CENTRE DE RECHERCHE COMMUN
INRIA MICROSOFT RESEARCH

INVENTEURS DU MONDE NUMÉRIQUE

A-Brain and Z-CloudFlow: Scalable Data Processing on Azure Clouds

Lessons Learned in 3 Years and Future Directions

A-Brain Project PIs: Gabriel Antoniu, Bertrand Thirion
Contributors: Alexandru Costan, Benoit Da Mota, Radu Tudoran and the Microsoft Azure team from EMIC

10th JLPC Workshop, NCSA, Urbana
25 November 2013

The Data@Exascale AssociateTeam (2013-2015)

Partners

- INRIA: Gabriel Antoniu, Matthieu Dorier, Radu Tudoran, Shadi Ibrahim, Alexandru Costan
- ANL: Kate Keahey, Rob Ross, Tom Peterka, Dries Kimpe, Franck Cappello
- ANL/UIUC: Marc Snir
- UIUC: Rob Sisneros, Dave Semeraro

Focus

- Open issues related to storage and I/O in HPC and clouds, data visualization and analysis

Data@Exascale: 2013 Highlights

Data@Exascale

Summer Internship of Matthieu Dorier at ANL

In Situ Visualization of HPC Simulations using Dedicated Cores (Damaris)

- People involved: Matthieu Dorier, Gabriel Antoniu, Tom Peterka, Roberto Sisneros, Dave Semeraro, Lokman Rahmani (recently hired as a PhD student at INRIA/KerData)
- **Results**: joint paper published at IEEE LDAV 2013, demo and poster at the Inria booth@SC13

Mitigating I/O Interference in Concurrent HPC Applications

- People involved: Matthieu Dorier, Gabriel Antoniu, Shadi Ibrahim, Rob Ross, Dries Kimpe
- Results: paper submitted to IPDPS 2014

Data@Exascale: 2013 Highlights

Data@Exascale

Summer Internship of Matthieu Dorier at ANL

In Situ Visualization of HPC Simulations using Dedicated Cores (Damaris)

- People involved: Matthieu Dorier, Gabriel Antoniu, Tom Peterka, Roberto Sisneros, Dave Semeraro, Lokman Rahmani (recently hired as a PhD student at INRIA/KerData)
- **Results**: joint paper published at IEEE LDAV 2013, demo and poster at the Inria booth@SC13

Mitigating I/O Interference in Concurrent HPC Applications

- People involved: Matthieu Dorier, Gabriel Antoniu, Shadi Ibrahim, Rob Ross, Dries Kimpe
- Results: paper submitted to IPDPS 2014

To learn more…

… **attend Mathieu's talk tomorrow at 9am! ☺**

Data@Exascale: 2013 Highlights

Summer Internship of Radu Tudoran at ANL

Evaluating Streaming Strategies for Event Processing across Infrastructure Clouds

- People involved: Radu Tudoran, Kate Keahey, Gabriel Antoniu, Alexandru Costan, Sergey Panitkin
- **Results**: joint paper submitted to IEEE CCGRID 2014

Data@Exascale: 2013 Highlights

Data@Exascale

Summer Internship of Radu Tudoran at ANL

Evaluating Streaming Strategies for Event Processing across Infrastructure Clouds

- People involved: Radu Tudoran, Kate Keahey, Gabriel Antoniu, Alexandru Costan, Sergey Panitkin
- **Results**: joint paper submitted to IEEE CCGRID 2014

To learn more…

… attend Kate's talk on Wednesday at 9:30am! ☺

The A-Brain Project: Data-Intensive Processing on Microsoft Azure Clouds

Application

- Large-scale joint genetic and neuroimaging data analysis

Goals

- Application: assess and understand the variability between individuals
- Infrastructure: assess the potential benefits of Azure

Approach

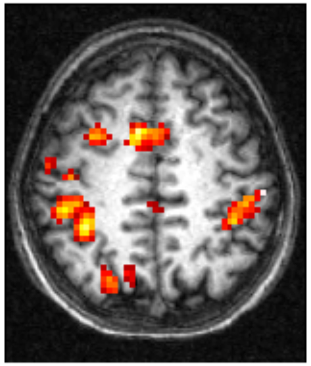
- Optimized data processing on Microsoft's Azure clouds

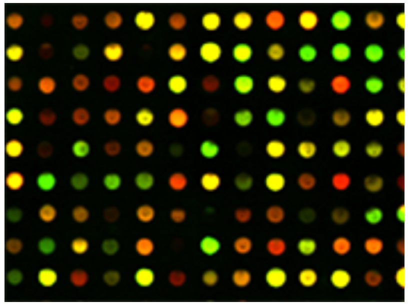
Inria teams involved

- KerData (Rennes)
- Parietal(Saclay)

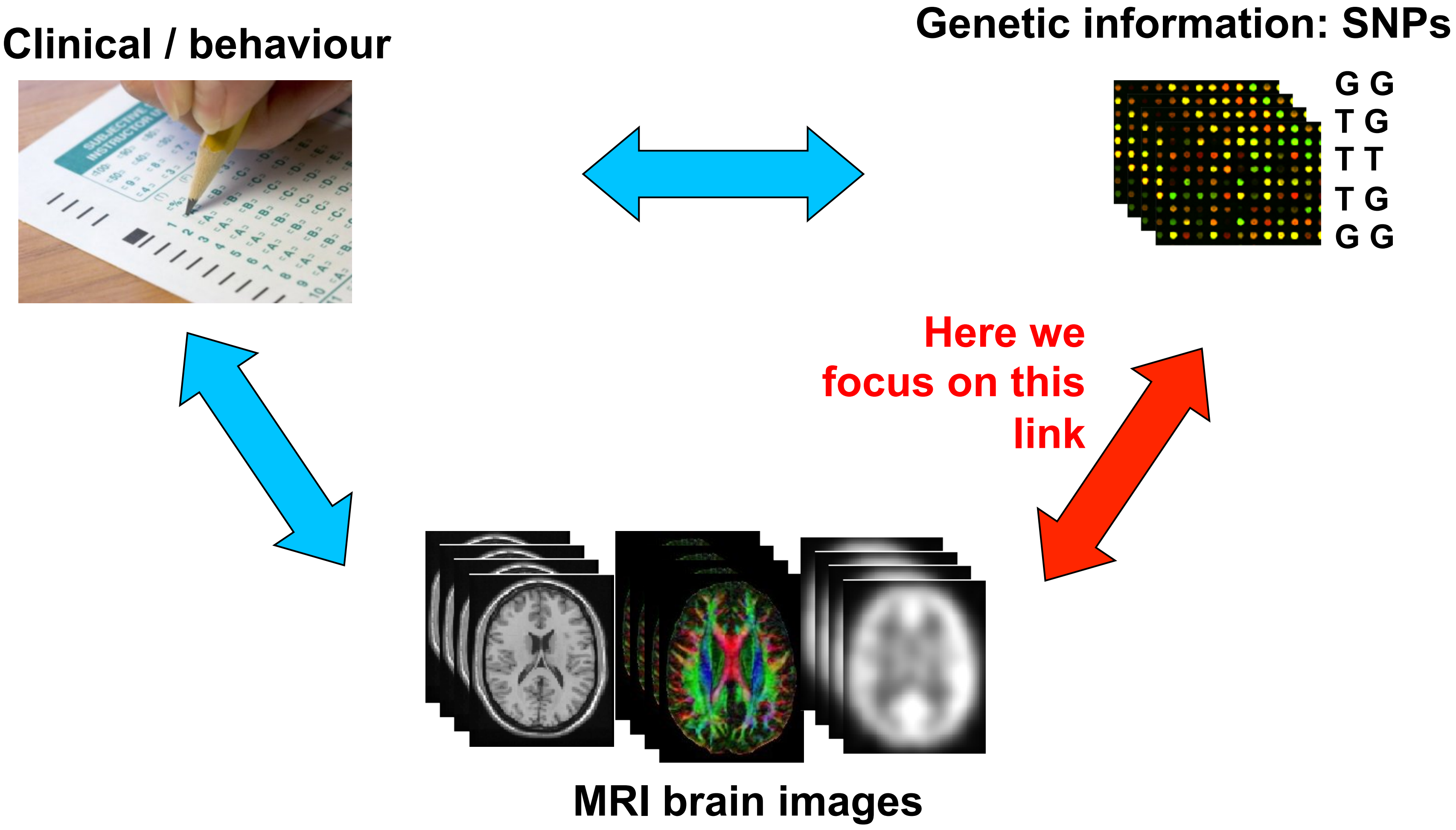
Framework

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





Motivating Application:
The Imaging Genetics Challenge



Clinical / behaviour

Genetic information: SNPs

Here we focus on this link

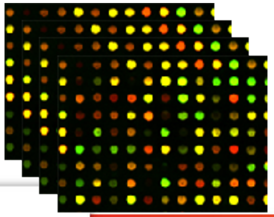
MRI brain images

Neuroimaging-Genetics Studies

- Objective: find correlation between brain markers and genetic data to understand the behavioral variability and diseases

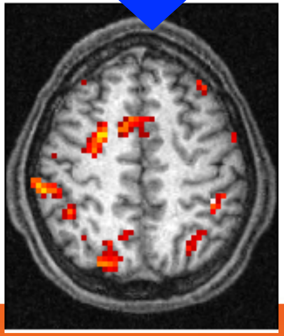
genetics

~10^6 Single nucleotid polymorphisms



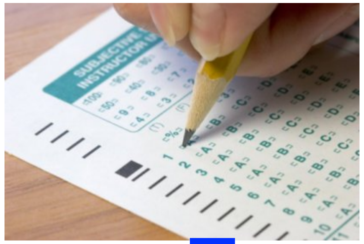
G G
T G
T T
T G
G G

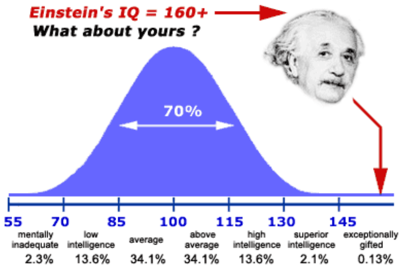
?



