IRIT
CNRS - INPT - UPS - UT1 - UT2J

Capitoul

La plateforme

OSIRIM

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

19 octobre 2017

cnrs INP TOULOUSE UNIVERSITÉ TOULOUSE III PAUL SABATIER UNIVERSITÉ TOULOUSE 1 CAPITOLE UNIVERSITÉ TOULOUSE Jean Jaurès

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 (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 tweets mondiaux (streaming), depuis septembre 2015.

- Partager des outils logiciels, par exemple pour l'évaluation de technologies :
  - Hadoop, Spark, Deep Learning, …

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 faire 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.
- Préciser les limites d'utilisation de la plateforme :
  - Plateforme dédiée à de l'expérimentation.
  - Aucun backup des données (pas d'engagement sur la conservation des données).

Projets hébergés

- Travaux de recherche des équipes :
  - SIG : intégration, Gestion NoSQL, Recherche, Fouille et Analyse dans les mégadonnées numériques, textuelles ou multimédias pouvant être structurées, semi-structurées ou non structurées.
  - IRIS : indexation et recherche d'informations dans de grandes masses de textes.
  - 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 et Compréhension d'Images.
  - …
- Projets :
  - QUAERO (terminé) : innovation sur l'analyse automatique et l'enrichissement de contenus numériques, multimédias et multilingues (IRIT/IRIS et SAMOVA, IRISA, Exalead (Dassault)).
  - SemDis: création de bases distributionnelles de référence pour le français.
  - CAIR: recherche agrégative de données (IRIT/IRIS, LIRIS).
  - Petasky : techniques de partitionnement de données issues du domaine de la cosmologie (LIRIS).
  - POLEMIC : analyse du comportement des utilisateurs dans les réseaux sociaux (IRIT/SIG, UAM Mexico).
  - COMPUBIOMED : Meta mining pour la recommandation en biosanté (IRIT/SIG, INSERM).
  - Tweet Contextualization : Contextualisation de tweets autour d'évènements (IRIT/SIG, Univ. Avignon).
  - LISTIC : projet visant à confronter les réseaux sociaux numériques aux réseaux sociaux reels
  - …

