Bull
atos technologies

Atos
Worldwide IT Partner

# Atos

# Agenda – Intervention Bull Atos

| | |
|---|---|
| ▶ **Présentation générale de Bull/Atos** | *Stéphane Carbonneau & Denis Rongier (10min)* |
| ▶ **Introduction, état de l'art du big data**<br>• Technologies Software (Datalake BDCF)<br>• Architectures Hardware | *Frank Marendaz & Benoit Pelletier (30 min)* |
| ▶ **Vision  sur la convergence HPC Big Data** | *Pascale Rosse-Laurent (30 min)* |
| ▶ **Scheduling** | *Michael Mercier (20 min )* |
| ▶ **Stockage Hadoop over Lustre** | *Eric Morvan (20 min )* |
| ▶ **DeepLearning** | *Guillaume André (20 min)* |
| ▶ **Contexte de workflow**<br>• Illustration avec le cas OMICS | *Pascale Rosse-Laurent (20 min)* |
| ▶ **Echanges**<br>• Commentaires, questions & réponses | *Tous (30 min)* |

# Présentation générale de Bull/Atos

Stéphane Carbonneau & Denis Rongier

# Atos en quelques mots

## Atos est un leader international dans les Services Numériques

**Faits**

- chiffre d'affaires annuel de **12 milliards d'euros**
- **100,000 employés** dans **72 pays**
- **5000 brevets**

**Nous renouvelons l'expérience client**

**Nous permettons la réinvention du Business**

**Nous fournissons un environnement sécurisé**

**Nous assurons l'excellence opérationnelle**

**Solutions clés**

Fournir à nos clients une chaine de valeur complète d'offres et de services, avec une forte expertise industrielle

- Consulting & Systems Integration
- Big Data & Cyber Security
- Managed Services
- Cloud Computing services through Canopy
- E-payment transactional services through Worldline

**Avec un écosystème partenaires de classe mondiale**

**Clients**

**Aujourd'hui pour les Jeux Olympiques, demain pour vous!** Atos est le Partenaire IT **Mondial pour les Jeux Olympiques & Paralympiques**

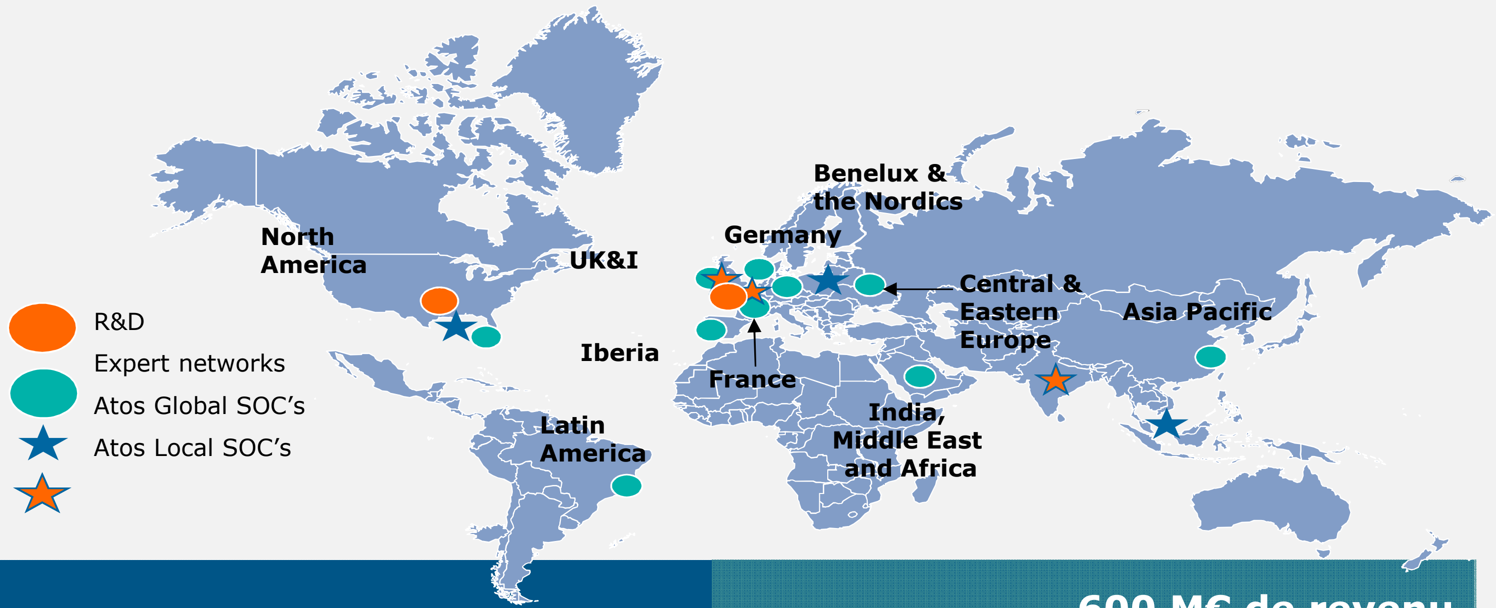
# Activités BDS

**3500 experts dans le monde aux services des plus grands clients publics et privés**
**1900 brevets**

***Une activité équilibrée autour de 2 "core-business"***

**Big Data**
- HPC
- Big Data Entreprise

**Securité**
- Cyber Securité
- Mission Critical Systems

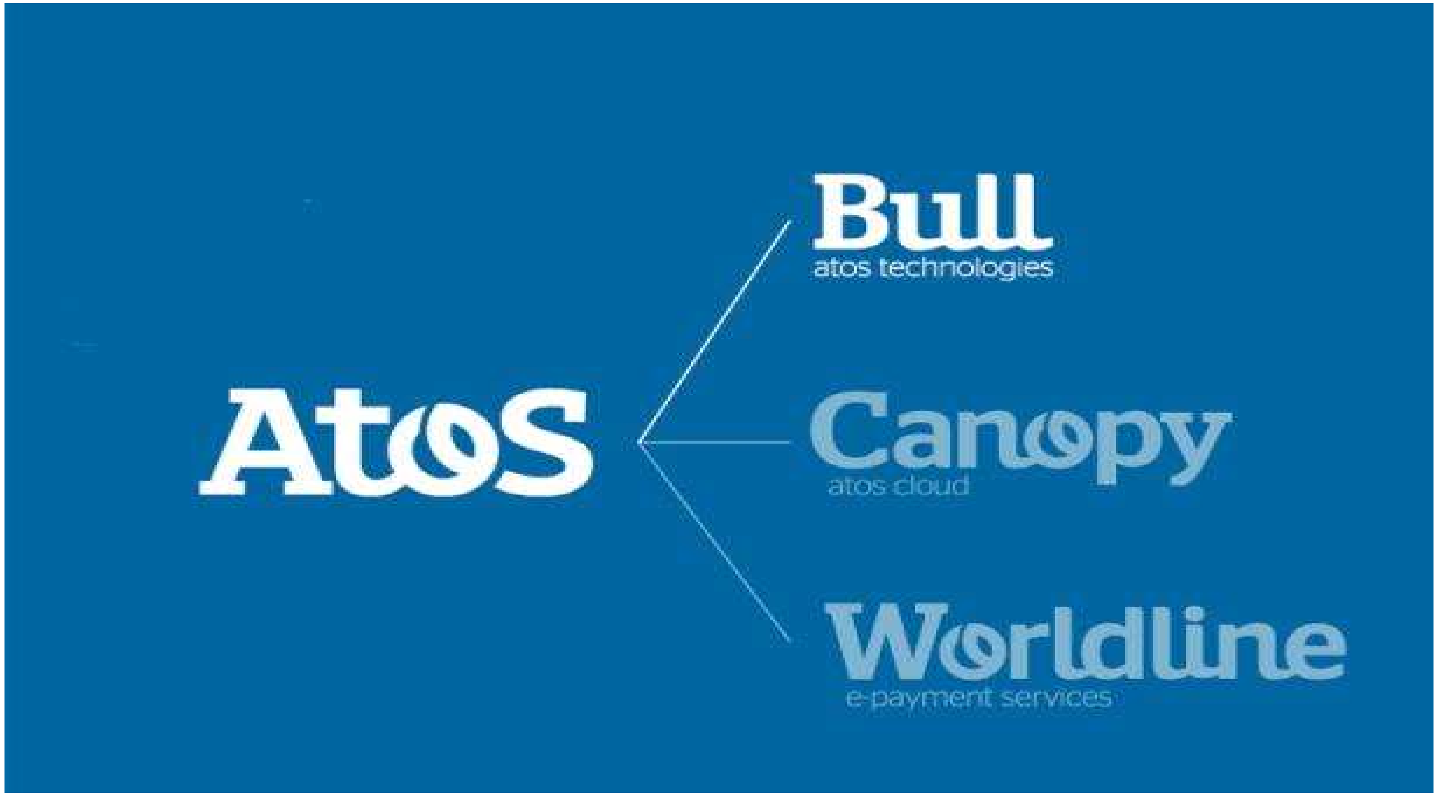



## Chiffres clés

- 1er acteur Européen dans le Big Data,
- leader Européen dans la Cyber Securité,
- reconnu dans les systèmes critiques

**600 M€ de revenu**
**au sein d'Atos (10 B€)**
**3500 experts**
**parmi 85.000 collaborateurs Atos**

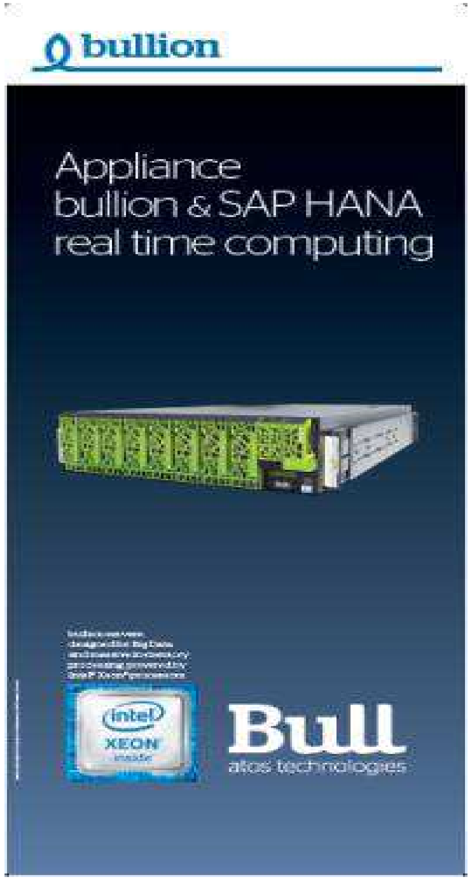
# Bull, Atos Technologies: la marque des produits matériels et logiciels d’Atos

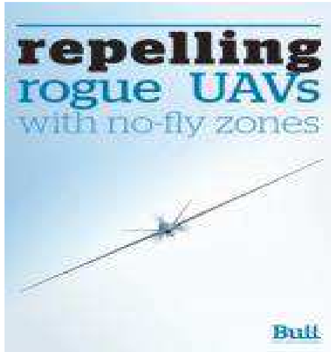




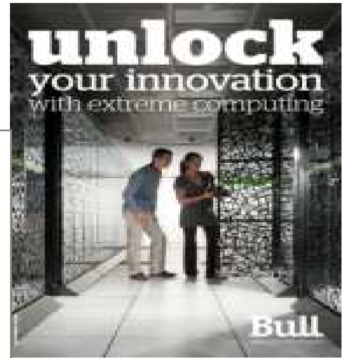
the expert brand for Atos Technologies, hardware and software products

# Nos technologies actuelles pour le Big Data & la Sécurité





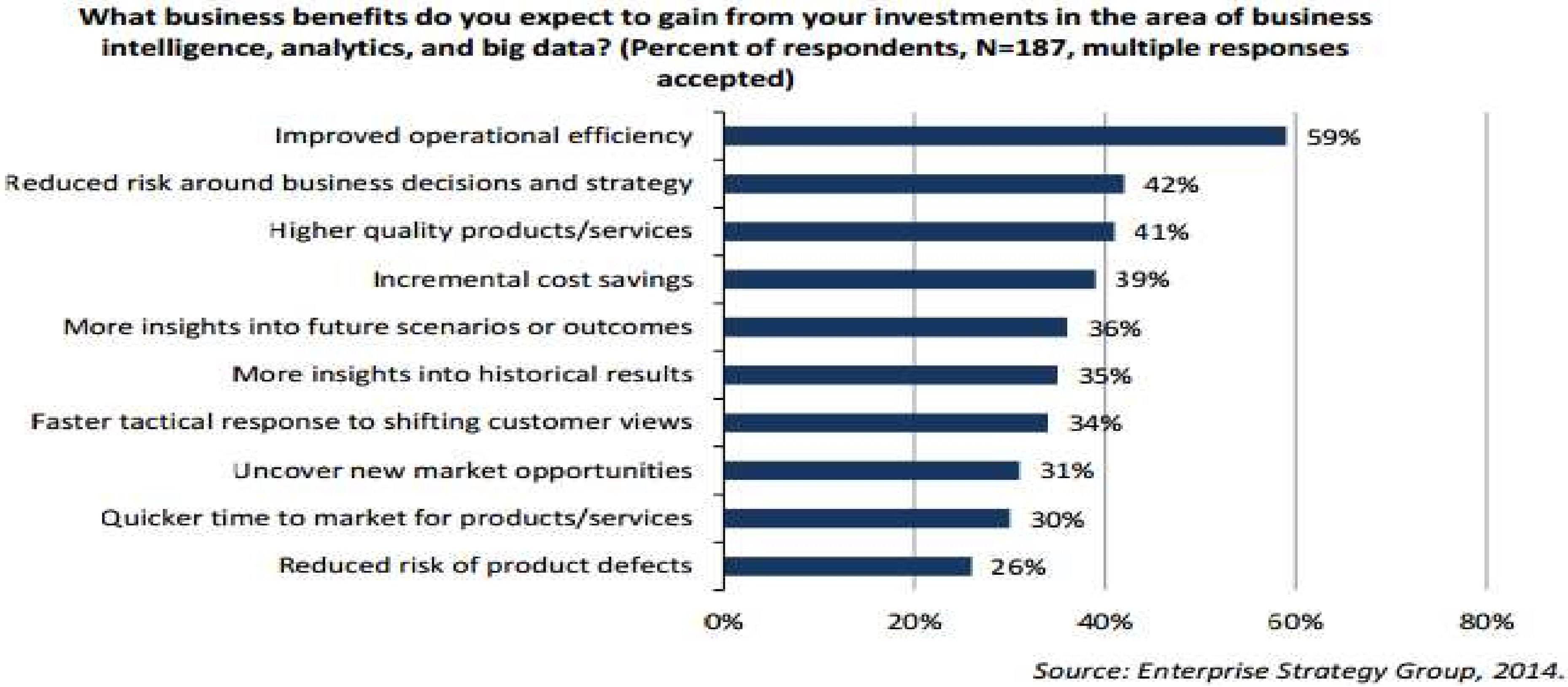


# Introduction, état de l’art du big data

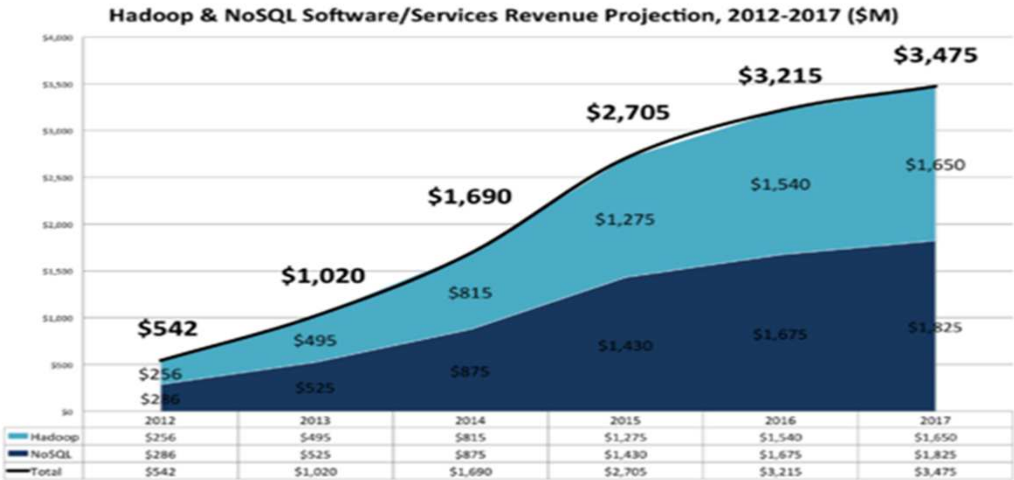
*Frank Marendaz & Benoit Pelletier*

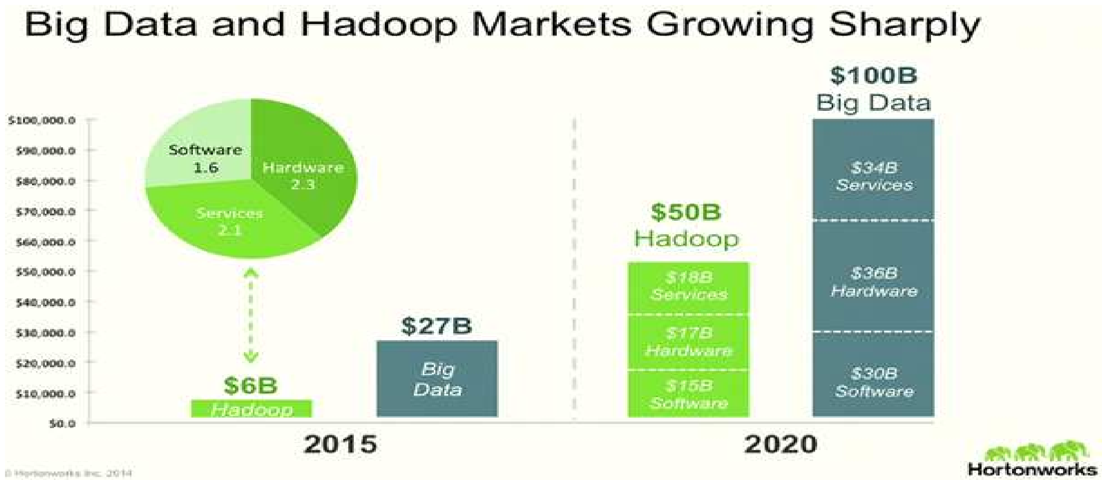
# What are organizations expectations

Regarding analytics & big data (data lake) investments



# Data Lake market




source: wikibon

Whichever the analyst, Data Lake represent a huge opportunity

# Big Data at the heart of Customer Transformation Challenges

Analyst say : Infocentric corporations will **outperform their industry peers financially by more than 20%** *. This is a clear game changer.

Big Data impacts organizations through numerous and diverse use cases

**TRUST & COMPLIANCE**
- Fraud Prevention
- Threat Management ★
- Surveillance ★

…

**CUSTOMER EXPERIENCE**
- Customer 360° ★
- Churn Detection ★
- Dynamic Offering ★

…

**OPERATIONAL EXCELLENCE**
- Predictive Maintenance ★
- Supply-chain Optimization ★
- Dynamic Forecasting ★

…

**BUSINESS REINVENTION**
- Advertising scheduling ★
- Grow core business with new products or service

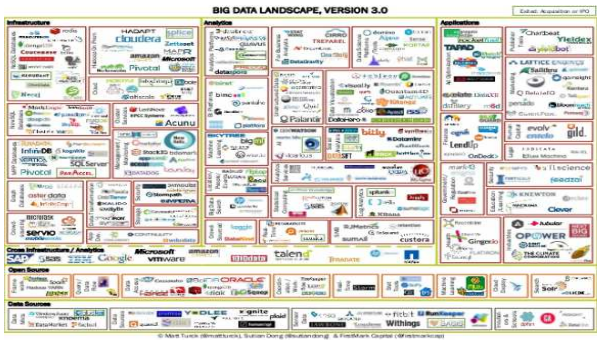
…

★ Existing use cases developed by Atos datascientists team

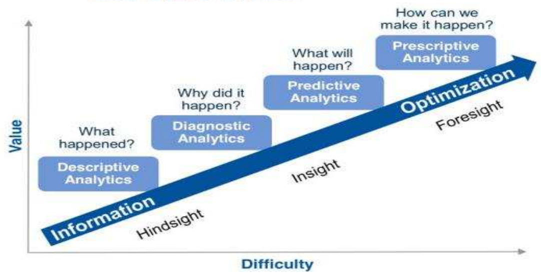
To benefit from it: need to implement a **Datalake with analytics**

# BigData context

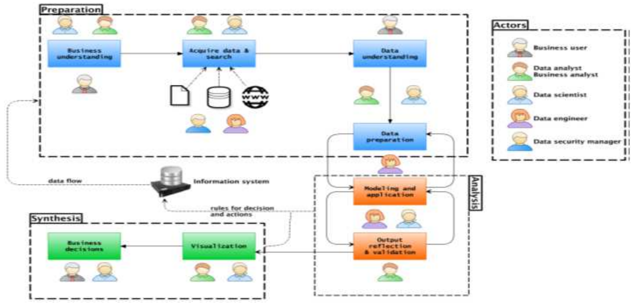
▶ Dynamic ecosystem (frameworks, algos)



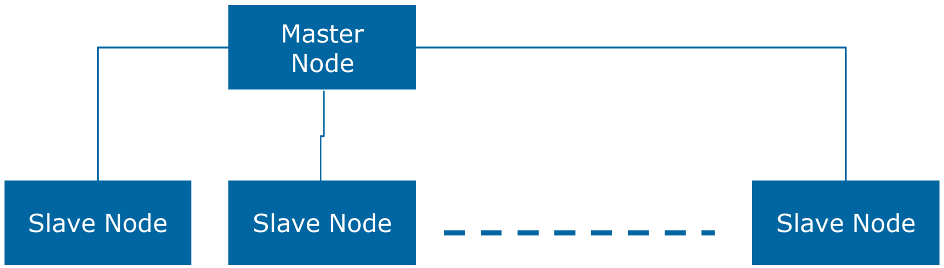





▶ Complex analytics workflow




▶ Distributed architecture




**Operation management is a nightmare**

# Datalake & Analytics: the challenges

## Different challenges for the different involved populations

- Despite many Big Data initiatives, **only 8% use Big Data in Production*.**
- Main it impacts different populations

Business User

- Access a comprehensive set of information (despite applications silos, limits of the BI solutions) with IT global orchestration

Data Scientist

- New player, the one which will « do the magic » with raw data
  - Need to free time to create value:

    80% of the time used to prepare/admin environments vs 20% creating value

IT

- limited resources to deliver quickly new application environments, services catalog to match business requests

- **NEED** for a plug & play platform that allows to focus time on insight and value creation, rather than wasting valuable resources on operations.

