Evolution & Perspectives du Mésocentre de l'Université Savoie Mont Blanc

Frédérique Chollet[1], Cécile Barbier[1], Sylvain Garrigues[1], Mathieu Gauthier-Lafaye[2], Muriel Gougerot[1], Stéphane Jézéquel[1], Nadine Neyroud[1], Philippe Séraphin[1]

[1] LAPP, Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS/IN2P3, Annecy

[2] LAPTh, CNRS, Université Savoie Mont Blanc, Annecy

Contenu

- Présentation du mésocentre MUST
  - Caractéristiques
  - Communautés d'utilisateurs et Usages
  - Ressources, Infrastructure, Equipe Technique
- Perspectives après dix ans d'existence
- Conclusion

MUST : Mésocentre de Calcul et de Stockage Université Savoie Mont Blanc

- 10 ans au service d'une dizaine de laboratoires de l'Université Savoie Mont Blanc (USMB)
- situé sur le campus d'Annecy-le-Vieux,
- opéré par le CNRS (équipe mutualisée du LAPP et du LAPTh)
- ouvert sur la grille européenne EGI / WLCG

2005 : Réponse à l'appel d'offre ministère pour la création d'un mésocentre national

> Convergence des besoins et des moyens de l'Université Savoie Mont Blanc et du LAPP (Laboratoire d'Annecy de Physique des Particules)

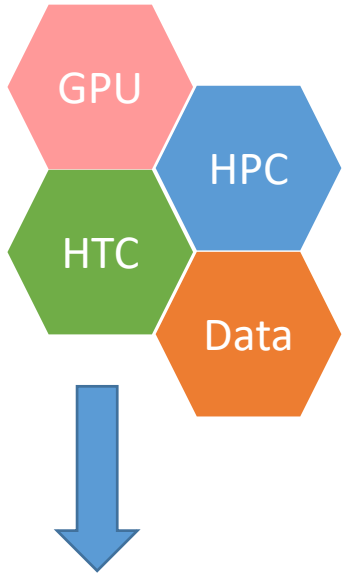
2007 : Inauguration et officiellement Tier 2 pour le traitement des données du LHC (ATLAS, LHCb)

2013 : Nouvelle salle informatique (170 m2) dans la Maison de la mécatronique

2017 : Labellisation « Plateforme numérique IN2P3 »

Caractéristiques du mésocentre MUST

- Une seule et même infrastructure pour des communautés & usages différents :
  - Tier 2 WLCG (Worldwide LHC Computing Grid) / LCG-France
  - Site EGI (European Grid Infrastructure) / France Grilles
  - Ferme de calcul batch pour les laboratoires Université Savoie Mont Blanc
- Une seule et même infrastructure pour des besoins multiples :
  - Intégration de ressources diverses (HTC, HPC et GPU) au sein d'un même cluster
  - Prise en compte des problématiques de stockage et de gestion des données
  - Accès local (batch local) ou distant (grille)
  - Logiciels commerciaux (MATLAB, MATHEMATICA, ABAQUS, MAPLE) avec gestion propre à chaque communauté
- Support au Calcul Scientifique
- Gouvernance (comité de pilotage)
- Ouverture aux partenaires économiques : étude d'opportunité avec la Fondation USMB (Pfeiffer Vacuum)



Optimisation de l'utilisation des ressources

Communautés d'Utilisateurs & Usages *(accès grille)*

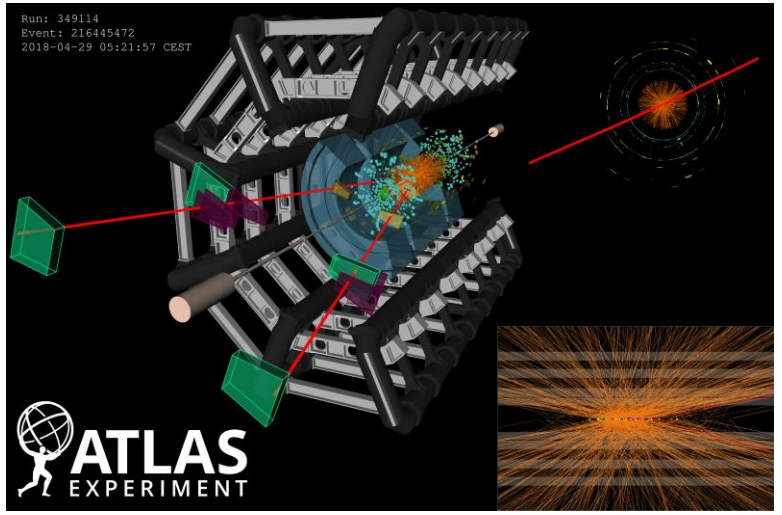
LAPP : Physique des particules et Astroparticules

Collaboration WLCG : Traitement des données LHC ATLAS, LHCb

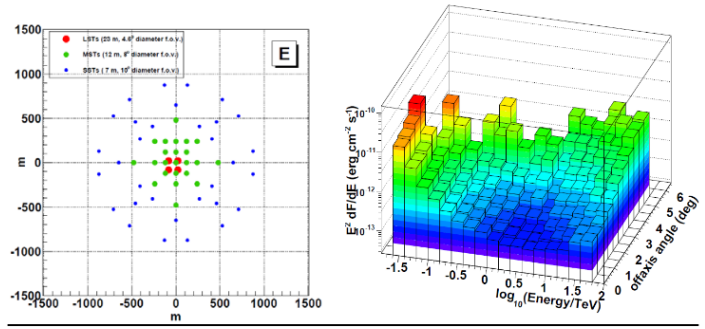
- Simulation et reconstruction des évènements (organisées)
  - Gestion centralisée de workflows complexes et de lots de données multiples : brutes, simulées, réduites
  - CPU intensif et/ou I/O intensif
- Analyse des données (organisée ou individuelle)

Communautés EGI :

- Astronomie gamma : HESS, CTA (Cherenkov Telescope Array)
- Astronomie, Cosmologie : LSST (Large Survey Synoptique Telescope)
- Futurs collisionneurs linéaires : Collaboration CLIC-ILC
- Interaction particules - matière : GEANT4 (Simulation Monte Carlo)



ATLAS Experiment @ 2018 CERN



Crédit : CTACG hal-00653017

Communautés d'Utilisateurs & Usages *(batch local)*

LAPTh : Physique théorique

- Calcul parallèle
- Cosmologie (simulation de modèles mathématiques)

LEPMI : Electrochimie et Physico-Chimie des matériaux

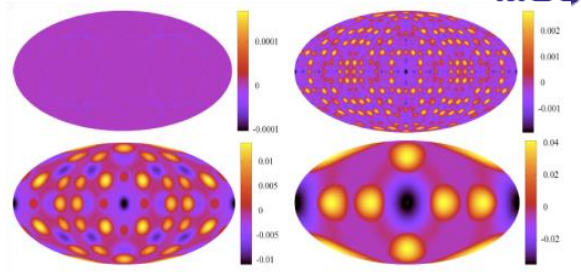
- Calcul parallèle
- Modélisation moléculaire

