High-Performance Big Data Management Across Cloud Data Centers

Radu Tudoran

PhD Advisors
Gabriel Antoniu INRIA
Luc Bougé ENS Rennes

KerData research team
IRISA/INRIA Rennes
Doctoral Work: Context

Exponential Growth

130 exabytes

20,000 exabytes

2005 2020

BIG DATA

VOLUME

VARIETY

VELOCITY

VERACITY

VALUE

Cloud Computing

Data Center

Internet Data Center

2012 Report
Geographically-Distributed Processing

- CERN ATLAS
  - PB of data distributed for storage across multiple institutions

- Ocean Observatory
  - Data sources located in geographically distant regions

- Large IT Web-Services
  - Data processing exceeds site limits
Doctoral Work in a Nutshell

Data management for MapReduce & Workflow

Scaling the processing

Multi-site MapReduce

Enabling scientific discovery

Enabling large-scale scientific processing

Data management across sites

High-Performance Big Data Management Across Cloud Data Centers

Transfer-as-a-Service

Cloud-provided Transfers Service

High-performance streaming

Dedicating compute nodes for storage

Optimize intersite transfers

Streaming across cloud sites

Configurable cost-performance tradeoffs
Doctoral Work in a Nutshell

Multi-site MapReduce

Enabling scientific discovery

Enabling large-scale scientific processing

Data management across sites

High-Performance Big Data Management Across Cloud Data Centers

Scaling the processing

Data management for MapReduce & Workflow

Transfer-as-a-Service

Cloud-provided Transfers Service

High-performance streaming

Optimize intersite transfers

Streaming across cloud sites

Dedicating compute nodes for storage

Configurable cost-performance tradeoffs
A Big Data Case Study: The A-Brain Application

Value: find the correlation between brain markers and genetic data in order to understand the behavioral variability and diseases

Subject 1
Subject 2
... Subject n

Correlations?

Genetic data \( \rightarrow \) dimension \( n_{\text{snps}} \times n_{\text{subjects}} \)

Image data \( \rightarrow \) dimension \( n_{\text{voxels}} \times n_{\text{subjects}} \)

Variety
Multi-modal joint analysis
Data contains outliers from acquisition process

Veracity
Biologically significant results and false detection control requires \( 10^4 \) permutations

Volume
\[ n_{\text{voxels}} = 10^6 \]
\[ n_{\text{snps}} = 10^6 \]
\[ n_{\text{subjects}} = 10^3 \]
Data space potentially reaches TB to PB level

Velocity
...not the case ...
But other examples are coming in a few slides
Data Management on Public Clouds

Cloud-provided storage service

Cloud
Compute Nodes

How about data locality?
Our approach: TomusBlobs

- Collocate computation and data in PaaS clouds by federating the virtual disk of compute nodes
- Self-configuration, automatic deployment and scaling of the data management system
- Apply to MapReduce and Workflow processing
Leveraging TomusBlobs for MapReduce Processing

- New MapReduce prototype (no Hadoop at that point on Azure)
- Adopt BlobSeer as storage backend
Scenario: 100 nodes deployment on Azure
Comparison with an Azure Blobs based MapReduce
TomusBlobs is 3x-4x faster than the cloud remote storage
Beyond MapReduce: Map-IterativeReduce

- Unique result with parallel reduction
- No central control entity
- No synchronization barrier
The Efficiency of Full-Reduction

The Most Frequent Words benchmark

A-Brain initial experimentation

- **Experimental Setup:** 200 nodes deployment on Azure
- **Map-IterativeReduce** reduces the execution timespan to half
TomusBlobs for Workflow Processing

Exploit workflow specificities:
- Data access patterns
- File manipulation
- Batch processing
TomusBlobs for Workflow Processing

- **Multiple transfer** solutions: FTP, In-Memory, BitTorrent
- Adapt the transfer to the **data access pattern**
- Adaptive **replication** strategies for higher performance
- Integration with Microsoft Generic Worker
**Workflow Processing on Cloud**

