Etat du projet GRIF

Aurélien Bailly-Reyre, Frédéric Derue, Victor Mendoza
gestionnaire : Hager Baalouchi

http://lpnhe.in2p3.fr/grif/

Réunion projets du LPNHE, 2 février 2021

HTC ou High Throughput Computing, aka « grille » :

Environ 20 000 coeurs de calcul (dont 4560 au LPNHE),
~9 peta-octets d'espace disque (dont 3.3 Po au LPNHE)
Tier 2 (ressources dédiées) au LPNHE pour ATLAS, LHCb, CTA, ISC-PIF
Tier 3 (best effort) au LPNHE pour ATLAS, LHCb, CTA, HESS, COMET, CMS, france-grilles….

Cloud Computing :

Environ 8000 coeurs de calcul (dont 400 au LPNHE)
~400 To de stockage sur disques (dont 30 To au LPNHE)

Les grilles de calcul (HTC)

Dissémination dans de nombreux instituts de la puissance de calcul, des données et des logiciels de manière efficace et automatisée
Les utilisateurs sont regroupés au sein d' Organisation Virtuelle (VO), suivant leurs collaborations. Chaque VO a accès à un certain nombre de sites

Principales grilles mondiales

- **European Grid Infrastructure**
  - 325 centres dans 56 pays
  - ~240 VOs, 19 clouds
  - ~1.1M cpu, ~1.6 M jobs/j
  - ~740 Po de stockage
- **Open Science Grid**
  - le pendant US de EGI
- **World Lhc Computing Grid**
  - spécifique au LHC
  - hiérarchisation des centres/Tiers
    - Tier 0 au CERN
    - 13 Tier 1 cpu/bandes/disques
    - 160 Tier 2 cpu/disques
    - Tier 3 (ressources non pledgées)

En France

- **France Grilles**
  - GIS pour la grille et le cloud
  - interdisciplinaire
  - réunions régulières grille et cloud
- **LCG France**
  - spécifique au LHC
  - le Tier 1 + 10 sites Tier 2 CNRS/IN2P3 (+le T2 CEA/IRFU)
  - >90% du cpu et du stockage de France Grilles dans ces sites
  - réunions mensuelles techniques et comité de direction

La France représente ~12% du cpu utilisé depuis 1 an par EGI (~9% EGI+OSG)

Le cloud (académique) computing

Le modèle « Infrastructure-as-a-Service (Iaas) »

- l'infrastructure (e.g. machines, réseau) est virtualisée
- sépare la maintenance du hardware du setup des logiciels spécifiques
- le cycle de vie de cette infrastructure virtuelle est gérée par un système de cloud :
  - gestion des images des machines virtuelles (VMs)
  - l'utilisateur peut mettre en route et gérer les VMs
  - le stockage peut etre attaché à ces VMs

Technologies de cloud

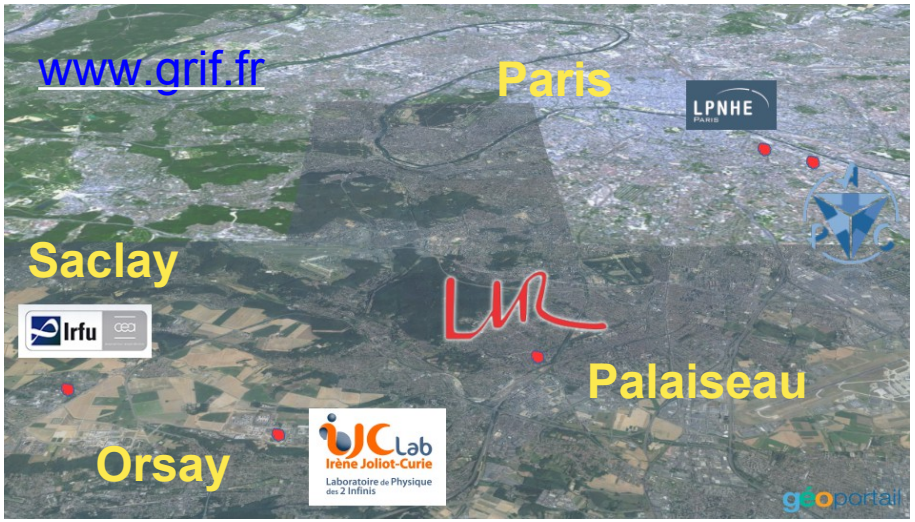
- **OpenStack**
  - environnement IaaS open source
  - utilisé par grandes entreprises (IBM, HP ...)
  - API standarisé (Amazon AW)
  - accepté par les grands centres HPC
  - communauté HEP active
- **Docker/Singularity**
  - alternatives à une virtualisation complète
- **Des outils HEP prêts**
  - CERNVM : machines virtuelles
  - CERNVM-FS : système de fichiers
  - HT-Condor : système de batch
  - DIRAC/VM-DIRAC : système de batch et gestion de données de LHCb

