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

Compte rendu
HEPiX de printemps 2016
Zeuthen, 18-22 avril 2016

LABORATOIRE DE L'ACCÉLÉRATEUR LINÉAIRE

LAPP 40 ans

CCIN2P3

25th Anniversary

HEPiX

Zeuthen Spring 2016

cnrs

lafa

- 11 participants français issus de 4 laboratoires
  - CEA/Irfu :
    - Pierre-François Honoré et Joël Surget
  - LAL
    - Valérie Givaudan et Michel Jouvin
  - LAPP
    - Frédérique Chollet
  - CCIN2P3
    - Foudil Bretel, Pierre-Emmanuel Brinette, Bernard Chambon, Nicolas Fournials, Sébastien Gadrat et Fabien Wernli

Plan de la présentation

- Le workshop en quelques chiffres
- Les tendances
- Les points forts par tracks :
  - Site reports (15)
  - End users and services (5)
  - Security and networking (7)
  - Grid, cloud and virtualization (10)
  - Storage and filesystems (12)
  - IT Facilities and business continuity (6)
  - Computing and batch systems (6)
  - Basic IT services (7)
- Prochains HEPiX

Le workshop en quelques chiffres

- 119 participants (record HEPiX@Oxford avec 134)
  - 12 participants d'Amérique du Nord
  - 84 d'Europe
  - 11 d'Asie
  - 12 de compagnies privées

- L'Asie participe de plus en plus
- Participation plus variée (en dehors de la HEP)
- 70 présentations (programme toujours dense)
  - Programme détaillé

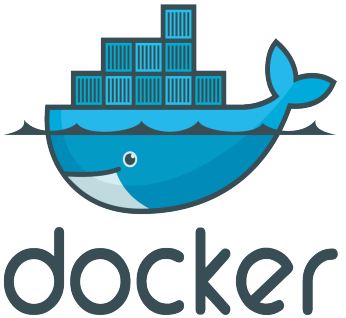
- *Disclaimer : présentation non exhaustive...*

Les 25 ans d'HEPiX...

- Calcul : grille et cloud
  - HTCondor
  - OpenStack
- Conteneurs/Docker
- Stockage
  - CEPH
  - (OpenZFS, surtout en backend de Lustre)
- Monitoring
  - Stack ELK (ElasticSearch, Logstash, Kibana)
  - Grafana
  - InfluxDB, Flume
- Gestion de configuration
  - Puppet/Foreman

Les tendances (qui montent) – Impressions très subjectives

- Calcul
  - HTCondor+HTCondorCE
- Conteneurs
  - Docker
- Stockage
  - EOS (utilisé et testé en dehors du CERN)
- Orchestration avec conteneurs et micro-services
  - Apache Mesos
  - Google Kubernetes
  - Docker Swarm/compose
  - Rancher http://rancher.com
- Réseau : Software Defined Networking
- Monitoring : Metadata http://metrics20.org
- Développement : GitLab/GitLabCI



- The State of OpenAFS - Stephan Wiesand
  - « This is not an « official » OpenAFS presentation »
  - Prochaine branche stable attendue 1.8, mais
    - Baisse importante du nombre de commits depuis 1-2 ans
    - Plus que 4 contributeurs réguliers
    - Le développeur qui s'occupait du portage Linux est parti et n'est toujours pas remplacé
      - Pas de version stable kernel 4.4, aucun travail sur kernels 4.5 et 4.6
  - Mais projet en survie… HEP était le principal utilisateur...
- CERN Site Report - Arne Wiebalck
  - Retrait d'AFS décidé, pas d'échéance fixe, mais complétion durant le LS2 (2019)
    - Pas de solution unique de remplacement
      - CVMFS (logiciels)
      - EOS(+Fuse) et CERNbox couvre la majorité des cas
      - Discussion toujours en cours pour le $HOME

- FERMILAB : Scientific Computing Storage Architecture
  - Projet de consolidation et d'évolution des différents moyens de stockage et d'accès aux données
  - Pendant du projet Fermilab HEPCloud (voir plus loin)
  - Suppressions des accès POSIX (fichiers) depuis les WNs
    - Suppression des systèmes de montages distants NFS
    - Adapter les workflows de traitement des expériences pour utiliser indifféremment des ressources locales, grille et clouds commerciaux
  - Grille d'analyse pour catégoriser :
    - les types de stockage (MSS, disque, sandbox, ...)
    - les mécanismes d'accès aux fichiers (depuis les jobs)
    - les types de fichiers en entrée et en sortie
    - les protocoles utilisés
  - Accompagnement au changement expérience par expérience
    - Exemple réussi : MINERvA

- Spectrum scale : Nouveautés dans la version GPFS 4.2
  - Une CLI d'administration unifiée et une GUI
  - Support stockage objet
  - Support natif de hadoop
  - Compression de données « froides »
  - Débordement du stockage sur un cloud transparent basé sur des règles LWE.
- NDGF : Efficient Nordic dCache Interface to TSM (ENDIT)
  - Interface d'interconnexion entre dcache et TSM basée sur un plugin natif.
  - Pas d'appels à des scripts
- Status report of TReqS
  - Un astucieux outil d'optimisation de l'accès aux bandes magnétiques

