

GPFS (General Parallel File System)

Généralités

- <http://fr.wikipedia.org/wiki/GPFS>

Version GPFS couvertes par cette documentation :

- 3.2.1-2
- 3.3

Documentation

- Documentations officielles IBM :
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=%2Fcom.ibm.cluster.gpfs.31.advanceadm.doc%2Fbl1adv_chnamip.html

Scénario

Matériel.

- 4 serveurs et 4 clients.
- 4 serveurs dont 1 serveur principal et 1 serveur secondaire GPFS qui sont tous quorum, et 4 postes membres du cluster sans être quorum (mais cela pourrait et serait conseillé).


Les serveurs ont accès aux volumes en accès SAN alors que les postes accèdent aux volumes via le réseau ethernet.

Toutes les étapes suivantes nécessite d'être root sur les machines et décrivent dans l'ordre la création d'un cluster complet.

Mise en place des clés SSH

Le bon fonctionnement du cluster GPFS repose sur les communications SSH en clés privées / clés publiques en root. Il faut donc générer déployer les clés SSH de la manière suivante.

Tableau disposition clés SSH

 tableau disposition clés SSH

- Serveur principal : clés publiques de toutes les machines du cluster y compris du serveur secondaire et de lui-même dans son ~/.ssh/authorized_keys.
- Serveur secondaire : clés publiques de toutes les machines du cluster y compris du serveur

principal et de lui-même dans son `~/.ssh/authorized_keys`.

- autres noeuds du cluster : clés publiques uniquement du serveur principal et secondaire dans les `~/.ssh/authorized_keys`.

Générer le couple clé publique/clé privée SSH

```
ssh-keygen -t rsa -b 4096
```

```
Generating public/private rsa key pair.
```

```
Enter file in which to save the key (/root/.ssh/id_rsa):
```

```
Enter passphrase (empty for no passphrase):
```

```
Enter same passphrase again:
```

```
Your identification has been saved in /root/.ssh/id_rsa.
```

```
Your public key has been saved in /root/.ssh/id_rsa.pub.
```

Ne pas entrer de passphrase parce que les connexions doivent être directes entre les machines du cluster.

Les clés privées/publiques sur chaque lame sont stockées dans `/root/.ssh/` dans les fichiers suivants.

- `id_rsa` pour la clé privée.
- `id_rsa.pub` pour la clé publique.

Vérifier et éventuellement positionner les droits 600 pour la clé privée et 644 pour la clé publique.

```
chmod 600 /root/.ssh/id_rsa  
chmod 644 /root/.ssh/id_rsa.pub
```

Copier ensuite le contenu des fichiers `/root/.ssh/id_rsa.pub` de chaque lame dans le fichier `/root/.ssh/authorized_keys` des machines correspondant aux serveurs primaire et secondaire et valider la connexion ssh.

Sur toutes les machines autres que le serveur principal et secondaire.

```
ssh-copy-id -i /root/.ssh/id_rsa.pub root@srvprincipal.domaine.fr
```

```
ssh-copy-id -i /root/.ssh/id_rsa.pub root@srvsecondaire.domaine.fr
```

```
ssh root@srvprincipal.domaine.fr (Répondre yes sur la question éventuelle,  
une fois la connexion réussie, faite "Ctrl + D" pour se déconnecter).
```

```
ssh root@srvsecondaire.domaine.fr (Répondre yes sur la question éventuelle,  
une fois la connexion réussie, faite "Ctrl + D" pour se déconnecter).
```

Copier le contenu des fichiers `/root/.ssh/id_rsa.pub` des serveurs primaire et secondaire dans le fichier `/root/.ssh/authorized_keys` des machines correspondantes aux autres serveurs faisant parti du cluster GPFS et sur lui-même.

Depuis le serveur primaire du cluster GPFS (srvprincipal).

```
ssh-copy-id -i /root/.ssh/id_rsa.pub root@<machines>.domaine.fr
ssh-copy-id -i /root/.ssh/id_rsa.pub root@srvsecondaire.domaine.fr
ssh-copy-id -i /root/.ssh/id_rsa.pub root@srvprincipal.domaine.fr
```

Depuis le serveur secondaire du cluster GPFS (srvsecondaire).

```
ssh-copy-id -i /root/.ssh/id_rsa.pub root@<machines>.domaine.fr
ssh-copy-id -i /root/.ssh/id_rsa.pub root@srvprincipal.domaine.fr
ssh-copy-id -i /root/.ssh/id_rsa.pub root@srvsecondaire.domaine.fr
```

Création du cluster GPFS

Maintenant que les clés publiques sont partagées, le cluster GPFS peut être créé.

Attention : vérifier au préalable que le nom de la machine (sur les machines du cluster) n'est pas déclaré dans le fichier `/etc/hosts` sur la ligne localhost, sinon enlever le nom du fichier. Il est nécessaire à cette étape que les nodes à intégrer au cluster soient accessible par leurs noms DNS.

