Analyse des résultats des tests de recopie de données de HPSS vers dCache

(Test de reprocessing de CMS)

Relevé des Conclusions

21 mars 2008 – 14h

Présents :

- Andreï Moskalenko [AM]
- Lionel Schwarz [LS]
- Jonathan Schaeffer [JS]
- Philippe Gaillardon [PHG]
- Ghita Rahal [GR]
- Farida Fassi [FF]
- Nelli Pukhaeva [NP]
- Catherine Biscarat [CB]
- Jean-René Rouet [JRR]
- Fabio Hernandez [FH]

Président : FH

Secrétaire : FH

Agenda : http://indico.in2p3.fr/conferenceDisplay.py?confId=800

Début de réunion : 14h05

1. Introduction

Cette réunion a pour objet l'analyse des résultats observés lors de l'exercice de reprocessing effectué par CMS en février 2008, nécessitant la copie des données sur bande (gérées par HPSS) vers les disques (gérés par dCache) et la préparation d'un exercice comparable envisagé par Atlas.

FH rappelle les résultats observés par CMS et présentés par FF lors d'une réunion précédente et présente les spécifications des dérouleurs de bande utilisés par HPSS pour le stockage de données de CMS. Il en résulte que les performances de ce matériel, en plus des temps de montage, d'amorçage, de rembobinage et de démontage, sont fortement dominées par le temps de positionnement sur la cartouche. Le temps effectif de lecture d'un fichier de 1 GB correspond à 5% du temps total de l'opération, lorsque le montage spécifique d'une cartouche est nécessaire.

2. Résumé des discussions

Après discussion sur la validité du modèle théorique présenté, il paraît clair qu'afin d'optimiser les performances lors de la lecture des données sur cartouche magnétique, il est fortement souhaitable de regrouper la lecture de plusieurs fichiers lors d'un seul montage d'une cartouche.

Plusieurs solutions pour atteindre cet objectif ont été évoquées :

- Faire en sorte que les cartouches contiendront dans la mesure du possible des fichiers d'un même type (ex. RAW, AOD, ESD, ...), afin d'augmenter la probabilité que lors de la lecture massive de plusieurs de ces fichiers, un seul montage de la cartouche les contenants soit nécessaire.
  Ceci suppose que lors de la lecture de ces fichiers, une requête regroupant plusieurs fichiers à copier sur disque soit adressée à HPSS (par dCache, en l'occurrence) afin d'exploiter ses capacités d'ordonnancement, ce qui devrait permettre d'optimiser le nombre nécessaire de montages de cartouches.
  Des mécanismes pour obtenir l'écriture organisée de fichiers du même type ont été évoqués, notamment en transférant de données de dCache vers HPSS par tranches de temps.
- D'autre part, des expériences précédentes ont montré qu'il n'est pas réaliste de s'appuyer sur des conventions d'usage établies avec des utilisateurs finaux pour obtenir l'écriture organisée des données sur cartouches. Il a néanmoins été remarqué que dans le cas qui nous concerne, l'écriture des données dans HPSS est déclenchée par dCache, dont nous avons une certaine maîtrise pour l'adapter à nos besoins et à nos contraintes.
- La possibilité de modifier le dialogue entre dCache et HPSS a été aussi évoquée. Il s'agit dans ce cas, de faire en sorte que dCache interroge HPSS pour obtenir la localisation physique des fichiers qui devront être copiés sur disque (identifiant de la cartouche et position physique sur la cartouche) et formule une demande triée de lecture groupée de fichiers, en optimisant le nombre de montages de cartouches à effectuer par HPSS. Cette solution paraît plus compliquée à mettre en œuvre, d'une part, et tendrait à se substituer aux capacités d'optimisation déjà existantes dans HPSS.
- Une discussion sur l'intérêt de faire cette optimisation, compte tenue des capacités de disque dCache prévues pour les expériences LHC a suivi. Les experts des expériences ont rappelé que l'un des rôles centraux des tier-1s dans le modèle de traitement des données du LHC est de reprocesser les données récentes et celles des campagnes d'acquisition précédentes, qui seront stockées sur bande. La possibilité de maintenir sur disque pendant une période de 1 à 2 mois les données en provenance du tier-0 a été discutée. Dans ce modèle, la nécessité de copier sur disque les données sur cartouche serait limitée aux campagnes de reprocessing massif. Par ailleurs, il est rappelé que dans la configuration actuelle et pour les besoins de maintenir la disponibilité de l'activité de réception de données brutes en provenance du tier-0, le buffer de réception de ces données est séparé de celui utilisé pour alimenter les jobs de reprocessing s'exécutant localement au tier-1.
- Les experts dCache soulignent les difficultés prévisibles à organiser et grouper les requêtes d'écriture de données à destination de HPSS, par des critères temporels et par type de données. D'autre part, le nombre de flux simultanés entre dCache et HPSS devra être ajusté aux capacités des serveurs de disque et à leur connectivité réseau.
- Il est aussi rappelé que dans dCache les données d'une même expérience sont généralement organisées en arborescences et par type de données.
- Des questions sur le débit nécessaire pour satisfaire les activités de reprocessing des 4 expériences LHC ont été posées. La réponse à cette question dépend du nombre de jobs en simultanée et donc de la capacité CPU qui sera consacrée à cette activité. Une fourchette de valeurs pourrait être néanmoins estimée.

3. Conclusions

Suite à ces discussions, les actions ci-dessous ont été déclenchées :

- ACTION [Experts dCache] Evaluer l'adéquation du composant dCache existant qui permettrait de regrouper l'écriture des données (dCache → HPSS). L'ordonnanceur de ce système serait programmé pour intégrer les paramètres qui nous intéressent afin d'optimiser l'écriture et la lecture de données pour l'activité de reprocessing.

- ACTION [Experts dCache + experts HPSS] Réaliser un test de ce mécanisme avec une instance de test HPSS pour évaluer l'impact de cette solution sur les performances.
- ACTION [Experts dCache] Instrumenter les outils de copie de données entre dCache et HPSS afin de collecter des données sur les performances observées d'écriture (dCache → HPSS) et de lecture (HPSS → dCache) de données. Ces chiffres seront utilisés pour alimenter le système d'information en cours de développement (sous la responsabilité de JRR) et constitueront un cas d'utilisation concret.
- ACTION [FH et JRR] Collecter les informations disponibles actuellement permettant de quantifier les performances de HPSS, en particulier les fichiers de comptabilité HPSS et les fichiers d'accounting de RFIO. L'objectif est d'identifier et d'en extraire les informations pertinentes permettant de mesurer les performances de HPSS et d'autres métriques d'intérêt opérationnel pour la chaine d'accès aux données. Ces informations chiffrées serviront également à alimenter le système d'information.

Fin de réunion : 16h25