La plateforme

OSIRIM

Observatoire des Systèmes d'Indexation et de Recherche d'Information Multimédia

Plan

- Définition et objectifs
- Modalités d'usage d'OSIRIM
- Les règles d'utilisation (la charte)
- Architecture physique et offre de services logicielle
- Projets hébergés et perspectives
- Un focus sur la baie EMC ISILON
- Slurm
- Hadoop
- Contraintes d'évolution de la plateforme

Définition

- Plateforme matérielle localisée à et administrée par l'IRIT.

- Un instrument scientifique qui met à disposition des utilisateurs une architecture matérielle et logicielle pour soutenir des activités scientifiques liées à l'analyse ou l'exploitation de grands volumes de données.

- A été réalisée dans le cadre du Contrat de Plan Etat Région (CPER) 2007-2013.

- A été financée par :
  - le fonds européen de développement régional (FEDER),
  - le gouvernement français,
  - la région Midi-Pyrénées et
  - le Centre National de la Recherche Scientifique (CNRS).

- Est opérationnelle dans sa version actuelle depuis début 2014, administrée par 1 IR CNRS (Noemi mai 2015) et 1 CDD IE CNRS 18 mois (octobre 2015), avec l'appui du service informatique de l'IRIT

Objectifs

- Héberger des projets scientifiques nécessitant :
  - le stockage et
  - le partage de plusieurs téraoctets de données

  ⇒ pour réaliser des expérimentations sur de grands volumes.

- Partager des corpus de référence.
  - Exemple : 1% des twitts mondiaux (streaming), depuis fin septembre

- Partager des outils logiciels, par exemple pour l'évaluation de technologies.

Modalités d'usage d'OSIRIM

- OSIRIM est ouverte :
  - Aux chercheurs et étudiants de l'IRIT travaillant sur des sujets liés au traitement de grands volumes de données.
  - À la communauté informatique et autres domaines scientifiques souhaitant utiliser ses moyens matériels ou logiciels sous certaines conditions.
- Administration :
  - Un projet est un espace d'hébergement de données et de logiciels partagés par plusieurs utilisateurs. Il est placé sous la responsabilité d'une personne.
  - les utilisateurs d'OSIRIM sont rattachés à un ou plusieurs projets
- Comment héberger un projet sur OSIRIM :
  - soumettre la demande d'hébergement via le site web «http://osirim.irit.fr», examinée par un comité de pilotage mensuel.
  - accepter la charte d'utilisation de la plateforme

Les règles d'utilisation (la charte)

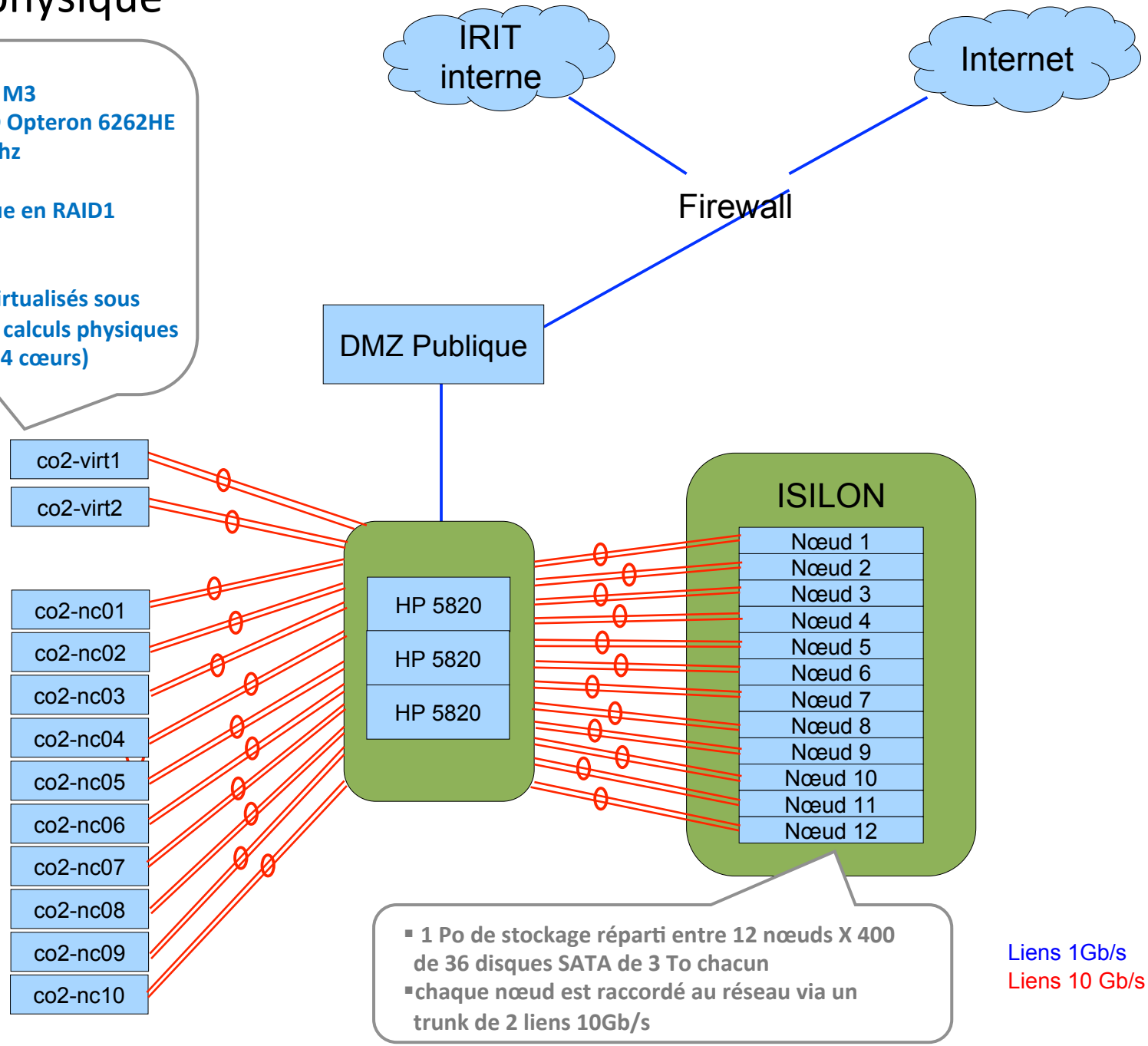
- Fixer les utilisations acceptables de cette plateforme :
  - les résultats produits directement par l'exploitation de la plateforme doivent revêtir un caractère scientifique.
  - l'utilisation des ressources de calcul doit respecter certaines règles sur un dispositif partagé.
  - l'utilisation de la plateforme par un utilisateur est soumise à autorisation du responsable de projet.

