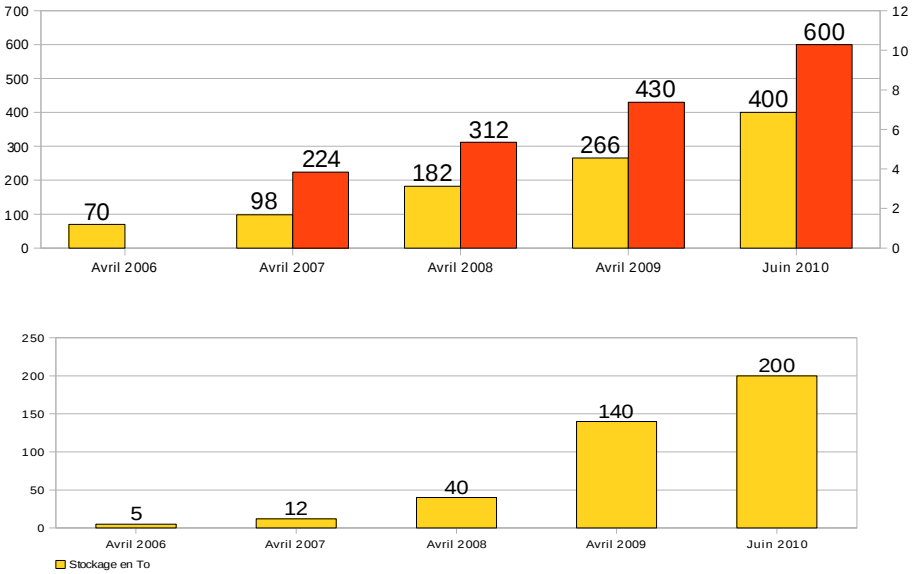
Support de la VO Alice à SUBATECH

- Présentation rapide du site
- Architecture services grille
- Résultats et Utilisation des Ressources
- Principaux travaux 2009-2010
- Questions à l'étude
- Projets et perspectives

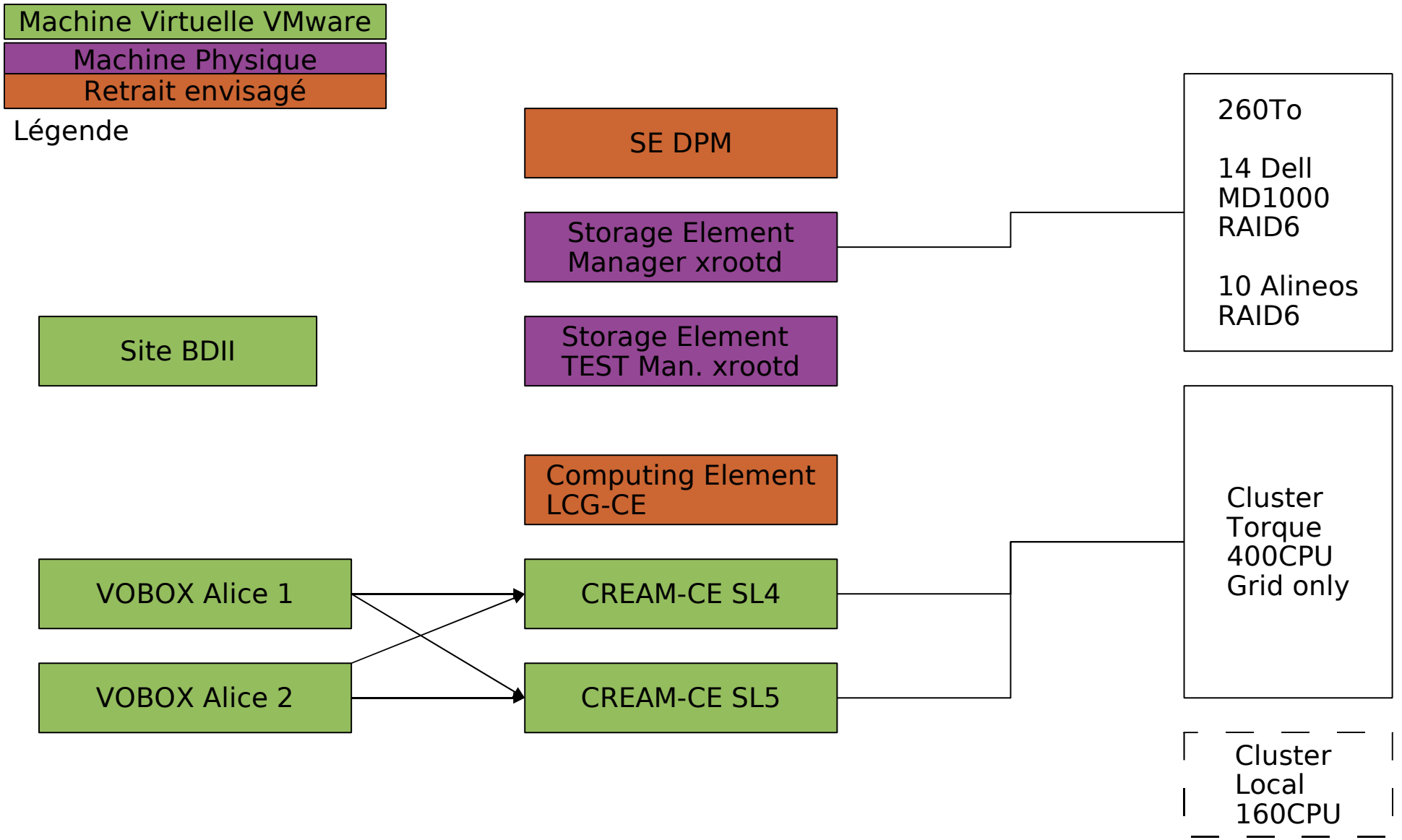
Présentation du site Subatech

- Une seule VO LHC : Alice
- Ressources Humaines : 1.4 ETP (permanents CNRS)
- Financement sur fonds propres,région P.Loire et LCG-France
- Technologies :
  - Virtualisation VMWare (CE, CREAM, VOBOX,sBDII)
  - Quattor (WN,SiteBDII,xrootd)
  - Nagios local
- Ressources 2010 (pledged)
  - 2400 SPEC-HEP06 (600KSI2K)
  - 200To Disque

(*) ETP = Equivalent Temps Plein



Architecture Actuelle (Juin 2010)



Disponibilité/Fiabilité

EGEE Availability and Reliability Report for VO OPS

Region Summary - Sorted by Name — **April 2010**

Data from SAM and Gridview

https://twiki.cern.ch/twiki/pub/LCG/GridView/Gridview_Service_Availability_Computation.pdf

Availability = Uptime / (Total time - Time_status_was_UNKNOWN)
Reliability = Uptime / (Total time - Scheduled Downtime - Time_status_was_UNKNOWN)
KSI2K : Installed capacity of the site measured in kilo specInt 2000 (KSI2K)
Reliability and Availability for Region - Weighted average of sites in the Region (supporting this VO) based on installed capacity

