INVENTORS FOR THE DIGITAL WORLD

Big Data Management in the Clouds and HPC Systems

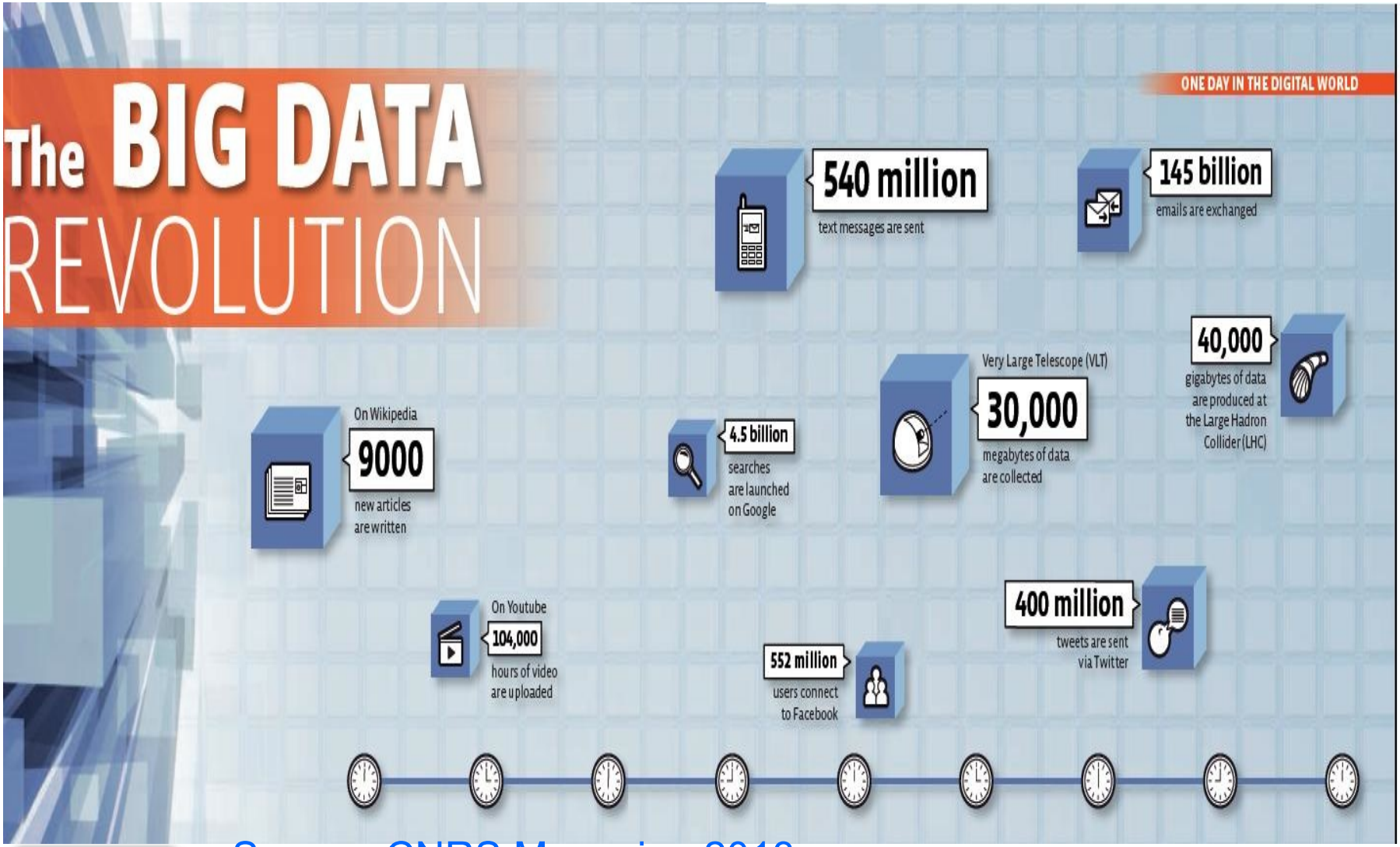
Hemera Final Evaluation

Paris 17th December 2014

Shadi Ibrahim

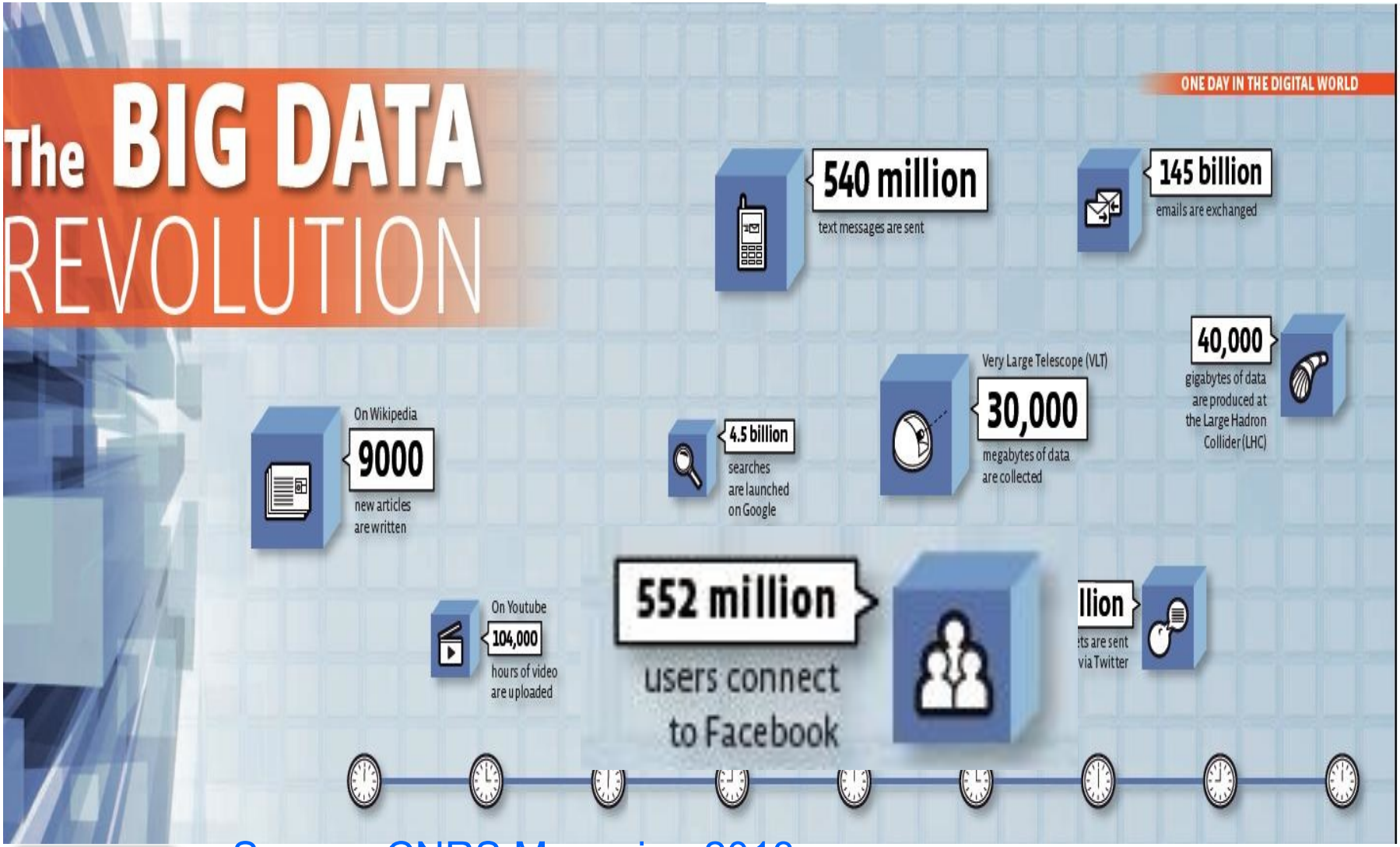
Shadi.ibrahim@inria.fr

Era of Big Data!