LISTIC : Traitement de l'Information et de la connaissance

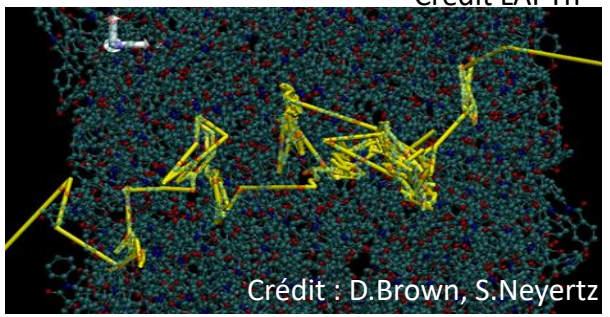
- *Deep Learning* sur GPU
- Traitement et Analyse d'images de télédétection

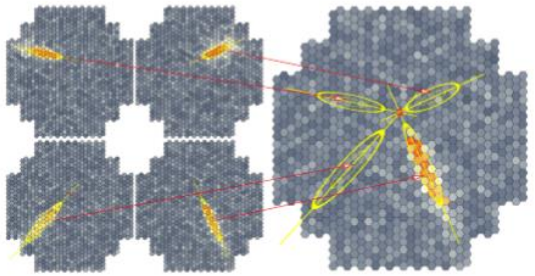
Collaboration LAPP-LISTIC :

- Application du *Deep Learning* à la reconstruction stéréoscopique des images CTA (Cherenkov Telescope Array)



Crédit LAPTh





Crédit K.Bernlohr (HESS figure)

Communautés d'Utilisateurs & Usages

Activité de calcul 2017

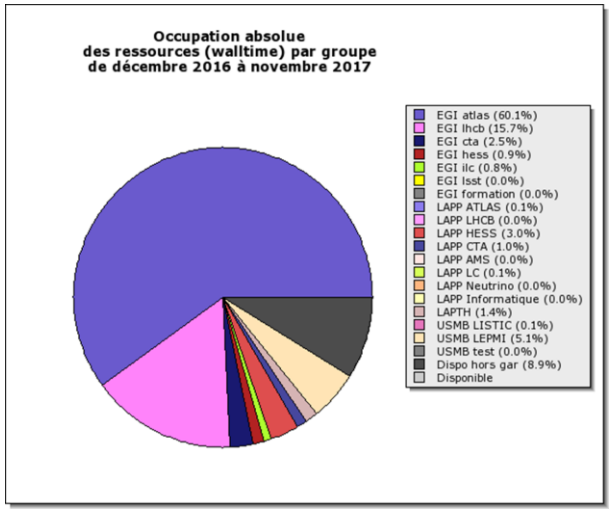
Grille 80 %

- Tier 2 WLCG (Atlas, LHCb) : **75 %** «***pledges***»
- Communautés EGI : **5 %**

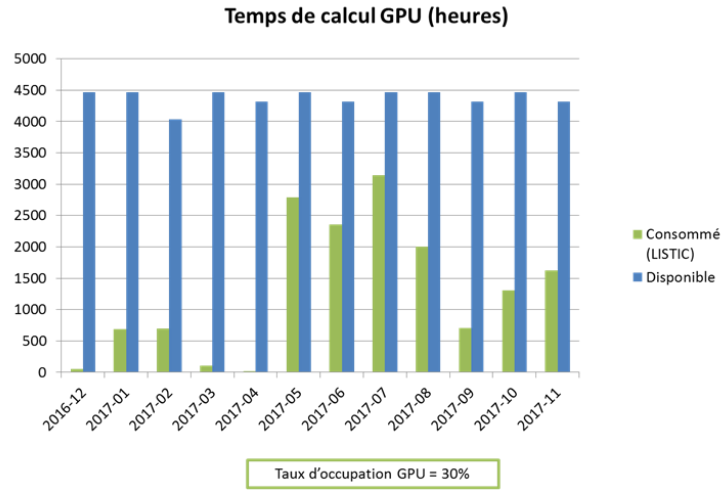
Batch local 11 %

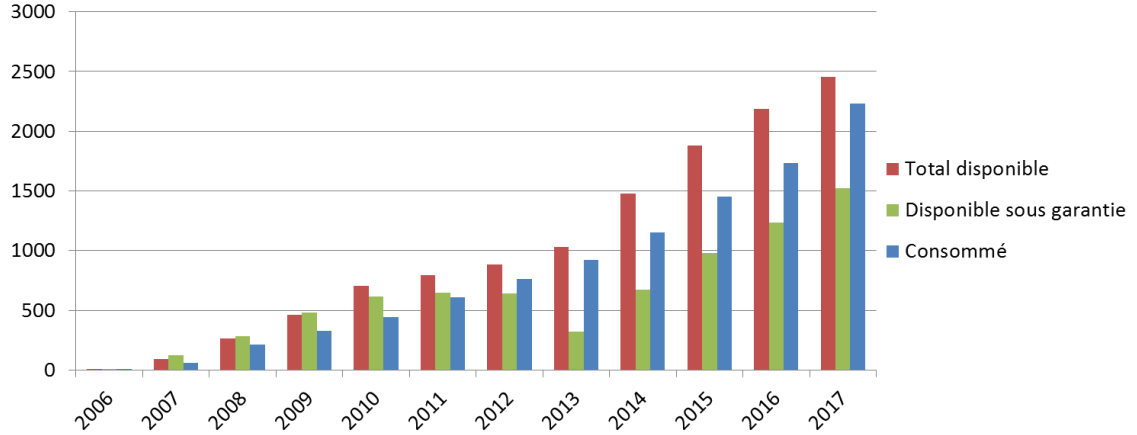
- Université Savoie Mont Blanc

Taux d'occupation CPU (2017) : **91%**



ATLAS : 540000 jours.CPU
8 % du temps calcul fourni à ATLAS par la France





Evolution des ressources

CPU, Stockage : aujourd'hui,

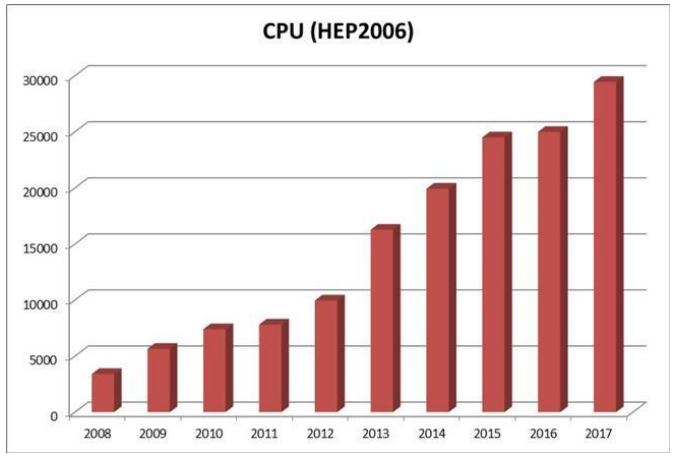
- 34500 HEP-SPEC06 (HS06) ~27 Tflops, 3300 jobs simultanés
  - 10,5 HS06 2Go (RAM) 20 Go min. par job slot
- 2,2 Po stockage disque