- Préciser la responsabilité de l'utilisateur :
  - l'usage des ressources informatique auxquelles il a accès.
  - la protection des informations enregistrées sur la plateforme.
  - la déclaration de la tentative de violation de son compte et de façon générale, toute anomalie qu'il peut constater.
  - aucun backup des données (pas d'engagement sur la conservation des données).

Architecture physique

- 12 serveurs IBM X3755 M3
  - 4 Processeurs AMD Opteron 6262HE de 16 cœurs à 1,6 Ghz
  - 512 Go de RAM
  - 2 x 300 Go de disque en RAID1
  - réseau 2 x 10Gb/s

- Répartis en 2 nœuds virtualisés sous VMWare et 10 nœuds de calculs physiques (10 x 512 Go de RAM et 64 cœurs)



- 1 Po de stockage réparti entre 12 nœuds X 400 de 36 disques SATA de 3 To chacun
- chaque nœud est raccordé au réseau via un trunk de 2 liens 10Gb/s

Au niveau logiciel …

Une offre de services articulée autour de deux approches :

- Un gestionnaire de job SLURM (Simple Linux Utility for Resource Management) permettant la distribution de traitements réalisés avec des langages / logiciels mutualisés : MATLAB, PYTHON, JAVA, …
- Une distribution Hadoop avec son écosystème applicatif : Hive, Pig, Hbase, Flume, …

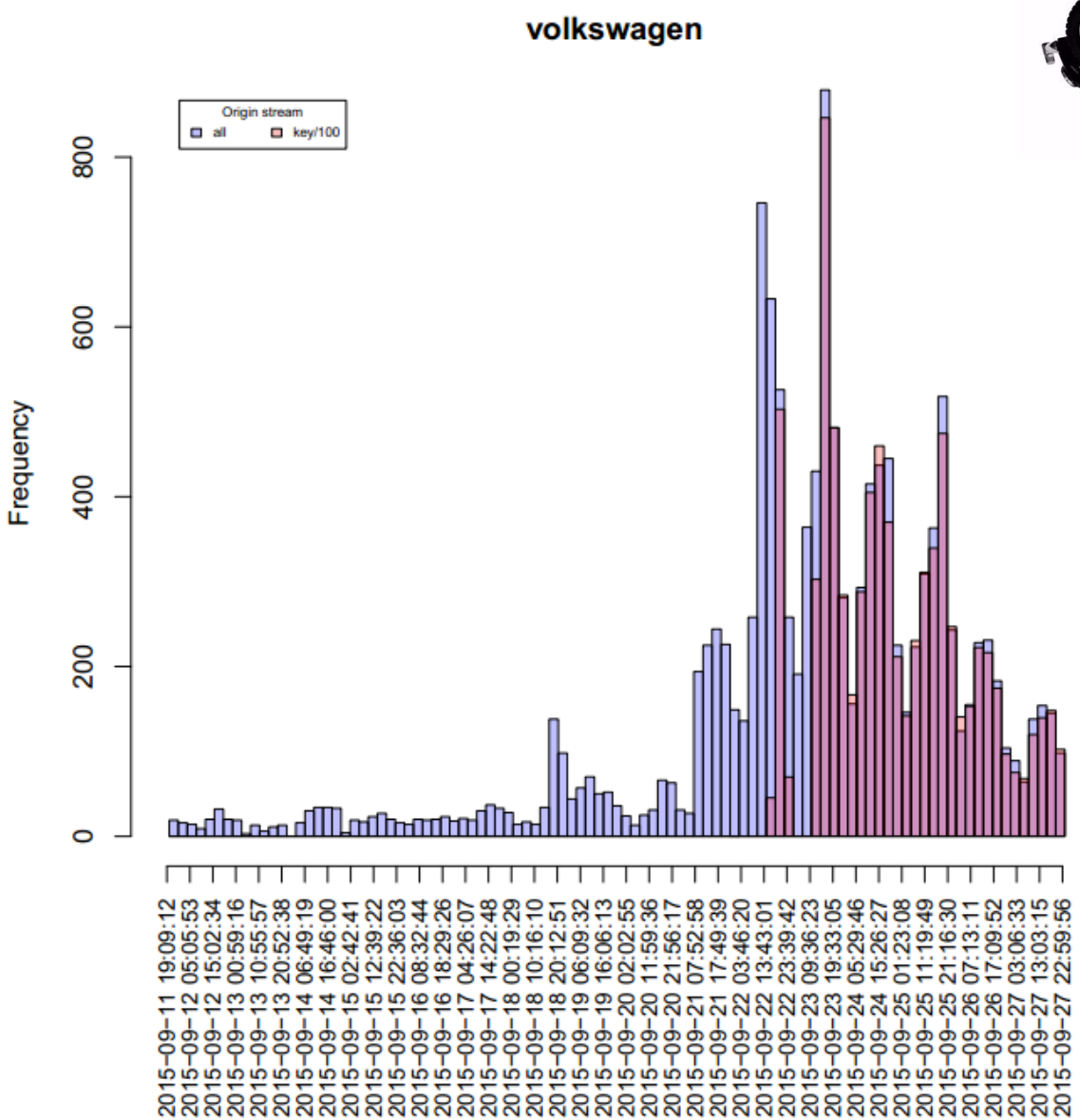
Des documentations pour apprendre à utiliser OSIRIM (intranet)

- Architecture de la plateforme
- Caractéristiques techniques de la plateforme
- Connexion à la plateforme
- Lancement de jobs sur le cluster
- FAQ

Où en sommes nous ?