- Lustre/ZFS Development - Walter Schoen
  - Quelques avantages du FS
    - 128 bits (stockage 2^{64} plus important que les FS actuels 64 bits)
    - Intégrités des données et correction d'erreur
      - Protection contre la corruption « silencieuse »
    - Intégrité et réparation pour les volumes montés par scrub
    - Instantanées
  - ZFS utilisé en backend de Lustre (production > 1 an)
    - Bonne expérience : stable et fiable
  - Développement d'une version optimisée du ZRAID basée sur la vectorisation CPU
  - Analyse des logs de Lustre par un modèle de Markov caché (automate de Markov à états cachés), Hidden Markov Model
    - But : prédiction de LBUG (gèle un thread noyau et nécessite un reboot) à partir des appels système
    - Mise en évidence de motifs dans les appels système

- Storage Chamäleons - Xavier Espinal
  - CASTOR : évolue vers un système *high-throughput*
    - But : *high-throughput* de la DAQ vers les *tapes*
    - Migration de RAID1 vers RAID60 : 350 MB/s/stream (contre 100)
    - Plus grand stockage de bandes au monde : 138 PB, +500M fichiers
  - EOS : principal stockage au CERN
    - Mise à l'échelle ok, performant
    - Désigné pour les besoins des utilisateurs
      - Utilisé en dehors du CERN (Fermilab, Russia-T1, EsNET, NERSC, …)
        - NERSC 1 PB pour ALICE xROOTd (PDSF Site Report)
    - Fédération/synchronisation de diverses instances (ex : CERNbox, slide suivant)
  - CEPH : principalement derrière OpenStack et S3
    - Contribution au développement
  - But : facilité l'accès aux données via des points de montage fuse disponibles sur toutes les plates-formes (du laptop au WN)

Besoin en stockage au CERN

Reliable

Fast Processing
DAQ Feedback loop

DAQ to CC
8GB/s+4xReco ALICE

Hot files

WAN aware
Tier-1/2 replica, multi-site

High throughout to tape
350+MB/s/drive - 12GB/s Pb-Pb

back-up

Filesystem 'feeling'
$HOME, SW-dist, Data

Consistent

∞

Few fast streams
CDR 2x40Gbps

Non-LHC and Local
Less structured, small communities
Unexpected usage Catalogue=Namespace

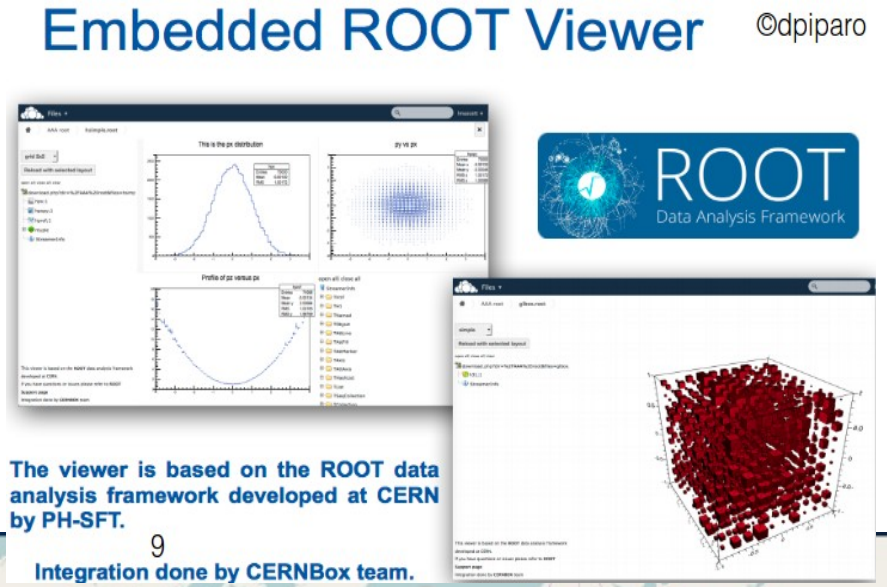
disk and gc?

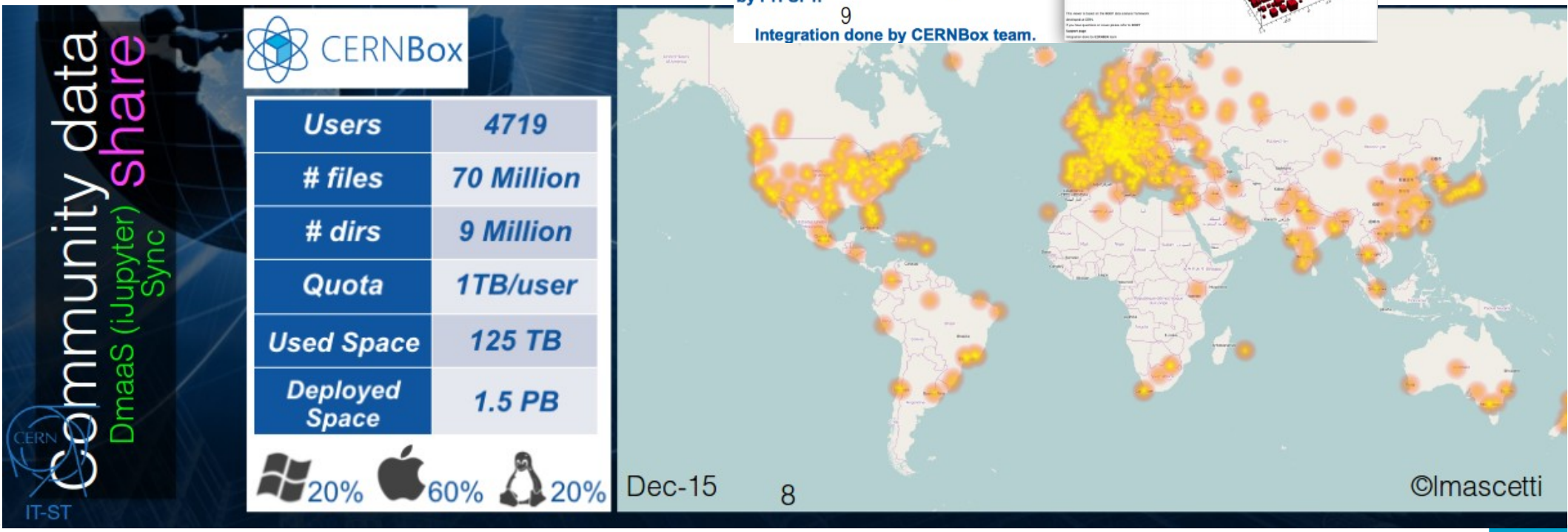
Many slow clients
Repro, reco, analysis constant >20k CMS