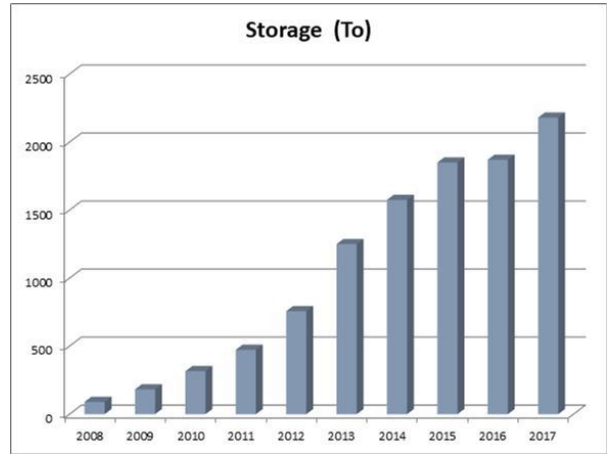
Croissance continue et régulière :

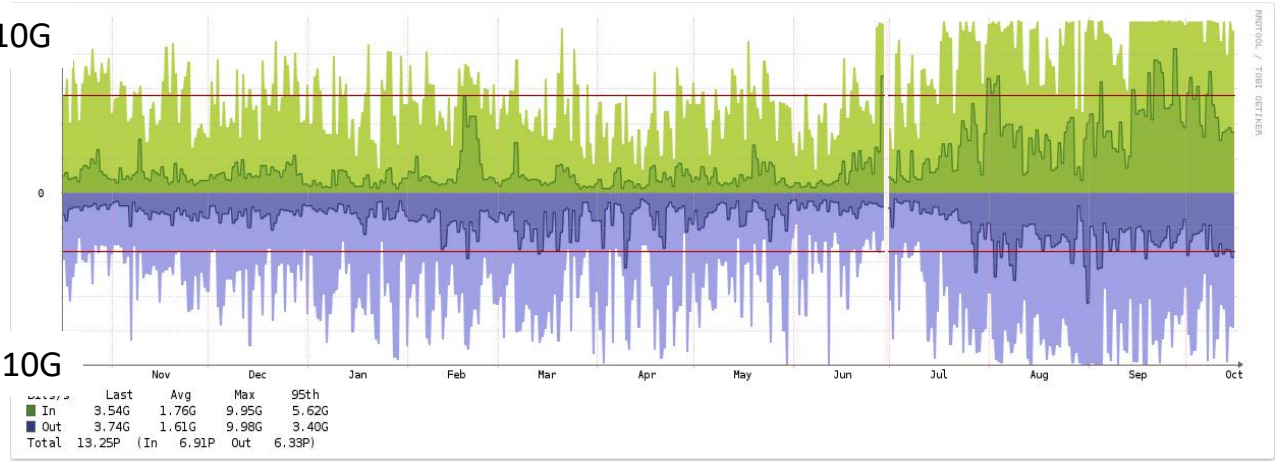
- suit l'évolution des besoins WLCG
- x 2 entre 2013-2017 à budget constant
- Engagement Tier 2 (pledges) : env. 2 % de la totalité des besoins Tier 2 ATLAS

Réseau :

- Depuis 2013 : 10 Gb/s vers le CC-IN2P3 (LHCONE) sur une fibre AMPLIVIA
- Ressource à part entière dont l'évolution devient stratégique

