Statut du cloud, CC-IN2P3

JI IN2P3-IRFU

26-29 sept. 2016

Au menu

Les types de cloud au CC-IN2P3

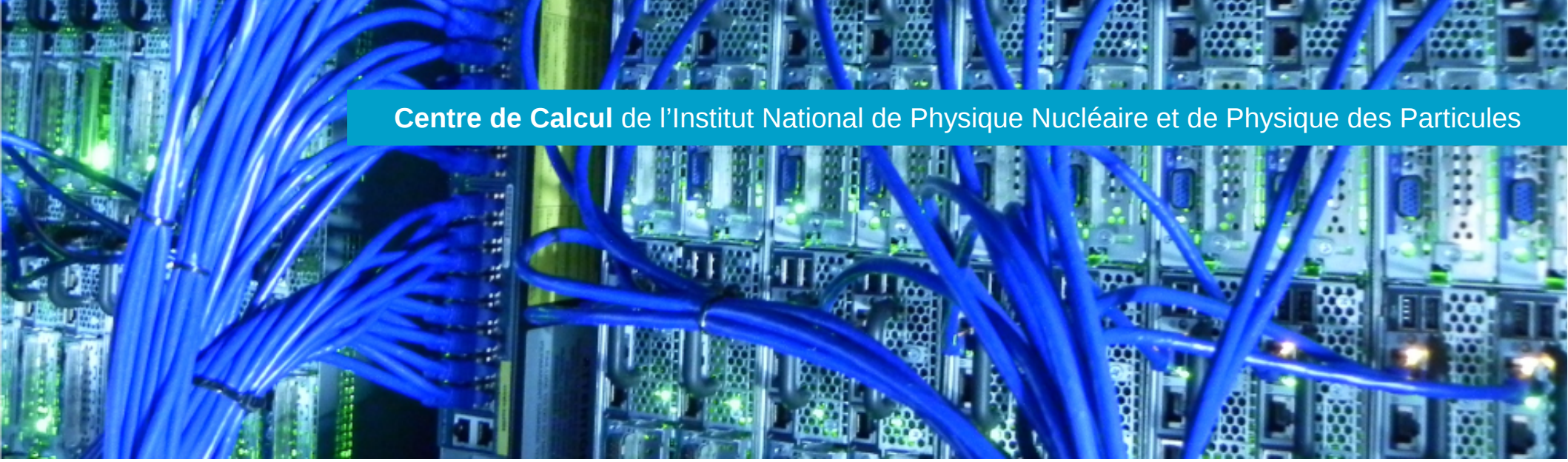
- R & D
- Production
- HPC
- Hébergés

L'infrastructure

- Déploiement
- Schémas & topologie
- Développement
- Consolidation du monitoring
- Benchmarks & performance
- Réseau et OpenStack Neutron

Le futur

- A court terme
- A long terme



LES TYPES DE CLOUD AU CC-IN2P3

Les clouds , n°1 – R & D

Pour qui ?

- Les équipes du CC-IN2P3
- Les laboratoires et expériences

Pour(quoi) ?

- En dehors des services de production
- Tester de nouvelles technologies facilement
- Des ressources en quantité (RAM, CPU, HDD)
- Performance (I/O, CPU et RAM)
- Sauvegarde (snapshot)

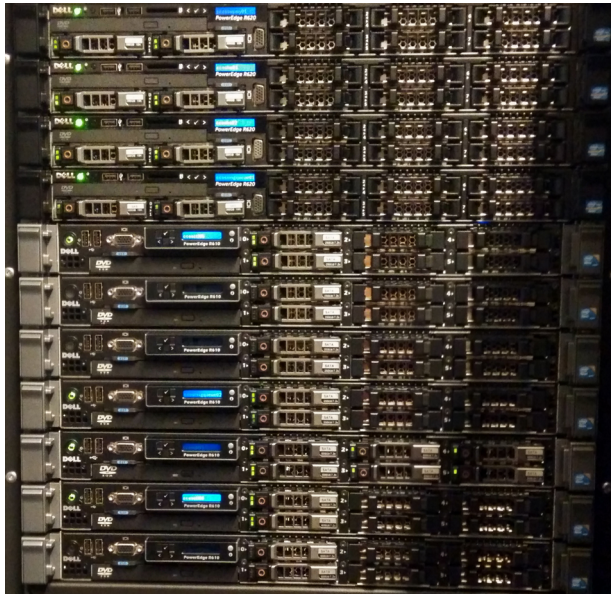
Quelques exemples de services.

- Pre-production (upgrade…)
- Tests de Spark
- Développements TreqS
- CI
- Formations (Puppet, COLOSS…)

Les clouds, n° 2 - Production

Pour qui ?

- Les équipes du CC-IN2P3

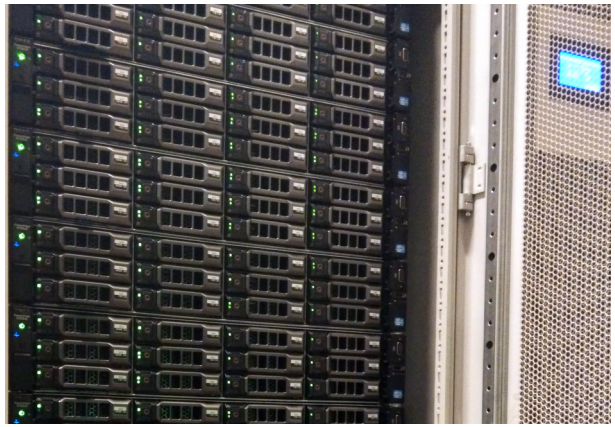


Pour(quoi) ?

- Décommissioner VMWare (coût)
- Flexibilité de déploiement (à la demande)
- Performance (I/O, CPU et RAM)
- Haute disponibilité (cluster GPFS, live mig.)

Quelques exemples de services.

- ElasticSearch
- Kibana
- Grafana
- Kerberos5
- Services Grid
- Puppet Server
- +80 autres



Les clouds, n°3 - HTC

Les motivations du passage au cloud

- Accéder à des ressources de façon opportuniste
- Des environnements prédéfinis et connus (OS, softwares...)
- Faciliter le déploiement des logiciels
- Une politique de scheduling précise et personnalisable
- Mise en place d'un modèle de computing plus facilement

Les alternatives au modèle cloud

- Évaluation du scheduler HTCondor et Dirac
- Utilisation des interfaces d'API (EC2/Nova...), Fair Share Scheduling
- Faciliter l'instanciation de nœuds de calcul (Puppet)

Les expériences utilisant le cloud

- Large Synoptic Survey Telescop (http://www.lsst.org)
- Euclid (http://www.euclid-ec.org/)
- Atlas (MC simulation)
- Bioaster (http://www.bioaster.org)
- ELISA (www.elisascience.org)

Les clouds, n°4 - Hébergés

Pour qui ?