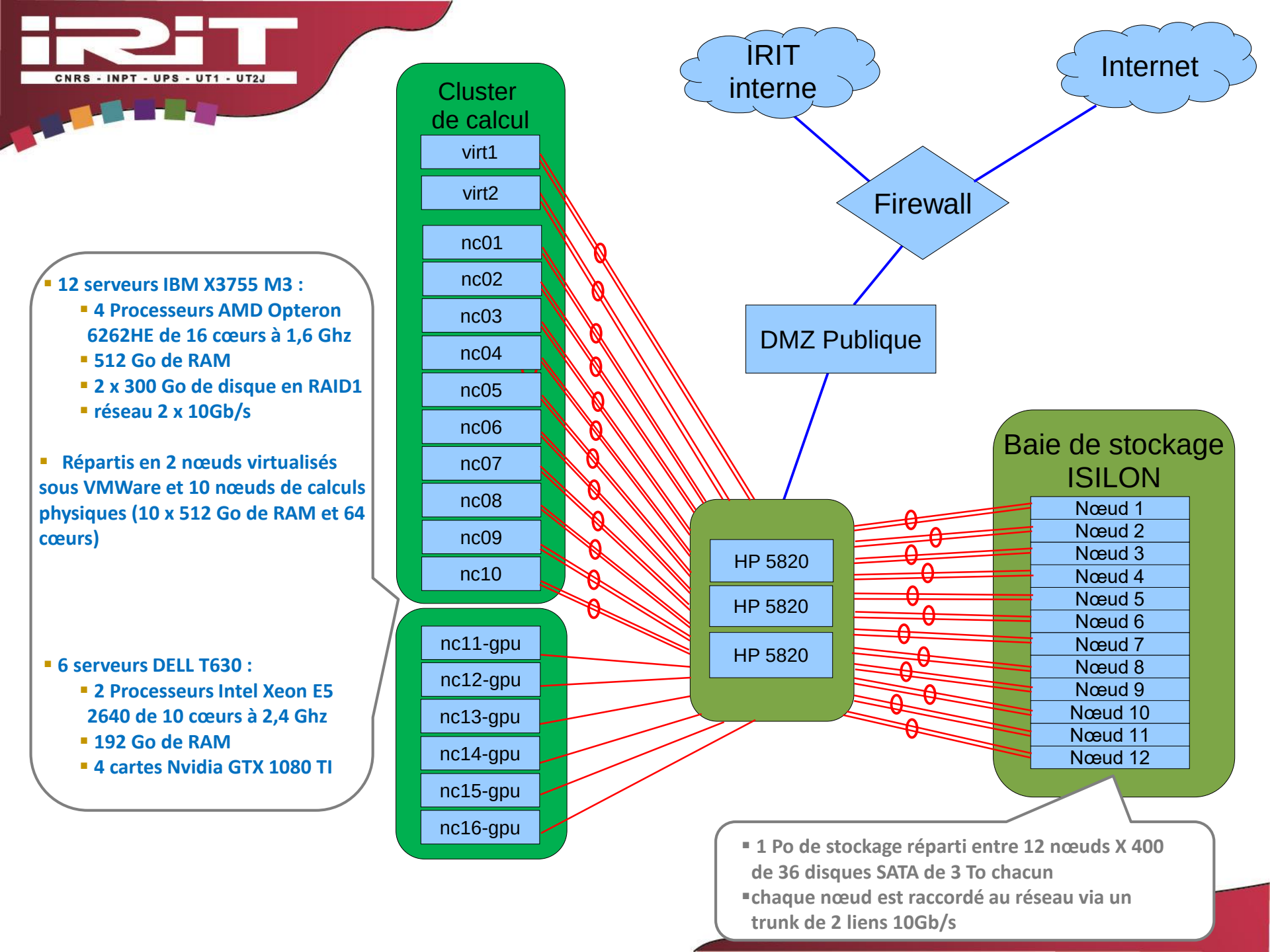
Mais aussi …

- Participations aux campagnes d'évaluation de systèmes de recherche d'informations :
  - TREC (Text Retrieval Conference), INEX (XML Retrieval), CLEF (Cross Language Evaluation Forum), TrecVid (TREC Video Retrieval Evaluation), mais aussi OAEI (Ontology Alignment Evaluation Initiative).

- Soutien pour l'initiation à la recherche dans des formations de master :
  - Master SID Université Toulouse 3 : apprentissage de technologies Hadoop (Hive).
  - Master M2 IT/ Enseeiht : Fouille de tweets.

- Accompagnement d'évènements spécifiques :
  - Hackday CORIA/CIFED 2016.

- Mise à disposition d'un espace de stockage pour Grid 5000



- 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)
- 6 serveurs DELL T630 :
  - 2 Processeurs Intel Xeon E5 2640 de 10 cœurs à 2,4 Ghz
  - 192 Go de RAM
  - 4 cartes Nvidia GTX 1080 TI