Endpoint Mounts
ie. /atlas in the WNs

CERN
IT-ST

- CERNbox : données distribuées, partagées, synchronisées





Panorama des systèmes de soumission

- HTCondor
  - Michel a présenté un CR du workshop à Barcelone
    - ClassAds
    - Multicore « partitionable »
    - Docker (docker universe)
    - Interfaçage avec le cloud via EC2 (grid universe)
    - Un nouveau workshop prévu au printemps 2017
  - DESY@Hambourg
    - HTCondor+ARC-CE (280 nodes/12k slots)
  - CERN
    - HTCondor+HTCondorCE (15k slots), extension prévue clouds commerciaux
  - Liste de diffusion sur HTCondorCE : htcondor-ce@hepix.org
- SLURM
  - Souvent utilisé pour des clusters HPC
  - DESY@Hambourg : 89nodes/4600 cores, réseau infiniband
  - En test à IHEP et au BNL
- Autres : UGE 8.2 DESY @Zeuthen, LSF @KEK

- Conteneurs
  - Docker
  - Orchestration : Docker Swarm/Compose, Google Kubernetes, Shifter (NERSC), NavOps (Univa GE)
- Applications
  - Docker@CERN : comparaison entre Docker Swarm/Compose et Kubernetes
    First Experiences with Container Orchestration in the CERN Cloud

	Docker Swarm / Compose	Kubernetes
Docker API	Yes	No
Expose Port	Yes	Yes
Load Balancing	No	Yes
Failover	Experimental	Yes
Node Scaling	Yes	Yes
Container Scaling	Manual	Auto
Cluster Network	Yes	Yes
Rolling Upgrade	No	Yes

- Using Docker container virtualization in DESY HPC environment - S. Yakubov
  - Docker dans SLURM (DESY @Hambourg)
    - Aucun overhead observé, Docker est prêt pour le HPC !

Examples – HPCG/HPL Benchmarks

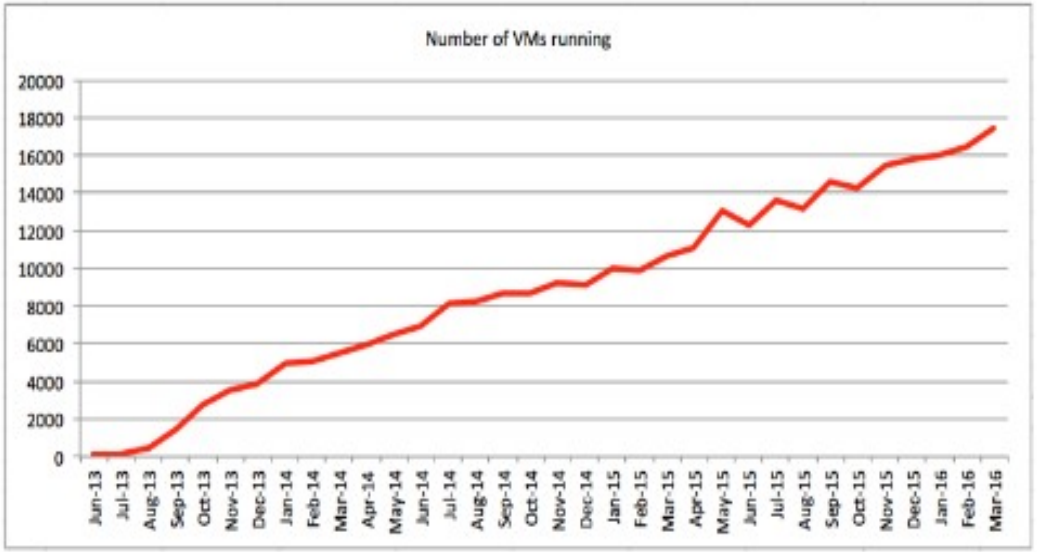
> Maxwell HPC cluster (using Intel tuned binaries)

