A-Brain: Using the Cloud to Understand the Impact of Genetic Variability on the Brain

Gabriel Antoniu, Bertrand Thirion
INRIA Rennes Bretagne Atlantique
INRIA Saclay–Île-de-France

Joint project with Microsoft Azure teams
within the
Microsoft Research – INRIA Joint Centre

INRIA-MSR forum, 12 April 2011

The A-Brain Project in a Nutshell

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 INRIA's BlobSeer data management system

INRIA teams involved

- KerData (Rennes)
- PARIETAL (Saclay)

Framework

- Joint MSR-INRIA Research Center
- MS involvement: Azure teams, EMIC

Posté le 28/10/2010 | **Institutionnel**

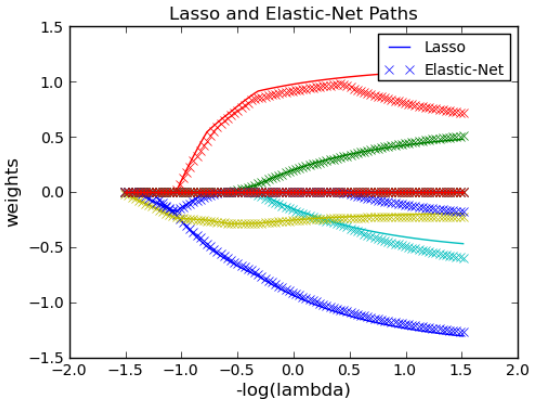
Microsoft et l'INRIA annoncent un partenariat autour du Cloud Computing

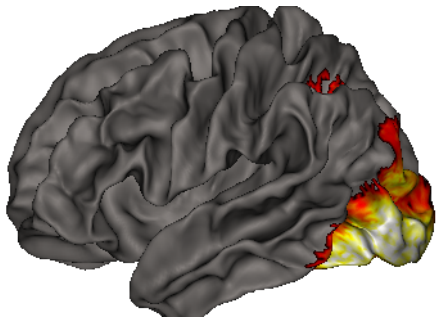
Pendant 2 ans, le projet AzureBrain, au service de la neuro-imagerie, va bénéficier des solutions Microsoft Windows Azure

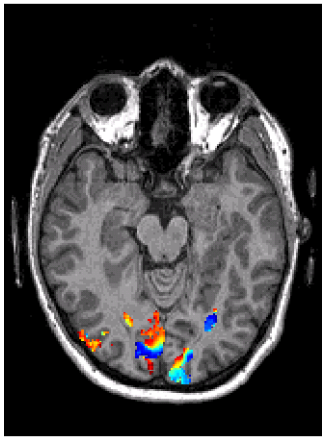
Issy-les-Moulineaux, le 28 octobre 2010 – Microsoft et l'INRIA (Institut National de Recherche en Informatique et en Automatique) renforcent aujourd'hui leur collaboration avec le lancement du projet de recherche AzureBrain qui sera réalisé au sein du centre de recherche commun INRIA-Microsoft Research. AzureBrain a pour objectif de permettre des avancées précieuses dans le domaine de la neuroscience et de la neuro-imagerie. Microsoft met concrètement au service de ce projet, des ressources de Cloud Computing qui permettent d'accélérer le rythme des recherches, offrant des puissances de calcul et de traitement sur-mesure, à travers les datacenters.

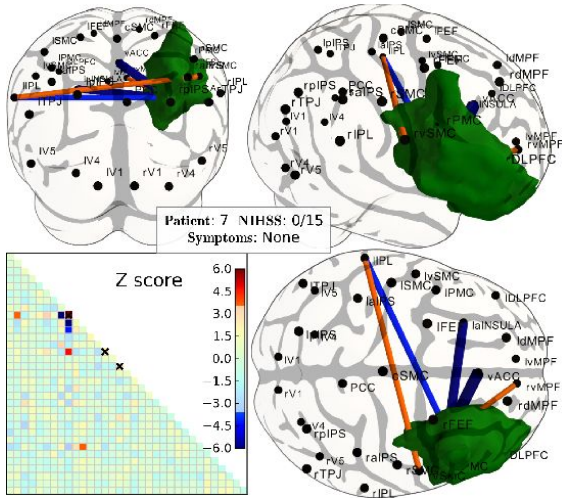
Parietal

- INRIA team, created in 2009, ~10 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