- Hébergement d'une quinzaine de projets :
  - Travaux de recherche des équipes :
    - SIG : indexation de grandes masses de données hétérogènes. SIG participe à la campagne TREC sur plusieurs tâches :
      - COntext (quelques tera de données), classé premier l'année dernière
      - Social Search (idem, tweets)
      - Kba ()
    - SAMOVA : évaluation d'outils d'indexation de contenus musicaux, indexation de grands volumes d'enregistrements d'émissions de télévision internationales.
    - MELODI : analyse de corpora textuels et ontologies.
    - TCI : Traitement complexe d'Image.
    - …
  - Projets :
    - QUAERO : innovation sur l'analyse automatique et l'enrichissement de contenus numériques, multimédias et multilingues.
    - IRIM_at_TRECVID : collaboration de plusieurs équipes scientifiques françaises, regroupées dans le consortium IRIM, pour leur participation à la campagne internationale TRECVID
    - RayWarps**:** Edition et contrôle interactifs et intuitifs d'images de synthèse
    - SemDis: création de bases distributionnelles de référence pour le français.
    - CAIR: recherche agrégative de données
    - Petasky : techniques de partitionnement de données issues du domaine de la cosmologie
    - POLEMIC : analyse du comportement des utilisateurs dans les réseaux sociaux,
    - …

18/9/2015: Public announcement by EPA of order to recall 2009–2015 cars



Projet POLEMIC : A. Quirin, F. Sedes

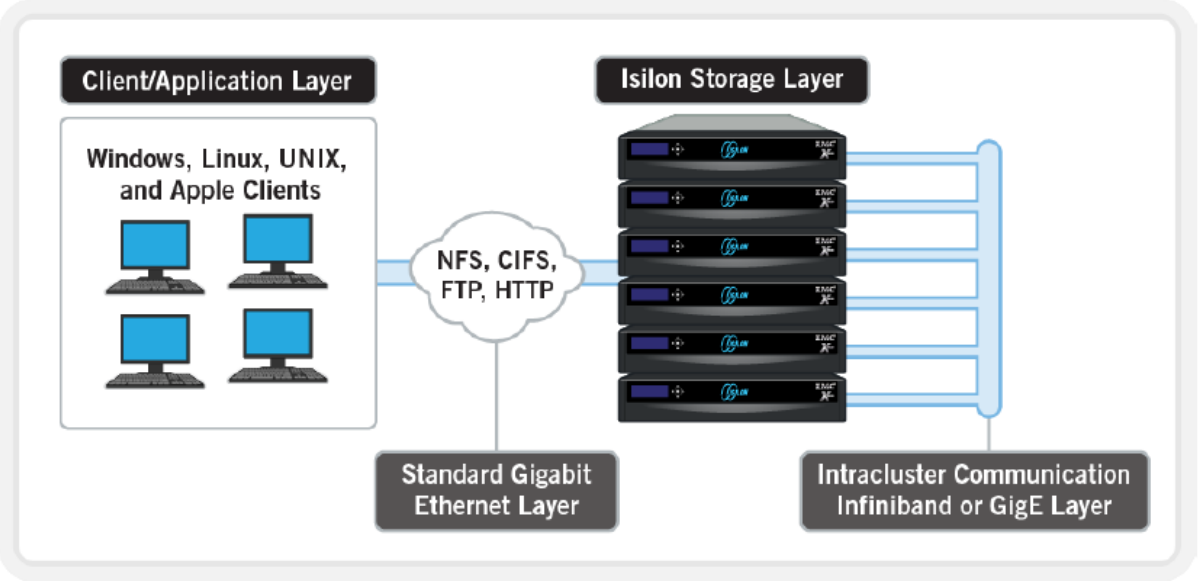
Quelques chiffres

- Nombre d'utilisateurs :
  - Plus de 100 utilisateurs.
- Espace utilisé :
  - ≈ 150 To utilisés.
- Jobs lancés depuis la mise en production de la nouvelle plateforme :
  - > 900 000 jobs slurm.

Perspectives

- Mise à disposition d'un espace de stockage conséquent pour le cluster de calcul Grid5000
- Déploiement de Mongodb, puis changement de version Hadoop (Hortonworks HDP 2.3 qui inclue Spark)
- Hébergement de projets de taille plus importante
  - Partenariat avec l'école nationale supérieure de Police (montage de projets H2020 2015-2016)
  - Nutrition / Santé (montage de projets H2020 2017)
- Mini séminaires et formations pour l'accompagnement des chercheurs
- Soutien pour l'initiation à la recherche dans des formations de master

Un Focus sur la baie EMC ISILON (1)