*Based on Gartner Estimation

# Data Lake & Analytics Factory by bullion

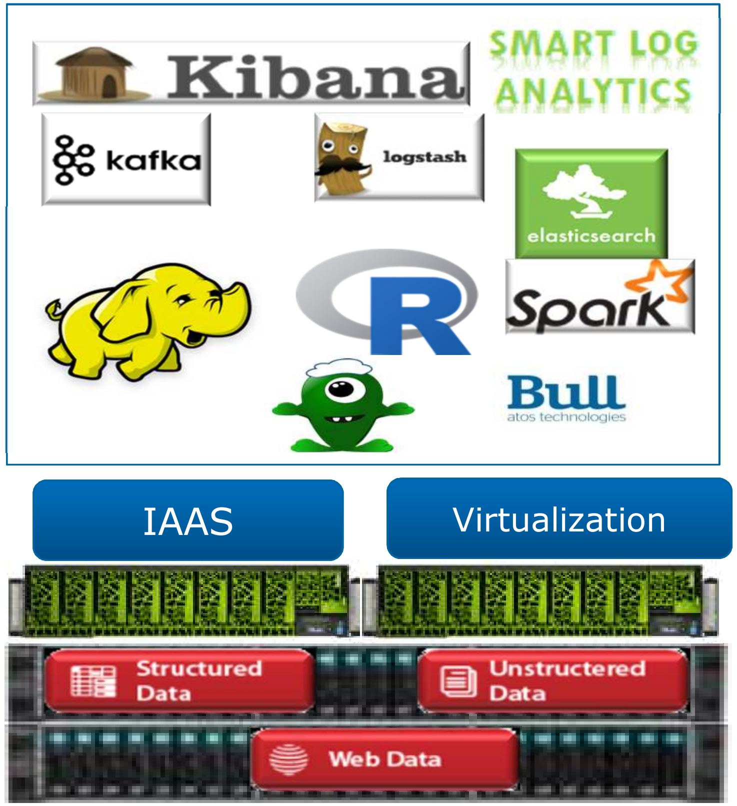
An fully integrated appliance

**Software suite**

- Analytics
- Data collection
- Data Science
- Hadoop ecosystem
- Alien4Cloud

**Infrastructure**

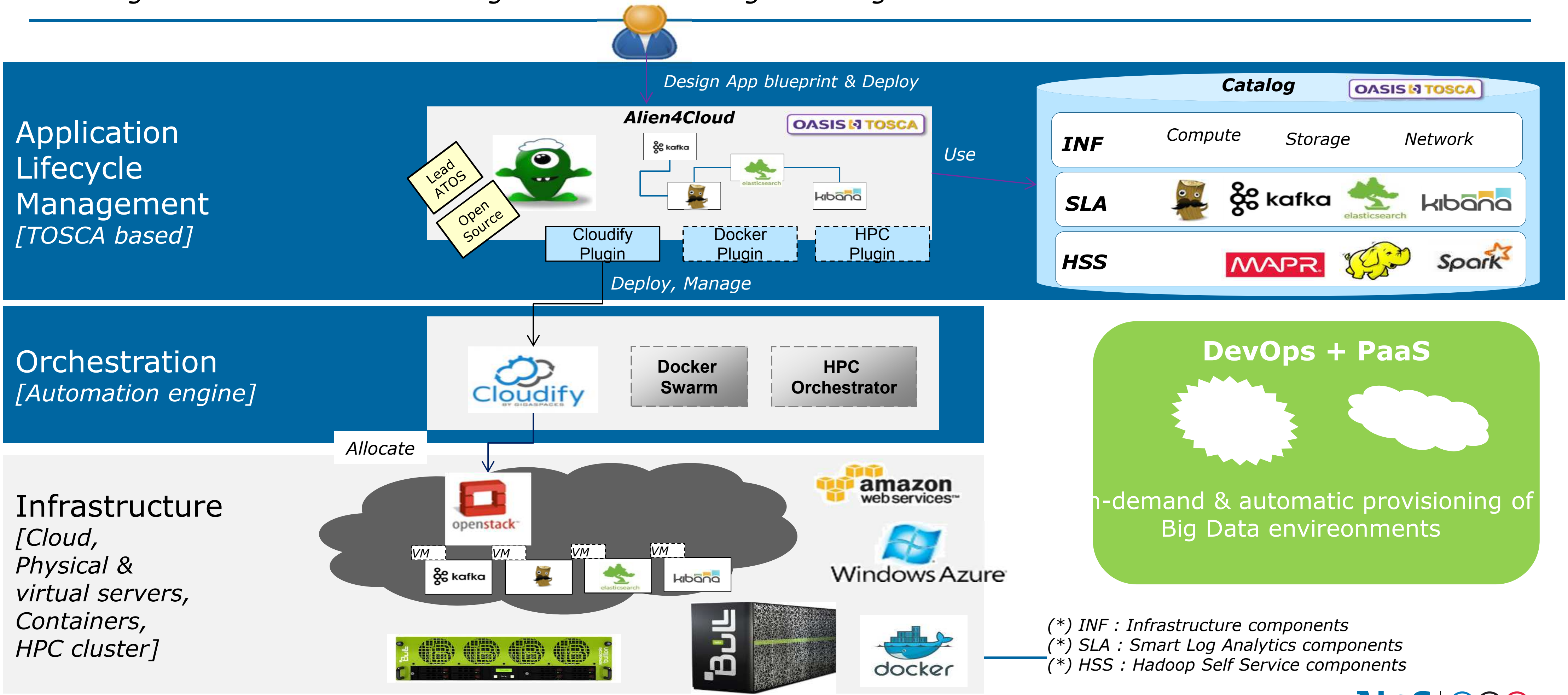
- **Hardware**




**Data Lake & Analytics Factory**

# Big Data Capabilities Framework

*« The right infrastructure for the right data with the right tooling »*




*(*) INF : Infrastructure components*
*(*) SLA : Smart Log Analytics components*
*(*) HSS : Hadoop Self Service components*

# Analytics components

To address end-user needs

**Analytics**

Geo positioning:
*Get geographic perspective*

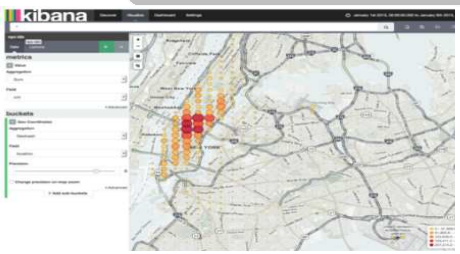
DataViz:
*Create Dashboard and get perspective*

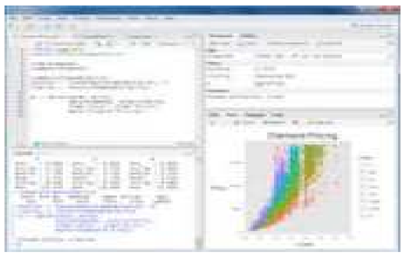
Search:
*Query data efficiently*

Analyze:
*Make sense of machine data*

## Comprehensive software suite

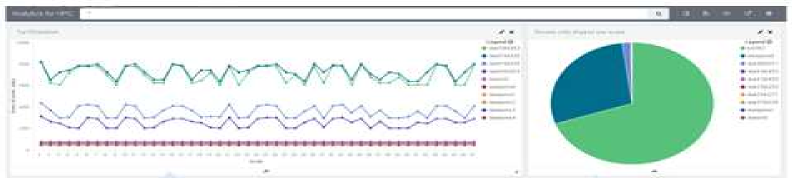
- Most popular open-source, benefit from the best technologies
- Complete integration



## IoT Ready

- SmartLog Analytics provides end to end search and analytics tool to conduct an investigation into machine data/IoT.



*Network traffic evolution with the top 10 equipments that generates the most of the traffic*

*Top 10 equipments with highest dropped cells*

# Data Lake & Analytics Factory

To enable data scientists "magic"

Dev

**Drill down:**
Dig deeper in data stack

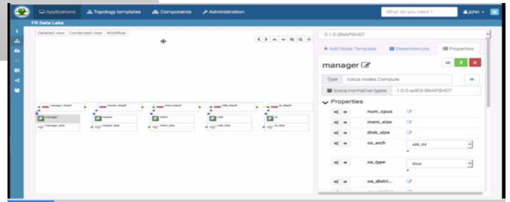
**Collect:**
Connect to many sources (+200)

**Time series:**
Messaging platform to send the time series

**Real-Time:**
Lambda-architecture abilities

## + Studio for App design

- Ability to allow users to upload and create their own datasets/datalake
- Expose, share or resell… your house-made apps



## + Save data scientist time

- Automated tools to manage deployment and lifecycle of Hadoop / data science related tools
- Save time for data scientists and thus… enable them to unleash the value of data

# Sound Data Lake & simple Analytics deployment

To make IT and operations life easier

**Deploy**

**Repository:**
Distributed storage system: data lake

**Deploy:**
On appliance or cloud ressources

**Scale:**
Grow as needed

**Time:**
Deploy in minutes

## Open Appliances

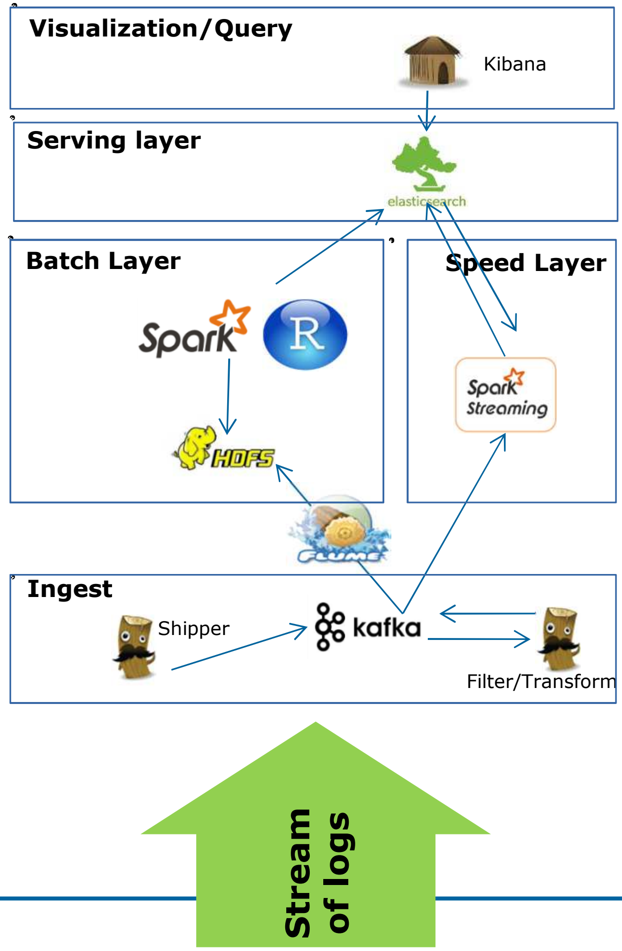
- Benefit from the best technologies
- Choose your Hadoop distribution: MapR, next Hortonworks and later Cloudera
- Open for other component

## Virtualization Capabilities

- Resource utilization: Isolate VM to create multiple data lakes for multiple orgnization (mutli-tenant hadoop), and enfore organizational security boundaries.
- DataCenter efficiency: Create pool of cluster scale-up/scale-out on demand.

# Example of complex analytics workflow deployment (lambda architecture)

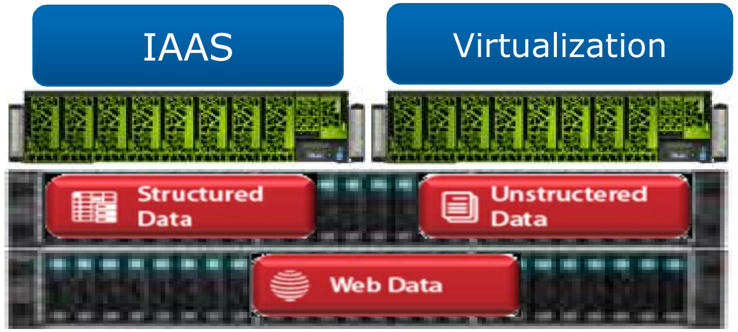
**<u>Outliers detection</u>**:
*throw an alert if the behavior is far from the prediction*




# A robust, enterprise proof architecture

To further push the Data Lake & Analytics advantages

- **Based on bullion server**
  - Enable high **Quality of service**
    - RAS features
    - Blade (I/O & RAM) hot swap/add
  - Built for virtualization
    - Compute density enables better TCO
    - QoS enable better consolidation
    - Deployment of analytics in VMs
  - Scale at will
    - scale-up to add compute/analytics ressources
    - scale-out to add SLA and grow




- **Appliance integration**
  - One stop shop, one stop support
  - Pre-integrated, pre loaded
  - Predictable performance and behavior

# Data lake & Analytics Factory

## Les principaux bénéfices

| | |
|---|---|
| **Flexibilité des ressources et évolutivité** | Capacité à **ajouter facilement** des ressources pour l'analytique (compute) et le stockage *data lake* (stockage - EBOD) |
| **Maitrise de la complexité et réduction des coûts** | **Ressources partagées** pour l'analytique et le *data lake*<br>Architecture et exploitation **simplifiée** |
| **Respect des niveaux de services** | **Virtualisation permet la migration des VMs**<br>**Amélioration de la QoS (RAS, blade add/swap, sensors…)** |
| **Efficacité opérationnelle<br>Capacité Mémoire<br>Performance** | **Facilité de provisionnement et de réallocation des ressources**<br>Catalogues **self service** et intégration complète |

# Le serveur bullion :  sa modularité et son évolutivité linéaire

## S2

**2CPU**

- up to **36 cores**
- up to **3 TB RAM**
- 7 PCI-e
- Active / Passive Power Supply*

## S4

**4CPU**

- up to **72 cores**
- up to **6 TB RAM**
- 14 PCI-e
- Active / Passive Power Supply*
- HW Partitioning

## S8

**8CPU**

- up to **144 cores**
- up to **12 TB RAM**
- 28 PCI-e
- Active / Passive Power Supply*
- HW Partitioning

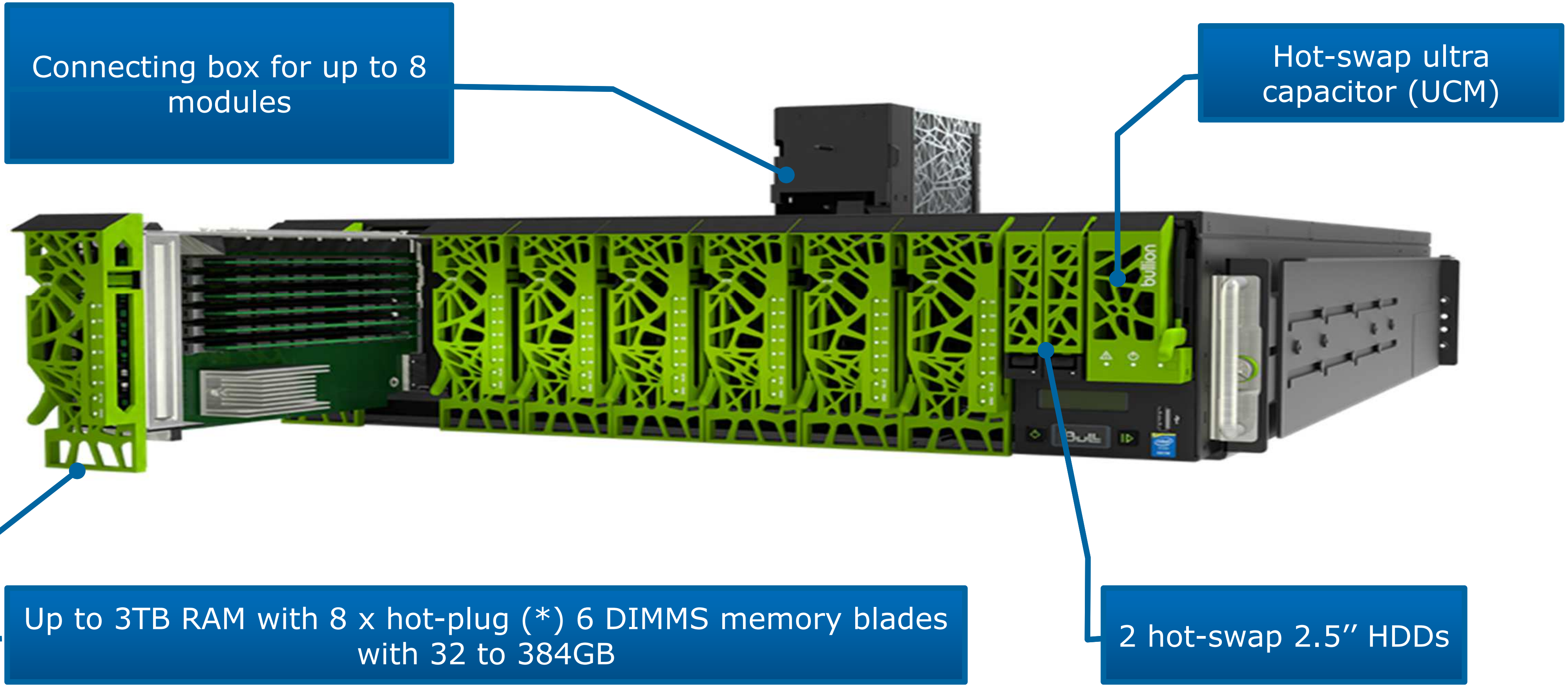
## S16

**16 CPU**

- up to **288 cores**
- up to **24TB RAM**
- 56 PCI-e
- Active / Passive Power Supply*
- HW Partitioning

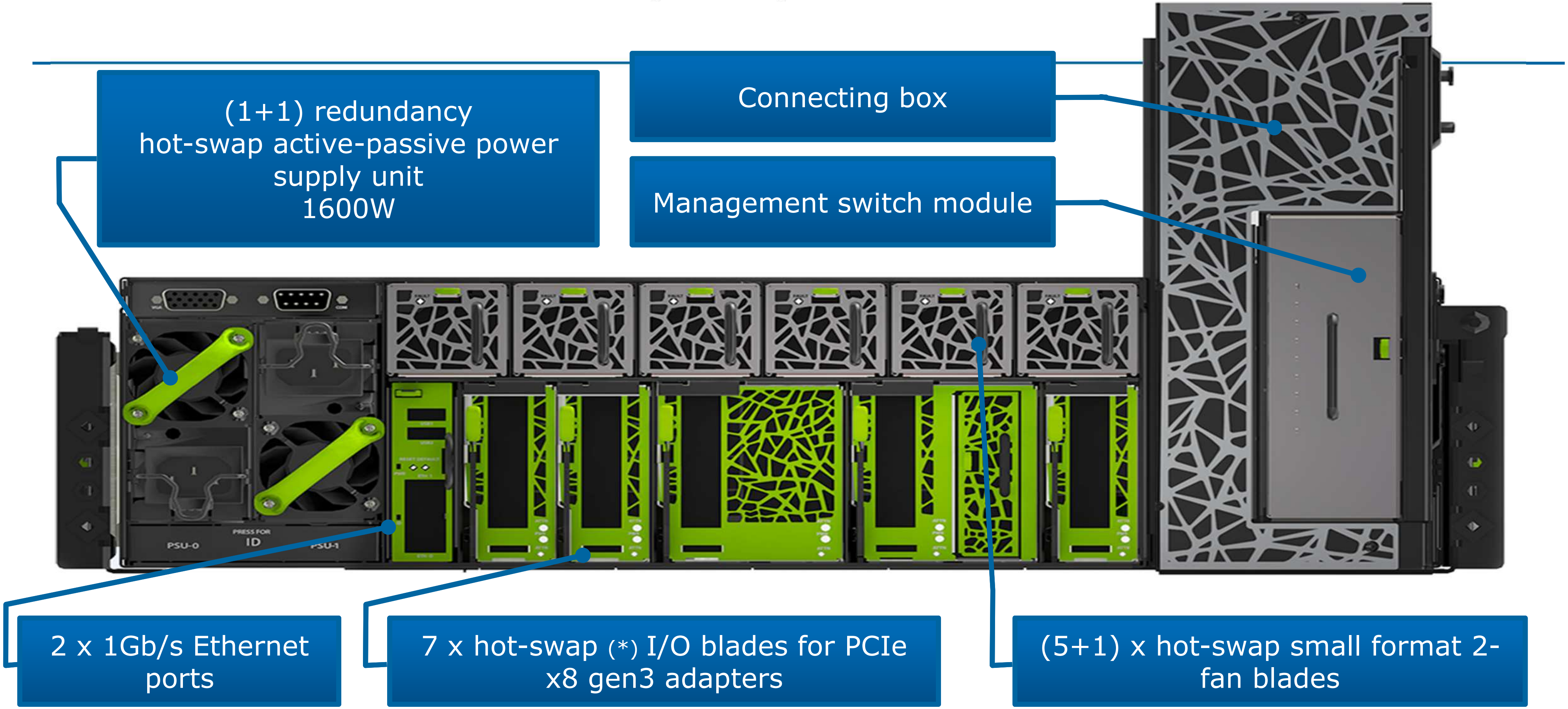
* *optional*

# L’architecture du bullion S




(*): hot-plug memory is depending on OS/hypervisor

# L'architecture du bullion S (suite)

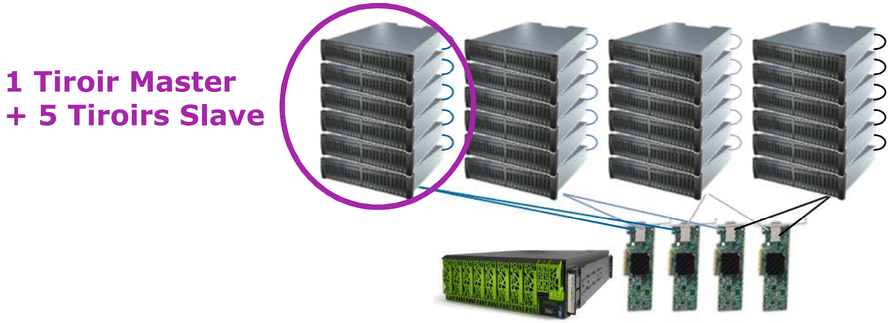



(*): hot-swap is depending on OS/hypervisor

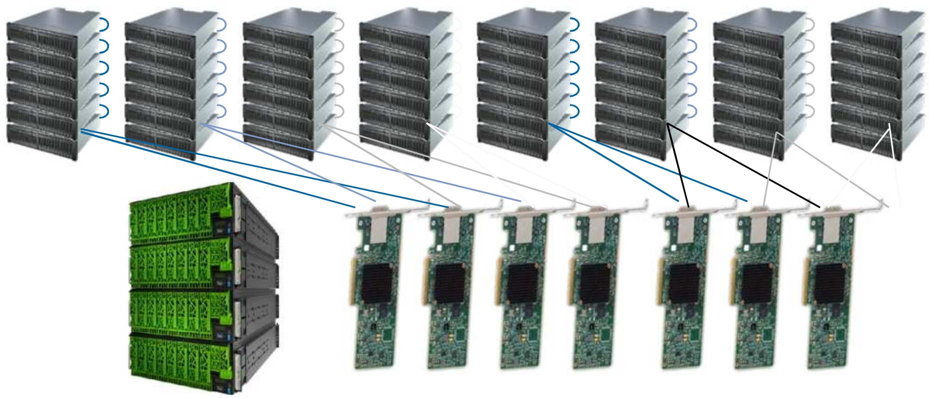
# bullion Storage Box : comment optimiser votre stockage ?