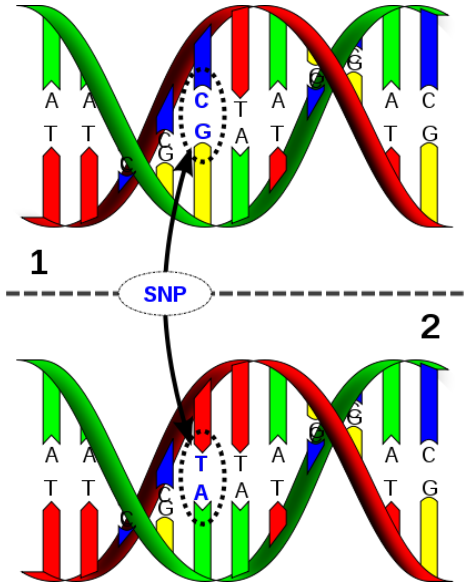






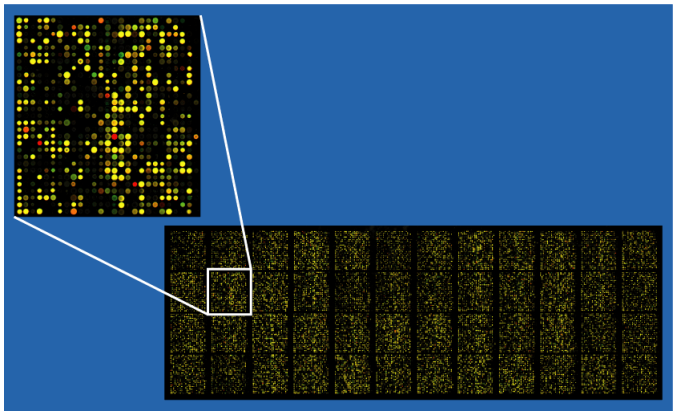


Neuroimaging-genetics: the Problem



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

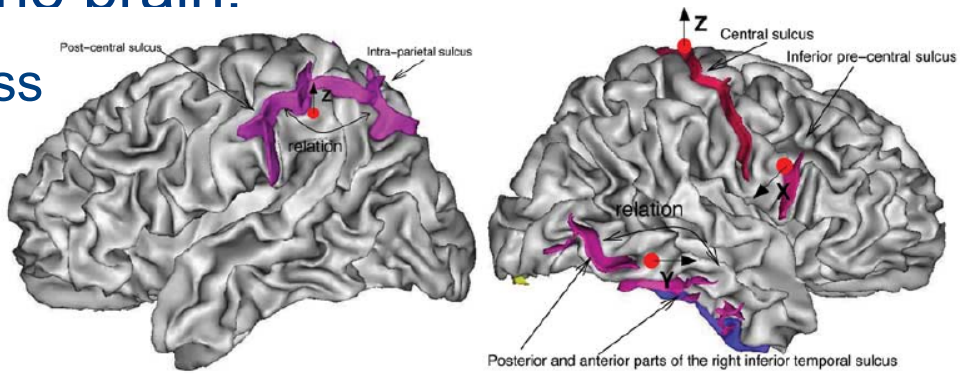
- Currently: large-scale studies to assess the relationships between diseases and genes: typically 10^4 patients per study + control groups
- Genetic variability captured in DNA microarray data