?

behaviour

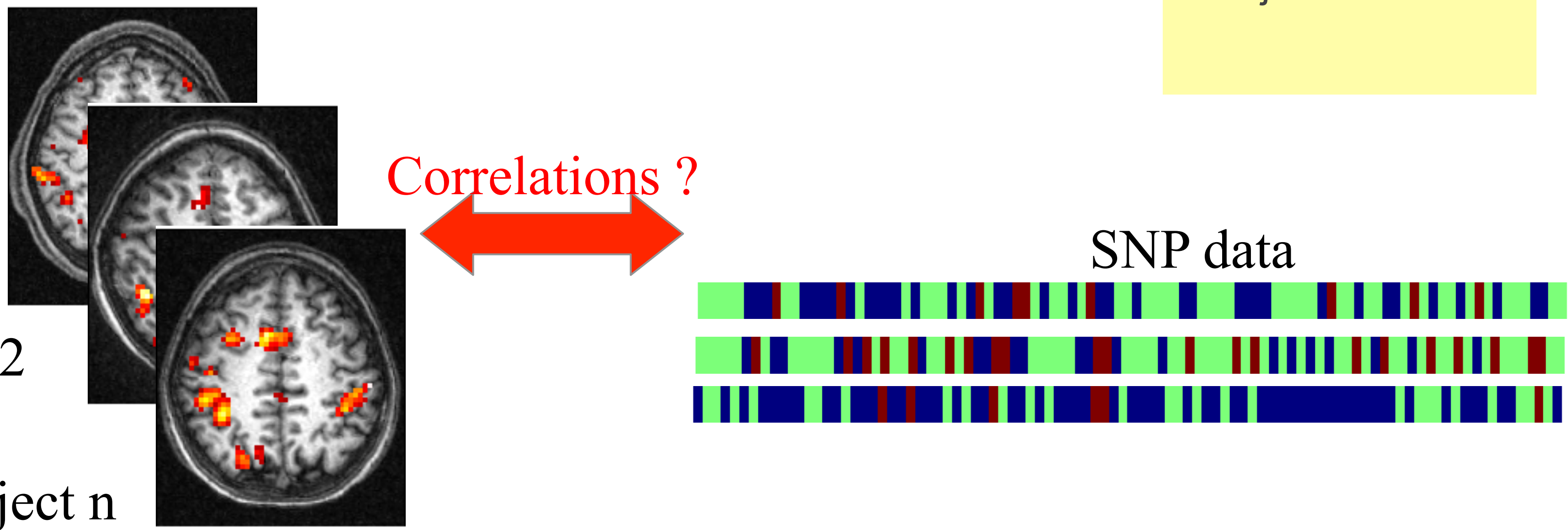




Statistical Analysis for Large-scale Neuroimaging-Genetics

- Image data → 4D to 2D, dimension $n_{voxels} \times n_{subjects}$
- Genetic data → dimension $n_{snps} \times n_{subjects}$
- Statistical question

$n_{voxels} = 10^6$
$n_{snps} = 10^6$
$n_{subjects} = 10^3$



Subject 1

Subject 2

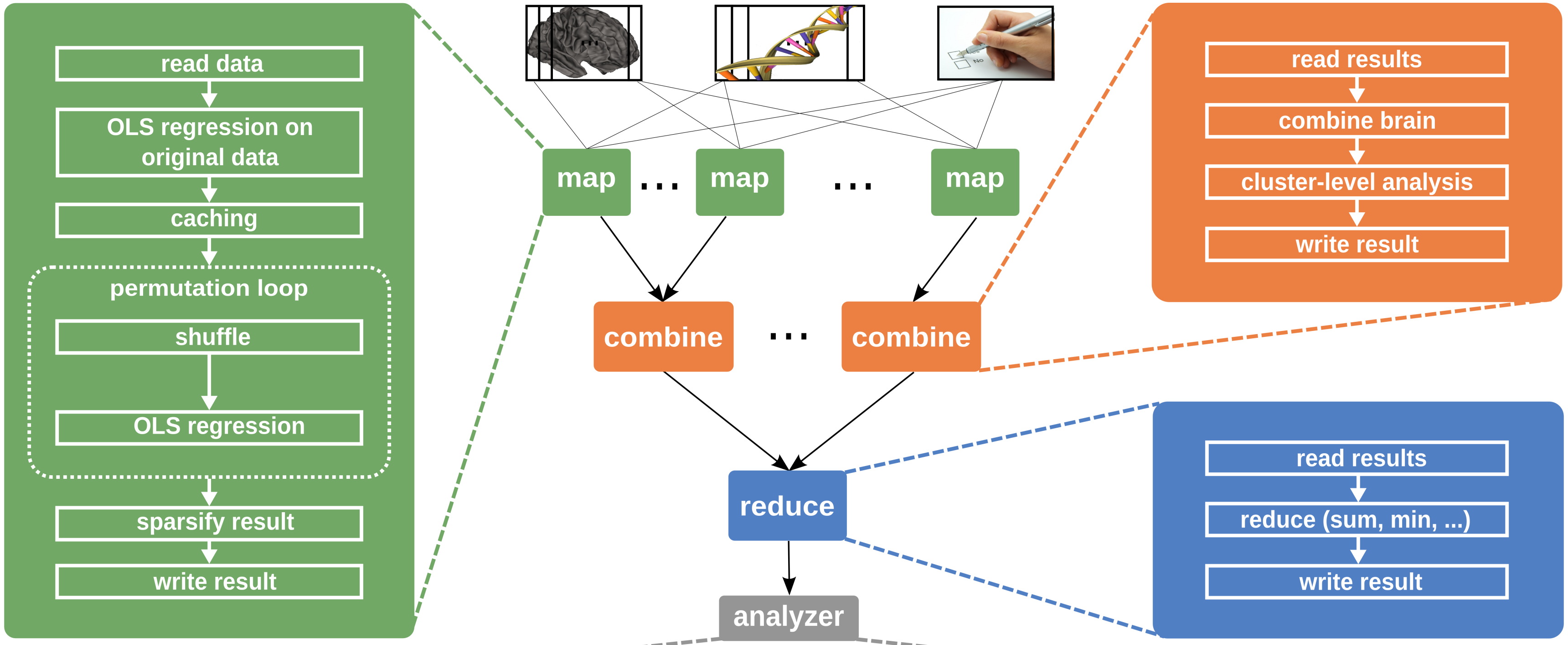
...

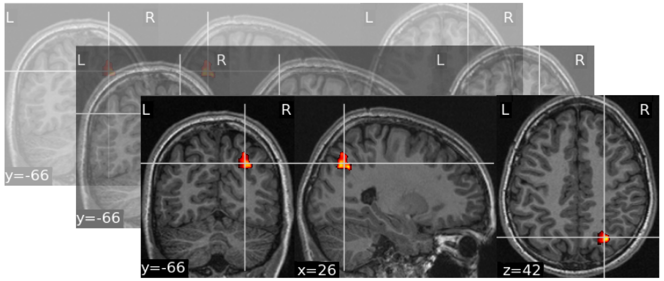
Subject n

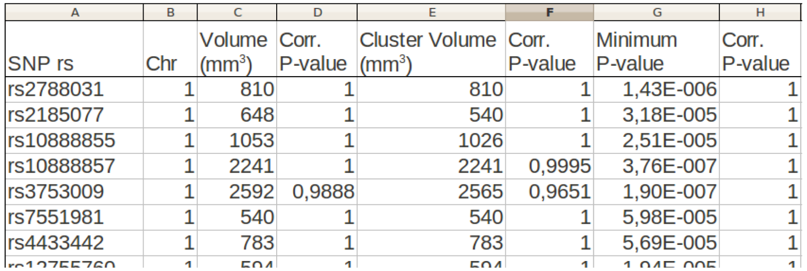
Correlations ?

SNP data

Approach: A-Brain as Map-Reduce Processing







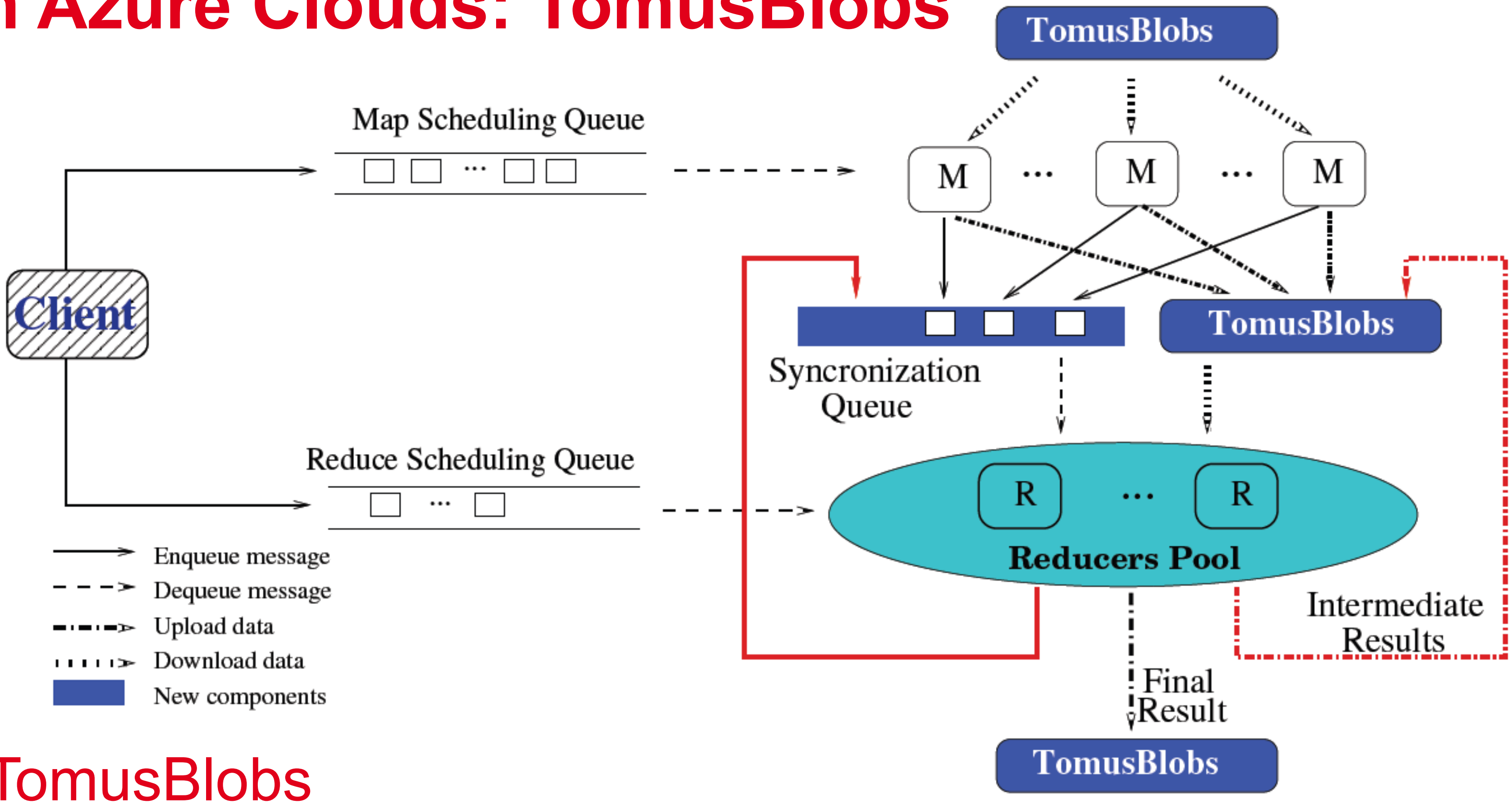
A	B	C	D	E	F	G	H
SNP rs	Chr	Volume (mm^3)	Corr. P-value	Cluster Volume (mm^3)	Corr. P-value	Minimum P-value	Corr. P-value
rs2788031	1	810	1	810	1	1,43E-006	1
rs2185077	1	648	1	540	1	3,18E-005	1
rs10888855	1	1053	1	1026	1	2,51E-005	1
rs10888857	1	2241	1	2241	0,9995	3,76E-007	1
rs3753009	1	2592	0,9888	2565	0,9651	1,90E-007	1
rs7551981	1	540	1	540	1	5,98E-005	1
rs4433442	1	783	1	783	1	5,69E-005	1
rs12755760	1	594	1	594	1	1,94E-005	1