Site de cloud

- **Le site fournit**
  - configuration des machines
  - allocation du stockage
  - « Scalabilité » des ressources allouées
  - configuration du réseau
- **Cloud public et commercial**
  - Amazon Web Service, etc …
  - déjà utilisé en HEP
  - bien sur il faut payer …
  - ok pour le calcul mais pour le stockage massif ?
- **Cloud privé et académique**
  - CERN-cloud …
  - sites France-Grilles/Cloud
  - projets européens

GRIF un Tier 2 distribué en Ile de France

- **GRIF fédère (depuis 2005) les efforts des labos HEP en Ile de France**
  - Tier-2 des 4 expériences LHC dans le cadre de WLCG/LCG-FR et d'autres expériences majeures (Belle II, CTA)
  - fournit des ressources grille pour d'autres expériences (VOs) des laboratoires et des VOs interdisciplinaires (biomed, complex-system *ISC-PIF*)
  - fournit aussi des ressources cloud
  - chaque sous-site est responsable de son projet scientifique (VO, fair-share etc.)
  - travail technique effectué en commun (dont réunions régulières)



- **Ressources humaines**
  - 1 représentant scientifique par laboratoire
  - ingénieurs : 10 syst admins représentant ~6 FTEs (e.g ~1.3 FTEs au LPNHE)

- **Agences de financement (personnel, matériel, infrastructure, ...)**
  - CEA/IRFU et CNRS/IN2P3 (à travers LCG-FR)
  - mais aussi Ecole Polytechnique, Université Paris Saclay, Sorbonne Université, …
  - + ressources propres (laboratoires, groupes, autres (e.g DIM-ACAV))

Salles machines

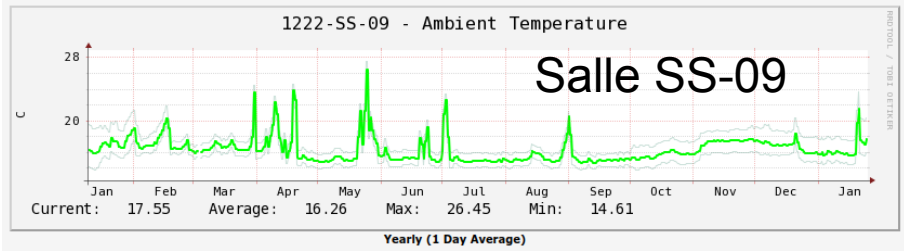
GRIF a 4 salles serveurs : IJCLab, IRFU, LLR et LPNHE

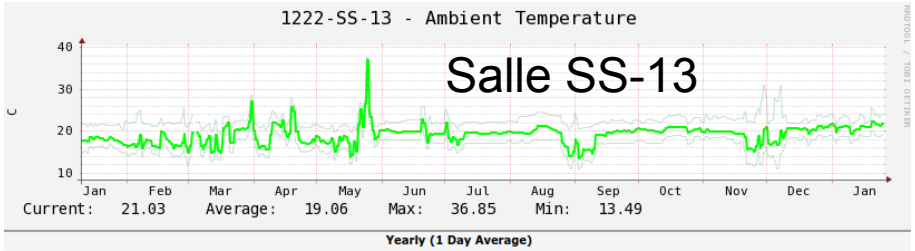
→ financements locaux (mise à niveau, électricité, fluides)

- **Au LPNHE**
  - 2 salles de 65 m^2 , 12 châssis par salle

    → GRIF est dans SS-13, SI+autres labos dans SS-09
  - climatisation par eau glacée de 200 kW (1 seule arrivée)

    → arrêts intempestifs → arrêt automatique des serveurs de calcul permet d'éviter de trop fortes montées en SS-13 (pas en SS-09) → pertes de performances mais pas d'arrêt du site !

