ACTES

2e colloque LCG-France
14 et 15 mars 2007
Clermont-Ferrand

Site LCG-France :
http://lcg.in2p3.fr

Site du Colloque :
http://indico.in2p3.fr/conferenceDisplay.py?confId=82

Rédactrice en chef : Frédérique Chollet
Notes des sessions : Karim Bernardet, Frédérique Chollet, Michel Jouvin, Pierrick Micout
Mise en page : Gaëlle Shifrin
Editeurs : Frédérique Chollet, Fabio Hernandez, Fairouz Malek et Gaëlle Shifrin

REMERCIEMENTS

Aux 70 participants inscrits
Aux membres du Comité d'Organisation
Aux présentateurs et aux présidents de session
A nos hôtes du LPC Clermont, à Dominique Pallin et aux membres du comité d'organisation local
Aux participants à la rédaction et à la relecture de ces Actes

SOMMAIRE

PREAMBULE

Le deuxième colloque LCG-France s'est tenu à Clermont-Ferrand les 14 et 15 mars 2007.
Ces journées, organisées par l'IN2P3 et le Dapnia, ont réuni près de 70 participants, acteurs de la grille de calcul LHC en France. Ce colloque a été l'occasion d'échanges entre représentants des sites (français et étrangers) et représentants des expériences. Il a permis de faire le point sur la mise en place du calcul pour le LHC dans le cadre du projet mondial W-LCG, d'évoquer les actions en cours et les perspectives pour 2007 et au-delà.

COMITE D'ORGANISATION

Président : Dominique Pallin, LPC-Clermont

Laurent Aphecetche (ALICE, CD LCG-France), Jean-Michel Barbet (Tier-2 Subatech-Nantes), Giacomo Bruno (Tier-2 CMS Belgique), Othmane Bouhali (Tier-2 CMS Belgique), Claude Charlot (CMS, CD LCG-France), Jean-Claude Chevaleyre (Tier-2 LPC-Clermont), Hélène Cordier (Tier-1 CC-IN2P3), Fréderic Derue (Tier-2 LPNHE-Paris), Christophe Diarra (Tier-2 IPN-Orsay), Eric Fede (Tier-3 LAPP-Annecy), Pierre Girard (Tier-1 CC-IN2P3), Fabio Hernandez (Responsable technique LCG-France), Michel Jouvin (Tier-2 LAL-Orsay), Pascal Van Laer (Tier-2 CMS Belgique), Cal Loomis (Tier-2 LAL-Orsay), Eric Lançon (ATLAS, CD LCG-France), Christine Leroy (Tier-2 Dapnia-Saclay), Liliana Martin (Tier-2 LPNHE-Paris), Fairouz Malek (Responsable scientifique LCG-France), Vania Maraganzova-Martin (LIG et GRID 5000), Pierrick Micout (Tier-2 Dapnia-Saclay), Thierry Ollivier (Tier-3 IPN-Lyon), Yannick Patois (Tier-3 IPHC-Strasbourg), Ghita Rahal (ATLAS, CC-IN2P3).

Etat des lieux pour le calcul LHC en France

Chair : Ghita Rahal (CCIN2P3), Dominique Pallin (LPC Clermont)

Le calcul LHC en France s'appuie sur une infrastructure de grille qui répond aux besoins des modèles de calcul des expériences. Ces modèles prévoient la distribution des traitements (acquisition, reconstruction, simulation, analyse) et des lots de données (brutes, reconstruites, simulées, réduites) sur une hiérarchie de sites de capacité et de niveaux de services différents.

Cette session a permis d'apprécier l'état d'avancement du projet LCG-France initié par l'IN2P3 et le Dapnia dans le contexte du projet W-LCG et de mesurer les efforts engagés tant au niveau scientifique et technique qu'au niveau de la gestion de projet. Les besoins des expériences qui s'expriment non seulement en termes de puissance de calcul et d'espace de stockage répartis mais également en termes de bande passante réseau ont été clairement présentés. Les expériences sont plutôt satisfaites des prestations offertes par le Tier-1 français, de même que le bilan d'utilisation et la visibilité de la composante française sont plutôt satisfaisants. La montée en puissance de l'utilisation de la grille en 2007 devra nécessairement s'accompagner d'améliorations notoires. La stabilité des services de grille et la performance des accès aux données constituent deux enjeux majeurs. Les sites Tier2 vont devoir être impliqués plus fortement encore. La participation des Tier-3s n'en est pas moins indispensable pour l'analyse des données. Il reste que l'implication des personnels des différents sites et des expériences semble être un des points clef à améliorer cette année.

Infrastructure de grille LHC en France et ressources associées

Fairouz Malek (LPSC)

La mise en place d'un Tier-1 au Centre de Calcul de Lyon et l'intégration des sites Tiers-2 /Tiers-3 dans l'infrastructure W-LCG sont les deux objectifs du projet LCG-France. Fairouz Malek a mis en perspective la montée en puissance des sites attendue entre 2006 et 2010 en termes de CPU, stockage disque et stockage de masse. Typiquement, les capacités du Centre de Calcul doivent être au moins multipliées par 16 sur cette période.

L'une des difficultés du projet LCG-France est liée au fait que les ressources associées au calcul LHC doivent être planifiées sur 5 ans alors que les décisions budgétaires sont annuelles. De ce fait, il est difficile d'avoir une vision à long terme. La participation des groupes de physiciens au projet est assez inégale. ATLAS se distingue par une forte implication. Trente personnes au centre de calcul sont fortement impliquées dans le projet, ce qui représente la moitié des effectifs du CCIN2P3. Plusieurs d'entre eux sont sur des contrats à durée déterminée (EGEE en particulier). C'est pourquoi l'avenir d'EGEE (EGEE 3 et EGI) a une importance très grande. Il est bien évident que malgré un effort pour recruter un maximum de personnel, il y a un risque de perte de compétences.

Dans le cadre de l'infrastructure de grille mondiale W-LCG, LCG France a établi des relations avec des sites Tier-1s et Tier-2s étrangers. Parmi les éléments qui contribuent à la complexité du projet, Fairouz Malek relève aussi le nombre d'interlocuteurs (institutionnels, projet, expériences, sites français et étrangers…), la nécessité de réactivité par rapport aux problèmes de déploiement de la grille, l'analyse des risques. Il n'en demeure pas moins que le défi scientifique et technique, la visibilité du projet sont des atouts et que les efforts de tous ne sont pas vains. En effet, LCG-France et plus largement le calcul LHC contribuent à enrichir l'animation scientifique au sein de l'IN2P3, à la création d'une communauté et donnent un sens à la mutualisation des ressources de calcul en France dans notre discipline.

Mise en place du Tier-1

Fabio Hernandez (CCIN2P3)

Cette présentation a passé en revue les activités 2006 et les perspectives 2007. Parmi les objectifs du projet LCG-France, la contribution du Tier-1 LCG-France représente environ 10% des ressources requises par les expériences dans l'ensemble des sites Tier-1s. Force est de constater que les résultats 2006 issus des rapports d'utilisation EGEE sont conformes aux objectifs car la contribution du CCIN2P3 à l'effort global en temps CPUs (grille et non grille) de tous les Tier-1s pour les quatre expériences a été globalement de 10% : 23% pour Alice, 9% pour ATLAS, 5% pour CMS, 14% pour LHCb.

Fabio Hernandez nous a proposé un bilan exhaustif de l'année 2006, montrant la répartition entre les utilisations grille et non-grille du Centre, l'efficacité des travaux soumis par la grille (rapport du temps CPU sur le temps d'exécution total), le rapport du temps CPU effectivement consommé par rapport ce qui a été planifié, ainsi que la capacité effectivement délivrée compte tenu des interruptions de service notamment dues au problème de refroidissement en aout 2006. En 2006, globalement il n'y a pas eu de problème pour répondre aux besoins. L'allocation de la capacité CPU du CCIN2P3 est restée équitable entre expériences LHC (45%) et expériences non LHC (55%), alors que la consommation des ressources CPU par les expériences LHC ne représente que 20% du total du temps CPU réellement consommé. Coté stockage disque, le constat est différent puisque seulement 34 % de l'espace (soit 180 To sur les 520 To initialement prévus) a pu être mis en place en 2006. A contrario, la totalité du stockage de masse (soit 535 To) dont 73% sont utilisés, a été installée.