Sur le serveur primaire (srvprincipal), créer un fichier décrivant le cluster comme suit.

```
vi /tmp/nodes.txt
```

```
srvprincipal.domaine.fr:quorum
srv3.domaine.fr:quorum
srvsecondaire.domaine.fr:quorum
srv4.domaine.fr:quorum
```

Remarque sur le mode quorum et manager.

Type possible : quorum, quorum manager, manager

- Le quorum maintien le cluster actif. Pour qu'il le reste il faut que la moitié + 1 des machines quorum soit active sinon le cluster s'arrête.
- Une machine quorum participe au bon fonctionnement du cluster.
- Une machine manager supporte le fonctionnement d'un nsd (= un volume qui est supporté par un couple de noeud ou machines).
- Une machine peut être les deux (quorum-manager).

Créer le cluster GPFS via la commande suivante.

```
mmdircluster -N /tmp/nodes.txt -p srvprincipal.domaine.fr -s
```

```
srvsecondaire.domaine.fr -C "CLUSTER_PROD" -A -r /usr/bin/ssh -R /usr/bin/scp
```

Une fois le cluster créé, vérifier sa bonne création via la commande suivante.

```
mmlscluster
```

```
GPFS cluster information
```

```
=====
```

```
GPFS cluster name:      CLUSTER_PROD.domaine.fr
GPFS cluster id:        13882477815451679276
GPFS UID domain:        CLUSTER_PROD.domaine.fr
Remote shell command:   /usr/bin/ssh
Remote file copy command: /usr/bin/scp
```

```
GPFS cluster configuration servers:
```

```
-----
```

```
Primary server:  srvprincipal.domaine.fr
Secondary server: srvsecondaire.domaine.fr
```

```
Node  Daemon node name          IP address      Admin node name
Designation
```

```
-----
```

Node	Daemon node name	IP address	Admin node name
1	srvprincipal.domaine.fr	192.168.0.1	
	srvprincipal.domaine.fr	quorum	
2	srv3.domaine.fr	192.168.0.3	srv3.domaine.fr
	quorum		
3	srvsecondaire.domaine.fr	192.168.0.2	
	srvsecondaire.domaine.fr	quorum	
4	srv4.domaine.fr	192.168.0.4	srv4.domaine.fr

Démarrer le cluster via la commande.

```
mmstartup -a
```

```
mmstartup: Starting GPFS ...
```

Vérifier sur toutes les lames que GPFS tourne bien.

```
ps aux | grep gpfs
```

```
root      18353      1  0 06:36 ?          00:00:00 [gpfsSwapdKproc]
```

Sur le serveur principal, vérifier l'état de tous les nodes du cluster via la commande suivante.

```
mmgetstate -a -L
```

Création d'un nsd

Il faut maintenant créer et ajouter le volume à partager en GPFS. Sur le serveur primaire du cluster GPFS (srvprincipal), créer le disque correspondant au disque GPFS (LUN=2 sur le SAN qui correspond à /dev/sdd sur la machine) en suivant la procédure ci-dessous.

Créer le fichier /tmp/nsd.txt et ajouter la ligne suivante.

```
vi /tmp/nsd.txt
```

```
/dev/sdd:srvprincipal.domaine.fr:srvsecondaire.domaine.fr:dataAndMetadata:-1  
:data_work
```

Créer et ajouter le disque GPFS via la commande.

```
mmcrnsd -F /tmp/nsd.txt -v no
```

```
mmcrnsd: Processing disk sdd
```

```
mmcrnsd: Propagating the cluster configuration data to all affected nodes.  
This is an asynchronous process.
```

Vérifier la bonne création du disque GPFS.

```
mmlsnsd
```

```
File system   Disk name   NSD servers
```

```
-----  
(free disk)  data_work  srvprincipal.domaine.fr,srvsecondaire.domaine.fr
```