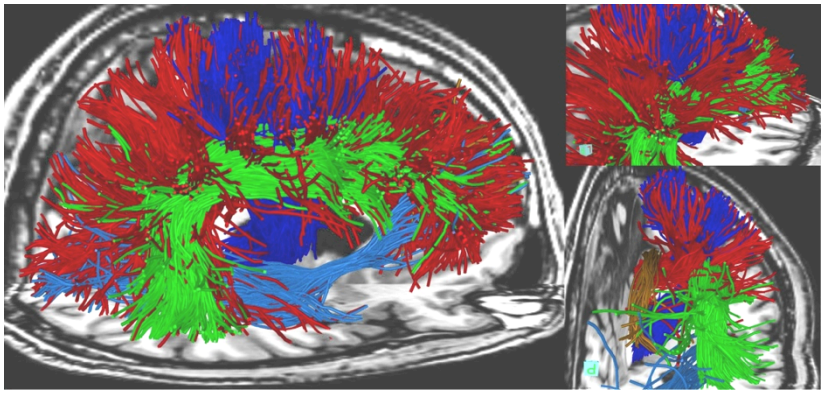


The Problem: Neuroimaging Data

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

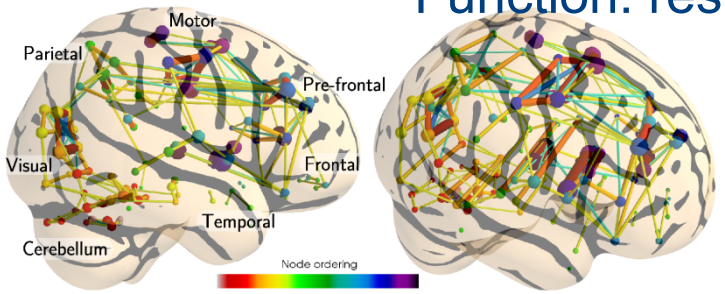
Morphology: shape, thickness

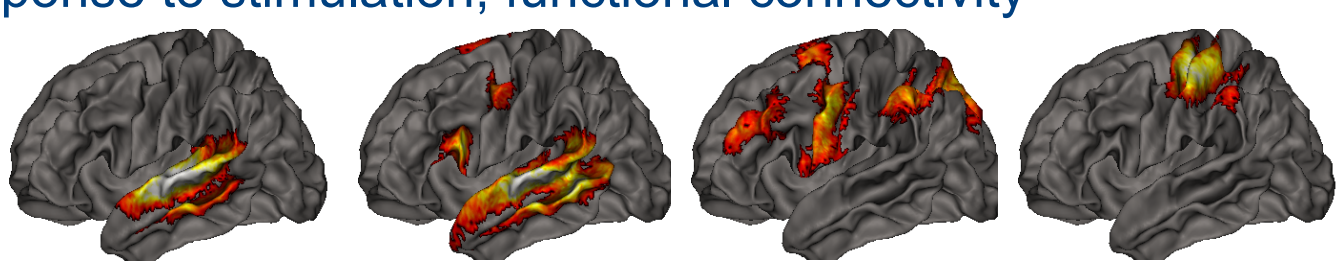




Structure: anatomical connectivity

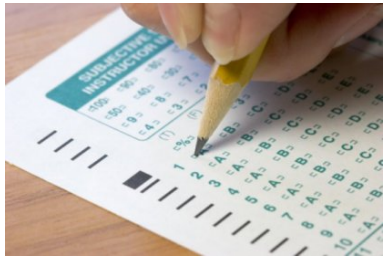
Function: response to stimulation, functional connectivity



