FG-iRODS : mutualisation d'expertise et infrastructure distribuée pour un accès transparent et hautement disponible aux données scientifiques

David Benaben, Catherine Biscarat, Yonny Cardenas, Hélène Cordier, Pierre Gay, Benoit Hiroux, Gilles Mathieu, Emmanuel Medernach, Jean-Yves Nief, Jérôme Pansanel, et al.

FG-iRODS : mutualisation d'expertise et infrastructure distribuée pour un accès transparent et hautement disponible aux données scientifiques

David Benaben
Centre de Bio-Informatique de Bordeaux
71, avenue Edouard Bouriaux – CS 20032
33882 Villenave d'Ornon Cedex

Catherine Biscarat
Laboratoire de Physique Subatomique et de Cosmologie
53, avenue des Martyrs
38026 Grenoble Cedex

Yonnis Cardenas
Centre de Calcul de l'IN2P3
21, avenue Pierre de Coubertin CS70202
69627 Villeurbanne Cedex

Hélène Cordier
Centre de Calcul de l'IN2P3
21, avenue Pierre de Coubertin CS70202
69627 Villeurbanne Cedex

Pierre Gay
Mésocentre de Calcul Intensif Aquitaine
351, cours de la Libération
33405 Talence Cedex

Benoît Hiroux
Mésocentre de Calcul Intensif Aquitaine
351, cours de la Libération
33405 Talence Cedex

Gilles Mathieu
INSERM
c/o Délégation Rhône-Alpes Auvergne
95, boulevard Pinel
69675 Bron Cedex

Emmanuel Medernach
Institut Pluridisciplinaire Hubert Curien
23, rue du Loess – BP28
67037 Strasbourg Cedex 2

Jean-Yves Nief
Centre de Calcul de l'IN2P3
21, avenue Pierre de Coubertin CS70202
69627 Villeurbanne Cedex

Jérôme Pansanel
Institut Pluridisciplinaire Hubert Curien
23, rue du Loess – BP28
67037 Strasbourg Cedex 2

Geneviève Romier
Institut des Grilles et du Cloud
c/o Centre de Calcul de l'IN2P3
21, avenue Pierre de Coubertin CS70202
69627 Villeurbanne Cedex

Résumé

Les instituts de recherche sont confrontés actuellement à un double problème : une augmentation conséquente des volumes de données issus des instruments qui doivent être analysés et stockés, et une pression constante pour réduire les ressources humaines et financières allouées au traitement de ces données.

C'est dans ce contexte compliqué que cinq laboratoires français, le Centre de Calcul de l'IN2P3 à Villeurbanne, l'Institut Pluridisciplinaire Hubert Curien à Strasbourg, le Laboratoire de Physique Subatomique et de Cosmologie à Grenoble, le Mésocentre de Calcul Intensif Aquitain et le Centre de Bio-Informatique de Bordeaux à Bordeaux, mutualisent depuis 2013 expertises et efforts pour fournir un service de gestion de données scientifiques innovant dans le cadre du projet FG-iRODS.

Ce projet, piloté par France Grilles, repose sur une instance iRODS (integrated Rule Oriented Data Management System) se composant de nœuds de stockage distribués géographiquement à Bordeaux, Grenoble et Strasbourg, ainsi que d'un catalogue de données hébergé à Villeurbanne. Il intègre également un ensemble de services pour répondre aux besoins des chercheurs des petites et moyennes communautés scientifiques dans les problématiques de gestion des données scientifiques (réplication, haute disponibilité, intégration dans les workflows, extraction automatisée de métadonnées, gestion fine des droits d'accès, ...) et leur offrir une infrastructure de niveau « production ».

De plus, un accompagnement complet est proposé aux utilisateurs dans le cadre de FG-iRODS, tels que formations, développements ou conseil, et permet de proposer une offre sur mesure répondant aux différentes problématiques de stockage et assure aux utilisateurs de rester maître de leurs données.

Mots-clefs

IRODS, DONNÉES SCIENTIFIQUES, STOCKAGE, MUTUALISATION D'EXPERTISE, SERVICE, INFRASTRUCTURE DISTRIBUÉE, BIG DATA

1 Introduction

Les communautés scientifiques acquièrent et traitent de plus en plus de données qui peuvent être produites expérimentalement ou être le résultat de calculs. Les besoins des chercheurs pour la gestion de ces données sont par exemple :

- traiter un grand volume de données, éventuellement distribuées sur plusieurs sites ;
- permettre la collaboration de sites partenaires équipés d'infrastructures et de matériels hétérogènes ;
- abstraire l'organisation physique des fichiers pour les utilisateurs ;
- rechercher des lots de données par métadonnées ;

– gérer finement les droits d'accès aux fichiers (ACL) ;

– accéder à distance aux données depuis différents types d'interface (PC, smartphone, API, ...).

