GRIF et le support de la VO ALICE

diarra@ipno.in2p3.fr

Le Tier-2 GRIF

- GRIF : Grille au service de la Recherche en Ile-de-France
  - Fédération de 6 sites en région parisienne

irfu institut de recherche sur les lois fondamentales de l'univers saclay

CEA, Saclay
1750 cores , 2600 job slots
1500TB

IPN INSTITUT DE PHYSIQUE NUCLÉAIRE ORSAY

Univ Paris-Sud, Orsay
1000 cores, 1160 job slots
240 TB

LAL LABORATOIRE DE L'ACCÉLÉRATEUR LINÉAIRE

Univ Paris-Sud, Orsay
1800 cores, 2100 job slots
240 TB

LPNHE PARIS

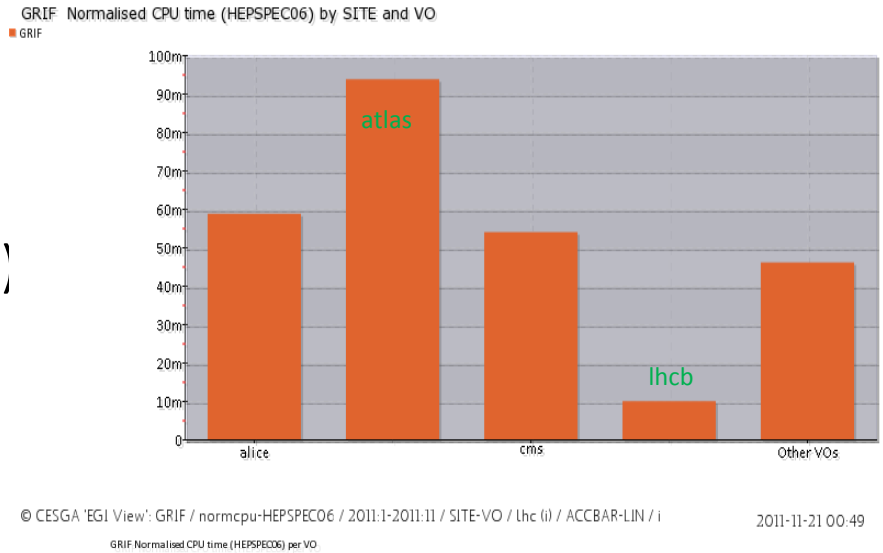
Univ Paris 7, UPMC, Paris
552 cores, 800 job slots
680TB

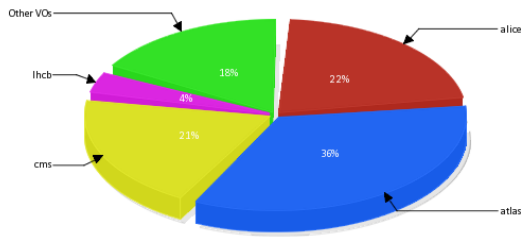
Ecole Polytechnique, Palaiseau
850 cores, 1500 job slots
750 TB

Univ Paris 7, paris
350 cores, 350 job slots
90 TB

Le Tier-2 GRIF les ressources

- GRIF : plusieurs sites vus comme un site grille unique
  - Une entrée dans le BDII
- 12 clusters dont 3 dédiés à MPI
- ~8500 job slots
- >3PB sous DPM (7 serveurs DPM)
- Services grille: CE, SRM, BDII, WMS, LFC, VOMS, MyProxy
- Plus de 40 VO supportés
- 4 VO LHC
- Autres VO : biomed, ILC, babar, dzero, fusion, …