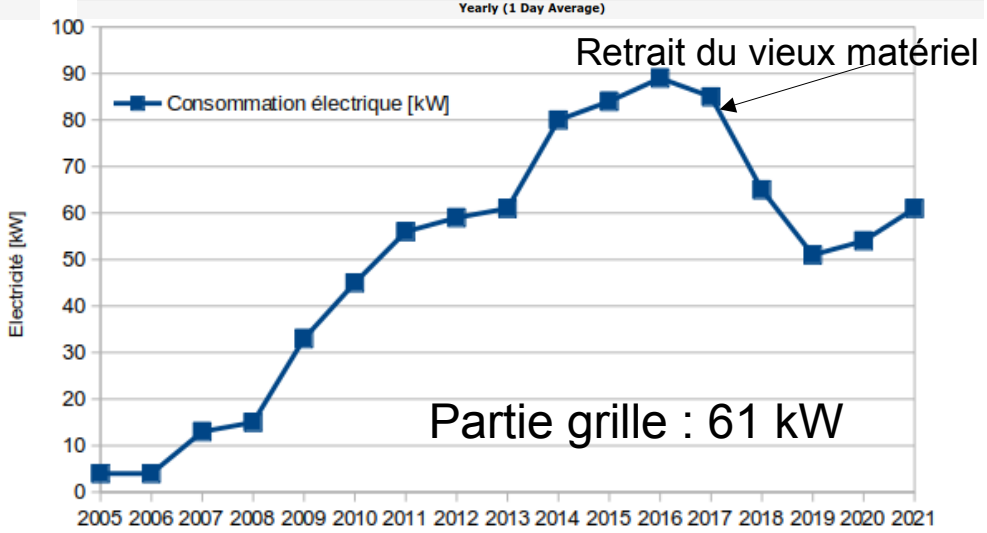




- **Consommation électrique**
  - alimentation ondulée (dont GRIF) : 130 kVA, non-ondulée : 80 kVA
  - onduleur de 200 kVA
  - consommation en 2021