Un des problèmes rencontrés en 2006 est lié aux délais de livraison des matériels et à leur installation. La nouvelle robotique à base d'unités à cartouches de haute capacité dont l'achat a été initié en 2006 est en cours d'installation. Outre l'augmentation de la capacité, l'évolution des infrastructures (puissance électrique, refroidissement…) et l'amélioration de la qualité de service sont également des enjeux pour le Centre de Calcul. Malheureusement, force est de constater que la qualité du middleware est encore insuffisante pour assurer la disponibilité des sites Tier-1 au niveau requis. Le bilan 2006 est complété par certains résultats illustrant de la participation du CCIN2P3 aux différents exercices de validation des expériences.

La mise à jour des infrastructures constitue le point numéro 1 du plan d'action 2007. L'augmentation des ressources va bien évidemment se poursuivre voire s'accélérer. En 2007, le Centre de Calcul va en effet provisionner les ressources planifiées pour 2007 plus une partie (env. 40 %) des ressources requises en 2008. La réflexion qui a commencé autour de la mise en place d'une facilité d'analyse doit se poursuivre et va nécessiter une implication des expériences pour bien comprendre les besoins. Enfin, 2007 doit voir le démarrage d'un autre projet d'envergure : la construction d'un nouveau bâtiment comprenant une nouvelle salle machine qui devrait être disponible à l'horizon 2009.

Les sites Tier-2s et Tier-3s

Frédérique Chollet (LAPP)

Outre le Dapnia et le CCIN2P3, l'infrastructure LHC implique aujourd'hui en France plus de onze laboratoires de l'IN2P3 avec une particularité au projet de Tier-2 GRIF (Grille de Recherche Ile de France) qui regroupe les efforts des laboratoires de la région parisienne. Trois Tier-2s (GRIF, LPC Clermont et SUBATECH) ainsi que la facilité d'analyse de Lyon (AF Lyon) sont inscrits dans le MoU WLCG. On a assisté à une augmentation du nombre de Tier-3s aujourd'hui au nombre de 4 (CPPM, IPHC, IPNL, LAPP). Les quatre expériences sont équitablement supportées dans les différents sites qui supportent également des expériences, projets ou domaines de recherche non LHC.

Globalement, la contribution des Tier-2s français et de la facilité d'analyse de Lyon représente 33% des ressources CPU proposées par LCG-France. Le critère d'élection des sites Tier-2s vis-à-vis de la collaboration WLCG n'est pas tant lié à la capacité proposée en termes ressources qu'à la qualité de service offerte à long terme. Dans le contexte LHC, la fourniture de la bande passante requise pour la gestion des données entre sites Tier-2s et Tier-1 est un point extrêmement important, totalement complémentaire à la fourniture des ressources de calcul et de stockage. Une étude sur l'interconnexion du site GRIF avec le Tier-1 et entre ses

diverses composantes est également en cours. Les Tier-3s ne figurent pas dans le MoU et n'ont pas d'engagement formalisé vis-à-vis de la collaboration W-LCG. Ces projets ne sont pas à négliger pour autant car leur nombre tant à croître. Ils sont par ailleurs utilisés par les expériences indifféremment dès qu'ils sont disponibles. Aujourd'hui, leur rôle est double offrant une contribution de petits Tier-2s aux expériences et de facilité d'analyse aux utilisateurs locaux. Les résultats obtenus en 2006 illustrent parfaitement la contribution des sites Tier-2s et Tier-3s au Service Challenge SC4 et aux exercices de validation.

L'objectif du groupe technique LCG-France T2-T3 créé en avril 2006 est de renforcer les collaborations entre tous les acteurs : à savoir administrateurs de sites, les experts du CCIN2P3, l'équipe du support opérationnel EGEE et les physiciens des expériences.

De nombreux points sont inscrits au plan d'action 2007 notamment : le déploiement de l'espace de stockage disque compatible avec le standard d'interface SRM (Storage Resource Management) de la grille, le déploiement du mécanisme d'autorisations de groupes et rôles VOMS (Virtual Organisation Management System), le suivi de la disponibilité et de l'efficacité des sites, la prise en compte de la composante grille dans la gestion de la sécurité informatique. Enfin, il est mentionné le début d'une réflexion visant à mettre en place un accord au sien de LCG-France entre l'IN2P3 et le Dapnia d'une part et les sites Tier-2s d'autre part. Cet accord formaliserait l'engagement des Tier-2s à satisfaire aux critères W-LCG et à pourvoir aux moyens matériels et aux ressources humaines pour la durée du LHC, l'engagement d'entraide mutuelle entre le Tier-1 et les Tier-2s ainsi que l'engagement de l'IN2P3 et du Dapnia pour le maintien des ressources financières et du personnel qualifié à partir de 2009 dans les Tier-2s .

Le calcul dans les expériences : ALICE

Patricia Méndez Lorenzo (CERN)

En 2006, le PDC06 d'Alice a été opérationnel en continu pendant plus de 9 mois avec plus de 40 millions d'événements produits. Alice se considère en Data Challenge continu. Il est mentionné les discussions en cours sur le rattachement au Tier-1 français de sites ALICE orphelins en Espagne et en Corée.

Le calcul d'Alice s'appuie sur l'environnement AliEn, «middleware» de grille léger, qui est installé dans tous les sites supportant Alice au niveau des VO-Boxes. Il existe un document de spécifications ainsi qu'une équipe support au CERN. En complément de la surveillance assurée par MonaLisa, des tests SAMs spécifiques à Alice vont être mis en place très prochainement.

Le bilan d'utilisation des sites français entre avril 2006 et mars 2007 montre que la contribution des Tiers-2 (51%) est essentielle. La politique d'Alice est d'envoyer les jobs où se trouvent les données et il n'y a pas de différence pour l'expérience entre Tier-1 et Tier-2 à part la qualité de service. Alice souhaite intégrer le plus rapidement possible les nouveaux sites français de Grenoble et de Strasbourg. Par ailleurs, le Centre de calcul de Lyon a été le Tier-1 Alice le plus stable au cours de l'exercice Tier-0/Tier-1. Globalement, pour Alice, il y a un grand besoin d'améliorer la stabilité des services au sein de la grille.

Le calcul dans les expériences : ATLAS

Eric Lançon (Dapnia)

Eric lançon s'est attaché à présenter le modèle de calcul Atlas et l'utilisation qui a été faite de la grille LCG, introduisant par là même le mode de fonctionnement en « nuage» des Tiers d'ATLAS.

La réévaluation par rapport au TDR qui a été faite cette année tient compte des nouvelles hypothèses LHC et revoit certaines estimations propres à l'expérience comme la taille des ESDs, la proportion de données simulées et le temps de simulation par événement. Cette réévaluation entraine une baisse de 20% des besoins pour les Tier-1s et de 10% pour les Tier-2s en 2008.

Eric Lançon tire les leçons des exercices réalisés en 2006 (production Monte Carlo et transferts de données). Le logiciel n'est pas encore stable. Il y a un manque d'alarmes et de personnel. La production est plutôt compliquée. Malgré tout, l'année 2006 a été un succès et le nuage FR a été très (ré)actif au cours des différents

tests. Le nuage français a contribué à hauteur de 16 % à la production Monte Carlo d'ATLAS effectuée sur la grille LCG, le Centre de Calcul de Lyon contribuant pour 57 % de cette part.