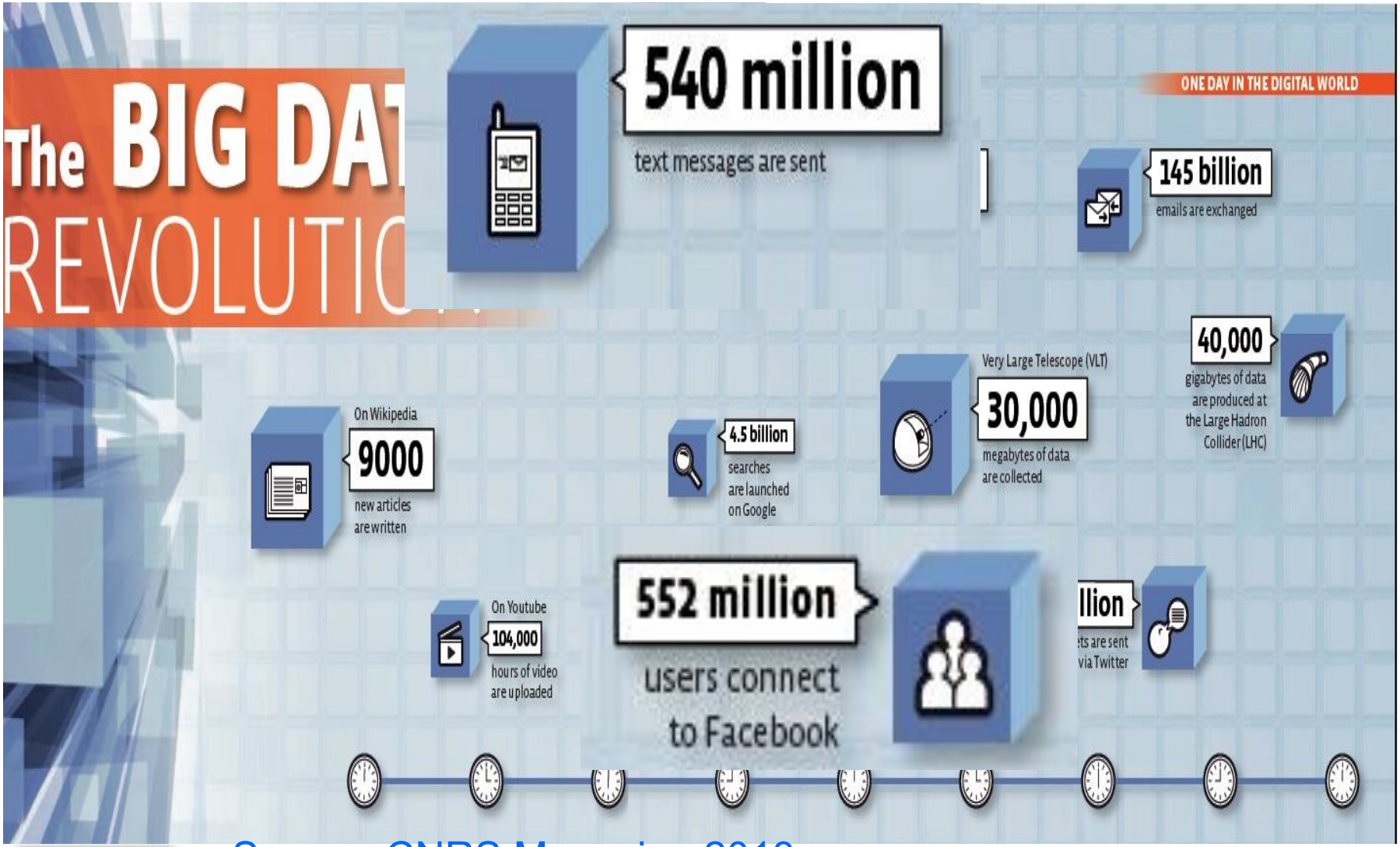
Source: CNRS Magazine 2013

Era of Big Data!



