Laboratoire d'Annecy-le-Vieux de Physique des Particules

Rapport de site LAPP Tier 2

Mésocentre MUST

Journées LCG France

22 mai 2019

Mathieu Gauthier-Lafaye

cnrs

UNIVERSITÉ SAVOIE MONT BLANC

Plan

- Mésocentre MUST
- Utilisateurs
- Équipes LAPP/LAPTh
- Infrastructures
- Ressources & Services
  - Calcul
  - Stockage
  - Réseau
- Supervision
- Projets en cours / futurs

Mésocentre de Calcul et de Stockage Université Savoie Mont Blanc

- Au service d'une dizaine de laboratoires de l'université
- Tier 2 (ATLAS, LHCb) depuis 2007, ouvert sur la grille EGI
- Plateforme labellisée par l'IN2P3/CNRS en 2017
  (V. Beckmann dans le comité de pilotage annuel)

Utilisateurs du mésocentre

- **Activités Grille**
  - **Tier-2 LCG : ATLAS, LHCb**
  - **VOs EGI :** HESS, CTA, LSST, CLIC-ILC, GEANT4, France Grilles
- **Calcul « local » Université Savoie Mont Blanc**

LAPP : Physique expérimentale

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 :

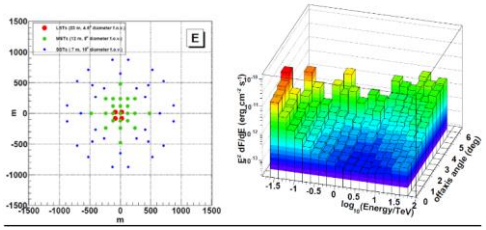
- Application du *Deep Learning* à la reconstruction stéréoscopique des images CTA

- **Gestion de données (iRODS)**

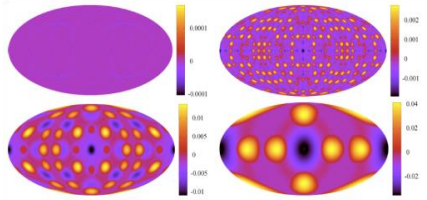
Edytem : Sciences de la montagne

- Étude des sédiments des lacs alpins

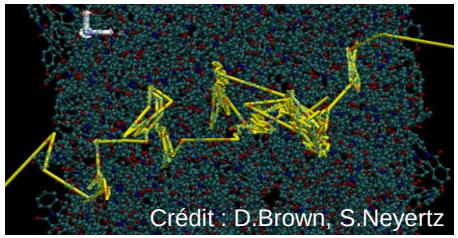
- **Ouverture au monde industriel via la Fondation USMB**

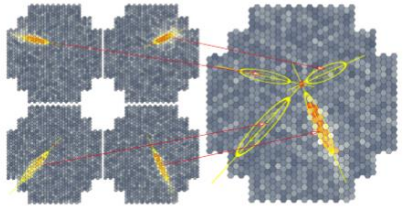


Crédit : CTACG hal-00653017



Crédit LAPTh

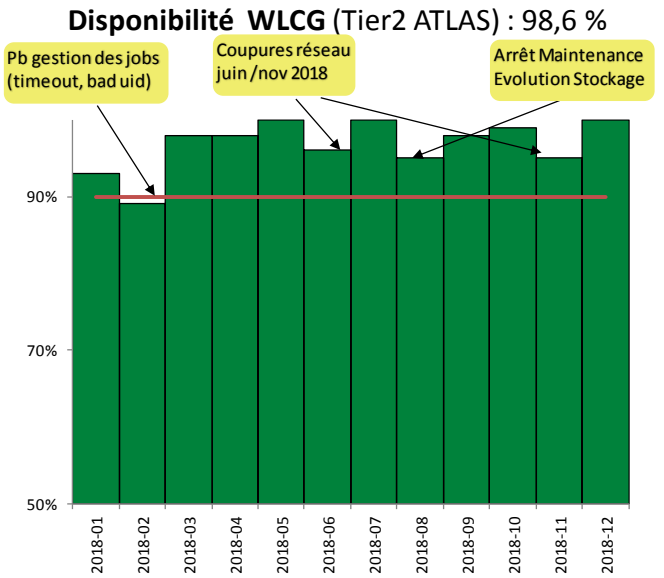




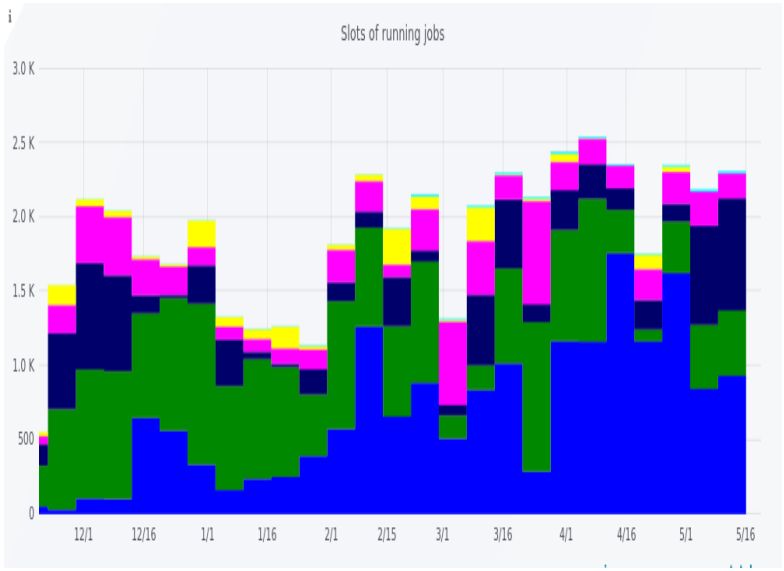
Crédit K.Bernlohr (HESS figure)