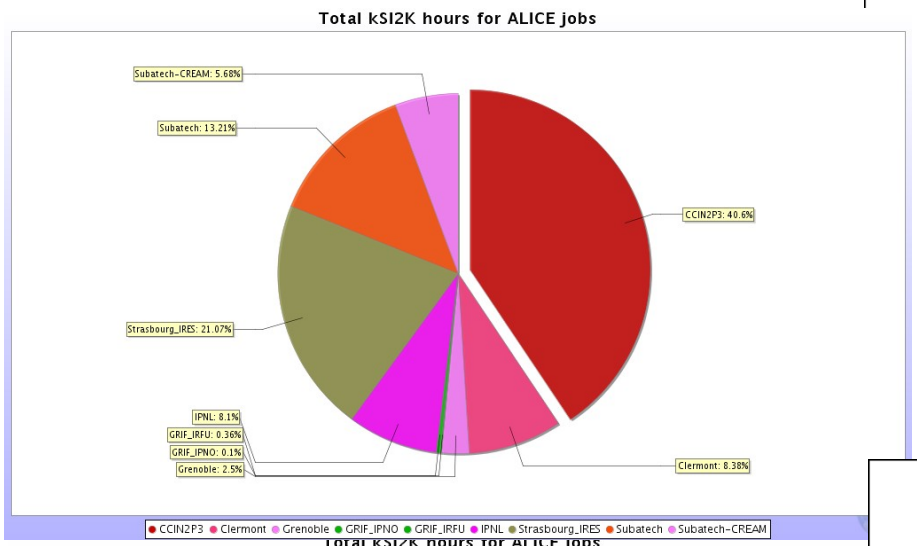
Colour coding : N/A | < 50% | < Target | >= Target

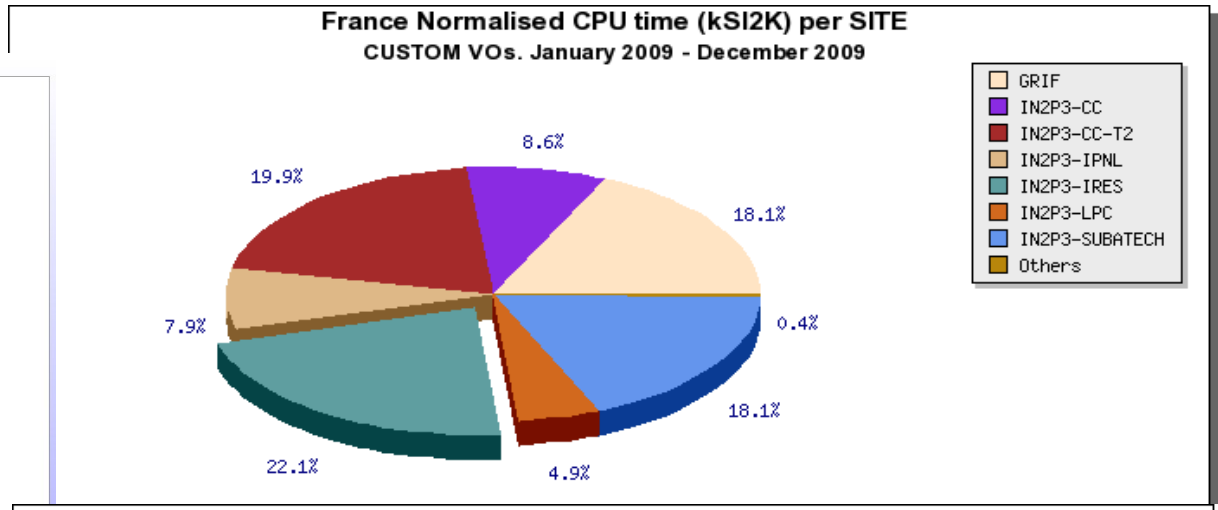
EGEE SLA Availability Target is 70 % and Reliability Target is 75 %

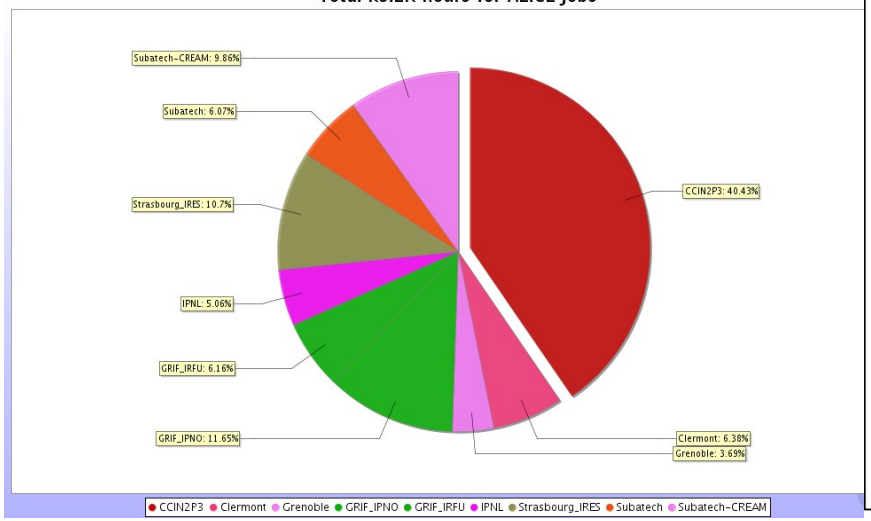
IN2P3-SUBATECH	115	460	880	99 %	99 %	0 %	98 %	98 %	99 %