Le Tier-2 GRIF Infrastructure/Opérations

- 6 sites interconnectés via un réseau privé (GRIFOPN) basé sur un VLAN
- Sur chaque LAN :
  - remontée des racks en 10Gb
  - serveurs de disques connectés en 10Gb ou via des liens agrégés
- Trafic GRIF sur un VLAN vers le CC
  - LAL, IPNO reliés à LHCONE (les autres suivront)
- Monitoring Nagios centralisé (alertes mail)
- CDB Quattor centralisée avec un serveur AII par site
- Des services en haute disponibilité (BDII pour l'instant) :
  - Fail-over du service sur un autre site en cas de problèmes
- Etroite collaboration entre sites autour des opérations
  - Réunion mensuelle du CT
  - Tours de garde hebdomadaire (rotation planifiée entre les sysadmin)
  - Wiki et système de ticket
  - Délégation de droits entre administrateurs (gsissh + sudo)
  - Deploy des arrata par une personne une fois par mois pour tout GRIF sauf si CVE critique
    - Rotation entre 4/5 personnes

ALICE et la grille en bref

- Le calcul de ALICE repose sur AliEn (Alice Environment)
- Tous les jobs grille sont gérés via une Central TaskQueue
- Pour en savoir plus sur l'analyse dans ALICE, voir la présentation de L. Aphecetche :
- http://indico.in2p3.fr/conferenceOtherViews.py?view=standard&confId=4340
- ALICE n'utilise pas le LFC LCG. Il a son propre FileCatalog
- ALICE n'utilise (presque pas) le WMS : la soumission se fait directement sur les CE
- Le stockage est basé essentiellement sur xrootd
- L'analyse peut se faire facilement avec pratiquement le même code en local, sur la grille LCG ou sur PROOF
- Les fichiers de sortie sont stockés en 3 replicas en principe (modifiable)
- L'utilisateur ne choisit pas les SE pour les fichiers de sortie. Le choix se fait automatiquement en fonction des résultats du monitoring du réseaux et des SEs.
- Les logiciels utilisés par les jobs sont : Root, AliRoot, Geant3, …
- Le Monitoring est centralisé et repose sur Monalisa (http://alimonitor.cern.ch)

ALICE : AliEn et la VOBOX

- Pour supporter ALICE, il faut une VOBOX
  - C'est une passerelle entre la grille AliEn et la grille LCG
- Les services AliEn (Alice Environment) tournent sur la VOBOX : AlienCE, AlienSE(inutilisé), PackMan (Package Manager), ClusterMonitor, Monalisa, …
- Le nom de la VOBOX, du site et des SEs doivent être déclarés dans une base LDAP d'ALICE
- La VOBOX remonte les informations de monitoring vers un service centralisé de ALICE
- Le AlienCE soumets les JobAgents (JA) sur le CE (LCG ou CREAM) tant que des slots sont libres et tant qu'une limite définie par ALICE n'est pas atteinte

ALICE : AliEn et la VOBOX (suite)

- Les JA récupère de la central TaskQueue du travail effectif à faire
- La priorité entre jobs est gérée au niveau de la central taskQueue d'Alien
- L'installation du software est faite sur la VOBOX dans une software area partagée entre la VOBOX et les WNs
- Le déploiement peut être manuel ou automatique (déclenché par ALICE)
- Les Mises à jour sont faites automatiquement par PackMan
  - Par exemple si un job réclame une version de logiciel non disponible

ALICE : AliEn et la VOBOX (suite)

- La VOBOX soumets les JA 24h/24
  - Il faut un proxy valide en permanence pour soumettre sur le CE (LCG ou CREAM) et pour interagir avec les services AliEn
  - Les services tournent avec les droits d'un utilisateur qui doit avoir le rôle VOMS « production » et/ou « lcgadmin »
  - Son proxy est enregistré sur la VOBOX et renouvelé en continue en utilisant le serveur MyProxy
- Dans GRIF, tous les jobs tournent avec l'identité du même utilisateur
  - C'est l'utilisateur ALICE avec le rôle VOMS lcgadmin
  - C'est sous l'identité de ce utilisateur que tournent les services Alien
  - C'est ce utilisateur qui est propriétaire de la software area

ALICE @GRIF

- ALICE est supporté par deux sites : IPNO et IRFU
- Il y a une VOBOX par site
  - IPNO : depuis fin Juillet 2006
  - IRFU : depuis Décembre 2006
  - Impossible d'avoir une seule VOBOX à cause de la swarea
- Les jobs tournent avec le proxy de quelqu'un de l'expérience (latchezar, lolass, … au choix) plutôt que ceux d'un sysadmin de GRIF
  - On n'est sûr qu'il n'y aura pas de problèmes d'autorisation
- Chaque VOBOX tourne sur une machine physique