ATLAS Tier 2 nucleus

- Résultat d'une bonne collaboration avec le computing ATLAS, en plus de
  - Capacité du stockage
  - Bon fonctionnement du site
- Activités de calcul variées « T1 like » (Data processing et MC reconstruction)
- Plus de types de données hébergées
- Plus de trafic réseau



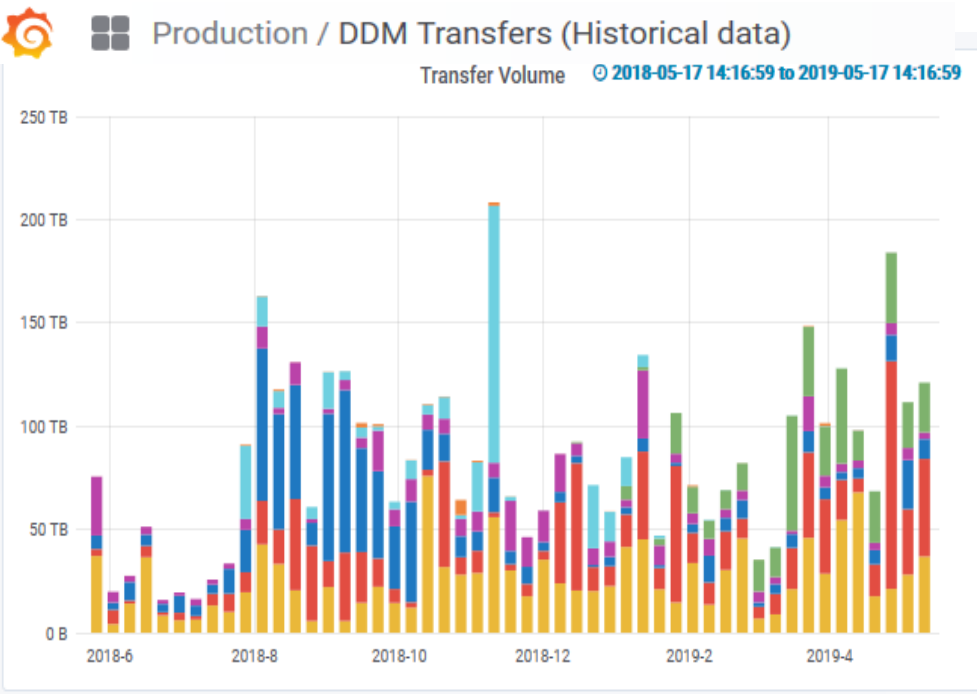
Types de jobs au LAPP durant les 6 derniers mois

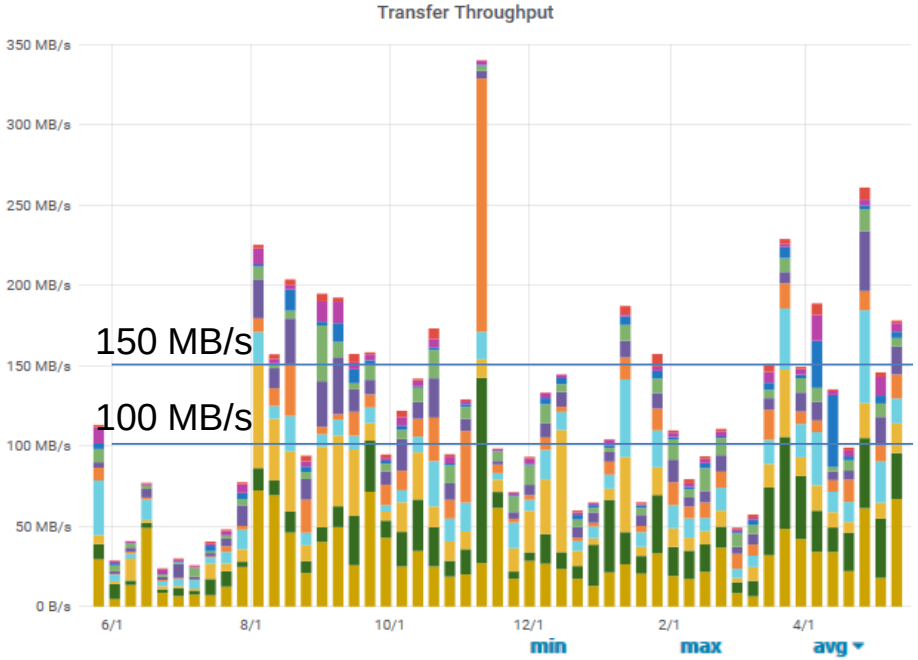


Analysis
MC Simulation
MC Reconstruction
Data Processing
Group Production

ATLAS Tier 2 Nucleus

Trafic plus important depuis le LAPP (source de données)





	avg	total
Data Consolidation	25.0 TB	1.3240 PB
Production Input	18.9 TB	1.0035 PB
Production Output	15.7 TB	834.1 TB
User Subscriptions	7.6 TB	400.5 TB
Analysis Input	7.2 TB	384.2 TB
Data rebalancing	6.4 TB	340.3 TB
Express	358 GB	19.0 TB

	min	max	avg
USA	0 B/s	72.3 MB/s	28.5 MB/s
UK	0 B/s	115.3 MB/s	19.0 MB/s
France	0 B/s	76.1 MB/s	16.9 MB/s
Germany	0 B/s	57.9 MB/s	13.9 MB/s
Italy	0 B/s	157.5 MB/s	11.9 MB/s
Switzerland	0 B/s	37.3 MB/s	10.2 MB/s
Canada	0 B/s	34.7 MB/s	7.0 MB/s
Netherlands	0 B/s	44.8 MB/s	3.9 MB/s
Spain	0 B/s	16.2 MB/s	3.7 MB/s
Japan	0 B/s	7.7 MB/s	2.0 MB/s

