

LCG-France Tier-1 & Analysis Facility

Relevé des Conclusions

29 Mai 2009 – 10h00-12h30

Présents : Xavier Canehan, Ghita Rahal, Catherine Biscarat, Luisa Arrabito, Nelli Pukhaeva, David Bouvet, Pierre Girard, Farida Fassi

Président et Secrétaire: Fabio Hernandez

Agenda : <http://indico.in2p3.fr/conferenceDisplay.py?confId=1994>

1. Introduction

Ce document est un résumé des discussions et des conclusions suite à la réunion consacrée à étudier le mécanisme mis en place pour l'installation du logiciel des expériences LHC via des jobs grille. Il ne retrace pas les discussions chronologiquement tel qu'elles se sont déroulées pendant la réunion.

2. Rappel du contexte

Le logiciel des 4 expériences LHC est installé par des jobs grille soumis par des personnes ayant un rôle VOMS leur donnant des droits particuliers. Nous avons choisi d'utiliser une zone AFS dédiée à chacune des expériences pour l'installation de ce logiciel et pour son utilisation par les jobs s'exécutant sur les worker nodes. Ceci permet de garantir que l'ensemble de worker nodes de la ferme utilise exactement le même logiciel. Le nombre important de versions ainsi que la taille de chacune d'entre elles, a conduit à mettre en place un mécanisme de réplication du logiciel installé, afin de répartir la charge sur plusieurs serveurs de disque AFS lors de son utilisation par des milliers des jobs en simultanée.

Le job manager BQS s'exécutant sur le CE qui réceptionne les jobs d'installation du logiciel (identifiés grâce à leur rôle VOMS) est chargé de leur octroyer un token AFS avec des droits d'écriture dans la zone correspondante à l'expérience. De plus, il positionne la variable d'environnement appropriée pour indiquer au job le répertoire racine à partir duquel il peut faire son installation. L'organisation des fichiers et de répertoires à l'intérieur de cette zone est la responsabilité de la VO.

A ce jour, les zones AFS utilisées par les 4 expériences LHC ainsi que les variables d'environnement pointant vers elles sont présentées dans le tableau ci-dessous :

Expérience	Variable d'environnement	Répertoire racine dans AFS	Rôle VOMS	Login AFS
ALICE	VO_ALICE_SW_DIR	/afs/in2p3.fr/grid/toolkit/alice	TBD	aligrid
ATLAS	VO_ATLAS_SW_DIR	/afs/in2p3.fr/group/atlas/sw	TBD	atlagrid
CMS	VO_CMS_SW_DIR	/afs/in2p3.fr/grid/toolkit/cms et	TBD	cmsgrid

		/afs/in2p3.fr/grid/toolkit/cms2		
LHCb	VO_LHCb_SW_DIR	/afs/in2p3.fr/grid/toolkit/lhcb	lcgadmin	lhcbgrid

3. Utilisation de AFS

Dans sa présentation, Xavier rappelle les avantages et les inconvénients de AFS en tant que système de fichiers pour l'installation du logiciel des expériences. Rappel des contraintes :

- Nombre élevé de fichiers accédés par un job (plusieurs centaines de bibliothèques dynamiques)
- Nombre élevé des jobs en exécution simultanée (plusieurs milliers) nécessitant accéder au logiciel de l'expérience,
- Volumétrie conséquente d'une version du logiciel (de l'ordre de 10 GB),
- Nombre important de versions installées (plusieurs dizaines),
- Accès en lecture principalement,
- Les composants de chaque version du logiciel de l'expérience sont configurés lors de l'installation avec le chemin complet du répertoire d'installation (non relogeables après installation)
- Procédure de mise à jour d'une version déjà installée est différente de celle de l'installation d'une nouvelle version.

Rappel des avantages de AFS :

- La taille du cache AFS côté client (c'est à dire, sur les worker nodes) est configurable. Elle est actuellement de 20 GB. Le cache manager de AFS assure la cohérence du cache local au worker node.
- Réplication des volumes en lecture seule sur plusieurs serveurs de fichiers distincts, afin de répartir la charge,
- Possibilité de créer des volumes AFS multiples (dans la terminologie AFS un volume est un conteneur de fichiers et de répertoires qui peut être manipulé comme une unité, dupliqué et monté à partir d'un répertoire),
- Possibilité de créer des volumes en lecture seule

Deux problèmes principaux doivent être résolus afin de permettre un job LHC d'installer le logiciel sur AFS et de le répliquer. Tout d'abord, l'opération de réplication de volumes AFS nécessite des privilèges particuliers d'administrateur de la cellule AFS. Elle ne peut pas être directement utilisée par les jobs grille d'installation du logiciel, puisque les comptes sous lesquels ils sont exécutés ne disposent pas (et ne peuvent pas disposer) des privilèges appropriés.