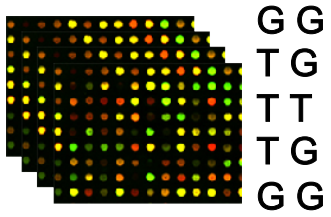


Neuroimaging: Intermediate Information between Genetics and Behaviour/Diseases

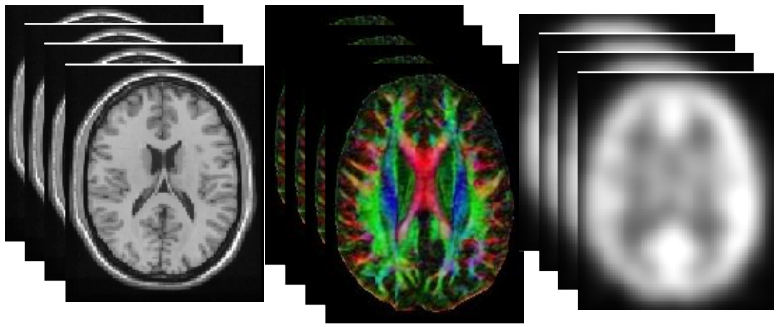
Clinical / behaviour



Genetic information: SNPs



Here we focus on this link



MRI brain images

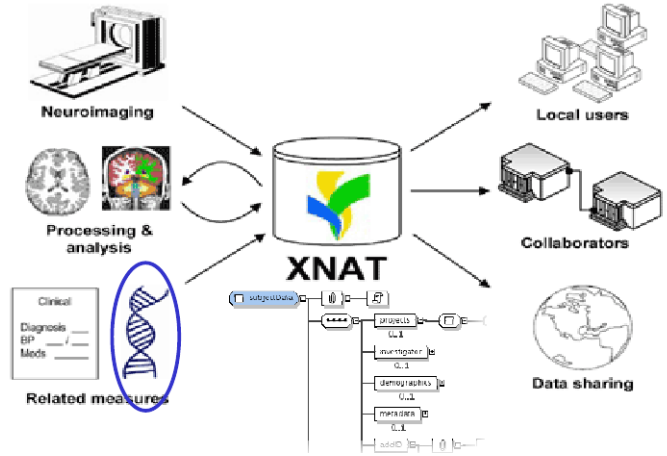
Hypothesis: brain images contain useful markers that relate genetics to behaviour/diseases

Neuroimaging-genetics:
Designing the Statistical Study

- Univariate studies: find a (SNP, neuroimaging trait) that are significantly correlated
  - e.g. the amount of functional activity in a brain region is related to the presence of a minor allele on a gene
- Regression studies: some sets of SNPs predict a neuroimaging/behavioral trait
  - e.g. a set of SNPs altogether predict a given brain characteristic
  - SNP set can be on a given gene or not
- Multivariate studies: an ensemble of genetic traits predict a certain combination of neuroimaging traits
  - Emphasis on the underlying network structure

Example of a dataset that might be studied in A-brain: The Imagen Database

- FP7 European project, 2007-2012
  [Schumann et al, Mol Psychiatry]
- Study neurobiological and genetic basis of reinforcer sensitivity, cognitive control and emotional reactivity + assess their relevance for mental disorder.
  - individual differences in brain responses x genotype may mediate risk factor in adolescents.
  - abnormalities in those brain processes are implicated in psychiatric disorder (addiction)
- Adolescents + longitudinal : predictive value of the assessment
- Database hosted and processed at Neurospin (CEA, DSV, LNAO)



Statistical Methodology

Model : allelic dosage model

For each (trait, SNP) pair $y = \mu + x\beta + z_c\beta_c + \epsilon$

Signal in one brain region

Number of minor alleles at a given genetic location

Confounding factors

$$p(\|\beta\|^2 | H_0) \sim \mathcal{F}_{\dim(\beta),\nu}$$

- The test is performed at each pair: n*p tests (up to 10^{12})
- Significance assessed by permutation test: 10^4 replications of the regression (correction for multiple comparisons)
- Efficient and standard, but…
- Very weakly sensitive method: only large effects can be detected

Statistical Methodology

- Instead, predict some brain characteristic using many genetic sites simultaneously

$$y = X\beta + \epsilon''$$

- Typically yields a rank deficient system, thus requires regularization

$$\hat{\beta} = \text{argmin}_\beta \|y - X\beta\| + \psi(\beta)$$

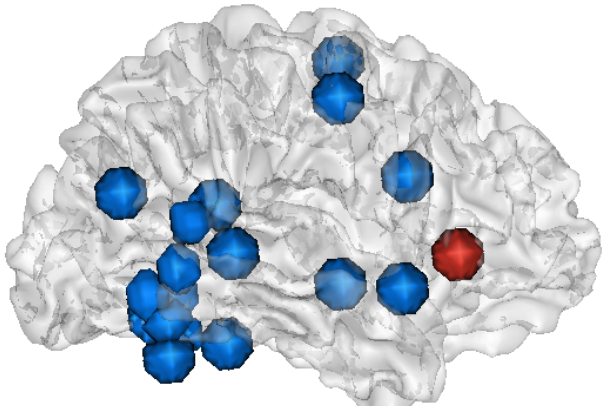
- State of the art: elastic net regularization: combination of L1 and L2 penalties.
  →sparse loadings