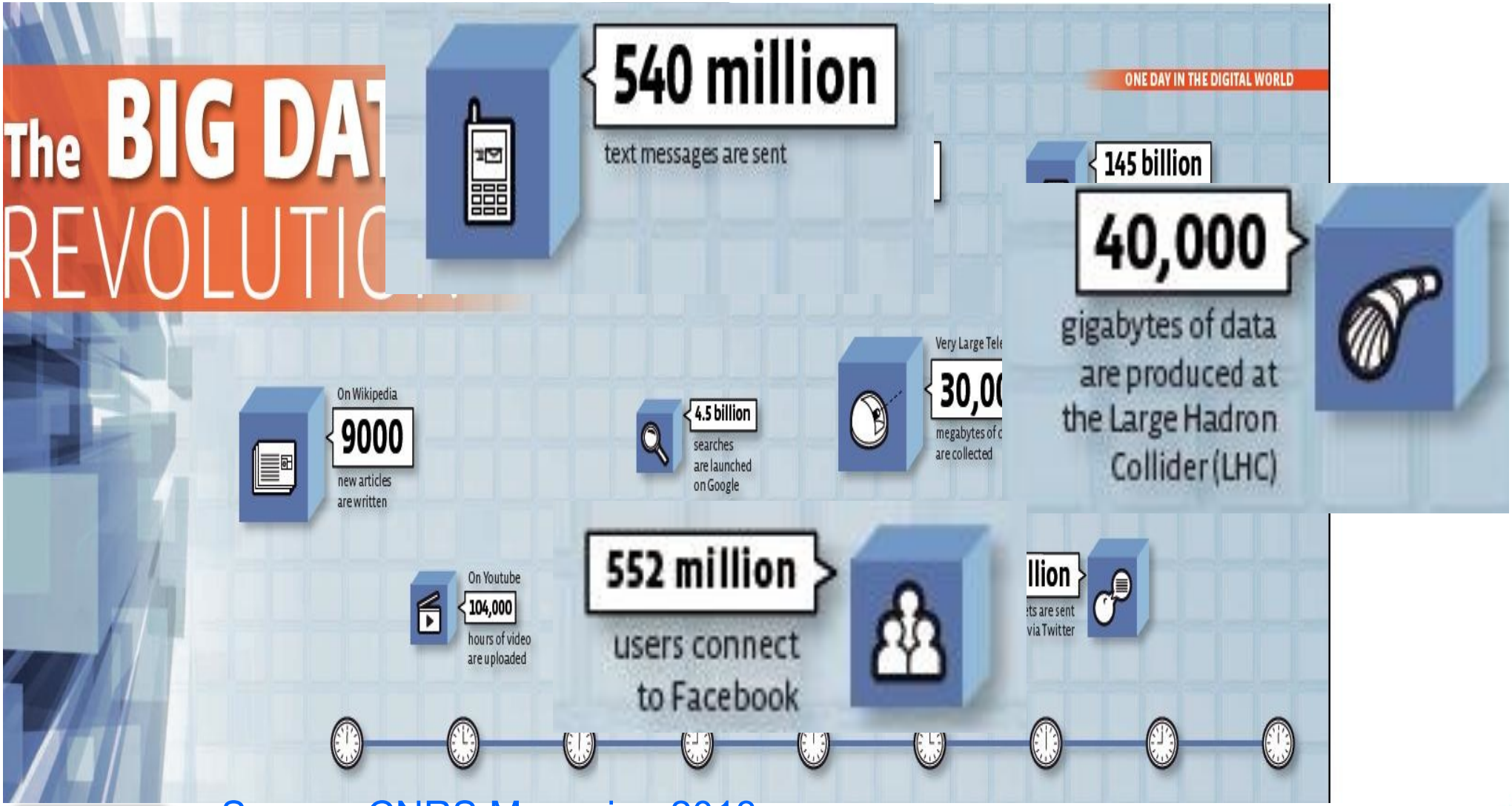
Source: CNRS Magazine 2013

Era of Big Data!



Source: CNRS Magazine 2013

Era of Big Data!



Source: CNRS Magazine 2013

KerData's Core Research Challenges

Applications

- Massive data analysis: clouds (e.g. MapReduce)
- Post-Petascale HPC simulations: supercomputers

Focus 1: Scalable big data management on IaaS and PaaS clouds

- *Challenge : understand how to reconcile performance, scalability, security and quality of service according to the requirements of data-intensive applications*

Focus 2: Scalable big data management on Post-Petascale HPC systems

- *Challenge: go beyond the limitations of current file-based approaches*

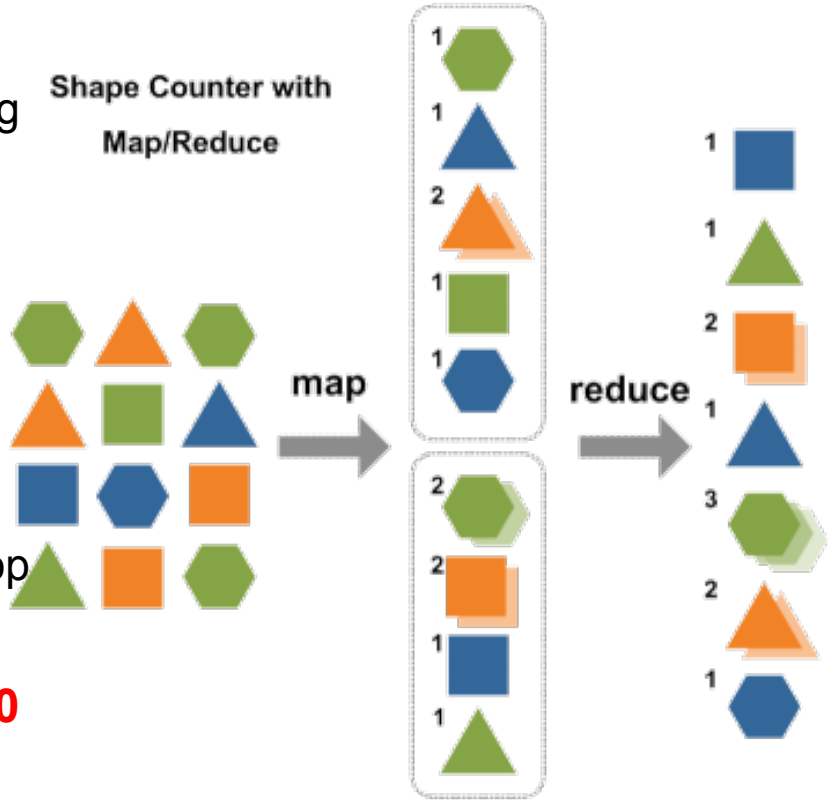
INVENTORS FOR THE DIGITAL WORLD

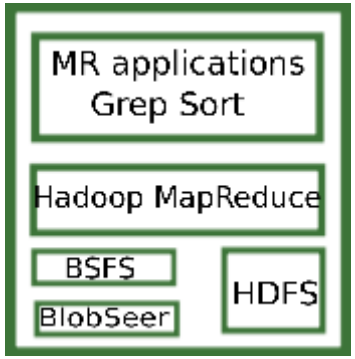
Focus 1:

Scalable Big Data Management in IaaS and Paas Clouds

Scalable Map-Reduce Processing

- ANR Project Map-Reduce (ARPEGE, 2010-2014)
- Partners: Inria (teams : KerData - leader, AVALON, Grand Large), Argonne National Lab, UIUC, JLPC, IBM, IBCP
  - Goal: High-performance Map-Reduce processing through concurrency-optimized data processing
  - URL: mapreduce.inria.fr
- **Some results**
  - Versioning-based concurrency management for increased data throughput (BlobSeer approach)
  - Efficient intermediate data storage in pipelines
  - Substantial improvements with respect to Hadoop
  - Application to efficient VM deployment
- **Intensive, long-run experiments done on Grid'5000**
  - Up to 300 nodes/500 cores
  - Plans: joint deployment G5K+FutureGrid (USA)
- **Papers**: JPDC, Concurrency and Computation Practice and Experience, ACM HPDC 2011 (AR:12.9%), ACM HPDC 2012, IEEE/ ACM CCGRID 2012, 2013, 2014, Euro-Par 2012, IEEE IPDPS 2013