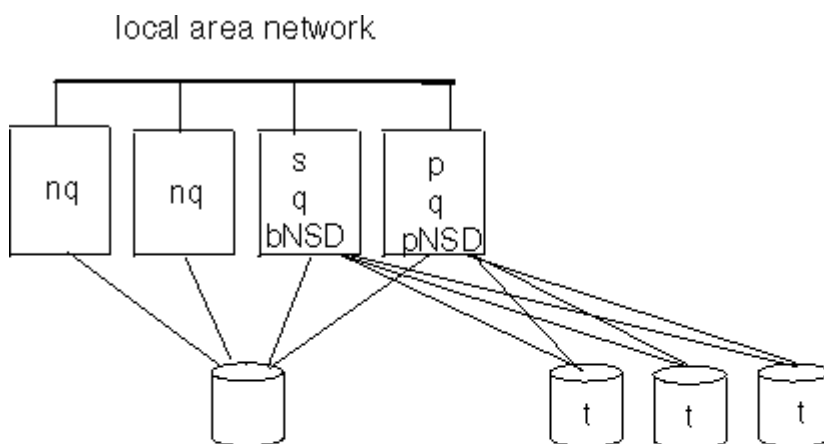
Ajout éventuel d'un tiebreaker disk

Le tiebreaker disk permet de s'affranchir de la limite du $(nb\ node/2)+1$ quorum en fonctionnement qui doit fonctionner pour faire tourner le cluster. Dans le cas d'un cluster GPFS qui est fonctionnel avec uniquement un serveur GPFS principal et un secondaire, seul deux nodes quorum maintiennent le cluster. Si on redémarre un des serveurs le cluster tombe. Pour éviter ce phénomène, on peut utiliser un tiebreaker disk qui permet de faire fonctionner le cluster avec un seul node quorum actif. Le tiebreaker se chargeant de maintenir le cluster avec le second node encore actif.

Pour ajouter un tiebreaker disk sur trois nsd (le maximum).

```
# mmchconfig tiebreakerdisks="nsd01;nsd02;nsd03"
Verifying GPFS is stopped on all nodes ...
mmchconfig: Command successfully completed
mmchconfig: Propagating the cluster configuration data to all affected
nodes. This is an asynchronous process.
```

Dans le schéma qui suit, GPFS reste actif avec le minimum de node quorum actif et deux tiebreaker disks disponibles.



p - primary cluster configuration server
 s - secondary cluster configuration server
 q - quorum node
 nq - non-quorum node

pNSD - primary NSD server
 bNSD - backup NSD server
 t - tiebreaker disk

Création du FileSystem GPFS

Maintenant que le disque GPFS est créé et partagé, il faut créer le filesystem correspondant.

Dupliquer le fichier `/tmp/nsd.txt` et renommer le fichier dupliqué en `/tmp/fs.txt`.

```
cat /tmp/nsd.txt > /tmp/fs.txt
```

Créer le filesystem via la commande.

```
mmdirfs /dev/data_work -F /tmp/fs.txt -A yes -T /data_work -v no
```

The following disks of data_work will be formatted on node srvprincipal.domaine.fr:

data_work: size 52428800 KB

Formatting file system ...

Disks up to size 111 GB can be added to storage pool 'system'.

Creating Inode File

Creating Allocation Maps

Clearing Inode Allocation Map

Clearing Block Allocation Map

Formatting Allocation Map for storage pool 'system'

Completed creation of file system /dev/data_work.

mmdirfs: Propagating the cluster configuration data to all affected nodes. This is an asynchronous process.

Vérifier la bonne création du filesystem GPFS data_work.

```
mmlsfs /dev/data_work
```

flag	value	description
-f	8192	Minimum fragment size in bytes
-i	512	Inode size in bytes
-I	16384	Indirect block size in bytes
-m	1	Default number of metadata replicas
-M	2	Maximum number of metadata replicas
-r	1	Default number of data replicas
-R	2	Maximum number of data replicas
-j	cluster	Block allocation type

```
-D nfs4          File locking semantics in effect
-k all          ACL semantics in effect
-a 1048576      Estimated average file size
-n 32          Estimated number of nodes that will mount file system
-B 262144      Block size
-Q none        Quotas enforced
               none        Default quotas enabled
-F 51712       Maximum number of inodes
-V 10.00 (3.2.0.0) File system version
-u yes         Support for large LUNs?
-z no         Is DMAPI enabled?
-L 4194304     Logfile size
-E yes         Exact mtime mount option
-S no         Suppress atime mount option
-K whenpossible Strict replica allocation option
-P system      Disk storage pools in file system
-d data_work   Disks in file system
-A yes         Automatic mount option
-o none        Additional mount options
-T /data_work  Default mount point
```

Vérifier que le fichier `/etc/fstab` contient bien la ligne suivante.

```
cat /etc/fstab
```

```
[...]
```

```
/dev/data_work    /data_work        gpfs
rw,mtime,atime,dev=data_work,autostart 0 0
```

```
[...]
```


Sur le serveur primaire du cluster, effectuer le montage GPFS sur la totalité du cluster.

```
mmmount all -a
```

Vérifier que le montage GPFS est bien monté sur chaque machine.

```
mount
```

```
/dev/data_work on /data_work type gpfs (rw,mtime,dev=data_work,autostart)
```

Sur le serveur primaire du cluster GPFS (srvprincipal), définir la plage de port de communication TCP.

```
mmchconfig tscCmdPortRange="35000-35200"
```

Vérifier que la plage de port TCP a bien été prise en compte sur chaque lame faisant parti du cluster GPFS

```
mm lsconfig
```

```
tscCmdPortRange 35000-35200
```

Afin de prendre en compte de façon effective la plage de port TCP défini ci-dessus, arrêter et redémarrer le cluster GPFS sur le serveur primaire (srvprincipal)

```
mmshutdown -a  
mmstartup -a
```

Log cluster

Le fichier de log du cluster se trouve dans le fichier suivant. C'est un lien symbolique qui pointe vers le dernier log.

```
/var/adm/ras/mmfs.log.latest
```

Ajout machines au cluster

Prenons le cas d'un poste qu'on veut ajouter au cluster GPFS créé préalablement afin qu'il accède au volume data_work.