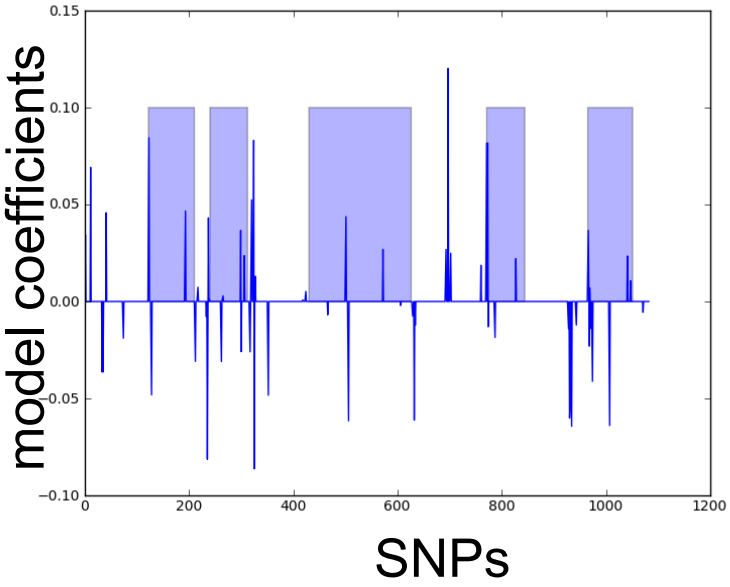
$$\psi(\beta) = \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$$

- Fit becomes expensive.
- Requires setting the parameters of the regularization (internal cross-validation)
- Performance evaluated using permutations

Statistical Methodology: Example on a Pilot Study

- In one region (45, 27, -3), about 10% of the asymmetry value is fit by the elastic net model
- Parameter tuning done by internal/ nested cross-validation
- Statistical test: the association strength is significant at $p<0.03$ corrected for multiple comparisons
- More sensitive than simple association study

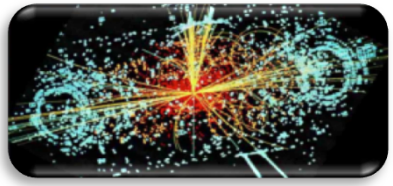




A General Problem Today: the Data Deluge

Experiments

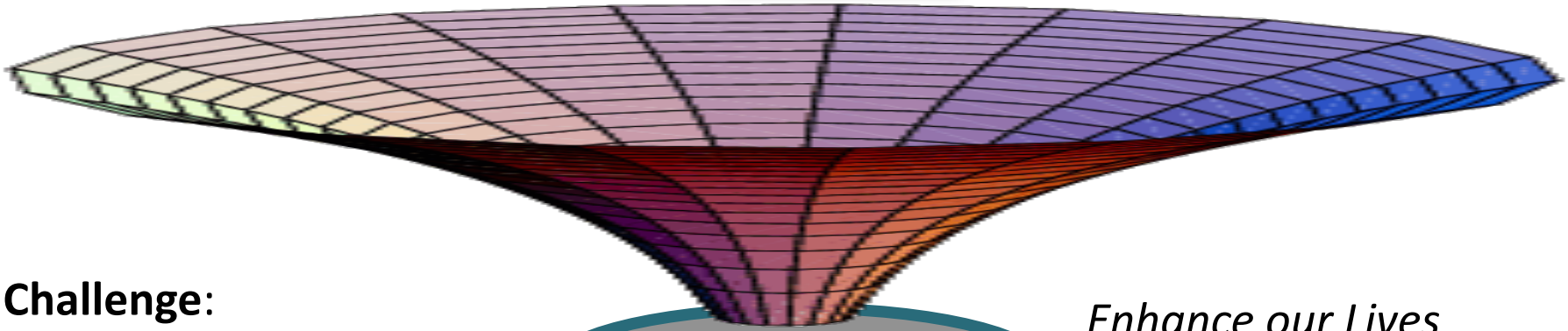
Simulations



Archives

Literature

Instruments



The Challenge:
Enable Discovery.
Deliver the capability to mine, search and analyze this data in near real time.