    → 61 kW pour la grille

    → 21 kW pour le cloud



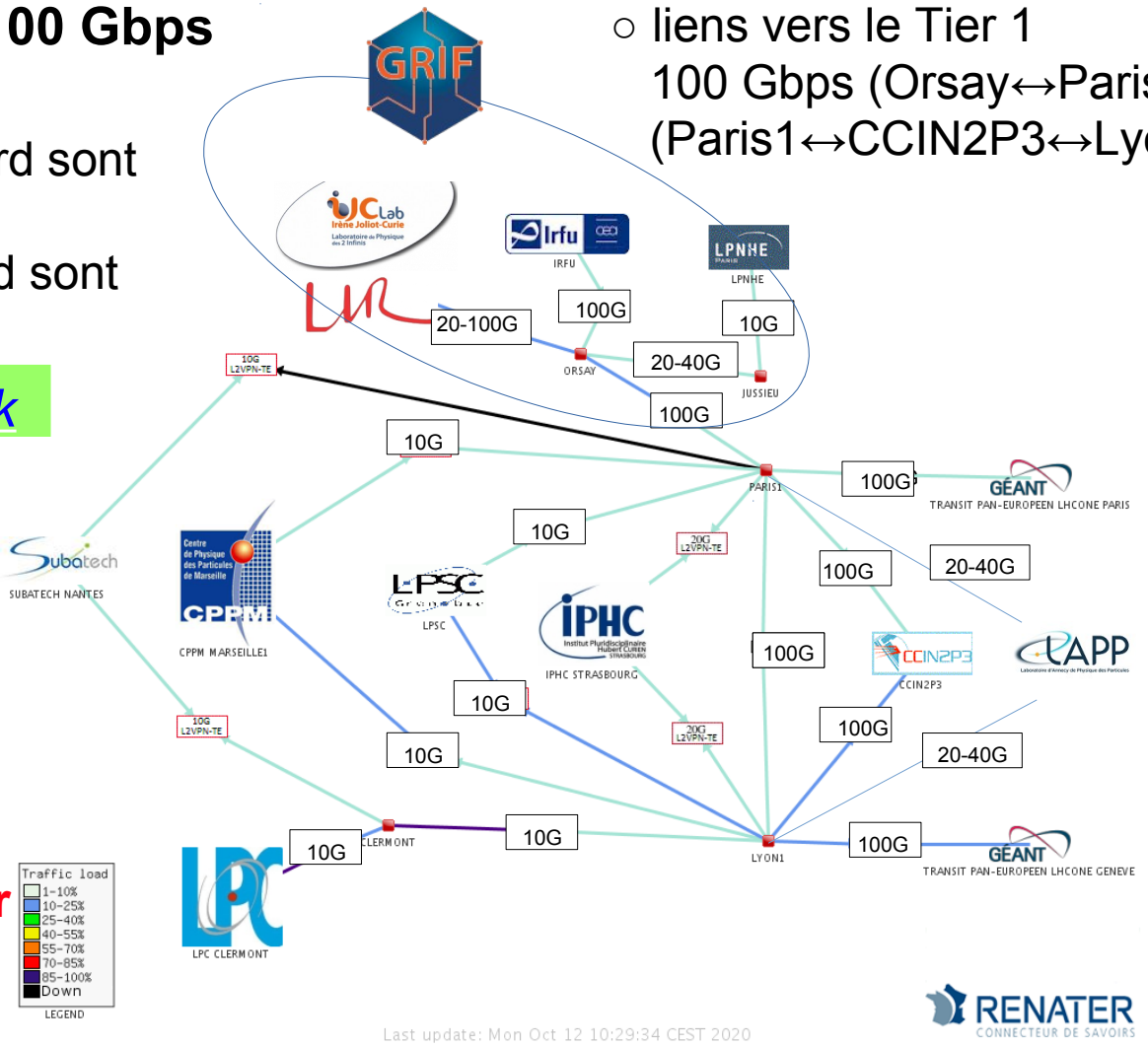
Le réseau

- **LHCOPN/GEANT @100 Gbps**
- **LHCONE**
  - liens entre les sites nord sont à 10-100 Gbps
  - liens entre les sites sud sont à 10-40 Gbps

Mises à jour prévues *link*

Site	2020	2021	2023
CPPM	10	20	40
IRFU	100	100	100
IJCLab	20	100	100
LAPP	20	40	40
LPC	10	20	40
LPNHE	10	20	40

Peut être insuffisant pour assurer à la fois la production et la R&D



- liens vers le Tier 1 100 Gbps (Orsay↔Paris1) et (Paris1↔CCIN2P3↔Lyon1)

- **Réseau interne**
  - achats (cher!) de plusieurs switchs évolutifs 10-40-100 Gbps
  - cœur de réseau 2*40 Gbps en fibre optique, stockage 25 Gbps, calcul 10 Gbps

VOs soutenues et ressources disponibles

- **VOs**
  - Tier-2 pour les 4 expériences LHC : ALICE (IJCLab, IRFU), ATLAS (IJCLab, IRFU, LPNHE), CMS (IRFU, LLR), LHCb (IJCLab, LPNHE)
  - Tier-2 pour CTA (IJCLab, IRFU, LLR, LPNHE
    - eCheops avec un seul SE au LLR (+IJCLab)
      → DIM-ACAV (J-Ph. Lenain, 2020)
  - autres VOs (pas sur tous les sous-sites) :
    HEP : Auger, Belle, Calice, Clas12, Comet, HESS, ILC, T2K
    non HEP : biomed, compchem, complex-system, france-grilles, vo.ipno

LPNHE T2 pour ATLAS, LHCb, CTA et ISC-PIF
+ ressources pour HESS, COMET, CMS, France Grilles