- Les utilisateurs des laboratoires de l'IN2P3
- Les expériences

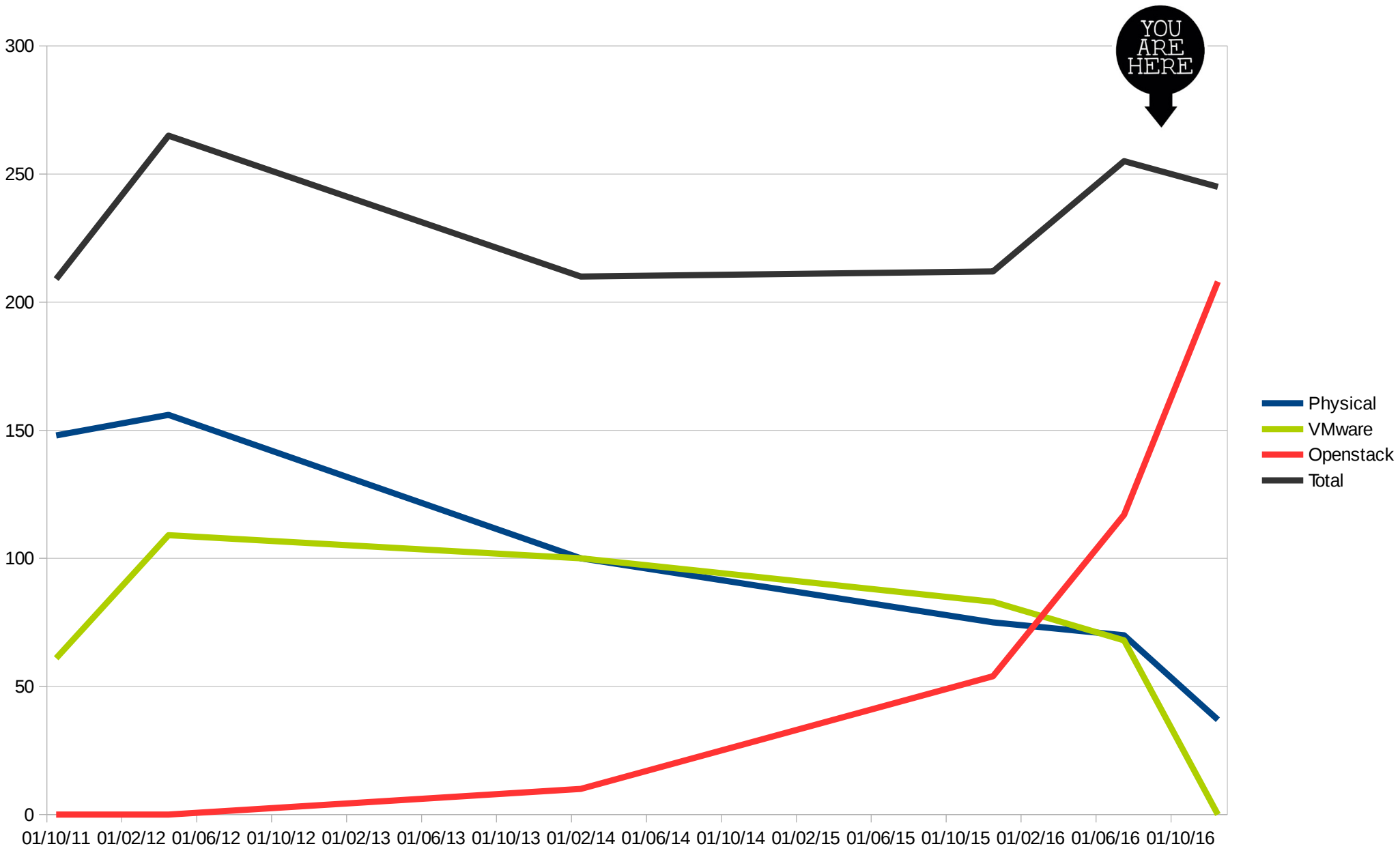
Pourquoi ?

- Profiter de l'expertise du CC-IN2P3
- Réduction des coûts de maintenance et d'exploitation
- Abstraction de la difficulté technique, démarrer le projet plus vite

Quelques projets hébergés

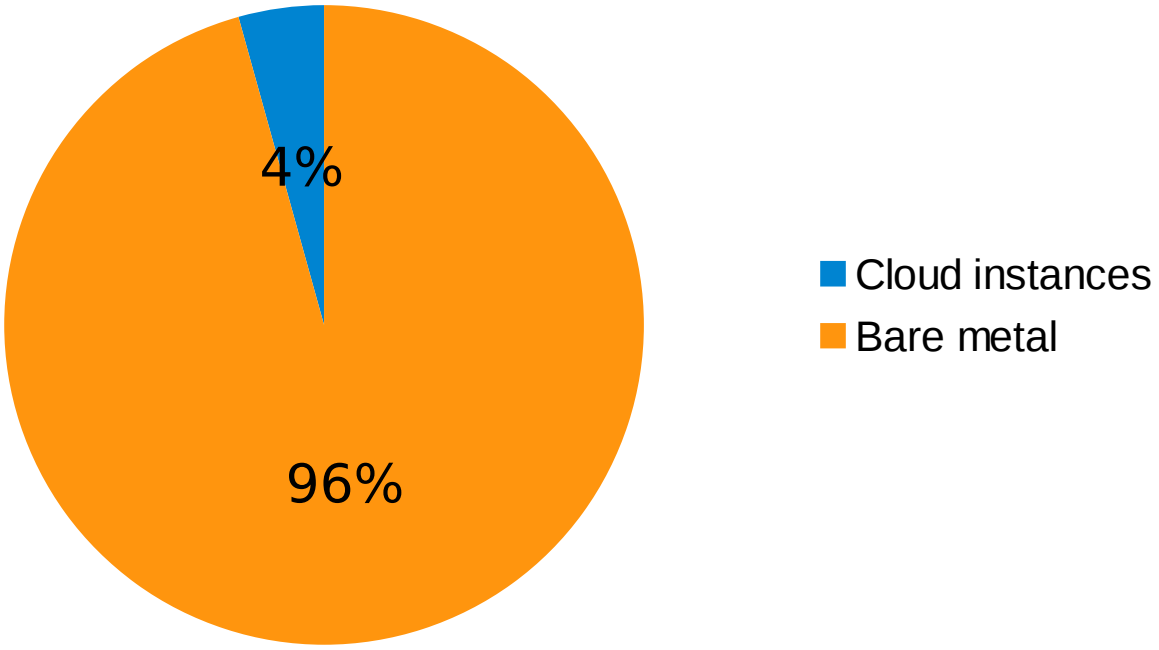
- AMI
- Bioaster
- eTRiKS
- CODEEN (CNES EUCLID)