Pour un bullion S2

- jusqu'à 24 modules / **576** drives
- jusqu'à **691To** SAS / **1,15PB** NL-SAS
- Possible mix of SSD, SAS and NL-SAS disks
  (2.5' and 3.5' enclosure in different loops)

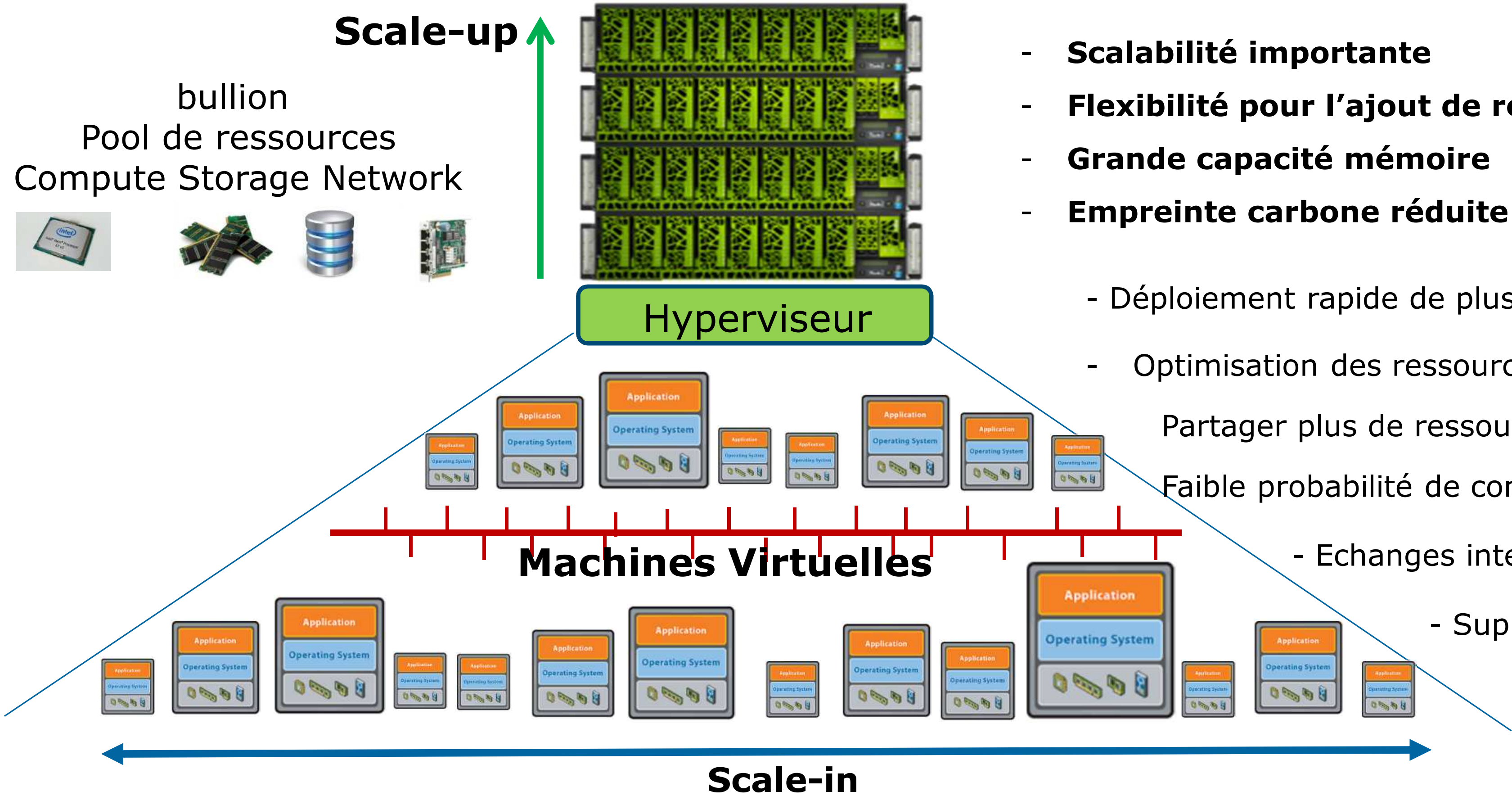



Capacité maximum avec 3.5" : **2,3Po**



**Même performance** que des disques
internes classiques

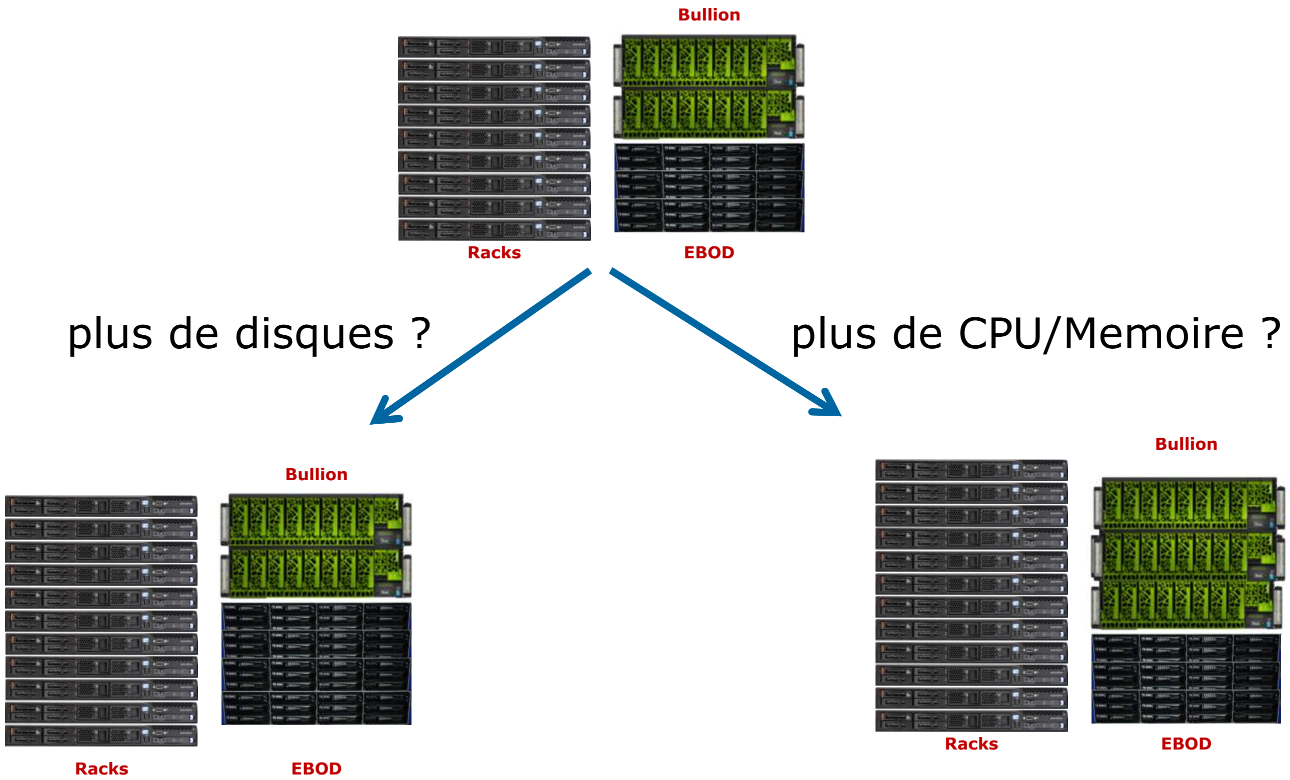
# bullion ou comment faire du Scale-in de VMs au sein une architecture HW Scale-up



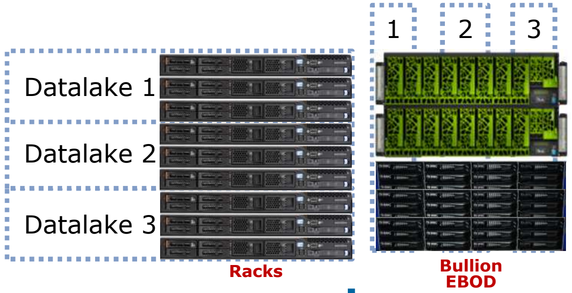

- **Scalabilité importante**
- **Flexibilité pour l'ajout de ressources physiques**
- **Grande capacité mémoire**
- **Empreinte carbone réduite**

- Déploiement rapide de plusieurs 100 de VMs
- Optimisation des ressources :
  - Partager plus de ressources physiques entre les VMs
  - Faible probabilité de contention CPU
- Echanges inter-VMs à faible latence
- Support de « grandes » VMs

# Rack vs Bullion+EBOD (scalabilité)

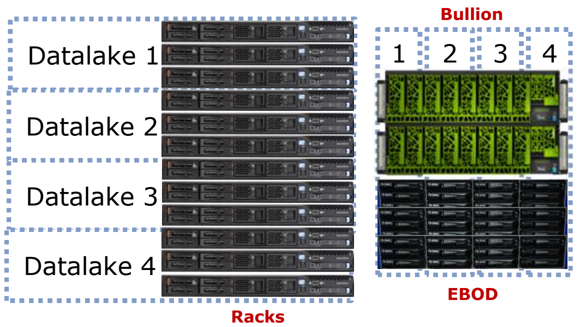


# Rack vs Bullion+EBOD (multi datalake)



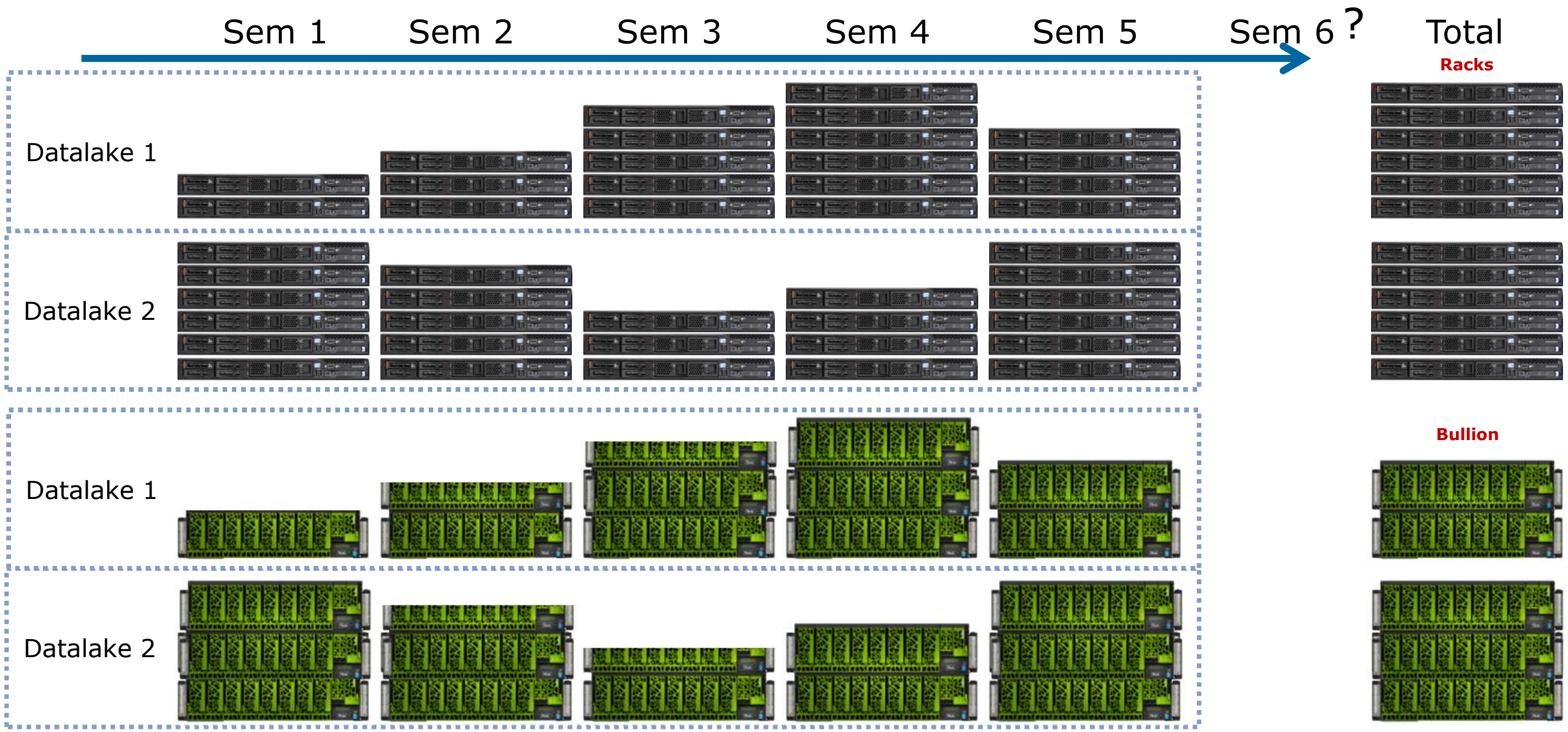

plus de datalake ?

exemple pour la résilience – Nombre de ports 10Gbits ports : 12




exemple pour la résilience – Nombre de ports 10Gbits ports : 4

# Rack vs Bullion+EBOD (flexibilité)
# 2 datalake qui n'utilisent pas toute la puissance en même temps



# Le serveur bullion : la synthèse

- **High End / High Performance :**
  16 CPUs/ 24 TB; ''E7-xxxx–EX'' H.E. Intel CPUs; 11600 SPECint*rate

- **Flexibilité :** 2 modules CPU**,** blades memoire, blades IO, partitionnement HW

- **Scalabilité :** de 2 CPUs/48GB jusqu'à 16 CPUs/24TB, storage box jusqu'à 2.3PB

- **Maintenabilité :** mécanismes blades mémoire & IOs**,** nombreux CRUs, System Management

- **Simplicité :** design basé sur l'expérience de la 1ère génération de bullion & support client / feedbacks du manufacturing
  => contrôleur de nodes (BCS) sur la carte mère, connecting box,…

- **Robustesse :** memory protection, …

- **Attractive Design, « conçu et fabriqué en France »**

# Une démarche d’accompagnement en 3 étapes majeures

**Cadrage : Identification des cas d’usages**

**Mise en place d’une plateforme big data ( « datalab »)**

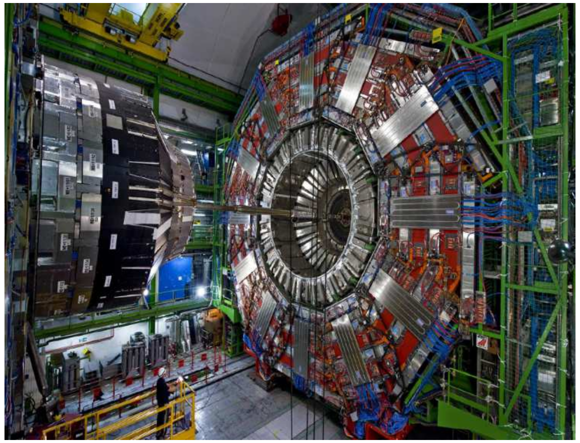
**Réalisation des cas d’usages métiers en mode itératif**

# Vision sur la convergence HPC Big Data

*Pascale Rosse-Laurent*

# Bigdata : Pourquoi maintenant ?
# En HPC « on connaît »

**LHC**



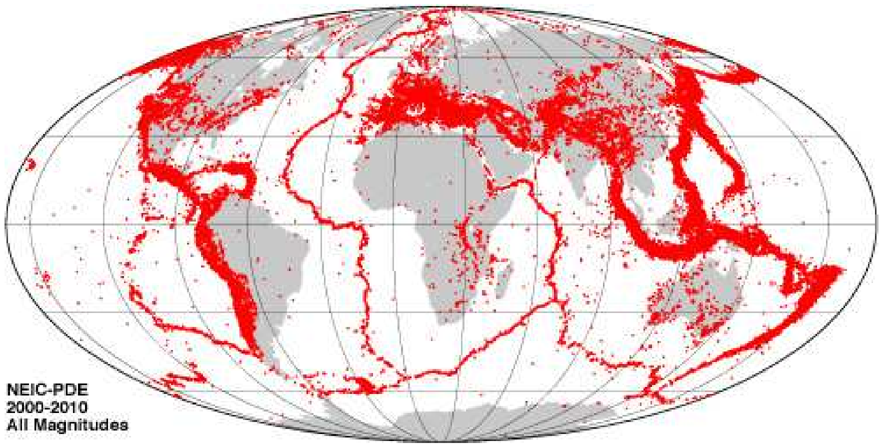
**Square Kilometer Array**

| Discipline | Duration | Size |
|---|---|---|
| HEP - LHC | 10 years | 15 PB/year |
| Astronomy - LSST | 10 years | 12 PB/year |
| Genomics - NGS | 2-4 years | 0.4 TB/genome |

# Earthquake detection

ABUNDANT SEISMIC ACTIVITY …

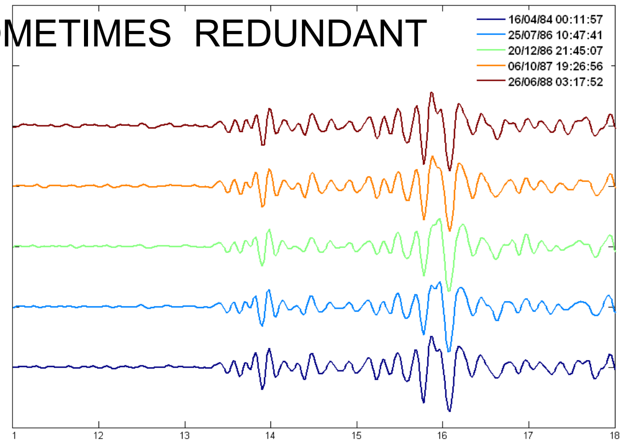
*~50.000 events / year*




**Different kinds of data**

- **Station/sensor metadata**
  - **100~1000 stations**
- **Reference patterns**
  - **20 million, 2GB**
- **Seismic records**
  - **Several million files, ~30TB**
- **Generated data**
  - **Several million files, ~30TB**

SOMETIMES REDUNDANT



# Les raisons d'un changement majeur: des nouvelles données....

- **De nouvelles sources de données en flux continu...**
  - la démocratisation des détecteurs, capteurs dans tous les domaines scientifiques
  - l'émergence d'objets intelligents dans notre environnement personnel (mobile , domotique, Internet of Things )
  - la digitalisation de l'entreprise (Industry 4,0)

- **Des données accessibles**
  - support du web : masse de données images et textes stockées
  - transfert de données fiables et réseaux fiables

mais surtout **des données exploitables :**

# Des données exploitables …

- **de nouvelles technologies matérielles**
  - **des capacités de calculs exponentielles**
    - une variété des entités calcul : CPU ,ARM, FPGA, GPU, PIM
      - Puissance de calcul adapte aux besoins
    - un Flops/watts/m² attractif
      - intégrable de l'objet intelligent au calculateur

  - **des capacités de stockage et des accès « temps réels »**
    - des mémoires vives larges (plusieurs TB) , mémoires non volatiles (2017),eSSD , sata , archive

    ⇒ des capacités d'accès « temps réels » à des données critiques.