Générer le couple clé publique/clé privée SSH

Se connecter en root sur le poste. Générer les couples clé privées/publiques SSH via la commande.

```
ssh-keygen -b 4096 -q -t rsa
```

```
Generating public/private rsa key pair.
```

```
Enter file in which to save the key (/root/.ssh/id_rsa):
```

```
Enter passphrase (empty for no passphrase):
```

```
Enter same passphrase again:
```

```
Your identification has been saved in /root/.ssh/id_rsa.
```

```
Your public key has been saved in /root/.ssh/id_rsa.pub.
```

Ne pas entrer de passphrase car les machines du cluster GPFS doivent pouvoir communiquer directement.

Vérifier et éventuellement positionner les droits 600 pour la clé privée et 644 pour la clé publique.

```
chmod 600 /root/.ssh/id_rsa  
chmod 644 /root/.ssh/id_rsa.pub
```

Copier ensuite le contenu du fichier `/root/.ssh/id_rsa.pub` du poste dans le fichier `/root/.ssh/authorized_keys` des serveurs principal et secondaire GPFS (srvprincipal et srvsecondaire) et valider la connexion ssh.

Sur le poste opérateur.

```
ssh-copy-id -i /root/.ssh/id_rsa.pub root@srvprincipal.domaine.fr  
ssh-copy-id -i /root/.ssh/id_rsa.pub root@srvsecondaire.domaine.fr  
ssh root@srvprincipal.domaine.fr (répondre yes sur la question éventuelle).  
ssh root@srvsecondaire.domaine.fr (Une fois la connexion réussie, faite  
"Ctrl + D" pour vous déconnecter).
```

Copier le contenu des fichiers `/root/.ssh/id_rsa.pub` des serveurs principal et secondaire dans le fichier `/root/.ssh/authorized_keys` du poste opérateur.

Depuis le serveur primaire du cluster GPFS (srvprincipal).

```
ssh-copy-id -i /root/.ssh/id_rsa.pub root@<nom_poste>.domaine.fr
```

Depuis le serveur secondaire du cluster GPFS (srvsecondaire).

```
ssh-copy-id -i /root/.ssh/id_rsa.pub root@<nom_poste>.domaine.fr
```

Intégration au cluster

Se connecter sur le serveur primaire du cluster GPFS (srvprincipal) en root.

Intégrer le poste opérateur au sein du cluster GPFS.

```
mmaddnode <nom_machine>.domaine.fr
```

Vérifier la bonne intégration du poste opérateur.

```
mmlsnode
```

```
Renvoie CLUSTER_PROD avec le nouveau ajouté en fin de liste.
```

Montage du volume

Sur le poste opérateur, vérifier le contenu du fichier /etc/fstab.

```
cat /etc/fstab
```

Le montage /dev/data_work doit apparaître en fin de fichier.

```
[...]  
  
/dev/data_work      /data_work          gpfs  
rw,mtime,atime,dev=data_work,autostart 0 0  
  
[...]
```

Sur le poste opérateur, démarrer GPFS et vérifier son bon démarrage.

```
mmstartup
```

```
ps -aef | grep gpfs
```

```
root      3674      1  0 12:05 ?          00:00:00 [gpfsSwapdKproc]
```

Sur le poste opérateur, monter le filesystem.

```
mmmount all
```

Vérifier le bon montage du filesystem /dev/data_work.

```
mount
```

Sur le poste opérateur, paramétrer le démarrage de GPFS au boot de la machine.

```
chkconfig gpfs on  
chkconfig --list gpfs
```

Sur le poste opérateur, vérifier dans la configuration GPFS l'option autoLoad.

```
mmlsconfig ----> l'option "autoload" doit être égale à "yes".
```

Ignorer le montage d'un volume pour certaines machines

Il est possible de sélectionner les volumes qui doivent être montés en fonction de la machine. Il arrive des cas où l'on ne souhaite pas monter certains volumes sur certaines machines qui font parties d'un même cluster et qui ont donc accès à de nombreux volumes.

Pour cela, se rendre sur chaque machine où l'on souhaite restreindre les montages et configurer les volumes (qui sont alloués à toutes les machines du cluster par défaut) pour qu'ils ne soient pas montés automatiquement.