- 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 …

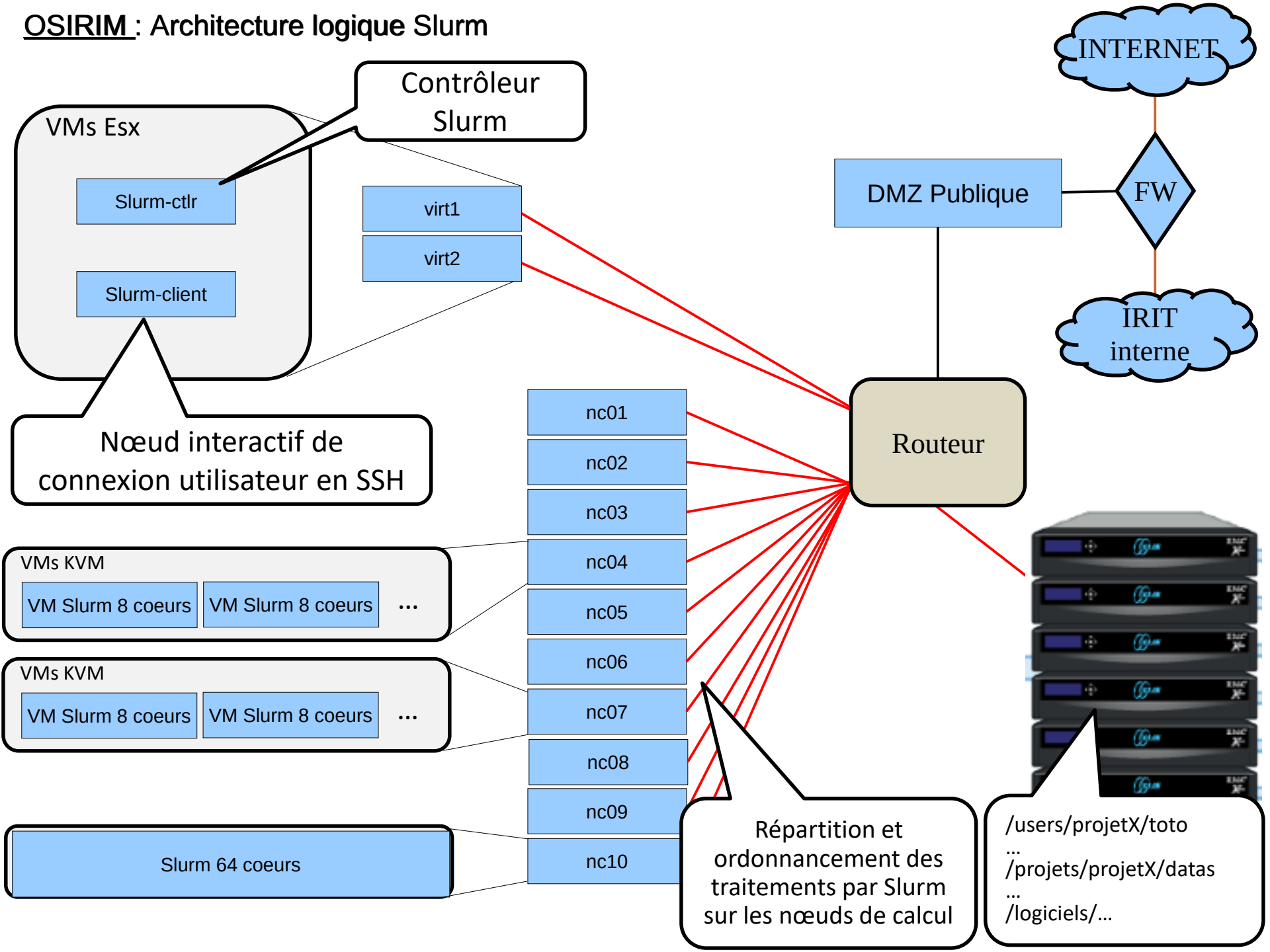
Une offre de services articulée autour de deux approches de distribution des traitements

- Un gestionnaire de jobs et de ressources SLURM (Simple Linux Utility for Resource Management) permettant la distribution de traitements réalisés avec des langages / logiciels mutualisés : C++, PYTHON, JAVA, R, …