# sequana, the open exascale-class supercomputer


- **Open and multi-technology**
  - **CPU, GPU, *FPGA, ARM***
  - **NVM**
  - **...**
- **Ultra dense and scalable**
- **Ultra-energy efficient**
- **Easy administration**
- **Multi-tier data organization**
  - NVM , Local flashs or disks
  - Embedded I/O access
- **Interconnect optimisation**

# Des données Valorisables ….

- **des méthodes traditionnelles optimisées pour l'utilisation des nouvelles ressources**

- **de nouvelles méthodes d'extractions, ETL**
  - **Bases de données no sql , etc…**

- **de nouveaux modèles d'exploitations et nouveaux algorithmes**
  - La distribution du traitement **Map reduce**

    implémentés pour des fouilles de textes

    *puis reconnaissance d'image puis graphes …..*
  - conjointe à la distribution des données :**File système distribué HDFS**
  - **etc …**

- **des environnements offrant des niveaux d'abstraction**
  - facilitant l'accès et la manipulation des données

# L’analytique temps réel versus Batch

# Les enjeux de l’analytique temps réel

**L’analytique temps réel, nécessite que soient adressés :**

- **PUISSANCE DE CALCUL**
  - Toujours fournir la **capacité de calcul nécessaire au temps réel**
  - Faire face à la croissance du volume de données (règlementaire et structurelle)
    - La Future Review of Trading Book (FRTB) va augmenter le nombre de scenarios de stress : prise en compte des profils de liquidité de 10 à 250 jours
- **LATENCE**
  - Pouvoir accéder très rapidement aux données : **du jour et en temps réel**
  - Pouvoir réaliser calculs multi dimensionnels et agrégations non linéaires
  - D’où nécessité de stocker et traiter la donnée **en-mémoire**
- **VOLUME DE DONNEES**
  - Augmentation structurelle du volume de données (20% d’augmentation du volume des données structurées selon les analystes)
  - Impact règlementaire du volume de données
    - La FRTB va multiplier les volumes (pour décomposer chaque vecteur de P&L par type de risque)

**En synthèse: besoin d’infrastructure disposant de grandes capacités de mémoire et de puissance de calcul toujours adaptée**

# Prescriptive Analytics

- **Unified architecture**
  **Based on Lambda architecture**

  **Full stream-based architecture**

  **+ Batch based + microbatch**

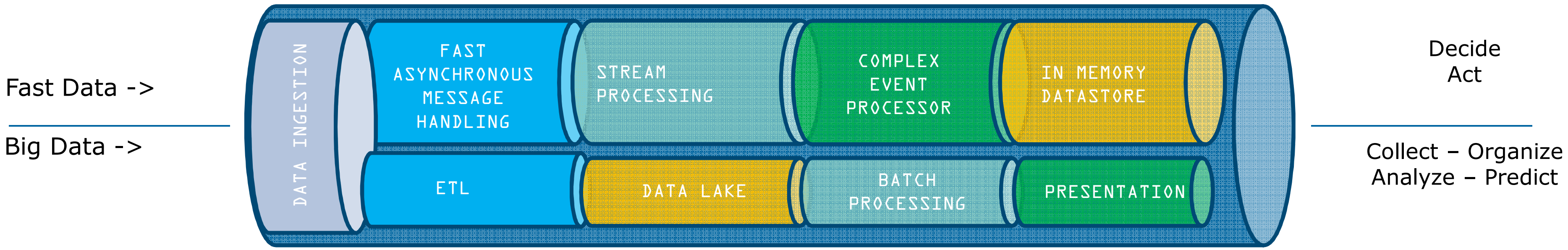
**Technology Enablers for Prescriptive Analytics**

Real-time data stream processing

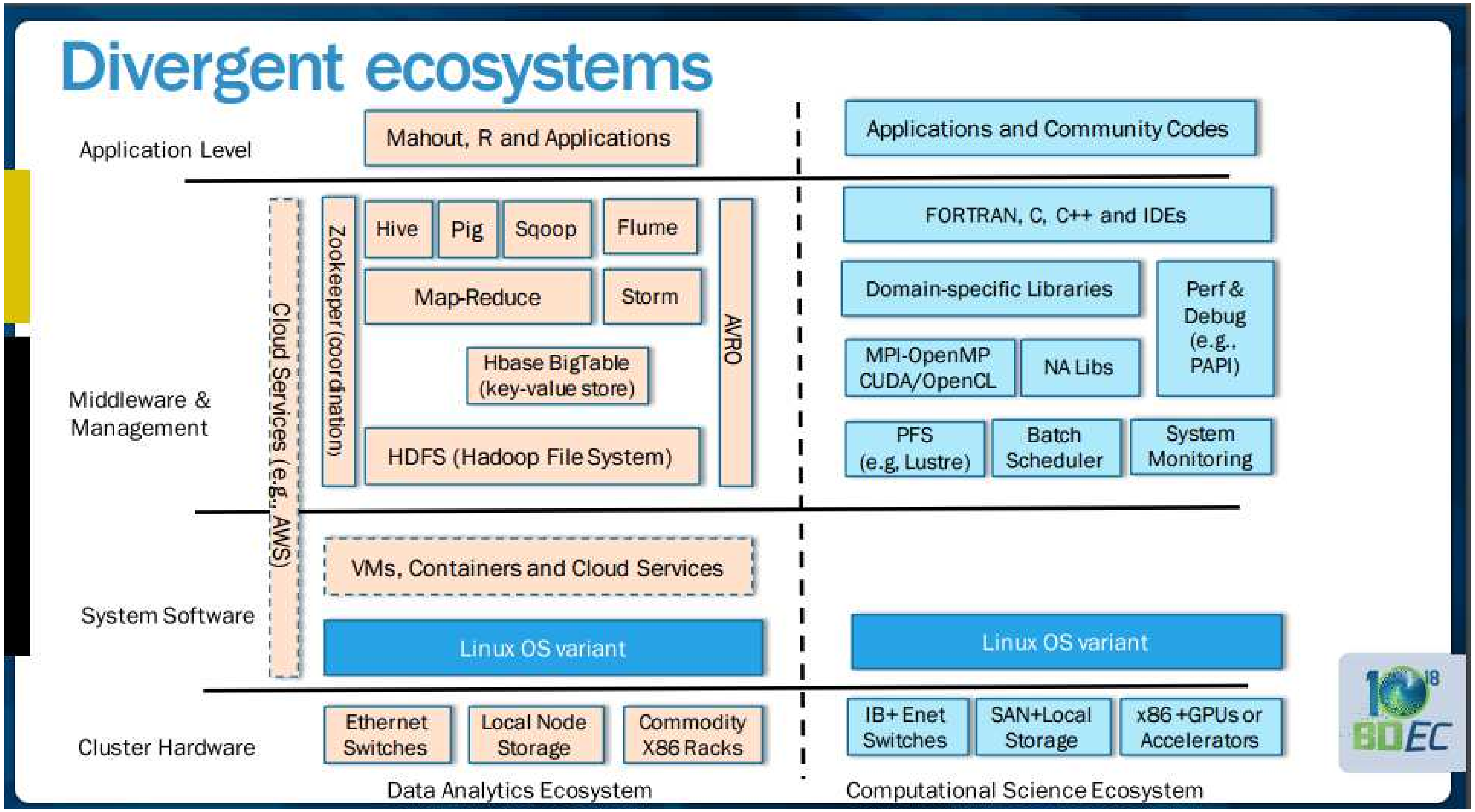
In-memory computing

Distributed analytics framework

Performant Machine-To-Machine communication



# Performance versus flexibility and portability



# ezHPC core : two phases approach

- **<u>Phase 1</u> :**

**Leverage existing computing resources  with optimized deployment of Big Data components on HPC cluster for real-time prescriptive & cognitive analytics**

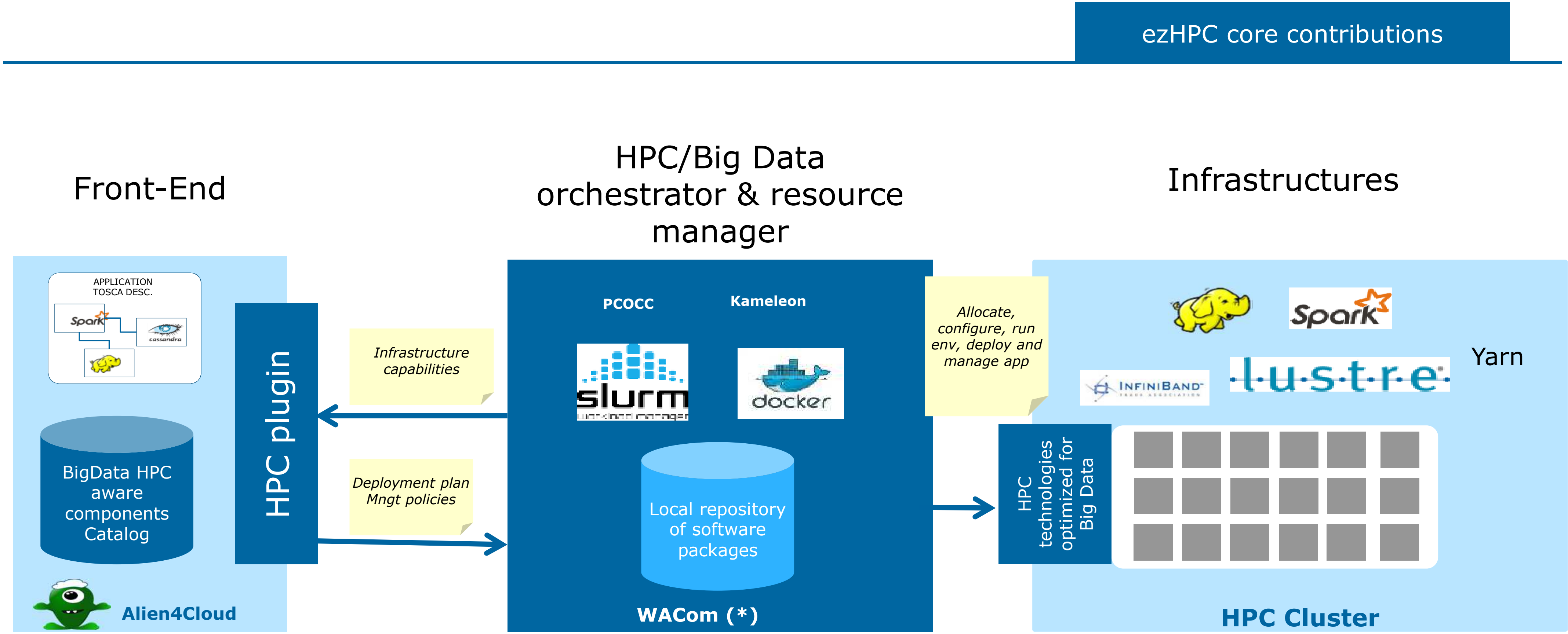
  - Enable « HPC-awareness » of Big Data components
  - Add HPC plug-in to application life cycle management tool
  - Coordinate resource allocation with the usual supercomputer tools
  - Guaranty seamless integration with HPC workloads

- **<u>Phase 2</u> :**

**Extend cloud infrastructures capabilities to optimally integrate HPC technologies and to burst on HPC infrastructures**

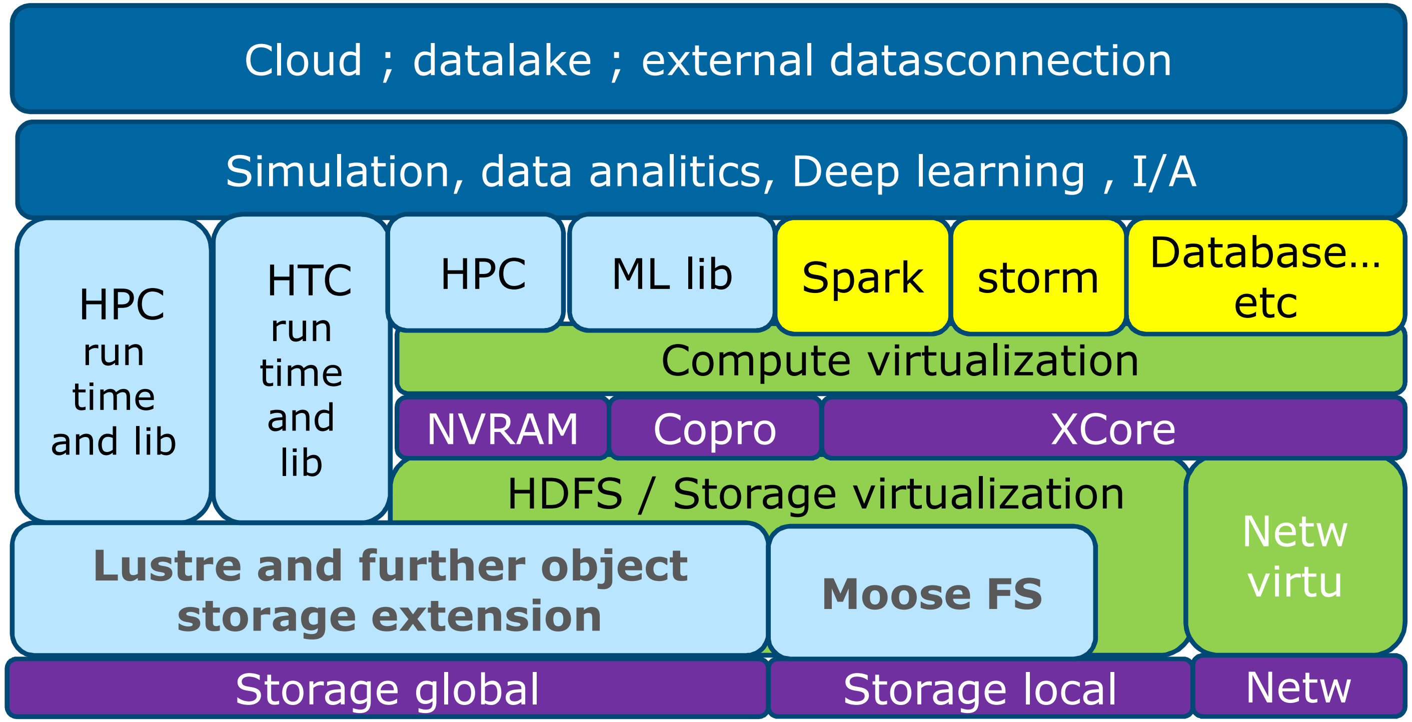
  - Support HPC technologies such as NVM, FPGA, GPU, fat-nodes in cloud infrastructures
  - Provide gateway for data import
  - Provide a unique Storage / File System space managing simultaneously Big Data, HPC data, other data
  - Develop a unified management including fined-grained resource management and on demand provisioning

# Phase 1: Architecture view




(*)WACom: Workload Adaptable Computing

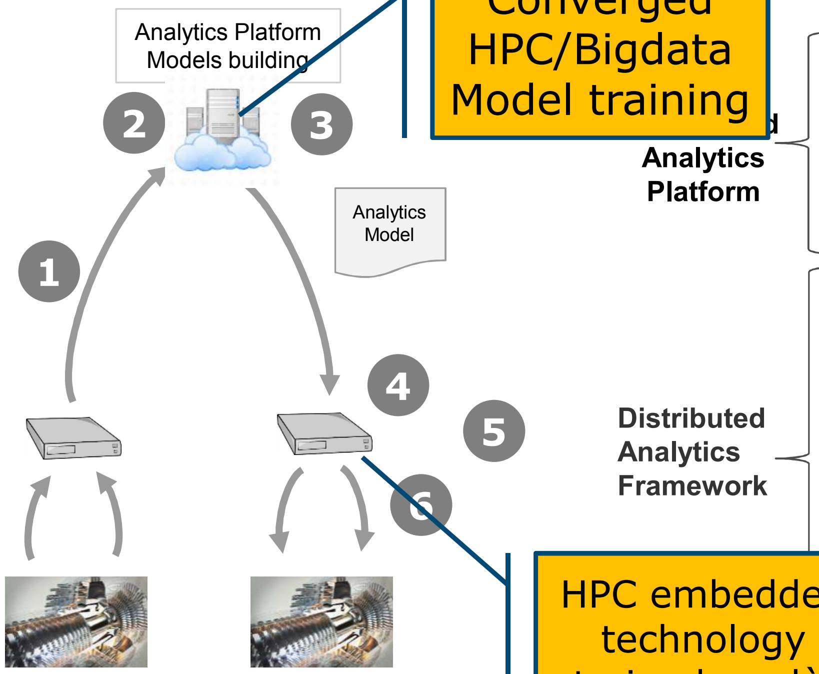
# Phase2: La convergence d'infrastructure matérielle et logicielle



# Les technologies et solutions de simulations s’intègrent dans l’entreprise de demain

*25 billion objects to be connected to the Internet by 2018 will generate vast quantities of data. New architectures are necessary to collect, aggregate and process data at IoE hubs, before sending only relevant data to the main computing centers.*

Converged HPC/Bigdata Model training




HPC embedded technology trained modèl

1. Sensors data are uploaded to the central analytics laboratory
2. Analytical models are defined and tested by the data scientists, and can be applied centrally for diagnostic and surveillance
3. Analytical models are translated into a standard exchange format, encrypted and digitally signed
4. Certified analytics streaming engines are integrated in the local control/command systems
5. The models are decrypted and validated
6. The trusted model is executed locally, in the local control/command system

**En**

"Increase focus on ETL Science needs data,

in the right place and format"

- Optimiser l'intégration des données existantes
  - bases relationnels internes,
  - bases externes ,
  - les fichiers structurés etc….
  - les formats de données métiers
- nécessaire d'étudier  la prise en compte des formats des données et de leurs organisations

"comprendre les phases des processus de traitement  "

# Scheduling

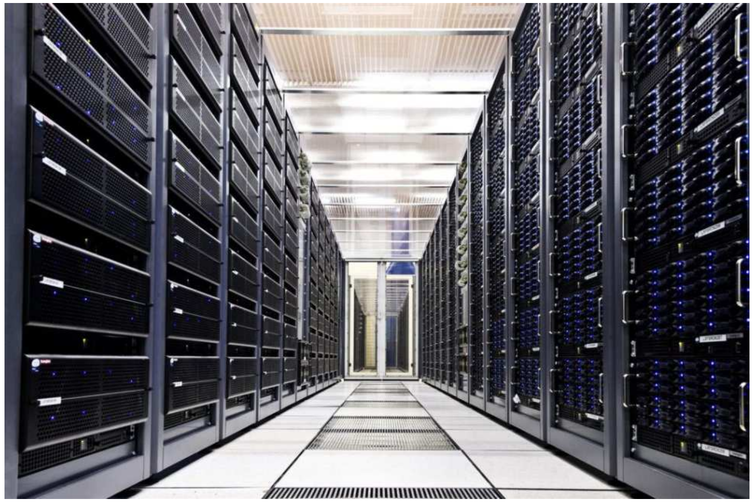
*Michael Mercier*

# HPC World

- Global computation driven resource manager
  - Managed essentially finite time jobs
  - Static resource assignment
  - Full cluster utilization (waiting queue)
- Best available hardware
  - Fast interconnect (infiniband, BXI)
  - Lots of memory
  - State of the art processors
- Datastore : dedicated parallel filesystem
  - Lustre, GPFS, …
  - Few (or absent) local storage
- Traditional usage
  - MPI applications
  - Simulations
  - Ad-hoc applications

# Big data world

- Data driven resource management
  - Big amount of data (from GB to PB)
  - Data aware job placement
  - Memory is more important than computation
- Datastore: distributed filesystem
  - No dedicated hardware, use nodes' local storages
  - HDFS, Hbase, Casandra, Elasticsearch, ...
- Made to scale-out on large cluster
  - Commodity hardware or virtualized compute
  - Fault-tolerant
  - Dynamic resource management
- Traditional usage
  - Map reduce
  - Graph processing
  - Machine learning
  - Data quering



# Big Data on HPC makes sense

High Performance Data Analytics (HPDA)

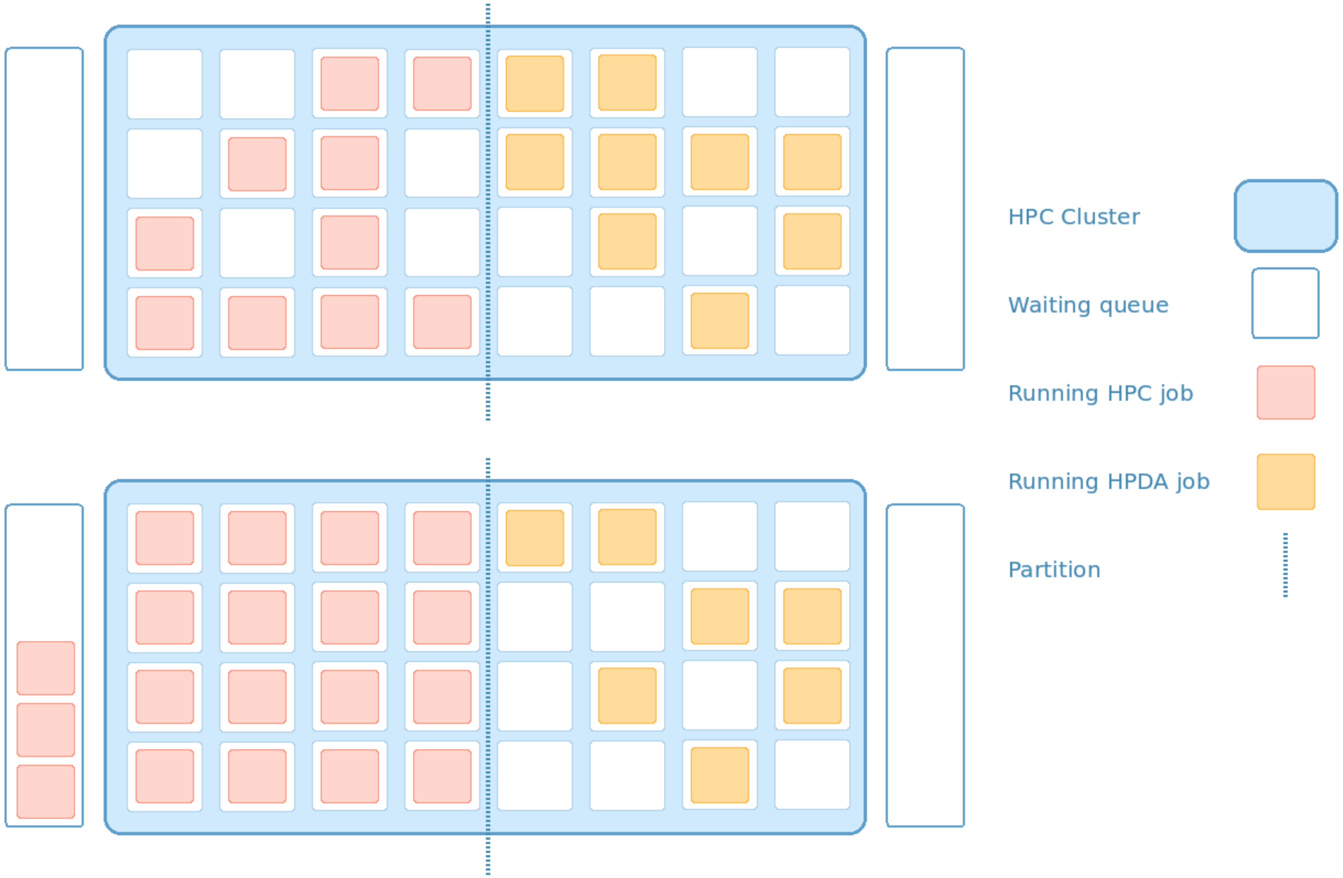
- **More and more data** to manage on HPC application
- Big Data workload can leverage **better hardware**
- Benefits of **big users and developers communities**
- Brings **dynamic resource management** and **fault-tolerance** applications to HPC