MAIN ACHIVEMENTS
ON THE INFRASTRUCTURE SIDE

Data-intensive Processing on Clouds: Challenges

- Computation-to-data latency is high
- Need scalable concurrent data accesses to shared data
- Need efficient Map-Reduce-like data processing
  - Hadoop is not the best we can get
  - The Reduce phase may be costly

Scalable Storage for Processing Shared Data on Azure Clouds: TomusBlobs



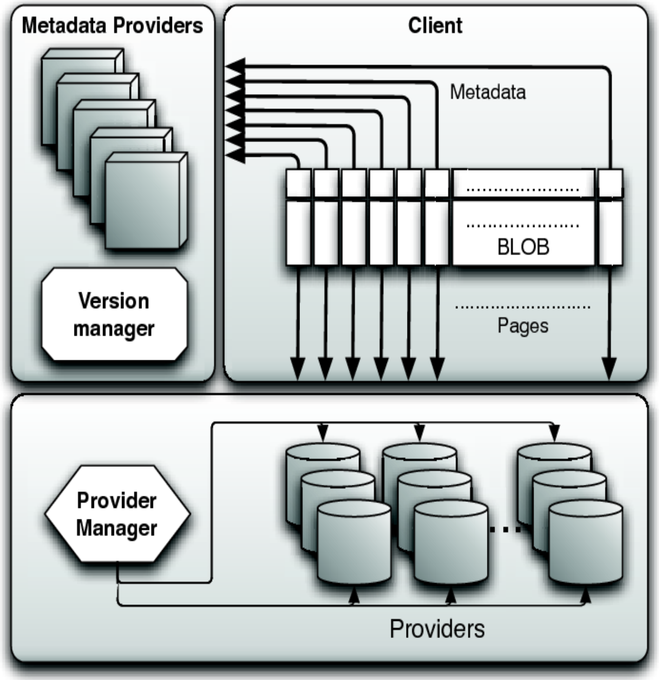
TomusBlobs

- Aggregates the virtual disks into a uniform storage service
- Relies on versioning to support high throughput under heavy concurrency
  - Leverages the BlobSeer data storage software (KerData)
- Transparent data chunk replication

Background: BlobSeer, a Software Platform for Scalable, Distributed BLOB Management

Started in 2008, 6 PhD theses (Gilles Kahn/SPECIF PhD Thesis Award in 2011)
Main goal: optimized for concurrent accesses under heavy concurrency

Three key ideas
Decentralized metadata management
Lock-free concurrent writes (enabled by versioning)
Write = create new version of the data
Data and metadata "patching" rather than updating



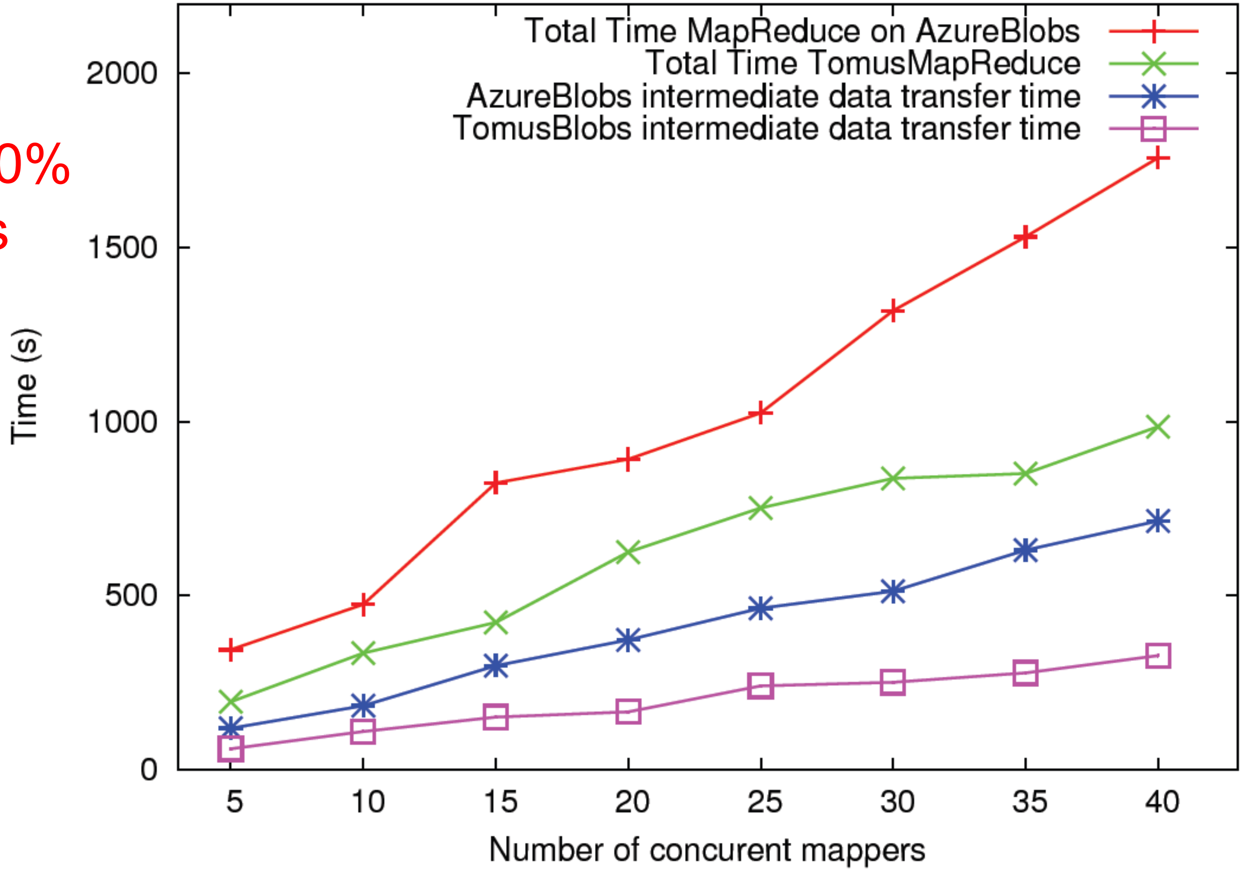
A back-end for higher-level data management systems
Short term: highly scalable distributed file systems
Middle term: storage for cloud services

Our approach
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/

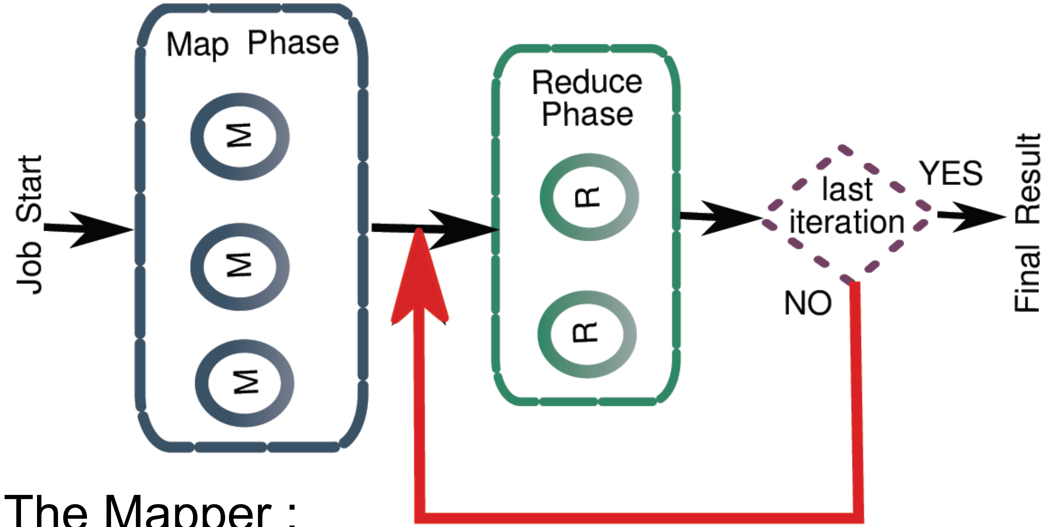
Using TomusBlobs for A-Brain: Results

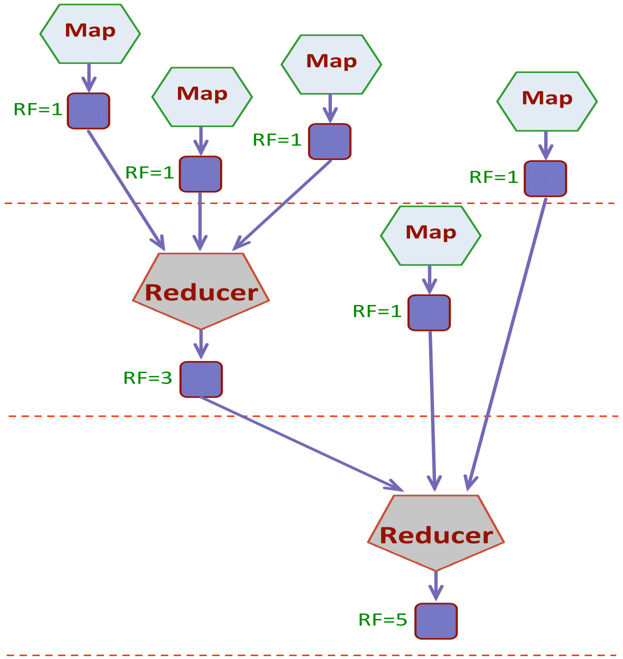
- Gain / Azure Blobs: ~50%
- Scalability: 1000 cores
- Demo available