Déploiements de services CC



En nombre de machines dédiées aux services CC

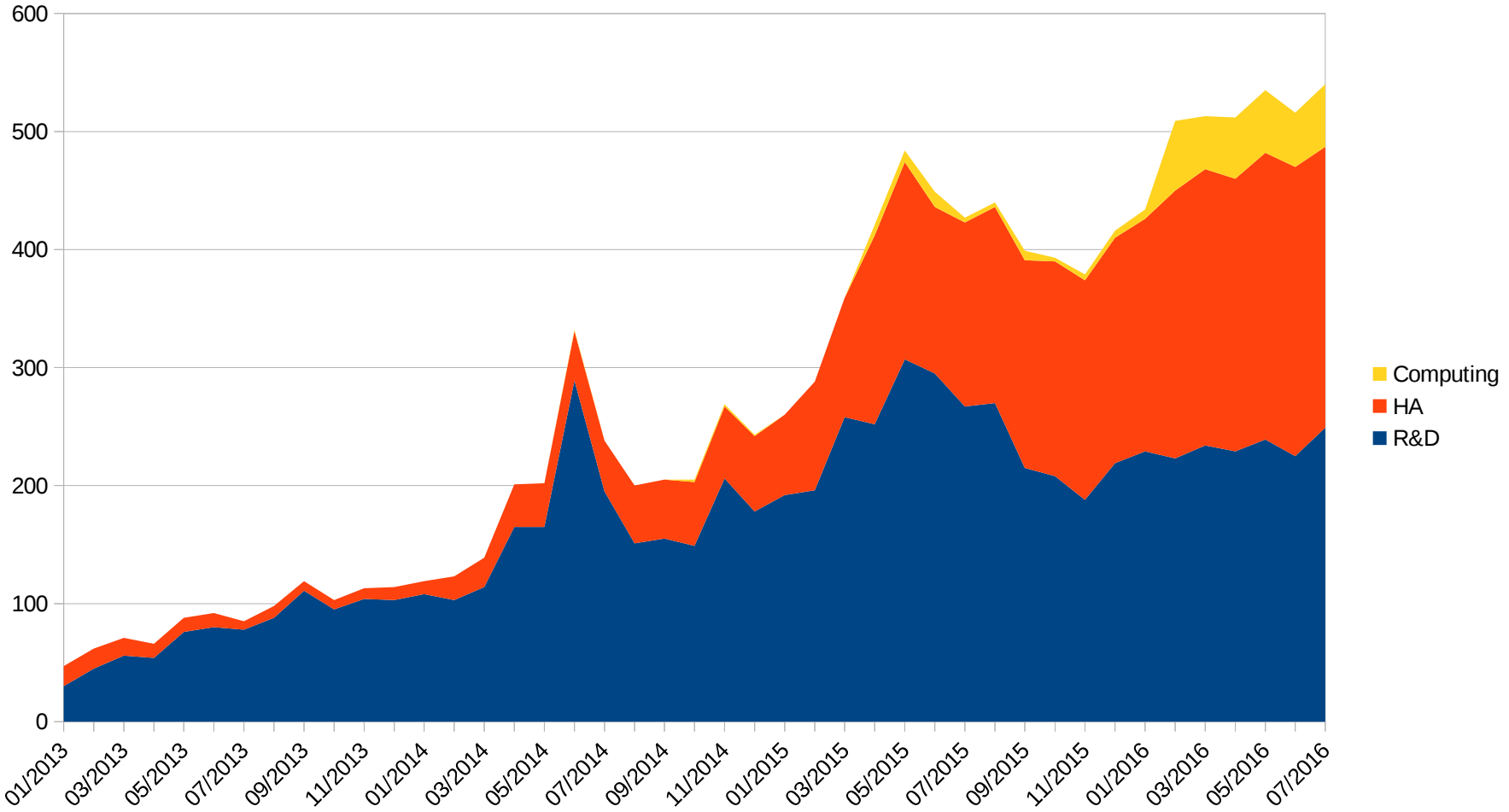
HTC, cloud vs bare



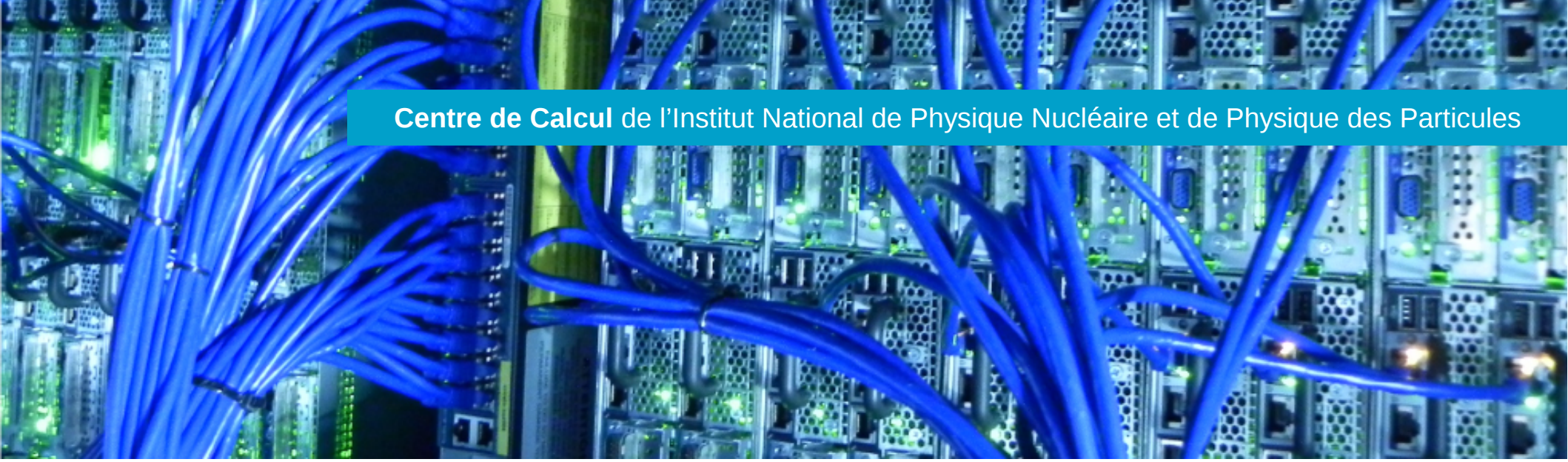
Comparatif HTC

- ~ 1056 cœurs HT en rapport des 25 000 de la ferme HTC (4.2%)
- Passage à 2160 cœurs HT bientôt (gain de 4.4%)