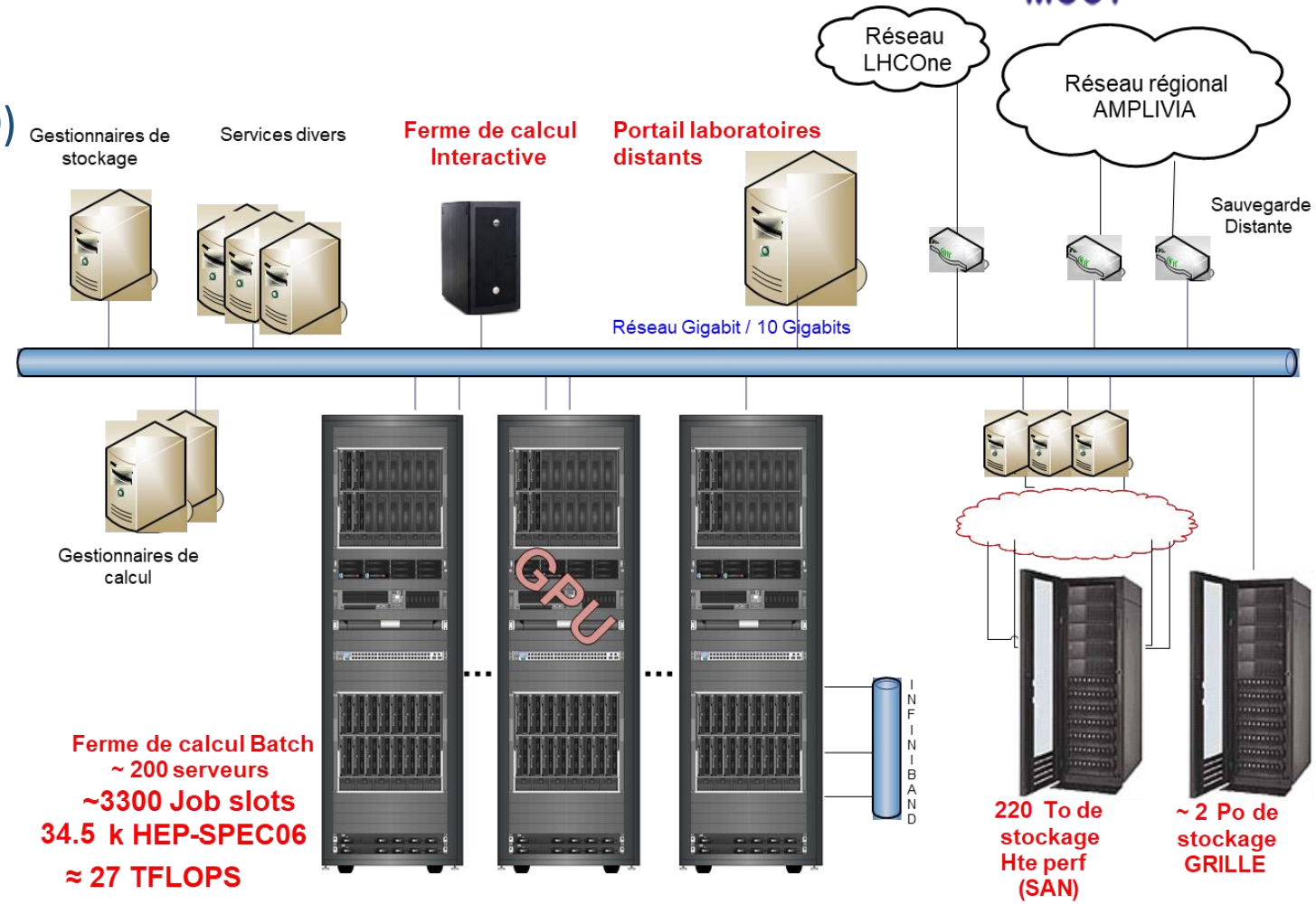






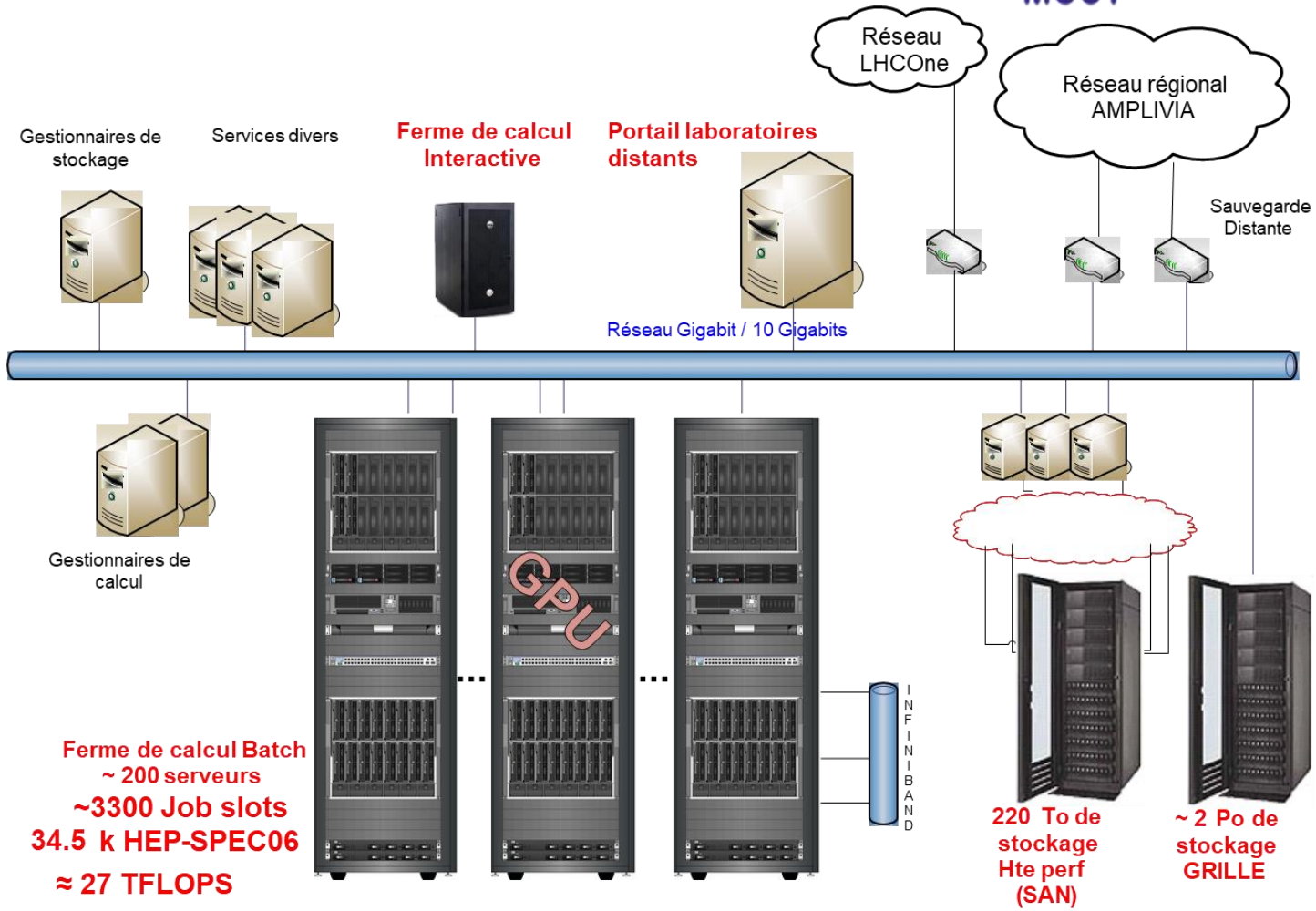
Infrastructure – Calcul scientifique

- GPU :
  - 4 serveurs (cartes Tesla K80 et V100)
- HTC : Solutions de type *blade center*
  - HP, Fujitsu
  - Benchmarks systématiques (HS06)
  - Modèle de coût LCG-France (€/HS06)
  - Achats 2018 : étude de l'offre DELL dans le cadre du marché MATINFO4
- HPC :
  - Config 16 serveurs (192 processeurs) + switch Infiniband
  - Fin de garantie en 2017
  - Evolution vers le Multi-core / Multithreading
  - Ordonnancement dynamique de jobs 8 cœurs de calcul



Infrastructure – Données scientifiques

- Stockage Haute Performance : ~200 To
  - Disponible sur tous les nœuds de calcul et machines interactives
  - SAN MD3820f de DELL avec 4 têtes
  - Stockage partagé GPFS : Evolution en cours suite au changement de politique tarifaire IBM
- Stockage capacitif : 2,2 Po sur la grille
  - Technologie DAS ; Solution DPM (Disk Pool Manager) développée par le CERN ; protocoles dédiés à la grille
  - 25 serveurs de disques
  - Double pile IPv4/IPv6
- Service iRODS : Stockage & Gestion de collections de données
  - Besoins au sein de l'USMB
  - Participation au projet de fédération pico2 (projet européen EOSCpilot)



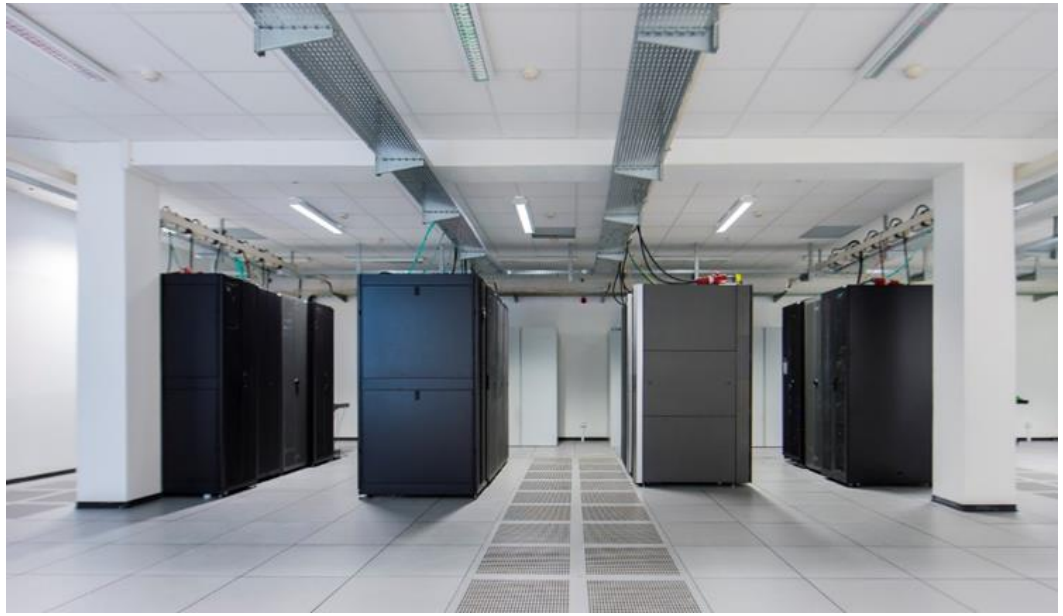
Infrastructure – Salles informatiques, Réseau, Services