Beaucoup d'améliorations restent à apporter qui relèvent à la fois des sites, du middleware de la grille et du logiciel de l'expérience. Il reste également beaucoup à faire en 2007, concernant le reprocessing au niveau du Tier-1 et l'analyse des données. Le format de bas niveau des données n'étant pas encore défini, la mise en place de la base de données de calibration reste à faire. Le modèle d'analyse doit être précisé. Il prévoit de spécialiser certains Tier-2s pour les analyses chaotiques mais tout cela s'appuie sur la définition, la configuration et l'utilisation des groupes et rôles qu'il convient de finaliser. L'analyse interactive n'entre pas dans ce modèle mais elle se fera typiquement dans les Tier-3s. L'année 2007 sera très importante car l'utilisation des sites va évoluer chaque Tier prenant progressivement la place qui lui est destinée au sein du modèle de calcul.

Le calcul dans les expériences : CMS

Claude Charlot (LLR)

La problématique essentielle du calcul pour CMS est une augmentation d'un ordre de grandeur du volume de données par rapport aux expériences en cours. Claude Charlot rappelle les données du problème et les principes de base du modèle envisagé pour le traitement des données.

La taille des données brutes est estimée à 1,5 Mo au démarrage. Cette taille initiale prend en compte un facteur lié à la connaissance du détecteur et devrait donc diminuer avec le temps. Néanmoins, l'augmentation de la luminosité compensera partiellement cette diminution. Avec un taux de déclenchement nominal de 150Hz, le volume total des données brutes est d'environ 4,5 Po/an. Le modèle d'analyse prévoit en plus de la copie située au CERN une seconde copie des données brutes répartie entre les grands sites (sites Tier-1). L'analyse se fera sur des données réduites ou AODs dont chaque site Tier-1 disposera d'une copie (environ 2 Po/an). Il est prévu de produire un nombre d'événements Monte Carlo en quantité approximativement équivalente au nombre d'événements réels acquis par l'expérience.

CMS considère que l'essentiel est de disposer d'un code de reconstruction rapide et de traiter les données selon des priorités appropriées, établies par l'expérience surtout au début de la prise de données aux LHC.

Claude Charlot évoque les fonctions des différents Tiers pour CMS :

- Le CERN en qualité de Tier-0 est chargé de l'archivage des données brutes et de la première reconstruction rapide.
- Les Tiers-1 au nombre de sept reçoivent chacun une copie d'une partie des données brutes et des données issues de la première reconstruction. Ils sont en charge du reprocessing des données réelles comme des données simulées, de l'archivage de la production Monte Carlo faite sur les Tier-2s et de la distribution des données d'analyse vers les autres Tier-1s ainsi que les Tier-2s.
- Enfin, les sites Tier-2s sont dédiés à l'analyse et à la production Monte Carlo et associés à un Tier1.

A noter qu'il est prévu trois reprocessings par an et que chaque Tier1 disposera de la totalité des données (AODs) nécessaires à l'analyse. La facilité d'analyse mise en place au CERN sera ouverte à l'ensemble de la collaboration CMS et équivaudra à un Tier 1 pour la partie stockage de masse et de 2 Tier-2s pour les parties CPU et disque.

CMS prévoit une forte acquisition dès 2007 et les nouvelles hypothèses concernant le LHC affectent peu les besoins intégrés sur la période 2008-2010. Claude Charlot met en avant une progression agressive en termes de ressources (CPU, disque et MSS) qui doivent être multipliées par un facteur entre 2 et 3 sur ces 3 ans. Concernant la distribution des données et les associations Tier-1s/Tier-2s, Claude Charlot explique que CMS prévoit de s'appuyer fortement sur la définition de lots de données primaires par le système de déclenchement (estimés actuellement à 50). Le CCIN2P3 hébergera 6/50 de la copie distribuée des données brutes (au prorata de la capacité disque proposée). Il est prévu que le CCIN2P3 héberge les données de simulation produites à Lyon, au GRIF et aux Tier-2s belge et chinois. Concernant l'accès aux AODs disponibles à Lyon,

une charge de 20% supplémentaires est à prévoir pour l'accès depuis les Tier-2s non associés à un Tier-1 (ie n'ayant pas de Tier-1 national et n'ayant pas signé d'accord particulier avec un Tier-1 étranger).

En 2006, dans le contexte du CSA06, le CCIN2P3 a passé avec succès les tests de transferts du Tier-0 vers les Tier-1s avec un taux nominal de 25 Mo/s atteint ainsi que les tests de reprocessing. Le GRIF a participé au même titre que 24 autres Tier-2s CMS à une période de tests de transferts Phedex dont l'objectif nominal était de 5 Mo/s. En 2007, CMS se prépare au data taking, les objectifs du CS07 prévu à l'été 2007 sont fixés à 50 % de ceux du modèle de calcul LHC. Les activités du CCIN2P3 vont se recentrer sur les fonctions de Tier-1 avec une importance accrue de la participation des Tier-2s, GRIF en particulier. Claude Charlot conclut sur l'importance de la mise en place de la partie Tier-2 de Lyon pour les besoins de l'analyse de données dès le second semestre.

Le calcul dans les expériences : LHCb

Andrei Tsaregorodtsev (CCPM)

Le calcul LHCb prévoit :
- l'enregistrement en temps réel des données au niveau du Tier-0
- une chaine de traitements des données (reconstruction, simulation, analyse) totalement distribuée
- des traitements qui seront exécutés au plus près des données qui seront-elles même réparties.

La distribution des tâches de traitement s'accompagnent d'un scenario de distribution des données explicites. Ainsi les données brutes seront copiées depuis le CERN (Tier-0) vers les Tier-1s qui assureront la reconstruction, le stripping et l'analyse. La simulation s'effectuera dans les Tier-2s occasionnant un flux de données remontant des Tier-2s vers les Tier-1s. Les six Tiers se partageront sur leur espace de stockage de masse un jeu de données brutes, un jeu de données simulées et un jeu de données reconstruites. Les besoins en stockage sont estimés pour permettre de disposer de 7 copies des données reconstruites.

LHCb dispose d'un « middleware de grille léger » : DIRAC incluant un système de gestions des tâches et un système de gestion des données et qui fait appel aux services génériques LCG lorsque cela est possible. L'idée de base est de minimiser les interventions humaines au niveau des sites. Le résultat du développement mené au sein de LHCb est un environnement à base d'agents légers distribués. La gestion des jobs est centralisée et fait appel à des jobs « pilotes » déployés sur les nœuds de calcul (worker nodes) capables d'aller chercher (selon un fonctionnement en mode PULL) les jobs DIRAC dans une base où ils sont ordonnés. C'est donc au niveau de DIRAC et de la queue centralisée des jobs LHCb que sont appliquées les règles de scheduling et les priorités souhaitées par l'expérience. LHCb a déployé des VO-Boxes dans tous les sites Tier-1s qui présentent l'avantage d'améliorer l'équilibre de charge et la redondance entre sites. Du coté de la gestion de données, LHCb a principalement développé un outil (service de transfert) de haut niveau au dessus des outils génériques de LCG capable de gérer les flux de données entre sites.
La contribution française en termes de CPU depuis décembre 2006 est de moins de 8 % avec deux sites visibles le CCIN2P3 (4,5 %) et LPC Clermont (3,3 %). La production Monte Carlo est efficace au point qu'il serait possible de fonctionner sur un mode opportuniste en vol de cycle à la LHCb@home. Par contre la distribution des données et la reconstruction pâtissent de l'instabilité des interfaces et mécanismes de stockage sur la grille. L'analyse utilisateur démarre bien que l'accès aux données demeure un point crucial.