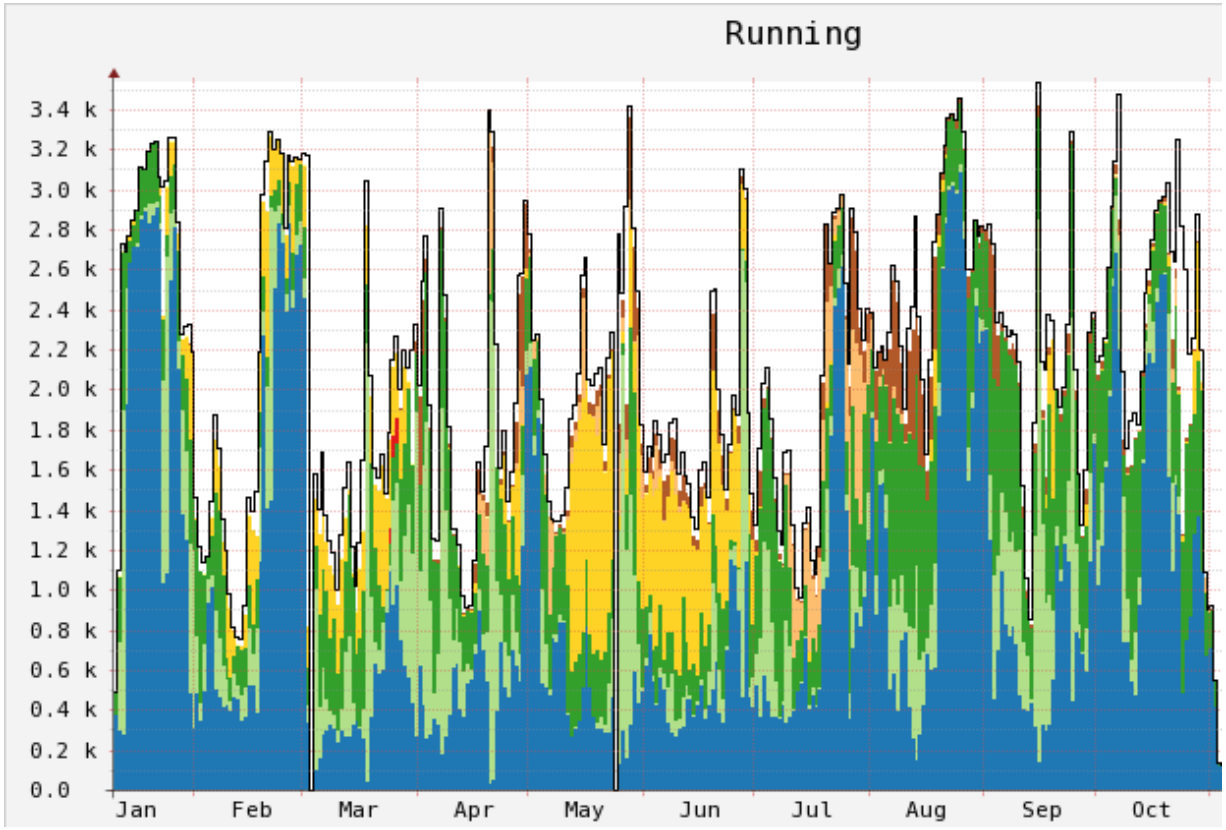
- **Ressources Tier 2 (pledgées, sous garantie)**
  - Pledge global (2021) pour les 4 expériences LHC : 130,000 HS06, 10,000 TB
  - ATLAS : 53,300 HS06 et 5,000 TB, LPNHE ~30% du total
    - 3 CE et SE (IRFU, IJCLab, LPNHE)
      → site de taille moyenne mais vu depuis ATLAS comme trois sites petits
  - LHCb : 16,900 HS06 et 300 TB, LPNHE ~30% du total
  - CTA : 6,000 HS06, 500 TB, LPNHE~20% total
  - ISC-PIF : 30,000 HS06, LPNHE~30% du total

LPNHE contribue à ~30% du total de chaque pledge T2

- **Ressources Tier 3 (non pledgées, hors garantie)**
  - e.g 3 ATLAS LOCALGROUPDISK >1 PB (essentiellement au LPNHE)
  - aussi du cpu (e.g pour ATLAS au LPNHE autant pledgé que non pledgé)
  - aussi pour les expériences non-LHC