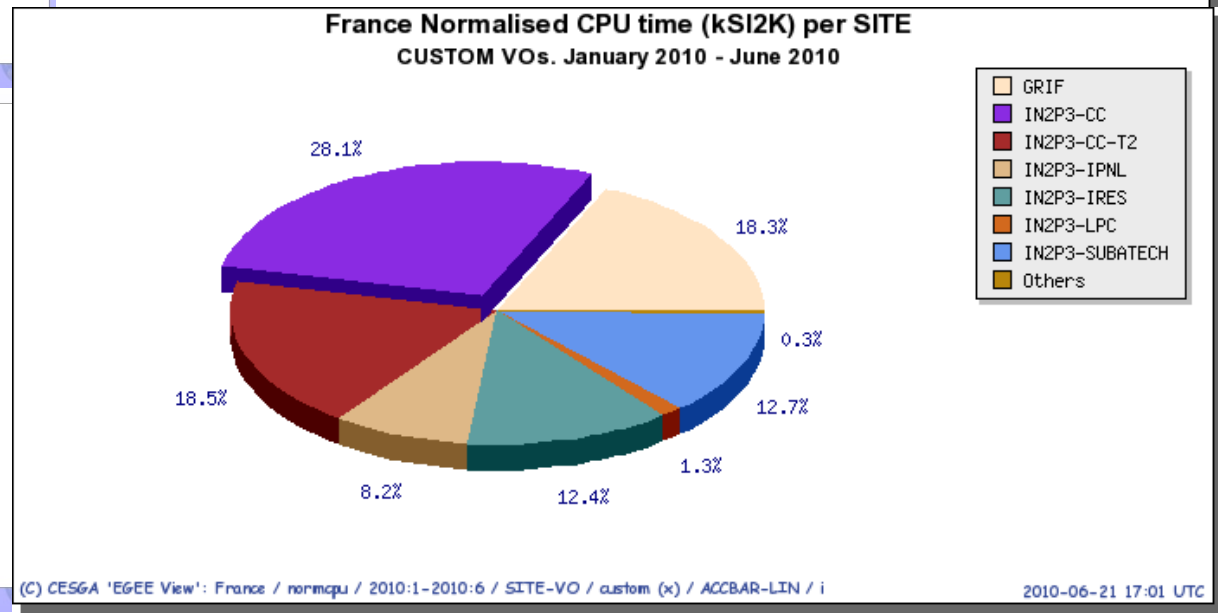
- 2008 : Reliability : 94.7 Availability : 94.7
- 2009 : Reliability : 96.75 Availability : 96.75
- 2010 : Reliability : 98.5 Availability : 98.5

Utilisation des ressources CPU(*)









(*) KSI2K

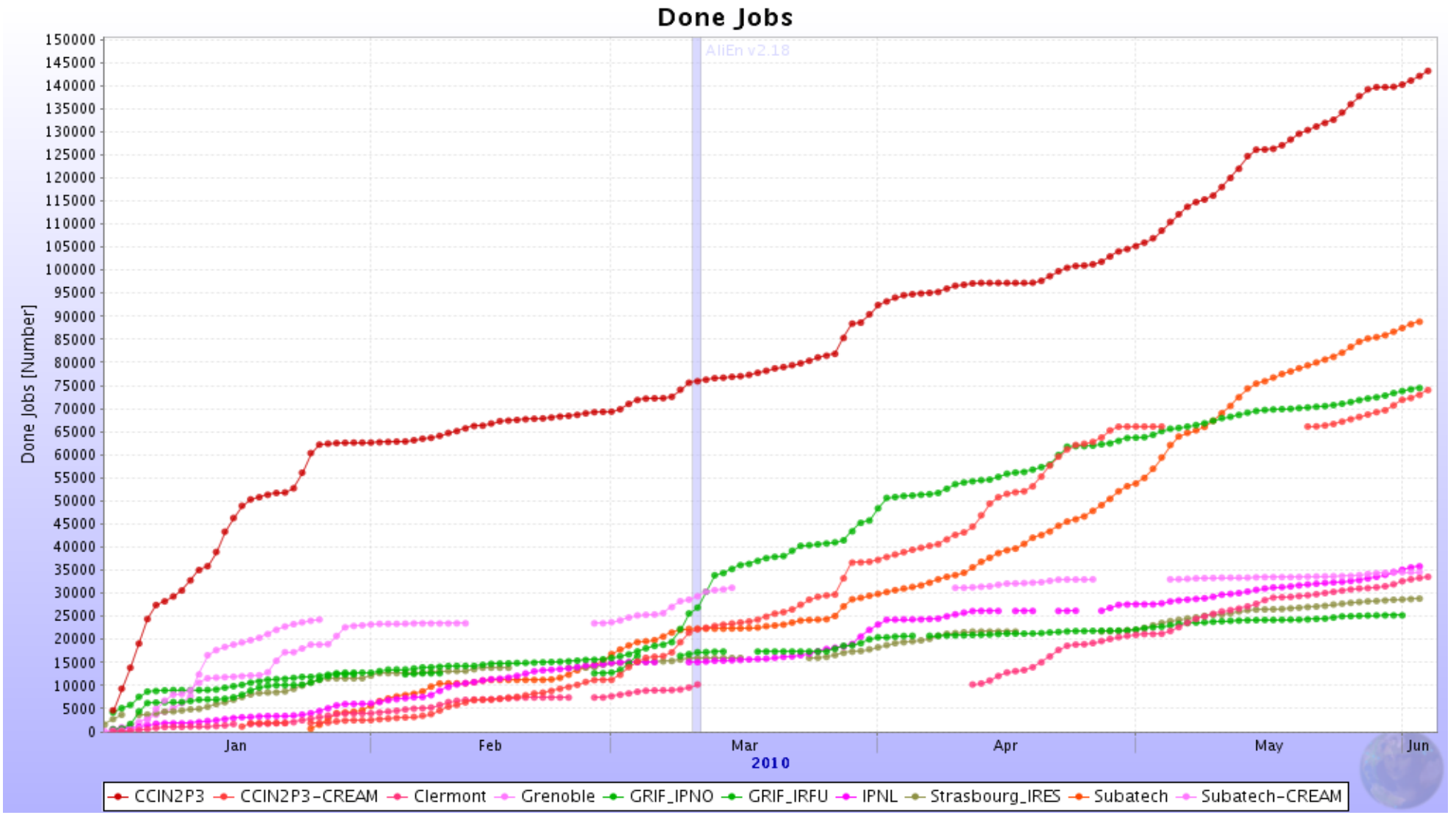
Utilisation des ressources CPU

CPU KSI2K (*)

- En 2009
  - 4% du total Alice
  - 18.77% du CPU fourni par la France (18.1% selon EGEE)
- En 2010
  - 2.51% du total Alice
  - 16% du CPU fourni par la France (12.73% selon EGEE)
- Relativement bonne corrélation MonaLisa/EGEE
  - A noter l'absence de GRIF_IPNO et GRIF_IRFU dans MonaLisa à cause d'un changement de nom

