HEMERA Large-Wingspan Action
Focus: Scalable Data Management

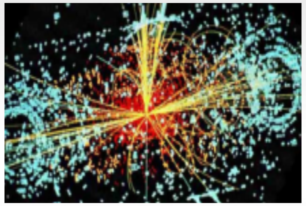
MapReduce Challenge
Data Management Working Group

Gabriel Antoniu, KerData Research Team

Context: the Data Deluge

Experiments

Simulations



Archives

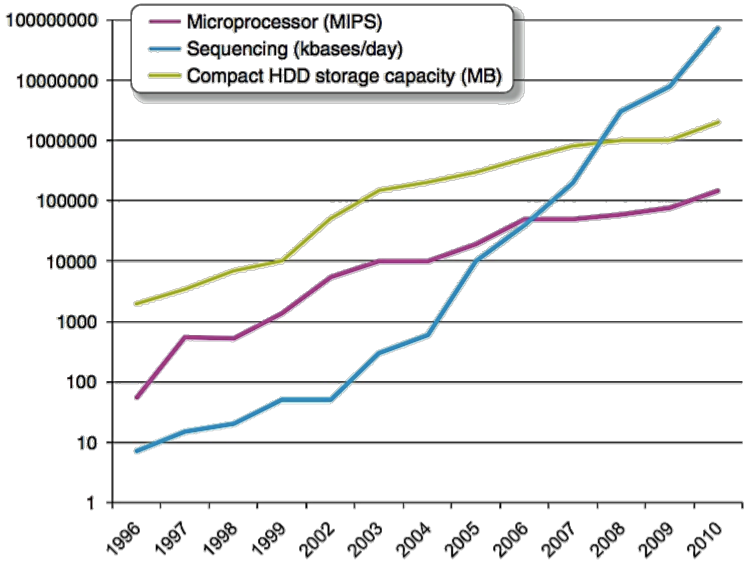
Literature

Consumer

The Challenge
Enable Discovery

Deliver the capability to mine, search and analyze this data in near real time

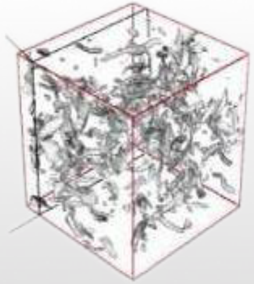
Petabytes
Doubling &
Doubling

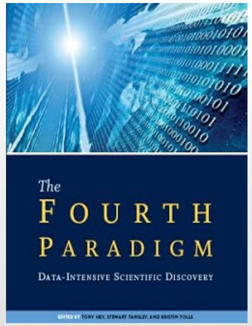


Credits: Microsoft

The Data Science:
The 4th Paradigm for Scientific Discovery

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$





Experimental

Thousand years ago

Description of natural phenomena

Theoretical

Last few hundred years

Newton's laws, Maxwell's equations…

Computational

Last few decades

Simulation of complex phenomena

The Fourth Paradigm

Today and the Future

Unify theory, experiment and simulation with large multidisciplinary Data

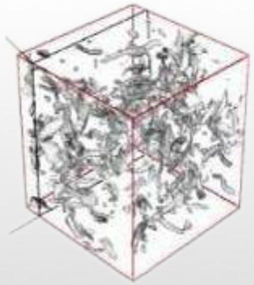
Using data exploration and data mining (from instruments, sensors, humans…)

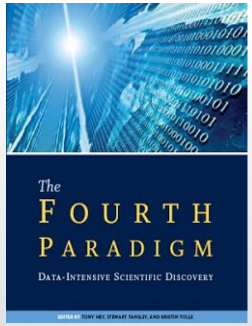
Distributed Communities

Crédits: Dennis Gannon

The Data Science:
The 4th Paradigm for Scientific Discovery

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$





Office of Science and Technology Policy
Executive Office of the President
New Executive Office Building
Washington, DC 20502

FOR IMMEDIATE RELEASE
March 29, 2012

Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE:
ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

The Fourth Paradigm

Today and the Future

Unify theory, experiment and simulation with large multidisciplinary Data

Using data exploration and data mining (from instruments, sensors, humans…)