Statistiques par usage

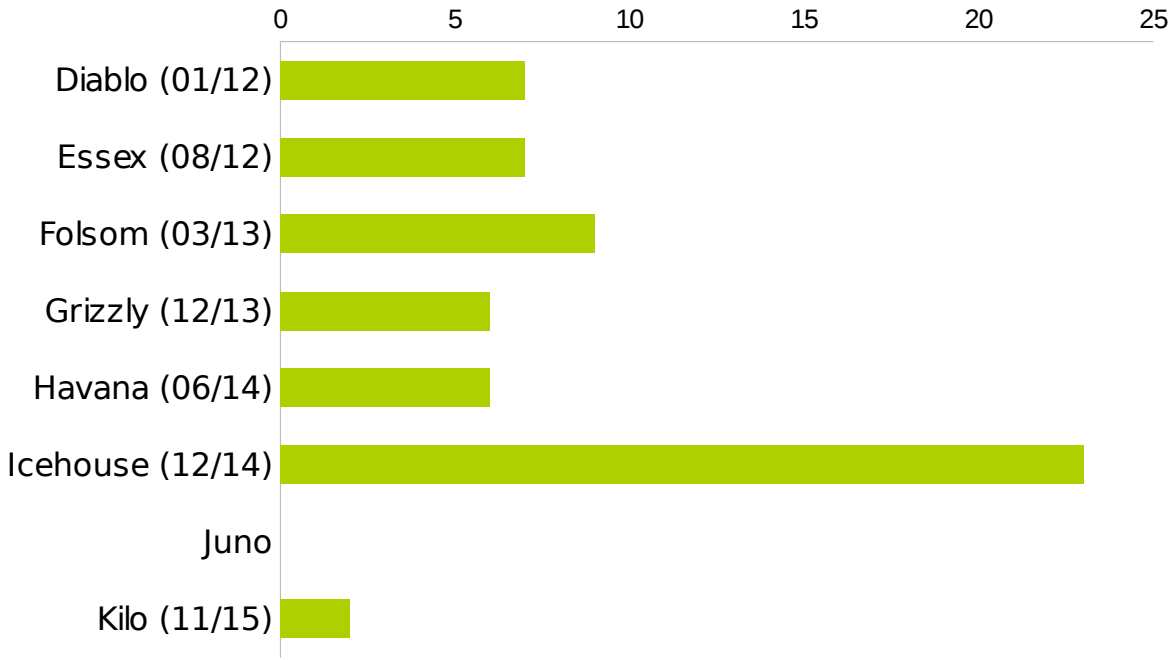


- En nombre d'instances actives
- projets 50
- utilisateurs 150
- computes 88
- agrégats 16



INFRASTRUCTURE

OpenStack au CC-IN2P3



En mois d'utilisation

Composants opérationnels :

Keystone
Glance
Nova
Horizon
Swift
Rally
Ceilometer
Cinder

Composants en cours :

Heat
Neutron

Le déploiement, les chiffres

Software

- CentOS 7
- Paquets RedHat RDO
- Déploiement et gestion de conf. par Puppet

Hardware (mai 2016)

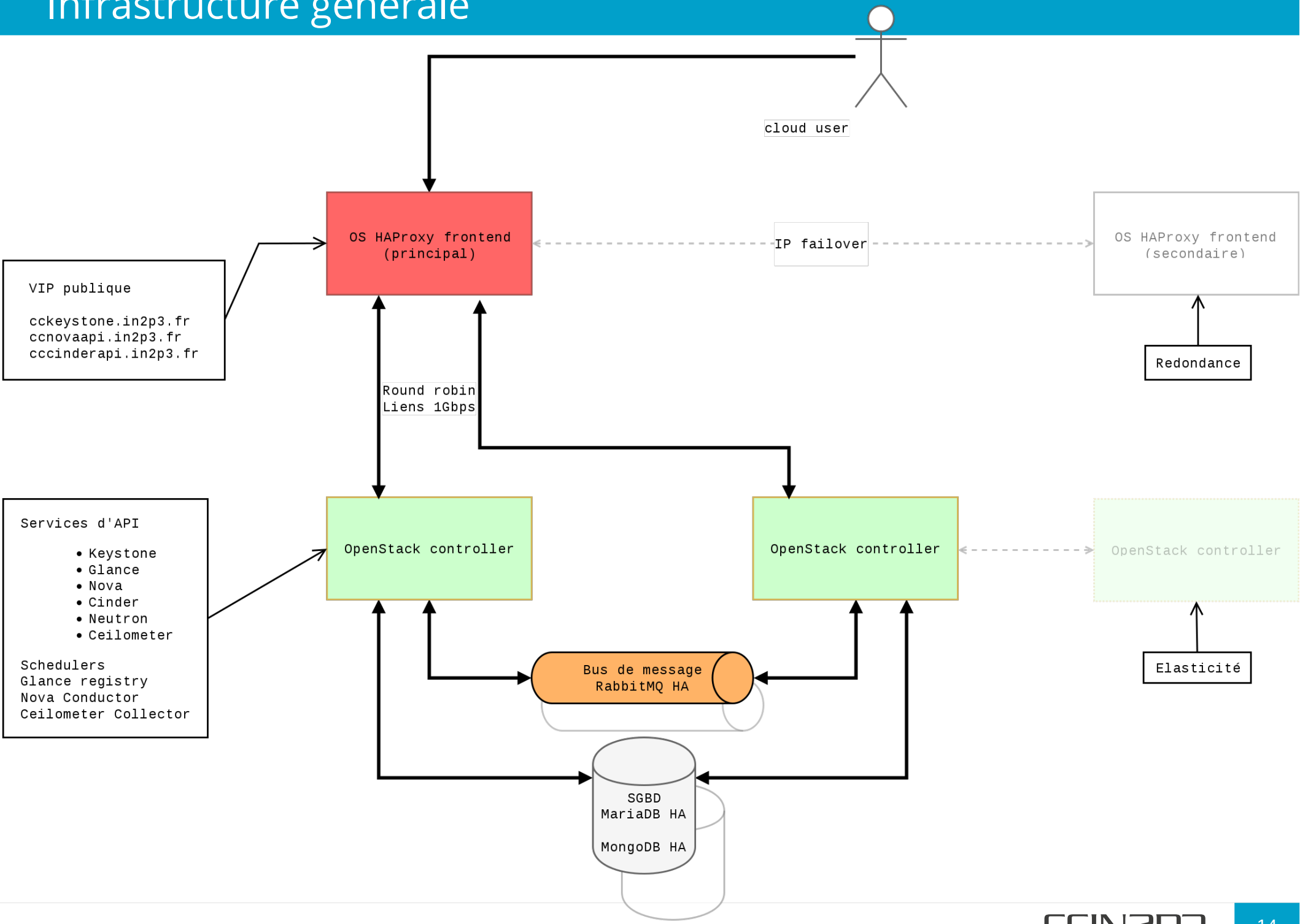
- R & D = 288 cœurs HT Xeon X56xx
- Prod. = 160 cœurs HT Xeon E5 26xx
- HTC = 1056 cœurs HT Xeon X56xx
- Héber. = 320 cœurs HT Xeon X56xx et E5 26xx
  - **Total de 1824 cœurs HT en production**