Petabytes
Doubling every 2 years

Enhance our Lives
Participate in our own heath care. Augment experience with deeper understanding.

Credits: Microsoft

KerData: Scalable Storage for Clouds and Beyond

A joint team at INRIA (Rennes) – ENS Cachan/Brittany, created in 2009, ~10 people

Focus: Scalable storage for new-generation, data-oriented high-performance applications

- Massive, unstructured data objects (Terabytes)
- Many data objects (10^3-10^6)
- High concurrency (10^3 concurrent clients)
- Fine-grain access (Megabytes)

Applications: distributed, with high-throughput requirements under concurrency

- Map-Reduce-based data-mining applications
- Governmental and commercial statistics
- Data-intensive HPC simulations
- Checkpointing for massively parallel computations

Target platforms: large clusters, clouds, Post-Petascale machines

http://www.irisa.fr/kerdata/

Our Current Focus: the BlobSeer Approach to Concurrency-Optimized Data Management

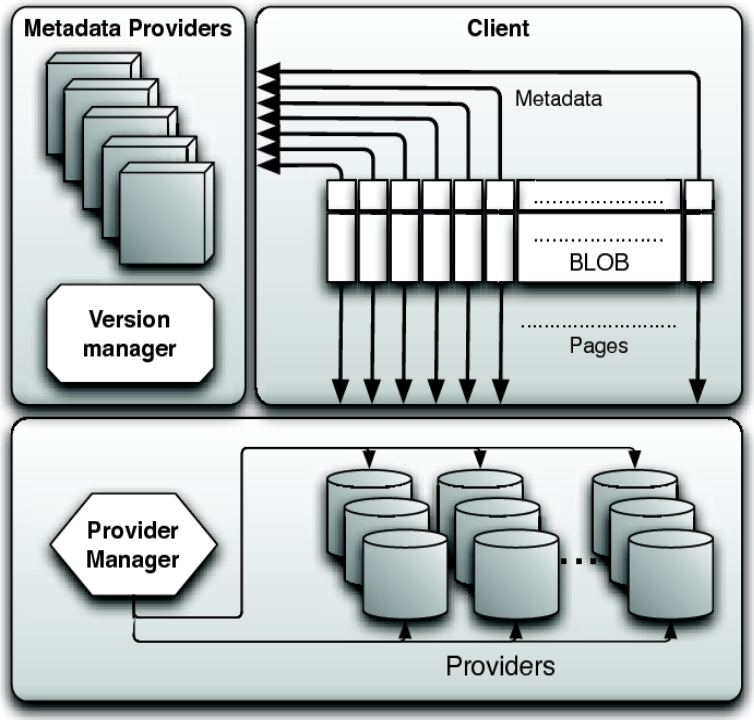
BlobSeer: software platform for scalable, distributed BLOB management

- 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

Validated on the ALADDIN-Grid'5000 experimental grid/cloud testbed



http://blobseer.gforge.inria.fr/

Leveraging BlobSeer on Clouds: MapReduce

- MapReduce: a simple programming model for data-intensive computing on clouds
- Typical problem solved by MapReduce
  - Read a lot of data
  - Map: extract something you care about from each record
  - Shuffle and Sort
  - Reduce: aggregate, summarize, filter, or transform
  - Write the results
- Approach: hide messy details in a runtime library
  - Automatic parallelization
  - Load balancing
  - Network and disk transfer optimization
  - Transparent handling of machine failures
- Implementations: Google MapReduce, Hadoop (Yahoo!)

