2009-07-03 TReqS - Etat d'Avancement et Plans

Résumé des conversations séparées entre Jonathan et Fabio d'une part, et Ghita et Fabio du 3 Juillet 2009 à propos de l'état d'avancement de TReqS et des plans pour la suite. Des modifications ont été apportées par Ghita.

- ## Intégration de TReqS et dCache
  - quelques pools dCache contenant des données ATLAS ont été activés pour utilisation du service TReqS cette semaine-ci. Néanmoins, peu d'activité observée, donc pas de conclusions.
  - des modifications ont été apportées au serveur TReqS pour améliorer sa stabilité
  - la configuration actuelle de TReqS en termes de nombre maximum de dérouleurs utilisables simultanément est comme suit:
    - ATLAS: 10 dérouleurs T10KA et 4 T10KB (chiffres à confirmer par JS)
    - CMS: 10 dérouleurs T10KA et 4 T10KB (chiffres à confirmer par JS)
  - actuellement le nombre maximum de dérouleurs par VO ne peut pas être modifié dynamiquement: il faut un redémarrage du serveur. Bien que cette opération n'ait pas d'impact sur les clients TReqS ni sur les requêtes en queue, il est souhaitable à l'avenir de pouvoir modifier ces paramètres plus facilement.
  - des pools dCache ont été configurés pour envoyer jusqu'à 200 requêtes de staging à TReqS chacun. Chacune de ces requêtes concerne le staging d'un fichier. Le serveur TReqS peut ainsi collecter toutes les requêtes de staging émises par l'ensemble des pools dCache afin de les ordonnancer. A titre d'exemple, pendant l'exercice de CMS STEP'09, chaque campagne journalière de staging demandait environ 2000 fichiers. En fonctionnement stable, on peut donc s'attendre à un nombre de requêtes de staging de l'ordre de 6000 pour les 4 expériences LHC.
  - le nombre total de requêtes RFIO servies simultanément par la machine qui sert de passerelle entre dCache et HPSS est actuellement de 80. C'est le nombre maximum de copies disque HPSS --> disque dCache à un instant donné. Cette limitation doit être ajustée en accord avec les autres paramètres de la chaîne. JS pense que, compte tenu de la durée de copie disque HPSS vers le disque dCache (de l'ordre de 5 à 10 secondes pour un fichier de taille typique), la valeur actuelle ne devrait pas être limitante.
  - JS est prêt à faire des tests en interne de staging en simultanée de ATLAS et CMS (et une troisième VO) la semaine du 6 juillet. L'intervention des experts ATLAS et CMS pour le choix d'un jeu de données représentatif est nécessaire. La suppression des fichiers utilisés pour le tests du disque dCache et HPSS sera effectuée.
  - synthèse du principe de fonctionnement entre dCachet et TReqS:
    - du point de vue du pool dCache, le staging de données de HPSS se passe en 3 phases:
      1. envoi de la requête de staging au serveur TReqS qui

la met en queue et l'ordonnance par rapport aux autres requêtes en queue, en particulier celles nécessitant la même cartouche
2. polling régulier de la base de données de TReqS pour connaître le moment où le fichier demandé a été stagé et se trouve sur le disque cache HPSS
3. copie du fichier depuis le disque cache HPSS vers le pool dCache demandeur (via "rfcp")

- ce mode de fonctionnement suppose qu'un nombre important de connexions simultanées sont nécessaires à la base de données de TReqS, pour satisfaire tous ses clients. La configuration actuelle permet 4500 connexions simultanées maximum (nombre à confirmer par JS).

Monitoring

- l'ensemble initial d'informations pour le monitoring temps réel du comportement de TReqS qui sont souhaitables avant de démarrer l'exercice de staging en interne sont:
  - évolution dans le temps du nombre de cartouches en cours de traitement (par VO et aggregée)
  - évolution dans le temps du nombre de cartouches en attente de traitement (par VO et aggregée)
  - évolution dans le temps du nombre de fichiers restant à stager (par VO et aggregée)
  - évolution dans le temps du volume de données restant à stager (par VO et aggregée)
- ces informations se trouvent dans la base de données du serveur TReqS. JS va faire le nécessaire pour les extraire afin de les visualiser via Smurf ou SYMOD, ce qui sera le plus rapide/facile.
- en fonction de l'expérience acquise, d'autres métriques seront rajoutées à cet ensemble

Collecte de données pour l'analyse à posteriori du comportement de TReqS

- *[ces informations ont été extraites d'une conversation avec Ghita]*
- d'après l'expérience pendant STEP'09, les données nécessaires pour l'analyse à posteriori sont:
  - via ACSLS
    - heure de montage et de démontage de chaque cartouche (information fournie par Suzanne)
  - via TReqS, pour chaque fichier demandé en staging
    1. identifiant de la cartouche
    2. nom du fichier
    3. taille du fichier (en MB)
    4. position du fichier sur la cartouche
    5. temps d'envoi de la requête à HPSS
    6. temps de fin de traitement (fichier copié sur le disque HPSS)
  - via TReqS, pour chaque bande impliquée dans une requête de staging
    1. le temps total nécessaire pour lire tous les fichiers

demandés de la bande
- via dCache (où tourne le client TReqS), pour chaque fichier demandé en staging
  1. le temps de service total, c'est à dire, à partir du moment où dCache émet la demande de fichier jusqu'au moment où le fichier est effectivement copié et disponible sur le pool dCache

- **Plans à court terme**
  - JS va documenter et briefer Yvan et Andrés sur le fonctionnement de TReqS afin de faire en sorte qu'il y ait des personnes capables d'intervenir pendant son absence, en particulier pendant les congés d'été
  - réaliser les tests internes de staging en simultanée ATLAS et CMS (semaine du 6 juillet)
  - réaliser les tests de staging en simultanée ATLAS et CMS, avec l'intervention des experts de ces expériences externes au CC (mois d'août, date à fixer en fonction des disponibilités des experts nécessaires à l'opération)
- **Plans à moyen terme**
  - JS pense qu'une ré-écriture complète de TReqS est nécessaire pour des besoins de maintenance du produit. Un nombre significatif de modifications ont été apportées au produit initial, y compris l'intégration d'une base de données, qui rend son architecture plus complexe que nécessaire. Le temps nécessaire pour faire ce travail est estimé à 1 à 2 mois.
  - JS souhaite profiter des connaissances de Andrés Gomez pour ce travail de refonte lorsqu'il sera entrepris.