http://www.irisa.fr/kerdata/doku.php?id=abrain

Extending the MapReduce Model: MapIterativeReduce





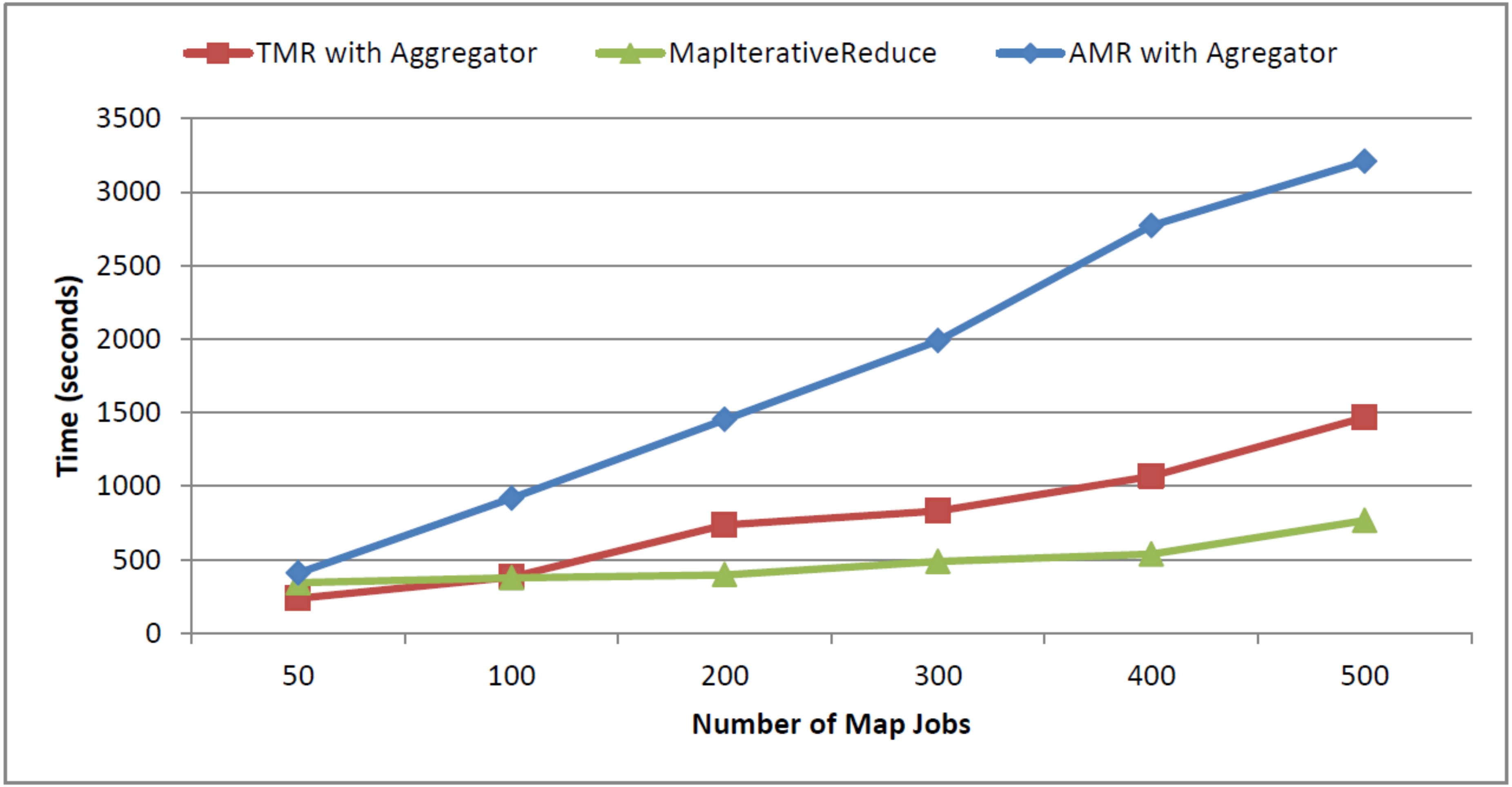
The Mapper :

- Classical map tasks

The Reducer

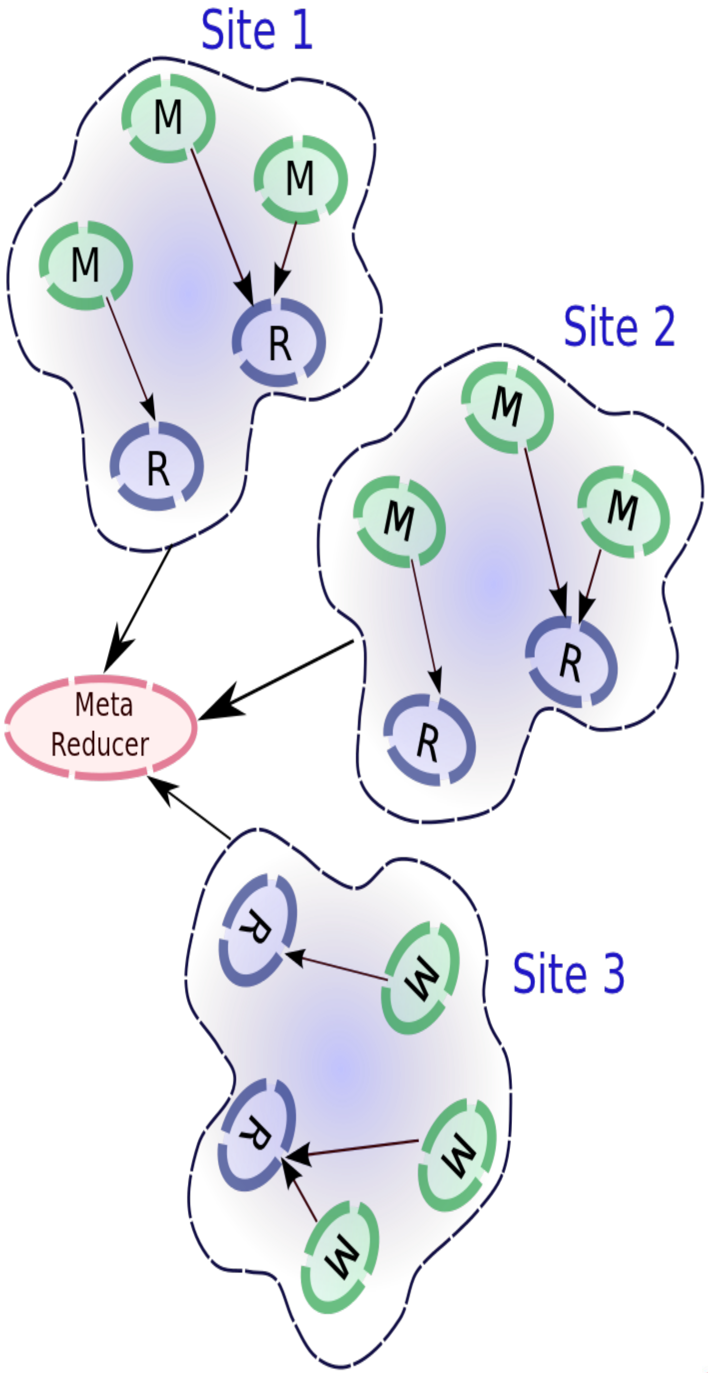
- Iterative reduction in two steps:
  - Receive the workload description from the Clients
  - Process intermediate results
- After each iteration, the termination condition is checked

Impact of MapIterativeReduce on A-Brain



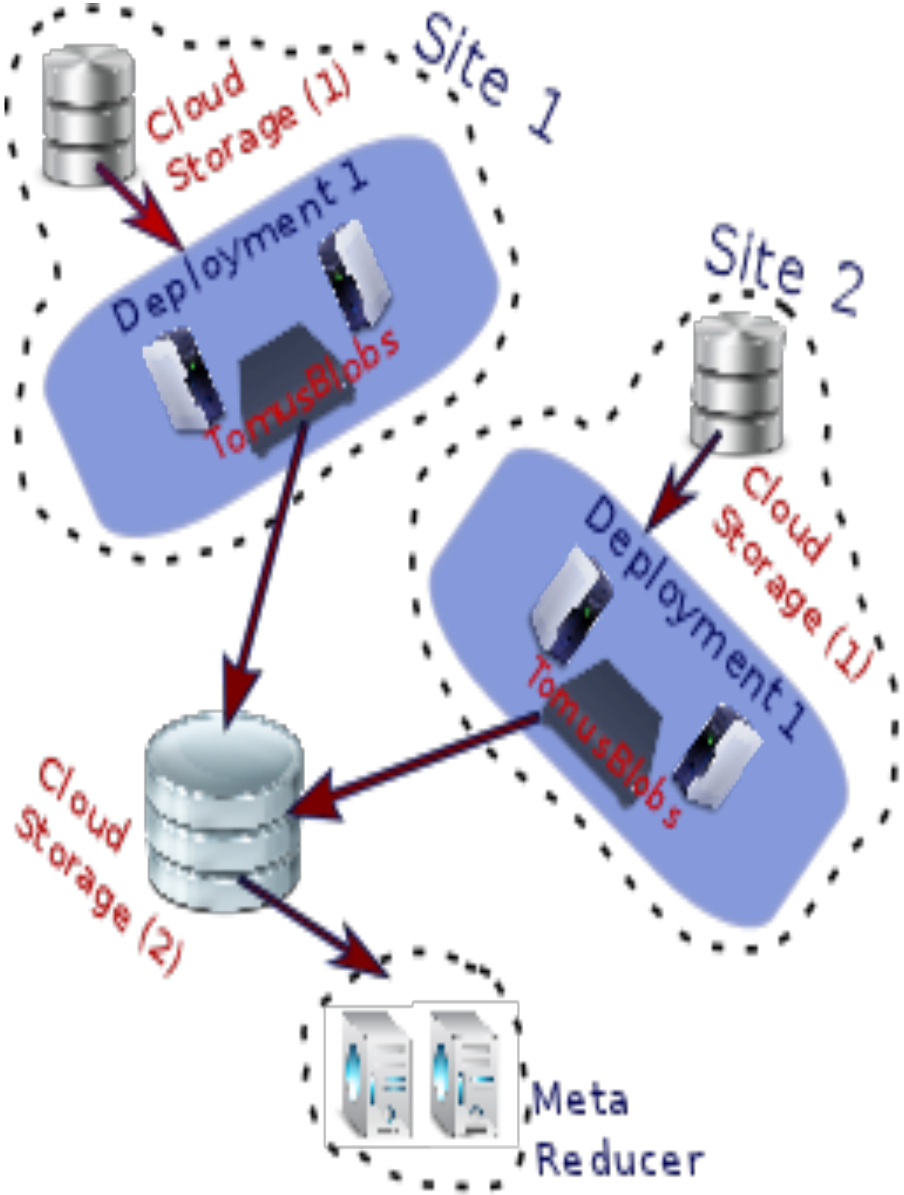
Beyond Single Site processing

- Scenario: data is produced in different locations or constrained (e.g. confidentiality)
- Problem: data movements across geo-distributed deployments are costly
- Goal: minimize the number of transfers and volumes of transferred data
- Constraint: single-site deployments work as independent services
- Approach: collaborative mechanism across datacenters to reach the common goal