- Deux salles disponibles distantes de 150 mètres
  - 60 m2 : sécurité incendie, onduleurs (80 kVa), climatisation 2*40 kW

  Héberge essentiellement les services du LAPP
  - 170 m2 : Sécurité incendie, onduleur (500 kVa), climatisation 240kW (extensible) , taux de remplissage de 16/50 racks, Puissance consomée 120kW, PUE : 1,6

  Héberge le mésocentre MUST + des services du LAPP et un rack de l'USMB.

- Un cœur de réseau « bicéphale » sur deux salles qui s'appuie sur la technologie VSS
- VLANs dédiés au mésocentre MUST qui convergent dans un cœur de réseau (Extreme Networks X670) IPv4 / IPv6
- Déploiement piloté par Quattor (de l'ordre de 300 profils de machines)
- Ordonnancement basé sur Torque/Maui
- Plateforme de virtualisation : hyperviseurs sous Proxmox



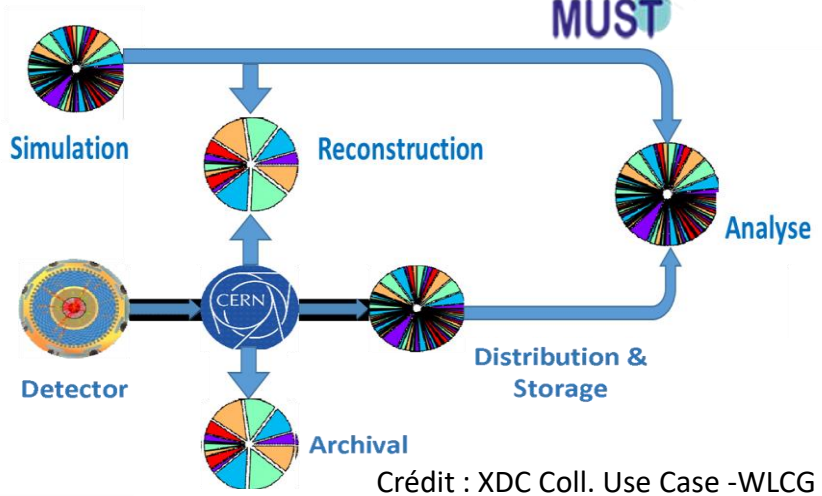
Equipe technique

- ~ 2.8 FTE répartis sur 6 personnes (LAPP, LAPTh)
- Une des clés du bon fonctionnement du mésocentre :
  - Profils complémentaires, compétences multiples

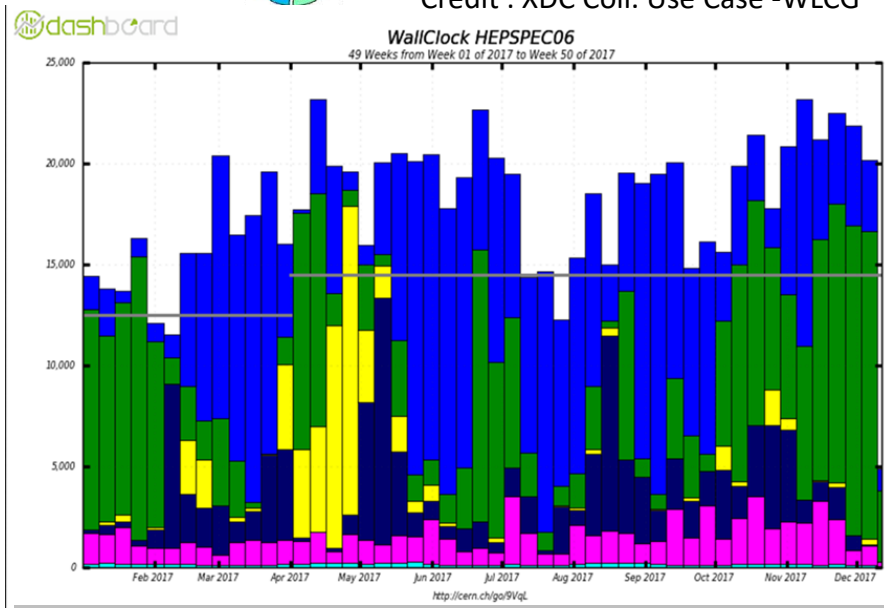
    ASR, expertise réseau, modèles de calcul des expériences, calcul scientifique, qualité, sécurité
  - Proximité avec les utilisateurs et les communautés utilisatrices
  - Equipe organisée et proactive
    - Fonctionnement 24/24 7/7 sans astreinte (rôle « Lanceur d'alerte » pendant les périodes de fermeture du laboratoire)
    - Projets de fond : déploiement IPv6, remplacement GPFS, support des conteneurs…
    - Surveillance assidue des services et des activités des expériences
    - Respect des engagements pris auprès de WLCG / LCG-France, EGI / France-Grilles
    - Proximité avec les experts Computing CTA et ATLAS
- Evolution notoire du métier ASR et des compétences associées
- Equipe à même de s'impliquer dans des activités R&D

Après dix ans d'existence

- Traitement des données LHC : composante la plus importante en terme quantitatif
- Les besoins et les évolutions venant de WLCG sont toujours évalués dans un contexte plus large
- Computing WLCG (vue simplifiée) :
  - Plus de 50 % du CPU pour la simulation (CPU intensif)
    impliquant sites Tier 2 et ressources opportunistes (cloud, HPC, BOINC projets @HOME)
  - L'essentiel du CPU restant consacré à la reconstruction (CPU et I/O intensif)
  - Le reste est utilisé pour l'analyse
- En 2016, MUST devient, avec le CC-IN2P3, l'un des 2 sites français de type « Nucleus » :
  - Implication dans quasiment tous les workflows de traitement dont la reconstruction
  - Stockage de données primaires (1 seul replica)
- Spécificité orientée Données
  - transverse à 2 communautés utilisatrices de MUST (WLCG, CTA)



Crédit : XDC Coll. Use Case -WLCG



Perspectives - Contexte & Objectifs

Contexte WLCG : *cf. présentation Eric Fede (CC-IN2P3)*

- Démarrage du LHC Haute Luminosité (HL-LHC) en 2026
- Problématique : des besoins croissants notamment pour le stockage, au-delà de ce qui est envisageable
  - à budget constant
  - avec une évolution technologique (+15 à +20 % /an)
- Stratégie WLCG portant sur tous les aspects du Computing (avril 2018)
  - Issue de la feuille de route communautaire de la HEP Software Foundation (HSF)
  - Mise en place du projet de R&D DOMA
- Réflexion globale coûts / performances via un groupe de travail conjoint HSF - WLCG

Réflexion initiée localement avec un double objectif :

- Initier des actions de R&D selon un point de vue adapté au contexte de MUST
- Etre partie prenante d'initiatives coordonnées : niveaux national, européen,international via WLCG, ATLAS, CTA, projet européen XDC (eXtreme DataCloud)

Perspectives - Orientations

Maintenir MUST en tant que Tier 2 « Nucleus » ATLAS pour les 5 à 10 ans à venir: est-ce possible ?