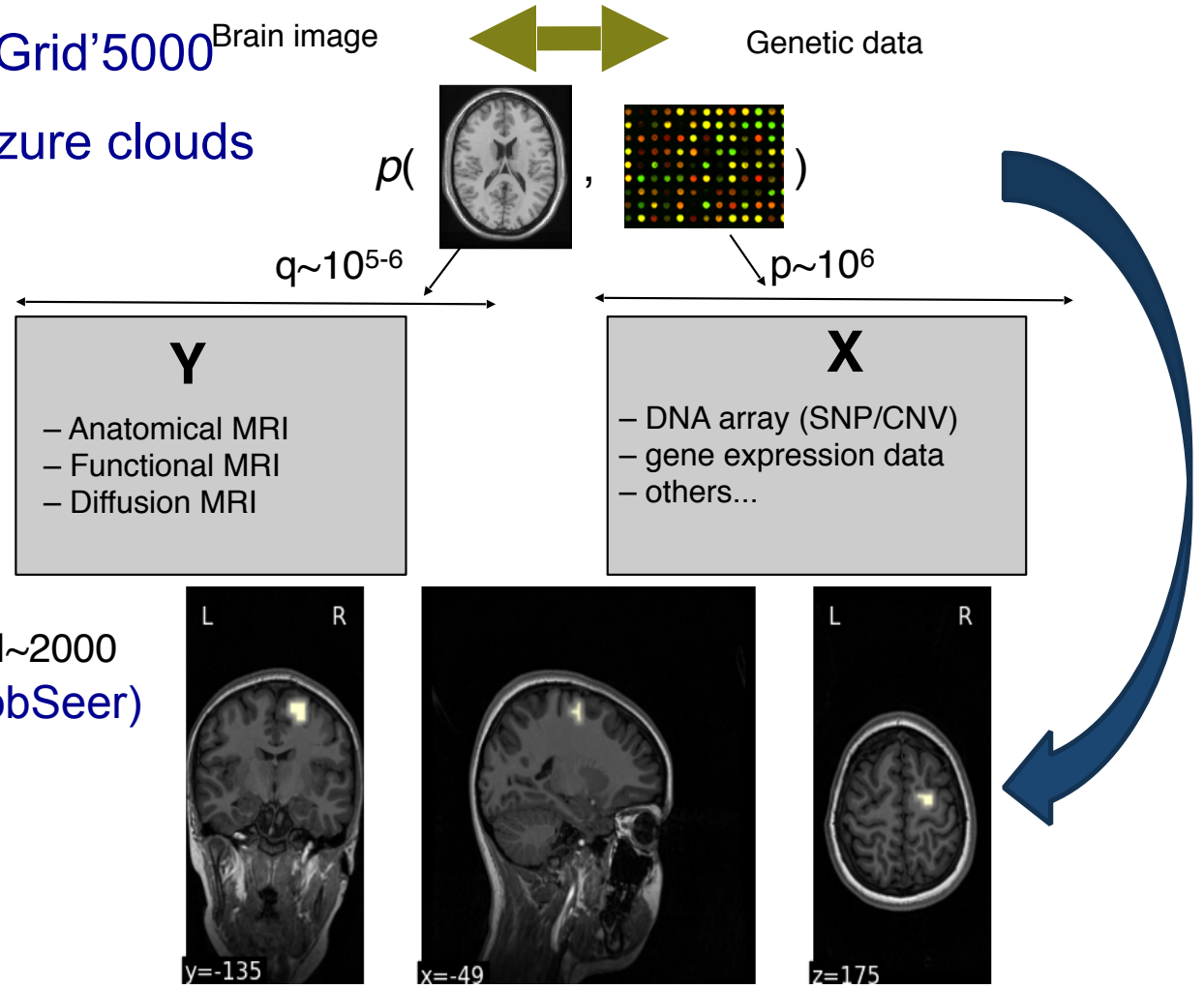




Impact: Transfer to Commercial Clouds
The A-Brain Microsoft Research – Inria Project

- Approach
  1. Preliminary experiments on Grid'5000
  2. Apply the approach to MS Azure clouds

- Partners
  - KerData, PARIETAL (Inria)
  - Microsoft ATL Europe

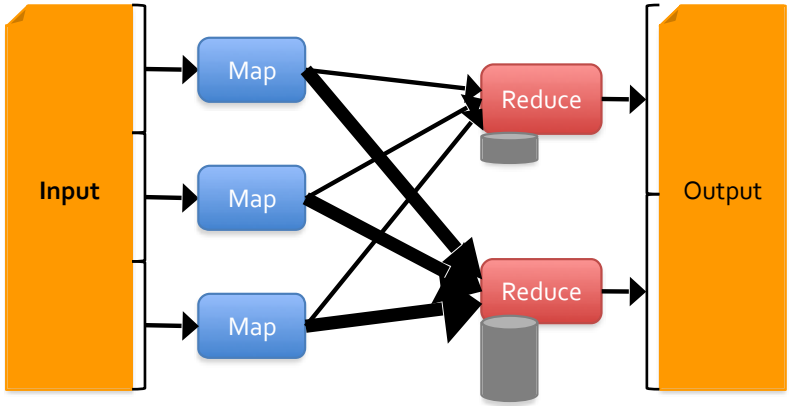


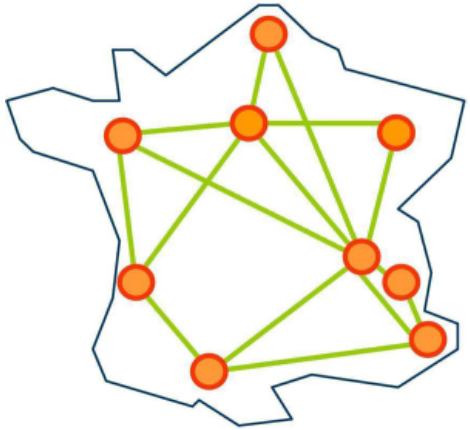
- TomusBlobs software (based on BlobSeer)
- **Gain / Blobs Azure : 45%**
- **Scalability : 1000 cores**
- Demo available!

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

Exposing Data Locality in MapReduce

- Data locality is crucial for Hadoop's performance
- Map scheduler ignores the state of the replication
  - e.g., 58% of Facebook's jobs achieve only 5% node locality and 59% rack locality (Eurosys 2010)
- Maestro: replica-aware map scheduling
- 35% performance improvement





100 nodes

"Maestro: Replica-aware map scheduling for mapreduce"
S Ibrahim, H Jin, L Lu, B He, G Antoniu, S Wu. CCGrid 2012

Energy-Efficient Big Data Processing

Why Energy efficient Hadoop?

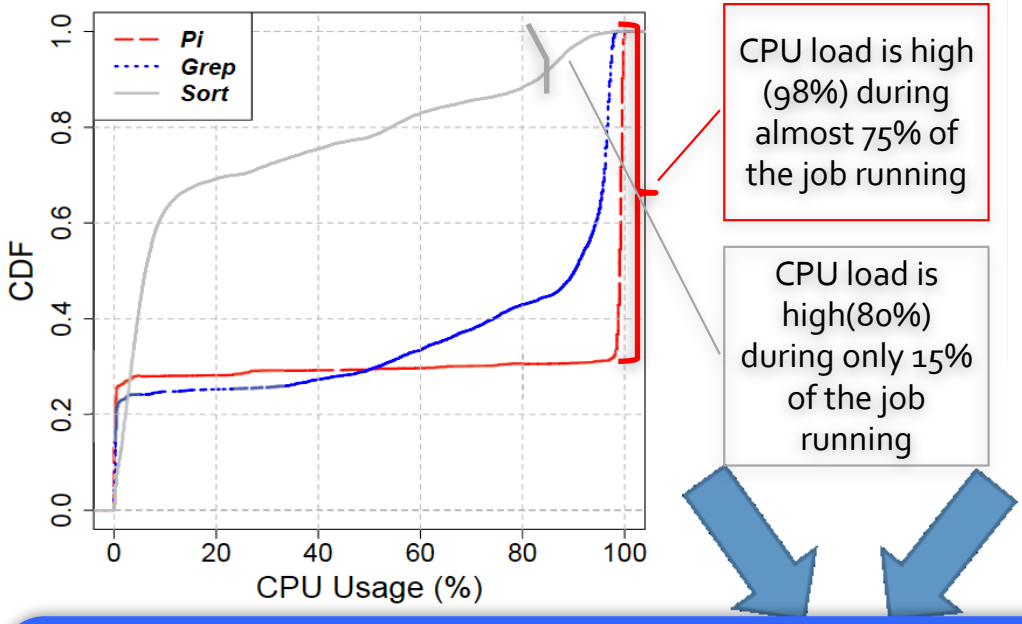
THE engine for **Big Data** processing in **data-centers**

Power bills become a substantial part of the total cost of ownership (**TCO**) of **data-centers**

It is essential to explore and optimize the energy efficiency when running Big Data application in Hadoop