Towards Geo-distributed TomusBlobs

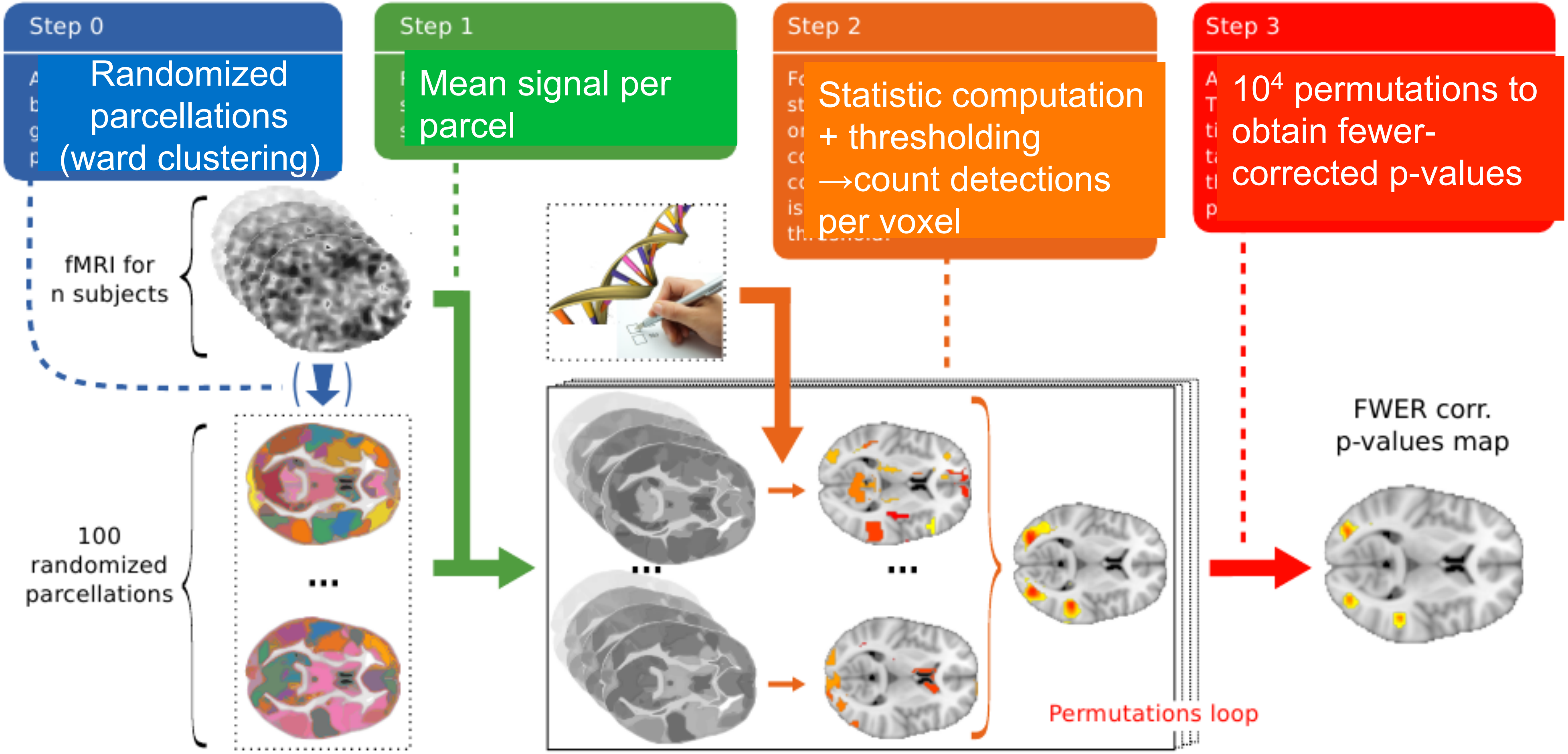
- TomusBlobs for intra-deployment data management
- Public Storage (Azure Blobs/ Queues) for inter-deployment communication
- Iterative Reduce technique for minimizing number of transfers (and data size)
- Balance the network bottleneck from single data center



MAIN ACHIVEMENTS ON THE APPLICATION SIDE

Contributions: RPBI – Improving Brain-wide Studies

Randomized-parcellation based inference



Step 0

Randomized parcellations (ward clustering)

Step 1

Mean signal per parcel

Step 2

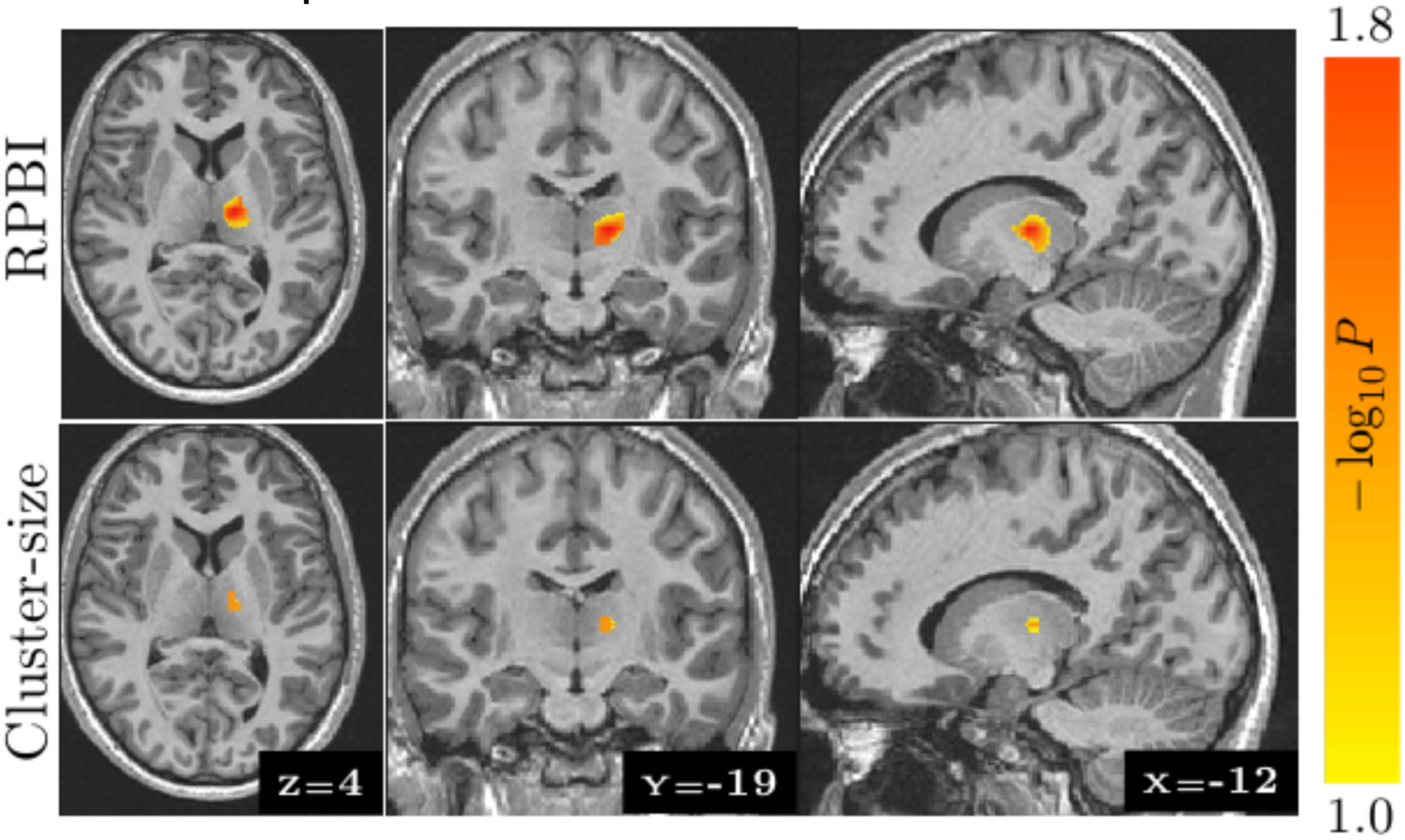
Statistic computation + thresholding →count detections per voxel

Step 3

10^4 permutations to obtain fewer-corrected p-values

Contributions: Results of RPBI

Experiment with a few SNPs of the ARVCF gene (close to COMT): fMRI signals upon motor response errors



RPBI uncovers a more significant association than traditional approaches

Contributions:
Improving Genome-wide Studies

Do not try to localize a few SNPs (among 10^6): rather assess the joint effect of all SNPs against brain variables (heritability)

- ➢ common variants are responsible of a large portion of heritability
- ➢ address the *missing variance* problem [Yang et al. Nat.gen. 2010]

Regress all the SNPs together against a given brain activation measure

FMRI signal in a subcortical region

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_1 + \boldsymbol{Z}\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

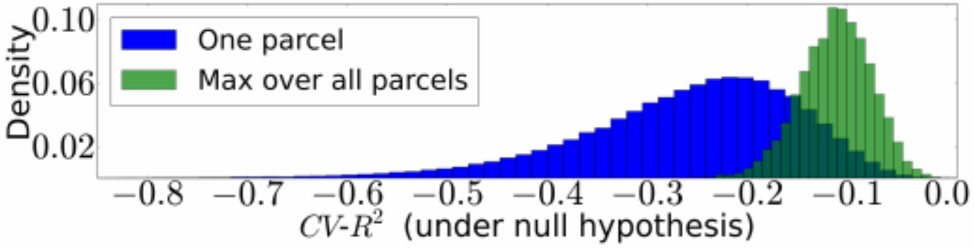
All SNPs

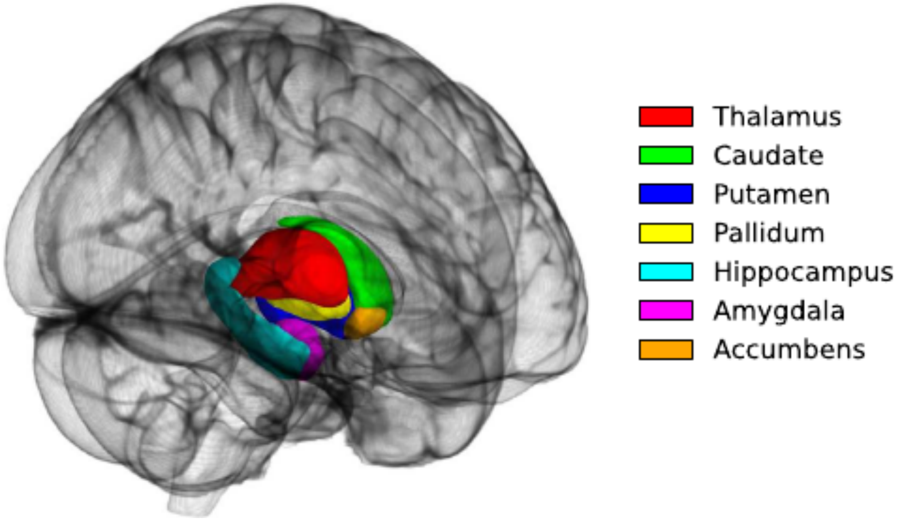
Other regressors (confounds)

[Da Mota et al., submitted to Frontiers in Neuroinformatics]