ALICE @GRIF : utilisation CPU

- Bonne occupation des CEs par ALICE
  - Tant qu'il y a des jobs à traiter et des slots libres, ALICE soumets
  - ALICE définit automatiquement des limites par site en fonction des ressources CPUs

```
[alis@node21 alienlog]$ more CE.log
…
Nov 22 15:58:28  info    According to the manager, we can queue max 150 and manage
    max 2500
…
Nov 22 15:59:29  info   JobAgents running, waiting: 1463,160
Nov 22 15:59:29  info   (Returning value from BDII)
Nov 22 15:59:29  info   Returning -10 free slots, with 1463 running jobs
```

ALICE @GRIF : utilisation CPU

Contribution honorable de GRIF

L'IPNO ne supporte pas beaucoup de grosses VO et donc ALICE peut exploiter toutes les ressources disponibles

A l'IRFU, le CE est splité entre 3 VO LHC et ALICE est favorisé en ce moment par la sous-utilisation de l'IRFU par les deux autres VO.

68 sites actifs

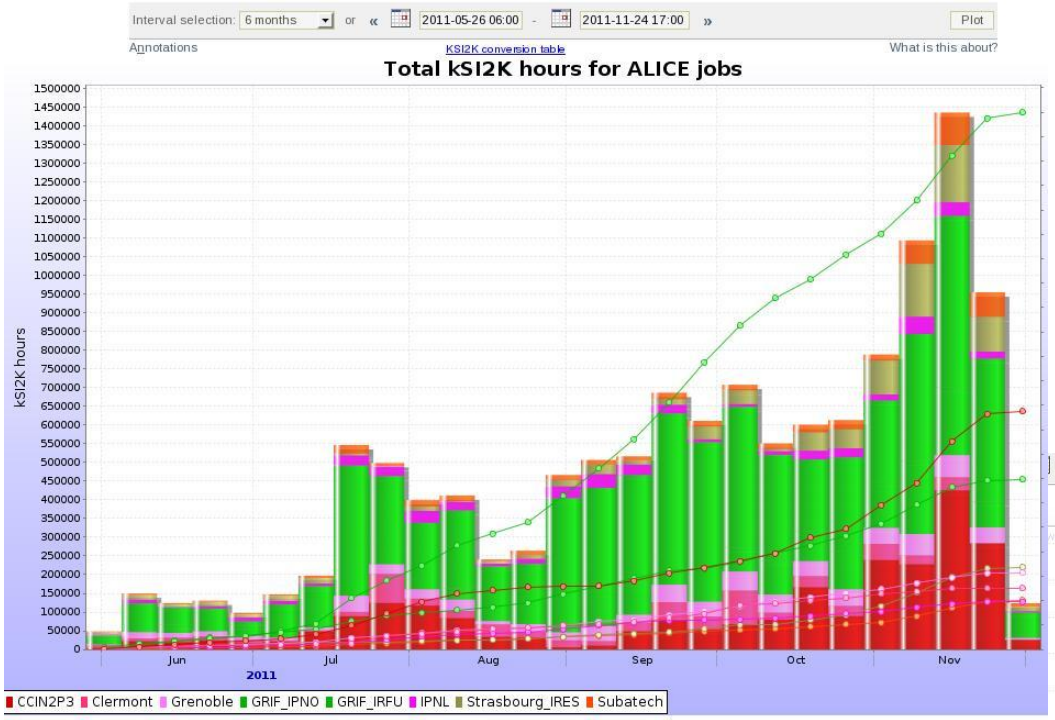
ALICE reports

What is this about?