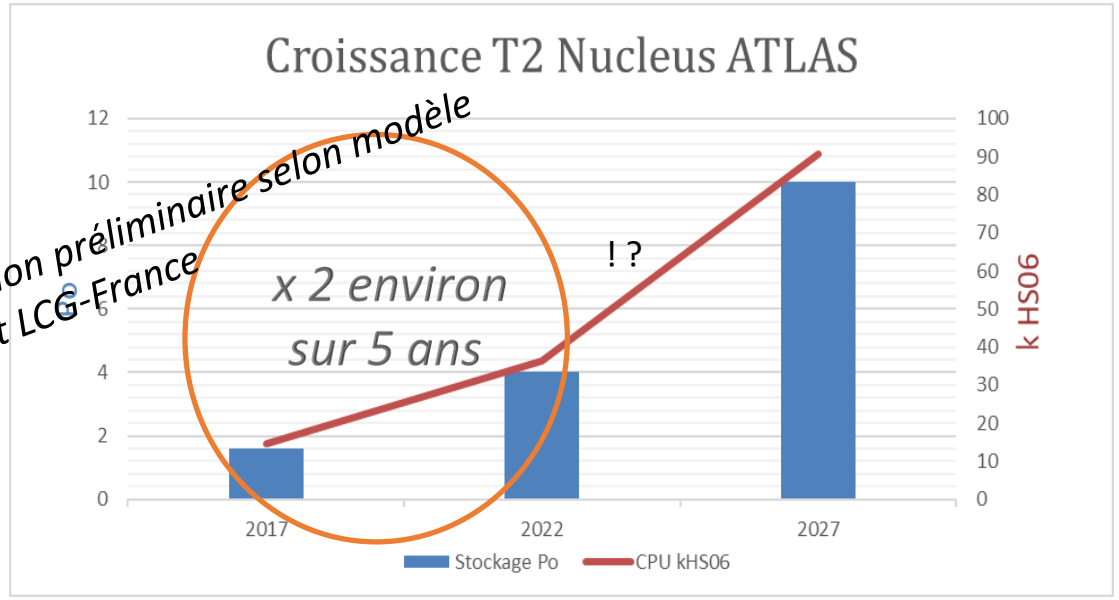
Quelques idées directrices :

- Développer la spécificité Stockage et Données de MUST :
  - Poursuivre la croissance du stockage à budget constant ou à coût maîtrisé
  - Développer les compétences « Gestion de données » acquises et la proximité avec l'expérience ATLAS
- Chercher à optimiser drastiquement l'utilisation du stockage et réduire les coûts opérationnels
- Envisager de nouvelles infrastructures de stockage : fédérée, multi-sites… *Data-lake* compatible
- Tirer parti de la proximité régionale : Annecy (MUST), Grenoble (LPSC), Lyon (CC-IN2P3)
- Travailler dans une optique non exclusivement LHC qui permette d'envisager un second projet en Astroparticules (CTA)

Perspectives - Estimation croissance Tier 2 «Nucleus» ATLAS

A l'horizon des 5 prochaines années, la projection de croissance Tier 2 Nucleus est comparable à celle de MUST entre 2013 et 2017 (même ordre de grandeur)

Les 5 prochaines années doivent être mises à profit pour la R&D et préparer le saut quantitatif à l'horizon de 10 ans



Estimations ATLAS (2017)

Nucleus	Now	5 year (2022)	10 year (2027)
Storage Capacity (PB)	2	5	12.5
Total CPU (kHS06)	40	100	250
LAN (Gb/s)	40	200	1000
WAN (Gb/s)	20	60	200

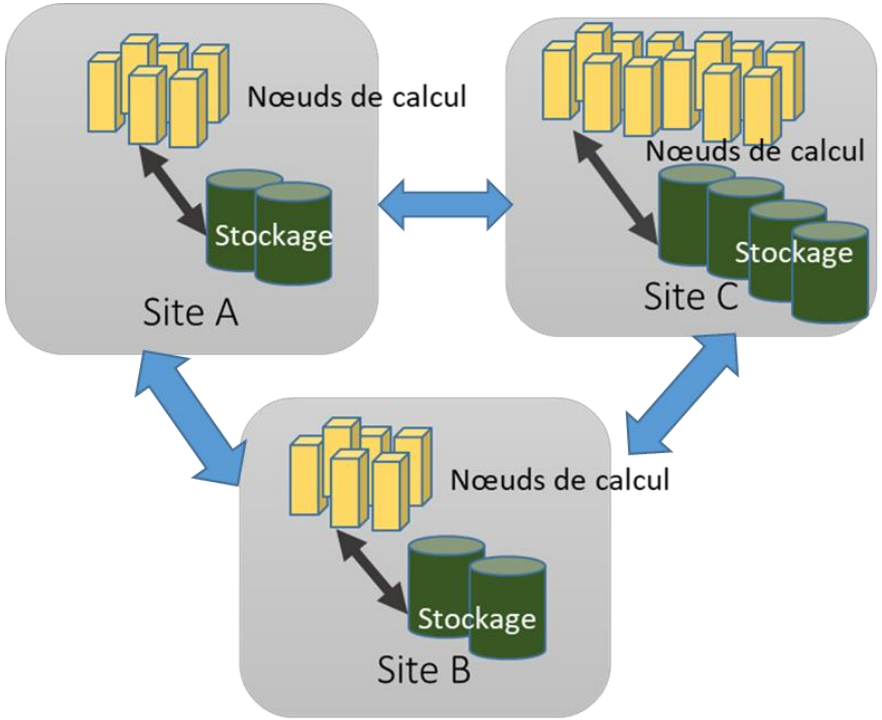
Source : ATLAS Network Requirements 20/06/2017
Alastair Dewhurst, Roger Jones, Shawn McKee on behalf of ATLAS

ATLAS T2 Nucleus à l'échelle de MUST	2017	2022	2028
Storage capacity (PB)	1.75	4	10
CPU (kHS06)	15	36	90
LAN (Gb/s)	10/40	10/40/100	10/40/100
WAN (Gb/s)	10	6 x 10	2 x 100

Perspectives - Etudes envisagées

Problématique : Gestion de données, Services de cache, Fédération des sources de données

Optimiser l'utilisation du stockage disque ; Réduire la duplication des données de l'expérience ATLAS