Data consolidation : création d'une 2nde copie des données dérivées produites (DAOD) vers un site non nucleus

Équipes LAPP/LAPTh

- Opéré par une équipe mutualisée du LAPP et du LAPTh (CNRS / USMB)
- Équipe de 8 personnes pour 4 ETP
  - Expertises complémentaires
  - Redondance
- L'équipe MUST s'appuie sur le travail des autres membres du support général informatique

Support général informatique

- Salle informatique LAPP : 70m2
- Réseau :
  - Infrastructure réseau (cœurs, …)
  - Postes de travail (filaire et WIFI)
- Services virtualisés sur notre plateforme « CLAP » :
  - Accès réseaux (DNS, DHCP)
  - Gestion des comptes et authentification (AD)
  - Serveurs de calculs interactifs
  - Cloud de stockage (ownCloud)
  - Web, Gitlab, Antivirus, …
  - Supervision (Nagios, Cacti, Prometheus…)
- NAS (NetApp, Freenas)

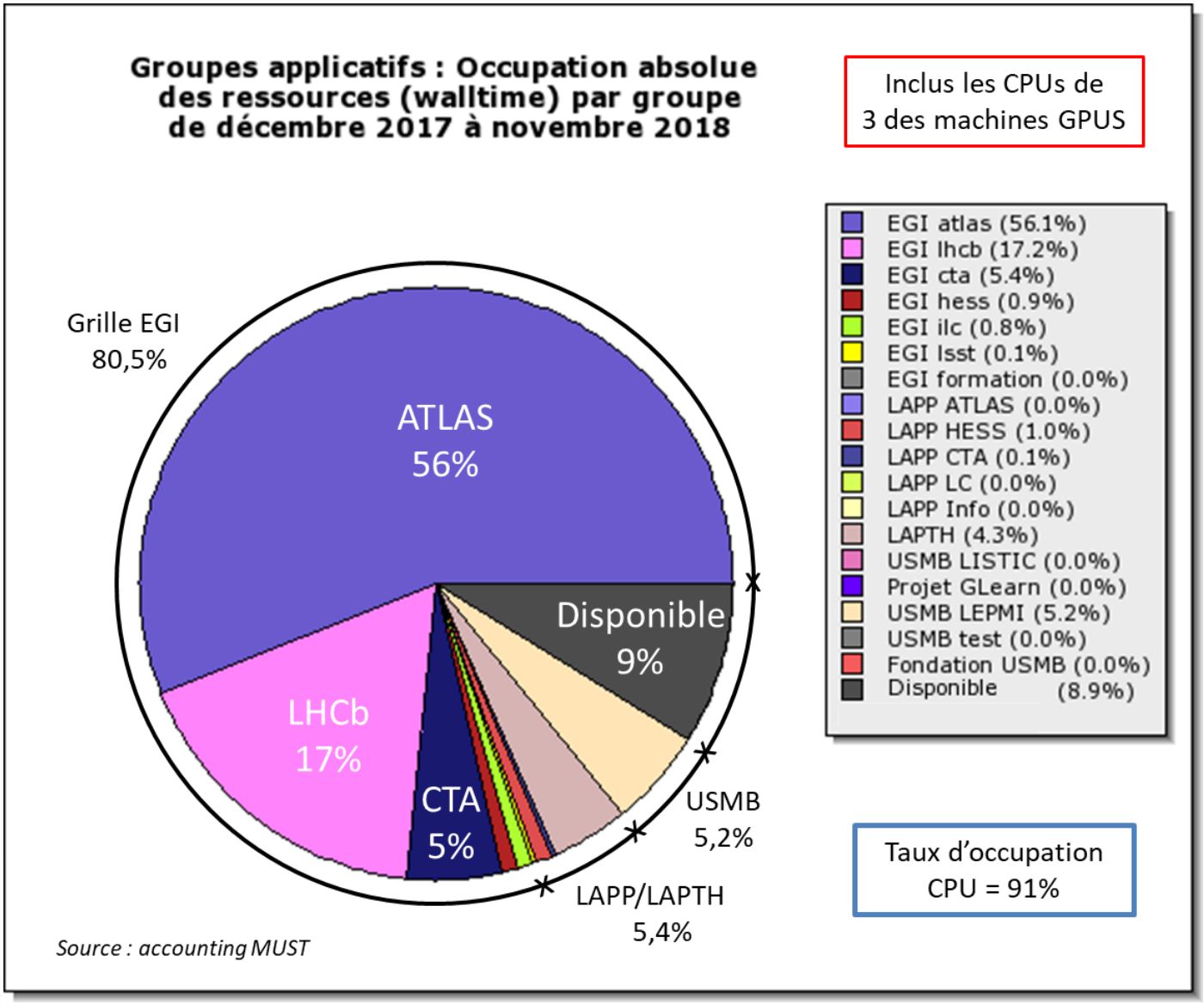
Infrastructures MUST

- Salle Mécatronique
  - Surface : 170 m2
  - Sécurité incendie
  - Onduleur : 500 kVa
  - Climatisation : 240kW (extensible)
  - Taux de remplissage : 14 racks (limite à 50)
  - Puissance consommée : 120kW
  - PUE : 1,6

Calcul

- Le cluster en chiffre :
  - 4000 jobs simultanés (10,9 HS06/job)
  - 42800 HS06
  - 71% de la capacité totale « pledgés »
  - 67% sous garantie
  - 5 WN GPU
- Utilise Torque / MAUI
  - RPM SL6 utilisé sur SL7
- Nouveautés 2018/2019 :
  - Changement de la politique d'achat :
    - Blade HP / Fuji vers des DELL C6400 / C6420
  - Passage en CentOS 7
  - Mise en place de Singularity