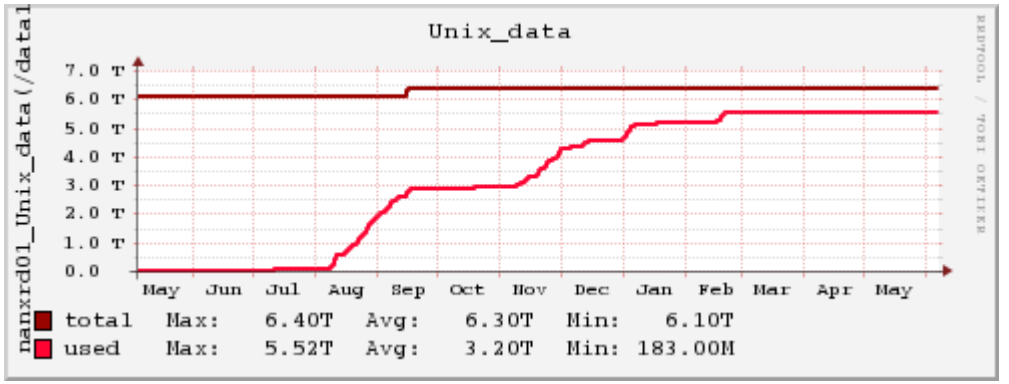
(*) Source Monitoring Alice MonaLisa

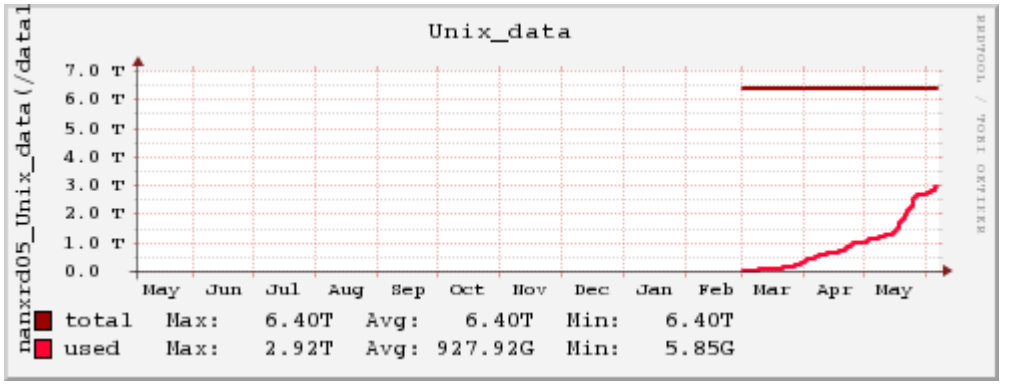
Jobs Sites Français 2010



Utilisation des ressources disque

- 4 serveurs 23.6To (*)
  - Soit ~ 94.4To
  - Installés en Avril 2009
  - Remplis en 6 mois
- 3 serveurs 23.6To
  - Soit ~ 70.8To
  - Installés en Mars 2010
  - Utilisés à 57% (22/06/2010)
- 10 serveurs 10.8To
  - Soit ~ 108To
  - En cours d'installation
- Total : ~ 273To

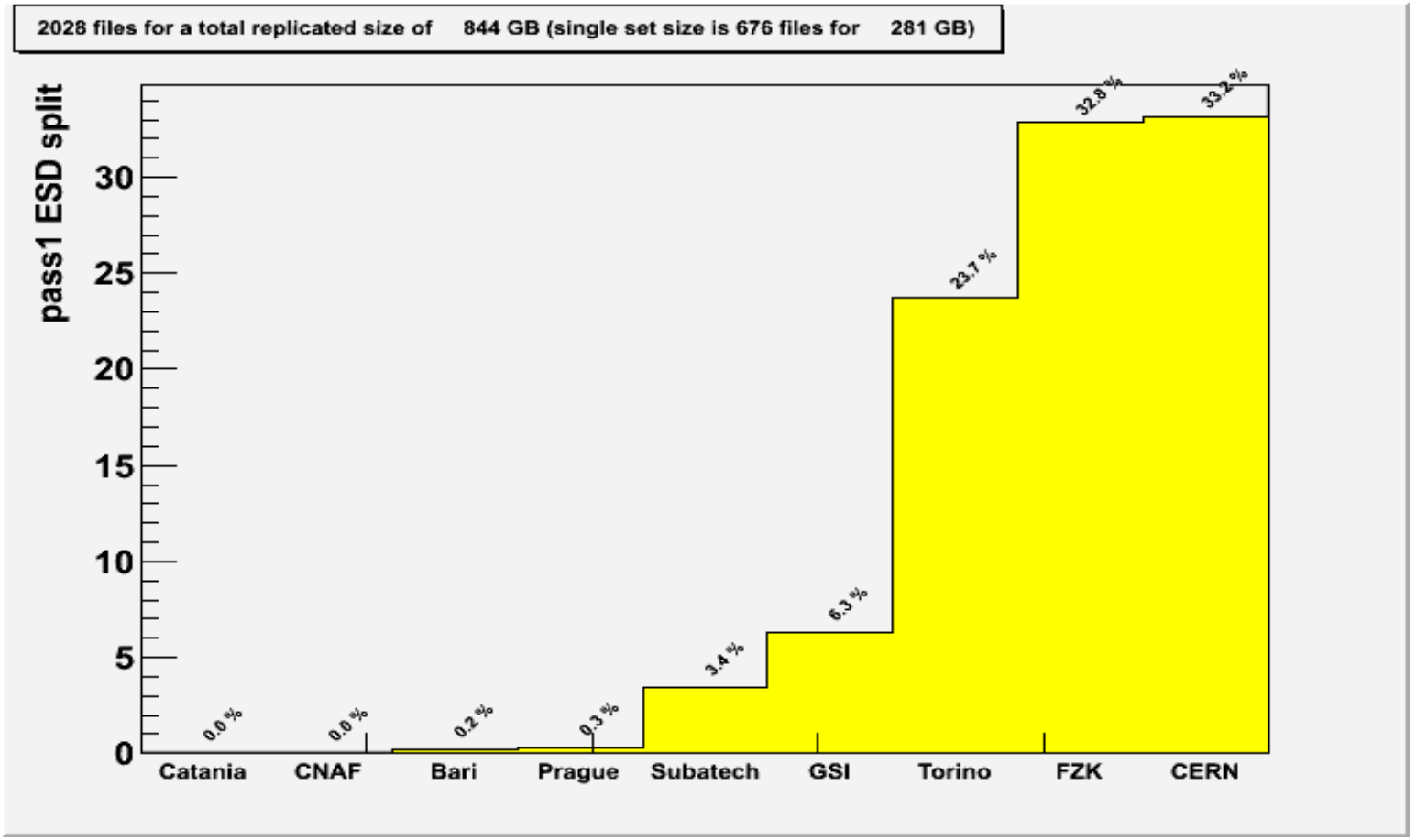




(*) TeraOctets Utiles

Premières Données Novembre 2009

- Subatech parmis les 5 sites sélectionnés pour stocker les premières données reconstruites le 23/11/2009 à 18:35

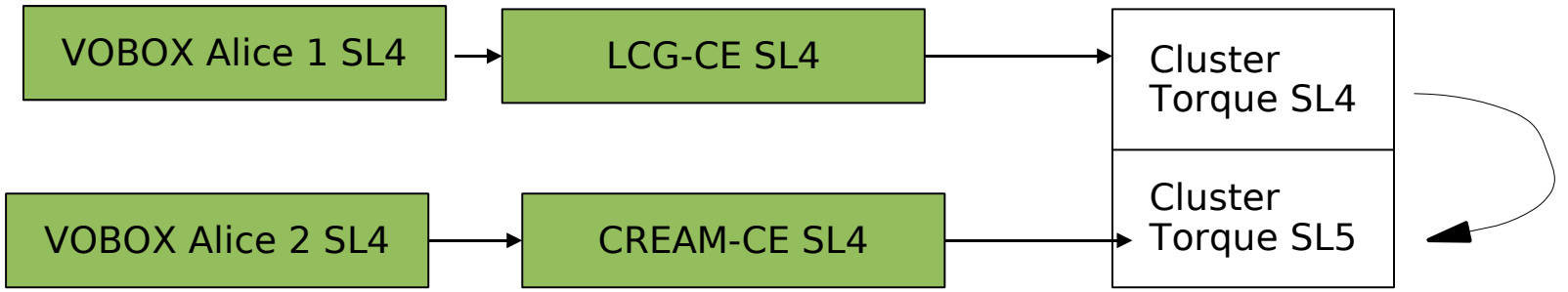


Principaux Travaux 2009-2010

- Migration SL5
- Installation et gestion stockage xrootd natif
- CREAM-CE
- Publication correcte des ressources CPU