Gestion et exploitation des grilles de calcul

Chair : Christine Leroy (Dapnia), Pierre Girard (CCIN2P3)

Cette session a été préparée et conduite avec un objectif très précis : avoir un échange de points de vue et aménager un espace de discussion entre les organisations virtuelles ou VOs incarnées par les expériences et les sites participant à l'infrastructure de calcul LHC. Dans cette optique, il était demandé aux expériences de présenter les mécanismes de soumission et de gestion des jobs de calcul sur la grille et aux sites d'aborder des aspects liés à l'exploitation et à la gestion de l'infrastructure à l'échelle globale de la grille comme à l'échelle locale des sites.

Force est de constater que chacun s'est plié avec bonne volonté à l'exercice : « Les VOs parlent aux sites » et «Les sites parlent aux VOs ». Faisant parfaitement écho aux conclusions provisoires et bilans présentées par les 4 expériences le matin même, la formule originale proposée par Christine Leroy et Pierre Girard s'est révélée parfaitement adaptée et extrêmement fructueuse.

« Les VOs parlent aux sites »

Chacune des quatre expériences a présenté les spécificités des travaux soumis sur la grille, l'essentiel concernant aujourd'hui la production Monte Carlo qui, en règle générale, est gérée de façon centralisée. Par contre, les environnements et mécanismes de soumission sont très différents.

Ce tour d'horizon était d'autant plus important que les expériences utilisent des environnements et des services spécifiques. Bien comprendre l'environnement et les mécanismes utilisés par les expériences permet de bien comprendre les demandes des expériences vis-à-vis des sites. Le fait que les VOs n'utilisent pas en règle générale les outils standards de EGEE/LCG est un sujet de préoccupation pour les sites qui craignent dans une certaine mesure de perdre la visibilité et le contrôle de l'utilisation de leurs ressources. Ces solutions ont pour but de contourner les faiblesses du middleware et sont donc présentées par les VOs comme une condition nécessaire au bon fonctionnement de leurs activités de production.

Cette situation, comprise par les sites, est tolérée faute de mieux. Mais il est attendu unanimement par les sites que les VOs abandonnent leurs solutions spécifiques, coûteuses pour les sites qui doivent mettre en place autant de solutions que de VOs, dès lors que le middleware offrira une réponse générique à leurs besoins. Il est donc important de faire évoluer le middleware dans ce sens et, aux vues de leurs solutions, les VOs semblent être une force de proposition primordiale pour ce faire.

« Les sites parlent aux VOs »

Avec plus de 17 millions de jobs soumis en 2006, on peut qualifier la grille LCG/EGEE d'opérationnelle en dépit des problèmes que rencontrent encore les expériences LHC. Dans cette partie, deux aspects complémentaires sont abordés : d'une part, l'exploitation globale de la grille qui joue un rôle fondamental dans la stabilité de la grille ; d'autre part, le fonctionnement et l'exploitation locale des sites. L'objectif des présentations était de montrer aux VOs l'ensemble des efforts faits par les sites, que ce soit globalement ou localement, pour améliorer la qualité de la production sur la grille.

Autre sujet abordé au cours des échanges : la gestion des priorités entre jobs de différentes natures (production, analyse). On note que les fonctionnalités utiles peuvent être implémentées à différents niveaux et que les stratégies des VOs et des sites diffèrent un peu. Il existe des possibilités intéressantes au niveau du système de batch local utilisé, possibilité d'appliquer des quotas en définissant des objectifs pour chaque VO notamment. Par contre, compte tenu de la richesse potentielle des VOMS, propager ce qui sera décidé par les VOs jusqu'au scheduler des sites n'est pas évident. Par ailleurs, aujourd'hui il n'est pas possible de publier des informations suffisamment précises. Des améliorations sont attendues avec la prochaine version du schéma d'information (Glue Schema 1.3) qui devraient permettent de rendre lisible la stratégie des sites.

On comprend donc la stratégie alternative de LHCb de gérer l'ensemble de cette problématique à son niveau de manière centralisée.

Aujourd'hui, sites et VOs ont le même souci : l'amélioration de l'efficacité intrinsèque de la grille et le support des activités des expériences avec un objectif commun : Améliorer les choses encore et encore. A l'issue de ce workshop, il reste beaucoup de travail à l'interface des VOS/expériences et des sites.

Introduction

Pierre Girard (CCIN2P3)

Allant à l'essentiel, l'introduction a rappelé à tous que « *depuis 3 ans, les expériences mettent au point leur modèle de calcul sur la grille alors que les sites mettent au point le déploiement du middleware et la gestion de leur production, chacun des 2 acteurs ayant une perception partielle et une connaissance factuelle de la problématique de l'autre. Les connaissances sont souvent acquises de la moins bonne façon qui soit, c'est-à-dire au gré des problèmes rencontrés, dans l'urgence et la difficulté. A la veille de l'entrée en production, il était primordial que chaque partie apporte à l'autre les précisions nécessaires à une meilleure compréhension de son fonctionnement et donc, à un travail collaboratif plus efficace* ».

« Les VOs parlent aux sites »

Spécificités d'ATLAS

Jérôme Schwindling (Dapnia)

L'installation, la validation et le changement de versions des logiciels ATLAS sur la grille sont aujourd'hui totalement automatisés (chaque site peut demander le maintien de certaines versions). Les caractéristiques des jobs depuis la simulation Monte Carlo jusqu'à l'analyse par des utilisateurs sont très différentes. Les jobs sont regroupés par « tâches » (1 à 40000 jobs) de différentes natures : génération, simulation/digitisation, reconstruction.

Aujourd'hui, il y a très peu de production Monte Carlo effectuée « à titre personnel ». Celle-ci est découragée afin de garantir la reproductibilité de la simulation et l'accès aux données générées. Il convient de souligner l'effort français et l'implication des physiciens dans la gestion de la production Monte Carlo qui assure une visibilité au nuage français, c'est-à-dire à l'ensemble des sites Tier-2s et Tier-3s connectés au Tier-1 de Lyon. Un exécuteur a été installé sur une machine à Lyon (de type VO-Box). Celui-ci gère prioritairement les « tâches » françaises avec pour objectif affiché l'amélioration de l'efficacité globale sur l'infrastructure LCG-France.

Les aspects techniques liés à la sélection des sites et éventuellement leur radiation momentanée, la surveillance et le suivi de l'efficacité de la production ont été présentés. L'efficacité actuelle est de 70 à 80 % avec des problèmes essentiellement causés par l'accès aux fichiers d'entrée (stage-in). L'utilisation du mécanisme des jobs pilotes est également fortement envisagée par ATLAS. Un test comparatif Condor versus Cronus est en cours au sein d'ATLAS avec des résultats attendus pour la mi-mai. Rendez-vous est pris pour suivre l'évolution du système de production d'ATLAS.

Spécificités de CMS et d'Alice

Artem Trunov (CCIN2P3)

Les outils spécifiques utilisés par Alice (services AliEN) et CMS (Phedex qui doit a priori tourner sur tous les sites , CRAB…) ont été présentés.

L'environnement AliEN d'Alice est indépendant de l'implémentation du middleware gLite/LCG. Il est composé de services centralisés gérant le catalogue, la mise en queue des travaux soumis et les autorisations et de services locaux présents sur les sites et hébergés au niveau de la VO-Box. Le modèle de « Job agent » mis en place au sein d'AliEn consiste en un job éclaireur qui est soumis par la voie standard LCG et qui contacte la base de données centrale pour l'envoi des « vrais jobs » Alice. Le rapport entre « jobs éclaireurs » et « vrais jobs » semble difficile à estimer. L'accès au stockage se fait par le protocole natif xrootd dont les performances semblent bonnes avec Castor et dCache mais qui n'est pas encore supporté par DPM. A noter également l'intérêt des deux expériences pour PROOF déjà mis en place au niveau de la facilité d'analyse du CERN, utilisé par Alice mais qui n'est compatible qu'avec ROOT.

Spécificités de LHCb

Andrei Tsaregorodtsev (CCPM)

Comme Alice et CMS, l'expérience LHCb dans un souci de performances et de fiabilité des expériences exploite un mécanisme de soumission en mode PULL non encore disponible dans la suite Middleware standard LCG.