**Experimental Setup:** 100 Azure nodes, Generic Worker engine

**TomusBlobs** adaptively chooses each time the best strategy

---

**Synthetic workflow**

**BLAST scientific workflow**

---

- **DirectLink Transfers**
- **Torrent Transfers**
- **TomusBlobs**
- **AzureBlobs**
Multi-site MapReduce

Enabling scientific discovery

Enabling large-scale scientific processing

Data management across sites

High-Performance Big Data Management Across Cloud Data Centers

Scaling the processing

Data management for MapReduce & Workflow
- Adaptive data management for collocated storage
- A MapReduce framework

High-performance streaming

Optimize intersite transfers

Configurable cost-performance tradeoffs

Dedicating compute nodes for storage

Cloud-provided Transfers Service

Streaming across cloud sites
Subjects ~2000
voxels $\sim 10^{5-6}$
SNPs $\sim 10^6$

Find Correlations
Single-Site Computation on the Cloud

Timespan estimation for single core machine: 5.3 years

Parallelize and execute on Azure cloud across 350 cores using TomusBlobs

Achievements:
- Reduced execution time to 5.6 days
- Demonstrated that univariate analysis is too sensitive to outliers

New challenge: adopt a more robust analysis which increases the computation requirements to 86 years (for single core)
Going Geo-distributed

- Hierarchical multi-site MapReduce: Map-IterativeReduce, Global Reduce
- Data management: TomusBlobs (intra-site), Cloud Storage (inter-site)
- Iterative-Reduce technique for minimizing transfers of partial results
- Balance the network bottleneck from single data center
Executing the A-Brain Application at Large-Scale

- Multi-site processing: East US, North US, North EU Azure Data Centers
- Experiments performed on 1000 cores
- Experiment duration: ~ 14 days
- More than 210,000 hours of computation used
- Cost of the experiments: 20000 euros (VM price, storage, outbound traffic)
- 28000 map jobs (each lasting about 2 hours) and ~600 reduce jobs
- Data transfers more than 1 TB

**Scientific Discovery:**
Provided the first statistical evidence of the heritability of functional signals in a failed stop task in basal ganglia
Doctoral Work in a Nutshell

Data management for MapReduce & Workflow
- Adaptive data management for collocated storage
- A MapReduce framework

Multi-site MapReduce
- Map-IterativeReduce
- Hierarchical two-tier MapReduce cloud framework

Enabling scientific discovery

Enabling large-scale scientific processing
- 2-week 1000 CPUs multi-site experiment
- Pioneering discoveries for brain-genes correlations

Scaling the processing

High-Performance Big Data Management Across Cloud Data Centers

Transfer-as-a-Service

High-performance streaming

Optimize inter-site transfers

Streaming across cloud sites

Cloud-provided Transfers Service

Dedicating compute nodes for storage

Configurable cost-performance tradeoffs

21
To Cloud or Not to Cloud Data?

Limitations:
- No (or weak) SLA guarantees
- High-latency and low throughput transfer
Addressing the SLA Issues for Inter-Site Transfers

Three design principles:
- Environment awareness: model the cloud performance
- Real-time adaptation for data transfers
- Cost effectiveness: maximize throughput or minimize costs
Sampling method
– Estimate the cloud performance based on the monitoring trace

Average and variability estimated for each metric:
- Updated based on weights given to fresh samples: from 0 (no trust) to 1 (full trust)

- Predictive transfers: express transfer time and cost
- Dynamically adjust the transfer quotas across routes
Leverage network parallelism:
Aggregate inter-site bandwidth through multi-path transfers

Addressing Inter-Site Transfer Performance:
Multi-Path Transfers
Addressing Inter-Site Transfer Performance: Multi-Hop Transfers