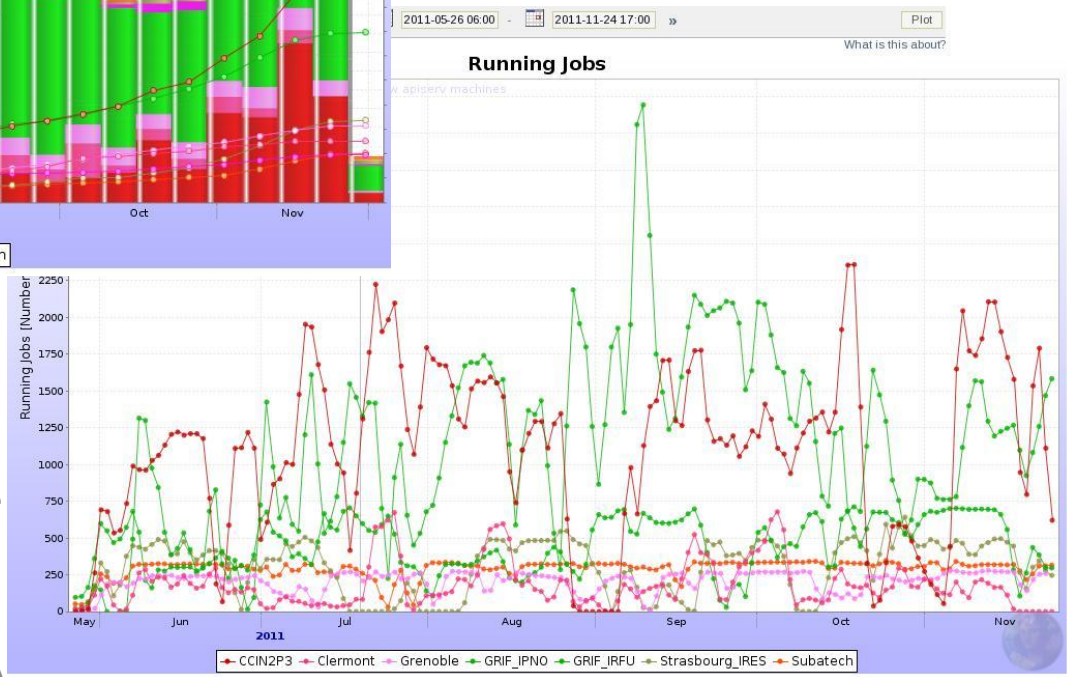
Report on ALICE sites' activity (23.05.2011 - 22.11.2011)