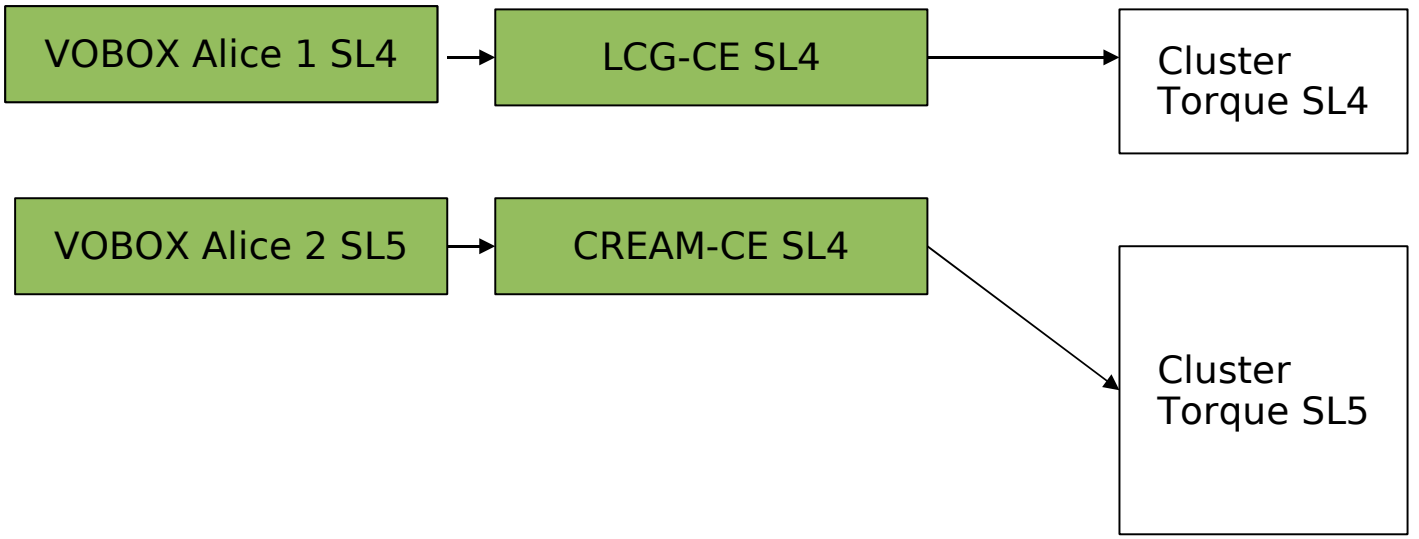
Migration SL5 : Phase 1

- Création d'un nouveau cluster SL5 + CREAM-CE SL4
- Seconde VOBOX SL4 soumet au CREAM-CE
- Passage progressif des workers de SL4 vers SL5



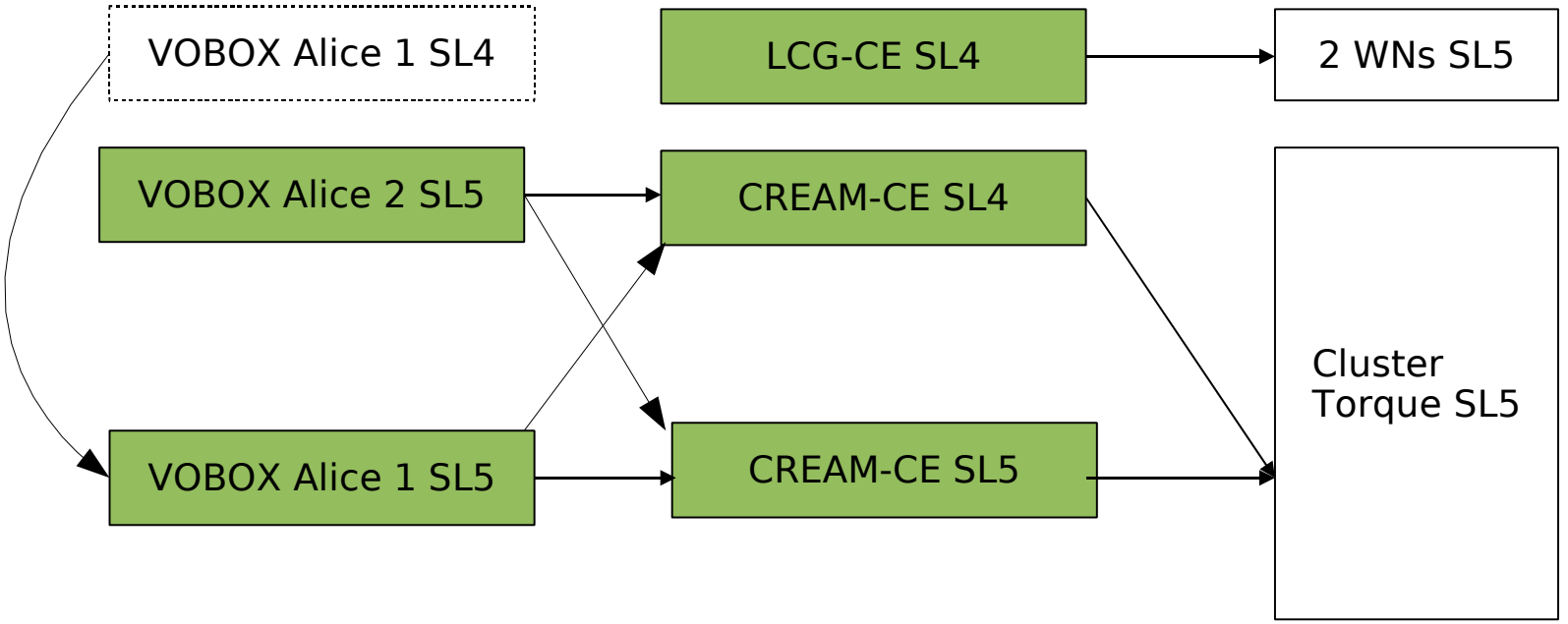
Migration SL5 : Phase 2

- Reinstallation VOBOX-2 en SL5 dès certification
- La majorité des workers est en SL5