– Actuellement **10 à 15 %** du stockage persistant de chaque site héberge des **copies temporaires (durée de vie ~ 15 jours)**

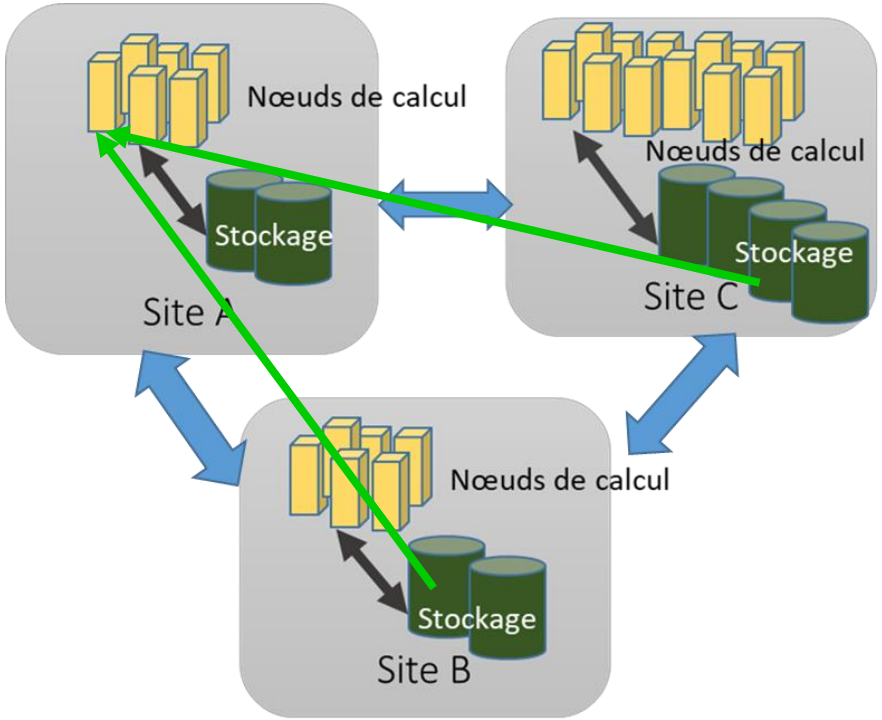
– Réplicas supplémentaires gérés par Rucio, l'orchestrateur de « data flows » configurable d'ATLAS (évalué dans le cadre XDC)

⟺ Duplication par transfert asynchrone (FTS)

⟷ Accès local *(Analysis Download Direct IO, Production Download, Production Upload)*

Perspectives - Etudes envisagées

- Activation des accès distants en lecture à une échelle régionale
- Les nœuds de calcul de chaque site accèdent à distance aux données stockées sur disque dans les autres sites
- Campagne d'évaluation en vraie grandeur (en cours) sur l'infrastructure de production ATLAS



– Accès distants sans cache entre 3 sites (faible latence) : Annecy (MUST), Grenoble (LPSC), Lyon (CC-IN2P3)

Duplication par transfert asynchrone (FTS)

Accès local *(Analysis Download Direct IO, Production Download, Production Upload)*

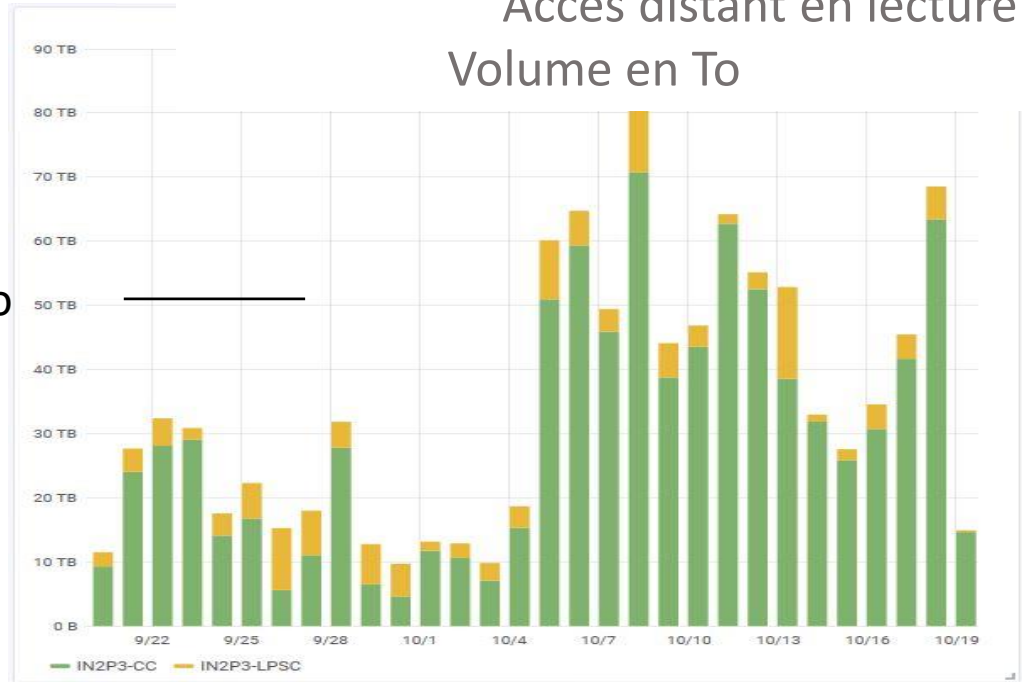
Accès "*remote*" en lecture uniquement

Perspectives - Etudes envisagées

- Gains potentiels : meilleure utilisation du stockage, meilleure réactivité pour l'accès aux données
- Impacts sur les performances : efficacité CPU, I/O, bande passante réseau
- Seuil à partir duquel l'accès à distance pénalise les I/O et nécessite la mise en place de cache au niveau des sites
- Amélioration des outils de *monitoring*

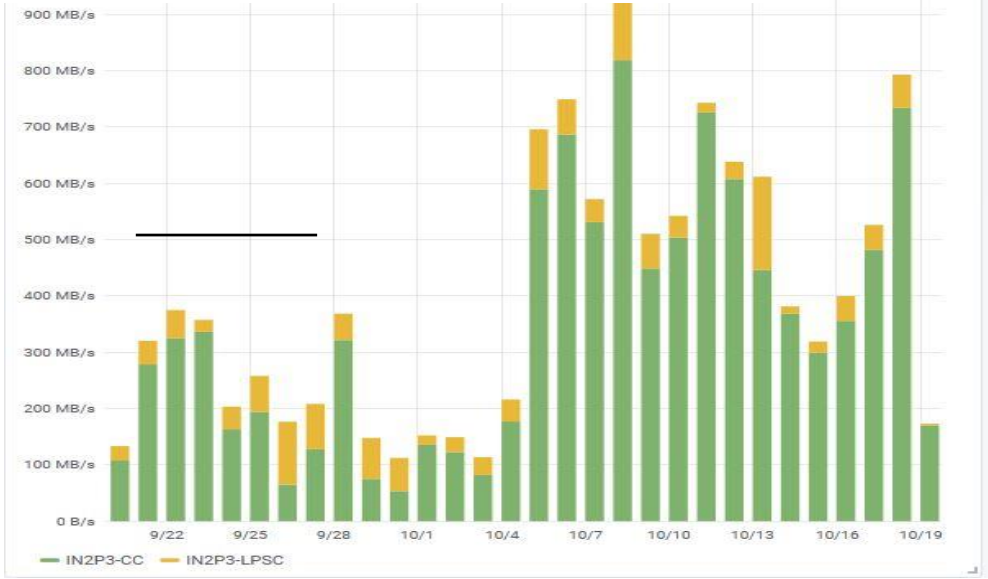
Accès distant en lecture par les nœuds de calcul de MUST