Le deuxième problème est que le chemin d'accès au répertoire en écriture (à utiliser par le job d'installation) est différent de celui qui doit être utilisé en lecture par les jobs « normaux ». Par exemple, dans le cas de ATLAS, le job d'installation doit utiliser un chemin de la forme /afs/.**in2p3.fr**/group/atlas/sw/... (notez le . avant in2p3.fr) et les jobs utilisant ce logiciel doivent y accéder en utilisant un chemin de la forme /afs/**in2p3.fr**/group/atlas/sw/... qui correspond aux copies. D'autre part, les différents composants du logiciel installé doivent être relogés pour utiliser le chemin des espaces dupliqués.

Xavier présente la solution mise au point en collaboration avec Alessandro di Salvo, chargé de l'installation du logiciel ATLAS sur l'ensemble de sites de la collaboration. Deux cas d'utilisation sont adressés par ce mécanisme :

- L'installation d'une nouvelle version
- La mise à jour d'une version préalablement installée

Le mécanisme conçu est relativement simple d'utilisation pour l'expérience mais n'est pas transparent. Il consiste en 2 commandes qui doivent être explicitement invoquées par le script d'installation du logiciel de l'expérience:

- `/usr/local/shared/bin/afs_create`
- `/usr/local/shared/bin/afs_release`

Ces 2 commandes sont accessibles dans le PATH des jobs grille d'installation. L'utilisation attendue pour un job d'installation d'une nouvelle version est dans l'ordre:

1. `afs_create <chemin d'installation>`
Cette commande crée un volume AFS de taille appropriée pour l'installation de la nouvelle version. A titre d'exemple, le chemin d'installation pour ATLAS sera de la forme `/afs/.in2p3.fr/group/atlas/sw/...` Il est construit par le job d'installation à partir de valeur de la variable d'environnement `VO_ATLAS_SW_DIR`.
2. Installation du logiciel proprement dite dans le répertoire `<chemin d'installation>`. Cette installation peut être un processus simple (tar par exemple) ou plus compliqué (basé sur RPM pour la gestion de dépendances). L'expérience a toute latitude pour implémenter le mécanisme le plus adapté pour elle.
3. Remplacer les références à « **.in2p3.fr** » par « `in2p3.fr` » dans tous les composants logiciels installés.
4. `afs_release <chemin d'utilisation>`
Le chemin d'utilisation pour ATLAS sera de la forme `/afs/in2p3.fr/sftgroup/atlas/...` Cette commande provoquera la réplication du volume AFS correspondant sur plusieurs serveurs de fichiers. Les jobs utilisant le logiciel récemment installé, utiliseront l'une des copies read-only gérées par le cache manager de AFS.

Dans le cas d'une mise à jour d'une version du logiciel déjà installée, l'étape 1 n'est pas nécessaire puisque le volume AFS correspondant est déjà créé. Elle est remplacée par un simple `cd <chemin d'installation>`. Les étapes 2, 3 et 4 sont identiques.

Le nombre de copies read-only réalisées pour un volume AFS donné est pré-configuré. Actuellement, 3 copies sont réalisées pour chaque version du logiciel de ATLAS. Les droits d'utilisation des commandes `afs_create` et `afs_release` sont gérés via `prvc`. Pour interroger les utilisateurs ayant le droit d'utiliser ces commandes faire :

```
prvc query -a afsad rocreate rorelease
```

Depuis la date de la réunion, le cas d'utilisation de suppression d'une version installée a été identifié.

Il est important de noter que pour des raisons techniques liées au fonctionnement du cache manager AFS (chemins complets en RO versus des segments en RW), ce framework ne peut être utilisé que pour l'installation du logiciel dans la zone sous la hiérarchie `/afs/in2p3.fr/sftgroup`. Un sous-répertoire sera créé pour toute expérience utilisant le framework. Dans le cas de ATLAS, une phase de transition sera nécessaire entre les 2 espaces: lorsque les versions appropriées du logiciel de l'expérience seront installées dans ce nouvel espace, une modification de la variable d'environnement `VO_ATLAS_SW_DIR` sera opérée pour la faire pointer vers le nouvel espace.