**We have two cluster management systems:**

**How to make them interact cleverly?**

# Why we need advanced scheduling?

## What is happening today



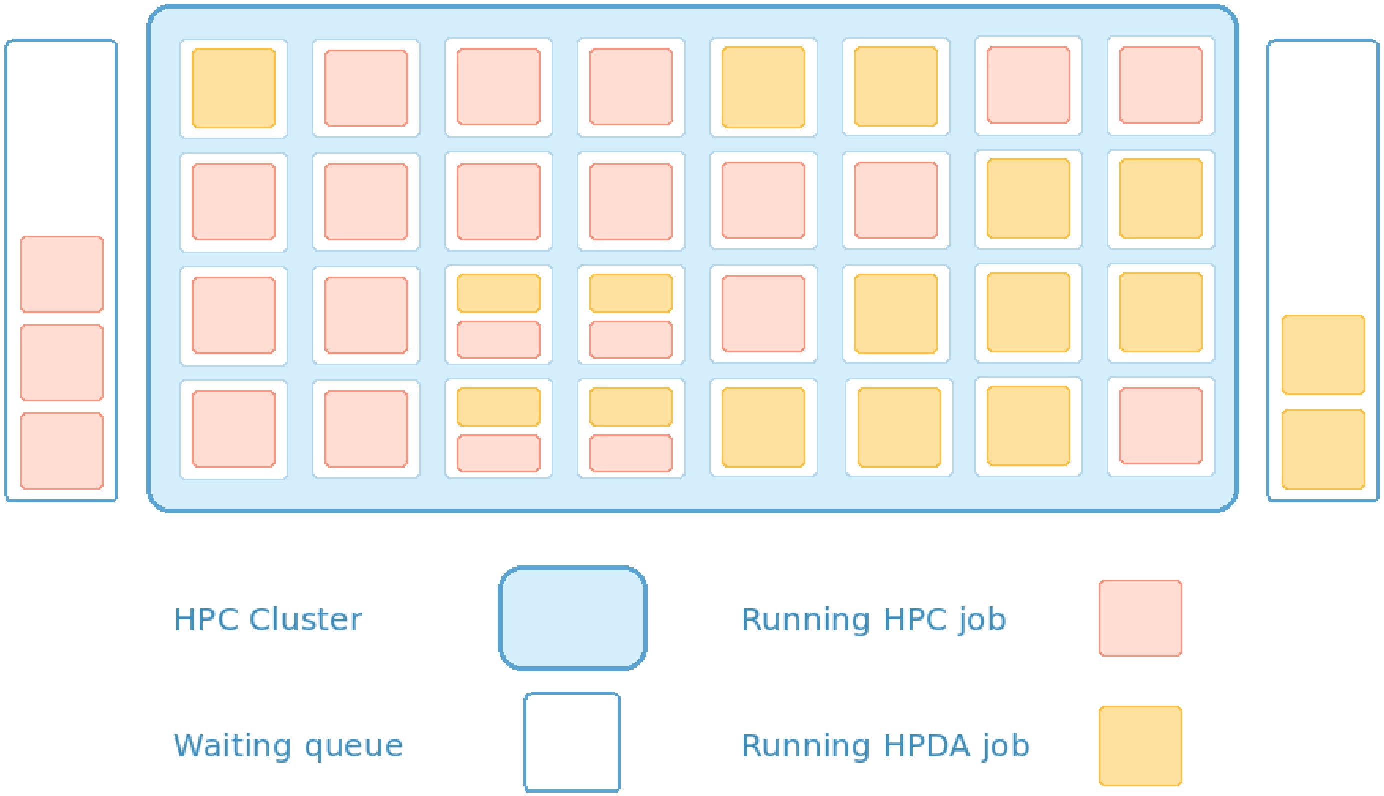

- **Static cluster partitioning**
  - Simple workload discrimination
  - No node sharing
  - Static distributed filesystem

- **Leads to resources waste!**
  - When one queue is filling up
  - Available nodes on the other partition are not used

# Why we need advanced scheduling?

What we want to achieve

- **Full cluster utilization**
  - no matter the share of different workloads
- Dynamic resource allocation
  - no static partition
- Mixed workload on the same resource
  - Co-scheduling make sense for example if:
    - one application IO bound
    - one application is CPU bound



# Different approaches for scheduling convergence

Batch submission

Some terms:

- Resources and Jobs Management System (**RJMS**)
  - The HPC workload management tool
  - Composed of:
    - a master node that takes the decisions
    - a daemon on each nodes to run and control the jobs
  - Examples: SLURM, OAR, PBS, …

- Big Data Analytics Framework (**BDAF**)
  - Big Data Analytics management tool
  - Similar to RJMS but for Big Data
  - Examples: Hadoop, Spark, Flink, …

- Batch submission: Submit one job to the RJMS
  1. Install and configure the BDAF
  2. Install and configure the distributed filesystem
  3. **Stage in** the input data into the distributed filesystem
  4. Submit to the BDAF the user application
  5. **Stage out** the results data from the distributed filesystem

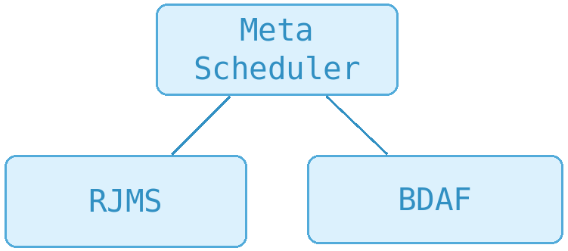
- **Not efficient** because of data movement
  - Stage in and out can take a really long time

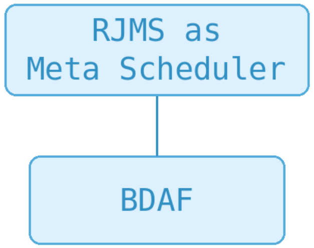
- Possible solutions:
  - Use the parallel filesystem instead of the distributed filesystem
  - Include staging into the scheduling policy

# Different approaches for scheduling convergence

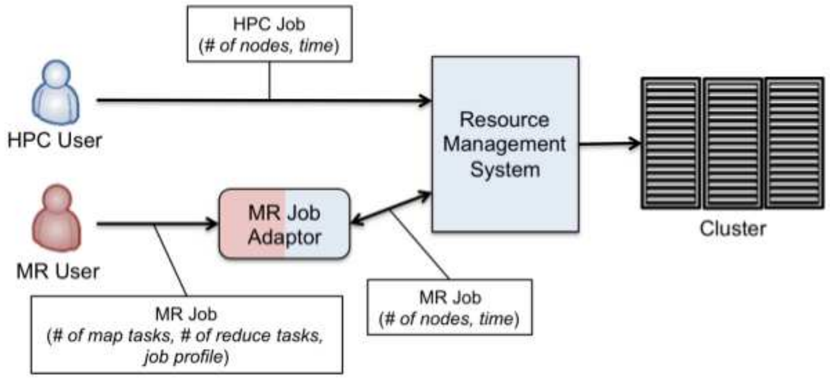
## Two-level scheduling / Scheduler adaptor

- Two-level scheduling: Use an upper level of scheduling
  - dynamically share the resources between workloads
  - use a third tool or a modified RJMS
  - need to implement an API to talk to the meta-scheduler
  - Example:
    - Mesos
    - Univia Grid Engine




- Scheduler adaptor: Adapt Big Data scheduler to talk to the RJMS
  - make the two management system cooperate
  - RJMS do scheduling decisions triggered by the BDAF
  - Example:
    - Intel Enterprise Edition for Lustre: YARN adaptor
    - Map Reduce adaptor from “Scheduling MapReduce Jobs in HPC Clusters.” Neves, Ferreto, De Rose, Euro-Par 2012

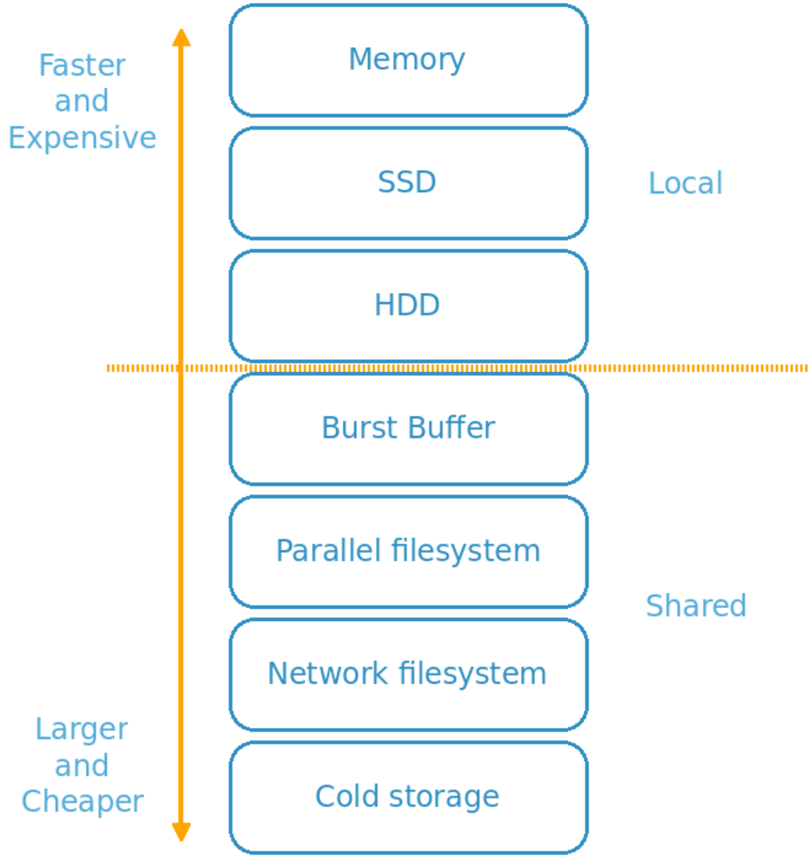


# The data management problem

- Data workflow
  1. Ingest new data
  2. Compute results
  3. Query previous results
  4. Use previous results to compute new results

⇒ We need **data persistence**

- Problem:
  – traditional HPC infrastructure do not provide local data persistence
- Solutions:
  – use the parallel instead of distributed filesystem
    - Example: LUSTRE instead of HDFS
    - From local to shared: Possible congestion problem
  – use local storage for better performance
    - Need data persistence for local storage
    - Use staging can be better sometimes



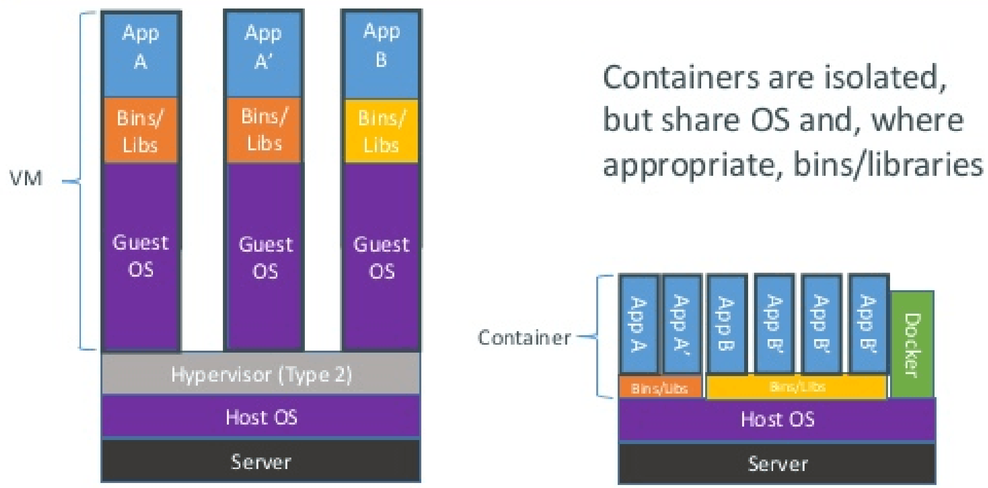

⇒ **RJMS need to manage data as resources**

# The software environment problem

- HPC software environment
  - Bare metal (no virtualization)
  - Statically setup by the administrators
  - RedHat based Linux (mostly)
  - Only userspace installation

- Big data software environment
  - Virtualized or bare metal
  - Dynamically setup by the users
  - Multiple Linux distribution support
  - Use a specific runtime version:
    - Mostly Java (JVM)

- Solution: **User defined software environment** for HPC
  - Using Linux containers **Docker**
    - Close to bare metal performances
    - Permit network isolation
    - **Problem**: not secure Docker daemon on each node
  - Using Linux containers images and **Shifter**
    - Real bare metal performance
    - No network isolation
    - No daemon
    - SLURM integration



# Futur works

Any questions?

- Determine Big Data workload characteristics
  - Running Big Data benchmarks
  - Profiling I/O, CPU, Network
  - **We are looking for real use case and workloads ☺**
- Test the different scheduling convergence approaches
  - Improve the batch scheduler simulator Batsim to fit our needs
    - Design and implement an IO model
  - Compare the different approaches
  - Determine the more promising one
- Implement a Proof of concept
  - Features
    - Dynamic resources allocation
    - Data awareness: staging and data locality
    - co-scheduling
- Design and run experiments on real conditions

# Stockage Hadoop over Lustre

*Eric Morvan*

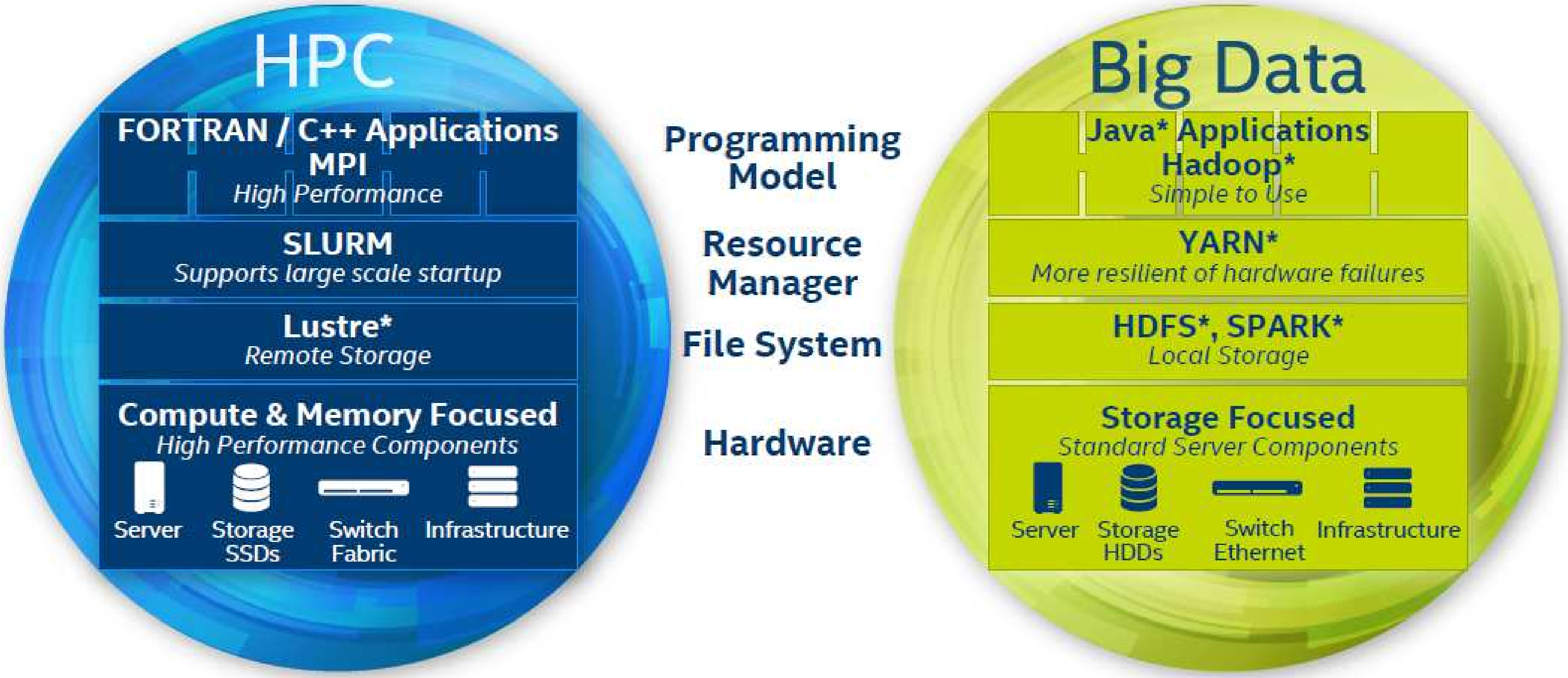
## Deploying Hadoop on Lustre.
## Overview

- **Challenges**
  - Data movements and storage consume 50%-70% of total system power (ScidacReview 1001)
  - Big Data problems are large, HPC resources can help brings more compute power
  - HPC storage can provide more capacity and data throughput

- **For which purpose ?**
  - For customers who would like to use Big Data applications on their HPC cluster
  - Sharing infrastructure
    - Deploy Hadoop/Spark on HPC cluster
    - Transparent deployment for the user through the resource manager (Slurm)
    - Fast storage access (IB) and fully supported

- **Why Lustre ?**
  - Lustre is POSIX compliant and convenient for a wide range of workflows and applications, while HDFS is a distributed object store designed narrowly for the write once, read many Hadoop paradigm

# Deploying Hadoop on Lustre.
Different Systems (Today)

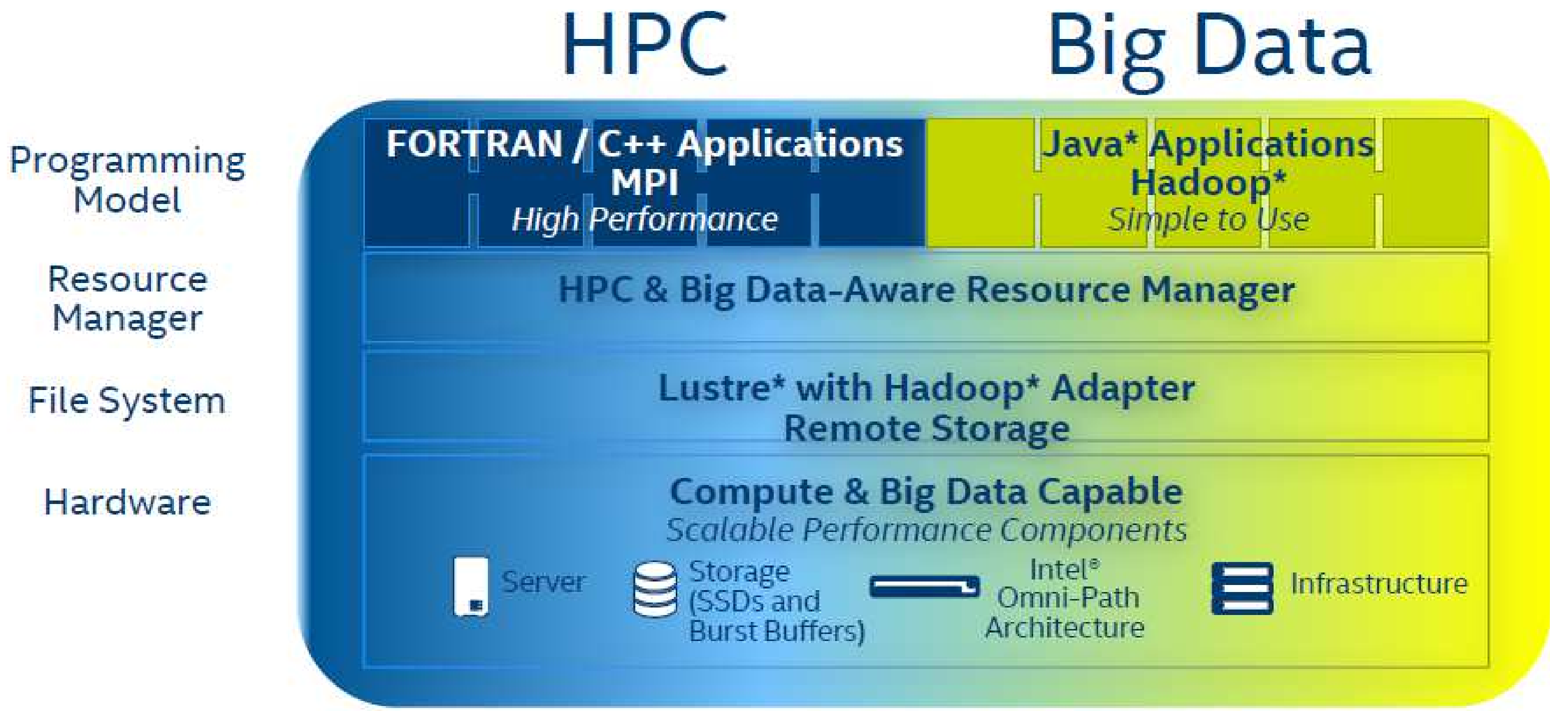


# Deploying Hadoop on Lustre.
# HPC vs Big Data

Differences

| LUSTRE | HDFS |
|---|---|
| Computations share the data | Data moves to the computation |
| Use Infiniband to access data | Uses HTTP for moving data |
| No data replication | Data replication (3X typical) |
| Centralized storage | Local storage |
| POSIX Compliant | Non-POSIX compliant |
| Widely used for HPC applications | Widely used for MR applications |
| Allows backup/ recovery using existing infrastructure (HSM) | Used during shuffle MR phase |

## **Deploying Hadoop on Lustre.**
Converged Architecture for HPC and Big Data



# Deploying Hadoop on Lustre.
# Lustre in the field

- High performance file system used for large-scale cluster computing
- Scalable to ~10 000 client nodes, ~10 petabytes, ~1 terabytes/s IO throughput

- ▸From Top500 fastest supercomputers in the world, since June 2005, used by
  - at least half of the top10
  - more than 60% of top100