La présentation d'Andrei Tsaregorodtsev a également permis de comprendre l'intérêt des jobs dits pilotes ou pilot jobs pour LHCb. Ces jobs soumis via le Ressource Broker permettent de s'assurer que le site de destination est correctement configuré et de réserver de la ressource de calcul pour les jobs de calcul qui sont ensuite soumis en mode pull et donc tirés vers le site selon les priorités de l'expérience à partir d'une queue centrale de jobs en attente d'exécution. En conséquence, LHCb a uniquement besoin de disposer d'une queue d'exécution longue au niveau de chaque site. LHCb est intéressée par la possibilité de sélectionner des sites selon ses propres critères ce qui devrait être possible avec la version gLite du WMS (Workload Management System).

Vus des sites, tous les jobs LHCb sont de même nature. Les sites doivent juste se préoccuper de la répartition des ressources entre LHCb et les autres VOs. La gestion des priorités des jobs et le suivi d'utilisation au niveau des groupes et des utilisateurs seront effectués par LHCb. Avec le mécanisme des jobs pilotes, les jobs LHCb restent anonymes contrairement aux jobs ordinaires qui se propagent avec l'empreinte du certificat utilisateur. Dans ce contexte, les jobs arrivant sur les sites appartiennent au groupe de gestion de la production, au propriétaire du service de soumission ou même au spokeman LHCb dans le cas extrême de l'analyse. Pour restituer l'appartenance des jobs, ne pas violer les règles d'utilisation de la grille et permettre aux sites d'effectuer un suivi, LHCb demande le support de glexec.

« Les sites parlent aux VOs »

Exploitation globale de la grille

Hélène Cordier (CCIN2P3)

Depuis l'entrée d'un nouveau site jusqu'au suivi de l'utilisation globale de la grille, les aspects opérationnels sont essentiels. Les outils disponibles (tests, outils de surveillance et alertes, astreintes…) permettent de connaître et de suivre l'état global de la grille notamment à partir des informations statiques de la base de connaissance générale (GOCDB) et des informations dynamiques publiées et mises à jour automatiquement dans le système d'information.

En complément, l'état (et donc la stabilité de chaque site) est établi par une batterie de tests fonctionnels soumis régulièrement dont les résultats sont utilisables par le système d'information pour filtrer les informations publiées par les ressources et services des sites en échec. Hélène Cordier a également présenté le système de support utilisateur GGUS (Global Grid User Support) et le site CIC (Core Infrastructure Centre). Il y a sans doute des fonctionnalités méconnues à destination des VOS, des utilisateurs et responsables de site sur le CIC (http://cic.gridops.org) : métriques perf. sites, contact utilisateur…Il est également mentionné l'importance du suivi d'utilisation pour les Tiers-2 (démarrage de la collecte officielle au 1er avril pour publication

annoncée au 1er septembre avec 6 mois d'historique).

Hélène Cordier a également dressé le plan d'action 2007. Pour le Tier-1, le CCIN2P3 a comme objectifs : la consolidation des services de grille au niveau des autres services, l'augmentation de la bande passante avec les Tier-2s et de la liaison avec les autres Tier-1s et FZK, enfin le lancement du projet de construction du nouveau bâtiment. L'amélioration de la disponibilité des sites Tier-2s et Tier-3s et la poursuite des tests de transfert de données est également au programme de l'année 2007. Enfin, faire en sorte que les administrateurs de site comprennent bien comment se fera l'accès aux données est un objectif qui doit être porté par l'ensemble de la communauté et donc par le projet LCG-France.

Suivi des jobs grille

David Bouvet (CCIN2P3)

La présentation fait le point sur la surveillance du site du CCIN2P3, les outils de suivi et de traçabilité en place, les problèmes rencontrés et les actions entreprises. Différents outils de surveillance (vision graphique de l'état des services OVAX, outil statistique sur le nombre de jobs en temps réel MRTG, surveillance de l'état des machines CCSMURF, WebRLS pour l'historique des problèmes) sont combinés à Nagios. Ces outils locaux sont complétés par un outil propre à LCG, proposé par ARDA encore appelé Dashboard et présenté comme un outil générique, point d'entrée unique adapté aux besoins des expériences, capable de collecter toutes les données de surveillance utiles. Il existe également au CCIN2P3 des outils de détection des jobs problématiques comme les jobs dits « slow » dont le rapport entre la consommation CPU et le temps de résidence en machine est faible, ou « early ended » qui ont une consommation CPU très faible. Les outils de suivi collectent des informations à tous les niveaux auprès de BQS, de la gatekeeper du Computing Element (CE) et même sur le Worker Node (WN) et sont associés des mécanismes d'alerte. En cas de problème, on se rend bien compte de la difficulté de constituer le diagnostic, d'identifier le job à travers l'ensemble de la chaine de soumission, de contacter l'utilisateur le cas échéant.

Surveillance et gestion d'incidents

Cécile Barbier (LAPP)

La mise en place d'outils de surveillance de l'infrastructure et de gestion des incidents est une préoccupation reprise au niveau de chaque site. Il est par ailleurs important que chaque site mette en place un outil local de suivi d'activité. Pour illustration, Cécile Barbier a présenté l'outil développé au LAPP.

Les besoins sont génériques mais les outils locaux doivent dans une certaine mesure s'adapter aux spécificités de chaque site. Plusieurs outils de surveillance des ressources assez répandus sont disponibles parmi lesquels Lemon, Ganglia, Nagios, Cacti et permettent entre autres de détecter certaines pannes ou dysfonctionnements. Les outils de suivi d'activité permettent de caractériser les services (éventuellement utile pour les redimensionner) et d'agir en conséquence, de visualiser l'utilisation des ressources et de vérifier que les priorités et la répartition des ressources sont respectées. Consulté par les administrateurs de site comme par les utilisateurs, le suivi d'activité devient alors un élément clé de la gestion de projet, surtout lorsque l'infrastructure est partagée entre différents partenaires ou communautés d'utilisateurs qui s'entendent sur des objectifs ou sur la répartition des ressources.

Job scheduling et tuning

Michel Jouvin (LAL)

Au sein de chaque VO, l'ordonnancement efficace des jobs doit permettre d'affecter un quota par type de tache (par ex ; la production Monte Carlo centralisée) selon les combinaisons rôle/groupe utilisée, de donner un accès rapide à certaines tâches d'analyse voire même dans certains cas bien particuliers d'autoriser une pseudo-interactivité.

L'ordonnancement des jobs sur la grille s'effectue à deux niveaux. Le choix d'un site et d'une queue est effectué par le service de Resource Broker selon les critères exprimés dans le fichier JDL de description du job. Au niveau de chaque site, l'ordonnancement dépend du système de batch local. Les possibilités offertes par le couple Torque/Maui permettent d'affecter une priorité à chaque job en attente en tenant compte à la fois d'un objectif cible par VO et de l'historique sur une période donné.

Chaque site est responsable de définir le quota de chacune des VOs selon ses propres engagements. Chaque VO est quant à elle responsable de définir le quota de chaque groupe/rôle à l'intérieur du quota de la VO. Mais la question de savoir si ces informations doivent être poussées vers les sites et si oui comment reste entière car le service GPBOX permettant de définir et de gérer dynamiquement ces quotas n'est pas disponible.

Gestion de l'infrastructure d'un site

Pierre-Louis Reichstadt (LPC Clermont)

Illustrant son propos avec le cas de Clermont-Ferrand, Pierre-Louis Reichstadt a passé en revue les points clés et parfois les embuches d'un projet de Tier-2.