Distributed Communities

Research Challenges

- A few applications
  - Massive data analysis on clouds (e.g. MapReduce)
  - Advanced cloud data services (adaptive replication, consistency)
  - Post-Petascale HPC simulations on supercomputers

- Focus 1: Scalable data analysis and storage on clouds
  - *Challenge : understand how to reconcile performance, scalability, security and quality of service according to the requirements of data-intensive applications*

- Focus 2: Scalable data I/O, storage and visualization on Post-Petascale HPC systems
  - *Challenge: go beyond the limitations of current file-based approaches*

Focus 1:

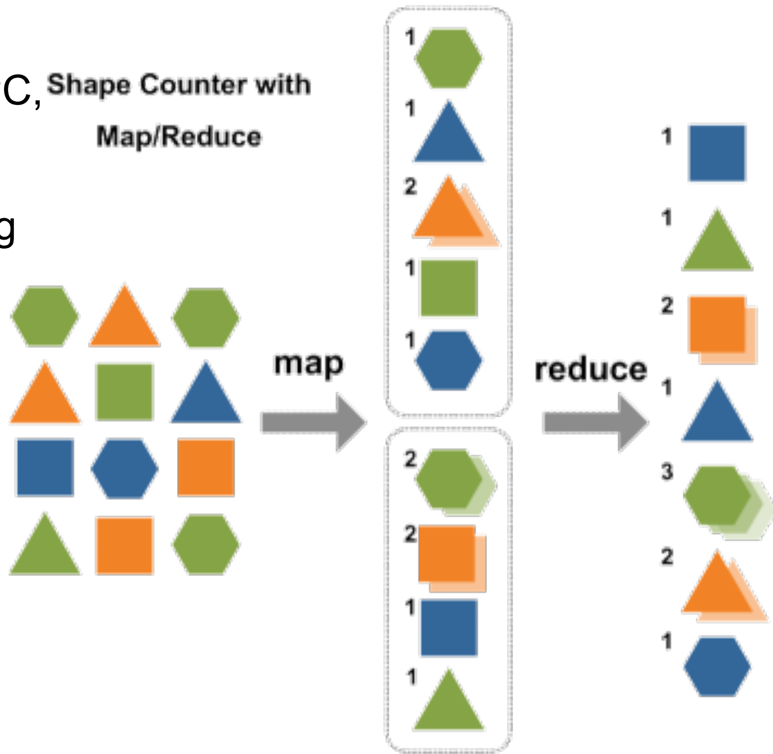
Scalable Data Analysis on Clouds

Scalable Map-Reduce Processing

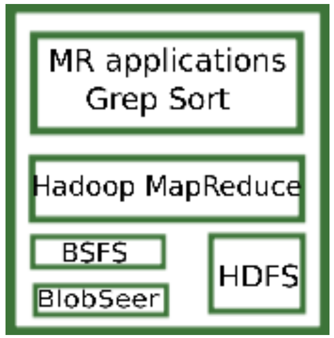
- ANR Project Map-Reduce (ARPEGE, 2010-2014) associated to the MapReduce HEMERA Challenge

  Partners: INRIA (teams : KerData - leader, AVALON, Grand Large), Argonne National Lab, UIUC, JLPC, IBM, IBCP, MEDIT

  - Goal: High-performance Map-Reduce processing through concurrency-optimized data processing
  - URL: mapreduce.inria.fr