Ressources de grille au LPNHE

- **En 2020 :**
  - ~3500 cpu/jobs et 2500 To de stockage
  - liaison 10 Gbit/s
  - GRIF LPNHE fournit >50% plus que ses pledges pour ATLAS, LHCb et CTA



NB : ATLAS envoie des jobs multicœurs (8) donc le max job=3500 (=# cpu) ne peut pas être atteint

>85% du calcul utilisé par ATLAS et LHCb

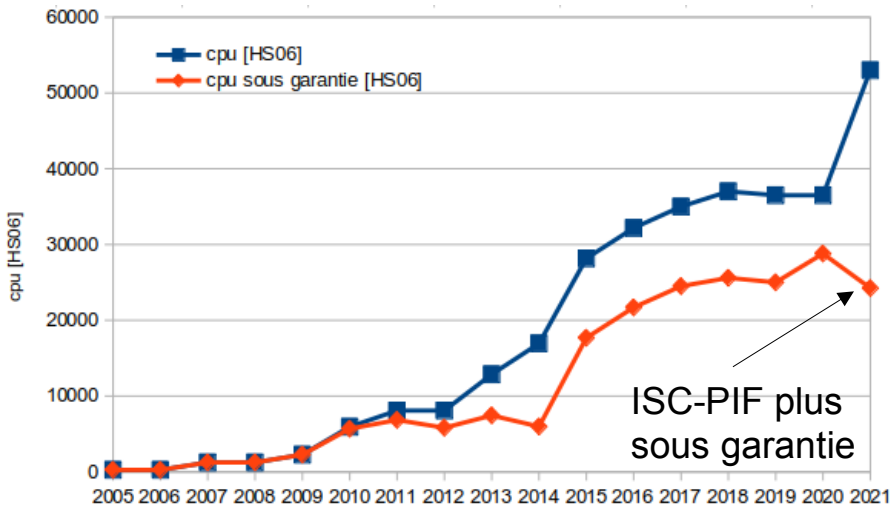
Implication importante d'ATLAS sur le suivi des performances des sites

cpu : ATLAS 68%, LHCb 18%, ISC-PIF 4%, CTA 4%, CMS 3%, HESS 3%, FR-grille <1%

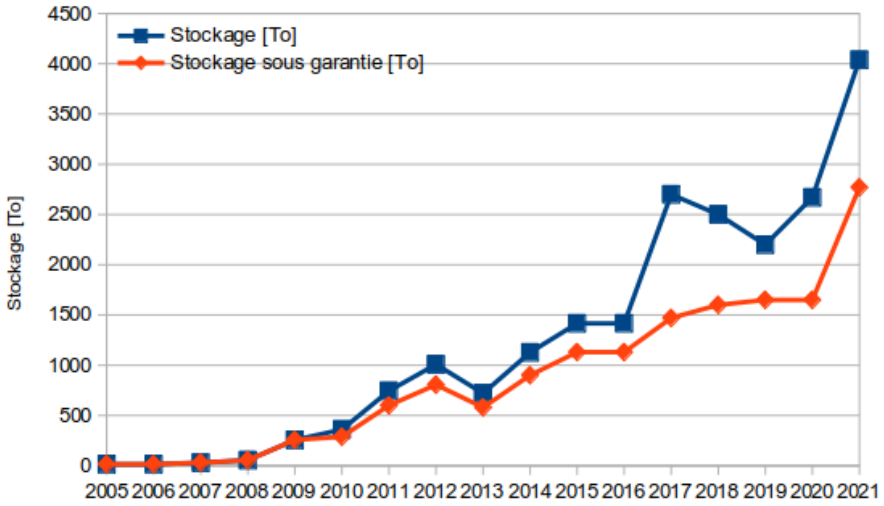
Stockage : ATLAS T2 ~1050 TB, ATLAS T3 ~1000 TB, LHCb T2 ~130 TB
CTA ~120 TB, HESS ~300 TB

Ressources de grille au LPNHE

- **En 2021 :**
  - 165 serveurs : 21 pour le stockage, 14 services et 130 calcul
  - 256 cpu et 4560 cœurs
  - 650 disques pour 3,3 Po utiles