En parallèle, depuis plusieurs années maintenant, les ingénieurs en informatique sont maintenus sous pression constante pour fournir des services de plus en plus évolués, avec une haute disponibilité et des enveloppes budgétaires en constante diminution. S'ajoute à cette équation une baisse régulière du nombre de personnels dédiés à l'informatique dans la plupart des laboratoires.

Afin de répondre aux besoins des chercheurs dans ce contexte difficile, plusieurs laboratoires interagissant dans le cadre du GIS France Grilles ont décidé de relever le défi en créant un nouveau service de gestion de données scientifiques hautement disponible basé sur iRODS : le projet FG-iRODS.

1.1 France Grilles

France Grilles est un groupement d'intérêt scientifique (GIS) créé en 2010, par 8 établissements français de recherche (CEA, CNRS, CPU, INRA, INRIA, INSERM, MESR, RENATER).

Piloté par l'Institut des Grilles et du Cloud, ses principales missions sont :

– établir et opérer une infrastructure distribuée nationale pour le traitement et le stockage de données scientifiques massives (grille [1] et *Cloud* de production [2]) ;

– contribuer avec les autres états membre impliqués au fonctionnement de l'e-Infrastructure européenne EGI ;

– favoriser les rapprochements et les échanges entre les équipes travaillant sur les infrastructures distribuées de production et de recherche.

France Grilles coordonne les opérations de 18 fournisseurs de ressources de calcul et de stockage géographiquement répartis en France, comme le montre la Figure 1. Ces ressources peuvent être de différents types, notamment du stockage iRODS.

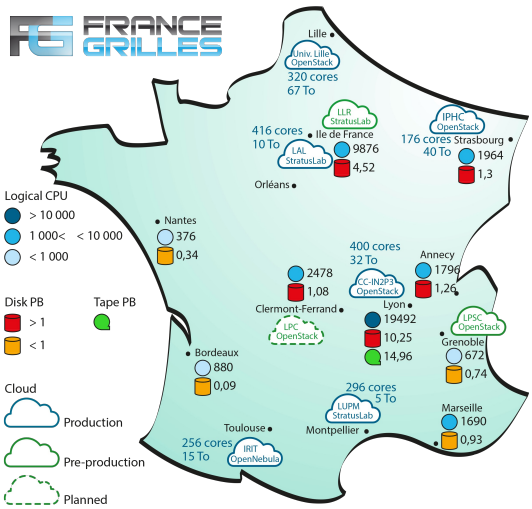

Figure 1 - Organisation de l'infrastructure de production opérée par France Grilles

1.2 iRODS

Le logiciel iRODS, créé en 2006 et diffusé sous licence BSD, permet de virtualiser le stockage et propose des fonctionnalités particulières :

- gestion de collection de données sur des ressources de stockage hétérogènes et géographiquement distribuées ;
- accès au catalogue comportant les informations relatives aux fichiers stockés, ainsi que les métadonnées qui y sont associées ;
- application de politiques de gestion de données et de règles prédéfinies ;
- accès à distance aux données, depuis différents types de terminaux ;
- possibilité de gérer de larges collections de données.

La liste de ces fonctionnalités n'est pas exhaustive, iRODS reposant sur un système de règles permettant d'automatiser la gestion des données et d'appeler des micro-services pour les traiter.

L'ensemble de ces fonctionnalités, le fait que le logiciel soit maintenu par une communauté active et que des instances soient déjà proposées par des partenaires de France Grilles, ont amené à la création du projet FG-iRODS.[3]

2 FG-iRODS

2.1 Objectifs

Le projet FG-iRODS a démarré en 2013 à l'initiative de quatre laboratoires, la coordination étant assurée par France Grilles :

- l'Institut Pluridisciplinaire Hubert Curien (IPHC) à Strasbourg ;
- le Centre de Calcul de l'IN2P3 (CC-IN2P3) à Villeurbanne ;
- le Mésocentre de Calcul Intensif Aquitain (MCIA) à Bordeaux ;
- le Laboratoire de Physique Subatomique et de Cosmologie (LPSC) à Grenoble.