- **8905GB de RAM en production**

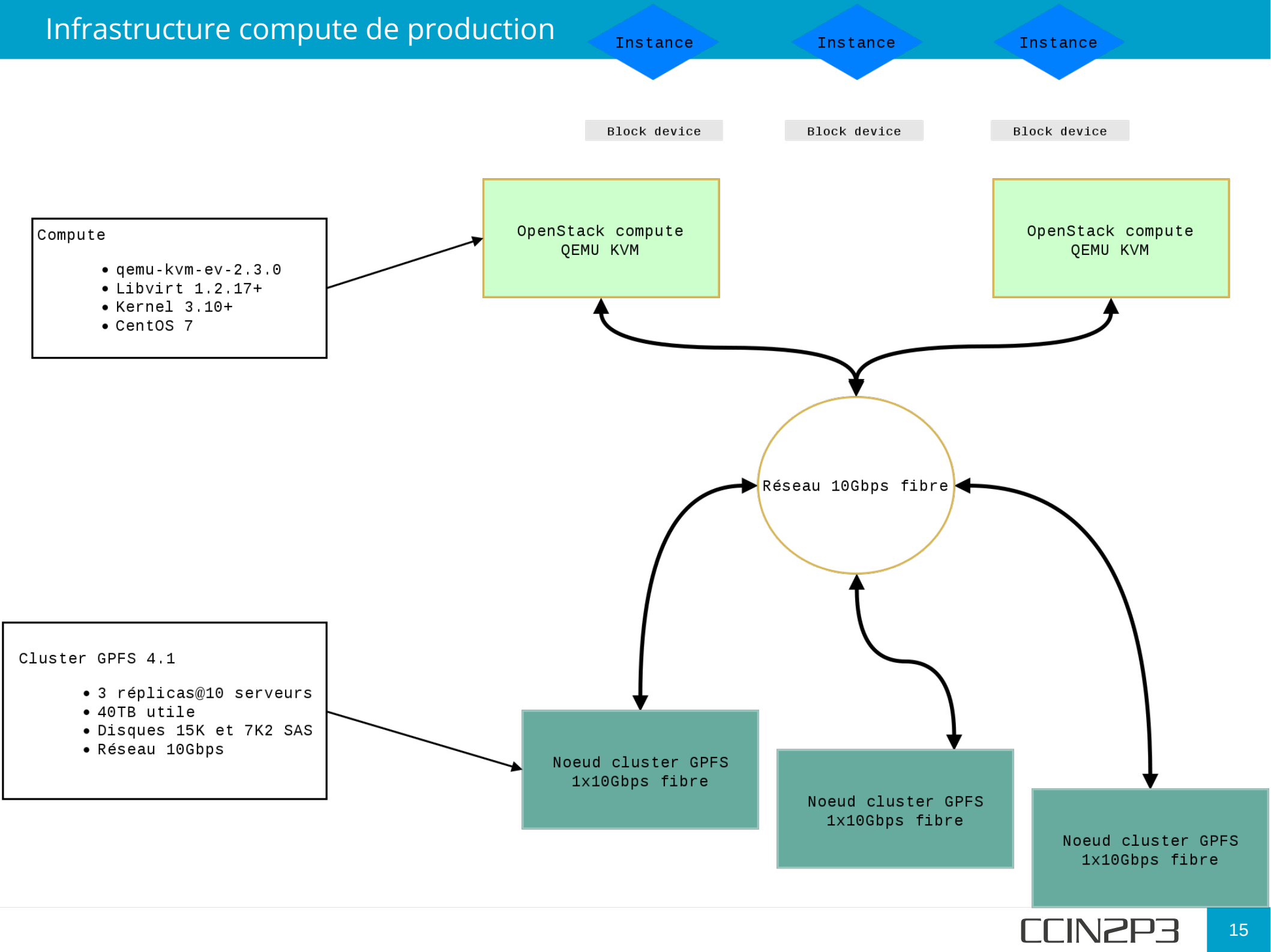
- 40TB de stockage distribué sur 10 serveurs GPFS
- 30TB de stockage orienté objet Swift
- 9TB de stockage rapide (15K) orienté block Cinder
- 24TB de stockage lent (7K2) orienté block Cinder
  - **Total de 113TB de stockage dédié au cloud**