Evolution du cpu



Evolution du stockage

après plusieurs années de stagnation, en 2021

cpu
ATLAS +30%
LHCb +50%
CTA +200%

stockage
ATLAS +50%
LHCb +25%
CTA +30%

Le cloud au LPNHE

- **Cloud Openstack**
  - 51 nœuds de calcul, 102 CPU, 408 cœurs, 51*16 Go de RAM
  - ressources prises sur du vieux matériel grille hors garantie
    → peu de choix de machines variées (e.g grande RAM)
  - le site cloud permet des tests de virtualisation
    - intéressant pour le service informatique
  - on peut l'utiliser (presque) comme un système de batch / stockage
    - quelques chercheurs déjà intéressés à faire des essais pour tourner leurs applications favorites et faire un retour d'expérience
  - autres utilisateurs
    - Jennifer2 (collaborateurs de M. Guigue, T2K/HK basé sur middleware Belle2/Dirac) depuis qqs jours

    ⇒ **augmenter le nombre d'utilisateurs au sein du laboratoire**

- **Autres projets**
  - les nouveaux projets portés par l'Union Européenne concernent le Cloud
    - le CC-IN2P3 et France-Grilles/Cloud sont déjà impliqués par exemple projet European Open Science Cloud EOSC (https://www.egi.eu/tag/eosc/)

  ⇒ **opportunités existantes (financement) à saisir avec le site maintenant en production**

Ressources humaines au LPNHE

- **Coté technique**
  - V. Mendoza (IR1 CNRS) : sur le projet depuis ~2007
    0.8 FTE sur GRIF : ~0.7 FTE sur la grille, 0.1 FTE sur le cloud
  - A. Bailly-Reyre (IGE Sorbonne Université) : sur le projet depuis 2017
    0.5 FTE sur GRIF : 0.3 FTE sur la grille, 0.2 FTE sur le cloud
  - Répartition typique du travail sur une année
    - 33% gestion, mises à jour → suivi des besoins de site mais aussi des demandes des expériences
    - 27% opération du site → non stop
    - 20% matériel et infrastructure → les achats se préparent/se font ~septembre
      le nouveau matériel est installé/testé de décembre à avril
    - 20% développement (maquette, tests, veille) → quand il reste du temps ….
  - **assez peu (! trop peu) de temps pour la R&D alors que celle-ci va devoir augmenter dans les mois/années qui viennent pour préparer la phase HL-LHC**

- **Coté scientifique**
  - F. Derue → resp. scientifique ~0.1 FTE sur GRIF
    aussi dans groupe Calcul ATLAS France (~0.5 FTE sur le calcul/stockage)

Ressources financières au LPNHE

- **Coté LHC**
  - budget LCG-FR suivant convention 2018-2022
  - réparti entre les labos au prorata du matériel à renouveller
  - couvre 70% des besoins de renouvellement du matériel sortant de garantie – 30% à trouver par ressources propres (labo, groupes)
  - ne couvre pas les besoins d'augmentation des expériences, ~10-20% par an ! qui représente autant (en k€) que le renouvellement
  - typiquement 25 k€/an par LCG-FR (hardware+missions)
  - 2020/21 première fois depuis 3-4 ans que le site grossit de manière importante !

- **CTA**
  - le matériel CTA arrive en fin de garantie cette année
  - heureusement DIM-ACAV obtenu en 2020 (eCheops porté par J-Ph. Lenain) permet(tra) le renouvellement et l'augmentation des ressources. Le LPNHE contribue pour 34% en complément de la région Ile de France

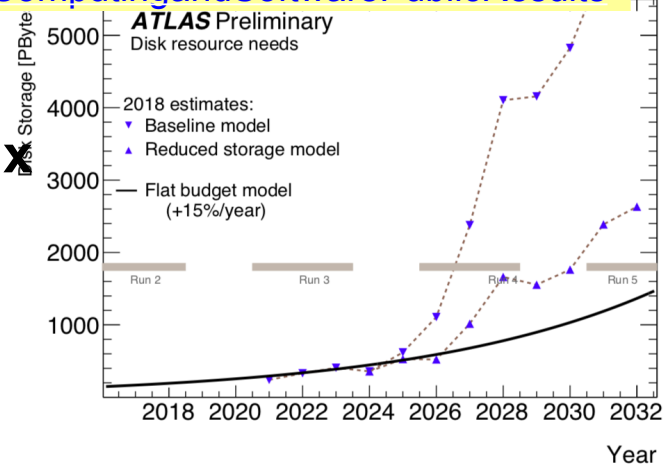
- **Cloud**
  - né et vit uniquement en utilisant du vieux matériel hors garantie issu de la grille
  - le site est petit et a peu de variété de matériel (#coeurs, RAM)