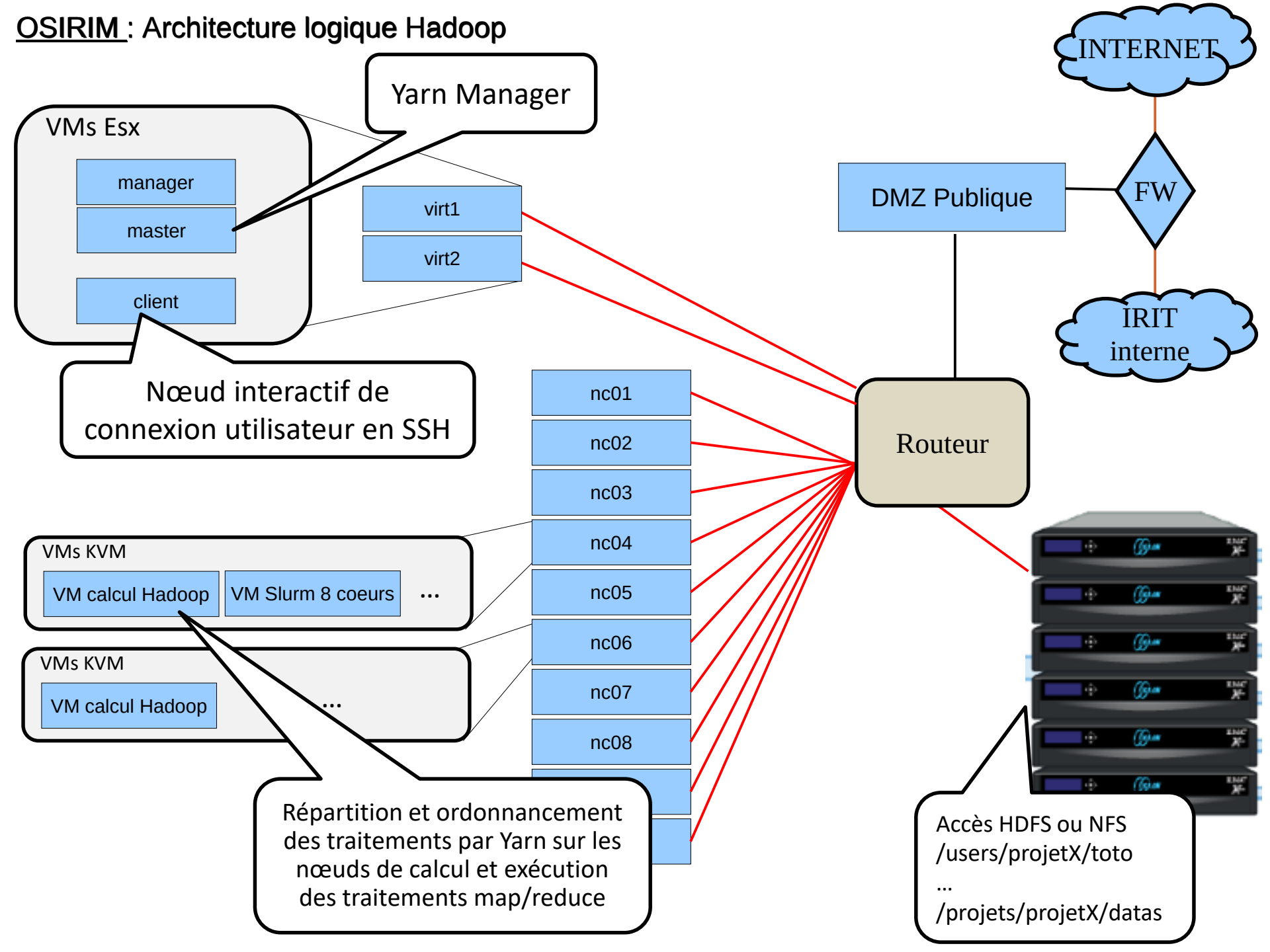
… et des données

- Une distribution HADOOP (Hortonworks HDP) avec son écosystème applicatif : SPARK, HIVE, PIG, HBASE, FLUME, …