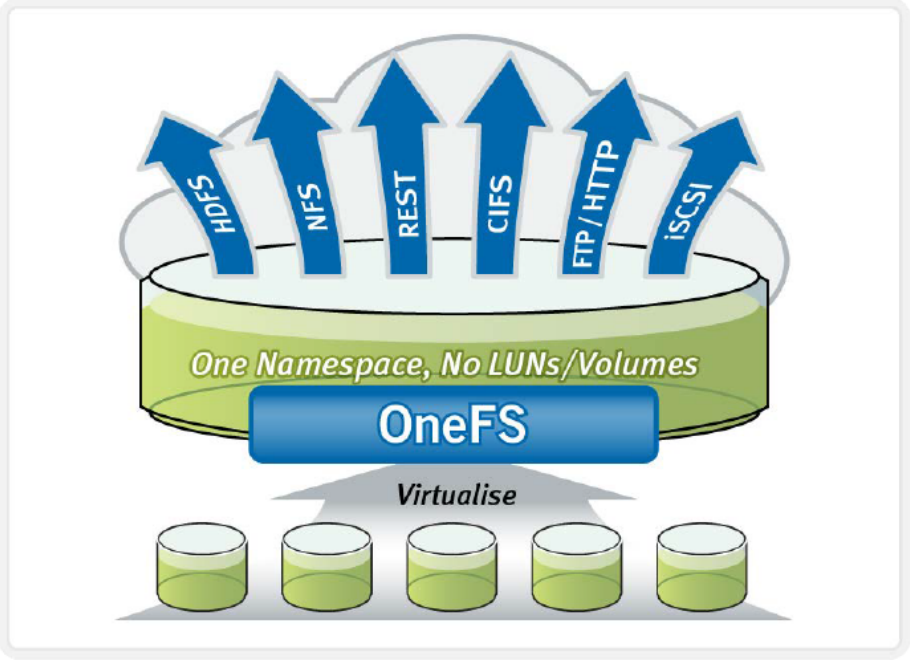


Un cluster Isilon est constitué de X nœuds qui apportent chacun au cluster leur capacité disque, cache mémoire, CPU et bande passante.

Le cluster fournit aux serveurs un file système unique dont la capacité peut évoluer en fonction des besoins.

La communication inter-nœud en infiniband repose sur un protocole propriétaire en unicast

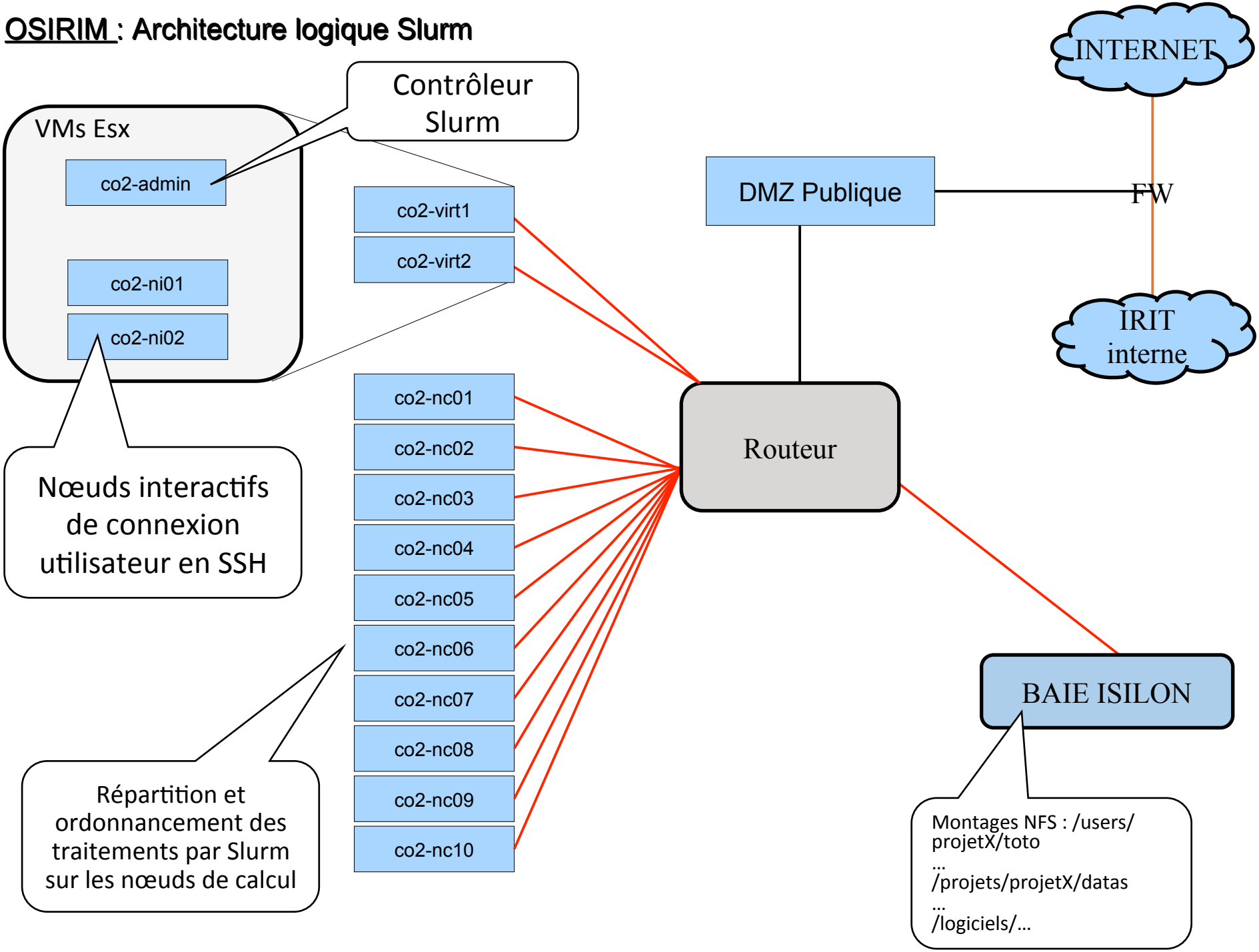
Un Focus sur la baie EMC ISILON (2)



OneFS est l'OS qui intègre à la fois le système de fichiers, la gestion de volume, et la sécurisation des données.