Investigate the Impacts of CPU-Frequencies Scaling on Power Efficiency in Hadoop

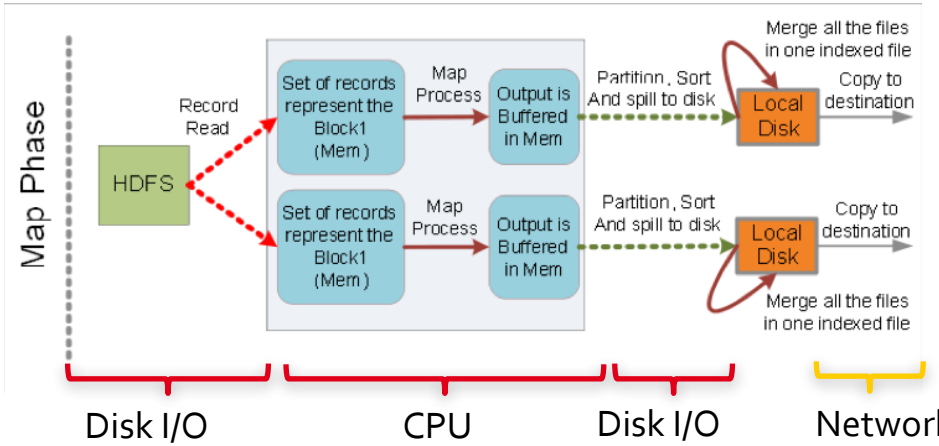
Diversity of MapReduce applications



CPU load is high (98%) during almost 75% of the job running

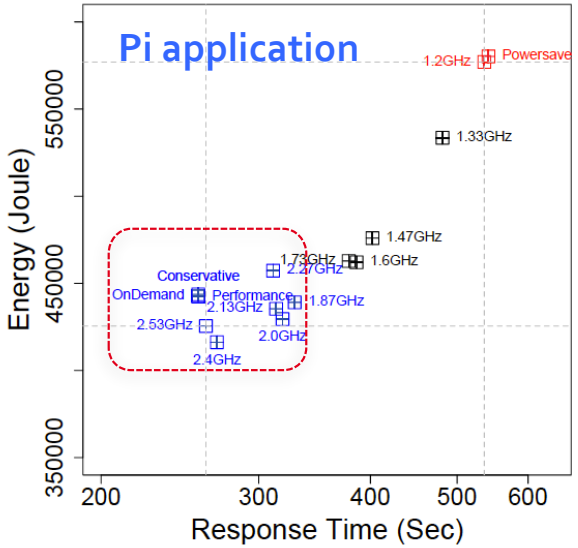
CPU load is high(80%) during only 15% of the job running

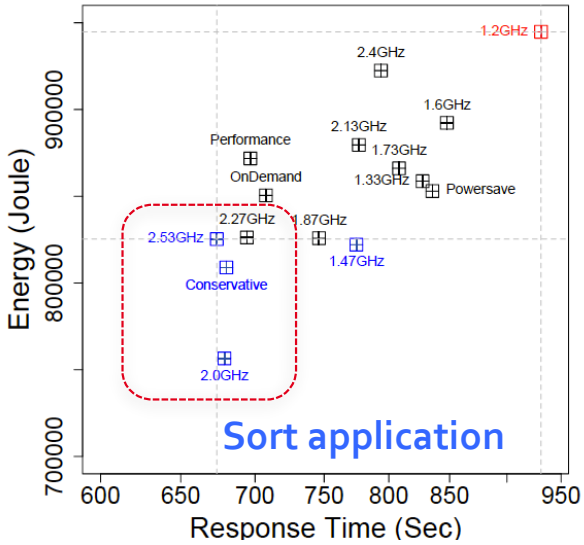
Multiple phases of MapReduce application



There is a significant potential of energy saving by scaling down the CPU frequency when peak CPU is not needed

Investigate the Impacts of CPU-Frequencies Scaling on Power Efficiency in Hadoop





We observe that different DVFS settings are not only sub-optimal for different MapReduce applications but also sub-optimal for different stages of the MapReduce application

Build dynamic frequency tuning tool specifically tailored to match MapReduce application types and execution stages

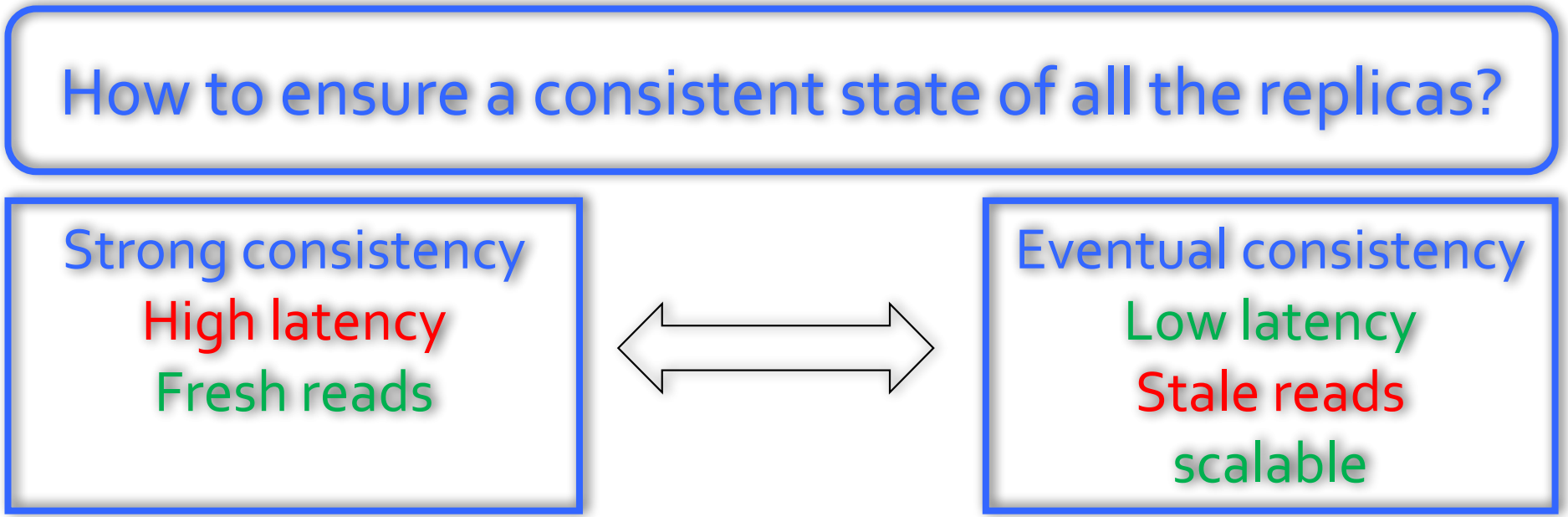
"Towards Efficient Power Management in MapReduce: Investigation of CPU-Frequencies Scaling on Power Efficiency in Hadoop", S. Ibrahim et al, ARMS-CC 2014.

Consistency Management in the Cloud

Replication has become an essential feature in storage system and is extensively leveraged in the cloud

- Fast access
- Enhanced performance
- High availability

How to ensure a consistent state of all the replicas?