Site	Group	Tier	Pledged KSI2K	Delivered CPU	Delivered Wall	Occupancy Wall/Pledged	Missing KSI2K Pledged - Wall	Efficiency CPU/Wall	Job statistics Assigned	Job statistics Completed ▲	Job statistics Efficiency
22. FZK	Germany	T1	1	2977	6595	659524%	-	45.15%	3295863	2152908	65.32%
12. CERN-CREAM	CERN	T0	-	951.2	1292	-	-	73.61%	4314142	1936813	44.89%
52. NIHAM	Romania	T2	1055	1477	2291	217.2%	-	64.49%	2133688	1585069	74.29%
13. CERN-L	CERN	T0	3500	640.8	970.7	27.73%	2529	66.02%	2867533	1455424	50.76%
15. CNAF	INFN	T1	2750	753.8	1040	37.85%	1709	72.42%	1718751	1165811	67.83%
61. Prague-CREAM	Czech Republic	T2	-	906.8	1644	-	-	55.13%	1521965	948973	62.35%
38. JINR	RDIG	T2	844	501.8	716.6	84.9%	127.4	70.03%	1130666	830998	73.5%
48. LLNL	US	T2	-	230.6	287.6	-	-	80.2%	1270977	696344	54.79%
24. GRIF_IPNO	IN2P3	T2	-	394.6	613.2	-	-	64.35%	799960	562282	70.29%
9. CCIN2P3	IN2P3	T1	5000	552.7	745.8	14.92%	4254	74.11%	762233	461650	60.57%
31. Hiroshima	Japan	T2	250	217.8	352.1	140.8%	-	61.85%	626444	448264	71.56%
69. Subatech	IN2P3	T2	430	110.6	185.7	43.19%	244.3	59.54%	536040	399755	74.58%
71. Torino	INFN	T2	937	423	684.1	73.01%	252.9	61.83%	537465	381485	70.98%
44. Kosice	Slovakia	T2	100	132.8	248.8	248.8%	-	53.38%	509472	368261	72.28%
63. RRC-KI	RDIG	T2	1015	329.5	456.4	44.97%	558.6	72.18%	552201	350952	63.56%
35. ISS	Romania	T2	1097	238.3	405.4	36.96%	691.6	58.79%	538684	349240	64.83%
47. Legnaro	INFN	T2	937	181.1	355.9	37.98%	581.1	50.88%	546035	343331	62.88%
46. LBL	US	T2	4020	620.5	840.9	20.92%	3179	73.79%	551925	317395	57.51%
27. GSI-CREAM	Germany	T2	-	110.2	156.2	-	-	70.54%	453408	307462	67.81%
41. KISTI_GSDC	Republic of Korea	T2	-	32.99	39.47	-	-	83.57%	400809	293666	73.27%
2. Bari	INFN	T2	937	148.1	252.8	26.98%	684.2	58.58%	338196	268455	79.38%
60. Poznan	Poland	T2	330	151.7	233.2	70.67%	96.78	65.03%	420739	267262	63.52%
68. Strasbourg_IRES	IN2P3	T2	700	187.8	257.9	36.84%	442.1	72.83%	359552	261800	72.81%
40. KISTI-CREAM	Republic of Korea	T2	150	76.69	170.6	113.7%	-	44.96%	443349	260749	58.81%
50. Madrid		T2	174	93.84	184.4	106%	-	50.9%	349539	219345	62.75%
14. Clermont	IN2P3	T2	654	143.1	240.5	36.77%	413.5	59.49%	264977	194343	73.34%
29. GSI-SGE	Germany	T2	-	46.3	143.5	-	-	32.26%	357085	183806	51.47%
39. KFKI	Hungary	T2	150	86.8	149.8	99.84%	0.243	57.96%	233960	177908	76.04%
23. Grenoble	IN2P3	T2	-	177.3	288	-	-	61.58%	246608	173174	70.22%
62. RAL	UK	T1	506	666.5	1187	234.7%	-	56.13%	372184	167869	45.1%
25. GRIF_IRFU	IN2P3	T2	-	1225	1785	-	-	68.63%	391799	163526	41.74%
6. Bratislava	Slovakia	T2	-	29.6	41.1	-	-	72.01%	212720	155155	72.94%

ALICE @GRIF : utilisation CPU



Très bonne contribution de GRIF au niveau français



Des vagues à cause de l'activité des autres VO

ALICE@GRIF : le stockage

- Chaque site de GRIF dispose d'un serveur DPM
  - MAIS ALICE recommande l'utilisation de xrootd natif
- Un plugin DPM-xrootd permet d'accéder en xrootd au stockage sous DPM
  - GRIF a donc choisi d'utiliser le plugin (mutualisation des ressources, peu d'effort à faire)
  - Mais pas de support de GSI avec le plugin : l'authentification se fait via les tokens ALICE
- Nov. 2008 : SE GRIF_IPNO::ALICE::DPM avec 16TB
  - Pool DPM alicePool ouvert à toutes les VOs (car pas de support GSI avec le plugin)
  - L'autre pools est dédié aux autres VO et privilégié par DPM pour ces VO
- Janv. 2010 : SE GRIF_IRFU::ALICE::DPM avec 123 TB
  - Utilisation de la nouvelle version du plugin : pool dédié ALICE,
  - choix possible d'une identité pour les créations de fichiers (utilisateur ALICE avec le rôle VOMS lcgadmin)
- Juin 2010 : SE GRIF_IPNO::ALICE::SE : 39 TB en xrootd natif
- Oct. 2011 : IPNO: support exclusif de xrootd natif
  - Abandon pour ALICE de DPM/xrootd et reconversion des disk servers DPM