L'ensemble constitue un unique système de fichiers distribué, avec un seul espace de nommage, qui a la capacité de présenter les données aux serveurs suivant plusieurs protocoles : NFS, CIFS, HDFS, Rest, HTTP, FTP, ISCSI

OSIRIM : Architecture logique Slurm



- Exemple de shell pour l'exécution d'un traitement via SLURM (test-matlab.sh)

```
#!/bin/sh
#SBATCH --begin=now+60      #(seconds by default)
#SBATCH --job-name=test-matlab
#SBATCH --mail-type=ALL      #(Notify user by email when certain event types occur. Valid type values are BEGIN, END, FAIL, REQUEUE, and ALL (any state change)
#SBATCH --mail-user=bob@irit.fr
#SBATCH --cpus-per-task=4    # (permet de specifier le nombre de cpus utilises si vous faites du multithreading)
#SBATCH --mem=768            #(real memory required per node in MegaBytes)
#SBATCH --nodes=2            #(Request that a minimum of 2 nodes be allocated to this job)
#SBATCH --output=test.out    #(Instruct SLURM to connect the batch script's standard output directly to the file name "test1.out")
#SBATCH --time=60            #(Set a limit on the total run time of the job allocation (minutes by default)
#SBATCH --tmp=200            #(Specify a minimum amount of temporary disk space(MB by default))
/logiciels/matlab/bin/matlab -nosplash -nodisplay -nodesktop -r simple_job
```

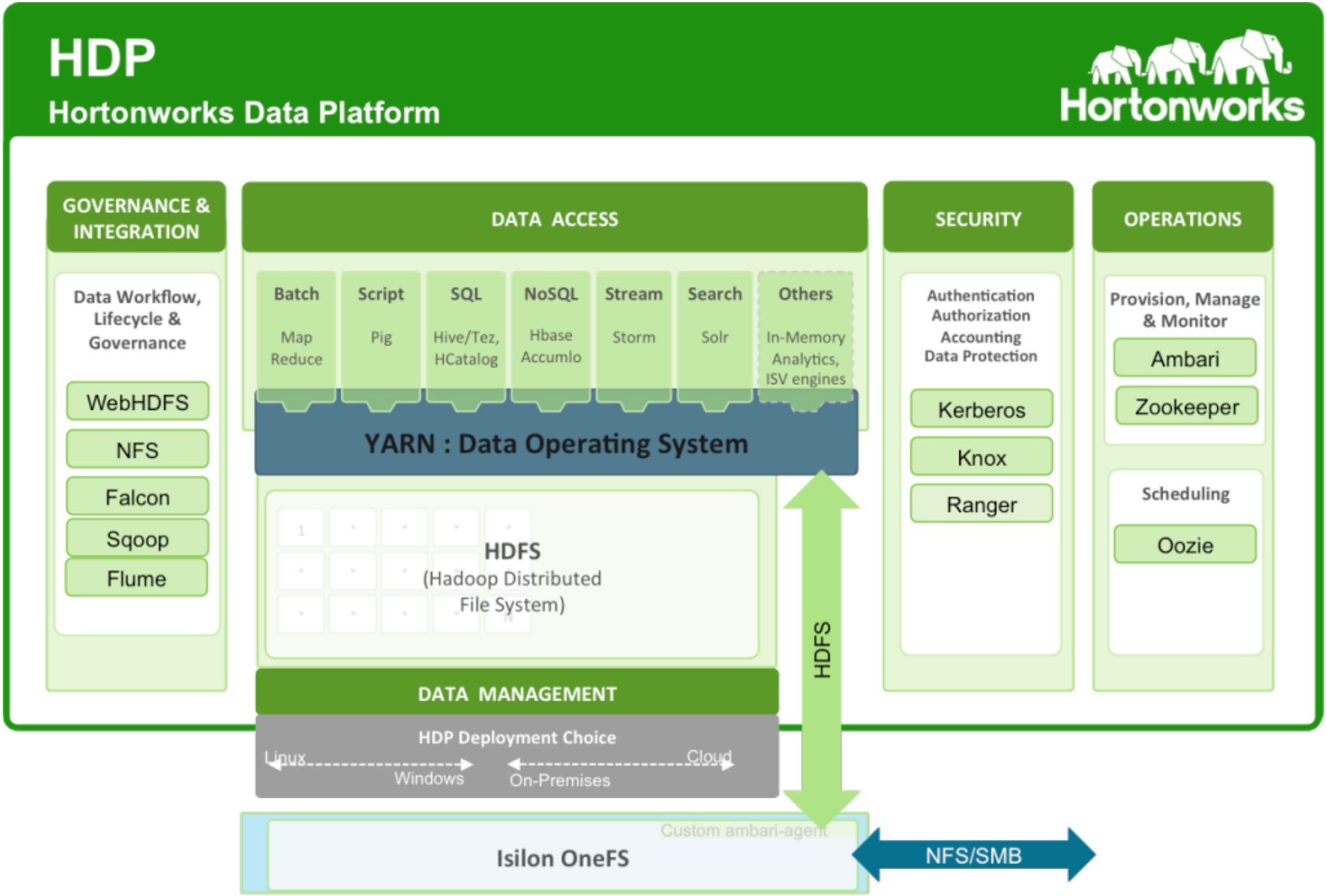
- Exécution du shell via Slurm

```
sbatch ./test-matlab.sh
```

Qu'est ce qu'Hadoop ?

- Hadoop est un framework qui permet de réaliser des tâches en parallèle sur des données stockées et distribuées dans un cluster

- Hadoop 2.0 est constitué de plusieurs composants :
  - **Hadoop Common**: le noyau, support aux autres modules
  - **HDFS**: Hadoop Distributed File System
  - **MapReduce**: assure le traitement parallèle et "scalable" de masses de données
  - **YARN**: gestion des ressources du cluster et gestion des jobs.
  - **Des applications de plus haut niveau** : Hbase, Pig, Hive, Oozie, …


HDP
Hortonworks Data Platform
Hortonworks
GOVERNANCE & INTEGRATION
Data Workflow, Lifecycle & Governance
WebHDFS
NFS
Falcon
Sqoop
Flume
DATA ACCESS
Batch
Map Reduce
Script
Pig
SQL
Hive/Tez, HCatalog
NoSQL
Hbase Accumlo
Stream
Storm
Search
Solr
Others
In-Memory Analytics, ISV engines
YARN : Data Operating System
HDFS
(Hadoop Distributed File System)
DATA MANAGEMENT
HDP Deployment Choice
Linux
Windows
Cloud
On-Premises
HDFS
SECURITY
Authentication
Authorization
Accounting
Data Protection
Kerberos
Knox
Ranger
OPERATIONS
Provision, Manage & Monitor
Ambari
Zookeeper
Scheduling
Oozie
Custom ambari-agent
Isilon OneFS
NFS/SMB