Migration SL5 : Phase 3

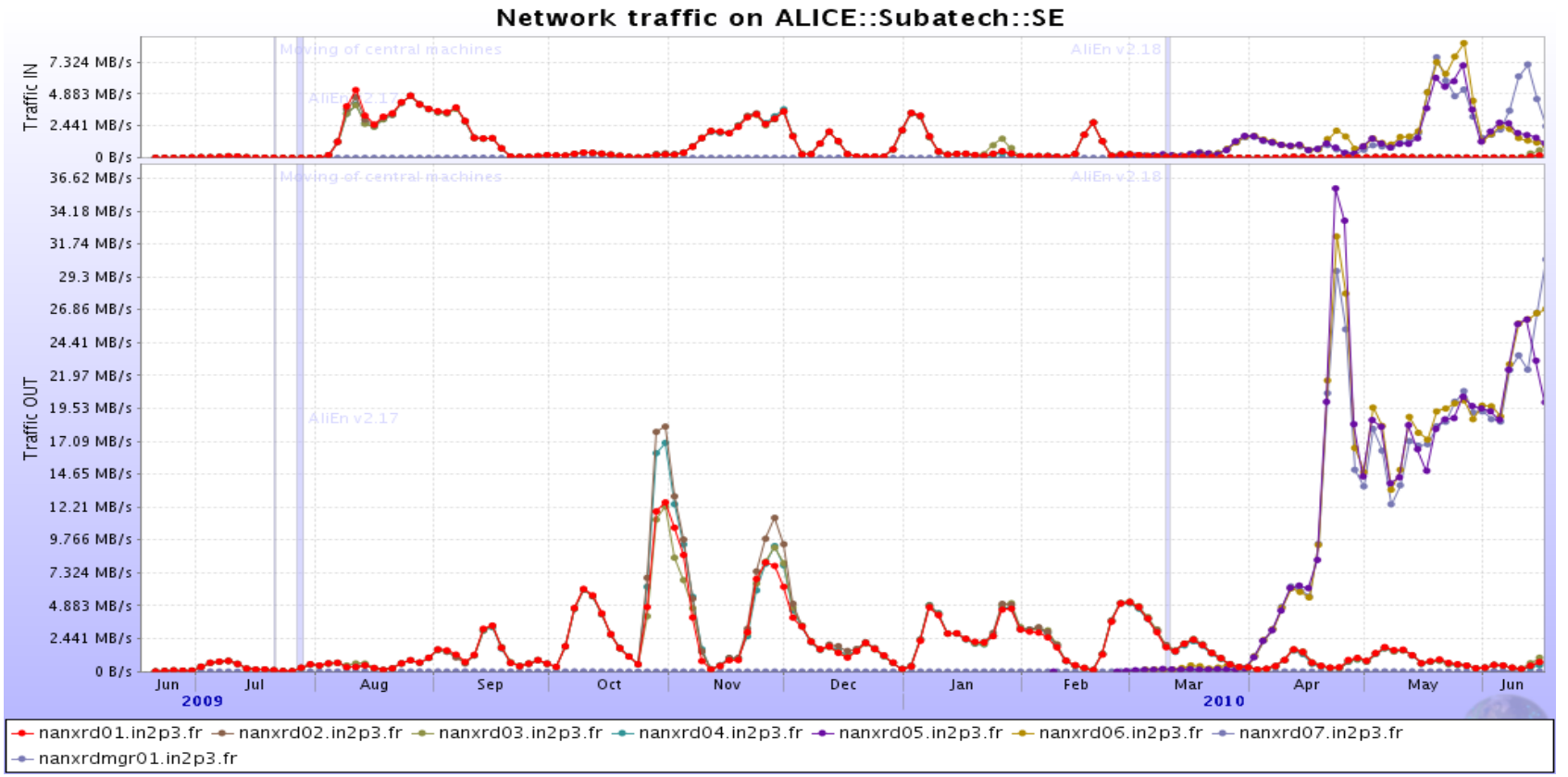
- LCG-CE : Cluster minimal 2 noeuds réinstallés en SL5
- LCG-CE n'est plus utilisé par Alice (tests SAM uniq.)
- Réinstallation VOBOX-1 en SL5 et CREAM SL5
- Soumissions croisée sur les 2 CREAM-CE



Xrootd

- Historique
  - Août 2008 Premier cluster xrootd de test
  - Sept.2008 Test des versions et du packaging RPM
  - Avril 2009 Installation de 100To sur 4 serveurs Dell/MD1000
  - Mars 2010 75To supplémentaires sur 3 serveurs Dell/MD1000
- Mises à jour
  - Xrootd-vmss-1.7
  - Xrootd-vmss-1.8 : Juin 2010
- Procédure de mise à jour :
  - Script de création du RPM binaire xrd-rpmer (contribution à l'amélioration de ce script)
  - Installation/configuration via Quattor