Integrating BlobSeer in the Hadoop Map-Reduce Framework

MapReduce: a natural application class for BlobSeer:

- Case study: Yahoo!'s Hadoop MapReduce framework
- Approach: use BlobSeer instead of Yahoo!'s Hadoop file system (HDFS)
- Motivation: HDFS has limited support for concurrent access to shared data

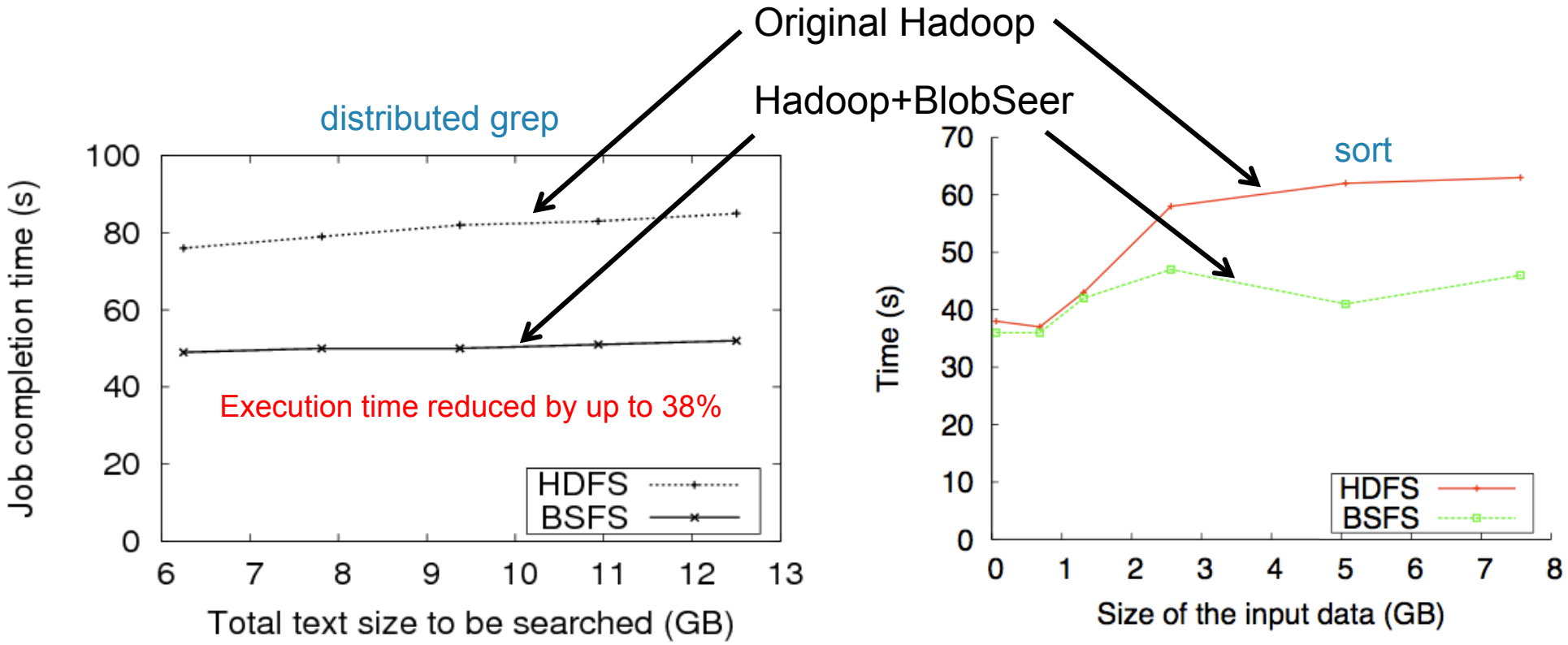
Implementing the HDFS API for BlobSeer

- Implements basic file system operations: create, read, write...
- Introduces support for concurrent append operations

BlobSeer File System (BSFS)

- File system namespace - keeps file metadata, maps files to BLOB's
- Client-side buffering: data prefetching, write aggregation
- Exposes data layout to Hadoop, just like HDFS

Highlight: BlobSeer Does Better Than Hadoop!