	Cores	HPL (TFLops)	HPCG (TFLops)
Maxwell	64 (2 nodes)	1.56	0.033
Maxwell+Docker	64 (2 nodes)	1.56	0.033
Maxwell	368 (15 nodes)	9.0	0.192
Maxwell+Docker	368 (15 nodes)	9.0	0.192

  - Docker dans UGE
    - NavOps=Docker+Kubernetes+Project Atomic+UGE
    - En beta (*early access*), *release* de production fin mai

- CPU Benchmarking - Manfred Alef
  - HS06 n'est plus adapté (archi. 32bits)
    - LHCb montre des différences importantes (30-40%)
      - Confirmées par ATLAS
    - Nécessité d'un nouveau benchmark
  - « Fast benchmark » (quelques mins) pour estimer des ressources « opportunistes »
    - Licence d'utilisation gratuite
    - La précision n'est pas une contrainte importante
  - Redémarrage du groupe de travail HEPiX sur le benchmark
    - Jérôme Pansanel et Emmanouil Vamvakopoulos y participent
- Low Power Processors in HEP - M. Michelotto
  - Les ventes des SoC sont boostées par les ventes de tablettes/mobiles
  - Principalement basée sur l'architecture ARM
  - Intel et AMD s'y intéressent (Intel Atom, et AMD Opteron)
  - Mais pas de châssis (haute densité) dédié...

- CERN Cloud Infrastructure Report
  - En chiffres
    - 5800 hyperviseurs
      - 155k cores
      - 18k VMs en moyenne
        - 1 VM crée ou détruite toutes les 10 s
    - CEPH en backend pour Glance et Cinder
      - 2700 images
      - 220 volumes
    - Évolution à venir : +57k cores au printemps 2016 !



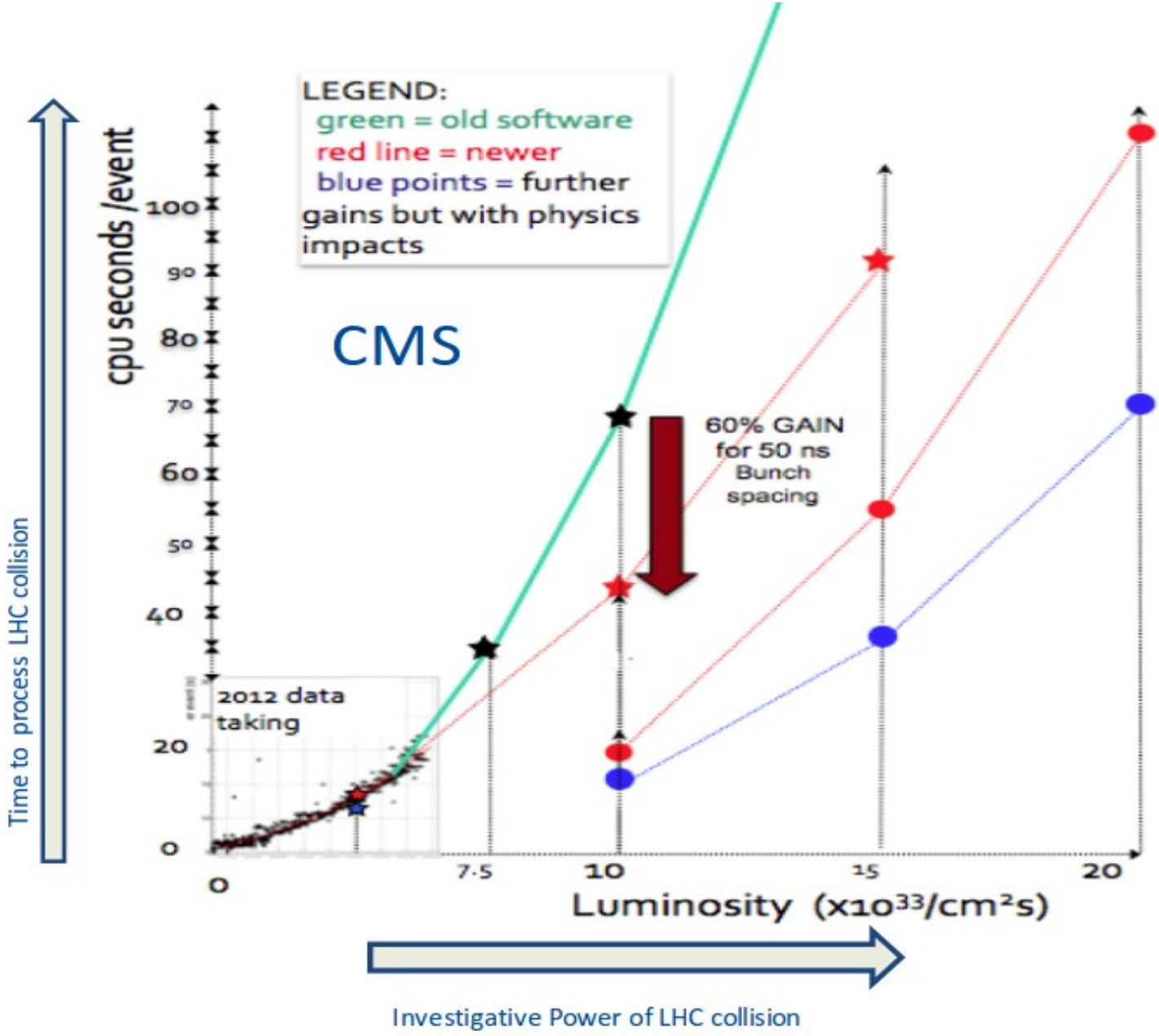
- CERN Cloud Infrastructure Report

  - *Overhead VM Vs bare metal :* détails (talk d'Arne@BNL)
    - NUMA paramètre critique pour l'optimisation ~3 % overhead restant (HS06)
    - EPT (extended page table for Intel), activé (meilleures performances) mais nécessité de pages « énormes » de 2MB
    - Infra mise-à-jour en ~2 mois
  - Nouveau HW avec SSD
    - Résolution de problème d'I/O
      - SSD caching
        - SSD caching ZFS ZIL/I2arc devant Cinder
        - b-cache Vs dm-cache, mais b-cache montre de meilleures performances