Calcul



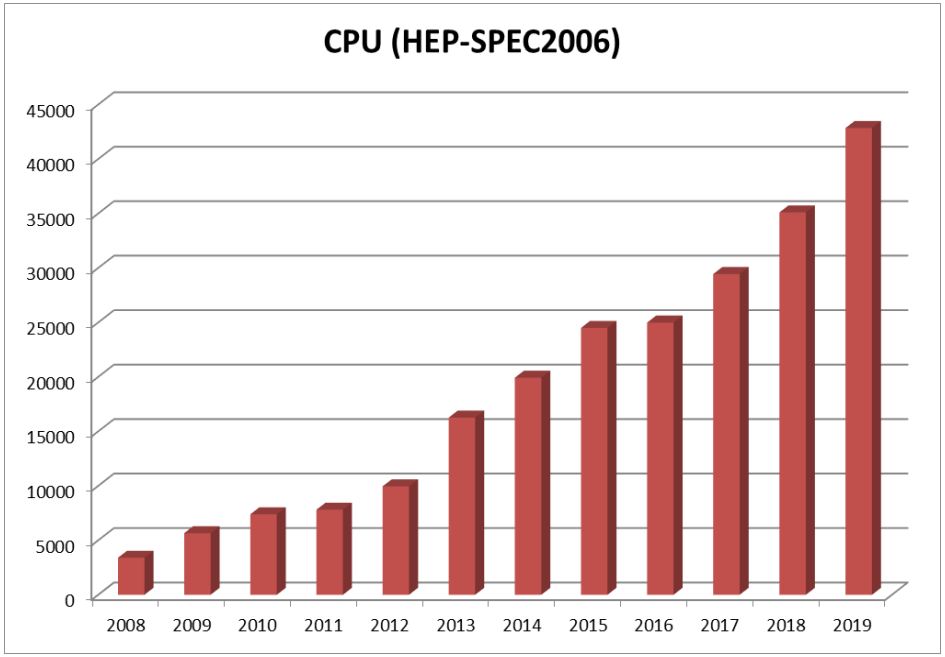
- **Grille : 80 %**
- **Local : 11 %**

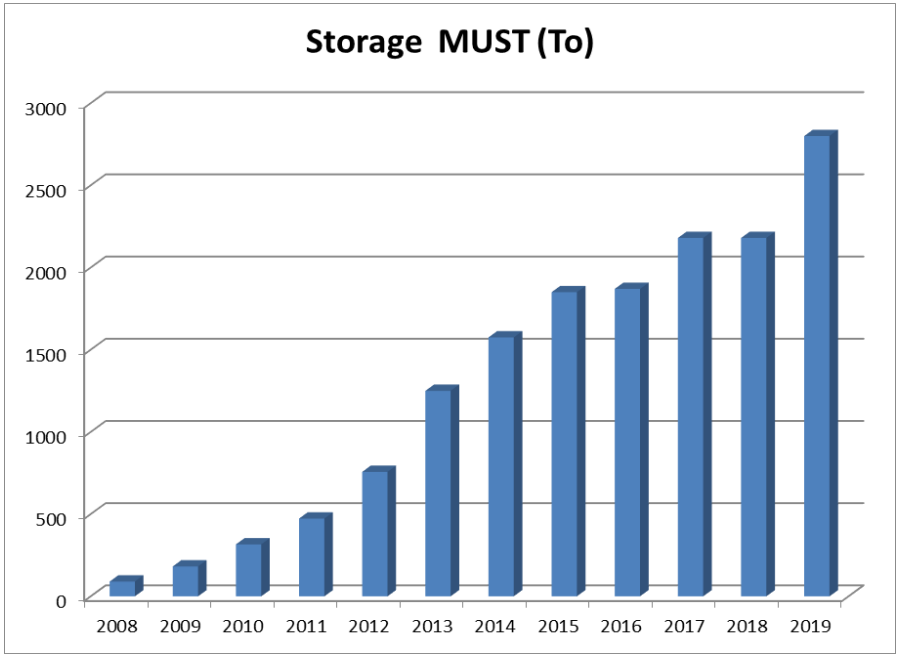
Taux d'occupation GPU = 30 %

Stockage, Gestion de données

- Stockage grille (DPM) : 2600 To
- Stockage partagé MUSTFS : 200 To
  - Home (université, vo, jobs…)
  - Logiciels (grille, local)
  - Données scientifiques (LAPP-LAPTh, Université)
  - Migration de GPFS à CEPHFS en cours
- Stockage et gestion de collections de données (iRODS)