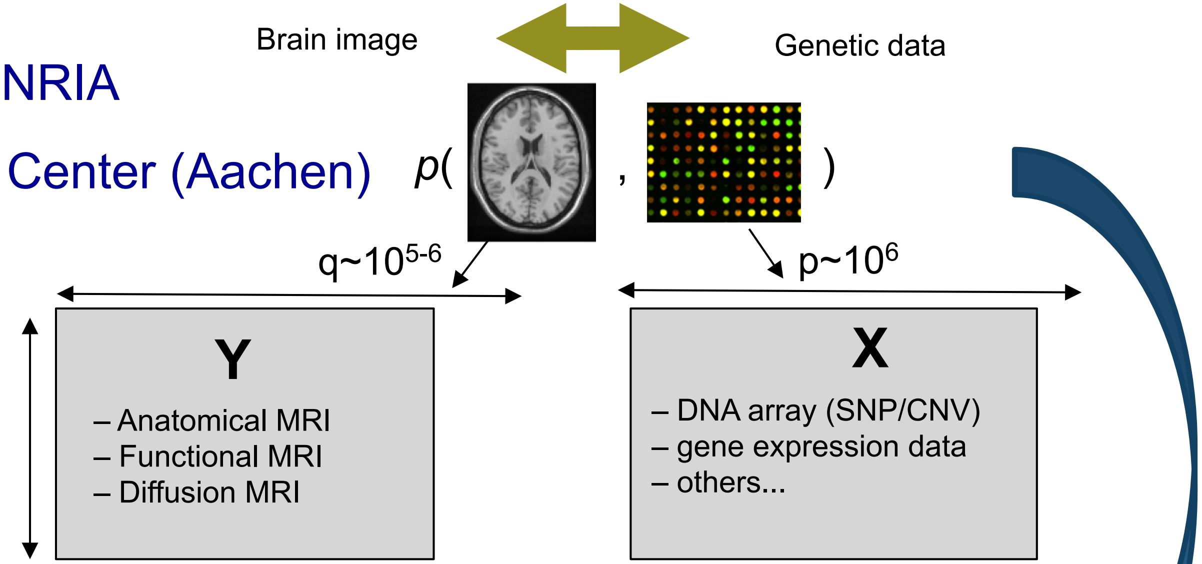
- **Idea**: Use BlobSeer as back-end storage for VM images and cloud application data
  - Versioning capability: lock-free access
  - Efficient intermediate storage in pipelines
- **Experiments done on Grid'5000**
  - Up to 300 nodes/500 cores
  - Plans: joint deployment G5K+FutureGrid (USA)



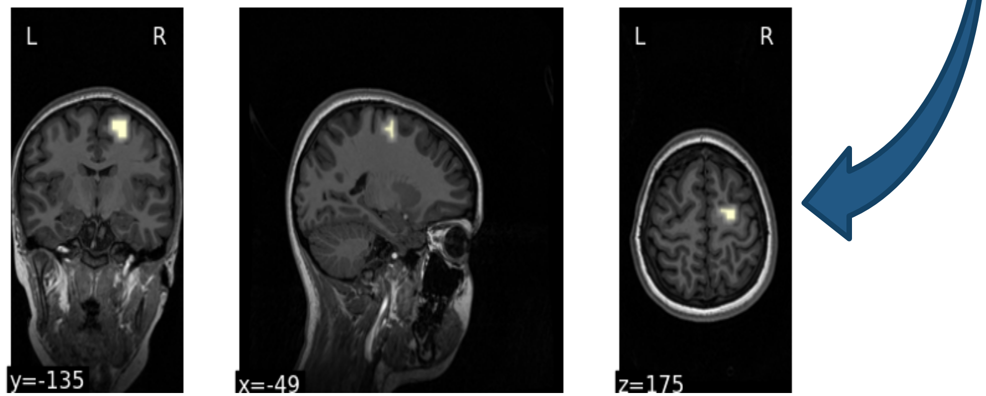
- **Papers**: JPDC, Concurrency and Computation Practice and Experience, ACM HPDC 2011 (AR:12.9%), ACM HPDC 2012, IEEE/ ACM CCGRID 2012 et 2013, Euro-Par 2012, IEEE IPDPS 2013

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

- KerData, PARIETAL teams at INRIA
- European Microsoft Innovation Center (Aachen)



N~2000



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

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

A Recent Result: Automated Self-Adaptive Consistency in Cloud Storage

- Replication has become an essential feature in cloud storage systems
- Issue: How to ensure a consistent state of data replicas?