Contributions: Results with Heritability

ROI name		$CV\text{-}R^2$	fwe corr. p-value
Thalamus	left	$\mathbf{0.026}$	1.10^{-4}
	right	$\mathbf{0.038}$	1.10^{-4}
Caudate	left	$\mathbf{0.003}$	2.10^{-4}
	right	-0.012	3.10^{-4}
Putamen	left	$\mathbf{0.019}$	1.10^{-4}
	right	$\mathbf{0.006}$	2.10^{-4}
Pallidum	left	$\mathbf{0.018}$	1.10^{-4}
	right	-0.010	3.10^{-4}
Hippocampus	left	$\mathbf{0.010}$	2.10^{-4}
	right	$\mathbf{0.020}$	1.10^{-4}
Amygdala	left	$\mathbf{0.016}$	1.10^{-4}
	right	$\mathbf{0.015}$	1.10^{-4}
Accumbens	left	$\mathbf{0.022}$	1.10^{-4}
	right	-0.002	2.10^{-4}

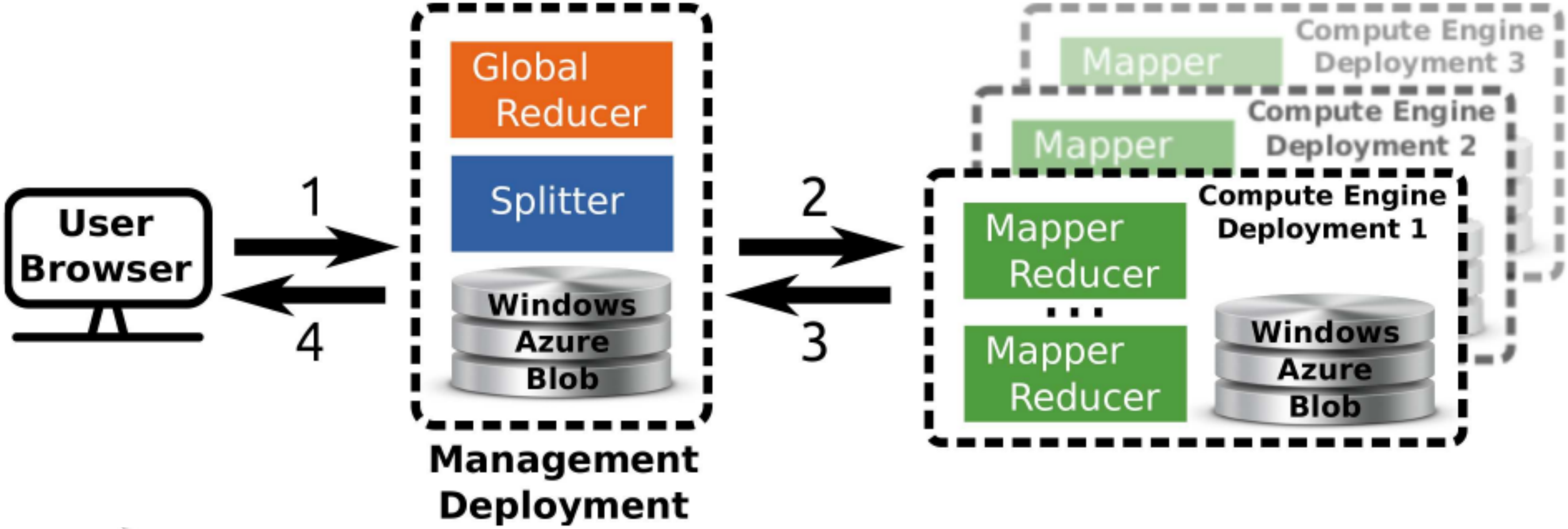




Experiment on the Imagen dataset: heritability of the stop failure brain activation signals in the sub-cortical nuclei: The signals are significantly more heritable than chance in all regions considered

What the Application Team Learned from A-Brain

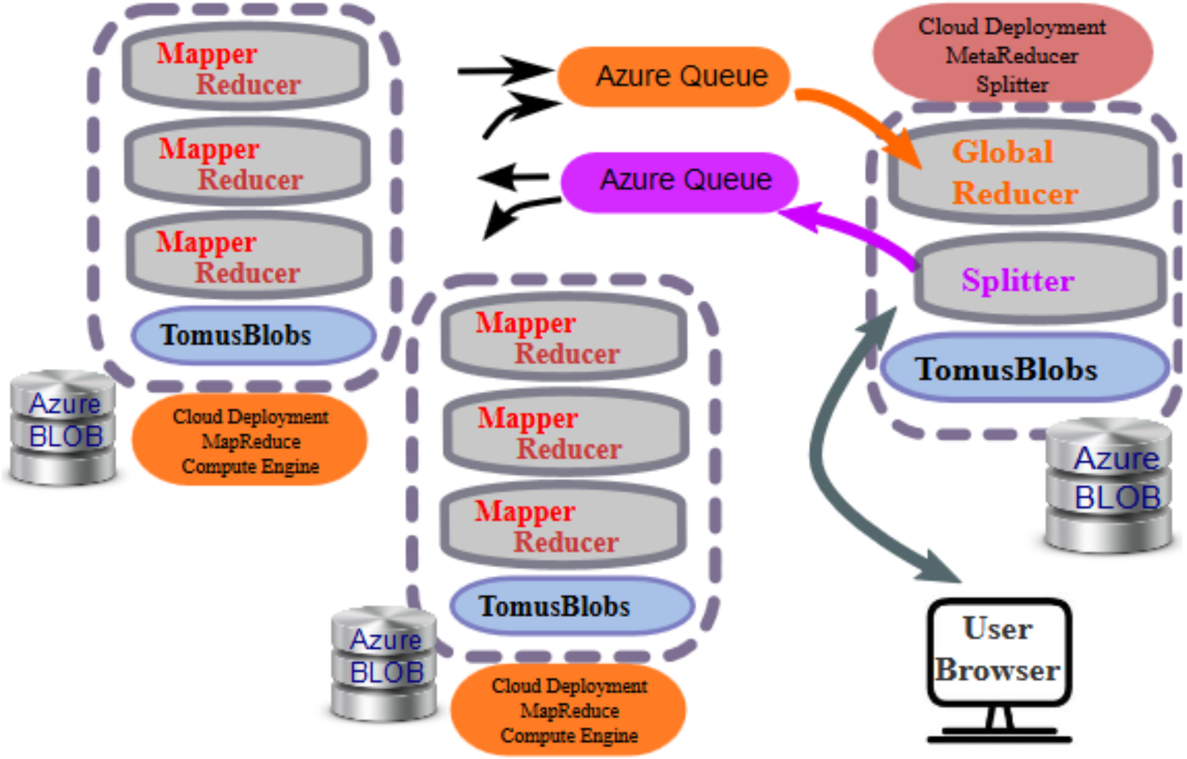
- Using the cloud can be advantageous:
  - Do not need to own the cluster
  - Resources rented until the end of the computation
  - Ease of use: execute the same code as the usual one
- Progress still needed to get closer to the power of a bare cluster



Our experience on Azure in the A-Brain project

- Experiments performed on 1000 cores
- Multi-site processing
- Data centers used: West US, North US, West EU, North EU
- Long running experiments:

  timespan for 1 experiment 1-2 days up to ~ of 15 days
- More than 300.000 hours of computation used



Application deployment times

- High deployment times: for each new or updated deployment on Azure, the fabric controller prepares the nodes
- Bigger problems reported for Amazon EC2:

  "The most common failure is an inability to acquire all of the virtual machine images you requested because insufficient resources are available. When attempting to allocate 80 cores at once, this happens fairly frequently. "

  Keith R. Jackson, Lavanya Ramakrishnan, Karl J. Runge, and Rollin C. Thomas. 2010. Seeking supernovae in the clouds: a performance study. In Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing (HPDC '10).

- The deployment time was reduced after the major update from November 2012
- Can still be a problem for real time/near real time scaling

Experience with VM failures

Regular failures

- The general knowledge about the cloud:
  - Commodity hardware will generate many failures
  - Fault tolerance mechanism for failures: watch dog, checkpointing, replication
- Only a very small fraction of the machines failed even during the course of very long running executions
  - During the 2 weeks experiments on several hundreds of nodes, only 3 machines failed (fail-stop-restart).

Experience with VM failures

Regular failures

- The general knowledge about the cloud:
  - Commodity hardware will generate many failures
  - Fault tolerance mechanism for failures: watch dog, checkpointing, replication
- Only a very small fraction of the machines failed even during the course of very long running executions
  - During the 2 weeks experiments on several hundreds of nodes, only 3 machines failed (fail-stop-restart).

Bad luck

During the 3 years experience in Azure 2 exceptional outages happened:

- 29 February 2012 – Demo in Saclay for Tony Hey
  - Azure down for leap year certificate problem
- 22 February 2013 – running the Big A-Brain Experiment
  - Azure fell down due to a failure in a security certificate

- Solution: Do not run experiments in February! ;-)

A-Brain: Two Things to Take Away

- The TomusBlobs data-storage layer developed within the A-Brain project was demonstrated to scale up to 1000 cores on 3 Azure data centers.
  - It exhibits improvements in execution time close to 50% compared to standard solutions based on Azure BLOB storage.

- The consortium has provided the first statistical evidence of the heritability of functional signals in a failed stop task in basal ganglia, using a ridge regression approach, while relying on the Azure cloud to address the computational burden.

Publications

Journals

- Alexandru Costan, Radu Tudoran, Gabriel Antoniu, Goetz Brasche. TomusBlobs : Scalable Data-intensive Processing on Azure Clouds. **Concurrency and Computation Practice and Experience**, Wiley, 2013. URL: http://onlinelibrary.wiley.com/doi/10.1002/cpe.3034/abstract.

- Benoit Da Mota, Virgile Fritscha, Gaël Varoquaux, Tobias Banaschewski, Gareth J. Barker , Arun L.W. Bokde, Uli Bromberg , Patricia Conrod, Jürgen Gallinat, Hugh Garavan, Jean-Luc Martinot, Frauke Nees, Tomas Pausl, Zdenka Pausova , Marcella Rietschel, Michael N. Smolka, Andreas Ströhle, Vincent Frouin, Jean-Baptiste Poline, Bertrand Thirion, the IMAGEN consortium. Randomized Parcellation Based Inference. **NeuroImage**, Elsevier, in Press.

- Benoit Da Mota, Radu Tudoran, Alexandru Costan, Gael Varoquaux, Goetz Brasche, Patricia Conrod, Herve Lemaitre, Tomas Paus, Marcella Rietschel, Vincent Frouin, Jean-Baptiste Poline, Gabriel Antoniu, Bertrand Thirion and the IMAGEN Consortium. Machine Learning Patterns for Neuroimaging-Genetic Studies in the Cloud. Submitted to **Frontiers in Neuroinformatics**.

Electronic Journals

- Gabriel Antoniu, Alexandru Costan, Benoit Da Mota, Bertrand Thirion, Radu Tudoran. A-Brain: Using the Cloud to Understand the Impact of Genetic Variability on the Brain. ERCIM News, April 2012.