Exemple classique de cluster Hadoop

Un cluster Hadoop est constitué de serveurs maîtres et esclaves :

- Les serveurs maîtres gèrent l'infrastructure
- Les serveurs esclaves hébergent les données distribuées et exécutent les traitements

Master Servers – NameNode, ResourceManager, Standby Name Node, HBase Master

Master Node 1
NameNode
Oozie Server
ZooKeeper

Master Node 2
ResourceManager
Standby NameNode
HBase Master
HiveServer2
ZooKeeper

Management Node
Ambari Server
Ganglia/Nagios
WebHCat Server
JobHistoryServer
ZooKeeper

Slave Servers – NodeManager, DataNode, HBase RegionServer

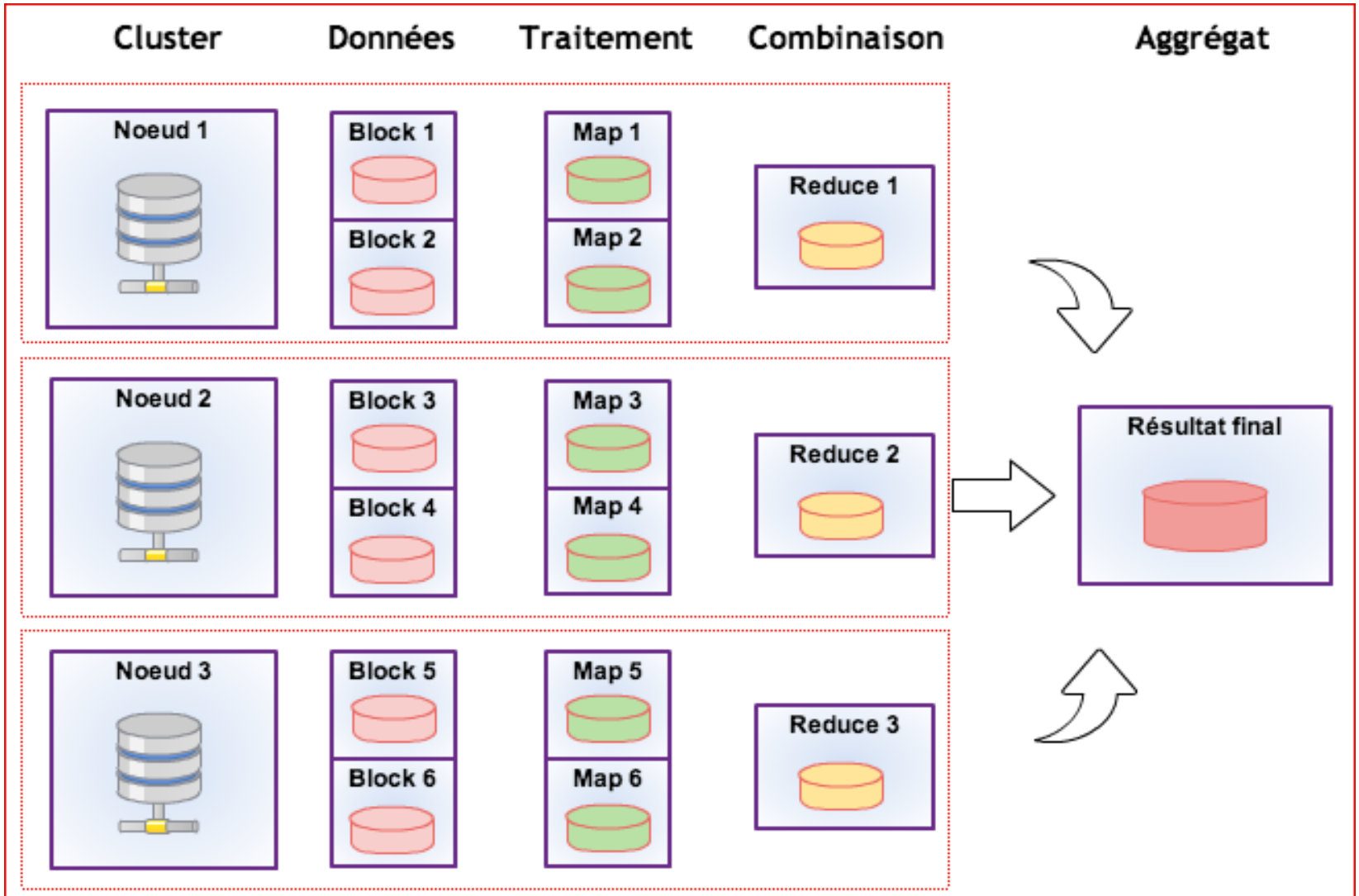
DataNode 1
DataNode
NodeManager
H RegionServer

DataNode 2
DataNode
NodeManager
H RegionServer

DataNode 3
DataNode
NodeManager
H RegionServer

...

DataNode n
DataNode
NodeManager
H RegionServer



Colocalisation des données et des traitements

Exemple de cluster Hadoop - Isilon

Un cluster Hadoop est constitué de serveurs maîtres et esclaves :

- Les serveurs maîtres gèrent l'infrastructure
- Les serveurs esclaves exécutent les traitements
- La baie ISILON héberge les données et les présente en HDFS

Master Servers – NameNode, ResourceManager, Standby Name Node, HBase Master

Master Node 1
Oozie Server
ZooKeeper

Master Node 2
ResourceManager
HBase Master
HiveServer2
ZooKeeper

Management Node
Ambari Server
Ganglia/Nagios
WebHCat Server
JobHistoryServer
ZooKeeper