Les objectifs initiaux étaient multiples. Nous souhaitions proposer une infrastructure composée de ressources de stockage distribuées géographiquement et présentées à l'utilisateur de manière unifiée. L'infrastructure se devait d'être de niveau « production » et nous devions assurer les formations pour faciliter l'adoption par les utilisateurs. De plus, l'ensemble de ces objectifs devait être atteint avec un investissement financier minimum.

Le projet a nécessité deux phases de développement consécutives pour que l'infrastructure et les services associés soient officiellement en production.

2.2 Infrastructure

Toute zone iRODS comporte un serveur iCAT (base de données relationnelle stockant les données relatives aux ressources et les métadonnées liées aux fichiers) et un ensemble de ressource de stockage. Dans le cas de FG-iRODS, le catalogue iCAT est hébergé au CC-IN2P3, alors que les ressources de stockage iRODS sont hébergées dans les trois autres laboratoires (IPHC, LPSC, MCIA), comme le détaille la Figure 2. Le service dispose actuellement d'une volumétrie de 100 To, idéale pour répondre aux besoins des petites et moyennes communautés scientifiques.

L'authentification peut être effectuée à l'aide d'un couple identifiant/mot de passe ou d'un certificat. Pour mémoire, un certificat est nécessaire pour l'utilisation des services grille. Ainsi, ces ressources sont accessibles depuis n'importe quelle machine cliente, en utilisant les clients natifs iRODS ou via une interface Web. Les clients sont déployés sur tous les nœuds de calcul de l'infrastructure de grille EGI en France. Ils peuvent également facilement être installés sur les serveurs du Cloud fédéré France Grilles.

Ainsi un utilisateur de France Grilles dispose d'un accès facile à travers des services grille ou *Cloud* à ses données sur FG-iRODS aussi bien que depuis sa propre machine. Les droits d'accès aux fichiers sont gérés de manière très fine en fonction des utilisateurs et des projets.

La surveillance des services est assurée à l'aide de Nagios, que ce soit au niveau des sites pour la supervision du matériel, ou au niveau de France Grilles pour les tests fonctionnels. Ainsi, la Figure 3 présente l'interface de surveillance accessible aux administrateurs du projet FG-iRODs pour vérifier que l'infrastructure fonctionne correctement. En cas d'erreur, les administrateurs ou les utilisateurs ont la possibilité d'ouvrir un ticket avec l'interface XGUS de France Grilles et de suivre la résolution du problème.

Afin d'assurer une réponse rapide aux problèmes pouvant survenir sur l'une des ressources de stockage, chaque administrateur d'une ressource FG-iRODS peut se connecter à l'un des autres sites pour diagnostiquer l'origine d'un problème et, si possible, le résoudre. La coopération entre les différents administrateurs est un élément essentiel de la synergie ayant permis à ce service d'exister, et ce, dans des conditions optimales.

Figure 2 - Détail des éléments composant l'infrastructure FG-iRODS

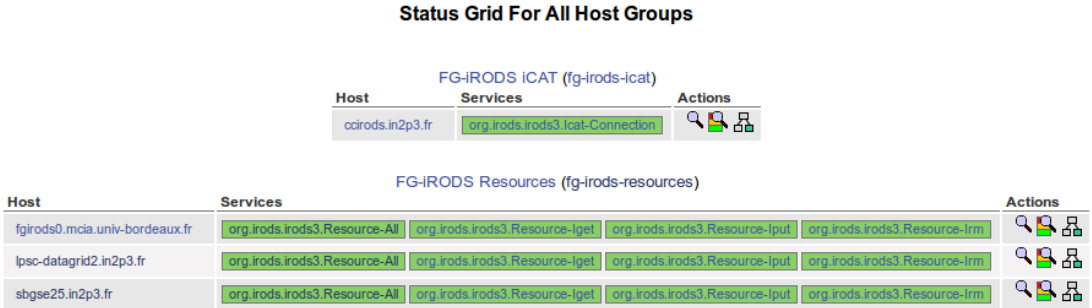

Figure 3 - Supervision du service FG-iRODS à l'aide de Nagios

2.3 Services aux utilisateurs

L'accueil des nouveaux utilisateurs suit une procédure, dont l'un des objectifs est de comprendre au mieux leurs besoins. Ainsi, la première étape de cette procédure est la collecte d'information, qui une fois étudiée par le groupe FG-iRODS, permettra de déterminer :

- si l'infrastructure FG-iRODS répond effectivement aux besoins de l'utilisateur ;
- quels sont les éventuels développements devant être réalisés pour répondre à ses besoins ;
- quelles sont les autres actions à réaliser pour accueillir l'utilisateur.