- Top sites with Lustre (Top500 November 2014)
  - 1.Tianhe-2, National Supercomputing Center
  - 2.Titan, Oak Ridge National Laboratory
  - 3.Sequoia, Lawrence Livermore National Laboratory
  - 4.K computer, RIKEN Advanced Institute for Computational Science
  - 6.Piz Daint, Swiss National Supercomputing Center
  - 26.Occigen, GENCI-CINES, delivered by Bull/Atos
  - 33.Curie, CEA/TGCC-GENCI, delivered by Bull/Atos

# Deploying Hadoop on Lustre.
# Lustre key components

- Clients
  - sees a unified namespace
  - standard POSIX semantics
  - concurrent and coherent read and write access

- Management Server (MGS)
  - stores configuration information of the file systems
  - contacted by Lustre targets when they start
  - contacted by Lustre clients when they mount a file system
  - involved in recovery mechanism
  - not a critical component

# Deploying Hadoop on Lustre.
# Lustre key components

- Object Storage server (OSS)
  - exports Object Storage targets (OSTs)
  - provides an interface to byte ranges of objects for read/write operations
  - stores file data
- Metadata server (MDS)
  - exports Metadata targets (MDTs)
  - stores namespace metadata: filenames, directories, access permission, file layout
  - controls file access
  - tells clients the layout of objects that make up each file
- OSTs and MDTs
  - uses a local disk file system: ldiskfs (enhanced ext4), ZFS
  - based on block device storage
  - usually hardware RAID devices, but works with commodity hardware

# Deploying Hadoop on Lustre.
# Resource Manager - Workflow

## HDFS

1. nodes allocation via the resource manager
2. SetUp Hadoop env on nodes
3. Data copy on HDFS File System
4. run job
5. Retrieving resulting files from HDFS
6. Uninstall Hadoop
7. Release nodes via ressource manager

## LUSTRE

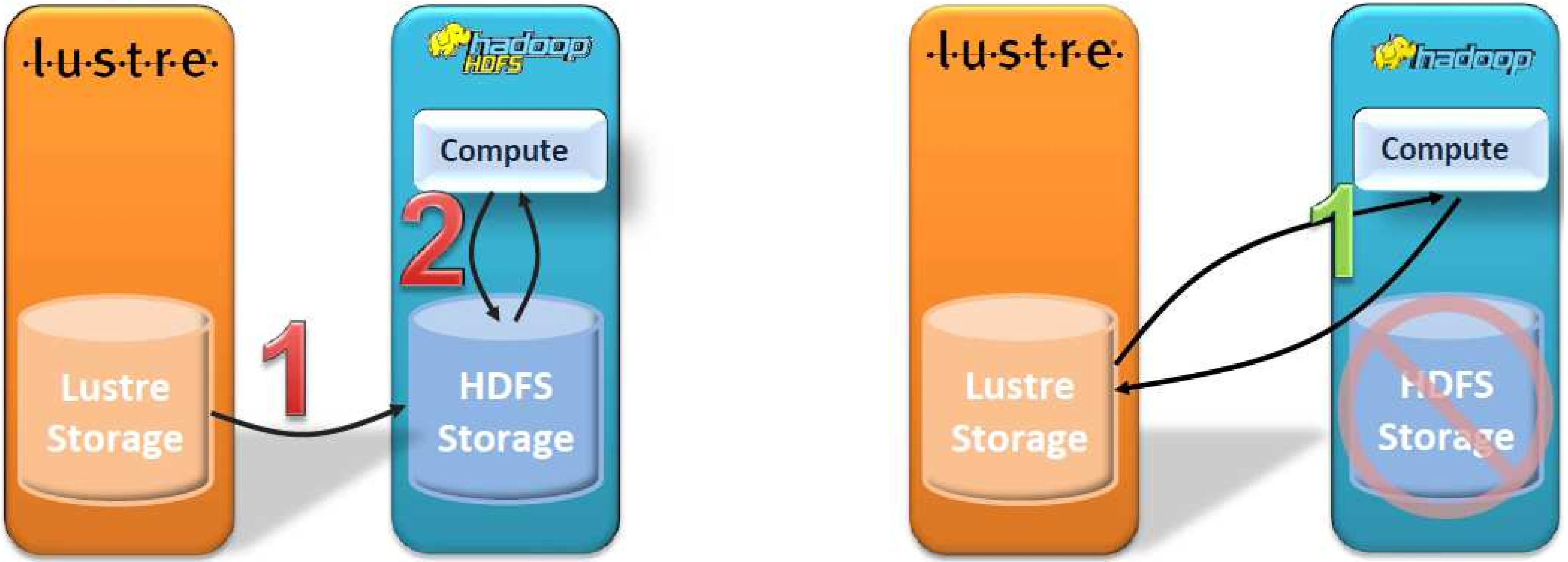
1. nodes allocation via the resource manager
2. SetUp Hadoop env on nodes
3. run job
4. Uninstall Hadoop
5. Release nodes via ressource manager

# Deploying Hadoop on Lustre.
# Workflow

## Ingestion of Data before Analysis

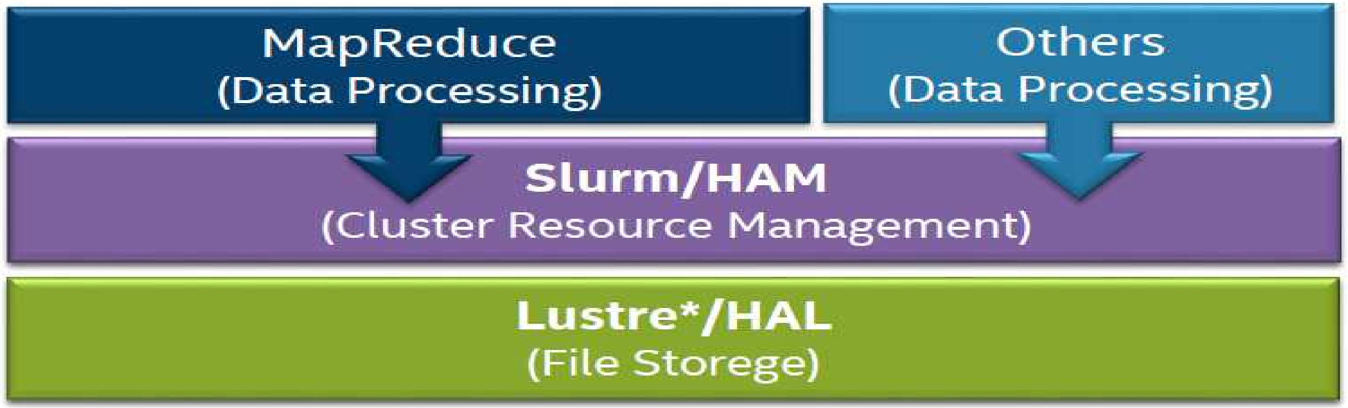
1) Transfer (and usually duplicate) data on HDFS
2) Analyze data

1) Analyze data on Lustre



# Deploying Hadoop on Lustre.

HAM & HAL




**HPC Adapter for Mapreduce/Yarn**

- Replace YARN Job scheduler with Slurm
- Plugin for Apache Hadoop 2.3 and CDH5
- No changes to applications needed
- Allow Hadoop environments to migrate to a more sophisticated scheduler

**Hadoop* Adapter with Lustre***

- Replace HDFS with Lustre
- Plugin for Apache Hadoop 2.3 and CDH5
- No changes to Lustre needed
- Allow Hadoop environments to migrate to a general purpose file system

# Deploying Hadoop on Lustre.

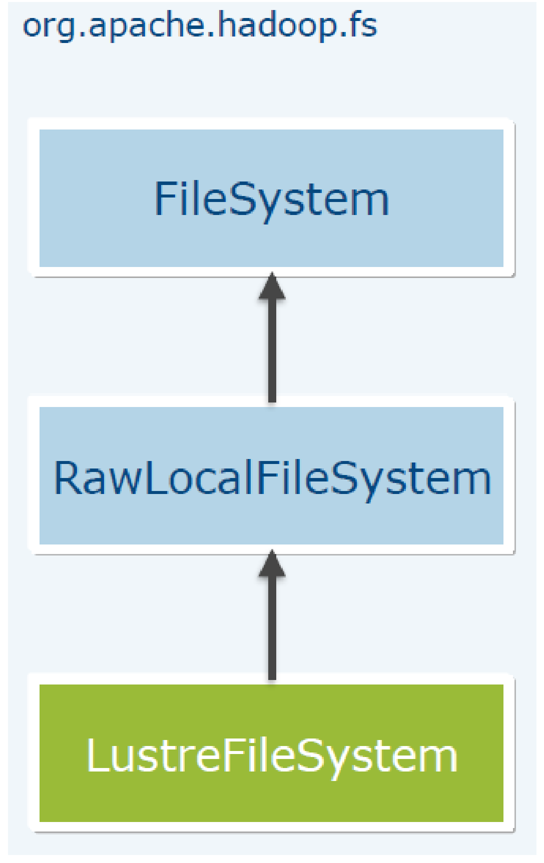
Hadoop over Intel EE for Lustre

- Hadoop uses pluggable extensions to work with different file system types

- Lustre is POSIX compliant
  - Use Hadoop's built-in LocalFileSystem class
  - Uses native file system support in Java

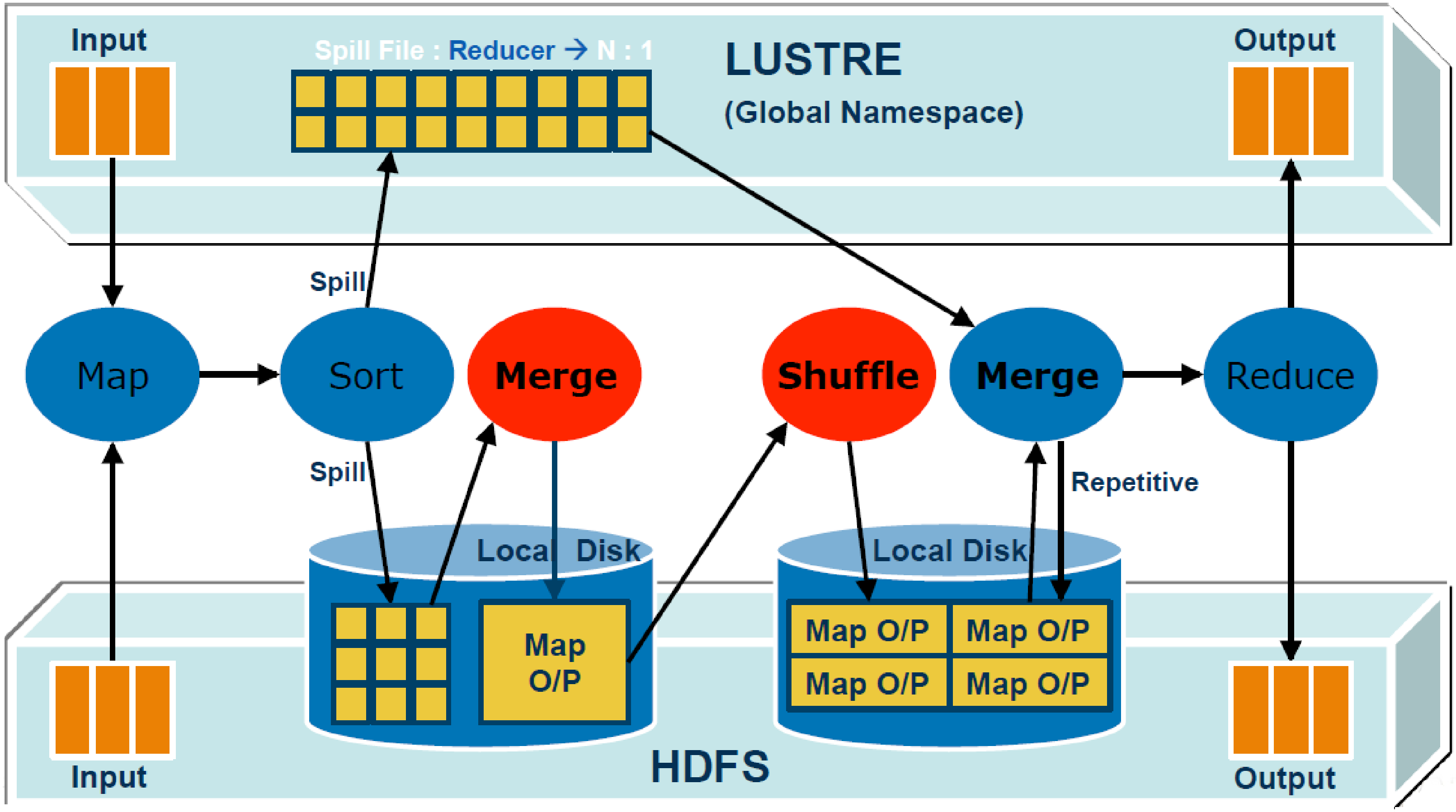
- Extend and override defaut beahavior: LustreFileSystem
  - Defines new URL scheme for Lustre – lustre:///
  - Controls Lustre striping info.
  - Resolves absolute paths to user-defined directory
  - Leaves room for future enhancement

- Allow Hadoop to find it in config files.



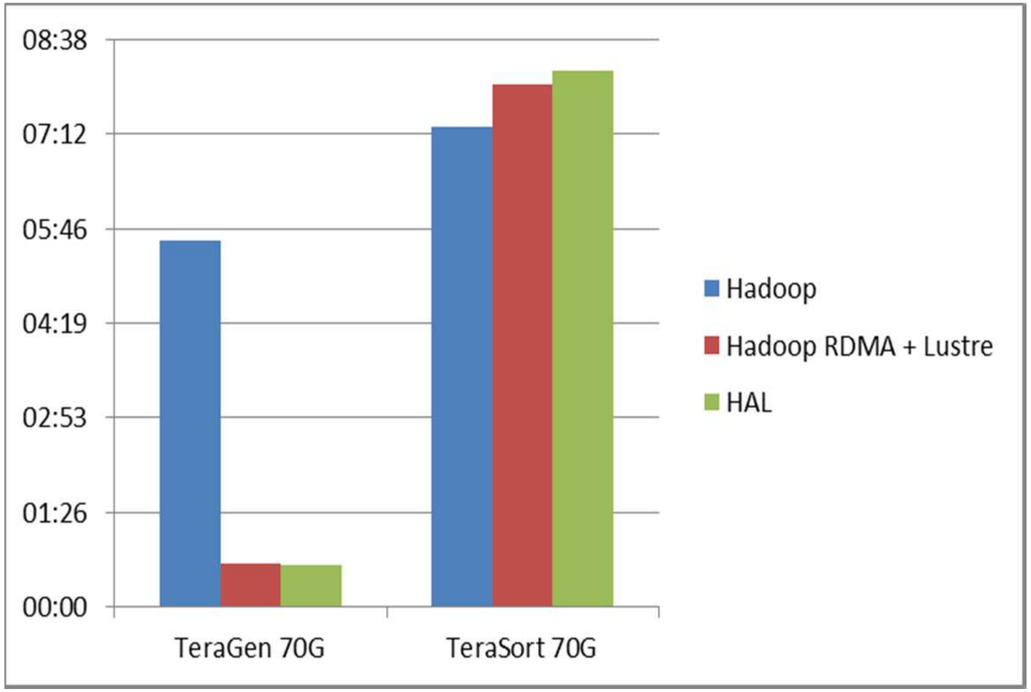
# Deploying Hadoop on Lustre.

Optimizing for Lustre: Eliminating Shuffle

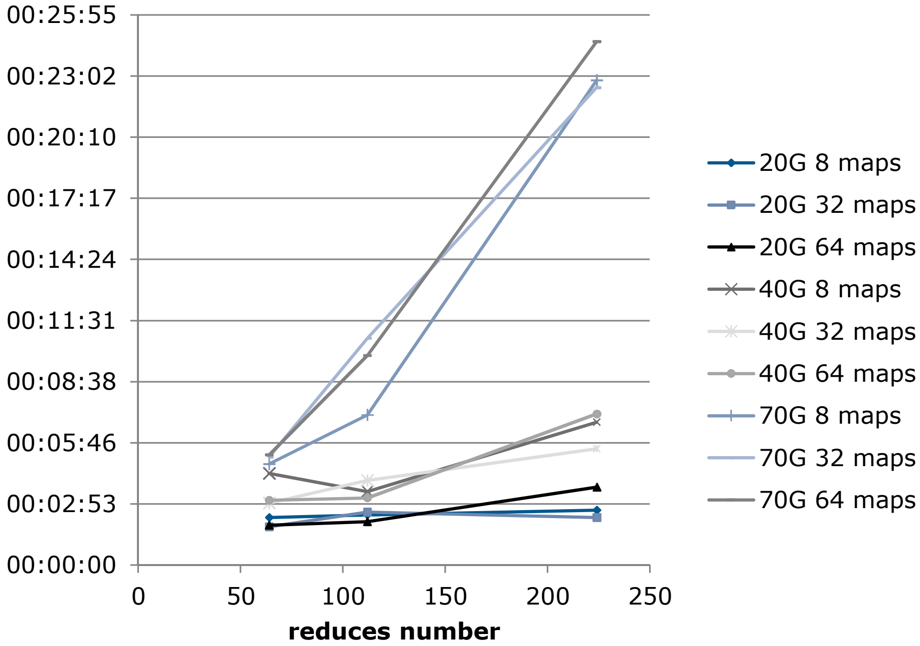


# Big Data solutions on an HPC cluster
## Performances

- 8 nodes (1 master & 7 slaves ):
  teragen & terasort workload

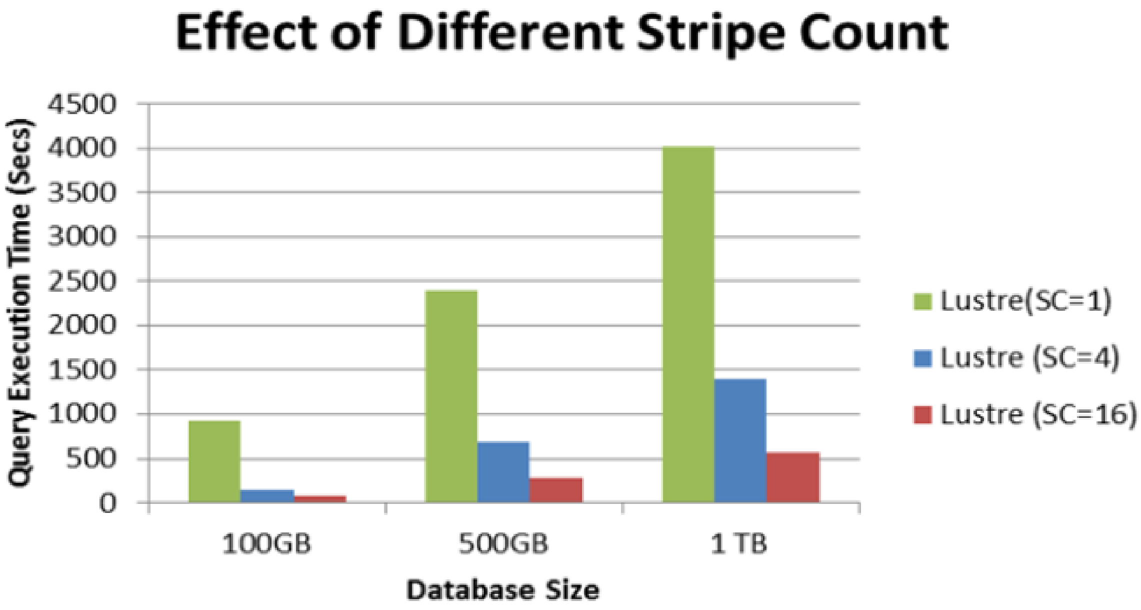



Terasort with different reduces



# Big Data solutions on an HPC cluster

Performances




# Deploying Hadoop on Lustre.
# Keys

## Performance

- Bring compute to the data: Run MapReduce* on Lustre* without code changes
- Run MapReduce* faster: Avoid the intermediate file shuffle with shared storage

## Efficiency

- Avoid Hadoop* islands in the sea of HPC systems
- Run MapReduce jobs alongside HPC workloads with full access to the cluster resources

## Manageability

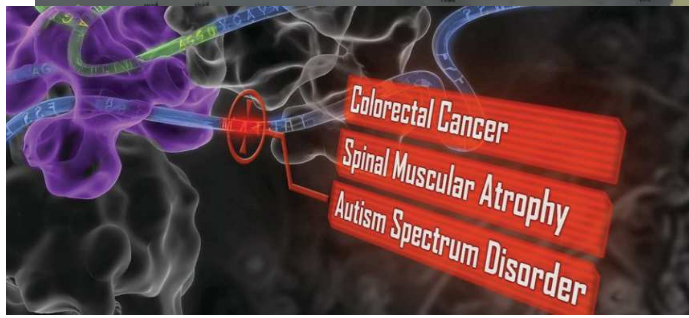
- Use the seamless integration to manage one common platform for Hadoop and HPC
- Develop with multiple programming models and deploy on shared storage

