A-Brain: Using the Cloud for Neuroimaging and Genetics Research

Gabriel Antoniu, Bertrand Thirion
MSR-INRIA Joint Research Centre
Joint work with Alexandru Costan, Benoit Da Mota, Radu Tudoran
and with the Microsoft Azure team from EMIC
The A-Brain Project: BlobSeer on Microsoft Azure Clouds

Application
- Large-scale joint genetic and neuroimaging data analysis

Goal
- Assess and understand the variability between individuals

Approach
- Optimized data processing on Microsoft’s Azure clouds based on the BlobSeer concurrency-optimized platform

Inria teams involved
- KerData (Rennes)
- Parietal (Saclay)

Framework
- Joint MSR-Inria Research Center
- MS involvement: Azure teams, EMIC
The Parietal Team

- INRIA team, created in 2009, ~12 people
- Involved in MRI-based brain imaging data analysis
- Emphasis on statistical methods, machine learning and computational anatomy.
- Situated on the main french platform for brain imaging, Neurospin (CEA)
- Many recent contributions on brain connectivity
Neuroimaging-genetics: The Problem

- Several brain diseases have a genetic origin, or their occurrence/severity related to genetic factors
- Genetics important to understand & predict response to treatment
- Identify risk and protective factors for brain diseases
- Brain: Huntington's disease, autism...

- Currently: large-scale studies to assess the relationships between diseases and genes: typically $10^4$ patients per study + control groups
- Genetic variability captured in DNA microarray data
The Problem: Neuroimaging Data

Brain images can be used to understand, model and quantify various characteristics of the brain:

Morphology: shape, thickness

Structure: anatomical connectivity

Function: response to stimulation, functional connectivity
The Imaging Genetics Challenge: Comparing Heterogeneous Information

Clinical / behaviour

Genetic information: SNPs

Here we focus on this link

MRI brain images
Imaging Genetics: Why?

- Study neurobiological and genetic basis for subject variability in behavior and cognitive functions
- Correlations between brain responses and genetic data may highlight risk factors in target populations
- Neuroimaging as the intermediate between genetics and behavior
- A major research direction of the Parietal INRIA team in Saclay

\[ p(\text{image} | \text{behavior}) \]

\[ p(\text{gene} | \text{image}) \]

Image \rightarrow Behavior

Gene \rightarrow Image
Imaging Genetics Methodological Issues

Brain image \leftrightarrow Genetic data

finding associations: \[ p(Y, X) \]

- Anatomical MRI
- Functional MRI
- Diffusion MRI

\[ q \approx 10^{5-6} \]

- DNA array (SNP/CNV)
- gene expression data
- others...

\[ p \approx 10^6 \]

N \approx 2000
Imaging Genetics Methodological Issues

- Multivariate methods: predict brain characteristic with many genetic variables
- Elastic net regularization: combination of $\ell_1$ and $\ell_2$ penalties $\rightarrow$ sparse loadings
- $O(p^3)$ complexity
- Parameters setting: internal cross-validation/bootstrap
- Performance evaluated using permutations

$$\hat{\beta}_{enet} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} (y_i - \sum_{k=1}^{p} x_{ik}\beta_k)^2 + \lambda_1 \sum_{k=1}^{p} |\beta_k| + \lambda_2 \sum_{k=1}^{p} \beta_k^2 \right\}$$
4D cluster size \( d \)
distribution under \( H_0 \)

A-Brain:
The goal is to reproduce this kind of study with \( 10^5 \) larger data
The Computational Problem

Neuroimaging data (voxels in each contrast map): $10^5$ to $10^6$

Genetic data: $10^6$ variables

Permutation tests: $10^3$

Around $10^{15}$ tests
Focus of the KerData Team: the BlobSeer Approach to Concurrency-Optimized Data Management

BlobSeer: software platform for scalable, distributed BLOB management
- Decentralized data storage
- Decentralized metadata management
- Versioning-based concurrency control
- Lock-free concurrent writes (enabled by versioning)

A back-end for higher-level data management systems
- Short term: highly scalable distributed file systems
- Middle term: storage for cloud services
- Long term: extremely large distributed databases

Methodology
- Design and implementation of distributed algorithms
- Experiments on the Grid’5000 grid/cloud testbed
- Validation with “real” apps on “real” platforms: Nimbus, Azure, OpenNebula clouds...

http://blobseer.gforge.inria.fr/
Highlight: BlobSeer vs. Hadoop

Execution time reduced by up to 38%

MapReduce: a natural application class for BlobSeer


A step further: application to VM multi-deployment and multi-snapshotting (HPDC 2011)
The MapReduce ANR Project (2010-2014)

Goal: an optimized Map-Reduce platform for cloud infrastructures
Total cost: 3,1M€, ANR funding: 827K€

Partners
- INRIA - KerData team (INRIA, Rennes) – leader
- INRIA - GRAAL team (Lyon), France
- Nimbus team, Argonne National Lab/University of Chicago, USA
- University of Illinois at Urbana Champaign, USA
- Joint UIUC/INRIA Laboratory for Petascale Computing
- IBM Products and Solutions Center, Montpellier, France
- Institute of Biology and Chemistry of Proteins, Lyon, France
- MEDIT (SME), Palaiseau, France
BlobSeer as a Storage Backend for Sharing Application Data: TomusBlobs
A-Brain: Application’s Throughput

- **Read:** $2.5x$
- **Write:** $3x$
A-Brain: Resources

Access to the Azure platform
• 2 million hours per year and 10 TBytes for storage will be available

Dedicated human resources
• 2 Postdoctoral fellows in Rennes and Saclay
• 1 PhD student in Rennes
A-Brain Timeline

Engagement “A-Brain” in M2:
1. Port and evaluation of BlobSeer on Azure: TomusBlobs+TMR
2. Evaluation of benefits of integrating BlobSeer with Microsoft Azure storage services: ongoing
3. Large-scale application experiments:
   So far: up to 200 with a toy application

Next steps:
• Paper on preliminary architecture (CCGRID 2012)
• Research in progress on M/R optimized scheduling
• Redesign of the Neuro application using a cloud-oriented programming model
  • Meetings PARIETAL-KERDATA
• Next milestone: demo in Spring
People Involved

Thank you!

- Gabriel Antoniu
- Bertrand Thirion
- Pierre Louis Xech
- Götz-Philip Brasche
- Benoit Da Mota
- Hakan Soncu
- Alexandru Costan
- Radu Tudoran
- Louis-Claude Canon