Expose the tight relation between **Consistency** and

- **Performance**
- **Monetary Cost**
- **Power Consumption**

Expose the tight relation between **Consistency** and

- **Performance**
- **Monetary Cost**
- **Power Consumption**

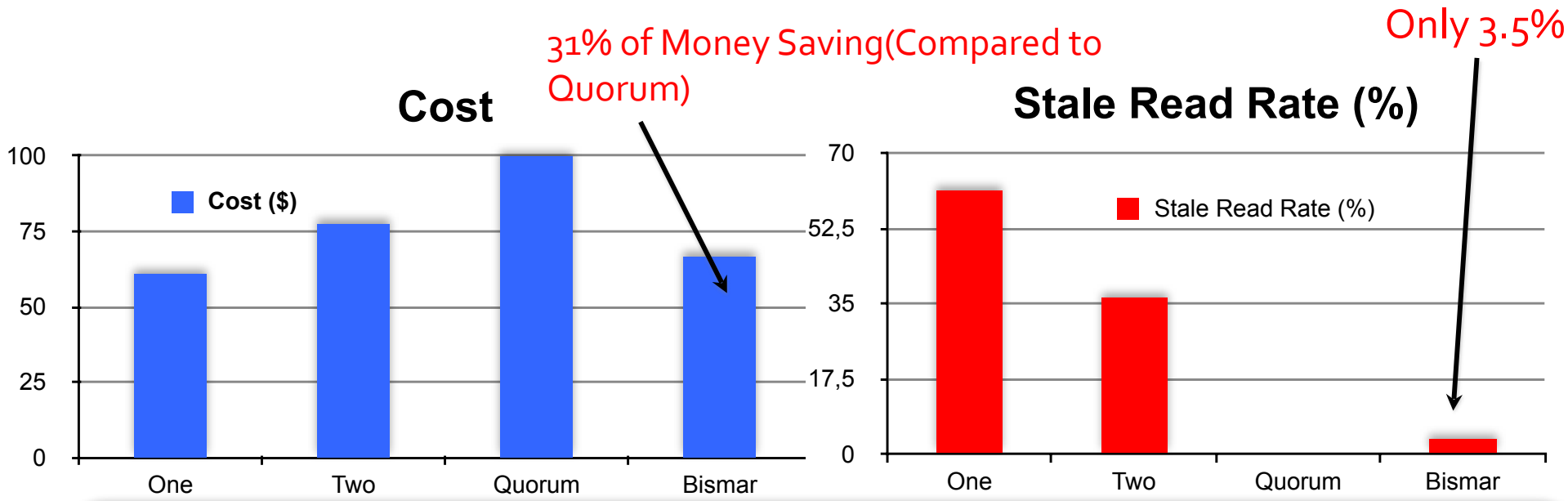
PNUTS

twitter

riak

facebook

Bismar: Cost-Efficient Consistency Management for the Cloud



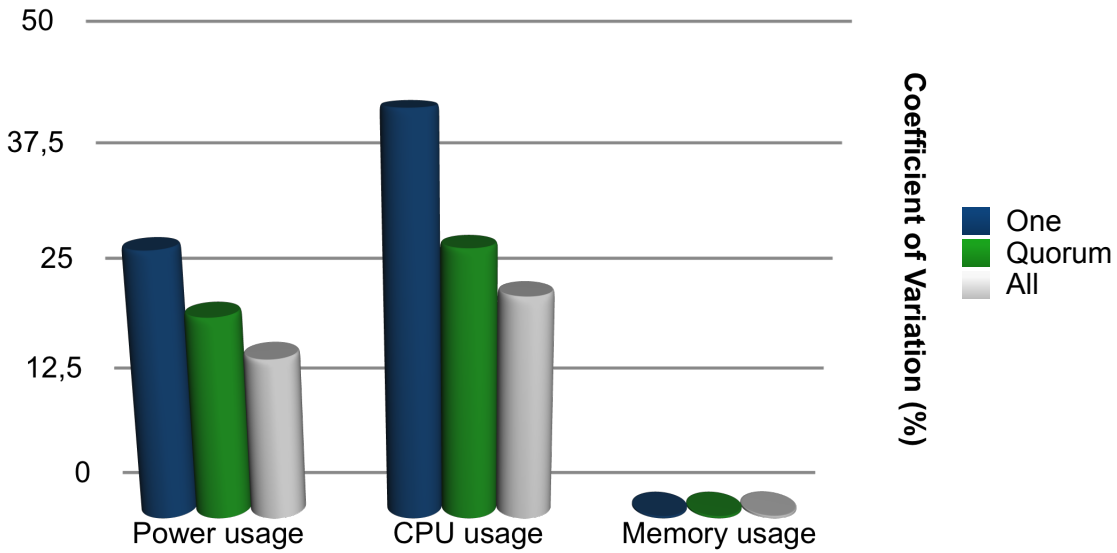
Bismar reduces the monetary cost by 31% while resulting in only 3.5% of stale data reads

"Harmony: Towards automated self-adaptive consistency in cloud storage" HE Chihoub, S Ibrahim, G Antoniu, MS Perez. CLUSTER 2012

"Consistency in the cloud: When money does matter!" HE Chihoub, S Ibrahim, G Antoniu, MS Pérez.CCGrid 2013

Energy vs Consistency

Exploring the tight relationship of Consistency *vs* Energy



Coefficient of Variation: high variation with low consistency levels

We observe that there are three main factors contribute to energy consumption in Cassandra cluster:

- Consistency Models
- Workload access patterns
- Degree of concurrency

Eventual Consistency introduces usage variability between storage nodes

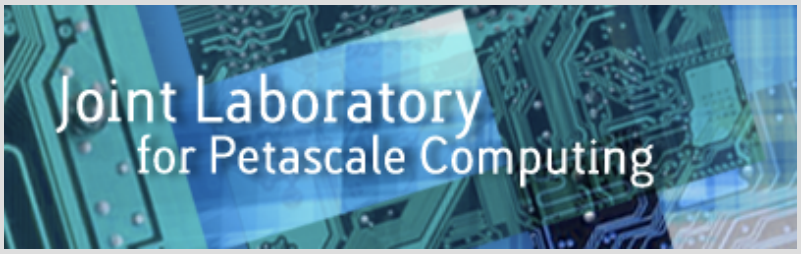
Adaptive Configuration of the Storage Cluster

INVENTORS FOR THE DIGITAL WORLD

- Joint Laboratory on Extreme Scale Computing
  - INRIA
  - University of Illinois at Urbana Champaign (UIUC)
  - Argonne National Laboratory (ANL)
  - Barcelona Supercomputing Center (BSC)

Focus 2:

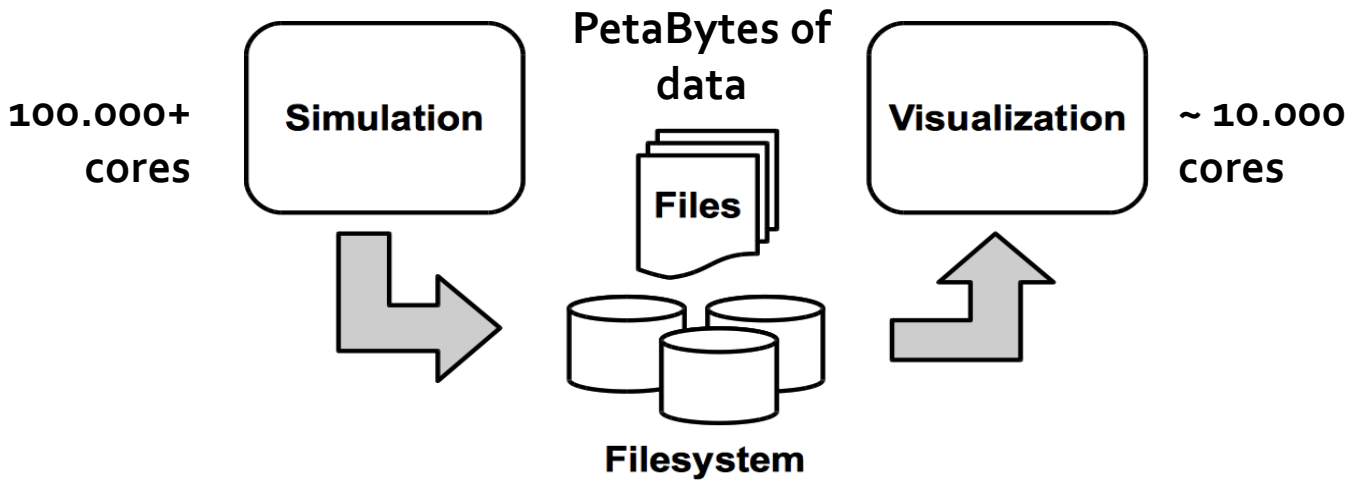
Scalable big data Management on Post-Petascale HPC systems