La simple présence d'un fichier (vide) avec la bonne syntaxe pour chaque volume suffit à désactiver au démarrage le volume nommé.

```
/var/mmfs/etc/ignoreStartupMount.<nom_volume>
```

Éventuellement ajouter dans le fichier un commentaire de la fonction du fichier.

```
Disable automount for <nom_volume>
```

Suppression d'un node (machine HS)

 cette méthode est soumise à caution et n'a pas été testée par mes soins.

S'assurer que la machine n'est plus sur le réseau (plus joignable), et qu'elle n'est pas quorum, ni manager de volumes.

Sur le serveur principal GPFS, supprimer le node.

```
mmdeinode -N <nom_du_serveur>
```

Une méthode pas des plus propre mais fonctionnelle permet de supprimer un node même lorsque la commande `mmdeinode` ne fonctionne pas.

Il faut supprimer le node en question du fichier `/var/mmfs/gen/mmsdrfs` sur le serveur principal et certainement redémarrer le cluster GPFS. Attention, action à réaliser avec prudence.

Après une réinstallation du poste, supprimer les anciennes clés SSH dans `authorized_keys` et le contenu des fichiers `known_hosts` qui concernent la machine défailante, puis rejouer la création et le positionnement d'une nouvelle clé générée.

Suppression d'un NSD et d'un Filesystem

Ce paragraphe a été créé lors de la nécessité d'intervertir deux volumes SAN utilisés par deux nsd GPFS. Lors de la création une inversion a été opérée. Les tailles des volumes n'était pas bonnes.

Arrêter les services qui se servent des volumes GPFS sur les machines en question.

Sauvegarder l'existant.

```
mmlsfs all >> ./save_fs.txt  
mmlnsd >> ./save_fs.txt
```

Sur le serveur principal GPFS.

```
mmumount all -N all
```

Vérifier sur chaque noeud que cela est bien démonté avec la commande mount.

```
mmdelfs /dev/DATA  
mmdelfs /dev/WSreq
```

```
mmdelnsd DATA  
mmdelnsd WSreq
```

Créer un fichier de création des nsd correspondant à ce qu'on veut obtenir.

```
vi DATA.nsd  
  
/dev/sdc:<nom_noeud1_fqdn>:<nom_noeud2_fqdn>:dataAndMetadata: -1:DATA
```

Remarque. Mettre autant de noeud (machine) qu'on souhaite si elles ont besoin d'avoir accès au volume.

Création du nsd.

```
mmcrnsd -F DATA.nsd
```

Création du filesystem GPFS DATA sur trois noeuds (machine).

```
mmcrfs /DATA /dev/DATA -F DATA.nsd -n 3 -A yes
```

Pour changer le point de montage en /DONNEES.

```
mmchfs /dev/DATA /DONNEES
```

Remonter le volume sur le cluster.

```
mmmound all
```

Cas particulier de réinstallation ou migration d'un serveur

Lors de la réinstallation ou de la migration d'un serveur (ex : passer de la RHEL4 à la RHEL5 sans "casser" le cluster GPFS et les données des NSD), son intégration dans GPFS doit être refaite. Or il est connu du cluster et existe dans un ou plusieurs NSD.

Il doit sortir des NSD dont il est le serveur primaire ou backup. Puis il doit sortir du cluster.

Enfin on réintègre dans le cluster et on modifie les NSD dont il doit être le serveur primaire ou backup.

Sortir un serveur d'un NSD

Démonter de tous les serveurs le NSD.

```
mmumount DATA -a
```

où DATA est le nom du NSD concerné.

Modifier le fichier desc_file.NSD pour remplacer le "nom du serveur" par un autre serveur (s'il est primaire) ou le retirer (s'il est backup).

Lancer la commande pour modifier la configuration du NSD.

```
mmchnsd -F desc_file.nsd
```

Vérifier la bonne prise en compte.

```
mm\lsnsd
```

Arreter GPFS sur tous les serveurs du cluster. Se positionner pour cela sur le serveur de cluster.

```
mmshutdown -a
```

Sortir le "nom du serveur" du cluster

Pour empêcher la communication entre le serveur de cluster et le "nom du serveur" que l'on souhaite installer, il faut arrêter les services réseau sur le "nom du serveur" à installer.

```
/etc/init.d/network stop
```

Sortir "nom du serveur" du quorum

Si le "nom du serveur" fait partie du quorum, il faut le sortir du quorum. Se positionner pour cela sur le serveur de cluster.

```
mmchconfig designation=nonquorum <nom_du_serveur>
```

Suppression d'un node (machine toujours online)