- CERN Cloud Infrastructure Report

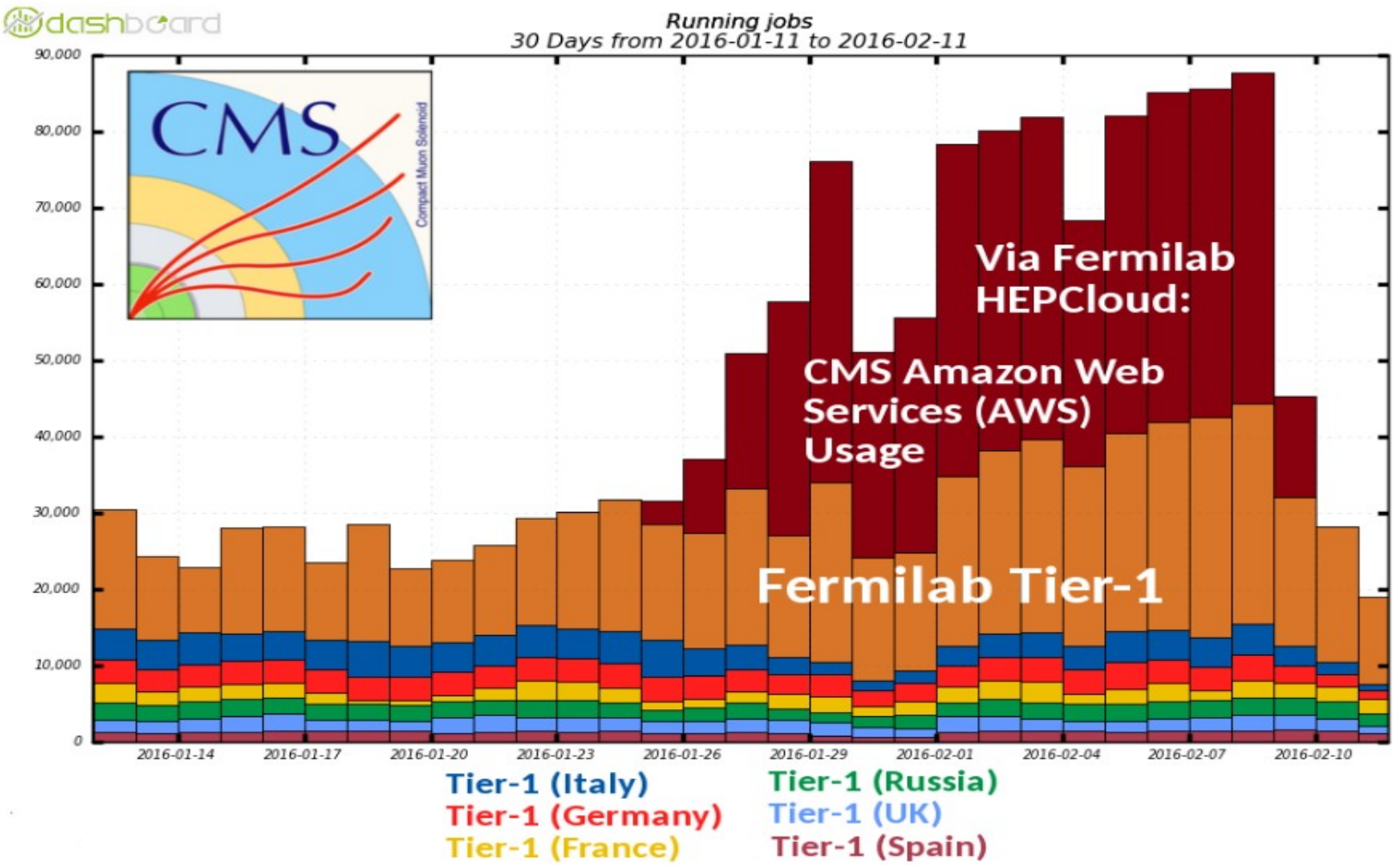
  - Migration à Neutron
    - Nova-network bientôt obsolète...
    - Fonctionnalités intéressantes
  - Keystone
    - Accès EduGAIN via Horizon
  - Conteneurs via Magnum
    - Container Orchestration in the CERN Cloud
  - Futur...
    - Bare metal provisioning par Ironic
    - Remplacement de Hyper-V par qemu/kvm
    - Pour suivre les travaux en cours : le blog d'OpenStack du CERN

The Fermilab HEPCloud Facility - Anthony Tiradani



- CPU insuffisant pour satisfaire les besoins à venir (voir aussi Helge's talk Helix Nebula)
- Extension possible dans un cloud commercial
- « Academic grant » pour tester l'intégration de ressources AWS dans le cloud du FNAL (atteint 60k cores, ~1/4 des ressources utilisées par CMS lors du test)

- The Fermilab HEPCloud Facility - Anthony Tiradani



Le coût d'utilisation d'un cloud commercial (ici AWS) reste 2 à 3 fois supérieur à celui de l'utilisation de ressources locales, mais diminue très rapidement ! (un facteur 10 il y a quelques années...)