Les développements à réaliser sont principalement liés à de nouvelles règles et micro-services. Ils permettront par exemple d'extraire automatiquement des métadonnées lors du dépôt de fichiers ou de garantir le respect des politiques de gestion des données.

Les utilisateurs sont invités à signer la charte (Acceptable Use Policy ou AUP dans la dénomination EGI) France Grilles comme pour les autres services ainsi que les conditions d'utilisation du service iRODS. Ces conditions prennent en compte la particularité du stockage de données.

Des sessions de formation pour les utilisateurs sont également assurées une à deux fois par an. En plus de fournir toutes les informations nécessaires au bon usage des ressources, elles permettent également de renforcer les liens entre utilisateurs et administrateurs de ressources, de contribuer à l'intégration des utilisateurs dans la communauté France Grilles.

Une documentation collaborative a été mise en place comme pour les autres services de façon à guider les utilisateurs dans leur pratique et de contribuer à la constitution d'une communauté d'utilisateurs.

Nous soulignons également qu'une procédure est disponible pour les utilisateurs souhaitant récupérer leurs données. Ainsi, les utilisateurs gardent la maîtrise de leurs données durant tout le cycle de vie du projet.

3 Exemples d'utilisation

3.1 Stockage de données pour l'analyse et la sauvegarde

L'un des groupes utilisant l'infrastructure FG-iRODS a adopté ce service pour :

- sauvegarder certaines données expérimentales de son laboratoire ;
- utiliser ces données sauvegardées dans des analyses exécutées sur les nœuds de grille accessibles aux utilisateurs de la VO vo.france-grilles.fr.

L'avantage principal pour le groupe utilisateur est la possibilité à la fois de sauvegarder des données hors site, et de les utiliser facilement dans des analyses, l'ensemble des nœuds de grille EGI français disposant du client iRODS (l'accès est transparent grâce à l'utilisation du certificat).

3.2 Stockage massif et partage d'expertise

Un autre exemple d'utilisation de l'infrastructure est celui d'un groupe souhaitant héberger 1000 To de données. Ne disposant pas des ressources nécessaires, le choix a été fait de les accompagner dans le déploiement de leur infrastructure iRODS avec la possibilité d'échange d'espaces de stockage pour effectuer certaines duplications de données hors site. Cette solution très flexible permet à la fois au groupe de démarrer rapidement l'acquisition des données en utilisant l'infrastructure FG-iRODS et de poursuivre sereinement en utilisant leur infrastructure, une fois qu'ils auront acquis les compétences d'administration du système.

3.3 Suivi de l'utilisation

Le suivi de l'utilisation des ressources est réalisé régulièrement à l'aide d'une interface développée et hébergée au CC-IN2P3. Une revue annuelle est également effectuée avec chaque groupe utilisateur dans le cadre d'une volonté d'améliorer régulièrement la qualité du service.

4 Conclusion

Depuis son lancement en 2013, le projet FG-iRODS a permis de créer un service complet pour la gestion des données scientifiques, de niveau « production » adossé aux services de traitement de données (grille et cloud). Il offre actuellement ressources et compétences à 5 groupes de recherche.

Bien que les partenaires du projet aient réussi à créer une dynamique constructive qui a permis de réussir dans le projet initial, les défis pour la suite sont importants. En effet, citons par exemple la question de la pérennité des données ou bien celle de l'évolution des ressources pour répondre au besoin croissant des utilisateurs. Des éléments de réponse ont été trouvés et intégrés à l'offre de service, mais ne sont pas encore suffisants pour répondre correctement à ces questions.

Bibiographie

[1] G. Wormser. L'Institut des Grilles. Dans Actes des Journées Mésocentre, Paris, Février 2008. http://calcul.math.cnrs.fr/Documents/Journees/fev2008/institut_des_grilles_meso.pdf

[2] G. Mathieu et al. Vers une fédération de Cloud Académique dans France Grilles. Dans Actes des Journées SUCCES, Paris, Novembre 2013. http://succes2013.sciencesconf.org/conference/succes2013/FG_Cloud_20131112.pdf

[3] C. Biscarat et al. Mise en place d'un gestionnaire de données léger, pluridisciplinaire et national pour les données scientifiques. Dans Actes des Journées SUCCES, Paris, France, Novembre2013. http://succes2013.sciencesconf.org/conference/succes2013/SUCCES2013_FG_iRODS_2013_11_14.pdf