Evolution du stockage for pour HL-LHC

- **Extrapolation HL-LHC**
  - extrapolation des ressources de stockage vont bien au-delà de ce qui est atteignable à budget constant
- **Modèle d'analyse du Run 2 est trop dispendieux**
  - formats de données plus petits dès le Run 3 ...
  - data tape carrousel dans les Tier 1s
- **Evolution de l'architecture**
  - trop de end-points du point de vue des manips
  - évolution vers une architecture dite « Data Lake »
    - stockage distribué autour de centres connectés par un réseau rapide et à petite latence
    - plus grande diversité de centres : gros, « disk-less », certains utilisant des caches, ou des disques de moindre qualité (si les données ne sont pas accédées si souvent)….
    - de nombreuses composantes sont déjà inclues (Rucio, FTS ….)

      ce qui permet des changements adiabatiques des infrastructures
  - problème additionnel
    - le système disk manager system (DPM) utilisé depuis des années par une grande partie des Tiers 2 (dont les français) amené à disparaître d'ici qqs années (durant le Run 3 ?)
    - une autre technologie de stockage doit être choisie (dcache, Eos, Echo …), avec les projets de R&D Data Lake (DOMA)

ComputingandSoftwarePublicResults



nécessite de la R&D à effectuer par les syst admins
(et suivi par des chercheur(e)s) tout en maintenant les sites en production

Evolution à court/moyen terme de GRIF

- **Ressources humaines**
  - organisation déjà existante dans GRIF et LCG-FR permet d'optimiser HR
  - certains sites (e.g LPSC) s'arrêtent d'ici 2023 par manque de personnel
  - besoin à la fois d'assurer le site en production tout en faisant de la R&D (en particulier pour le stockage)
- **Calcul (CEs)**
  - court terme : réduction à 2 pools pour le calcul
    1 pour le CNRS/IN2P3 (IJCLab, LLR, LPNHE) pour les 4 VOs LHC,
    1 pour le CEA/IRFU pour ALICE, ATLAS, CMS
  - ressources locales déjà inclues dans grille/cloud et un cluster batch (IRFU)
- **Stockage (SEs)**
  - passer de plusieurs end-points (un par sous-site) à un seul (idéalement) global pour permettre aux VOs d'accéder au stockage de manière plus simple
  - moyen terme (~été 2021) : premier démonstrateur de stockage unifié différentes technologies à étudier vs backend/frontend
- **Soutien aux VOs**
  - augmentation des pledges pour les 4 expériences LHC (LCG-FR et autres FAs)
  - continue le soutien aux VOs non-LHC, avec un stockage augmentant (e.g CTA)

Conclusion

- **GRIF**
  - donne des ressources à de nombreux différents projets (à travers la grille, le cloud) pour différentes collaborations (4 expériences LHC, Belle II, CTA et HEP, non-HEP) et incorpore des serveurs de calcul de projets non-HEP (ISC-PIF)
  - évolution à moyen/long terme est dirigée par les besoins des expériences LHC – mais pas seulement
  - ressources financières plus que limites pour assurer l'augmentation des ressources demandées par ATLAS et LHCb – 2020 a été malheureusement exceptionnel ...
  - l'utilisation du cloud a du mal à décoller

- **Evolution vers un Data Lake**
  - la gestion du stockage devra être modifiée en 2021 (moins de end-points) et se basera sur une augmentation de l'utilisation du réseau
  - GRIF installe déjà les différents outils demandés par les expériences (ATLAS/CMS) et pourra alors se baser sur la connaissance déjà acquise dans les sites français «sud » (ALPAMED) dans DOMA-FR pour tester un Data Lake