Volume en To



50 To

Bande Passante en Mo /s

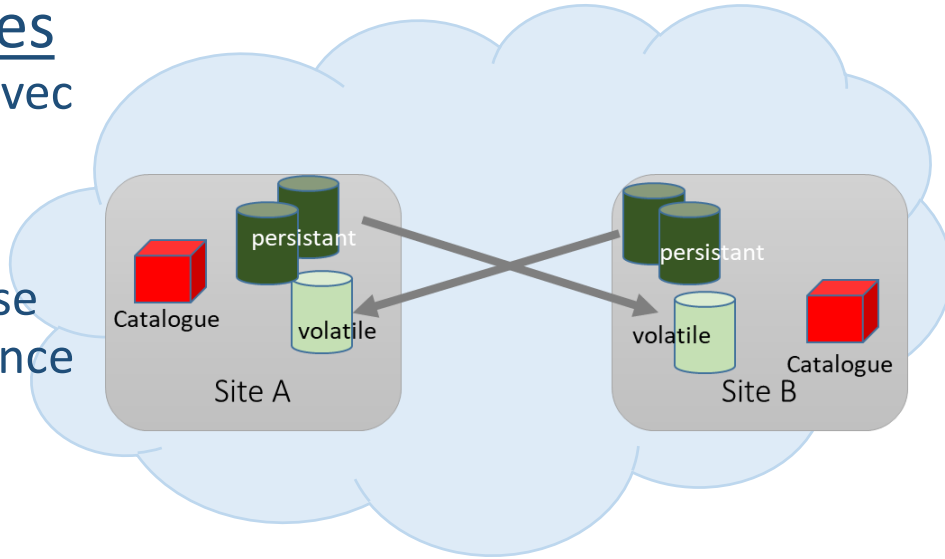


500 Mo /s

Perspectives - Etudes envisagées

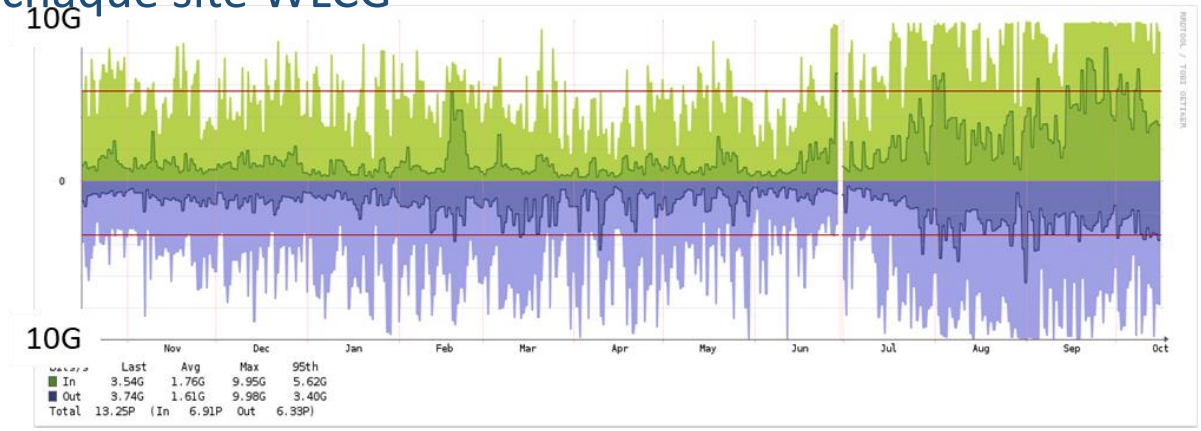
Fédérer les services de stockage à travers des caches

- Evaluation cache intégré « *Volatile pool* » de DPM en coll. avec les Tier 2 français
- Evaluation des solutions de caches « *standalone* » et des possibilités de caches fédérés
- Objectifs : Minimisation de la latence, acquisition d'expertise
- Intérêt à confirmer avec les workflows effectifs de l'expérience



Superviser l'utilisation du Réseau et évaluer les besoins futurs

- 2 instances de perfSONAR déployées sur chaque site WLCG



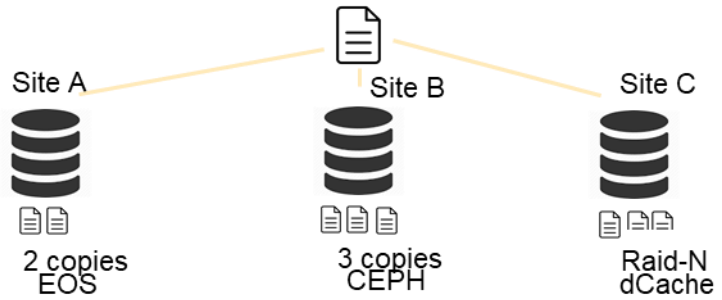
Perspectives - Etudes envisagées

Optimiser la mise en œuvre du stockage disque

- Continuité du service de calcul en cas d'arrêt pour maintenance du stockage
- Evaluation de stockage fédéré (s'appuyant sur DPM) en coll. avec les Tier 2 français
- Ceph comme alternative à DPM, composant middleware spécifique
- Interopérabilité à long terme avec les standards des futurs services de stockage type Data-Lake

Evoluer vers une infrastructure distribuée, multi-site, fédérée… annonçant ce que pourrait être une architecture type *Data-Lake*

- Services de stockage moins nombreux mais plus importants
- Infrastructure intégrant une partie de l'intelligence et de l'orchestration
- Introduction de QOS : classes de service (latence des accès, niveau de fiabilité, de persistance, coût etc..)
- Sécurité de la donnée : réévaluation de la redondance locale et globale



Crédits : WLCG (CERN)

Conclusion

- Le mésocentre MUST présente la particularité de s'inscrire dans trois contextes différents : universitaire, national et international
- Approche intégratrice pour l'accès à des ressources et services divers au sein d'une même plateforme
- Continuum de compétences au sein de l'équipe technique de MUST
- Calcul des expériences LHC (WLCG) reste une « *driving force* » mais ce n'est plus la seule
- Une spécificité « orientée Données » qui est un atout pour le futur
  - Problématique transverse aux communautés : WLCG, CTA…
  - MUST partie prenante d'activités de R&D Gestion de données & Stockage
- 5 ans devant nous pour optimiser la mise en œuvre et l'utilisation du stockage ainsi que les accès aux données scientifiques
- A la clé, des économies d'échelle