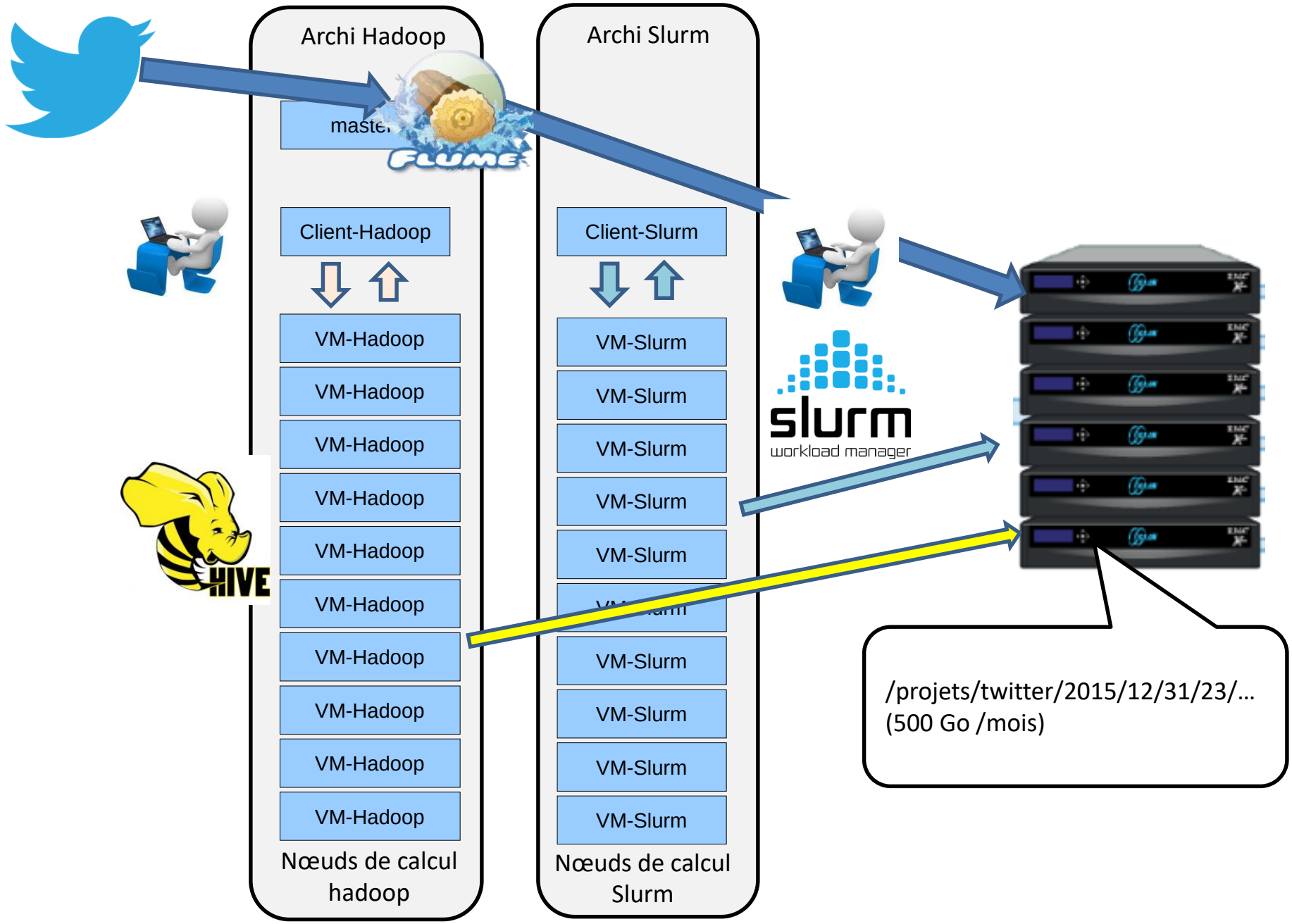
OSIRIM : Architecture logique Slurm



OSIRIM : Architecture logique Hadoop



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

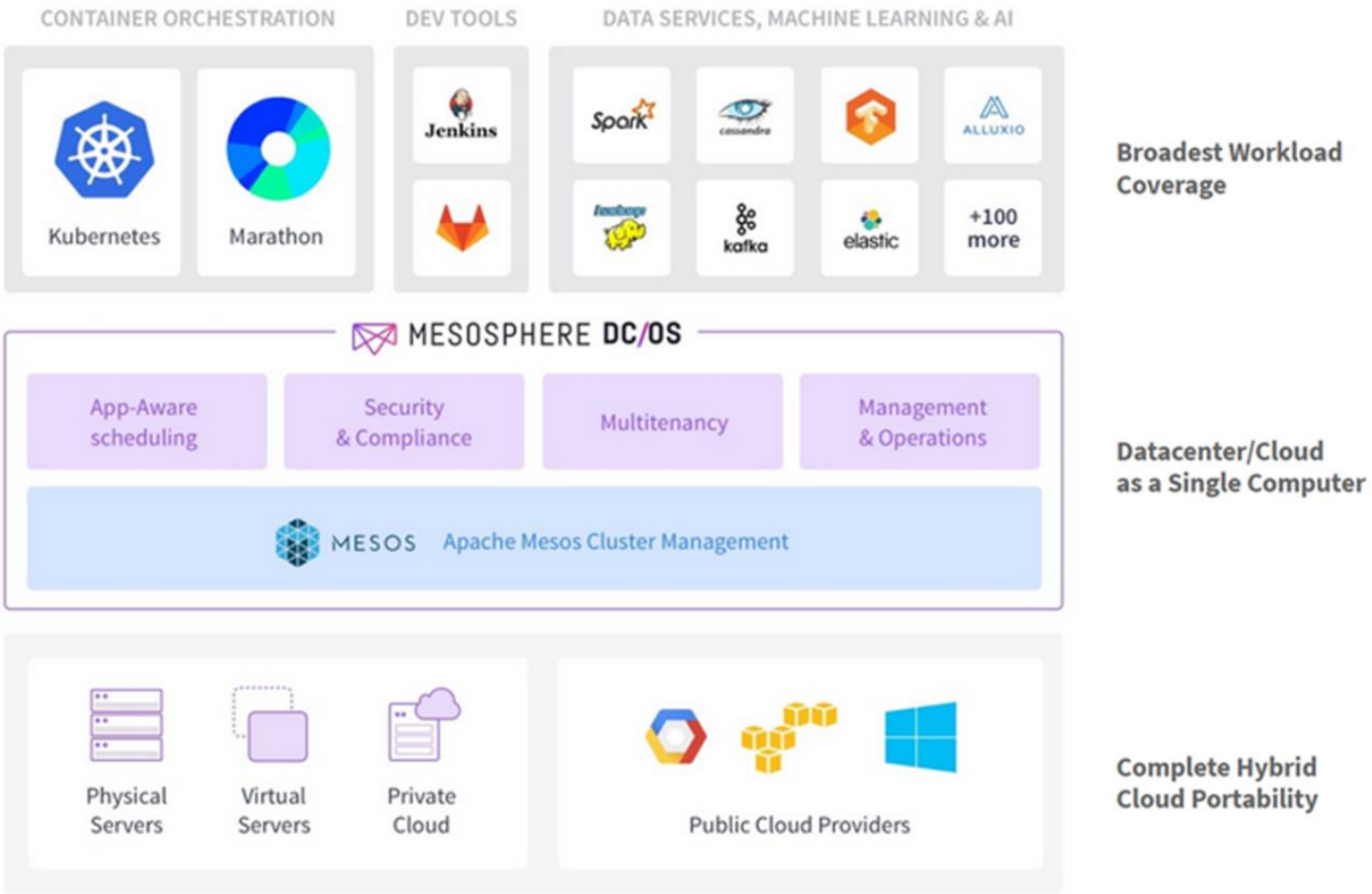


Mais aussi …

- Hébergement de demandes spécifiques par déploiement de machines virtuelles dédiées :
  - Projet LISTIC (Architecture lamp + moissonneur Twitter DMI-TCAT)
  - Neo4j
  - MongoDB
  - …
- Mise à disposition de frameworks de DeepLearning : Caffe, TensorFlow, Theano, Torch
  - Sous forme de conteneurs Docker.
  - Traitements exécutés dans une partition Slurm dédiée contenant l'ensemble des ressources de calcul GPU

A terme …

- Faire évoluer Osirim vers une architecture proposant des services sur étagères …
- Maquette DC/OS



Merci de votre attention

- Questions ?

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