- **Réseau 1/10Gbps Ethernet non bloquant**

Infrastructure générale



Infrastructure compute de production

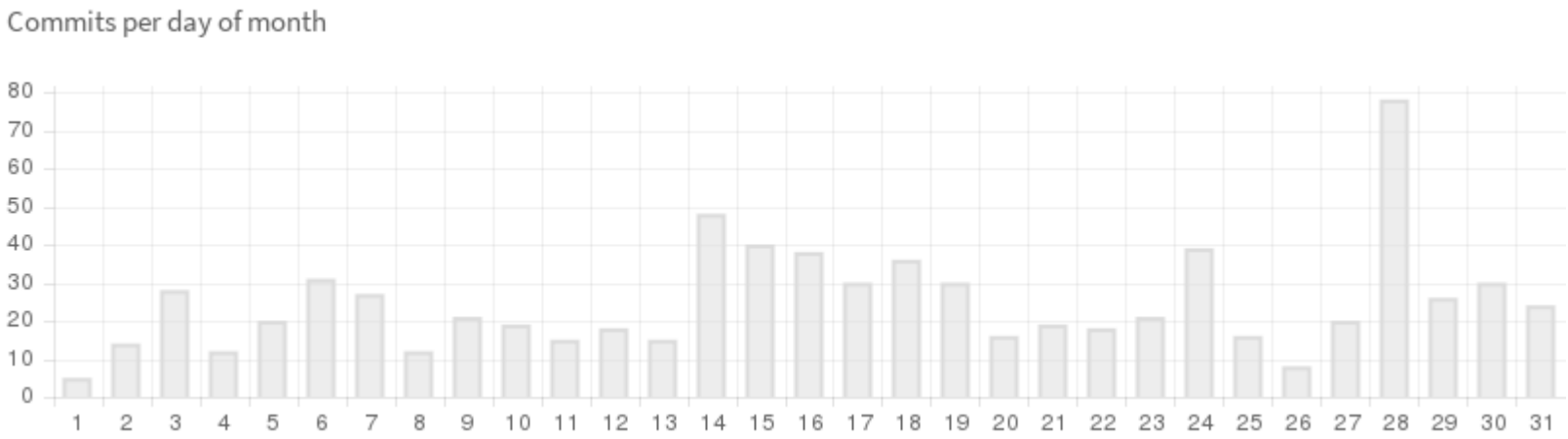


Méthodologie

- Workflow git-flow pour toute l'équipe
  - Master – develop – feature - hotfix
- Toutes les données dans hiera (YAML)
  - Secret chiffré par GPG
- Un module site_openstack pour tous le déploiements
- Un ENC (smurf) maison pour qualifier les machines
  - par usage (openstack)
  - et rôle (compute, controller, networking node...)
- Semantic versioning (vX.X.X)

Commit statistics for **master** Nov 05 - Sep 20

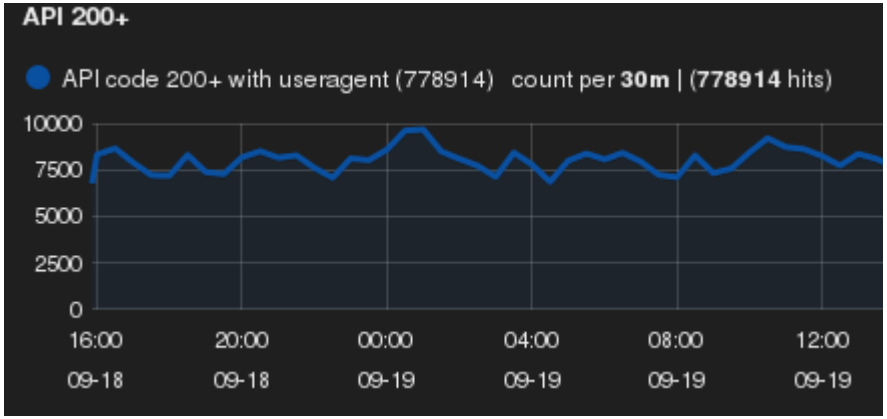
- **774** commits during **685** days
- Average **1.1** commits per day
- Contributed by **7** authors



Infrastructure de monitoring

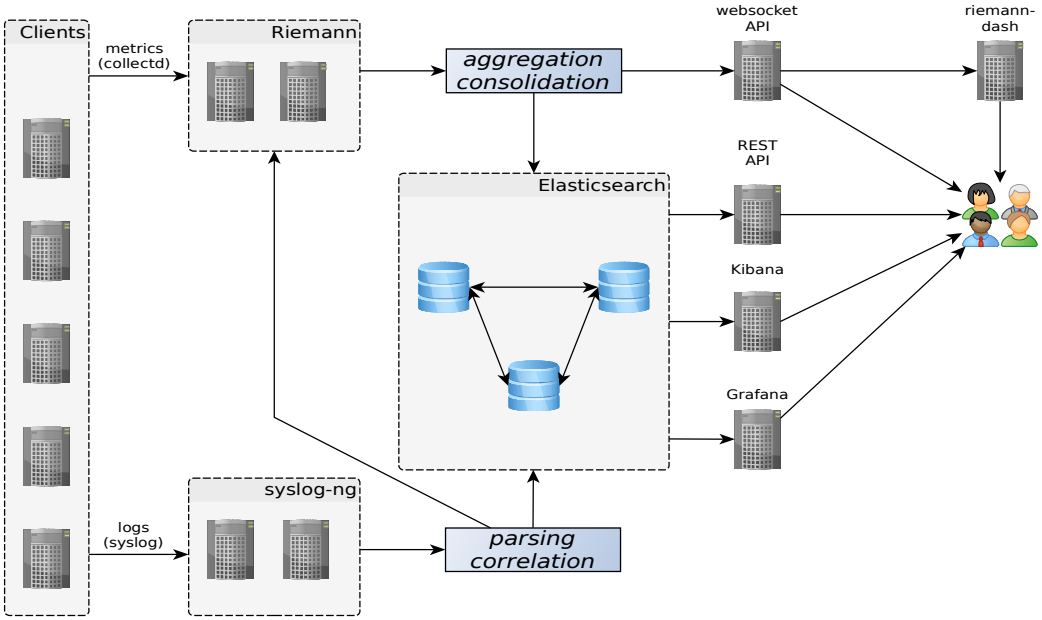
Gestion des logs

- Collecte et envoi par syslog-ng
- Format JSON, indexé par clef
- Stocker dans ElasticSearch
- Visualisation dans Kibana
- Patch Oslo.log pour tout formater en JSON



Gestion des métriques

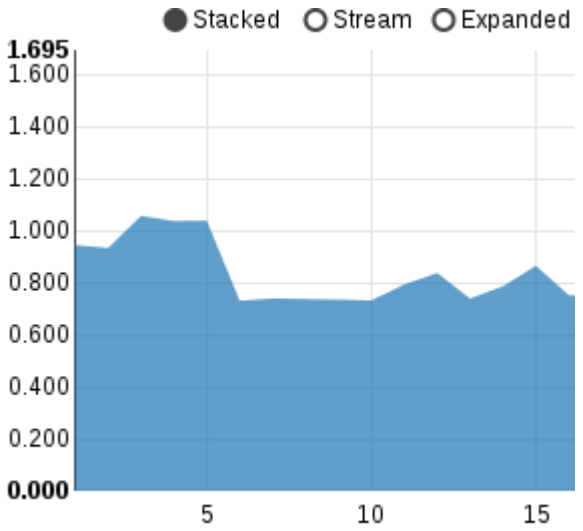
- Collecte et envoi par collectd
- Manipulation par Riemann
- Stocker dans ElasticSearch
- Visualisation dans Grafana