The need for Scalable, Smart, yet Efficient I/O Management

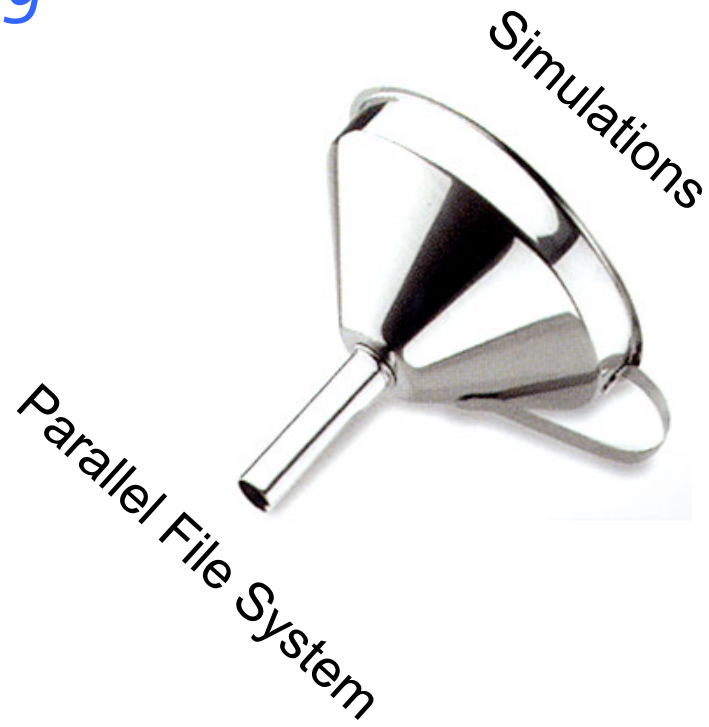
Tomorrow's supercomputer

- Millions of cores
- Increasing gap between compute and I/O performance



Efficient I/O Using Dedicated Cores

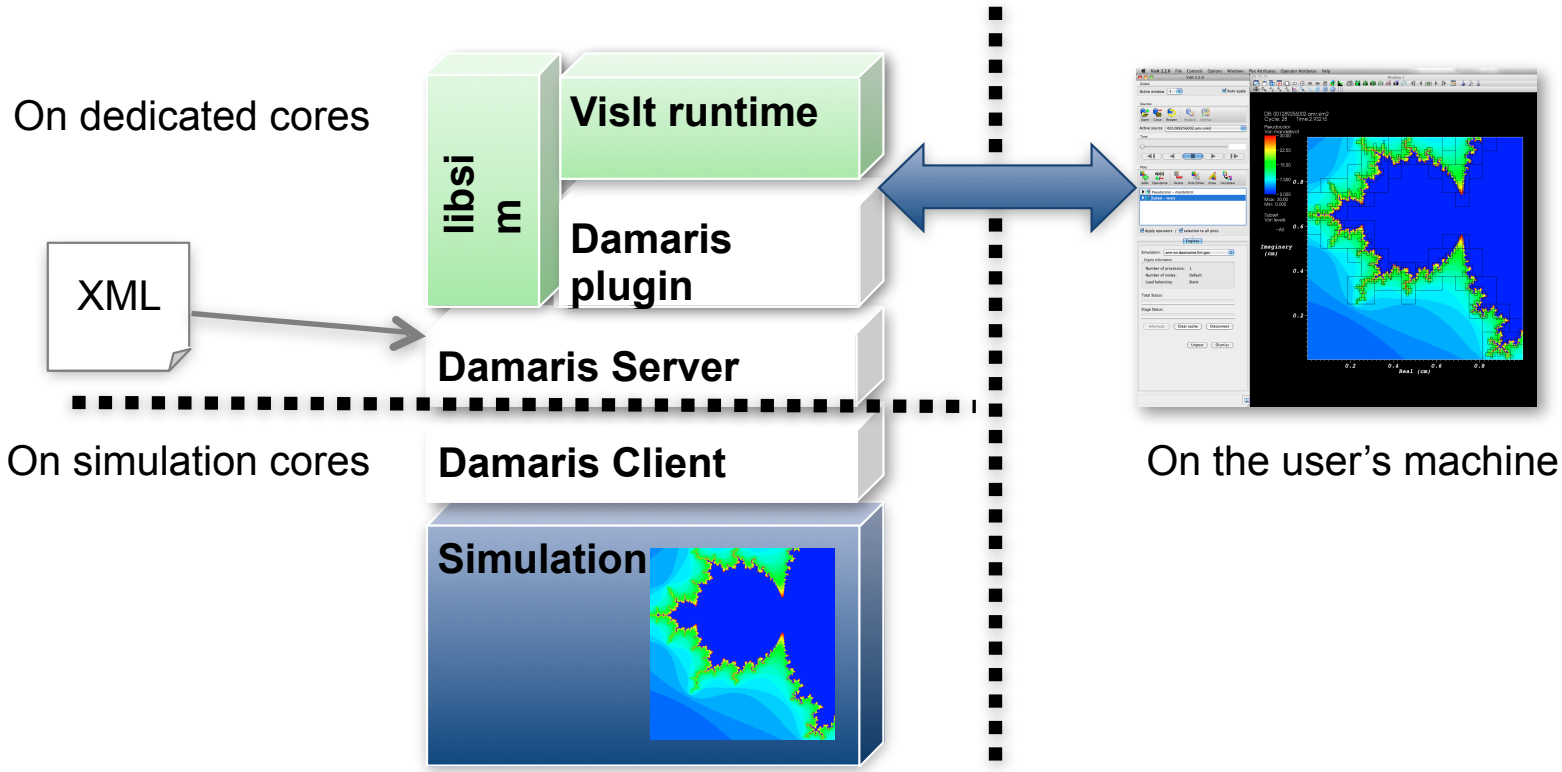
- **Damaris:** *Dedicated Adaptable Middleware for Application Resources Inline Steering*
- **Main idea:**
  - Dedicate cores in each SMP node for data management