Further increase network parallelism:
Avoid network throttling by considering alternative routes through other data centers
How much are you willing to pay for performance?
How much is it actually worth paying?
Comparing to Existing Solutions

- **Experimental setup**: up to 10 nodes, Azure Cloud
- Transfers between North Central US to North EU Azure data centers
Doctoral Work in a Nutshell

Data management for MapReduce & Workflow
- Adaptive data management for collocated storage
- A MapReduce framework

Multi-site MapReduce
- Map-IterativeReduce
- Hierarchical two-tier MapReduce cloud framework

Enabling scientific discovery

Enabling large-scale scientific processing
- 2-week 1000 CPUs multi-site experiment
- Pioneering discoveries for brain-genes correlations

High-Performance Big Data Management Across Cloud Data Centers

Transfer-as-a-Service

Cloud-provided Transfers Service

Scaling the processing

High-performance streaming

Optimize inter-site transfers
- Multi-route transfers for aggregating inter-site bandwidth
- Cost-Performance transfer optimization

Streaming across cloud sites

Dedicating compute nodes for storage

Configurable cost-performance tradeoffs
To Stream or Not to Stream?

**Average Compute Rate per VM**

*Graph showing the average compute rate per VM with data points for VMs ranging from 1 to 32.*

**Average Data Rate per VM**

*Graph showing the average data rate per VM with data points for VMs ranging from 1 to 32.*

- **Stream&Compute**
  - Stream Processing Application
  - VM
  - Cloud Data Center

- **Copy&Compute**
  - Stream Source
  - Cloud Data Center
  - VM Data Node
  - Persistent Attached Storage

**Logos and Badges**

- Argonne National Laboratory
- NIMBUS
- openstack
- Future Grid
- Windows Azure
Towards Dynamic Batch-based Streaming

Latency (L) modeled based on stream context: event size, throughput, arrival rate, routes, serialization/de-serialization technology, batch size
JetStream for MonALISA

- 1.1 million events; North US to North EU Azure data centers
- Automatically resource optimization
- Optimizing the latency and transfer rate tradeoff

Independent event streaming takes 30,900 seconds compared to 80 seconds for JetStream
Variable Streaming Rates

**Elastic scaling of the resource based on load**

**Environment-aware → self-optimization**
Doctoral Work in a Nutshell

- **Multi-site MapReduce**
  - Map-IterativeReduce
  - Hierarchical two-tier MapReduce cloud framework

- **Enabling scientific discovery**

- **Enabling large-scale scientific processing**
  - 2-week 1000 CPUs multi-site experiment
  - Pioneering discoveries for brain-genome correlations

**High-Performance Big Data Management Across Cloud Data Centers**

- **Data management for MapReduce & Workflow**
  - Adaptive data management for collocated storage
  - A MapReduce framework

- **Scaling the processing**

- **Transfer-as-a-Service**

- **High-performance streaming**
  - Evaluate the performance of cloud streaming
  - Context-aware latency modeling

- **Optimize inter-site transfers**
  - Multi-route transfers for aggregating inter-site bandwidth
  - Cost-Performance transfer optimization

- **Streaming across cloud sites**

- **Cloud-provided Transfers Service**

- **Dedicating compute nodes for storage**

- **Configurable cost-performance tradeoffs**
Transfer Options on the Cloud

The Default Option

- No (or minimal) configuration
- High-latency and low throughput transfer
- Fixed price scheme

Multi-path transfers

- Aggregate inter-site throughput
- Fixed price scheme
- Managed, configured and administrated by users
How About a Transfer as a Service?

**Asymmetric cloud service**
- Federated clouds
- No transparent communication optimizations

**Symmetric cloud service**
- Same cloud vendor
- Allows any number of communication optimizations
Is TaaS Feasible Performance-wise?