- Deploying services with Mesos at a WLCG Tier 1 (RAL)

  - Mise-à-jour par rapport à la présentation de Andrew@BNL

  - Déploiement de nouveaux services simplifié par la gestion de la configuration, mais encore (trop) des tâches manuelles
  - Applications sont « conteneurisées »
  - Marathon pourrait remplacer (partiellement) le système de gestion de configuration
  - Monitoring aisé
  - Problème pour l'insertion des « secrets » au sein des conteneurs

- Security Update – Liviu Vâlsan

  - *Malvertising* à grande échelle : une des principales sources d'infection actuelle
    - Utiliser un ad-blocker pour bloquer la plupart des contenus à risque
    - Ne pas utiliser de plugins faibles (Flash, Silverlight, Java), ou à minima les utiliser avec un "clic pour activer"
  - Phishing encore très actif :
    - Taux de « click » malencontreux 22 % lors d'une campagne de test au CERN !
  - Ransomware du moment : Locky et Petya
    - Petya s'attaque au MBR (diffusion via dropbox, solution de déchiffrement existe).
    - Mac OSX commence à être visé.
  - Mobiles : Android n'est pas en reste, et iOS n'est pas invulnérable
    - 20 K apps véhiculant des malwares identifiées sur GooglePlay, dont des clones de Facebook, Candy Crush, WhatsApp...

- Security Update – Liviu Vâlsan

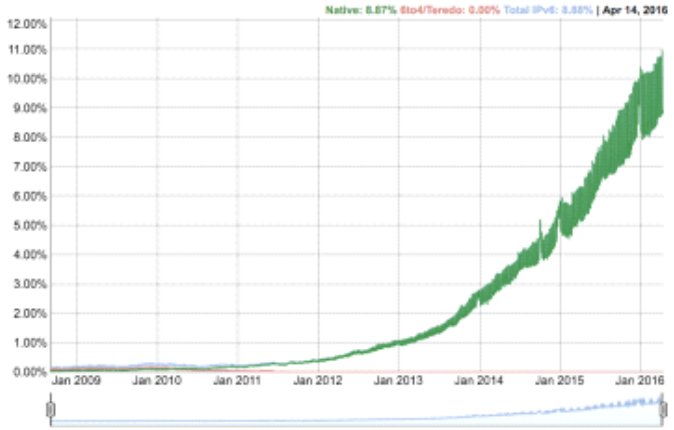
  - Accent particulier sur la sécurisation de l'infrastructure Windows :
    - en commençant par les comptes administrateurs locaux et de domaines, désactiver le support des protocoles LM et NTLv1…
  - SOC (Security Operations Center) :
    - Mise en place au CERN d'une plateforme capable de collecter, corréler, et analyser tous les événements et logs utiles à la détection d'anomalies / menaces
    - Proposition : "WLCG SOC Working Group" mise en commun idées, designs, outils ; mise en réseau équipes
    - *Traceability & Isolation WG* : nouveaux mécanismes de virtualisation et « conteneurisation » -> nécessaire collaboration entre VOs et Sites

- # Recent work of the HEPiX IPv6 WG – David Kelsey
  - Le groupe http://hepic-ipv6.web.cern.ch existe depuis 2011
  - Aujourd'hui, ~10 % du trafic internet est IPv6
  - En Europe, la Belgique fait figure d'exception avec 40 % de taux d'adoption Ipv6
  - **Enjeu pour WLCG :** exploiter du CPU et des ressources Cloud opportunistes IPv6 only, dans "un futur proche" (au Canada, en Asie…)

HEPiX

Google IPv6 stats

https://www.google.com/intl/en/ipv6/statistics.html



19-Apr-2016 HEPiX IPv6 5

Planning WLCG :

- **2015 :** Monitoring réseau (perfSONAR) en dual-stack Ipv4/IPv6
- **2016 :** Passer le stockage en double-stack
- Disque DCache en double-stack pour CMS @PIC Tier-1
- **2017 :** IPv6 only CPU

