sizes.
Summary

• Structured pipelines simplify creation, execution and sharing of complex analyses
  – Hosted as a service can further reduce barriers
• By outsourcing pipeline execution on the Cloud we can reduce overhead and costs
  – Previously we took weeks to process ~100 scans
    o Using this approach < 5 cents a subject ($5 for 1 hour)
• What's next?
  – Can we deliver this as a service?
    o Billing, security, paradigm shift, interactive tools …
  – Developing toolsheds for sharing tools and pipelines
Acknowledgements

- Mike Vannier, Xia Jiang, Farid Dahi
- Globus Online
  - Ian Foster, Steve Tuecke, Rachana Ananthakrishnan
- Globus Genomics
  - Ravi Madduri, Paul Dave, Dina Sulakhe, Lukasz Lacinski