GPFS au CC-IN2P3

Loïc Tortay, 8 février 2008

dapnia
cea
saclay

CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

GPFS : General Parallel File System

- Produit IBM, système de fichiers de cluster avec des fonctionnalités « larges » (haute disponibilité & HPC)
- Système de fichiers « officiel » des machines parallèles IBM (SP, p690 & plus, BlueGene)
- Certifications pour clusters de bases de données Oracle et DB2 (IBM)
- Beaucoup de sites HPC/recherche publique, parmi lesquels : FZK, NDGF (NorduGrid), NERSC, DEISA (IDRIS, MPG Juelich, MPG Garching, Cineca), TeraGrid (SDSC, NCSA, ...)

GPFS : terminologie/concepts

- Cluster : ensemble de machines sous une même autorité administrative qui accèdent aux mêmes système de fichiers
- Nœud : machine du cluster
- NSD (*Network Shared Device*) : volume disque rendu accessible par le réseau par GPFS
- I/O Node : nœud qui a des disques et les rend accessibles pour GPFS
- Quorum : attribut permanent d'un nœud pour la gestion des métadonnées, les N nœuds *quorum* votent et doivent atteindre le *quorum* (N/2+1)
- Manager : attribut semi-dynamique d'un nœud pour la gestion d'un système de fichier (verrous) (distributed token manager -- 3.x)

GPFS : Communication inter-nœuds

- GPFS choisit le chemin le plus court vers les données : accès direct aux disques ou réseau (NSD)
- Communication TCP/IP uniquement, indépendant de la technologie
- Ethernet (Gb+), IP over Myrinet, Federation & BlueGene officiellement supportés (mais pas InfiniBand avant 3.2)
- Communication *administrative* inter-nœuds basée sur RSH (ou équivalent) *sans mot de passe* pour l'utilisateur ***root***

GPFS : Fonctionnalités (1)

- Espace de noms unique pour un système de fichiers
- Système de fichiers POSIX, accès transparent
- 32 systèmes de fichiers/cluster (256 avec 3.2)
- 200 To/système de fichiers
- « striping » des données sur les différents NSD (« round-robin » par défaut)
- Copies multiples optionnelles des méta-données (2)
- Copies multiples optionnelles des données (2)
- Support intégré de la haute-disponiblité des NSD (optionnel)

GPFS : Fonctionnalités (2)

- Configuration centralisée ***et*** distribuée
- ACLs (POSIX & NFSv4)
- Quotas explicites et par défaut pour utilisateur et groupe
- Authentification Unix (UID)
- Accès *distant* : « multi-cluster »
- Snapshots
- Interface DMAPI
- Client Linux : module noyau (OpenSource) + démon
- Supporte AIX & Linux (32 et 64 bits) dans un même cluster
- filesets (3.x)
- Clustered NFS (3.2)

GPFS : Architectures supportées

- HA SAN simple : deux ou trois nœuds, typiquement SGBD
- HA SAN : > 3 nœuds
- HA sans SAN : 3 nœuds ou plus (DAS)
- HPC « globale » : full-SAN/Federation
- HPC « intermédiare » : (minis-)SAN pour les I/O nodes
- HPC distribuée : sans SAN, DAS (*3ware @NERSC*)
- HPC mixte : SAN + DAS
- Multi-clusters

GPFS : Versions

- Une version majeure pour l'ensemble du cluster
- Les versions mineures peuvent coexister dans un cluster
- 2.3 : CC-IN2P3, stable
- 3.2 : mars 2008
  - pools & filesets
  - ILM (gestion du cycle de vie des données), langage SQL-like
  - quotas sur les filesets
  - Liaison transparente vers HPSS basée sur les fonctionnalités ILM de 3.1 + un peu de DMAPI

GPFS : Licences & coût

- GPFS n'est pas OpenSource
- Gratuit pour les sites « éducation/recherche » *sisi* délivrance de diplômes (éligibilité *ScholarPack*)
- Sinon, IBM facture au CPU
- Licence site IN2P3 (+DAPNIA), coût forfaitaire jusqu'à 1500 nœuds pris en charge par le CC
- Formation IBM requise pour accéder au support

GPFS : @CC-IN2P3

- 3 clusters :
  - « test » : 3.2, douzaine de nœuds (SL3, SL4, 32 & 64 bits), DAS
  - « production » : 2.3.0-12, 1300 nœuds, 14 I/O nodes, 2 serveurs quorum only, 140 To de disques, serveurs SL4 x64, clients SL3 32 bits et SL4 32 & 64 bits, 60 millions de fichiers
  - « LCG » : 2.3.0-12, demi-douzaine de nœuds, SL3, HA sans SAN
- Objectif mars 2008 :
  - « production » : 3.2
  - 35 I/O nodes, 3 serveurs quorum only, 1500 noeuds,
  - 940 To de disques
  - ILM (pools/filesets)
  - Monitoring SNMP agentless (3.x)
- Plus tard : 3.2 & liaison transparente avec HPSS

GPFS : Bonnes pratiques

- 7 nœuds quorum au maximum
- quotas à 90%
- méta-données sur du RAID-1 ou disques rapides (FC ou SAS, >= 10 krpm)
- Communication administrative inter-nœuds :
  - SSH par clefs
  - modèle standard (*all-to-all*)
  - modèle restreint (*few-to-all*), *non supporté*
  - modèle pseudo-distant/segmentation (NERSC)

GPFS vs Lustre

- Commercial vs OpenSource + support payant
- Client poids moyen vs Client lourd (noyau spécifique)
- Client poids moyen vs *Client léger (module noyau+ zéro configuration)*
- Modèle d'administration intégré vs Modèle initialement peu clair (fichier de configuration XML à partager par NFS), maintenant serveur de « management » ?
- X Po vs Y Po (Y > X)
- Lien HPSS à venir vs lien HPSS (à venir)
- TCP/IP (ou Federation ?) vs TCP/IP ou LNET (InfiniBand, Quadrics Elan, ...)
- 200 To/système de fichiers vs *X* Po/système de fichiers (1.3 Po @DAM)

GPFS vs Lustre (2)

- Centaines à millier de nœuds/cluster vs dizaines de milliers de nœuds/cluster (BlueGene)
- Authentification Unix vs Authentification Unix ou Kerberos (à venir)
- Accès distants (WAN) disponible et utilisé (TeraGrid) vs à venir (roadmap + Carriocas)
- Méta-données distribuées (quorum+manager+disques) vs MDS (multiples à venir)
- Réplication intégrée optionnelle des données & méta-données vs rien
- Distribution des données (striping) intégrée vs à venir
- Rien vs Lustre *routers* (et interfaces/réseaux multiples)
- IBM vs Sun

Configuration CC-IN2P3

- 2 serveurs se partagent ~20 To de disques : mini-SAN 2 nœuds (2 x 1 Gbps + 2 HBAs FC2 par serveur);
- chaque serveur est serveur primaire pour la moitié des disques et secondaire pour l'autre moitié (et vice versa);
- 7 couples de serveurs;
- > 1000 NSD;
- granularité 230 Go ou 1.45 To (4 To)
- 10 systèmes de fichiers (3 To à 43.5 To)
- 60 jobs de Planck : ~480 Mo/s; 600 jobs de BaBar : ~400 Mo/s, 120 jobs de SNLS : ~640 Mo/s
- A venir : 2 serveurs se partagent 80 To de disques (10 Gbps + 2 HBAs FC4/serveur) (x10)