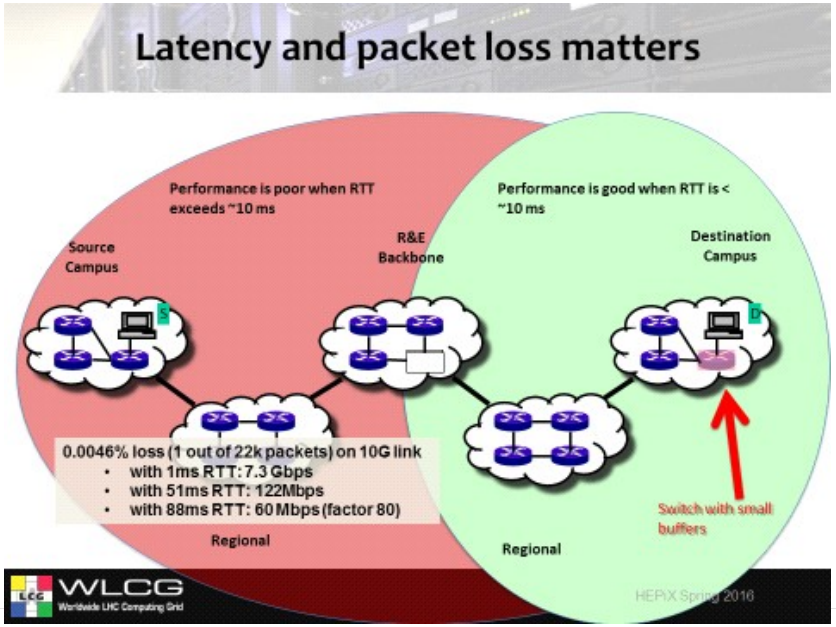
Aujourd'hui 5 % des services WLCG sont en double-stack ; les transferts de données Ipv6 sont une réalité.

Travail en cours côté sécurité : élaboration d'un guide de bonnes pratiques

Réunion le 7 juin au CERN : http://indico.cern.ch/event/394830/

- # PerfSONAR Status in WLCG/OSG – Shaw McKee

  - Avancées du groupe de travail WLCG « Network and Transfer Metrics »
  - PerfSONAR : outil de supervision réseau multi-domaine
    - Près de 250 instances déployées dans le cadre WLCG/OSG dont une vingtaine en France
    - Topologie, Mesures & tests de bout en bout (traceroute/tracepath, latence et bande passante iperf3)
  - Infrastructure de collecte des métriques pour analyse et prédictions
  - **Objectifs :** identifier et comprendre les (vrais) problèmes réseau, introduire de l'intelligence réseau au niveau des applications pour optimiser les workflows et les transferts de données



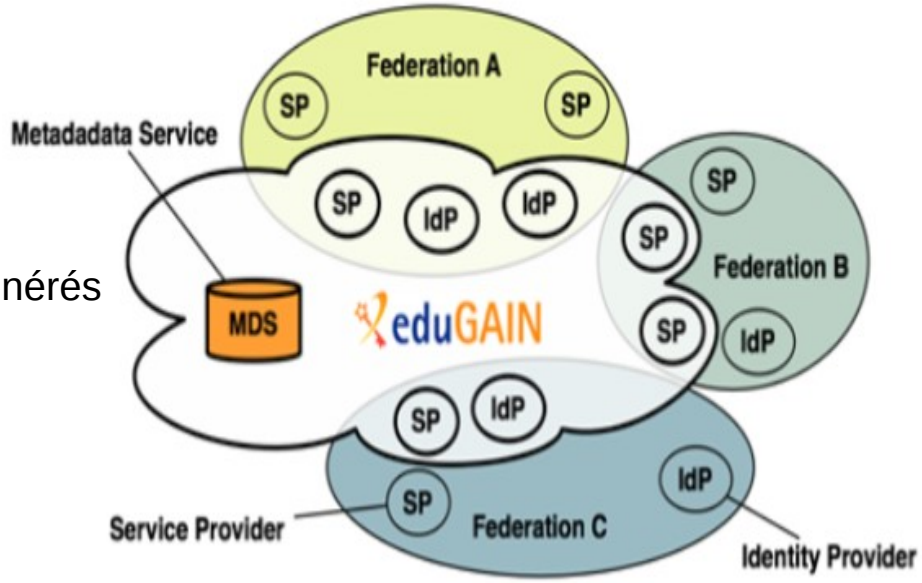
- **Performances liens 10G :**
  - Impact des pertes de paquets et de la latence très important au delà d'un RTT de 50 ms

    RTT=50 ms 0,0046 % loss (1 paquet / 22K)

    => 122 Mb/s max
  - Intérêt des mesures de latence et de pertes de paquets moins couteuses que les mesures de bande passante
- **En cours** : Plateforme analytique ATLAS (Flume + Elastic Search) intégrant les métriques perfSONAR
- **Au-delà** : Premiers tests US de solution SDN (Software Defined Network) via des contrôleurs Openflow