Evolution des ressources

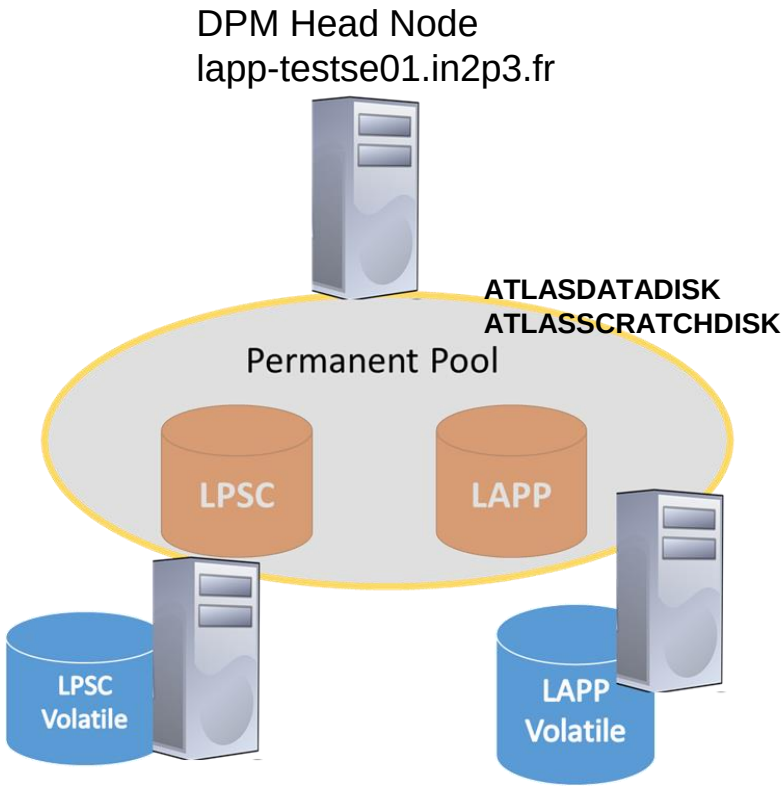




R&D ATLAS LAPP – LPSC

- Testbed FR-ALPES : démarré en février 2019
- Stockage DPM réparti entre le LAPP et le LPSC
  - Support des VOs : ATLAS, dteam
- Motivation
  - Collaboration LAPP-LPSC : administration conjointe d'un stockage distribué sur 2 sites
  - Gain opérationnel et consolidation du stockage possibles à terme
  - Evaluation pertinence des volatiles pools DPM par ATLAS

Cible



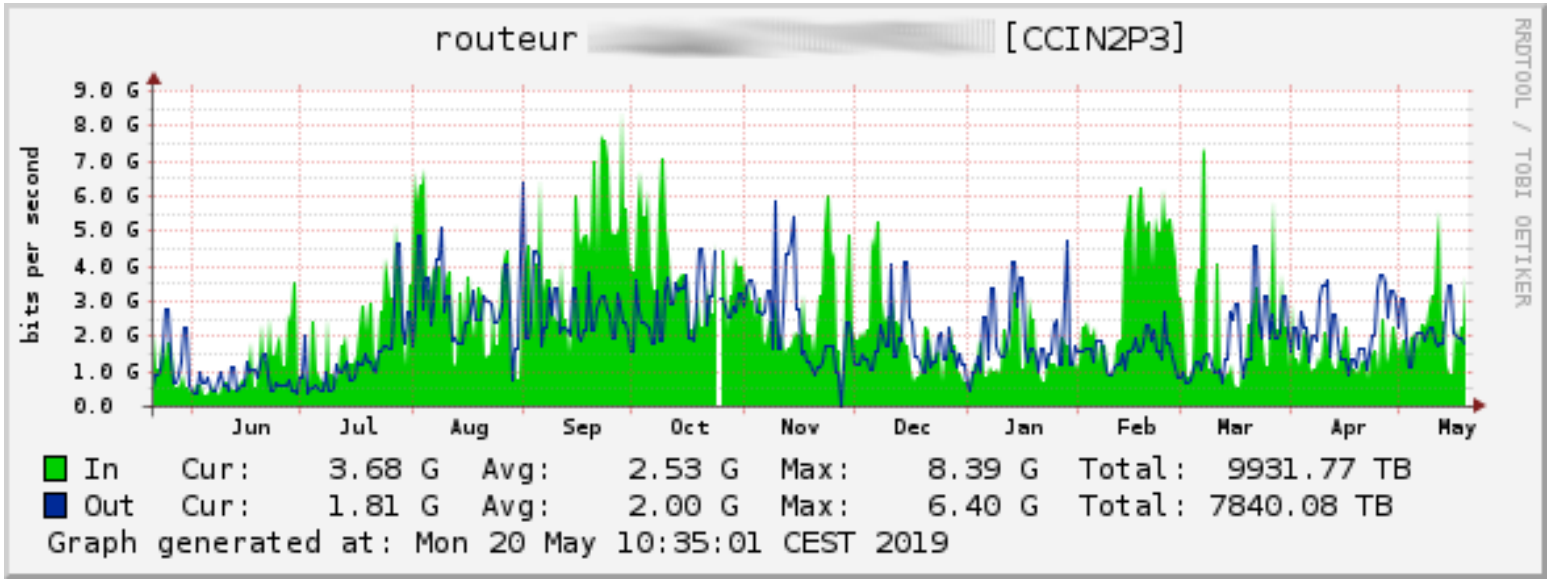
R&D ATLAS LAPP – LPSC

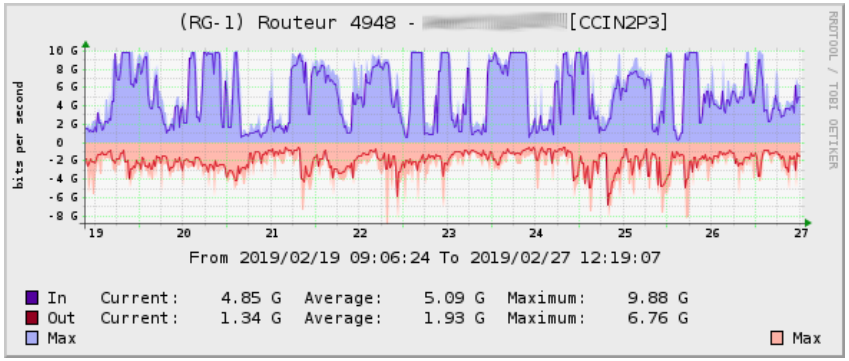
- Valeur ajoutée par rapport aux évaluations antérieures (T2 Italiens notamment) :
  - Collaboration étroite entre sysadmins et computing ATLAS
  - Testbed FR-ALPES totalement intégré dans l'environnement ATLAS
    - SAM Test ATLAS
    - Transferts FTS
    - HammerCloud (Template ABC)
    - Interaction avec Rucio
- Etapes :
  - Mise en place de l'infrastructure distribuée
  - Migration et activation de Dome (DPM v1.12)
  - Évaluation des volatiles pools