Slave Servers – NodeManager,, HBase RegionServer

Node manager 1
NodeManager
H RegionServer

Node manager 2
NodeManager
H RegionServer

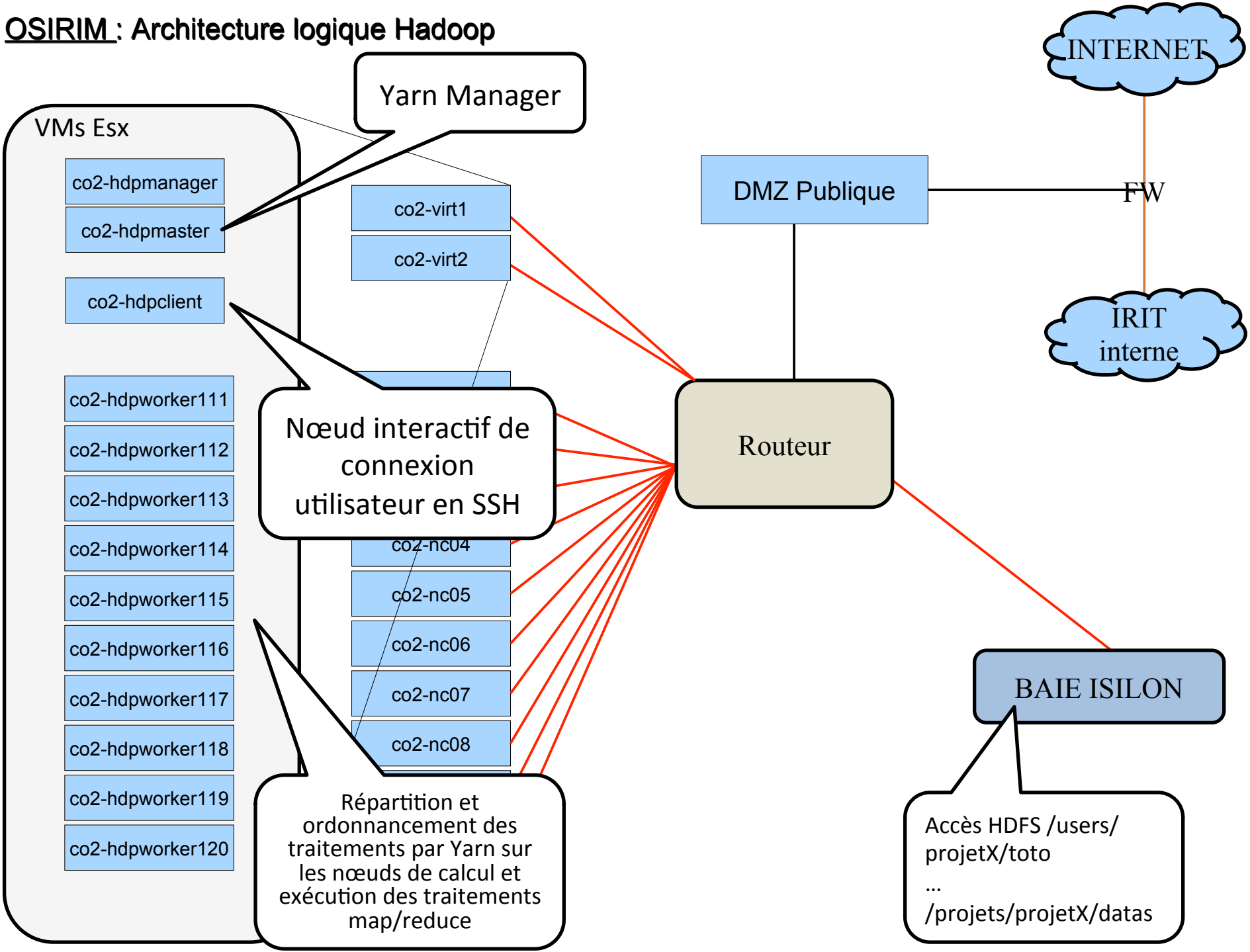
Node manager 3
NodeManager
H RegionServer

...

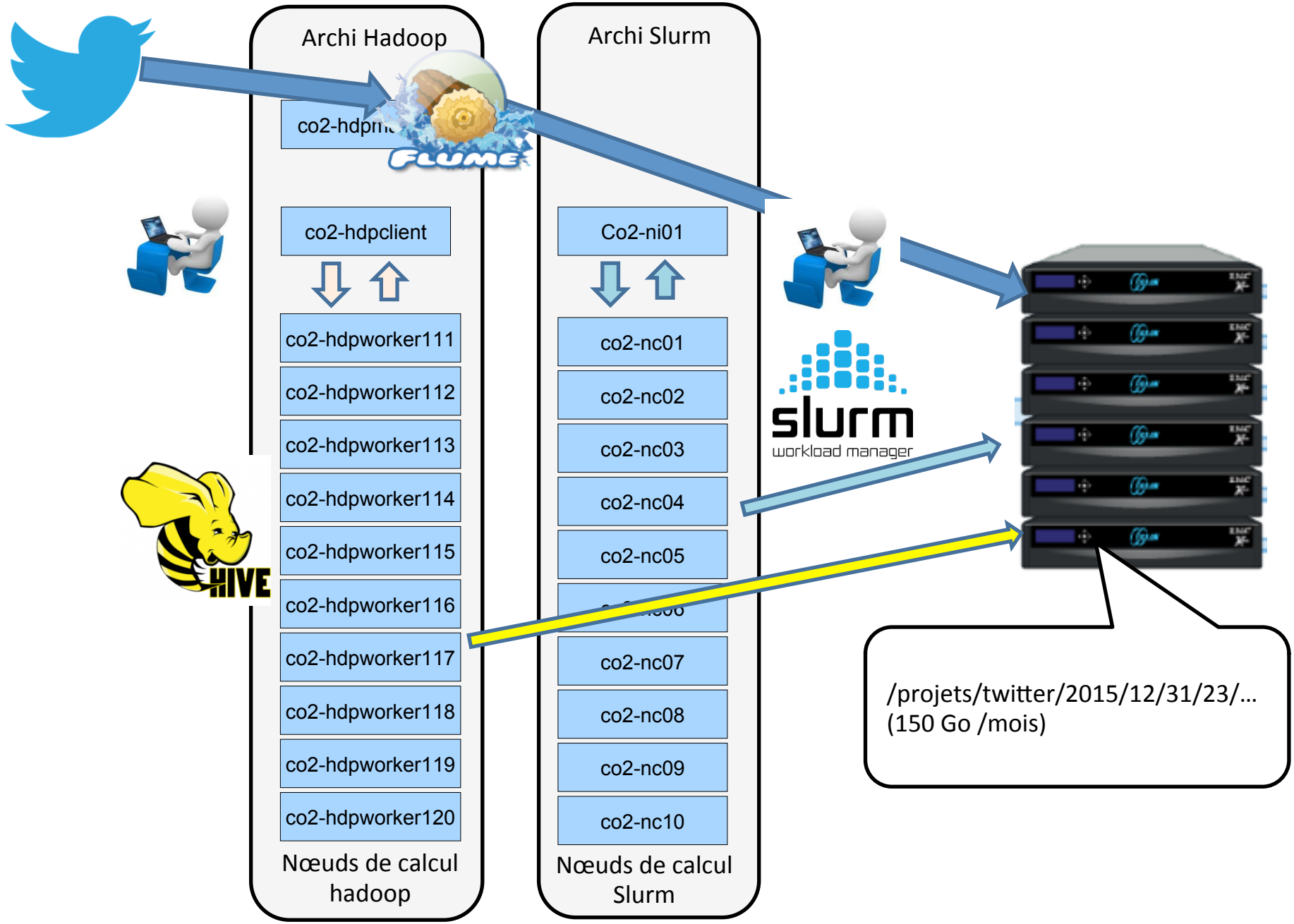
Node manager n
NodeManager
H RegionServer