« *Un Tier 2, c'est un vrai projet*, nous confie Pierre-Louis Reichstadt, *qui implique une gestion administrative lourde et un enchaînement logique (qui peut être récursif et/ou itératif) : définir le périmètre du projet, réaliser les études (fonctionnelles et techniques) écrire les dossiers de présentation, rechercher les financements (et les obtenir !), réaliser les appels d'offre (écriture des documents de consultation, lancement, suivi, choix...) enfin suivre les livraisons et installations, tout ceci avant de mettre en route l'infrastructure et les services de grille !* »

Les financements d'origines multiples, les interlocuteurs variés, le choix de l'organisme porteur ont un impact sur le suivi administratif et financier du projet. Les problèmes d'infrastructure ne sont pas à négliger sur le plan technique comme sur le plan financier. De la conformité et la fiabilité des installations dépendent la sécurité, le bon fonctionnement et la fiabilité du site. Il faut également savoir que les surcoûts sont élevés si les locaux se révèlent inadaptés. Ainsi, Pierre-Louis Reichstadt donne de bonnes indications sur le chiffrage d'une salle typique : partie matériel et partie infrastructure.

Communication LCG-France

Chair : Gaëlle Shifrin (CCIN2P3)

Il est proposé de mettre en place une communication LCG-France qui s'inscrive dans le cadre de la communication LHC. Par communication LCG-France, on entend communication externe du projet LCG-France. Cette activité aurait clairement comme objectif d'améliorer la visibilité de la contribution française au projet W-LCG. Aujourd'hui, Gaëlle Shifrin, chargée de communication au CC-IN2P3 (notamment de LCG), en collaboration avec Perrine Royole-Degieux en charge de la communication LHC à l'IN2P3, se propose de contribuer à cet effort en travaillant étroitement avec la direction LCG-France.

Gaëlle Shifrin se déclare prête à coordonner les actions dans ce domaine et propose d'identifier un contact par site. Cependant, avant d'évoquer les actions possibles (plaquettes, sites web, charte graphique, support à l'organisation d'événements locaux, communiqués de presse….), tous les acteurs du projet doivent se prononcer sur :
- ce qu'ils attendent d'une communication 'LCG France'
- les éventuelles initiatives propres à chaque site
- les cibles prioritaires et les messages fondamentaux .

La proposition est reçue favorablement. La discussion porte sur l'orientation principale à donner à la communication LCG-France : association forte avec le LHC, mention du projet EGEE…Il semble possible de combiner efficacement la communication LCG-France et la communication un peu plus spécifique aux sites. Frédérique Chollet verrait bien une plaquette LCG-France projet (format Recto-verso A4) utilisable par tous déclinée dans une version où l'on conserve un recto projet et un verso site par ex.

Gestion des données grilles

Chair : Lionel Schwarz (CCIN2P3), Stéphane Jézéquel (LAPP)

Les expériences du LHC ont un modèle similaire pour le mouvement des données. A l'exception d'une particularité pour LHCb, les mouvements de données prévus entre sites sont bien identifiés:
- la migration des données brutes et des premières données reconstruites du CERN vers les Tier-1s
- les transferts des données réduites entre sites Tier-1s et Tier-2s (réplication entre Tier-1s, distribution vers les Tier-2s)
- enfin, les transferts des données Monte Carlo produites par les Tier-2s vers les Tier-1s

Les expériences ont rencontré les mêmes problèmes pendant leurs tests de transfert : difficulté d'avoir tous les sites opérationnels en même temps (problèmes avec les éléments de stockage, pannes électrique, problèmes SRM). Elles ont développé leurs propres outils de transfert pour pallier certains de ces problèmes ou bien pour interfacer plusieurs infrastructures de grilles (EGEE, OSG par ex.). La nouvelle version de SRM devrait en régler certains mais la majorité de ces problèmes sont dus au changement d'échelle. Le CC-IN2P3 incite les Tiers français à réaliser des tests de transferts intensifs.

L'accès aux données est également dépendant des problèmes liés à l'instabilité des SE (SRM, ...). On peut également mentionner la difficulté que représente l'intégration de l'ensemble des composants du système de gestion des données (FTS, SRM, gridFTP, LFC…). La plupart des VOs privilégient un accès local des données du Worker Node vers le SE local via des protocoles comme dcap, rfio ou root (ce qui n'est pas le cas d'ATLAS pour le moment pour la production Monte Carlo). Xrootd donne entière satisfaction à Alice. Une implémentation de xrootd dans DPM est à l'étude. Notons qu'il y a une forte expertise de xrootd au CC-IN2P3. Aujourd'hui, les sites sont plutôt préoccupés par la gestion des jobs mais ce point devra être très vite pris en compte.

Point sur l'infrastructure réseau et stockage des sites

Lionel Schwarz (CCIN2P3)

En introduction et avant de laisser la parole aux expériences, Lionel Schwarz présente l'infrastructure réseau actuelle s'appuyant sur Renater ainsi que les différentes infrastructures de stockage mises en place dans les sites français. Le CCIN2P3 en tant que Tier-1 utilise dCache alors que tous les autres sites Tier-2s et Tier-3s utilisent DPM. Trois sites utilisent également GPFS comme système de fichiers.

Transferts massifs : ALICE

Patricia Méndez Lorenzo (CERN)

Alice utilise un «middleware de grille léger» : ALIEN. ALIEN possède plusieurs services centraux. Un de ces services est FTD (File Transfer Daemon) pour le transfert de fichiers. Il est déployé sur tous leurs sites pour la réplication des fichiers. Il nécessite des éléments de stockage de type SRM et utilise FTS. FTS est installé sur la VO-Box. Le mode «pull» est utilisé : ce qui signifie que c'est le SE (Storage Element) de destination qui initie les transferts. Les tests se poursuivent entre le CERN et les Tier-1s pour vérifier la stabilité de FTS et son intégration avec FTD. Pendant ce type de tests, il est difficile d'avoir tous les T1 opérationnels en même temps (en général 3 sur 5 l'étaient). Entre le CERN et le CC-IN2P3, le taux de transfert était de 40 Mo/s en moyenne (la cible étant 60 Mo/s). De manière générale, le support des experts a été excellent. Le CC-IN2P3 a été le site le plus stable.

Transferts massifs : ATLAS

Vincent Garonne (LAL)

Atlas a un modèle pour la gestion des données similaire à celui d'Alice. Le débit du T0 vers l'ensemble des Tier-1s serait de l'ordre de 1 GB/s et 20 Mo/s d'un T1 vers ses T2 associés. On estime à 300 000 le nombre de fichiers qui seront enregistrés par jour (données brutes, Monte Carlo, reprocessing, répliques) avec un facteur 2 ou 3 possible.

Atlas doit fonctionner avec trois grilles (LCG, OSG et NorduGrid). DDM est l'outil développé par Atlas pour la gestion des fichiers produits par ces grilles. C'est une couche au dessus du middleware de ces grilles. DDM est constitué de services centraux et d'agents installés sur les VO-Boxes (une dans chaque Tier-1). Il fonctionne également en mode «pull» et utilise FTS pour les transferts. Les sites LCG sont organisés en nuages : par exemple, le nuage français est constitué du CC-IN2P3 et de ses Tier-2s et Tier-3s associés. Le catalogue LFC utilisé est celui du Tier-1. Aujourd'hui, l'accès aux *Storage Elements* se fait via SRM, bien que la question se pose pour l'analyse.

Comme pour Alice, pendant les tests de transfert T0-T1 d'ATLAS, il a été très difficile d'avoir tous les T1 fonctionnant en même temps. Le CC-IN2P3 a été un des meilleurs sites. De plus, comme d'autres VO faisaient des tests similaires en même temps, les débits chutaient. Un des principaux problèmes est le manque de stabilité des SEs.

Transferts massifs : CMS

Artem Trunov (CCIN2P3)

Artem Trunov rappelle quels sont les échanges de fichiers prévus entre sites et les taux de transferts nominaux estimés :
- échange de données brutes et des données issues de la première reconstruction du CERN vers les Tier-1s (~30 Mo/s) ;
- échange d'AODs entre sites Tier-1s (entrée ~50 Mo/s sortie ~50 Mo/s) ;
- distribution de données réduites des Tier-1s vers les Tier-2s (~100 Mo/s) ;
- et transferts des fichiers MC des Tier-2s vers le Tier-1 associé (~10 Mo/s).