Infrastructure benchmark

Mise en place d'OpenStack Rally

- Déploiement Puppet
- Test de fonctionnalités, boot, delete, attach...
- Test de performance, durée d'exécution
- Historisation des comportements
- Test de Service Level Agreement



Total durations

Action	Min (sec)	Median (sec)	90%ile (sec)	95%ile (sec)	Max (sec)	Avg (sec)	Success	Count
nova.boot_server	12.548	12.826	17.823	18.731	19.639	14.552	100.0%	5
nova.reboot_server	4.48	4.523	4.853	4.914	4.976	4.63	100.0%	5

Gestion de la performance

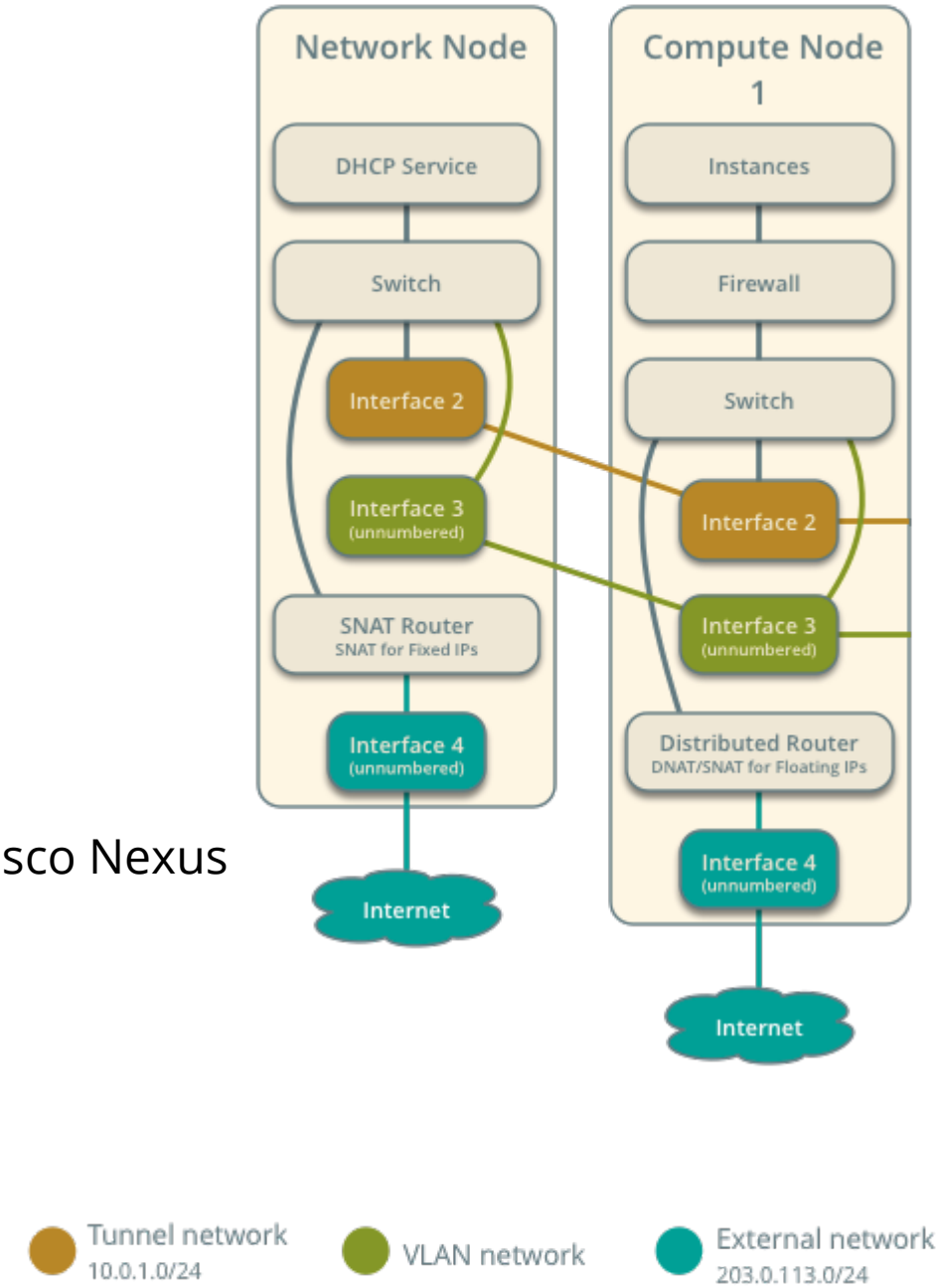
- MonteCarlo12 et HSPEC06
- Placement NUMA (bi-socket)
- Pinning de thread, pas de re-scheduling
- Hugepages de 2MB via TLBfs, - de hit/miss
- QEMU 2.3+
- Corrélation entre l'HT et la gestion cache L2
- Intel CMT & CAT pour le cache L3
- Intel Xeon E5-2680 v2 Sandy Bridge