ALICE@GRIF : le stockage taux d'utilisation

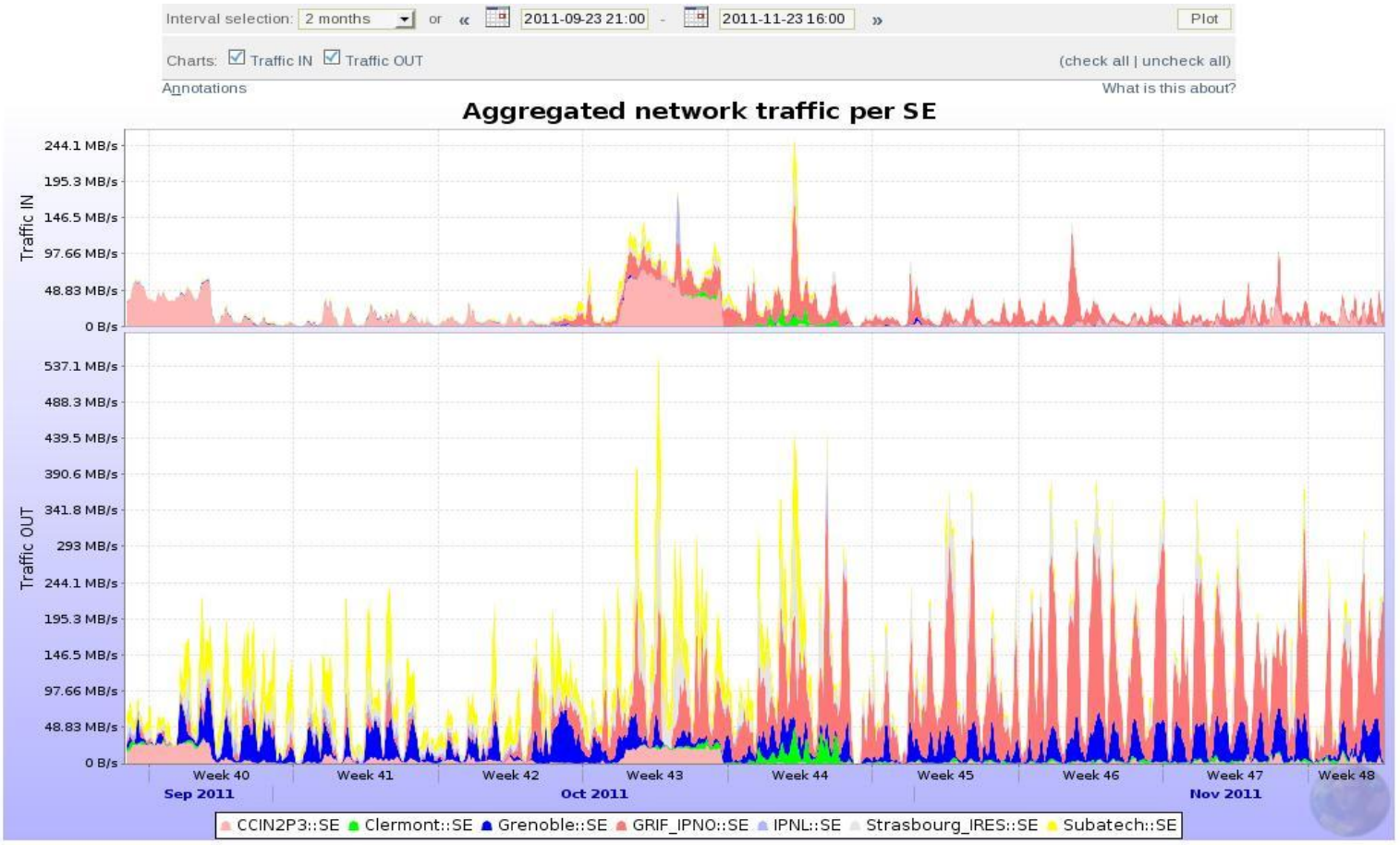
- Bon taux de remplissage des SE (peu d'espace libre)