Règles à suivre pour la suppression d'un node

- Un node à supprimer ne peut pas être le serveur principal ou secondaire GPFS à moins qu'on supprime le cluster en entier. Vérifier au préalable avec la commande `mm\sc\cluster`. Dans le cas où le node à supprimer est l'un des ces serveurs et qu'on veut conserver le cluster, il faut assigner un autre node en tant que principal ou secondaire à l'aide de la commande `mmhc\cluster` avant de supprimer le node.
- Un node à supprimer ne peut pas être un serveur primaire ou backup NSD pour un quelconque volume du cluster GPFS à moins qu'on souhaite supprimer le cluster. Vérifier que ce n'est pas le cas avec la commande `mm\l\snsd`. Dans le cas où le node à supprimer est primaire ou backup NSD pour un ou plusieurs volumes, déplacer les volumes sur un autre node du cluster à l'aide de la commande `mmchnsd` avant de supprimer le node.
- GPFS doit être arrêté sur le node qui doit être supprimé. Utiliser la commande `mmshutdown` pour cette opération.

Suppression du node

```
mmde\lnode -N <nom_du_serveur>
```

Rentrer le "nom du serveur" dans le cluster

Réactiver la communication entre le serveur de cluster et le "nom du serveur" que l'on souhaite installer.

```
/etc/init.d/network start
```

Si le "nom du serveur" fait partie du quorum, le repositionner en tant que tel.

```
mmaddnode -N <nom_du_serveur> :quorum:
```

Sinon

```
mmaddnode -N <nom_du_serveur> ::
```

Ajouter de nouveau le NSD au node

Modification de la structure du NSD pour intégrer "nom du serveur". Pour cela modifier le fichier `desc_file.NSD` pour rajouter le "nom du serveur" et appliquer le fichier avec la commande qui suit.

```
mmchnsd -F desc_file.NSD
```

Vérifier la bonne prise en compte.

```
mmlnsd
```

Relancer GPFS.

```
mmstartup -a
```

Changer IP d'un node

- Méthode pour GPFS 3.3 extraite de la doc IBM suivante :
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=%2Fcom.ibm.cluster.gpfs.31.advanceadm.doc%2Fbl1adv_chnamip.html

GPFS assumes that IP addresses and host names remain constant. In the rare event that such a change becomes necessary or is inadvertently introduced by reinstalling a node with a disk image from a different node for example, follow the steps in this topic.

If all of the nodes in the cluster are affected and all the conditions in step 2 below are met:

1. Use the `mmshutdown -a` command to stop GPFS on all nodes.
2. Using the documented procedures for the operating system, add the new host names or IP addressees, but do not remove the old ones yet. This can be achieved, for example, by creating temporary alias entries in `/etc/hosts`. Avoid rebooting the nodes until the `mmchnode` command in step 3 is executed successfully. If any of these conditions cannot be met, utilize the alternate procedure described below.
3. Use `mmchnode -daemon-interface` and `-admin-interface` to update the GPFS configuration information.
4. If the IP addresses over which the subnet attribute is defined are changed, you need to update your configuration by using the `mmchconfig` command with the `subnets` attribute.
5. Start GPFS on all nodes with `mmstartup -a`.
6. Remove the unneeded old host names and IP addresses.

If only a subset of the nodes are affected, it may be easier to make the changes using these steps:

1. Before any of the host names or IP addresses are changed:
 - Use the `mmshutdown` command to stop GPFS on all affected nodes.
 - If the host names or IP addresses of the primary or secondary GPFS cluster configuration server nodes must change, use the `mmchcluster` command to specify another node to serve as the primary or secondary GPFS cluster configuration server.
 - If the host names or IP addresses of an NSD server node must change, temporarily remove the node from being a server with the `mmchnsd` command. Then, after the node has been added back to the cluster, use the `mmchnsd` command to change the NSDs to their original configuration. Use the `mmlnsd` command to obtain the NSD server node names.
 - Use the `mmdelnode` command to delete all affected nodes from the GPFS cluster.
2. Change the node names and IP addresses using the documented procedures for the operating system.

3. If the IP addresses over which the subnet attribute is defined are changed, you need to update your configuration by using the `mmchconfig` command with the `subnets` attribute.
4. Issue the `mmaddnode` command to restore the nodes to the GPFS cluster.
5. If necessary, use the `mmchcluster` and `mmchnsd` commands to restore the original configuration and the NSD servers.

Configuration d'un cluster GPFS de grande taille

```
mmchconfig tscCmdPortRange=35000-35200,maxFilesToCache=10000,  
pagepool=1G,maxMBpS=1600,prefetchThreads=200,worker1Threads=200
```

Pour chaque poste opérateur.

```
mmchconfig pagepool=256M -N <nom_poste>
```

Attribuer le rôle "manager" à chaque machine.

```
mmchnode -N <liste des machines séparées par des virgules> --manager
```

NSD sur un couple de node manager

Par défaut les nsd sont gérés par le quorum mais on peut ajouter des managers qui auront cette tâche de gestion des volumes. De manière précise chaque volume est associé à un couple de deux machines manager, l'une primaire et l'autre backup.