L'outil de transfert utilisé par CMS est PHEDEX. Artem Trunov rappelle que c'est le seul outil spécifique qui doit être installé dans tous les sites CMS. Au cours des tests T0-T1, CMS a atteint 23 Mo/s en moyenne avec le CC-IN2P3. Le but était de vérifier la stabilité pendant un mois. Concernant les tests T1-T1 et T1-T2 , ils se sont révélés peu performants.

Comme les autres expériences, CMS mentionne que beaucoup de problèmes sont liés à l'instabilité de SRM. Artem Trunov analyse les difficultés actuelles : le modèle d'authentification GSI qui requiert l'ouverture de ports et fonctionne difficilement en mode NAT, les implémentations de l'interface SRM difficilement interopérables, le service FTS qui ne prend pas en compte la gestion de la bande passante et ne permet pas aux sites Tier-1s de contrôler les flux sortants.
Pendant le test T0-T1, un débit record à 250 Mo/s a été atteint entre le CERN et le CC-IN2P3 (absence de tests d'autres VOs), ceci notamment grâce aux experts du centre de calcul de l'IN2P3.

Transferts massifs : LHCb

Andrei Tsaregorodtsev (CCPM)

Les transferts du puits vers le Tier-0 (débit à 70 Mo/s soit un fichier de 2 Gb toutes les 30 secondes) et la réplication automatique vers les Tier-1s à la même vitesse seront réalisés avec DIRAC. Les requêtes de transfert sont insérées et mises en queue dans une base de données puis elles sont exécutées par des agents de transfert présents sur les VO-Boxes. Des précautions sont prises pour ne pas perdre de données (contrôle de l'intégrité des fichiers, les données sont supprimées du puits si par exemple une copie existe sur un Tier-1 ...).

Le système est en phase de test (test grande échelle en avril). Pendant les tests de réplication T0-T1, il était difficile d'avoir tous les T1 opérationnels en même temps et d'atteindre l'objectif fixé à 40 Mo/s. Concernant la distribution des données reconstruites qui doivent être présentes sur tous les Tier-1s pour analyse, le principal problème est l'instabilité des SE (source comme destination). A l'heure actuelle, la distribution des données entre les Tier-1s est impossible aux taux nominaux : 10-15 Mo/s entrant et sortant pour chaque Tier-1.

Accès aux données : LHCb

Andrei Tsaregorodtsev (CCPM)

Andrei Tsaregorodtsev détaille le mécanisme d'accès aux données mis en place par DIRAC qui s'assure que l'accès aux fichiers d'entrée est possible avant de démarrer l'application. Chaque job récupère la «meilleure» réplique (par exemple, la réplique locale si elle existe) et met le fichier à disposition (avec la commande lcg-gt) qui peut alors être accédé directement via dcap ou rfio. Mais dans le cas d'un job utilisant plusieurs fichiers d'input, il peut arriver que des fichiers disparaissent du cache (possibilité de réservation avec SRM 2.2).

Andrei Tsaregorodtsev détaille plusieurs problèmes d'accès aux données. La sémantique de la commande lcg-gt n'est pas la même selon castor (retourne le TURL après que le fichier soit stagée) ou dcache (retourne le TURL vers où le fichier devrait être stagé). L'accès dcap d'un fichier non stagé ne fonctionne pas. Le mécanisme de réservation n'est pas disponible actuellement. Ainsi quand le job utilise plusieurs fichiers d'input, certains peuvent être supprimés du cache. Andrei Tsaregorodtsev évoque également la surcharge des systèmes de stockage en cas d'accès concurrents. Rien que pour l'activité de reconstruction sur un Tier-1, l'infrastructure de stockage doit pouvoir supporter un débit de 100 Mo/s correspondant à une centaine de jobs concurrents. Par ailleurs, l'interface SRM répond très lentement quand la charge est trop importante (timeout). Enfin, les services de stockage montrent une certaine fragilité et instabilité et il est impossible de soumettre un ticket GGUS pour chaque problème occasionnel rencontré.

Face à ces problèmes, LHCb pourrait développer son propre service de staging afin que les fichiers soient stagés avant que le job arrive sur le site ou bien ajouter cette fonction de staging dans les applications (prototype) via par exemple la librairie GFAL et le support de la fonction srmBringOnline. Aujourd'hui, l'accès aux données reste problématique. Des améliorations sont attendues avec la version 2.2 de SRM mais la stabilité des services restera un point critique.

Accès aux données : CMS

Artem Trunov (CCIN2P3)

Les règles d'accès aux données énoncées par CMS prévoient l'utilisation de SRM et FTS pour les transferts entre sites, l'accès aux données localement via les protocoles rfio, dcap ou root en évitant de recopier les fichiers d'entrée sur le nœud de calcul, ainsi que la mise à disposition des résultats sur le SE local. Pour l'accès aux données de calibration et de condition, squid et le protocole http sont utilisés.
Pour établir la correspondance entre nom de fichier logique (LFN) et nom de fichier physique (PFN), CMS n'utilise pas de catalogue LFC ou toute autre base de données mais plutôt leur outil TFC (Trivial File Catalog). Il s'agit d'un fichier XML qui contient une série de règles pour convertir un LFN vers un PFN ou l'inverse.

Accès aux données : ATLAS

Vincent Garonne (LAL)

Vincent Garonne illustre les différentes étapes nécessaires pour l'analyse des données (sélection, localisation, accès et lecture des données). Les données d'ATLAS sont regroupées en datasets (ensembles de fichiers). Pour l'analyse, le job rejoint les données : le catalogue DDM est interrogé pour connaître les sites sur lesquels se trouve le dataset puis le job est soumis sur un de ces sites. AMI permet d'avoir les métadonnées associées à un dataset. Pour la localisation des données, DDM utilise des catalogues locaux (LFC pour LCG et chaque Tier-1). On compte l'enregistrement de 300 000 à 1 million de fichiers par jour. L'information des répliques

est locale aux sites ce qui pose un problème (requête lente) si le dataset comporte un grand nombre de fichiers.

L'un des problèmes posés par la grille LCG tient au fait que l'atomicité des datasets n'est pas garantie. Les appels LFC peuvent échouer. Une solution serait de rajouter l'information de dataset localement (développement en cours). L'interface d'accès SRM pose beaucoup de problèmes. ATLAS s'interroge sur la pertinence d'utiliser xrootd (en test à SLAC). Les jobs de production Monte Carlo copient localement les fichiers d'input (même si les fichiers sont distants). Concernant les jobs d'analyse, la situation est également délicate car il est prévu qu'ils utilisent un accès direct (dcap, rfio) mais l'accès via dpm ne fonctionne pas directement (patch à appliquer dans l'application). Aujourd'hui, la compatibilité des logiciels expérimentaux avec les outils utilisés pour générer les données sur la grille n'est pas assurée et c'est là la principale difficulté.

Accès aux données : ALICE

Patricia Méndez Lorenzo (CERN)

Alice utilise xrootd comme protocole d'accès aux données et obtient de bons résultats avec castor et dcache. Une implémentation DPM de ce protocole est en cours. xrootd a été testé avec plus de 2000 clients concurrents accédant au même serveur de disque et a démontré une grande stabilité sans problème particulier.

En production, la lecture des données brutes sera prédominante et les serveurs de données doivent être dimensionnés en tenant compte du nombre de nœuds de calcul et du trafic induit par la réplication des données depuis le CERN. PROOF (avec xrootd) est utilisé pour l'analyse : utilisation des disques locaux des nœuds de PROOF ce qui permet d'avoir le maximum de performance; toute autre méthode d'accès serait inefficace. La limitation vient des performances des serveurs de stockage et du réseau.