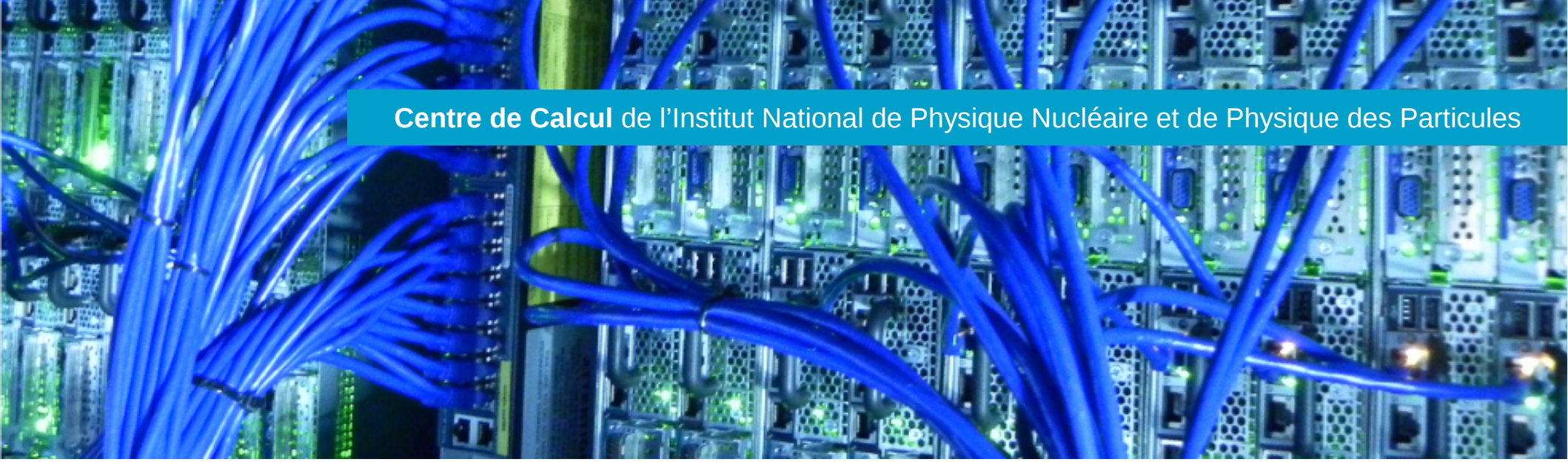
Migration de Nova-Network à Neutron

- Nova-network depuis ~ 2012, 30aine de VLAN
- Prévue pour Décembre 2016 sur une journée
- Passage à OpenVSwitch pour le L2
- Passage aux namespaces iptables pour le L3
- Même adressage et fonctionnalités

Déploiement Neutron

- D'ores et déjà Puppetisé
- Testé en pré-production
- Choix du mode Distributed Virtual Router
- IPs publiques routées directement par notre Cisco Nexus
- SNAT des IPs privées pour accès Internet
- Ségrégation des projets par VLAN ou VXLAN



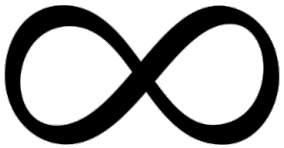

Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules

LE FUTUR

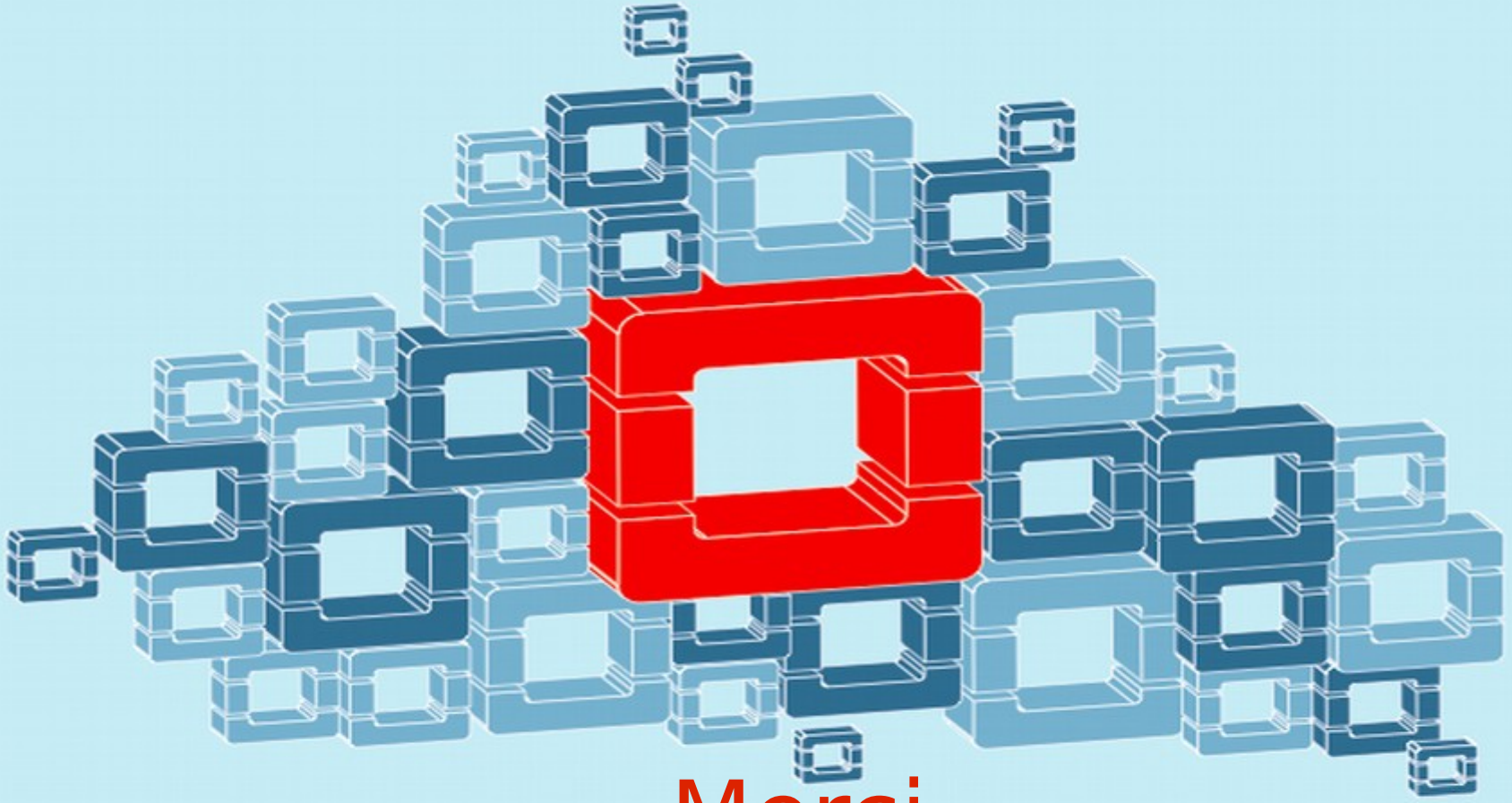
A court terme

- Fair Share Scheduling pour la fin 2016
  - Développement de la solution Synergy pour OpenStack
  - Projet européen INDIGO Datacloud
- Mise en production d'OpenStack Heat, template de déploiement
- Continuer la gestion de la performance HTC, overhead < 5 % et IO
- Finaliser la compatibilité des containers
  - Bug devicemapper
  - Ethernet LRO offloading et Intel IXGBE + ip_forwarding/bridge = timeout
- Intégration des métriques Ceilometer dans la plateforme de BI Symod

A long terme



- Récupération des ressources automatiquement (Freezer ?)
- IPv6 de bout en bout
- Étude et mise en place d'un nouveau backend Cinder
  - Probablement Ceph ou GPFS
  - Sensibilité trop importante avec LIO et ISCSI Kernel
- Travail autour de l'orchestration des containers
  - OpenShift
  - Kubernetes
- Consolidation de l'infrastructure d'images Glance
  - Séparer les flux de transfert de ceux d'API, management
  - Augmenter la capacité de transfert (snapshot, boot)
  - QoS, monitoring



Merci

Questions ?