MapReduce: a natural application class for BlobSeer

- Study: BlobSeer as a file system for Yahoo!'s Hadoop MapReduce framework
- Publications: JPDC(2010), IPDPS 2010, MAPREDUCE 2010

The MapReduce Project (2010-2014)

Goal: an optimized Map-Reduce platform for cloud infrastructures

Total cost: 3,1M€, ANR funding: 827K€

Partners

- INRIA - KerData team (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

CENTRE DE RECHERCHE COMMUN INRIA MICROSOFT RESEARCH

INRIA

From MapReduce to the A-Brain Project

Application

- Large-scale Joint Genetic and Neuroimaging Data Analysis

Approach

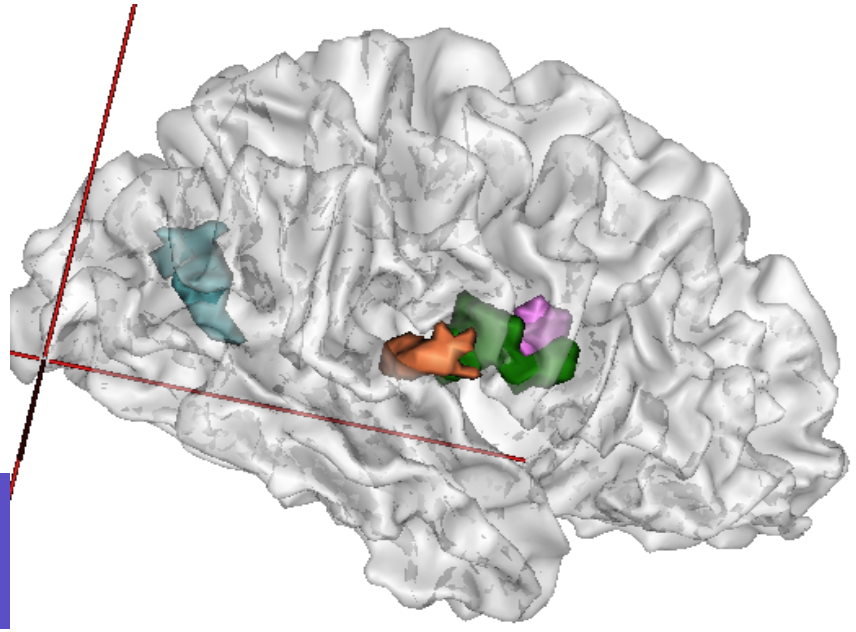
- Optimized data processing on Azure clouds

Bricks

- Application (INRIA/PARIETAL)
- BlobSeer data management system (INRIA/KerData)
- Platform: Microsoft Azure cloud

Posté le 28/10/2010 | **Institutionnel**

Microsoft et l'INRIA annoncent un partenariat autour du Cloud Computing



AzureBrain: Resources

Access to the Azure platform

- 2 million hours per year and 10 TBytes for storage will be available

Human resources

- Several researchers in two INRIA teams
  - KerData (Rennes): optimized cloud storage
  - Parietal (Saclay): neuroimaging and genetics
- Dedicated human resources (to be hired!)
  - Postdoctoral fellows and engineers both in Rennes and Saclay

Roadmap for A-Brain and Beyond

WP1: Application

- Task 1: Understand and extract relevant input data
- Task 2: Structure the application output
- Task 3: Redesign the application using a cloud-oriented programming model

WP2: Optimized cloud data management platform

- Task 4: Design a joint BlobSeer – Azure architecture for data processing
  - IN PROGRESS, preliminary prototype running!
- Task 5: Evaluate the benefits of integrating BlobSeer with Microsoft Azure storage services
  - Potential collaboration initiated with Geoffrey Fox, Indiana University
- Task 6: Evaluate the impact of using BlobSeer on Azure with large-scale application experiments

Possible follow-up

- Collaborative project on geo-replicated cloud storage with MSR Cambridge (Systems and Networking Group, ER proposal under evaluation), co-funding opportunities: INRIA, EIT ICT Labs

For more information…

- The KerData team at INRIA, Rennes: http//www.irisa.fr/kerdata
- The Parietal team at INRIA, Saclay: http://parietal.saclay.inria.fr/

Contacts

- gabriel.antoniu@inria.fr - advanced cloud data management
- bertrand.thirion@inria.fr - joint neuroimaging and genetics data analysis