Creation d'un nsd composé de deux volumes SAN physiques attachés.

Commande pour afficher les noms des devices locaux par rapport aux LUN attachées.

```
/opt/mpp/lsvdev
```

La commande précédente est utile pour récupérer les noms des devices locaux (`/dev/sdao` par exemple) qui correspondent aux numéros des LUN SAN attachées.

Après avoir attaché les deux volumes SAN, aux machines, noter les noms des devices sur la machines, ils sont utiles pour les commandes qui suivent.

Pour le filesystem `/data/save`.

```
vi save1.nsd
```

```
/dev/sdao:srv3.domaine.fr:srv4.domaine.fr:dataAndMetadata:-1:data_save_1
```

```
mmcrnsd -F ./save1.nsd -v no
```

```
vi save2.nsd
```

```
/dev/sdap:srv4.domaine.fr:srv3.domaine.fr:dataAndMetadata:-1:data_save_2
```

```
mmcrnsd -F ./save1.nsd -v no
```

```
vi save.fs
```

```
/dev/sdao:srv3.domaine.fr:srv4.domaine.fr:dataAndMetadata:-1:data_save_1  
/dev/sdap:srv4.domaine.fr:srv3.domaine.fr:dataAndMetadata:-1:data_save_2
```

```
mmcrfs /dev/save -F save.fs -T /data/save -B 1M -A yes
```

Les couples de serveurs primaires et backups choisis doivent être réparti de façon équitable pour gérer les multiples filesystems.

NSD composé de 10 volumes physiques

En prenant exemple d'un volume GPFS composé de 10 volumes physiques SAN. Ils constitueront un seul volume GPFS.

Les machines srv3.domaine.fr et srv4.domaine.fr doivent être manager.

Creation du nsd "/data/archive"

```
cat > /tmp/gpfs/data_archive.nsd <<EOF  
  
/dev/sdt:srv3.domaine.fr:srv4.domaine.fr:dataOnly::data_archive_1:bigdata  
  
/dev/sdu:srv4.domaine.fr:srv3.domaine.fr:dataOnly::data_archive_2:bigdata  
  
/dev/sdv:srv3.domaine.fr:srv4.domaine.fr:dataOnly::data_archive_3:bigdata  
  
/dev/sdw:srv4.domaine.fr:srv3.domaine.fr:dataOnly::data_archive_4:bigdata  
  
/dev/sdx:srv3.domaine.fr:srv4.domaine.fr:dataOnly::data_archive_5:bigdata  
  
/dev/sdy:srv4.domaine.fr:srv3.domaine.fr:dataOnly::data_archive_6:bigdata  
  
/dev/sdak:srv3.domaine.fr:srv4.domaine.fr:dataAndMetadata::data_archive_meta  
data1:  
  
/dev/sdal:srv4.domaine.fr:srv3.domaine.fr:dataAndMetadata::data_archive_meta  
data2:  
  
/dev/sdam:srv3.domaine.fr:srv4.domaine.fr:dataAndMetadata::data_archive_small  
files1:  
  
/dev/sdan:srv4.domaine.fr:srv3.domaine.fr:dataAndMetadata::data_archive_small  
files2:
```

```
EOF
```

Les couples de serveurs primaires et backups choisis doivent être réparti de façon équitable pour gérer les multiples filesystems.

Créer le nsd composé des multiples volumes.

```
mmscrnsd -F /tmp/gpfs/data_archive.nsd
```

Création du filesystem

```
mmscrfs /dev/data_archive -F data_archive.nsd -T /data/archive -B 1M -n 100 -S yes
```

Création d'une policy pour le filesystem

```
cat > /tmp/gpfs/data_archive.policy <<EOF
```

```
/* stockage de petits fichiers sur storage pool "system" */
```

```
RULE 'CfgFiles' SET POOL 'system'
```

```
WHERE NAME LIKE '%xml' OR NAME LIKE '%XML' OR NAME LIKE '%lock'
```

```
/* stockage fichiers METEO sur storage pool "system" */
```

```
RULE 'MeteoFiles' SET POOL 'system'
```

```
FOR FILESET('METEO')
```

```
/* Autres fichiers sous "bigdata" */
```

```
RULE 'default' SET POOL 'bigdata'
```

```
EOF
```

```
mmchpolicy data_archive /tmp/gpfs/data_archive.policy -I yes
```

Création d'un fileset pour METEO.

```
mmscrfileset data_archive METEO
```

Augmentation nombre d'inodes

 à compléter.