# DeepLearning

*Guillaume André et Matthieu Ospici*

# Introduction

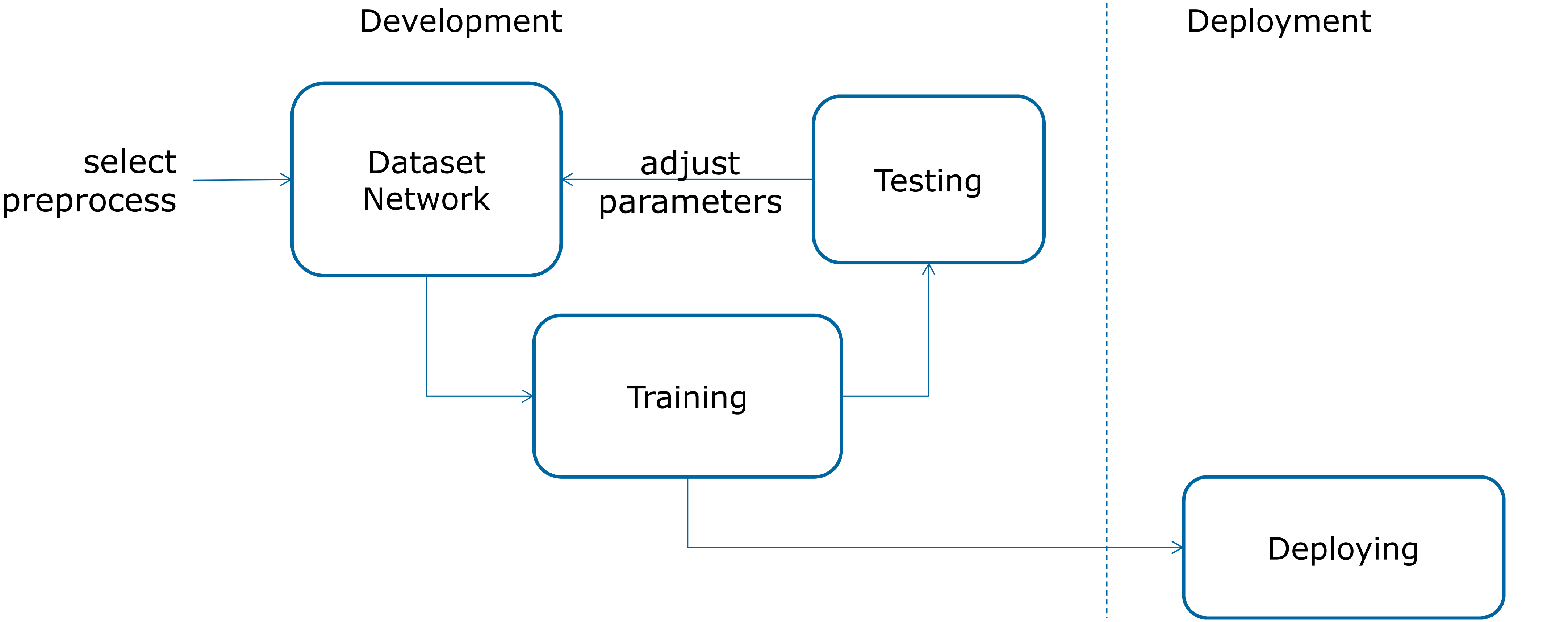
- branch of machine learning
- attempt to learn representation of data by using multiple processing layers
- various architectures exist :
  - Convolutional Neural Networks (CNN)
  - Recurrent Neural Networks (RNN)
  - …
- Many fields of application :
  - Automatic Speech Recognition : Microsoft Cortana, Google Now, Apple Siri
  - Natural Language Processing : sentiment analysis (eg Watson), spoken language understanding
  - Image Recognition
  - Autonomous vehicule
  - bioinformatics : predict gene ontology annotations and gene-function relationships (DeepGenomic)
  - …
- State-of-the-art result on various tasks

# Frameworks

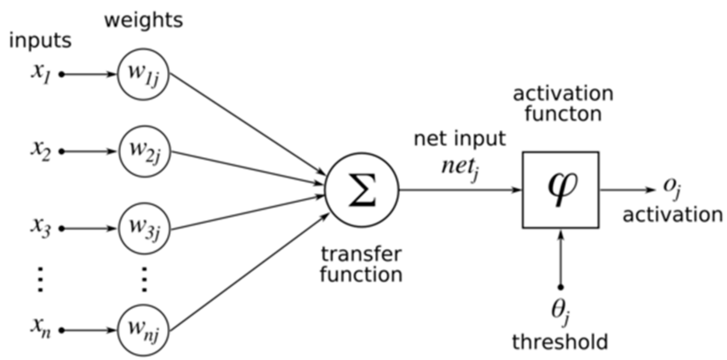
- Several open-sourced frameworks :
  - Caffe
    - used by companies as Facebook, Twitter, Yahoo … for image classification mainly and for deployment
  - Torch
    - developed by Google and Facebook, a great framework for research, used with Lua
  - Theano
    - developed by MILA at university of Montreal, mainly used for research, used with python
  - TensorFlow
    - developed and used by Google for deployment, recently open-sourced
  - …

# Deep Learning workflow

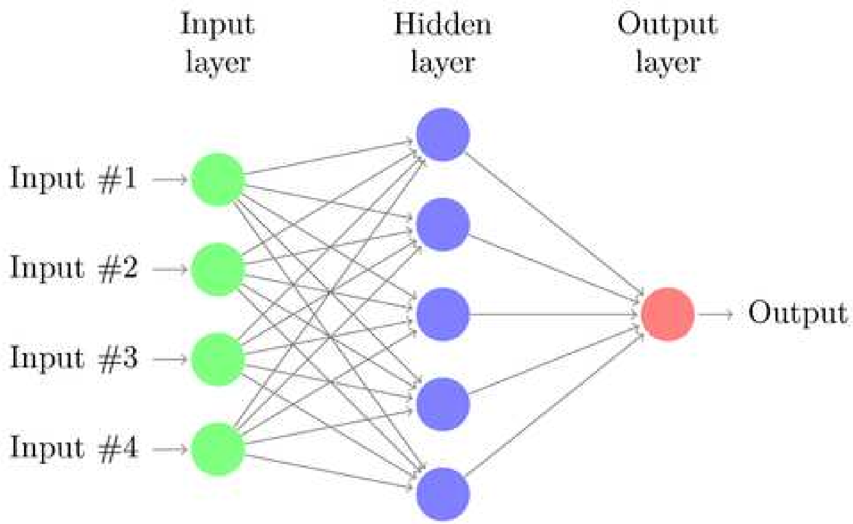


# Network

Set of processing layers
System of inter-connected neurons.
A layer is a group of neuron
Each neurons process information.
Each connection has a numerical weight.

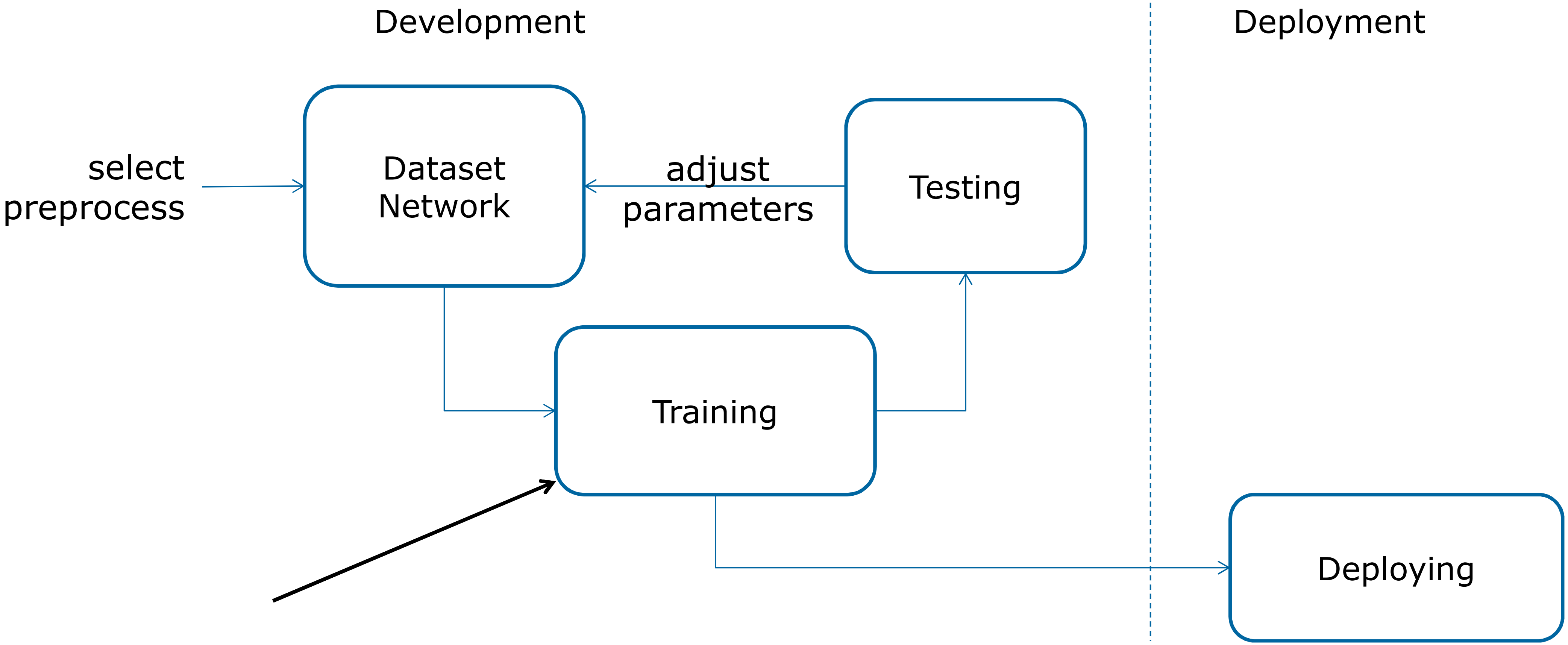



a neuron



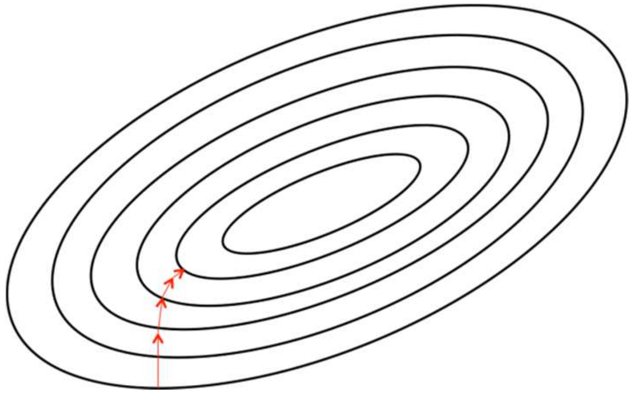

one hidden layer network / perceptron

# Deep Learning workflow



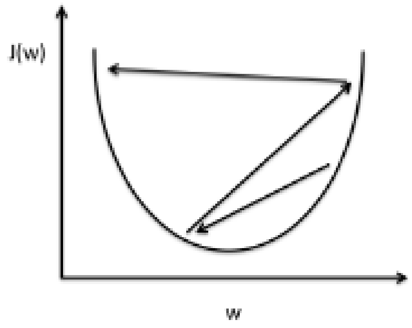
# Training

- based on forwarding and backwarding examples through the net
  - forward training examples
  - compare prediction and actual label
  - back-propagate the error to update the weights
- To update the weights, we try to minimize the cost function :
  - method used are based on gradient descent
  - ε is the learning rate and C the cost function

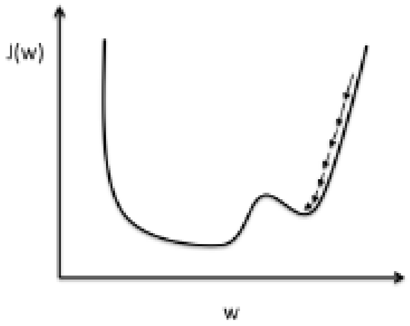


$$W_{t+1} = W_t - \varepsilon \frac{dC(W)}{dW}$$

gradient descent algorithm

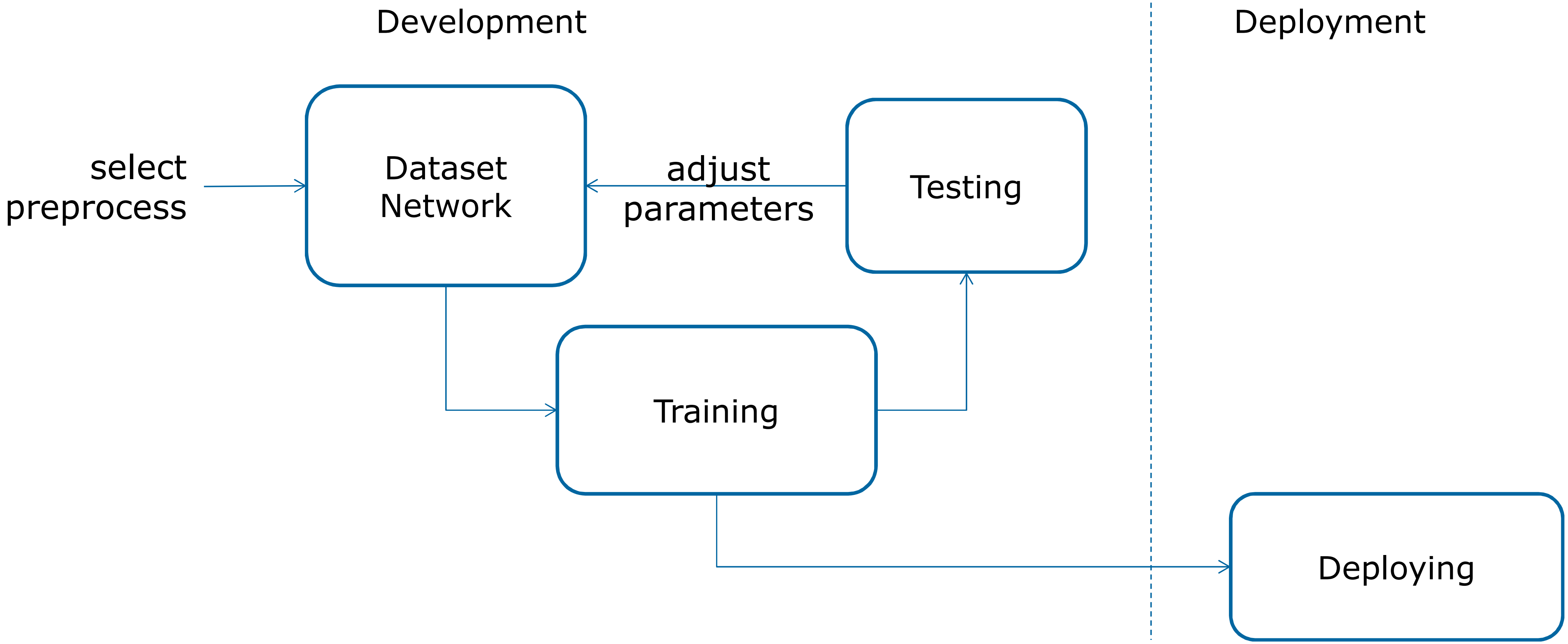



Large learning rate: Overshooting.



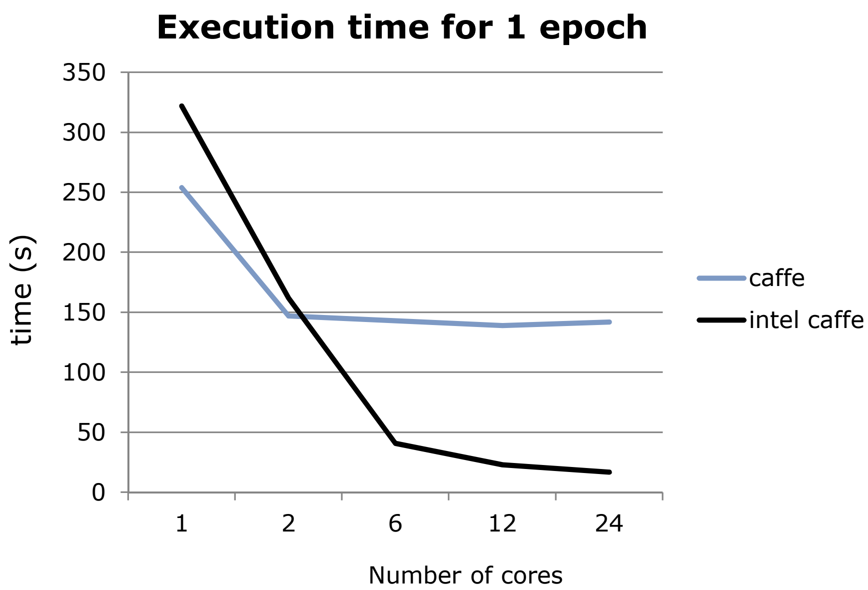

Small learning rate: Many iterations until convergence and trapping in local minima.

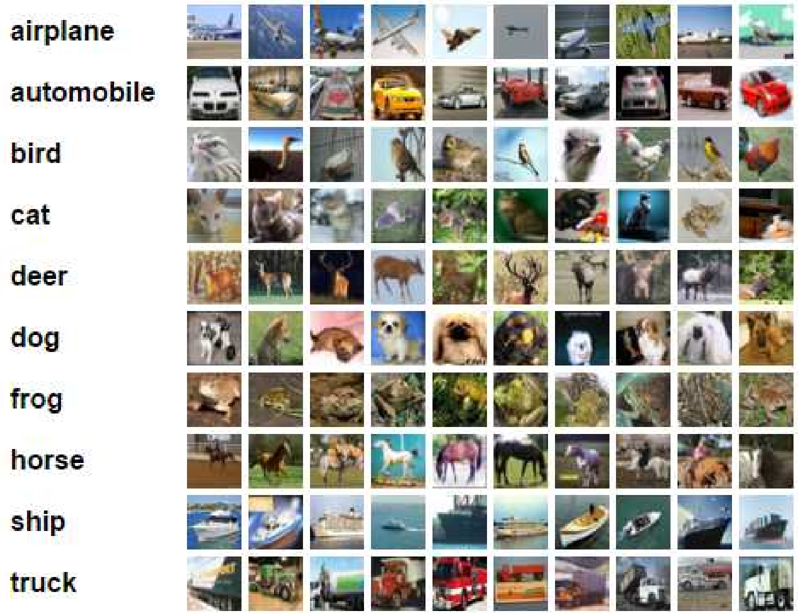
# Deep Learning workflow



# Caffe on CPU

- Train a small network on CIFAR10
  - CIFAR10: small test case 50000 + 10000 32x32x3 images
  - simple neural network (3 convolutions layers and 1 fully connected)
- CPU : 2 x Intel E5-2692v2 (2x12cores), 64GB memory

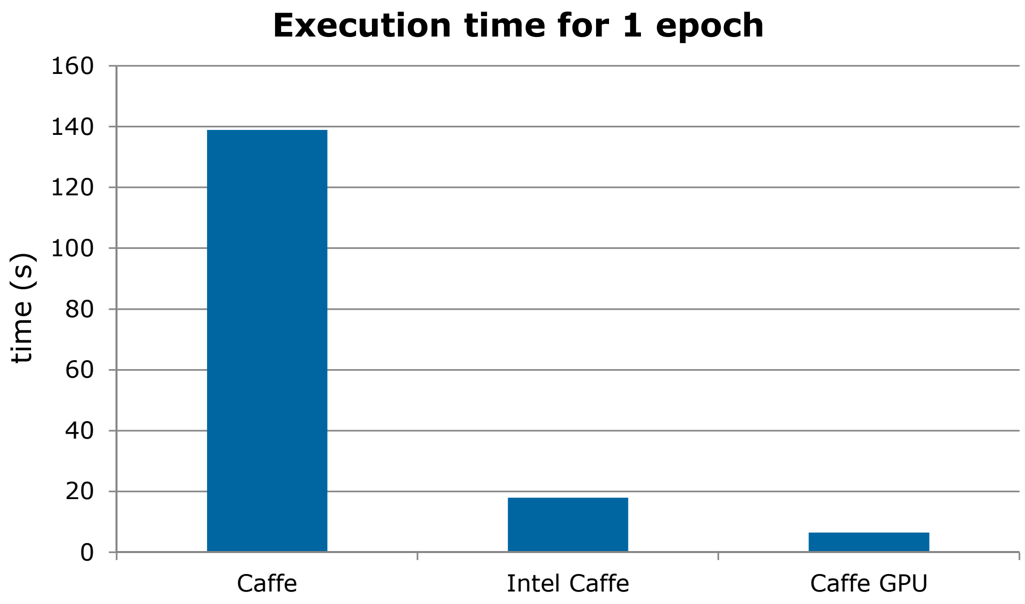


samples from Cifar10

# Caffe GPU/CPU

- GPUs : 2 x Nvidia K20X
- CPU : 2 x Intel E5-2692v2 (2x12cores), 64GB memory



# Deep Learning on HPC

- Working on distributing deep learning

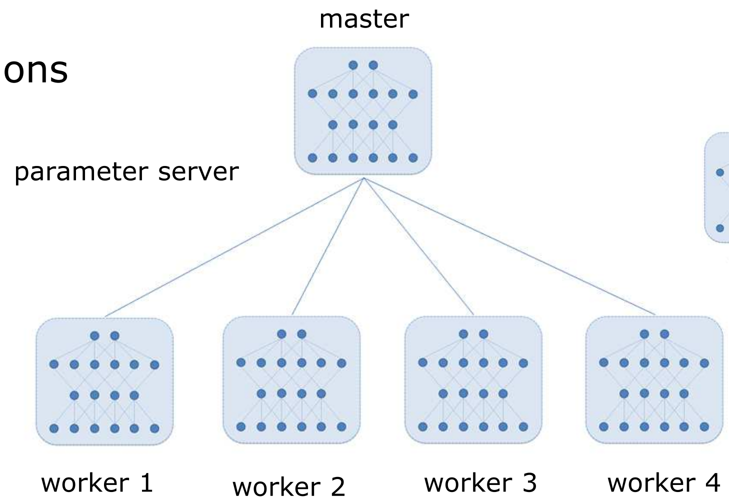
- Why ?
  - to use bigger network with more parameters
    - we tend to billions of parameter for some models
  - to accelerate learning of models
    - less than a week for the bigger models

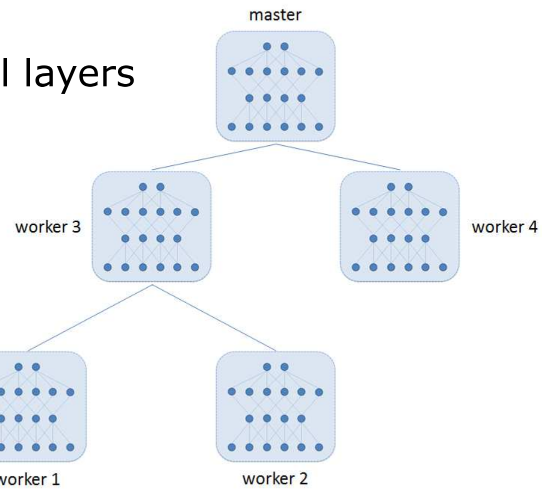
- How ?
  - data parallelism
  - model parallelism
  - → Hybrid parallelism

# Data parallelism

- distribution of the data on the nodes
- every nodes learn the same network
- every nodes have different data batch input
- better used with layers with more computation and less parameters as convolutional layers
- several models exits