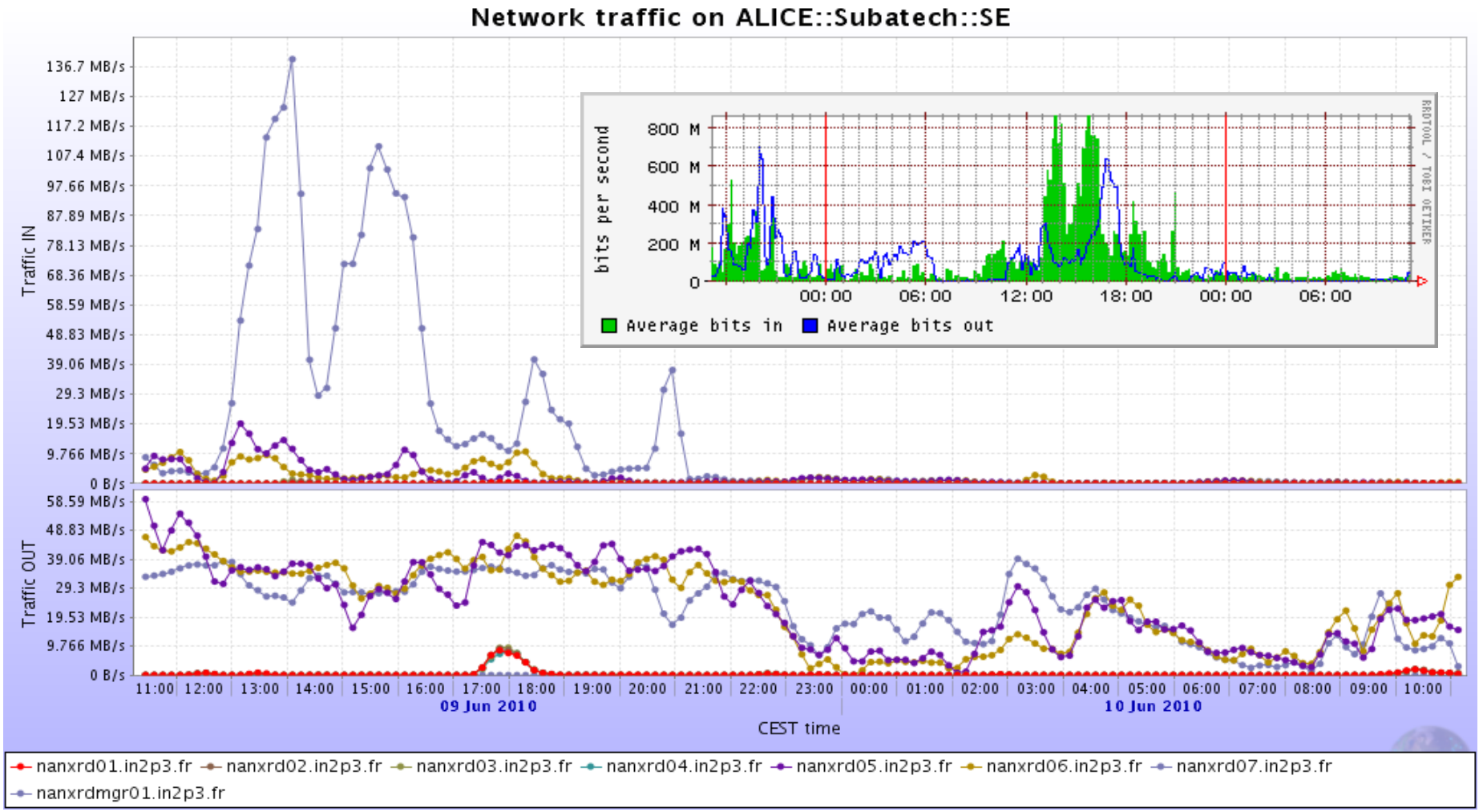
Xrootd



- Maximums mesurés (MonaLisa Mai-Juin 2010) :
  - Sur 4 serveurs : 250Mo/s entrée 930Mo/s sortie (*)
  - Par serveur : 235Mo/s entrée 235Mo/s sortie

(*) 28/10/2009

Xrootd natif : WAN access



CREAM-CE

- Un des premiers CREAM-CE en production
  - Décembre 2008
- Remontée et analyse des problèmes
  - Contribution à l'identification de problèmes,
    voir prés. P.Mendez GDB des 14/10/2009 et 11/11/2009
- Installation d'un CREAM-CE v1.6 en SL5
  - Installé le 13 Avril 2010
  - Mis en production le 26 Mai 2010 (après upgrade majeur)
  - Partage du même batch server entre les deux CREAM-CE
  - Utilisation du “old” BLAH Parser sur le serveur de batch
  - Le BLAH parser est dupliqué pour servir les 2 CREAM-CE

Publication des ressources CPU

- Suite modification APEL (glite-apel-core 2.0.11-0)
  - La valeur de référence pour APEL passe de GlueHostBenchmarkSI00 (valeur moyenne pour un cluster hétérogène) à GlueCPUscalingReferenceSI00 (puissance du CPU le plus faible)
  - Pour ne pas sous-estimer les ressources fournies, il faut utiliser le mécanisme de scaling factor de PBS
  - Testé et mis en oeuvre :

CPU (core)	KSI2K/core	Facteur
Clovertown 2.33GHz 5345	1650	1
Clovertown 2.50GHz E5420	1975	1.19
Nehalem E5520 (R410)	2950	1.78

Référence :
“Job Matching” Stephen Burke au GDB du 9 Juin :
http://indico.cern.ch/conferenceDisplay.py?confId=72057

Questions à l'étude

- Performance stockage et bottlenecks
  - Niveau worker node : 1Gbit/s pour 8 jobs
  - Niveau réseau entre worker nodes et serveurs stockage
  - Niveau serveur stockage : 2Gbits/s pour 50 worker nodes
  - Ex: 50x1Gb/s = 50Gb/s => 25 serveurs à 2Gb/s
  - Quelle est la configuration optimale ?
- Redondance
  - La disponibilité de certains services est essentielle
  - Introduire de la redondance ?
    - Vobox : Seconde VOBOX (demandée par Alice)
    - CE (prévu dans AliEn : soumission en parallèle)
    - Xrootd manager (fait à FZK)
    - Site BDII (alias DNS compatible avec le monitoring ?)

Rappels sur les actions Alice en cours

- Pour les sites :
  - Mettre à jour xrootd en version vmss-1.8
  - Fournir une seconde vobox
  - Publier les voboxes dans le BDII et la GOCDB
  - Avec CREAM 1.6, utilisation (uniquement à des fins de debugging) du serveur GridFTP sur le CREAM-CE. Le serveur GridFTP sur la vobox n'est plus nécessaire.
- A l'étude au niveau de la VO Alice
  - Passage des KSI2K au HEP-SPEC06 (retardé)
  - Torrent pour distribuer le software Alice
  - Publication BDII pour le stockage xrootd
  - glExec

Perspectives

- Réseau 10Gbits/s et reorganisation du réseau
  - Séparation noeud de grille du réseau du laboratoire
  - Déplacement progressif des services
- Glexec Argus et SCAS
  - Homogénisation du mapping des DN vers les comptes
- Analyse
  - Un équipement d'analyse interactive est le complément indispensable de la grille pour nos chercheurs
  - L'utilisation de la grille pour les activités d'analyse directement par les utilisateurs et la maîtrise de ces activités est maintenant le défi important
  - Voir la présentation de Laurent Aphecetche

Boite à Outils pour un site Alice

- Documentation :
  http://alien.cern.ch/twiki/bin/view/AliEn/Home
- Monitoring MonaLisa :
  http://pcalimonitor.cern.ch
- Alice LCG Task Force Meeting
  http://indico.cern.ch/categoryDisplay.py?categId=3l151
- WLCG Daily Operations Meeting
  https://twiki.cern.ch/twiki/bin/view/LCG/WLCGOperationsMeetings
- Alice Dashboard :
  http://dashboard.cern.ch/alice/index.html
- Tests Nagios spécifiques Alice :
  https://sam-alice.cern.ch/nagios/

BackUp Slides

A propos d'Accounting

2009

Site	CPUT EGEE	CPUT Alice	Ecart %	KSI2K EGEE	KSI2K/h	KSI2K Alice	KSI2K/h
CCIN2P3	517677			927611			
CCIN2P3-T2	1256428			2151031			
CCIN2P3 Total	1775105	1754000	1.2	4078642	1.73	3058000	1.74
GRIF_IPNO		8007				12910	
GRIF_IRFU		17180				21750	
GRIF Total	1119401	25187	191.2	1949298	1.74	34660	1.37
LPC Clermont	295273	432000	37.6	531167	1.79	607800	1.4
IPHC Strasbourg	980256	881000	10.67	2382021	2.4	1477000	1.67
IPN Lyon	565026	383500	38.28	850118	1.5	585200	1.5
LPSC	115386	106000	8.48	43847	0.38	181000	1.7
Subatech		606000				949600	
Subatech-CREAM		247700				416900	
Subatech Total	929403	853700	8.49	1952532	2.1	1366500	1.6

a) Ecart entre CPU time EGEE et Alice : de 1.2% à 38%
b) Rapport KSI2K/hour sur EGEE : de 0.38 à 2.4 !