- <http://publib.boulder.ibm.com/infocenter/clresctr/vrx/index.jsp?topic=%2Fcom.ibm.cluster.gpfs>

- [.doc%2Fgpfs_faqs%2Fgpfs_faqs.html](#)
- http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=%2Fcom.ibm.cluster.gpfs.v3r4.gpfs100.doc%2Fblins_maxnfile.html&resultof=%22maximum%22%20%22inode%22%20%22inod%22
- http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=%2Fcom.ibm.cluster.gpfs.v3r4.gpfs300.doc%2Fbladm_mmchfs.html&resultof=%22mmchfs%22%20%22mmchf%22%20%22-F%22%20%22f%22
- <http://fr.wikipedia.org/wiki/Inode>

What is the architectural limit of the number of files in a file system?

The architectural limit of the number of files in a file system is determined by the file system format:

For file system created with GPFS V3.4 or later, the architectural limit is 264.

The current tested limit is 9,000,000,000.

For file systems created with GPFS V2.3 or later, the limit is 2,147,483,648.

For file systems created prior to GPFS V2.3, the limit is 268,435,456.

Please note that the effective limit on the number of files in a file system is usually lower than the architectural limit, and could be adjusted using the `mmchfs` command (GPFS V3.4 and later use the `--inode-limit` option; GPFS V3.3 and lower use the `-F` option).

Méthode de calcul et modification avec `mmchfs` (extrait du man).

```
--inode-limit MaxNumInodes[:NumInodesToPreallocate]
```

`MaxNumInodes` specifies the maximum number of files that can be created. Allowable values range from the current number of created inodes (determined by issuing the `mmdf` command with `-F`), through the maximum number of files possibly supported as constrained by the formula:

$$\text{maximum number of files} = (\text{total file system space}) / (\text{inode size} + \text{subblock size})$$

[You can determine the inode size (`-i`) and subblock size (value of the `-B` parameter / 32) of a file system by running the `mmlsfs` command.]

If your file system has additional disks added or the number of inodes was insufficiently sized at file system creation, you can change the number of inodes and hence the maximum number of files that can be created.

For file systems that will be doing parallel file creates, if the total number of free inodes is not greater than 5% of the total number of inodes, there is the potential for slowdown in file system access. Take this into consideration when changing your file system.

NumInodesToPreallocate specifies the number of inodes that will be pre-allocated by the system right away. If this number is not specified, GPFS allocates inodes dynamically as needed.

The MaxNumInodes and NumInodesToPreallocate values can be specified with a suffix, for example 100K or 2M. Note that in order to optimize file system operations, the number of inodes that are actually created may be greater than the specified value.

This option applies only to the root fileset. When there are multiple inode spaces, use the `--inode-space` option of the `mmchfileset` command to alter the inode limits of independent filesets. The `mmchfileset` command can also be used to modify the root inode space. The `--inode-space` option of the `mmlsfs` command shows the sum of all inode spaces.

```
400Go
inode size : 512 bytes
Block size 524288
Maximum number of inode (actual value) : 100352
```

Calcul avec 400Go.

```
429496729600 / 512 + (524288 / 32)
429496729600 / 16 896 = 25 420 024
```

Calcul avec 100Go.

```
107374182400 / 512 + (524288 / 32)
107374182400 / 16 896 = 6 355 006
```

Changement de la valeur max inode. Cette commande fonctionne directement à chaud sans redémarrage.

```
mmchfs <filesystem_name ex: /dev/gpfs_data> -F <value>
```

Troobleshooting

Regarder log.

```
/var/adm/ras/mmfs.log.latest
```

```
gpfs.snap
mmfsadm
mmtracectl
```

mmfsck

- <http://blog.irwan.name/?p=2258>
- http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=%2Fcom.ibm.cluster.gpfs.v3r50-3.gpfs100.doc%2Fbl1adm_mmfsck.htm

Check filesystem GPFS sur /dev/<volname> Sur le serveur GPFS.

```
mmumount -a </dev/<device_name>
```

Si blocage du démontage car utilisé se rendre sur les machines qui utilisent le volume et entrer les commandes suivantes.

```
fuser -mv /<mount_point>
fuser -mkv /<mount_point>
Refaire le mmumount -a sur le serveur GPFS.
```

Sur le serveur GPFS.

```
mmumount /dev/gpfs_<volname> -a
mount
mmfsck /dev/gpfs_<volname> -v -n > mmfsck_check.out 2>&1
vi mmfsck_mpf_check.out
mmfsck /dev/gpfs_<volname> -v -y > mmfsck_fix.out 2>&1
vi mmfsck_mpf_fix.out
mmfsck /dev/gpfs_<volname> -v -y > mmfsck_fix2.out 2>&1
vi mmfsck_mpf_fix2.out
mount
mumount /dev/gpfs_uc_mpf -a
```

From:
<https://wiki.ouieuhoutca.eu/> - **kilsufi de noter**

Permanent link:
<https://wiki.ouieuhoutca.eu/gpfs>

Last update: **2021/01/21 21:42**