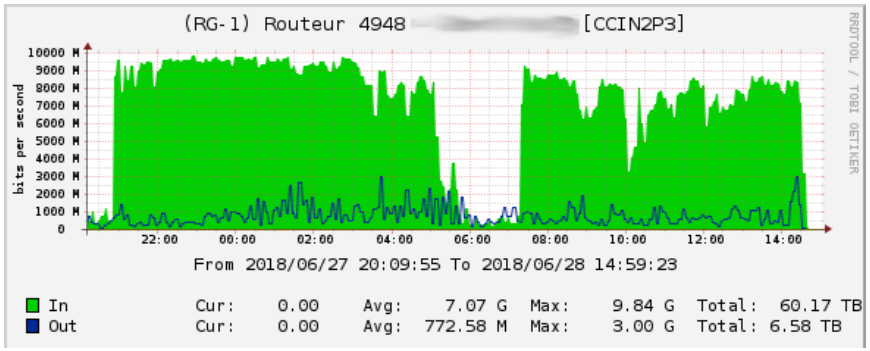
Réseau

- Connectivité WAN via le réseau régional AMPLIVIA
  - 10 Gb/s vers le CC-IN2P3 LHCONE depuis le CC-IN2P3
    - Pics d'utilisation du lien supérieurs à 9 Gbit/s pendant plus de 24h
      - Effets visibles par exemple sur l'utilisation des serveurs interactifs
    - 2 coupures de plus de 24h en juin et novembre 2018
      - Dont une coupure de fibre dans Annecy 48h !
    - LHCONE : l'augmentation de la bande passante sur le lien SUD a eu un effet positif (LHCB a repris la soumission)
  - 1 Gb/s avec routage AMPLIVIA :
    - Problèmes de fiabilité même pour le laboratoire
    - Non exploitable : Sous dimensionner pour lien de secours, pas d'IPv6
    - Bascule manuel en cas de coupure et utilise la même fibre optique
  - Évolution et sécurisation via un deuxième lien physique annoncé depuis 2013 (repoussé de 6 mois en 6 mois)
    - Q2 2018 : Convention CCIN2P3 / AMPLIVIA signée pour 4 X 10 Gb/s à partir du 1er trimestre 2019
    - Q4 2018 : Validée en comité de pilotage AMPLIVIA
    - Q2 2019: Convention retoquée par service juridique de la région

Réseau


routeur [CCIN2P3]
bits per second
9.0 G
8.0 G
7.0 G
6.0 G
5.0 G
4.0 G
3.0 G
2.0 G
1.0 G
0.0
Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May
RRDTOOL / TOBI OETIKER
In Cur: 3.68 G Avg: 2.53 G Max: 8.39 G Total: 9931.77 TB
Out Cur: 1.81 G Avg: 2.00 G Max: 6.40 G Total: 7840.08 TB
Graph generated at: Mon 20 May 10:35:01 CEST 2019


(RG-1) Routeur 4948 - [CCIN2P3]
bits per second
10 G
8 G
6 G
4 G
2 G
0
-2 G
-4 G
-6 G
-8 G
19 20 21 22 23 24 25 26 27
RRDTOOL / TOBI OETIKER
From 2019/02/19 09:06:24 To 2019/02/27 12:19:07
In Current: 4.85 G Average: 5.09 G Maximum: 9.88 G
Out Current: 1.34 G Average: 1.93 G Maximum: 6.76 G
Max
Max


(RG-1) Routeur 4948 [CCIN2P3]
bits per second
10000 M
9000 M
8000 M
7000 M
6000 M
5000 M
4000 M
3000 M
2000 M
1000 M
0
22:00 00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00
RRDTOOL / TOBI OETIKER
From 2018/06/27 20:09:55 To 2018/06/28 14:59:23
In Cur: 0.00 Avg: 7.07 G Max: 9.84 G Total: 60.17 TB
Out Cur: 0.00 Avg: 772.58 M Max: 3.00 G Total: 6.58 TB