Disk storage elements

	AliEn SE	Statistics						Xrootd info			
SE Name	AliEn name	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage
1. Bari - SE	ALICE::Bari::SE	893.4 TB	208.5 TB	684.9 TB	23.33%	5,520,284	File	2.14 PB	1.984 PB	159.4 TB	92.72%
2. BITP - SE	ALICE::BITP::SE	24 TB	2.381 TB	21.62 TB	9.921%	70,428	File	26.91 TB	6.634 TB	20.28 TB	24.65%
3. Bratislava - SE	ALICE::Bratislava::SE	112.8 TB	37.67 TB	75.13 TB	33.4%	1,085,507	File	94.59 TB	53.36 TB	41.22 TB	56.42%
4. Catania - SE	ALICE::Catania::SE	132.3 TB	156.3 TB	-	118.2%	3,334,031	File	132.3 TB	60.02 TB	72.27 TB	45.37%
5. CCIN2P3 - SE	ALICE::CCIN2P3::SE	440 TB	170.9 TB	269.1 TB	38.85%	3,822,856	File	-	-	-	-
6. CERN - ALICEDISK	ALICE::CERN::ALICEDISK	2.194 PB	1.101 PB	1.093 PB	50.17%	20,966,792	CASTOR	-	-	-	-
7. CERN - GLOBAL	ALICE::CERN::GLOBAL	-	0	1.863 TB	-	7,083	root	-	-	-	-
8. CERN - SE	ALICE::CERN::SE	20.49 TB	13.38 TB	7.112 TB	65.29%	4,030,917	File	20.46 TB	7.493 TB	12.97 TB	36.62%
9. CERN - SETEST	ALICE::CERN::SETEST	542.9 GB	19.75 GB	523.1 GB	3.639%	2,022	File	-	-	-	-
10. Clermont - SE	ALICE::Clermont::SE	179.9 TB	183.2 TB	-	101.8%	4,165,842	File	179.9 TB	175.8 TB	4.089 TB	97.73%
11. CNAF - SE	ALICE::CNAF::SE	873.3 TB	539.6 TB	333.7 TB	61.79%	10,659,311	File	873.3 TB	583.2 TB	290 TB	66.79%
12. CyberSar_Cagliari - SE	ALICE::CyberSar_Cagliari::SE	145.9 TB	38.94 TB	107 TB	26.69%	1,032,382	File	140 TB	93.44 TB	46.51 TB	66.76%
13. Cyfronet - SE	ALICE::Cyfronet::SE	10 TB	11.72 TB	-	117.2%	523,218	File	9.995 TB	9.805 TB	193.8 GB	98.11%
14. FZK - SE	ALICE::FZK::SE	1.254 PB	890.2 TB	393.8 TB	69.33%	15,936,601	File	1.252 PB	1.13 PB	125.4 TB	90.22%
15. Grenoble - DPM	ALICE::Grenoble::DPM	72 TB	6.215 TB	65.78 TB	8.633%	214,757	SRM	-	-	-	-
16. Grenoble - SE	ALICE::Grenoble::SE	31 TB	28.39 TB	2.614 TB	91.57%	677,406	File	31.47 TB	30.04 TB	1.427 TB	95.47%
17. GRIF_IPNO - SE	ALICE::GRIF_IPNO::SE	223.3 TB	182.3 TB	41.03 TB	81.63%	4,856,144	File	223.3 TB	195.8 TB	27.43 TB	87.71%
18. GRIF_IRFU - DPM	ALICE::GRIF_IRFU::DPM	171 TB	42.86 TB	128.1 TB	25.07%	793,483	SRM	-	-	-	-
47. Strasbourg_IRES - SE	ALICE::Strasbourg_IRES::SE	77.97 TB	67.78 TB	10.19 TB	86.93%	2,160,744	File	78.94 TB	77.27 TB	1.664 TB	97.89%
48. Subatech - SE	ALICE::Subatech::SE	270.3 TB	276.7 TB	-	102.4%	7,486,402	File	270.3 TB	264.9 TB	5.359 TB	98.02%