Publications

Conferences and workshops (2013)

- Radu Tudoran, Alexandru Costan, Ramin Rezai Rad, Goetz Brasche and Gabriel Antoniu. Adaptive File Management for Scientific Workflows on the Azure Cloud. IEEE International Conference on Big Data **(IEEE BigData 2013**), October 6-9, 2013, Santa Clara, CA, USA. **Acceptance rate: 17%**.

- Radu Tudoran, Alexandru Costan, Gabriel Antoniu. DataSteward : Using Dedicated Compute Nodes for Scalable Data Management on Public Clouds. In Proc. of ISPA 2013- 11th IEEE International Symposium on Parallel and Distributed Processing with Applications (**IEEE ISPA 2013**), Melbourne, Australia, July 2013.

- Benoit da Mota, Virgile Fritsch, Gaël Varoquaux, Vincent Frouin, Jean-Baptiste Poline, and Bertrand Thirion. Distributed High-Dimensional Regression with Shared Memory for Neuroimaging-Genetic Studies. in **Euroscipy 2013**.

- Benoit Da Mota, Virgile Fritsch, Gaël Varoquaux, Vincent Frouin, Jean-Baptiste Poline, and Bertrand Thirion. Enhancing the Reproducibility of Group Analysis with Randomized Brain Parcellations. In **MICCAI** - 16th International Conference on Medical Image Computing and Computer Assisted Intervention - 2013, Nagoya, Japan, June 2013.

- Virgile Fritsch, Benoit Da Mota, Gaël Varoquaux, Vincent Frouin, Eva Loth, Jean-Baptiste Poline and Bertrand Thirion. Robust Group-Level Inference in Neuroimaging Genetic Studies. In Pattern Recognition in **Neuroimaging**, Philadelphie, United States, May 2013.

Publications

Conferences and workshops (2012)

- Radu Tudoran, Alexandru Costan, Gabriel Antoniu, Hakan Soncu. “TomusBlobs: Towards Communication-Efficient Storage for MapReduce Applications in Azure.” In Proc. 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (**CCGrid'2012**), May 2012, Ottawa, Canada.

- Radu Tudoran, Alexandru Costan, Gabriel Antoniu, Luc Bougé. A Performance Evaluation of Azure and Nimbus Clouds for Scientific Applications. In Proc. **CloudCP** 2012 - 2nd International Workshop on Cloud Computing Platforms, Held in conjunction with the ACM SIGOPS Eurosys 12 conference, Apr 2012, Bern, Switzerland.

- Radu Tudoran, Alexandru Costan, Benoit Da Mota, Gabriel Antoniu, Bertrand Thirion. A-Brain: Using the Cloud to Understand the Impact of Genetic Variability on the Brain. 2012 **Cloud Futures** Workshop, Berkeley, May 2012.

- Radu Tudoran, Alexandru Costan, Gabriel Antoniu. MapIterativeReduce: A Framework for Reduction-Intensive Data Processing on Azure Clouds. Third International Workshop on MapReduce and its Applications (**MAPREDUCE'12**), held in conjunction with ACM HPDC'12., Jun 2012, Delft, Netherlands.

- Benoit Da Mota, Vincent Frouin, Edouard Duchesnay, Soizic Laguitton, Gaël Varoquaux, Jean-Baptiste Poline, Bertrand Thirion. A fast computational framework for genome-wide association studies with neuroimaging data. 20th International Conference on Computational Statistics (**COMPSTAT 2012**), Aug 2012, Lamissol, Cyprus.

- Benoit Da Mota, Michael Eickenberg, Soizic Laguittton, Vincent Frouin, Gaël Varoquaux, Jean-Baptiste Poline, Bertrand Thirion. A MapReduce Approach for Ridge Regression in Neuroimaging-Genetic Studies. Data- and Compute-Intensive Clinical and Translational Imaging Applications Workshop **(DCICTIA-MICCAI'12**), held in conjunction with the 15th International Conference on Medical Image Computing and Computer Assisted Intervention, Oct 2012, Nice, France.

People Involved

Gabriel Antoniu (INRIA, Project Lead)

Bertrand Thirion (INRIA, Project Lead)

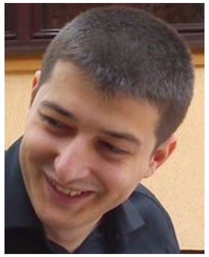
Pierre Louis Xech (Microsoft)

Götz-Philip Brasche (Microsoft Research)

Benoit Da Mota (INRIA)

Hakan Soncu (Microsoft Research)

Alexandru Costan (INRIA)



Radu Tudoran (INRIA)

WHAT'S NEXT?

Z-CLOUDFLOW: DATA-INTENSIVE WORKFLOWS IN THE CLOUD (2013-2016)

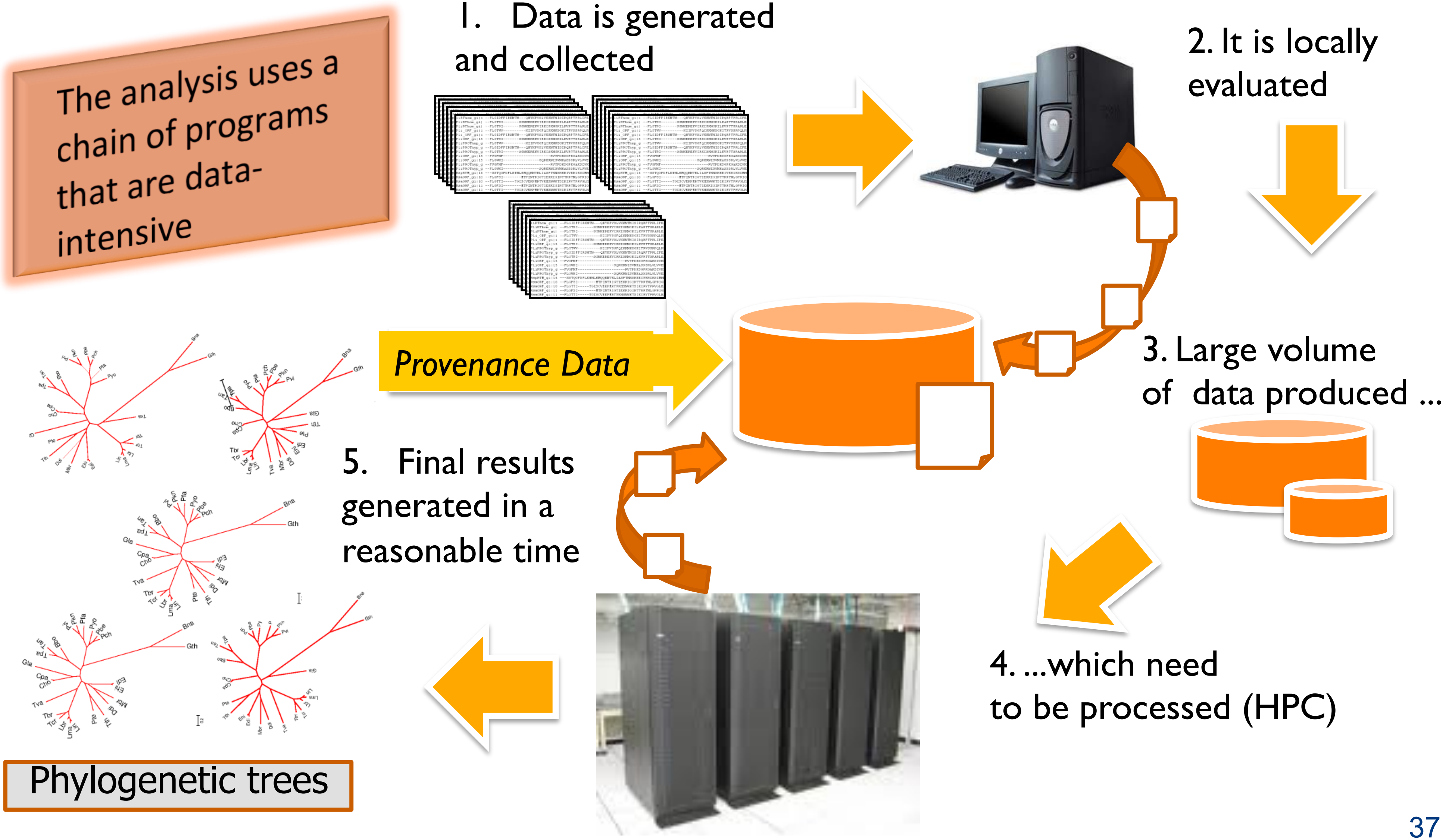
Joint project

Inria-Microsoft Research Center

KerData

Zenith

Scientific Workflow Scenario



The analysis uses a chain of programs that are data-intensive

1. Data is generated and collected

2. It is locally evaluated

3. Large volume of data produced ...

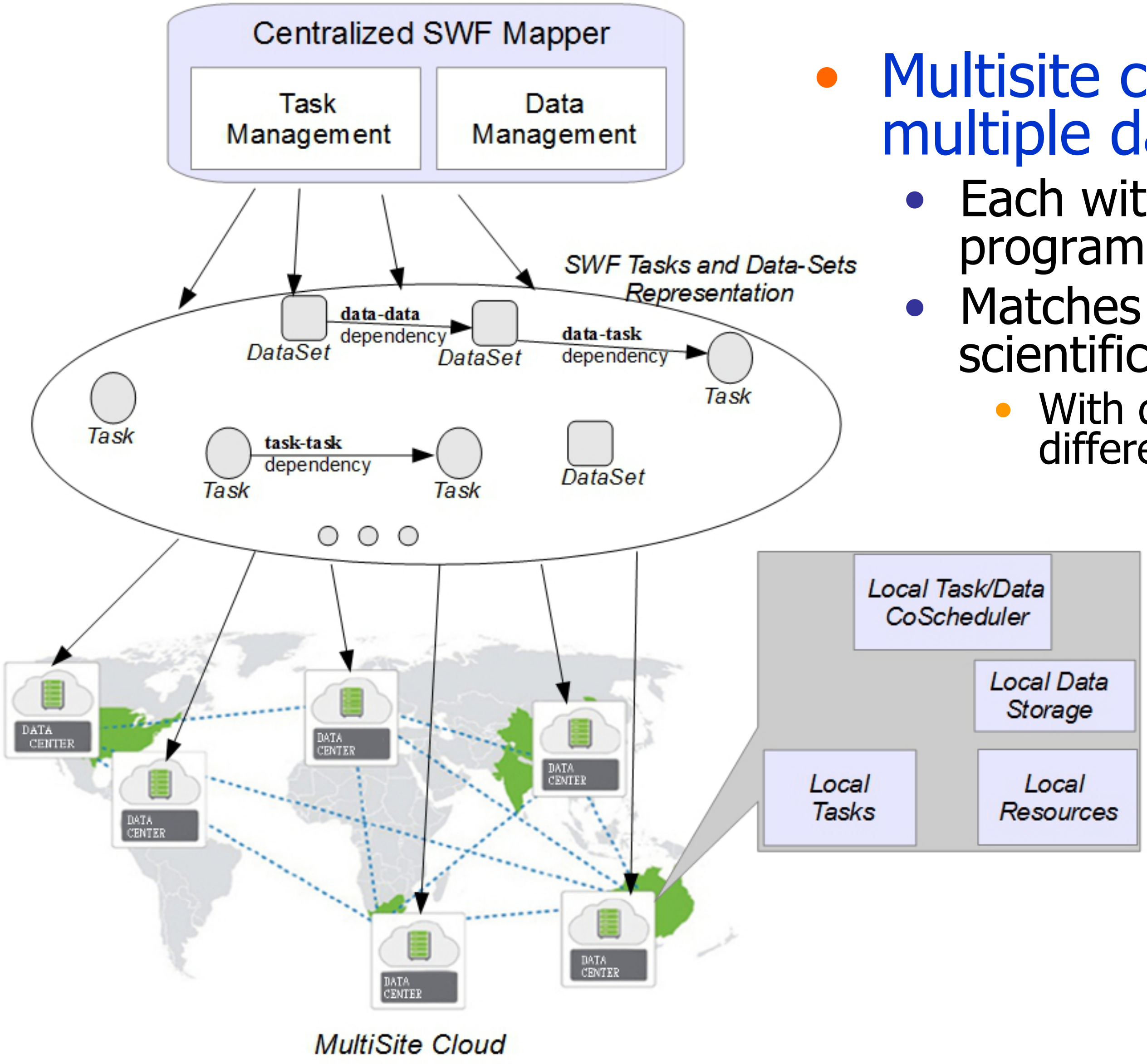
4. ...which need to be processed (HPC)