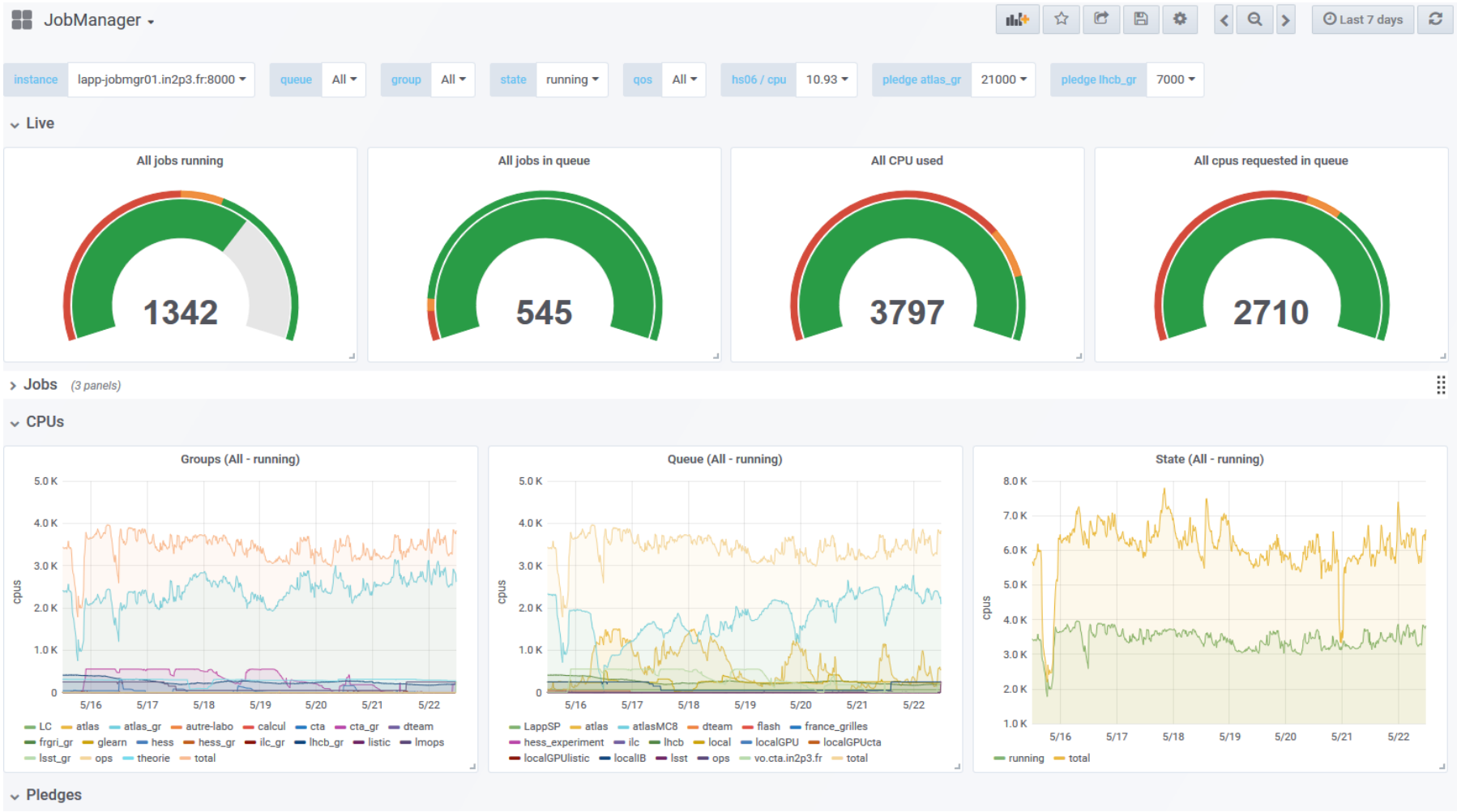
Réseau

- Réseau interne :
  - Cœur de réseau laboratoire
    - Haute disponibilité entre les deux salles (2016-2017)
  - Cœur de réseau MUST :
    - Responsable du routage du réseau MUST
    - 48 ports 10 Gb/s + 4x40 Gb/s
    - Secondé par l'ancien cœur de réseau (2 * 24 ports)
  - Gestion d'un réseau hétérogène :
    - Marques de matériel (Extreme Networks, Cisco, HP Procurve/H3C, Fujitsu)
    - Niveau de fonctionnalités réduites sur les anciens châssis de lame pour les WN
  - Actions 2018/2019 :
    - 2018: Mise en place de réseaux dédiés (DPM, WN, services) :
      - Simplification des ACL extérieures par regroupement
      - Préouverture des services sans solliciter le CCIN2P3
    - 2019 : Renforcement des liaisons optiques entre les 2 salles
    - 2019 : Doublement du cœur de réseau MUST (stack)

Supervision / Accounting

- Un certain nombre d'outils de supervision :
  - Permet d'être proactif
  - Aide à la compréhension des problèmes
- Métrologie :
  - Réseau :
    - Cacti
  - Système :
    - Ganglia
    - Grafana / Prometheus (remplacement de Ganglia à terme)
- Supervision :
  - Nagios
- Accounting MUST (application maison)
- Suivi de la consommation énergétique (application maison)