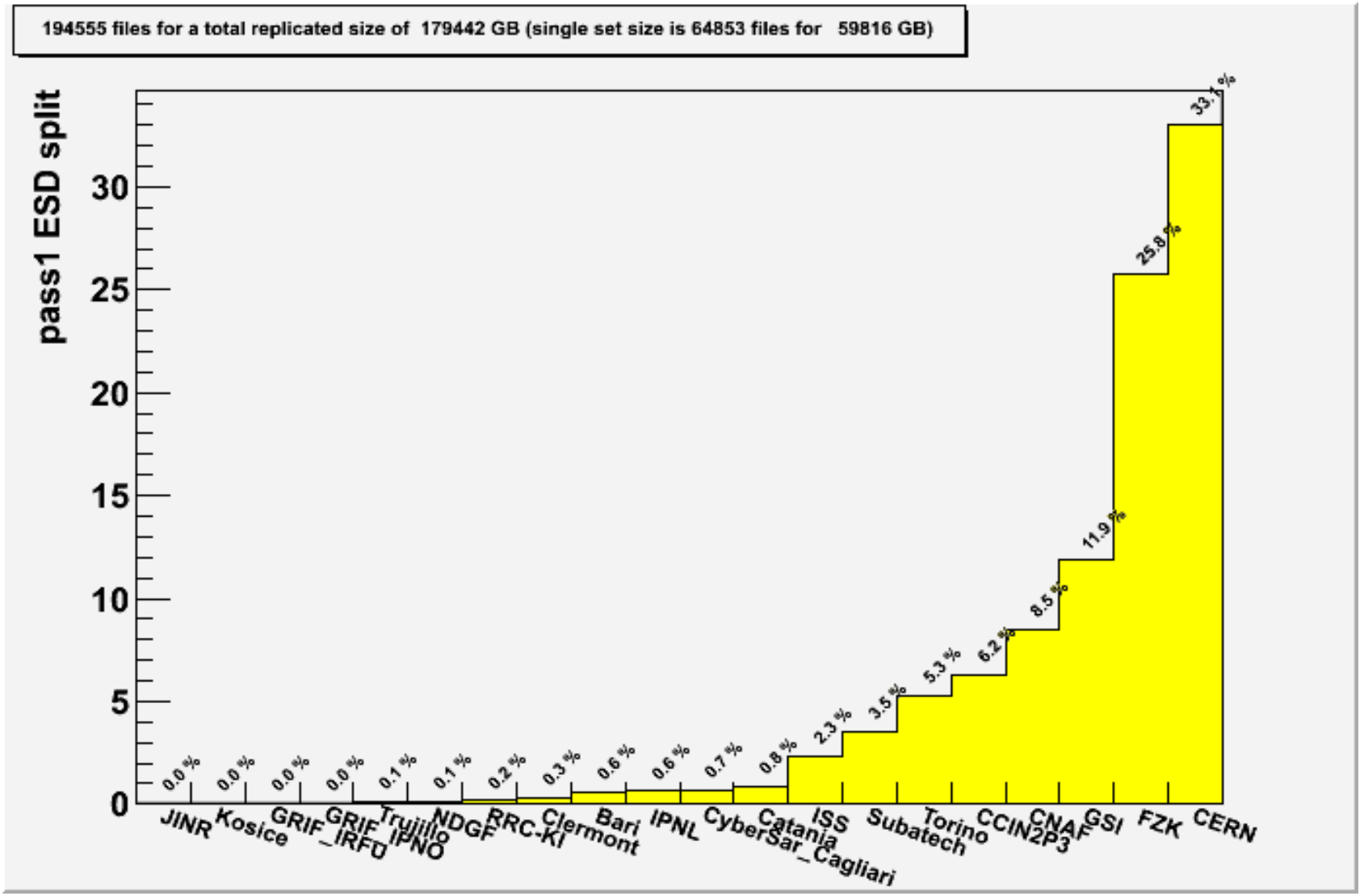
A propos d'Accounting

2010

Site	CPUT EGEE	CPUT Alice	Ecart %	KSI2K EGEE	KSI2K/h	KSI2K Alice	KSI2K/h
CCIN2P3	527016			1110591	2.11		
CCIN2P3-T2	356307			748654	2.1		
CCIN2P3 Total	883323	641200	31.76	1859245	2.1	749600	1.17
GRIF_IPNO		159800				217800	1.36
GRIF_IRFU		138400				116000	0.84
GRIF Total	346683	298200	15.04	715299	2.06	333800	1.12
LPC Clermont	27000	92080	109.3	53999	2	116500	1.27
IPHC Strasbourg	138595	185200	28.79	336786	2.43	196500	1.06
IPN Lyon	170756	93230	58.73	306848	1.8	91410	0.98
LPSC	33994	42030	21.14	12916	0.38	65140	1.55
Subatech		106100				107900	1.02
Subatech-CREAM		124200				190000	1.53
Subatech Total	275176	232300	16.9	480829	1.75	197900	0.85

a) Ecart entre CPU time EGEE et Alice : de 15% à 109%
b) Rapport KSI2K/hour sur EGEE : de 0.38 à 2.4 (les mêmes qu'en 2009)

Données : LHC10c Mai 2010



Est-ce que mon site marche bien ?

- La VOBOX
- Le comportement des jobs
- Le stockage

La Vobox

- Utilise un certificat proxy renouvelé à partir de myproxy.cern.ch pour soumettre des JobAgents (JAs) à un ou plusieurs CEs
  - Certificat proxy valide (y compris sur myproxy.cern.ch)
  - Service de renouvellement fonctionnel
- Soumission des jobs : AliEn-CE
  - Surveiller le log : alien-logs/CE.log
- Synchronisation de la Software Area : Packman
  - Surveiller le log : alien-logs/Packman.log
- Monitoring : MonaLisa

Comportement des Jobs

- Production MonteCarlo
  - Facile à caractériser : massive et CPU intensive
- Désormais l'analyse utilisateur est majoritaire
  - Arrivent par vagues
  - Grande variété de comportements
  - Plus d'I/O que de CPU, efficacité faible
  - Parfois pathologique (épuisement de la RAM)
- => Beaucoup plus difficile à surveiller, rend les comparaison inter-sites peu pertientes

Performances du stockage

- Le stockage est très sollicité par les travaux d'analyse
- La majeure partie du trafic est interne au site
- Détection des problèmes et performances :
  - Sur MonaLisa : traffic, unix load, sockets
  - Sondes locales (nagios) : unix load,
- Comparaison des mesures :
  - Nagios, ML et MRTG, autres ?