Strong consistency	⟺	Eventual consistency
High latency		Low latency
Fresh reads		Stale reads
		scalable

PNUTS

Solution: The Harmony Approach
Automated self-adaptive consistency tuning

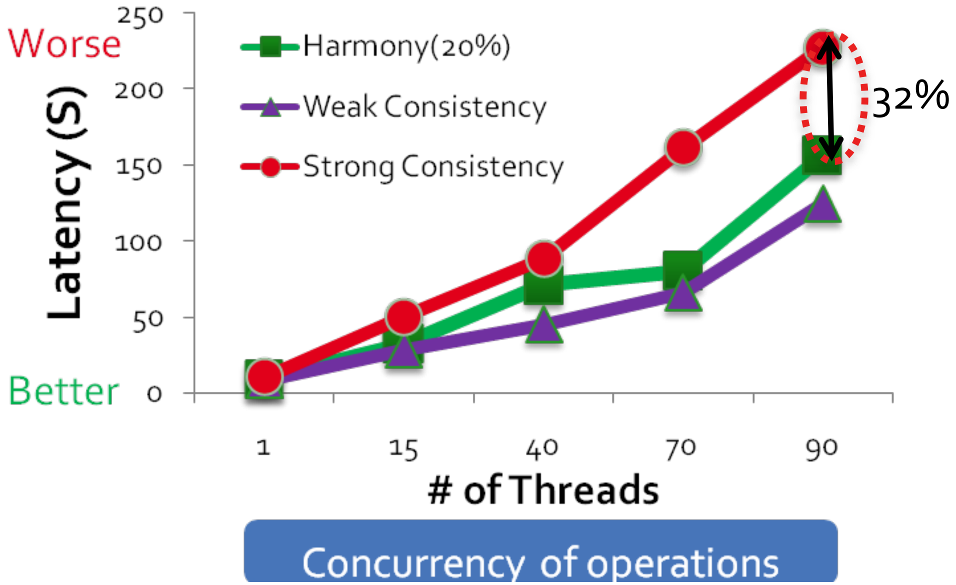
- Define the consistency level at run-time in terms of tolerable *Stale Reads* rate
  - Use a « smart » estimation model of stale reads based on the current application access pattern and on network latency
- Self-Adaptive Consistency: handle at run-time the trade-offs :
  - Consistency-performance
  - Consistency-availability
- Tune the consistency level: set the number of replicas involved in a read operation at run-time based on the stale reads rate that one application may tolerate

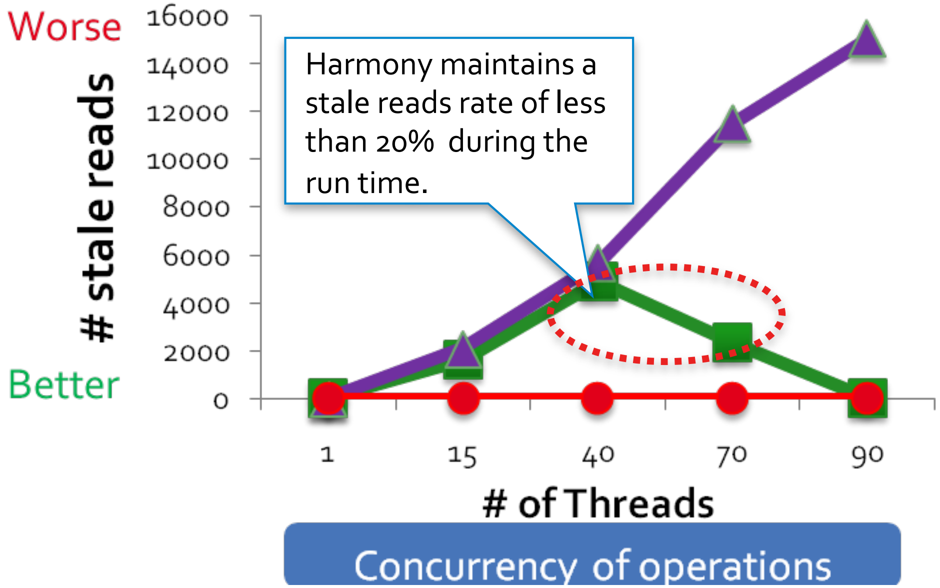
$$
\begin{aligned}
&\textit{If } app\_stale\_rate \geq \theta_{stale} \textit{ Then}\\
&\quad \textit{Choose Eventual Consistency } (Consistencylevel = 1)\\
&\textit{Else}\\
&\quad -\textit{Compute } X_n \textit{ the number of always consistent}\\
&\quad\;\; \textit{replicas necessary to have } app\_stale\_rate \geq \theta_{stale}\\
&\quad -\textit{Choose Consistency level based on } X_n\\
&\textit{End if}
\end{aligned}
$$