5. Final results generated in a reasonable time

Provenance Data

Phylogenetic trees

Why to Use Multi-site Clouds for Workflows?



- Multisite cloud = a cloud with multiple data centers
  - Each with its own cluster, data and programs
  - Matches well the requirements of scientific apps
    - With different labs and groups at different sites

Open Issues for SWfMS in the Cloud

- Adaptive scheduling in heterogeneous, dynamic infrastructures
  - Strong variations of performance because of shared resources
- Automatic optimization and parallelization
  - As with our workflow algebra
- Exploiting data provenance at runtime to deal with dynamic workflows
  - Workflows that react to external events such as human interaction and dynamic steering

MultiSite Cloud Data Management: Challenges

- What strategies to use and how for efficient data transfers?

- How to group tasks and datasets together to minimize data transfers?

- How to do workload balancing between datacenters to avoid bottlenecks?

Our Approach to Go Further…

- Adapt workflow processing to the multisite cloud environment
  - Exploit cloud capabilities
- Adopt an algebraic approach for specifying workflows
  - Eases parallelization, optimization and scheduling
- Process workflow execution plans efficiently by optimizing data transfers during execution
  - Rely on an efficient distributed storage layer
- Build an appropriate cloud storage framework addressing the challenges of multisite clouds

CENTRE DE RECHERCHE COMMUN INRIA MICROSOFT RESEARCH

A-Brain and Z-CloudFlow: Scalable Data Processing on Azure Clouds

Lessons Learned in 3 Years
and Future Directions

INVENTEURS DU MONDE NUMÉRIQUE

10th JLPC Workshop, NCSA, Urbana
25 November 2013

Thank you!

MultiSite Cloud Data Management: Challenges

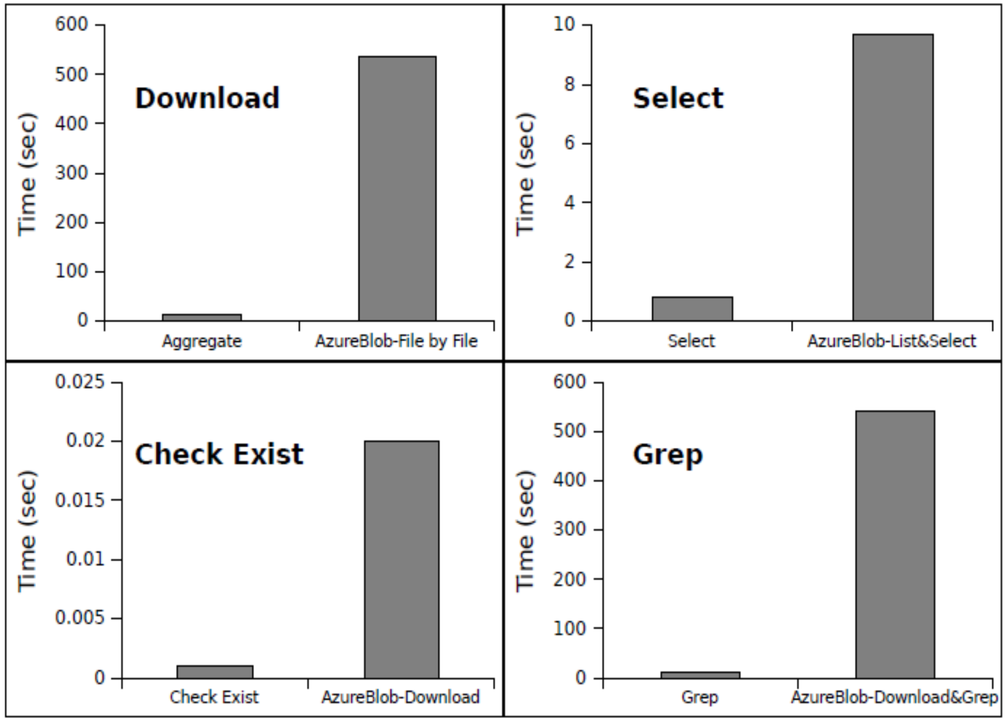
- What strategies to use and how for efficient data transfers?
  - Using monitoring-based performance modelling predict the best combination of protocols (e.g. memory-to- memory, FTP, BitTorrent) and transfer parameters (e.g. flow count, multicast enhancement, replication degree) to maximize throughput or minimize costs
- How to group tasks and datasets together to minimize data transfers?
  - Utilizing dependencies among datasets and tasks, enhance data locality through efficient data and task co-scheduling strategies
- How to do workload balancing between datacenters to avoid bottlenecks?
  - Cope with latency and performance variability due to multi-tenancy

Needed: Monitoring Services for BigData

Variable performance

- Mainly due to multi-tenancy
- Need to predict the behavior of the underlying network and end-systems, in order to optimize the transfers over federated datacenters and partition the computation

Cloud introspection as a service



Monitoring API

- Monitoring and logging services for BigData
- Current cloud storage APIs do not support even simple operations

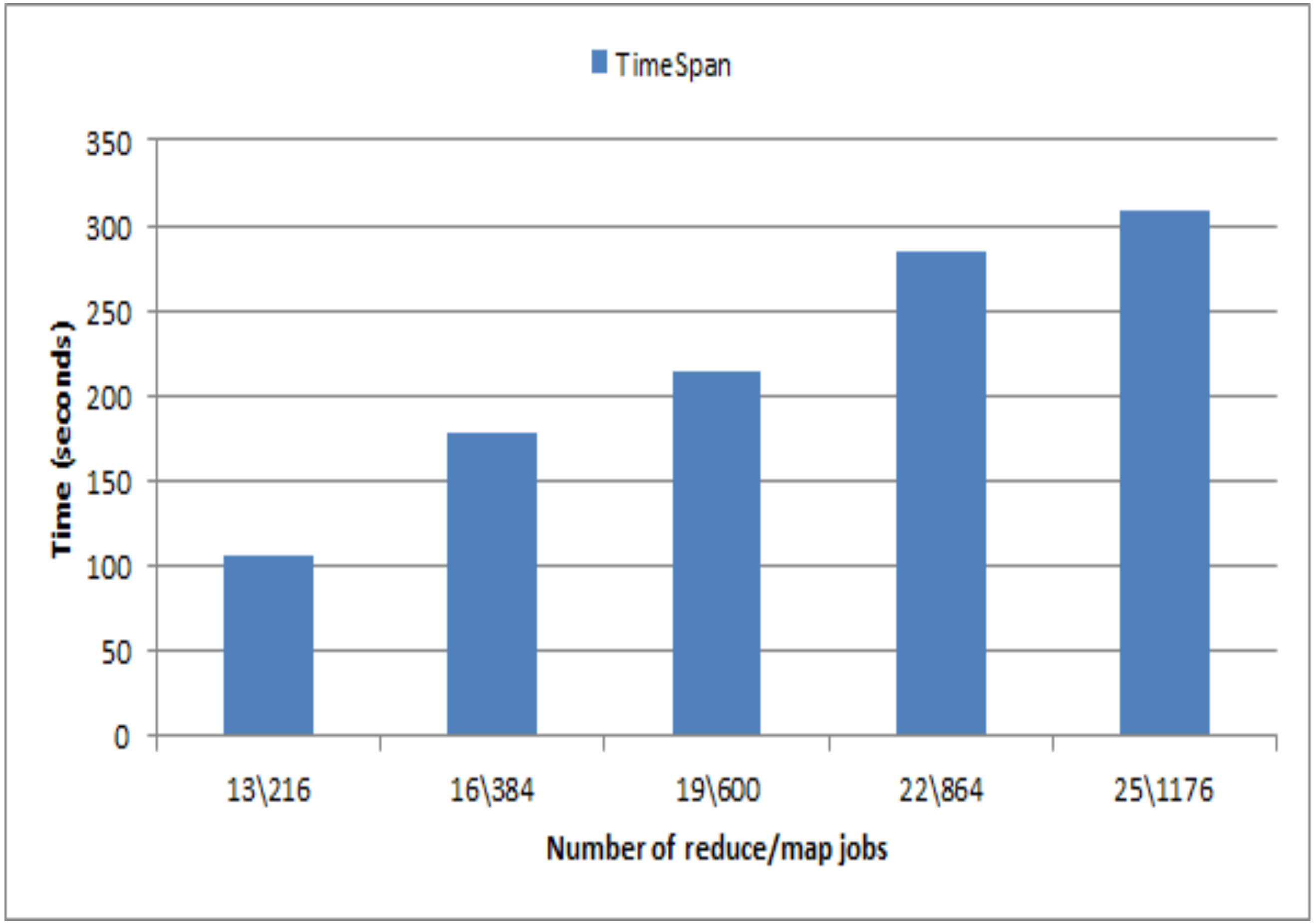
on multiple files/blobs (e.g. grep, select/filter, compress, aggregate)

Towards a scientific Big Data processing toolkit?

Application: Current Status

- Good method for brain-wide association RPBI
- Genome-wide associations: build on the ridge-based heritability estimate
  - Analysis at the level of pathways, genes
  - Robust version of ridge regression ?
- Progress still needed
  - Not enough data!
  - Need more precise hypotheses to test

Multi-Site MapReduce



- 3 deployments (NE,WE,NUS)
- 1000 CPUs
- ABrain execution across multiple sites