Multi-tenant service usage: performance degradations of 20% ... while the number of service nodes per app is decreased from 5:1 to 1:1
Scenario: Transfer large volumes of data across Azure sites

Cost: Cost margins for the service usage can be defined based on performance
Data transfer market: Flexible and dynamic pricing

=> **win-win** situation for cloud vendor and users

**Why?** Decrease price => to reduce idle bandwidth

Increase price => to decrease network congestion
Conclusions & Perspectives
Doctoral Work in a Nutshell

**Data management for MapReduce Workflow**
- Adaptive data management for collocated storage
- A MapReduce framework

**Scaling the processing**

**Multi-site MapReduce**
- Map-IterativeReduce
- Hierarchical two-tier MapReduce cloud framework

**Enabling scientific discovery**

**Enabling large-scale scientific processing**
- 2-week 1000 CPUs multi-site experiment
- Pioneering discoveries for brain-genes correlations

**High-Performance Big Data Management Across Cloud Data Centers**

**Data management across sites**

**Cloud-provided Transfers Service**
- Study architectural options for a transfer service
- Provide a flexible pricing schema

**Transfer-as-a-Service**
- Evaluate the performance of cloud streaming
- Context-aware latency modeling

**High-performance streaming**
- Multi-route transfers for aggregating inter-site bandwidth
- Cost-Performance transfer optimization

**Optimize inter-site transfers**
- High-performance streaming

**Streaming across cloud sites**

**Dedicating compute nodes for storage**
- Advanced data-related services
- Topology nodes layout discovery and storage node placement

**Configurable cost-performance tradeoffs**
Achievements

Publications

• 1 Book Chapter

• 3 Journal articles
  – Frontiers in Neuroinformatics 2014
  – Concurrency and Computation Practice and Experience 2013
  – ERCIM Electronic Journal 2012

• 7 International Conferences publications
  – 3 papers at IEEE/ACM CCGrid 2012 and 2014 (Cloud Cluster and Grid, rank A), Acceptance rates: 26%, 19%
  – IEEE SRDS 2014 (Symposium on Reliable Distributed Systems, rank A)
  – IEEE Big Data 2013, Acceptance rate 17%
  – ACM DEBS 2014 (Distributed Event Based Systems), Acceptance rate 9%
  – IEEE Trustcom/ISPA 2013 (rank A)

• 7 Workshops papers, Posters and Demos
  – MapReduce in conjunction with ACM HPDC (rank A)
  – CloudCP in conjunction with ACM EuroSys (rank A)
  – IPDPSW in conjunction with IEEE IPDPS (rank A)
  – Microsoft: CloudFutures, ResearchNext, PhD Summer School
  – DEBS Demo in conjunction with ACM DEBS

Software

TomusBlobs

PaaS data management middleware
• Available with Microsoft GenericWorker
MapReduce engine for the Azure cloud
• Cloud service for bio-informatics

Cloud Benchmark Service

SaaS for benchmarking the performance of data stage-in to cloud data centers
• Available on Azure Cloud

JetStream

Middleware for batch-based, high-performance streaming across cloud sites
• Binding with Microsoft StreamInsight

External Collaborators

• Microsoft Research ATLE, Cambridge
• Argonne National Laboratory
• Inria Saclay
• Inria Sophia Antipolis

43
Perspectives

• Multi-site workflow across geographically distributed sites
  Workflow data access patterns, self-* processing, cost/performance tradeoffs

• Cloud stream processing
  Management of many small events, latency constraints for distributed queries

• Diversification of the cloud data management ecosystem
  X-as-a-Service, uniform storage across sites, API for task orchestration

One size does not fit all!
Doctoral Work in a Nutshell

Multi-site MapReduce
- Map-IterativeReduce
- Hierarchical two-tier MapReduce cloud framework

Enabling scientific discovery

Enabling large-scale scientific processing
- 2-week 1000 CPUs multi-site experiment
- Pioneering discoveries for brain-genes correlations

Data management across sites

Data management for MapReduce & Workflow
- Adaptive data management for collocated storage
- A MapReduce framework