Identity Federation for HEP – David Kelsey

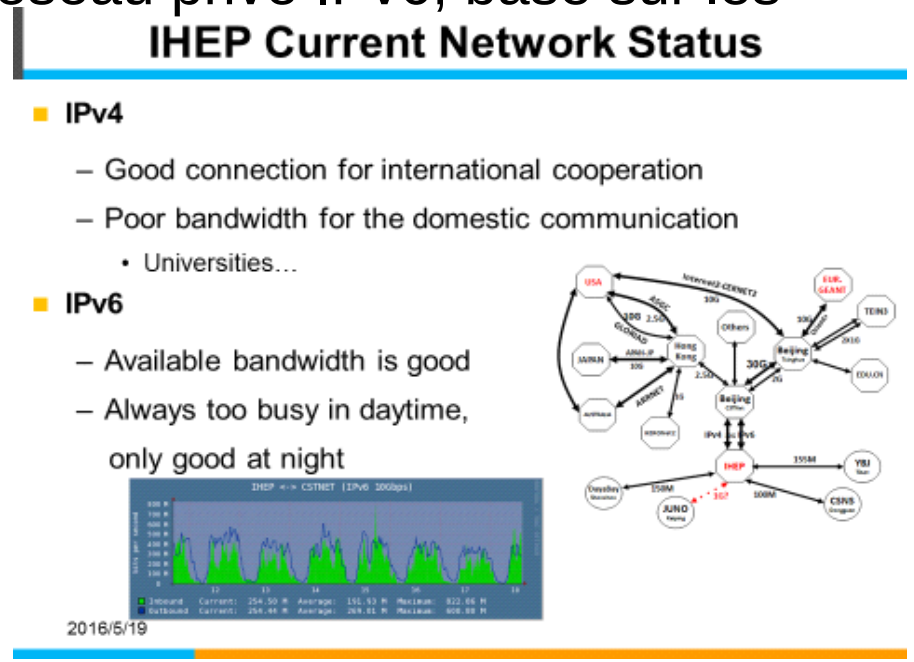
- Fédération d'identité World-wide : où en est-on ?
- Apport d'eduGAIN en tant que service interfedération/ interconnexion de fédération
  - Relations de confiance basée sur des politiques communes et un service de partage de métadonnées (MDS) entre fédérations
- Que faire en cas de compromission d'un compte d'IdP ?
  - REFEDS (refeds.org) tente de promouvoir un framework de confiance pour faciliter la gestion concertée en cas d'incident Security Trust Framework for Federated Identity (SIRTFI)
- AARC (Authentication & Autorisation for Research & Collaboration) projet H2020 https://aarc-project.eu/
- Dans le monde HEP :
  - Permettre la fédération des identités
  - Bâtir sur EduGain
  - Service Pilote WLCG : Certificats X509 générés de façon dynamique et transparente pour l'utilisateur via IOTA (Identifier-Only Trust Assurance) CA



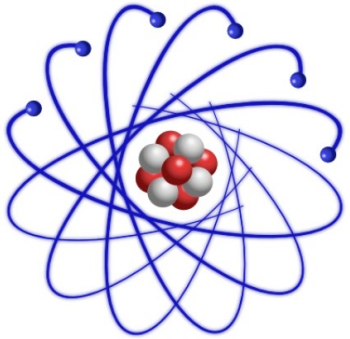
Crédit : GEANT Alessandra Scicchitano

- # A private network based on SDN for HEP - Zhihui Sun

  - Software Defined Network : Technologie de réseau programmable, un sujet récurrent qui dépasse le stade de la R&D ; des projets aux US
  - Use case de nos collègues chinois de IHEP dans un contexte un peu particulier (besoin de connectivité entre sites)
    - Faible bande passante en IPv4 excepté à l'international
    - Bonne connectivité domestique IPv6 mais en concurrence avec l'internet généraliste pendant la journée
  - Mise en place préliminaire d'un réseau privé IPv6, basé sur les principes du SDN.
    - Pour l'instant, 3 sites connectés, routes statiques



- Scientific Linux Update - Rennie Scott



  - SL5
    - Fin du support en mars 2017
  - Quelques nouveautés pour SL6 (RHEL 6.8 en beta)
  - Release de SL7.2
    - contextualisation

      http://ftp.scientificlinux.org/linux/scientific/7x/contexts/
    - Dépôt communautaire (all EL7 « flavour ») :
      http://ftp.scientificlinux.org/linux/scientific/7/repos/x86_64/
  - ...

- NERSC collection architecture - Thomas Davis
  - Nouveau centre de données, nouvelle infra de collection : très grand nombre de capteurs, voir PDSF Site Report pour les détails
  - Objectifs : analyse du flux en temps réel (adopter des standards : metrics20.org)
  - Infra basée sur Docker, Rancher (orchestration) et Consul (service discovery)
- Monitoring at scale: a needle in the haystack - F. Wernli
  - Architecture de grande envergure basée sur ElasticSearch
    - Entrée : collectd+Riemann et Syslog-ng
    - Dashboard : Kibana, Grafana, reimann-dash et accès dispo par REST
    - Objectifs : analyse temps réel et analyse post-mortem
      - Nécessité : définir des métadonnées, labelliser/standardiser les événements
- Création d'une liste de discussion autour du monitoring
  - hepix-sig-monitoring@hepix.org
  - Plate-forme d'analyse de logs en temps réel et prévention d'incidents

- GreenIT Cube - Jan Trautmann

- GreenIT Cube - Jan Trautmann

  - 6 étages, 128 racks par étage
    - Coût de l'infra actuelle 11,5 M€ (coût total : 16 M€)
  - Construction rapide : 12/2014, premier cluster en mars 2016
  - Côté énergie
    - 8 lignes complètement redondantes
    - 4x4 lignes connectées sur 1 ligne
  - Refroidissement
    - Bâtiment séparé
    - 2 circuits (1300m^3/h en circuit fermé)
    - Chaleur utilisée pour chauffer les bureaux et la cantine
  - PUE mesuré pendant un test < 1,1

Prochains HEPiX

- Fall 2016 : Berkeley
  - 17-21 octobre, dos-à-dos avec WLCG/CHEP qui aura lieu à San Francisco (8 au 14 octobre)
- Spring 2017 : Wigner, Budapest (à confirmer)
  - 24 au 28 octobre
- Fall 2017 : KEK, Tsukuba
  - 16 au 20 octobre

A New Focus – We Host Fall 2016 HEPiX