Supervision / Accounting



Supervision / Accounting

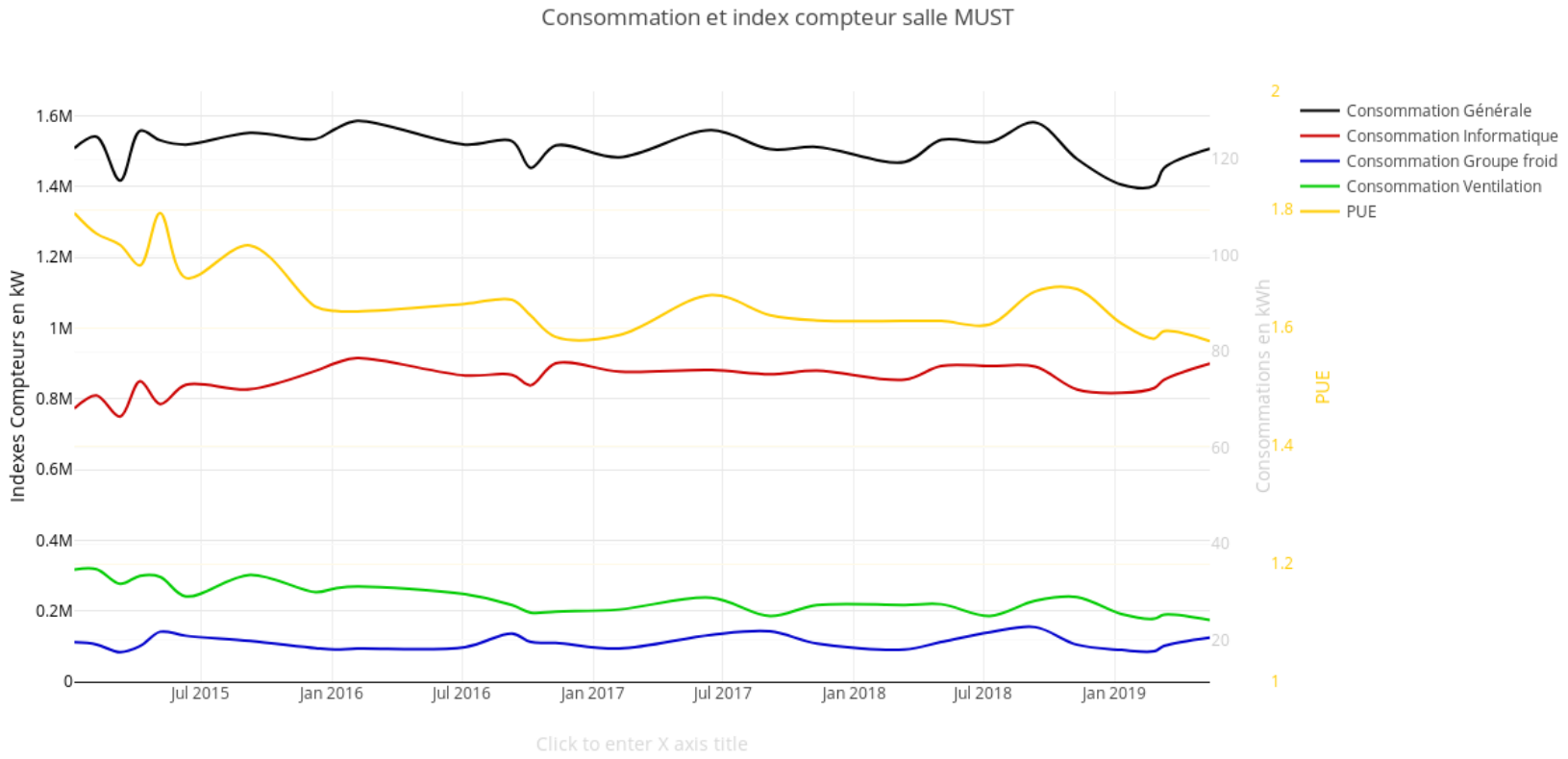
Suivi de la consommation énergétique

Service	Status	Last Check	Duration	Attempt	Status Information
nrpe-cvmfs-atlas	OK	21-05-2019 13:20:00	1d 10h 27m 23s	1/5	SERVICE STATUS: OK: repository revision 49191
nrpe-cvmfs-atlas_condb	OK	21-05-2019 13:27:27	1d 10h 59m 58s	1/5	SERVICE STATUS: OK: repository revision 6427
nrpe-cvmfs-cta	OK	21-05-2019 13:27:40	25d 17h 21m 17s	1/5	SERVICE STATUS: OK: repository revision 296
nrpe-cvmfs-dirac	OK	21-05-2019 13:17:48	40d 0h 40m 36s	1/5	SERVICE STATUS: OK: repository revision 65
nrpe-cvmfs-gridpp	OK	21-05-2019 13:17:49	7d 13h 42m 33s	1/5	SERVICE STATUS: OK: repository revision 14
nrpe-cvmfs-lhcb	OK	21-05-2019 13:18:00	15d 7h 46m 33s	1/5	SERVICE STATUS: OK: repository revision 67481
nrpe-cvmfs-lhcb_condb	OK	21-05-2019 13:18:08	40d 0h 40m 18s	1/5	SERVICE STATUS: OK: repository revision 3580
nrpe-cvmfs-lsst	OK	21-05-2019 13:18:17	25d 18h 23m 8s	1/5	SERVICE STATUS: OK: repository revision 35
nrpe-cvmfs-sft	OK	21-05-2019 13:18:21	24d 19h 12m 54s	1/5	SERVICE STATUS: OK: repository revision 14387
nrpe-disks_space	OK	21-05-2019 13:27:11	25d 19h 2m 55s	1/5	DISK OK - free space: / 10936 MB (76.89% inode=86%): /boot
nrpe-gpfs	OK	21-05-2019 13:27:37	6d 11h 32m 46s	1/5	GPFS OK
nrpe-jobs_escaped	OK	21-05-2019 13:27:21	6d 11h 29m 57s	1/3	JOBS ESCAPED OK: 0 process found.
nrpe-memory	OK	21-05-2019 13:27:24	5d 7h 32m 53s	1/5	MEMORY OK: 21.10% used (13523/64105 MB)
nrpe-mount	OK	21-05-2019 13:27:37	7d 7h 41m 44s	1/5	MOUNT OK: /mustfs/SOFTWARE.
nrpe-ntp_time	OK	21-05-2019 13:27:45	7d 7h 38m 42s	1/5	NTP OK: Offset -0.0007013082504 secs
nrpe-quattor	CRITICAL	21-05-2019 13:17:49	2d 1h 47m 1s	5/5	QUATTOR CRITICAL: 488 errors (487 chkconfig, 1 grub), 2 wa
nrpe-services	OK	21-05-2019 13:23:03	14d 20h 57m 56s	1/5	SERVICE OK: gmond, sssd, pbs_mom currently running.
nrpe-status	OK	21-05-2019 13:23:10	7d 7h 39m 5s	1/4	NRPE v3.2.1

Pour résumer : CVMFS, montage, utilisation mémoire, Quattor, processus échapés…

Supervision / Accounting

Suivi de la consommation énergétique



Projets en cours

- Par l'équipe technique :
  - R&D sur DPM avec le LPSC (FR-ALPES)
  - Remplacement de GPFS par CEPHFS (présentation à venir)
  - Remplacement de Quattor par Foreman / Puppet (long cours)

- En cours avec des "dépendances extérieures" :
  - 2019 : feuille de route pour l'évolution infrastructure salle Mécatronique en collaboration avec l'université
    - Sécurisation (électrique, climatisation…)
    - Urbanisation pour l'amélioration du PUE
  - Sécurisation et augmentation de la bande passante réseau WAN

Projets futurs

- Évolution des services de calculs :
  - JobManager : Torque / MAUI ⇒ HT Condor, SLURM ?
  - CE : Arc-CE ou HT Condor CE ?
- Nouveau site internet pour la présentation du mésocentre et la documentation utilisateur
- Nouvelle interface d'utilisation des serveurs de calculs interactifs (Jupyterhub / notebook)
  - Notebook « démo de soumission d'un job » sur le mésocentre.