ALICE@GRIF : le stockage taux d'utilisation

- Des taux de transfert satisfaisants



ALICE : les problèmes rencontrés

- La 1ère VOBOX de l'IPNO rebootait fréquemment : résolu avec changement de hardware
  - La machine réinstallée en WN n'a plus eu le problème
- ALICE ne supporte pas une cohabitation entre WN 32 et 64 bits
  - Cela a obligé à l'arrêt des quelques WNs 32 bits
- Un site peut recevoir des milliers de jobs en queue si le BDII dysfonctionne
  - ALICE se base sur la réponse du BDII pour savoir s'il reste des slots CPUs libres
- Parfois les jobs utilisateurs consomment toute la mémoire virtuelle du WN
  - Il faut alors rebooter les WNs
  - Problème devenu rare depuis l'été : ALICE a intégré un mécanisme de contrôle de l'utilisation de la mémoire : un job trop gourmand est tué
  - L'IRFU a mis un quota sur l'utilisateur de la mémoire. Même en mettant des valeurs raisonnables, cela semble affecter l'efficacité job du site sauf avec les jobs de production.
- Le site IPNO a été un trou noir à cause des WNs avec 16 job slots
  - ALICE avait une limitation à 14 jobs max par WNs
  - Donc les JA soumis étaient tués mais remplacés par d'autres puisque le BDII indiquait des slots libres
- Difficile de comprendre les raisons d'une mauvaise efficacité des jobs
  - Peut être lié à un dysfonctionnement du stockage
  - Dans le cas de l'IRFU, nous ne savons pas encore pourquoi l'efficacité est si différente de l'IPNO
  - Est-ce que les quotas mémoires mis en place par l'IRFU dans le système de batch affecte l'efficacité ?

ALICE : les problèmes rencontrés : Le stockage

- Crash fréquents des démons xrootd sur les disk servers
  - Solution: Mise en place d'une cron pour les redémarrer
  - Probablement résolu dans la dernière version de xrootd
- Le SE ne répond plus « No managers available » alors que le démon tourne bel et bien
  - Solution: Redmérarrer les démons xrootd sur le head node DPM
- Charqe I/O (MySQL) trop élevée sur le head node → timeout xrootd
  - En effet l'historique des requêtes est conservé
  - Solution: tuner MySQL et purger la base MySQL mais ça peut prendre plusieurs jours
  - Problème résolu depuis DPM 1.7.4 : purge continue paramétrable

ALICE : les problèmes rencontrés
Le stockage (suite)

- Timer expired sur les accès xrootd sur les servers DPM/xrootd
  - Pas forcement sous une forte charge
  - Arrive même si les bases sont purgées et MySQL bien paramétré
  - Les experts soupçonnent un problème réseau
  - Corrigé en principe dans la nouvelle version
- Lors d'un DPM drain, les fichiers volatiles sont supprimés
  - Pour une raison inexpliquée ALICE avait à l'IPNO des TBs de fichiers volatiles : il ont été perdus lors d'un drain
  - Réparation: donner la liste de ces fichiers perdus à ALICE pour qu'il les réplique
  - Solution: changer dans la base le type de fichiers avant de lancer un dpm-drain
- Le SE DPM/xrootd de l'IRFU est sous-utilisé : vide à 70%
  - Nous ne savons pas encore trop pourquoi

Conclusion

- GRIF contribue de façon importante aux ressources de ALICE
- ALICE est de nos jours une VO plutôt facile à supporter
  - Il y a beaucoup moins de problèmes qu'il y a quelques années
  - Il faut une VOBOX avec de préférence un SE xrootd natif
- L'équipe d'ALICE répond plutôt assez efficacement aux demandes des sites
- Au niveau de GRIF il nous reste à résoudre deux problèmes avec ALICE :
  - La sous-utilisation du SE/DPM de l'IRFU
  - La mise en place éventuelle d'un SE xrootd natif à l'IRFU
  - Comprendre la différence au niveau de l 'efficacité des jobs entre l'IPNO et l'IRFU
  - Voir comment faire face aux jobs trop gourmand en mémoire sans affecter l'efficacité des jobs