HEPiX FSWG : comparaison de solutions de stockage

Loïc Tortay, 24 juin 2008
Vidéo-conférence EGEE SA1 France

dapnia
cea
saclay

CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

HEPiX FSWG

- Groupe de travail constitué fin 2006 (IHEPCCC)
- Organisations participantes: CEA, CERN, DESY, FZK, CC-IN2P3, INFN, LAL, NERSC, RAL, RZG, SLAC, ...
- Objectif initial (atteint en 2007): étudier les solutions de stockage existantes et préconiser des choix et bonnes pratiques (principalement pour le contexte HEP)
- Objectif secondaire : évaluation concurrente de différents systèmes de stockage (pour les Tier-2)

HEPiX FWSG : Tests

- Deux types principaux de technologies :
  - Systèmes de fichiers distribués (accès POSIX transparent)
  - Applications spécifiques/systèmes de fichiers applicatifs
- Choix d'évaluer **sur le même matériel** avec **les mêmes jobs** :
  - AFS (référence), GPFS, Lustre
  - dCache, DPM, Xrootd

HEPiX FSWG : Environnement de tests

- Tests au CERN début 2008
- Environnement *typique* d'un Tier-2 :
  - ~10 serveurs de disques
  - ~500 jobs simultanés (~60 clients)
  - Réseau gigabit Ethernet (clients & serveurs)
- Programmes de tests fournis par G. Cowan (Edinbourg)
- Chaque application installée et configurée par des experts

HEPiX FSWG : Configuration des tests

- Chaque serveur fournit un espace « distinct » avec ~5k fichiers => ~50k fichiers de 300 Mo (AOD)
- ~450k petits fichiers créés en plus pour « ajouter du bruit »
- Distribution des clients sur les serveurs simple/statique (6 clients/serveur)
- Tests d'écriture (cf. supra), de lecture séquentielle, puis de lecture aléatoire
- Pour chaque test :
  - nombre de jobs simultanés croissant (10 à 480)
  - *runs* de 45 minutes

HEPiX FSWG : Résultats

- Tous les systèmes testés sont capables de recevoir les données (créées/écrites par les clients) à ~1 Gb/s par serveur (y compris AFS)
- Lecture séquentielle (blocs de 1 Mo) :
  - 10 jobs : AFS < DPM < dCache << Xrootd < Lustre < GPFS
  - 100 jobs : Xrootd < DPM < dCache < AFS < Lustre < GPFS
  - 480 jobs : Xrootd << dCache < DPM < AFS < GPFS < Lustre
- Lecture *aléatoire* (blocs de 10, 25, 50 et 100 Ko) :
  - 100 jobs : Xrootd (!) < dCache < DPM << AFS << Lustre < GPFS
  - 200 jobs : AFS < Xrootd < dCache < DPM << GPFS << Lustre
  - 480 jobs : AFS << Xrootd < dCache < DPM < GPFS << Lustre

HEPiX FSWG : Conclusions/recommandations

- Lustre & GPFS dominent en termes de performances
- L'accès POSIX simplifie beaucoup la vie des utilisateurs
- Le groupe de travail recommande Lustre pour les *gros* espaces de travail partagés
- SRM+Lustre (StoRM) peut être la solution de convergence

Liens

- HEPiX FSWG : http://hepix.caspur.it/storage/
- Rapport final :
  http://indico.cern.ch/contributionDisplay.py?contribId=28&sessionId=10&confId=27391