High-performance streaming
- Evaluate the performance of cloud streaming
- Context-aware latency modeling

Cloud-provided Transfers Service

Transfer-as-a-Service
- Study architectural options for a transfer service
- Provide a flexible pricing schema

Streaming across cloud sites

Optimize inter-site transfers
- Multi-route transfers for aggregating inter-site bandwidth
- Cost-Performance transfer optimization

Dedicating compute nodes for storage
- Advanced data-related services
- Topology nodes layout discovery and storage node placement

Configurable cost-performance tradeoffs

Scaling the processing
Backup slides
From few-large DCs to many small DCs

**Multi-site processing**
- Integrated MapReduce processes across sites
- Workflow orchestration
- Site cross-scheduling of tasks

**Multi-site data management**
- Uniform storage across data centers
- High-performance transfer tools – Transfer as a Service
- Usage and data access patterns

[Links]
Service Diversification

Handling Big Data grows in complexity
- Architectural design options for many items storage
- Enriched and diversified data-oriented services
- Smart replication strategies

Diversification of processing
- Customizable-user API: towards business workflows
- Solutions for providing the versatility of workflows and simplicity of MapReduce
Deployment start times

- For each new or updated deployment on Azure, the fabric controller prepares the nodes → High deployment times (though better after the update from Nov. ‘12)

- 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.”


Scheduling mechanisms for efficient data access.
Lessons learned: running BigData applications

A real need for advanced data management functionality for running scientific Big Data processing in the clouds

- Monitoring API
  - Monitoring and logging services for Big Data
  - Current cloud storage APIs do not support even simple operations on multiple files/blobs (e.g. grep, select/filter, compress, aggregate)

- Data management for geo-distributed processing
  - Cloud storage delivers poor performances → High performance alternatives
  - Inter-site data transfer is not supported → Transfer as a service
How much data can I transfer using 25 VMs for 10 minutes?

**Experimental setup:** up to 25 nodes, Azure Cloud
- Transfers between North Central US to North EU Azure data centers
Is it feasible performance-wise?

**Multi-tenancy**

**Impact of CPU Load on I/O**

**Service Access Concurrency**: performance degradations of ~20% when reducing the service nodes per application from 5:1 to 1:1

**CPU load on user transfer nodes**: performance degradation up to 40%
Doctoral Work in a Nutshell

Scaling the processing

MapReduce Workflow
- Adaptive data management for collocated storage & computation
- A MapReduce framework

Multi-site MapReduce
- Map-IterativeReduce
- Hierarchical two-tier MapReduce cloud framework

Enabling scientific discovery

Enabling large-scale scientific processing
- 2-week 1000 CPUs multi-site experiment
- Pioneering discoveries for brain-genes correlations

High-Performance Big Data Management Across Cloud Data Centers

Transfer-as-a-Service
- Study architectural options for a transfer service
- Provide a flexible pricing schema

Cloud-provided Transfers Service

High-performance streaming
- Evaluate the performance of cloud streaming
- Context-aware latency modeling

Optimize intersite transfers
- Multi-route transfers for aggregating inter-site bandwidth
- Cost-Performance transfer optimization

Streaming across cloud sites

Dedicating compute nodes for storage
- Advanced data-related services
- Topology nodes layout discovery and storage-node placement

Configurable cost-performance tradeoffs

Data management across sites
Beyond the Put/Get data management systems:
What is the good option to build advanced data management functionality?

What if some of the compute nodes become data nodes?
Design Principles

- Dedicate compute nodes for managing and storing data
- Topology awareness
- No modification to the cloud middleware
• Discover the virtualized topology → **Clustering approach**
  - Throughput measurements between VMs
  - Asserting the performance
• Maximize throughput between application nodes and storage nodes
Assessing the storage throughput

- **Scenario**: Cumulative throughput
- **Experimental setup**: 50 client nodes, 50 storage nodes
- **Transfer improvement** due to CPU and network management