Harmony: Results

- Harmony reduces the latency for strong consistency by 32%
- Harmony maintains the desired consistency level

Never exceeds the pre-defined tolerable stale reads rate (20%)





Best Presentation Award for Shadi Ibrahim,

Postdoc fellow funded by Hemera, at the Grid'5000 School (Dec 2012)

Scalable Storage on Clouds: Open Issues

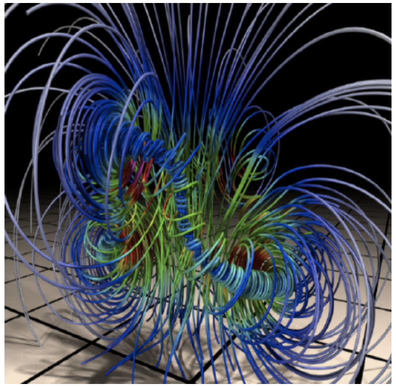
- Understanding price-performance trade-offs
  - Consistency, availability, performance, cost, security, quality of service, energy consumption
  - Autonomy, adaptive consistency
  - Dynamic elasticity
  - Trade-offs exposed to the user
- High performance variability
  - Understand it, model it, cope with it
- Deployment/application launching time is high
- Latency of data accesses is still an issue
- Data movements are expensive
- Cope with tightly-coupled applications
- Cope with various cloud programming models
- Virtualization overhead
- Benchmarking
- Performance modeling
- Self-optimization for cost reduction
  - Elastic scale down
- Security and privacy

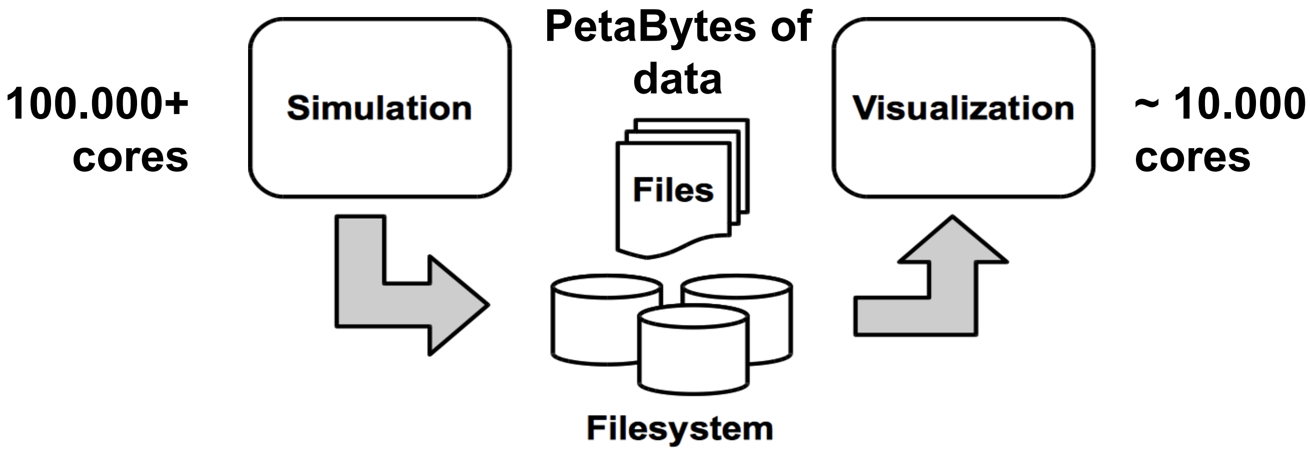
Grid'5000 and the Hemera Community are essential to make progress in these directions!