Centres d'analyse

Chair : Eric Lançon(Dapnia), Claude Charlot (LLR)

La situation a évolué par rapport au discours «La grille n'est pas prête pour l'analyse» entendu les années passées et aujourd'hui on constate que l'analyse sur la grille démarre. 2006 a vu des premiers exercices d'analyse dans toutes les expériences. En dépit de la diversité des solutions, il est clair que, comme D0 l'a illustré, l'analyse requiert un front-end pour faciliter l'utilisation de ressources complexes et l'inter-connexion avec les outils de gestion de données. Il existe deux approches : celle d'ATLAS et LHCb qui développent un front-end commun adaptable à leurs besoins spécifiques et celle d'Alice (et peut être de CMS) qui s'orientent vers une infrastructure spécifique (aujourd'hui hors grille) telle PROOF qui pourrait ne fonctionner qu'avec ROOT. Il faut donc conclure que de nombreux points cruciaux restent à définir précisément : format de base, protocole d'accès sans oublier qu'il faudra impérativement prendre en compte la spécificité des besoins des applications d'analyse notamment en entrées/sorties pour espérer atteindre les performances escomptées.

Introduction

Claude Charlot (LLR)

Bien qu'elle se situe en bout de la chaine de traitement, l'analyse doit idéalement intervenir aussi vite que possible après la prise de données. Le système dédié à l'analyse doit être flexible avec le moins de charge opérationnelle possible afin de permettre de nombreuses itérations rapides. L'analyse est basée sur la réduction des données (filtrage d'événements et réduction information). Le format de base pour l'analyse est un format de données réduites mais il y aura sans doute pas mal d'analyses faites à partir des formats de données brutes ou reconstruites. Dans le contexte de la grille LHC, les tâches d'analyse sont effectuées principalement dans les Tier-2s à l'exception de LHCb qui prévoit l'analyse dans les Tier-1s. Les expériences ont développé des outils pour aider les utilisateurs finaux à formuler et à soumettre leurs travaux d'analyse et pour s'interfacer avec le système de gestion des datasets.

Il faut prendre en compte les besoins d'analyse batch et interactifs. Pour les expériences qui utilisent le format Root, PROOF est probablement un bon candidat pour une ferme d'analyse interactive. L'autre question essentielle concerne la cohabitation avec la production Monte Carlo et les usages locaux.

D0 analysis farm

Tibor Kurča (IPNL)

Les flux de données D0 sont très similaires aux expériences LHC et l'infrastructure de gestions des données qui s'appuie sur SAM est déjà une infrastructure distribuée incluant un catalogue de meta-données. La grille D0, SAMGrid est interfacée avec LCG et OSG. Tibor Kurča rappelle les exigences de l'analyse de données très orientée entrée/sortie et dont il est assez difficile de prévoir les besoins en termes de ressources. Au-delà des ressources de calcul et de l'architecture des systèmes mis en place, d'autres critères sont à prendre en compte impérativement. La définition du format de données est essentielle pour l'efficacité de l'analyse. La facilité de prise en main des outils par les utilisateurs est également un facteur déterminant.

Some CMS grid analyses

Stijn De Weirdt (VUB)

L'exercice CSA06 réalisé par CMS a permis de tester les différents outils (DBS/DLS, CRAB) et différents scénarii d'analyse. L'objectif était à la fois de tester l'utilisation des Tier-2s pour l'analyse et d'entrainer les physiciens à l'utilisation des outils grille. Un processus (analysis workflow) très précis a été établi et suivi

(définition de filtres, transmissions aux Tier-2x, souscription des datasets et publication du filtre par les Tier-2s, réception des tâches à exécuter et exécution). Une phase de préparation interactive qui requiert des capacités Tier-3 moyennes est nécessaire. Des problèmes ont été identifiés notamment des problèmes de dépendances indésirables entre les différents composants (CRAB, DBS/DLS, reconstruction). L'objectif du futur CSA07 est d'améliorer l'implication des Tier-2s.

PROOF

Gerardo Ganis (CERN)

L'analyse consiste à traiter des lots d'événements indépendants qui peuvent être traitées en parallèle. Partant du constat que les outils d'analyse doivent être capables d'exploiter ce parallélisme intrinsèque tout en préservant l'interactivité, PROOF (Parallel ROOT Facility) permet de faire de l'analyse parallèle sur un cluster local. PROOF est un service capable de scinder les données, de paralléliser automatiquement les traitements sur ces données et de recombiner les résultats. PROOF s'appuie sur xrootd et est éminemment compatible avec ROOT. L'intégration de PRROF avec le software des expériences est possible et l'expérience d'Alice est très positive. Le déploiement effectuée au niveau de la CERN Analysis Facility montre que PRROF est une alternative à la grille dans les cas de traitements massifs où l'interactivité est importante.

Distributed analysis using GANGA

Dietrich Liko (CERN)

GANGA est un outil pour l'utilisateur final qui permet de définir et de gérer facilement les jobs sur la grille. GANGA est un co-développement d'ATLAS et de LHCb qui peut être interfacé avec les deux environnements expérimentaux (Gaudi/Dirac, Athena/DDM). Il permet également d'utiliser de manière combinée une ferme de calcul locale et la grille LCG. GANGA propose différentes interfaces (shell, script ou GUI). Chacune des deux expériences prévoit des plug-ins pour intégrer ses propres outils (AMI, DIRAC…). GANGA rassemble aujourd'hui près de 50 utilisateurs réguliers (300 personnes ont essayé GANGA !) issus des expériences ATLAS et LHCb mais également d'autres collaborations comme GEANT4, Compass…Dietrich Liko mentionne également l'activité au sein de OSG autour de Pathena/PANDA.

T2 set-up for end-users

Artem Trunov (CCIN2P3)

Un Tier-2 est typiquement dédié à l'analyse de données mais également destiné à servir plusieurs communautés d'utilisateurs. Aujourd'hui, si l'environnement grille ne satisfait pas pleinement l'utilisateur final, CMS a souhaité comprendre pourquoi et a réalisé une « enquête de satisfaction » auprès des utilisateurs. La majorité des utilisateurs utilisent la grille de préférence à leurs ressources batch locales mais en privilégiant un nombre de sites limités. L'outil CRAB (Cms Remote Analysis Builder) qui permet de soumettre sur la grille est utilisé par la plupart (6/10). Le Storage Element est clairement identifié comme première source d'erreurs mais l'efficacité globale de la grille est estimée entre 70 et 90 %. La moitié des personnes interrogées estime que l'étape de mise au point en local est utile mais la plupart se verrait bien accéder des ressources d'analyse hors grille. Dans un second temps, Artem Trunov développe différentes possibilités pour faciliter l'accès et la gestion des utilisateurs au niveau Tier-2/Tier-3. Depuis l'utilisation de gssissh, la mise en place d'espace de stockage répondant aux besoins des utilisateurs jusqu'à l'utilisation de xrootd en remplacement de SRM jugé trop lourd, Artem Trunov prône des solutions qui vont dans le sens de la simplification et de l'efficacité accrue des sites au bénéfice des utilisateurs finaux.

Conclusions du Colloque

Fairouz Malek remercie chaleureusement les participants, les orateurs et présidents de session, nos hôtes clermontois, Dominique Pallin en particulier ainsi que les membres du comité d'organisation. Elle fait remarquer que cette deuxième édition a été plus proche des utilisateurs et que la demande prononcée d'interaction entre sites et expériences a été visiblement accomplie. Elle espère que ce colloque aura profité à tous et aura permis de mettre en place les bases de travail pour avancer harmonieusement ensemble dans le but de rendre le calcul LHC efficace et l'analyse scientifique à la portée de tous avec des moyens et des outils performants.

François Lediberder tient à faire part de l'intérêt que la direction de l'IN2P3 accorde au projet de LCG-France et se réjouit de la participation, de la qualité de l'organisation, des interventions et des discussions.