- Slower convergence
- Waiting time during parameters communications



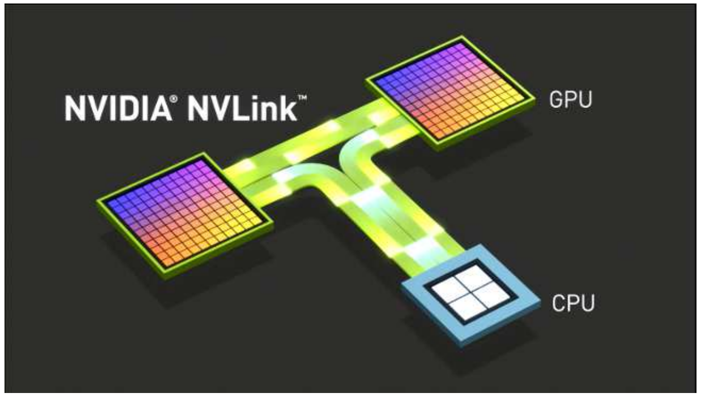



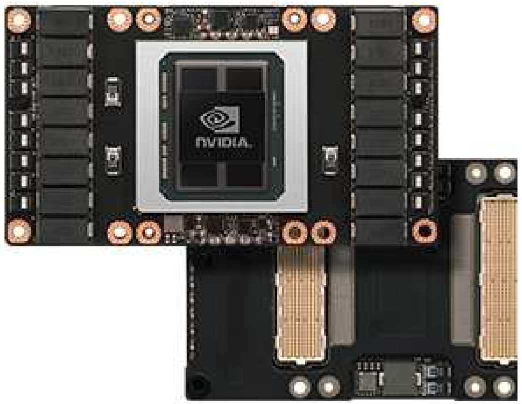

tree reduction

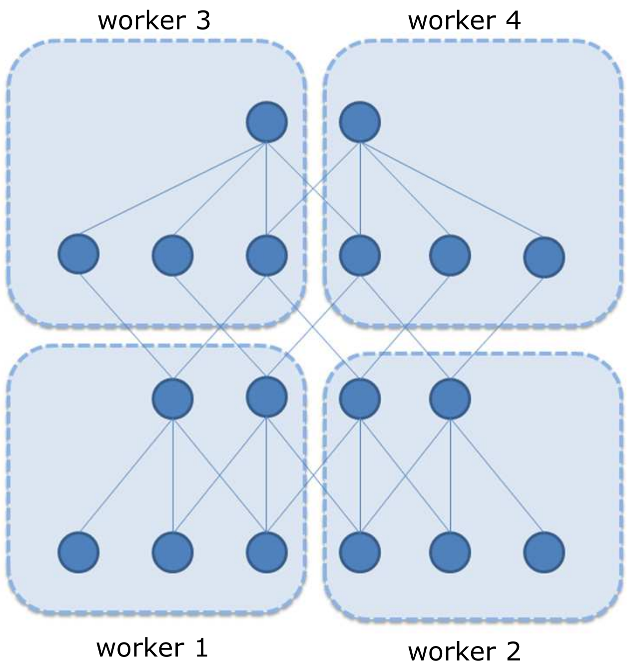
# Model parallelism

- distribution of the model on the nodes
- every nodes contains a part of the network
- every nodes use the same data batch input
- better used with layers with less computation and more parameters as fully-connected layers

- Need an efficient communication : large bandwidth and optimized latency
  - Coming this year, Pascal GPUs together with NVLink will help improving communication

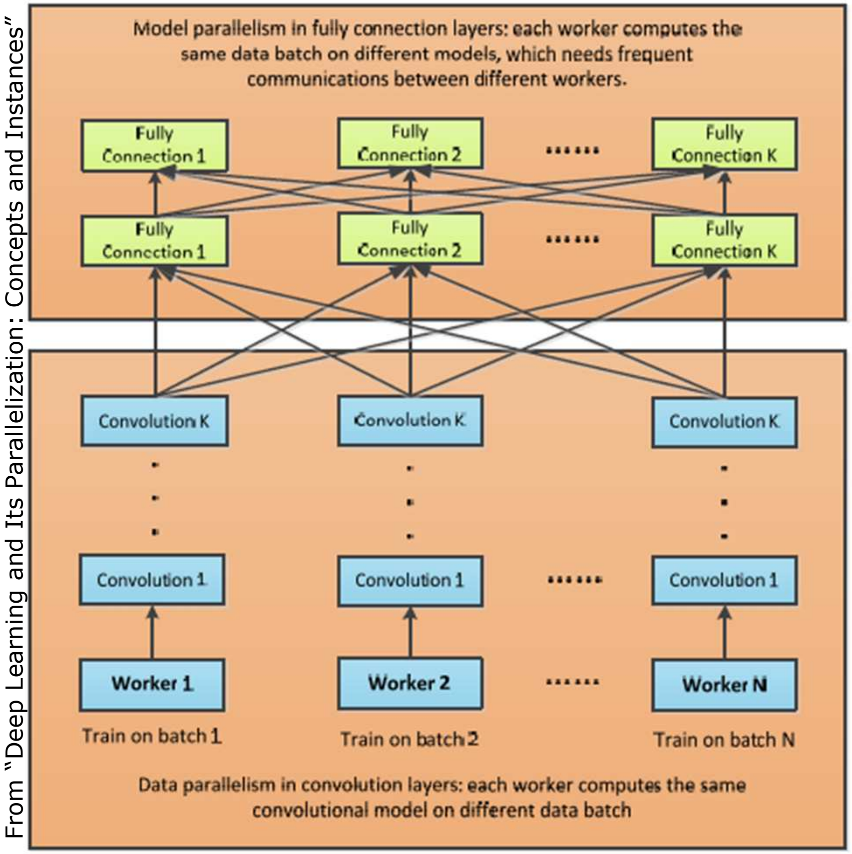


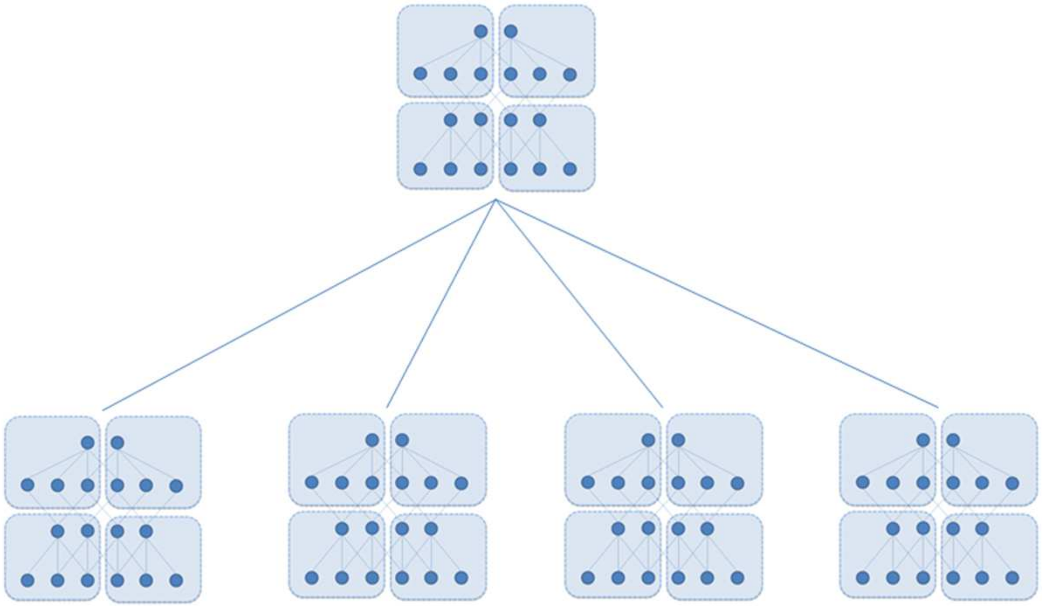



# Hybrid parallelism

From "Deep Learning and Its Parallelization: Concepts and Instances"



# Distributed deep learning frameworks

- Each company is developing its own solution
- Many framework were open-sourced recently

- We have begun to analyze some distributed framework :
  - CaffeOnSpark : developed by Yahoo!, use Caffe and Spark
  - IntelCaffe : developed by intel, optimizes Caffe on CPU
  - CNTK : developed by Microsoft, can be used on CPU or GPU cluster

# Workflow genomics et infrastructure calcul

*Pascale Rosse-Laurent*

# WHEN GENOMICS IS APPLIED, THESE ARE THE CHALLENGES…


**GENOMIC PIPELINE** (all linked tasks for genomic analysis)
**Difficulties to design it, execute it and interpret results from it**

### PRIMARY ANALYSIS – BASE CALLING

Converting light signal intensities into small string of nucleotide sequences around 100-200 bp long (reads).

**The millions reads equivalent to a whole human genome weighs around 150GB and normally, there is not a centralized storage**

### SECONDARY ANALYSIS – ALIGNMENT & VARIANT CALLING

Mapping DNA reads to an annotated reference sequence and determining the extent of variation from the reference

**Output from secondary analysis is not obtained at the required speed producing a bottleneck in the genomic analysis pipeline**

### TERTIARY ANALYSIS - INTERPRETATION

Analysing variants to assess their origin, uniqueness and likely functional impact

**This requires algorithms and software packages to query public databases & bioinformatics expertise**

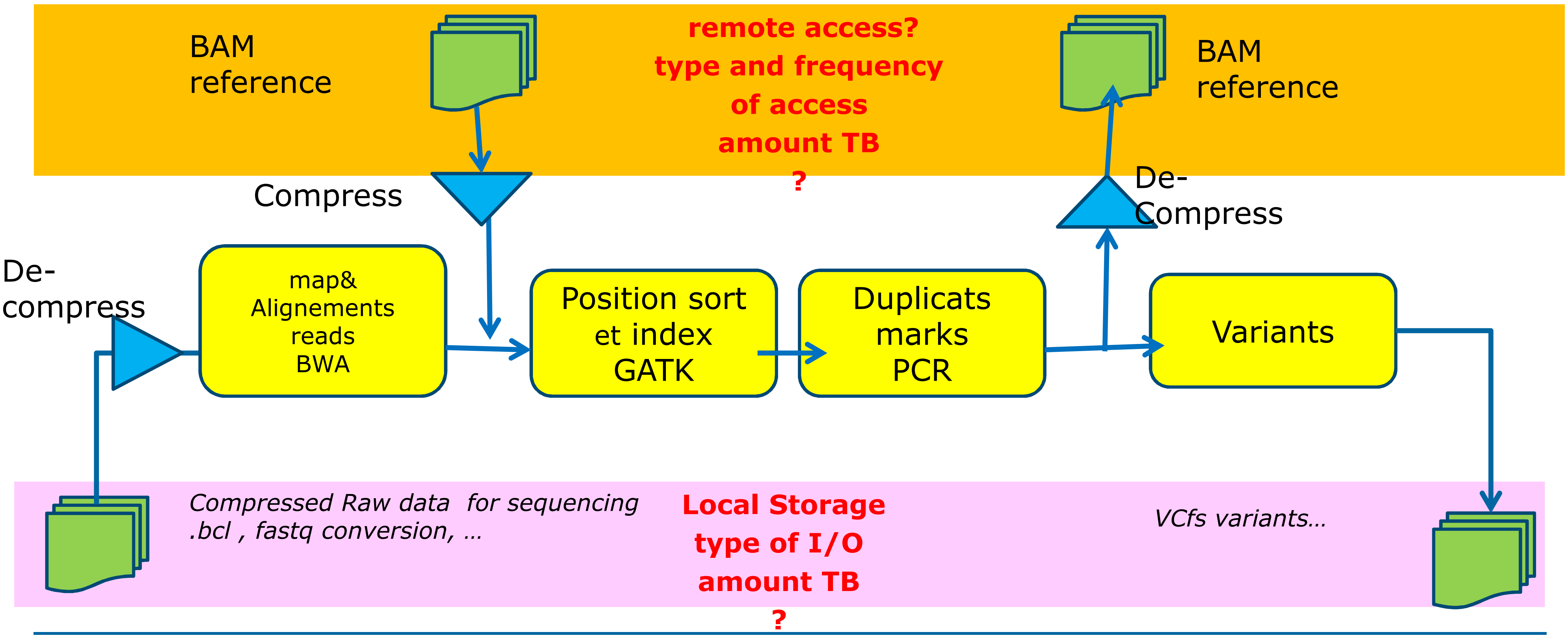
### INTERNAL & EXTERNAL DATA FLOW

Sequence data must be transferred off the sequencing machine, and may also need to be moved to several other locations for further analysis and safe archiving.

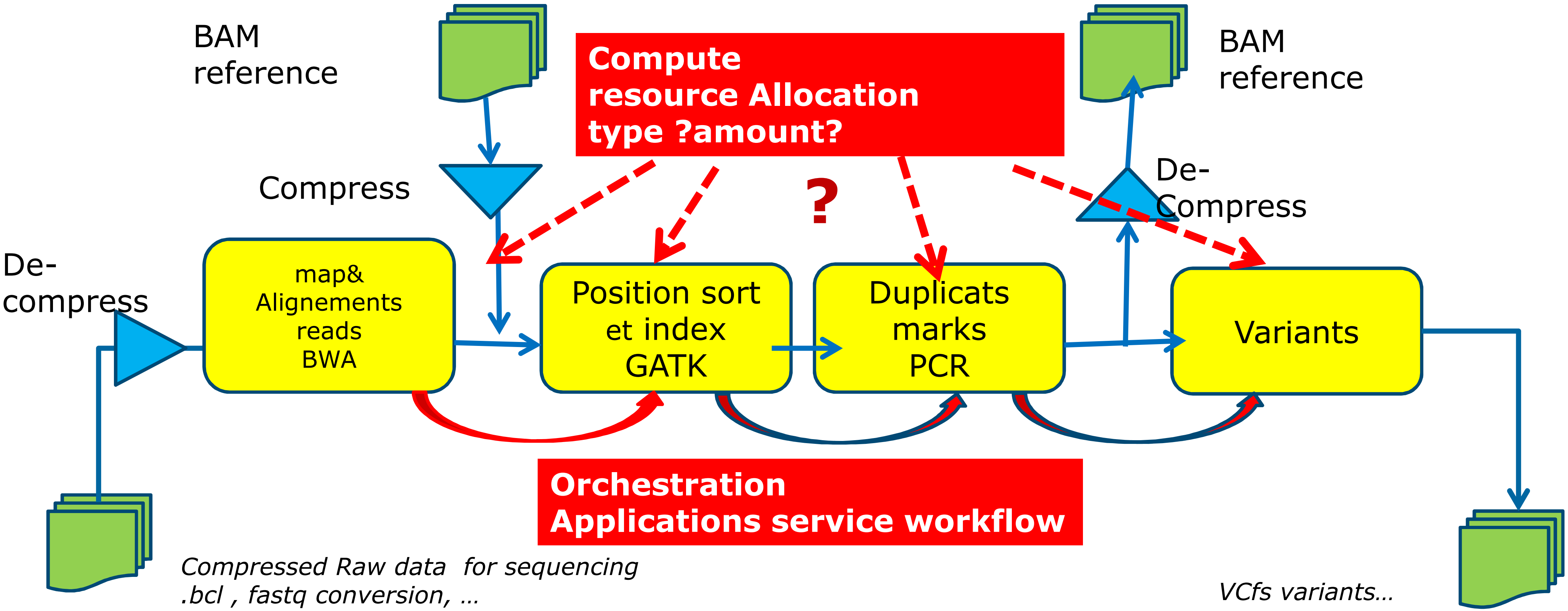
**Agile network communication with emphasis on I/O performance is required**

**SECURITY** An individual's genetic information is sensitive personal data.
**Technologies has to be put in place to guarantee confidentiality.**

# Schéma de traitement and data organisations



# Schéma de traitement: applications services organisation: ressources réservation et orchestration

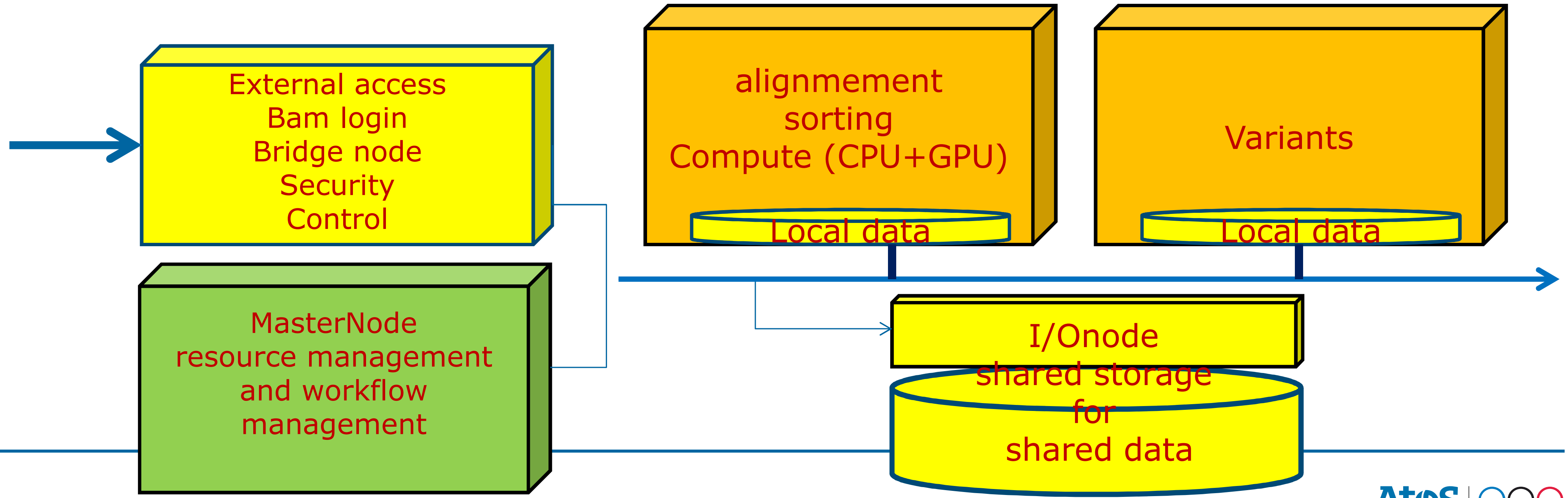


# Optimiser un workflow genomics sur un calculateur

- Analyse du workflow
    - Compréhension de toutes les phases et de leur interdépendance
        - Réserver des ressources adaptées (Slurm)
        - Identifier les services avec affinités, critères de ressources
        - Organiser le déploiement et configuration du workflow applicatifs
    - Optimiser la gestion des données
        - Identifier les données partagées au sein du
        - Organiser les espaces locaux
        - Organiser les accès distants
- Détermination des outils ou des indicateurs permettant l'analyse.
- Spécifier les méthodes de déploiement  optimal sur l'architecture matérielle et logicielle
- Gestion de l'anonymisation ? Sécurité des données  ?

# Constat

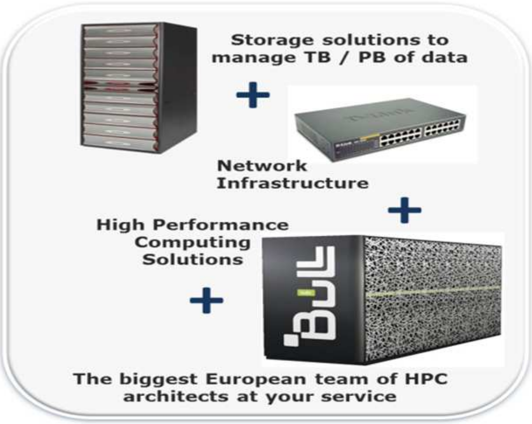
- Ordre d’importance
  1. la gestion des flux et accès aux données
  2. l’orchestration des phases de traitements
  3. les logiciels spécialisés , les librairies, la parallélisation, ….
  4. les solutions matériels HW accelerator, GPU ,FPGA, Xeon Phi



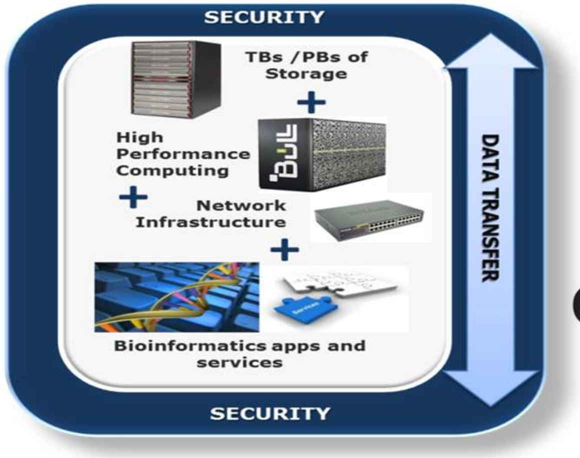
# CHALLENGES IN GENOMICS

- Lack of centralized genomics storage
- Bottlenecks in secondary analysis
- Bioinformatics expertise
- Difficulties in data sharing
- Data transfer painful

# SOLUTIONS




omics
Entry level




omics
Master

# REFERENCES OVERVIEW

FRANCE GÉNOMIQUE

NATIONAL CENTER FOR GENOMICS ANALYSIS

UNIVERSITY OF RIJEKA

IT4 INNOVATION

LEIDS UNIVERSITAIR MEDISH CENTRUM

HEALTHCARE RESEARCH INSTITUTE RAMÓN Y CAJAL

HEALTHCARE RESEARCH INSTITUTE 12 DE OCTUBRE HOSPITAL

PRINCE FELIPE BIOMEDIC RESEARCH CENTER

MEDICAL AND MOLECULAR GENETICS INSTITUTE

SPANISH NATIONAL CANCER RESEARCH CENTER

CENTER FOR GENOMICS AND ONCOLOGICAL RESEARCH

# Questions ?

- Exascale and extreme data prospective : Pascale.rosse-laurent@bull.net
- Equipe XBD (extreme bigdata) Benoit.pelletier@atos.net, Eric.Morvan@atos.net
- Equipe applications CEPP Xavier.vigouroux@atos.net

# Echanges

*Commentaires, questions & réponses*

**merci pour votre attention!**

Bull
atos technologies