Focus 2:

Scalable Data I/O, Storage and Visualization for Post-Petascale HPC

Context :
Data Management in Post-Petascale HPC Systems



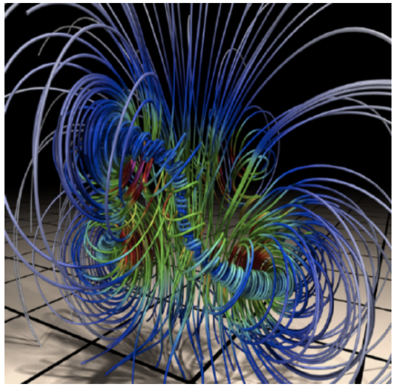


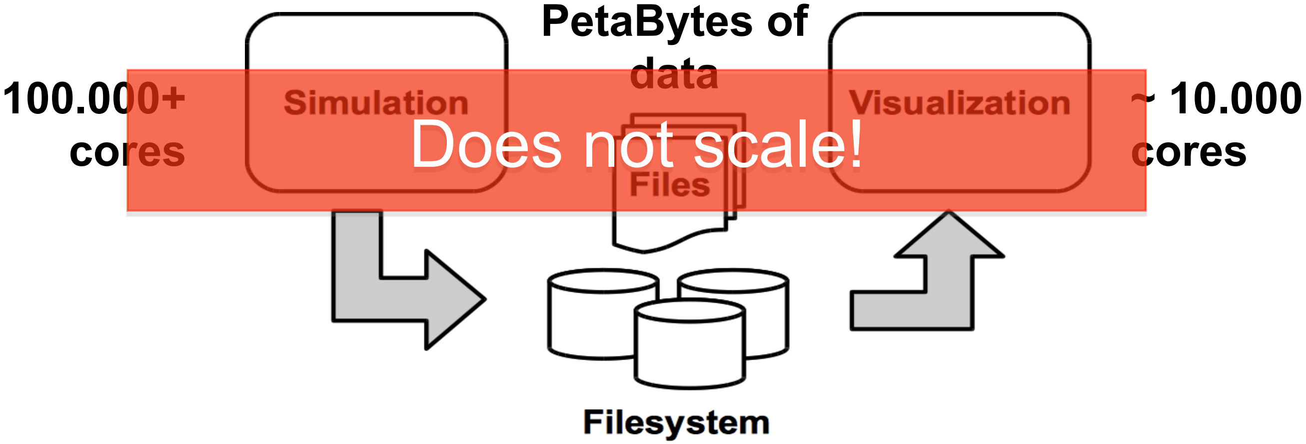
- ✧ Problem: simulations generate TB/min
- ✧ How to store et transfer data ?
- ✧ How to analyze, visualize and extract knowledge?

Joint Laboratory for Petascale Computing

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

Context:
Data Management in Post-Petascale HPC Systems





- ✧ Problem: simulations generate TB/min
- ✧ How to store et transfer data ?
- ✧ How to analyze, visualize and extract knowledge?

What is difficult?

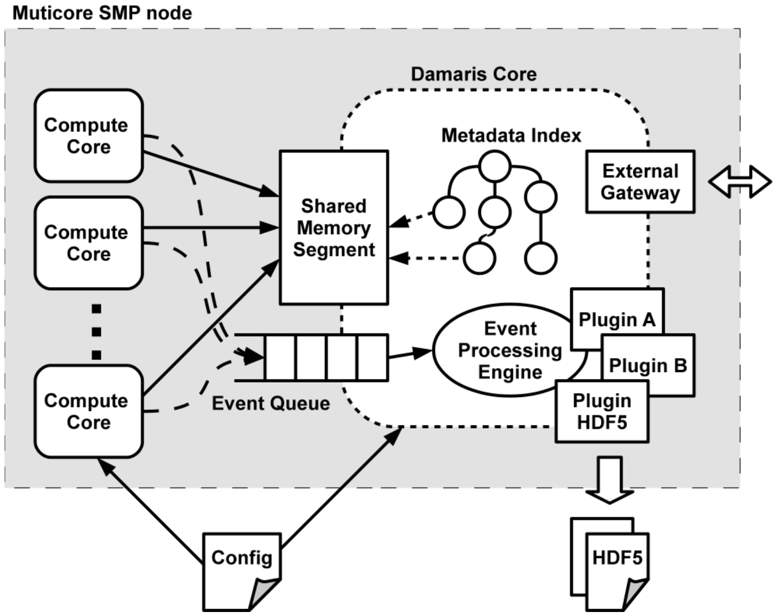
- Too many files (e.g. Blue Waters 100.000+ files/min)
- Too much data
- Unpredictable I/O performance