4. Etat des lieux de l'installation du logiciel des expériences LHC

L'état des lieux de ATLAS et LHCb en ce qui concerne l'installation du logiciel a été présenté. Les détails sont disponibles sur les présentations attachées à l'agenda de la réunion. Voici le résumé :

ATLAS

- Mécanisme d'installation complètement automatisé depuis 2006 (voir description dans le document attaché à l'agenda). L'installation proprement dite, la création de tags et la validation des versions installées sont automatisées. Il est possible pour un utilisateur de demander l'installation d'une nouvelle version (via une interface web), d'épingler une version particulière (afin qu'elle ne soit pas supprimé du site) et de notifier l'utilisateur.
- Le statut des installations est consultable à l'adresse : https://atlas-install.roma1.infn.it/atlas_install/list.php
- Au CCIN2P3, l'installation de chaque version est faite une seule fois et publiée pour les CE des 2 sites (tier-1 et tier-2).
- Le volume d'une version est d'environ 8 GB plus 1-2 GB par patch.
- Actuellement : 143 versions différentes (y compris les patches). Il faut noter que le CCIN2P3 est l'un des sites utilisé par ATLAS pour la validation du logiciel, ce qui suppose qu'un nombre plus important de versions est installé sur le site.
- 64 versions ont été installées en 2009. Aucune version n'a été supprimée.
- Catherine aimerait avoir un moyen d'interroger le nombre d'accès à une version donnée. La commande `AFS vos exa extended` pourrait être une solution à ce besoin : elle informe sur la date de dernière utilisation d'un volume AFS. Xavier fournira les détails à Catherine.

LHCb

- L'espace d'installation de LHCb ne contient que de bibliothèques et des exécutables. Aucune compilation n'est faite sur site.
- Actuellement 55 version actives pour les différents packages de LHCb (Boole, Brunel, Dirac, DaVinci, Gauss, ...).
- Espace utilisé actuellement : 25 GB. Estimation de l'espace requis : 50 GB.
- Un job de vérification de la version du logiciel installée est lancé quotidiennement, via SAM. Ce job installe des composants manquants, supprime des versions obsolètes et vérifie les permissions d'accès dans la software area (pointée par `VO_LHCB_SW_DIR`).
- LHCb n'a pas observé de problèmes avec la software area au CCIN2P3.
- Pas de réplication actuellement de cette zone pour LHCb. Le seul volume AFS est en read-write : c'est le même accédé par les jobs, bien qu'ils n'aient pas les privilèges pour modifier cet espace. Une perte de cette zone entraînerait un problème pour tous les jobs LHCb sur le site. Une réplication de la software area de LHCb paraît judicieux.

ALICE

- Pas de précisions concernant la situation de ALICE. L'installation du logiciel est faite à partir de la VO box du site. Nous ignorons si le processus d'installation est déclenché manuellement ou automatiquement via un mécanisme propre à ALICE.
- Il n'y a pas de réplication de la software area de ALICE. Une réplication semble judicieuse.

CMS

- Pas de précisions concernant la situation de CMS. CMS a la particularité que 2 zones (cms et cms2) sont utilisées sans comprendre réellement la motivation de cette configuration particulière.
- CMS utilise RPM et une base de référence privée pour vérifier les dépendances des composants installés.
- Il n'y a pas de réplication des 2 espaces de software area de CMS. Une réplication semble judicieuse.

5. Conclusions

- Un framework est en place pour l'installation du logiciel des expériences, avec possibilités de réplication de la software area. Il a été validé par ATLAS. Ce framework n'est pas transparent et suppose une adaptation du mécanisme d'installation utilisé par l'expérience pour 1) demander la création d'une zone d'installation de la nouvelle release, 2) « reloger » le logiciel installé, 3) demander la réplication de la zone récemment créée. L'étape 2), générée par la distinction entre le chemin d'installation et le chemin d'accès, est probablement la plus contraignante.
- Xavier et ses collègues administrateurs AFS disposent des outils pour surveiller la charge sur les serveurs de fichiers qui servent les espaces disque de software area des 4 expériences LHC.
- Le problème lié au manque d'un outil pour passer des commandes sur les worker nodes sans s'y connecter (pour le debug des jobs) a été à nouveau soulevé. FH va en discuter avec le responsable du groupe système.
- Basés sur cette expérience avec ATLAS, un framework générique pourrait être proposé à WLCG pour implémentation sur les sites qui le souhaitent. FH va coordonner cette action.
- La possibilité de distinguer les zones disque utilisées pour l'installation de celles utilisées par les jobs « normaux » a été discutée. Dans ce modèle, les zones d'installation et les zones d'utilisation pourraient avoir le même espace de nommage évitant ainsi la nécessité pour l'expérience de reloger systématiquement le logiciel installé. Des solutions basées sur NFS ou sur des machines virtuelles ont été évoquées et restent à explorer. Des worker nodes pour l'exécution des jobs d'installation pourraient être nécessaires : les jobs seraient étiquetés par le job manager comme job d'installation et BQS se chargerait de la mise en exécution sur les machines appropriées. Cette solution enlève un peu de la flexibilité actuelle où le job d'installation peut être mis en exécution sur un worker node banalisé.