"Damaris: how to efficiently leverage multicore parallelism to achieve scalable, jitter-free i/o"
M Dorier, G Antoniu, F Cappello, M Snir, L Orf.
CLUSTER 2012

http://damaris.gforge.inria.fr/

Towards smart in-situ Visualization



- Connect visualization software to simulation
- Perform visualization when the simulation runs
- Perform "smart" visualization, i.e. reduce image resolution when needed

http://damaris.gforge.inria.fr/

Energy Efficient I/O Management

Power bills become a substantial part of the total cost of ownership (TCO) of supercomputers

- A typical supercomputer of thousands of cores consumes several megawatt of power

Performance has long been the major focus of the HPC community

- No.1 supercomputer, Tianhe-2 : performance of 33.8 PFLOPS but with a 24 MW power consumption

Energy Efficient I/O Management

Power bills become a substantial part of the total cost of ownership (TCO) of supercomputers

- A typical supercomputer of thousands of cores consumes several megawatt of power

Performance has long been the major focus of the HPC community

- No.1 supercomputer, Tianhe-2 : performance of 33.8 PFLOPS but with a 24 MW power consumption

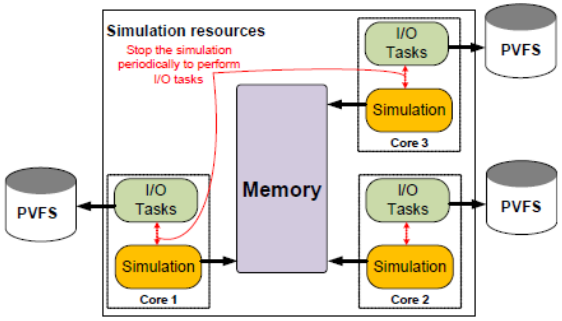
Energy will even increase as we reach the era of Exascale systems.

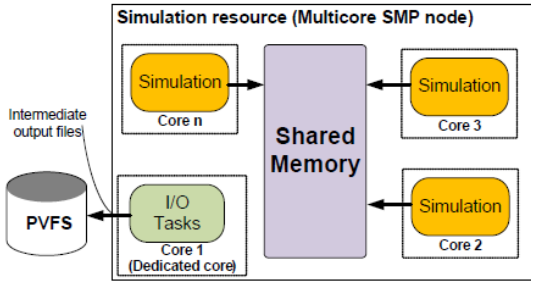
Energy Efficient I/O Management

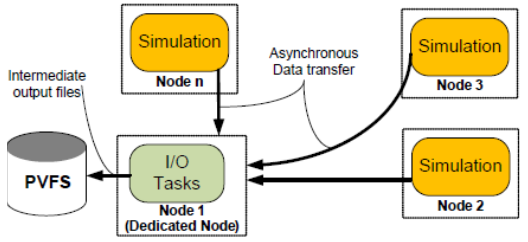
Evaluate the energy consumption of different I/O approaches based on dedicated cores, or dedicated nodes

Three factors at least contribute to such variations:

- The adopted I/O approach
- The output frequency
- The architecture of the system on which we run the HPC application

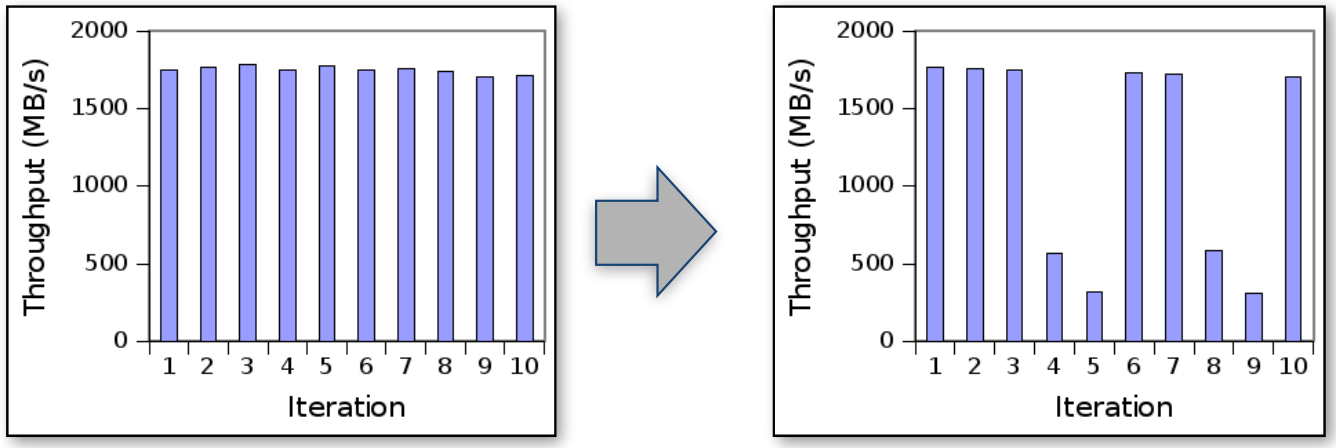






"A performance and energy analysis of I/O management approaches for exascale systems" O Yildiz, M Dorier, S Ibrahim, G Antoniu. DIDC 2014

Mitigating I/O Interference



CALCioM: Cross-Application Layer for Coordinated I/O Management

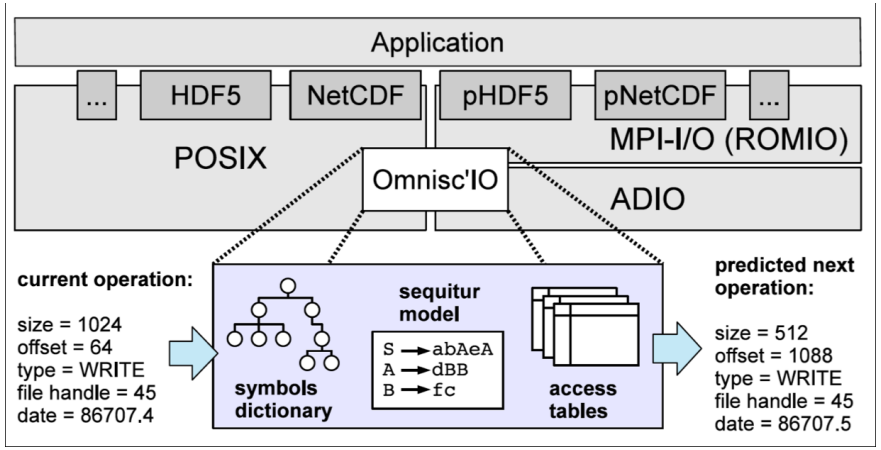
- **Goal:**
  - Make applications communicate their I/O behavior to one another
  - Make them coordinate to avoid interfering
  - Choose best coordination strategy dynamically
- Experiments with Grid'5000 and Surveyor (BlueGene/P, ANL)

"Calciom: Mitigating i/o interference in hpc systems through cross-application coordination."M Dorier, G Antoniu, R Ross, D Kimpe, S Ibrahim. IPDPS 2014

Predicting I/O Patterns

Goal: predict the spatial and temporal I/O patterns
Omnisc'IO: use context-free grammars

- Predictions:
  - Where, when, and how much
  - At run time
  - With negligible overhead
  - And negligible memory footprint
- Results:
  - With CM1, Nek5000, GTC and LAMMPS
  - On Grid'5000



"Omnisc'IO: a grammar-based approach to spatial and temporal I/O patterns prediction"
M Dorier, S Ibrahim, G Antoniu, R Ross - SC'14

Thank you!

Shadi Ibrahim

INRIA research Scientist

KerData Team

Shadi.ibrahim@inria.fr