Baie ISILON

OSIRIM : Architecture logique Hadoop



OSIRIM : Exemple d'exploitation d'un corpus de tweets



Contraintes d'évolution de l'offre de services

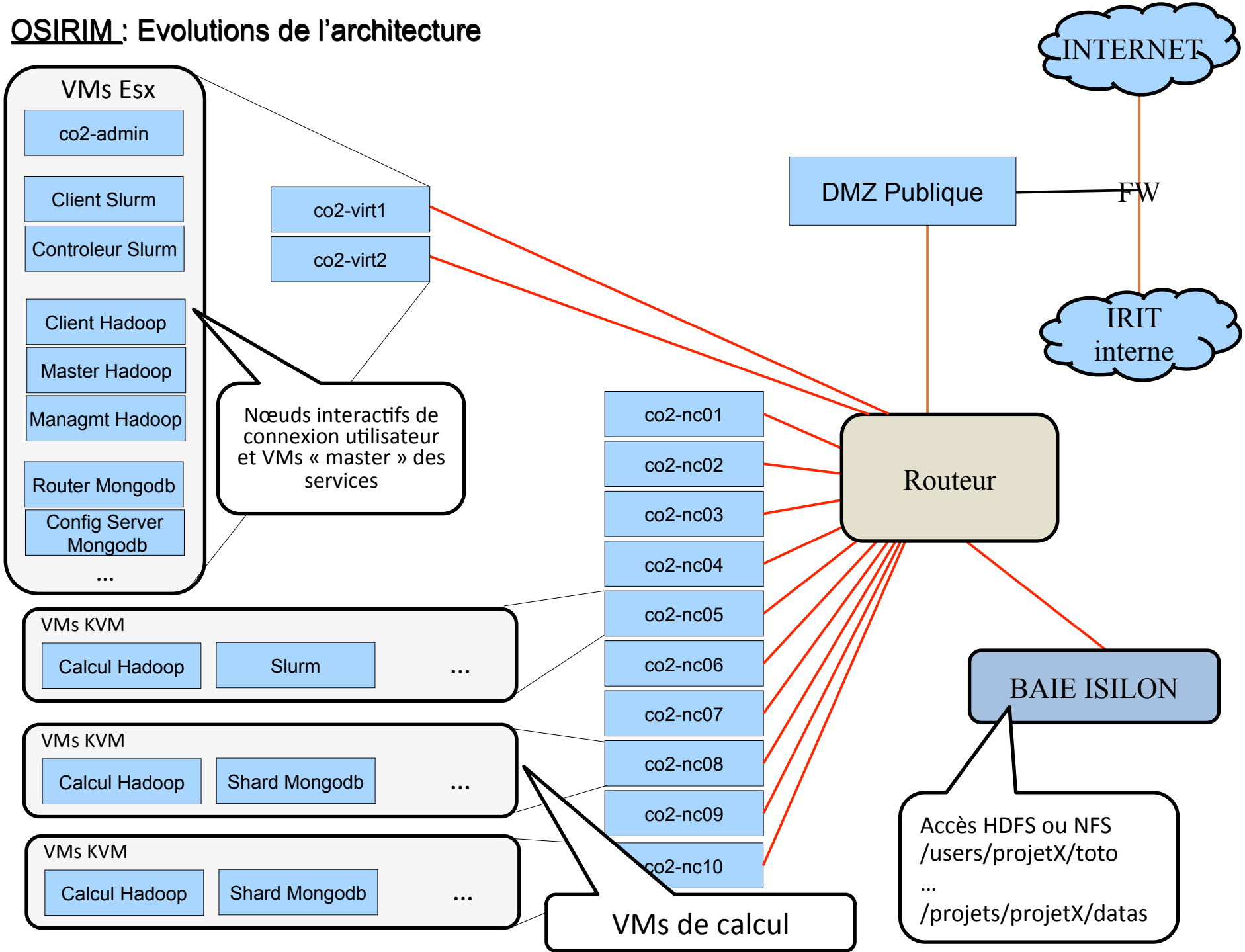
- Faire cohabiter des offres logicielles diverses
  - Slurm, Hadoop, Mongodb, Spark, …

- Ajuster le dimensionnement des services en fonction des demandes utilisateurs

=> Actions en cours :

- Virtualiser progressivement les noeuds de calculs existants sous KVM
- Activer ou désactiver à la demande les VM en fonction des contraintes de charge

OSIRIM : Evolutions de l'architecture



Merci de votre attention

- Questions ?

- Pour tout contact et demande d'hébergement :
  - http://osirim.irit.fr
  - osirim@irit.fr