Joint Laboratory for Petascale Computing

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

Damaris: A Middleware-Level Approach to I/O on Multicore HPC Systems

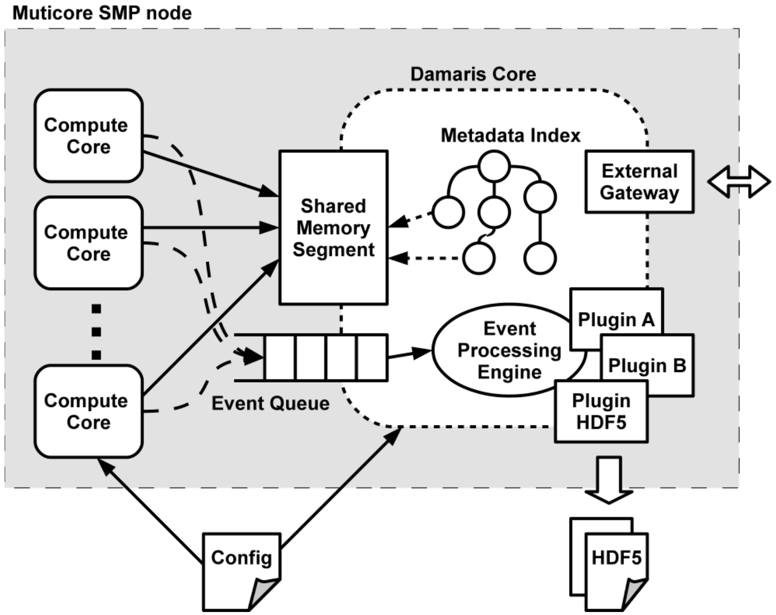
- **Idea** : one dedicated I/O core per multicore node
- **Originality** : shared memory, asynchronous processing
- **Implementation**: software library
- Application: Tornado simulation (Blue Waters)
- **Preliminary experiments on Grid'5000**



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

Damaris: A Middleware-Level Approach to I/O on Multicore HPC Systems

- **Idea** : one dedicated I/O core per multicore node
- **Originality** : shared memory, asynchronous processing
- **Implementation**: software library
- Application: Tornado simulation (Blue Waters)
- **Preliminary experiments on Grid'5000**



- **Scales on 10,000+ cores on Kraken (11th of Top500)**
- **Scales on 16,000+ cores on Titan (1st of Top500)**
- x12 less files
- Overhead-free compression
- Predictable performance

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

Damaris: Early Impact

- First results of the Joint Lab for Petascale Computing transfered to Blue Waters (2011)

- 2nd Award for Matthieu Dorier at the ACM Student Research Competition (ICS 2011)

« *This work is practically very useful (and novel) to the field of computational meteorology and probably fluid dynamics. I think **Damaris** is going to be the best I/O option for these unprecedented supercomputing simulations.*»

Leigh Orf, atmospheric scientist, Central Michigan University

Scalable Storage and I/O on HPC Systems: Open Issues

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

- Explore dedicated I/O cores for *in-situ* visualization and processing
  - Automatic analysis and image generation
  - Adaptive image generation (recognize « interesting » data subsets)
  - Interactive visualization
- Coupling with alternative approaches (I/O forwarding, etc.)

- Collaboration with application communities
  - Data@Exascale Associate Team (2013-2015)
    - Joint Lab for Petascale Computing
    - INRIA, Argonne National Lab, University of Illinois at Urbana-Champaign

To go further…

Grid'5000 is a unique and essential tool for investigating open issues!

- Control over the infrastructure
- Root access
- Customized environment
- Multi-cluster deployments
- Multi-site deployments
- Shared expertise within the Hemera community
- A strong asset in collaborative International and European projects
- A pioneer project serving as a model and as a "seed": FutureGrid (USA), BonFire (Europe)