IUT Nancy-Charlemagne Université Nancy 2
Licence Pro Asrall

Tuteur : Maître de Conférences : Lucas Nussbaum

Systèmes de fichiers distribué : comparaison de GlusterFS, MooseFS et Ceph avec déploiement sur la grille de calcul Grid'5000.

Jean-François Garçia, Florent Lévigne,
Maxime Douheret, Vincent Claudel

Nancy, le 17 mars 2011

Table des matières

Partie 1

Introduction

1.1 Contexte

Étudiants en licence professionnelle ASRALL (Administration de Systèmes, Réseaux, et Applications à base de Logiciels libres), notre formation prévoie une période de x mois à mis temps pour la réalisation d'un projet tuteuré.

Le projet que nous avons choisis consiste à comparer diverses solutions de systèmes de fichiers distribué.

1.2 Système de fichiers distribué

Un système de fichiers (file system en anglais) est une façon de stocker des informations et de les organiser dans des fichiers, sur des périphérique comme un disque dur, un CD-ROM, une clé USB, etc. Il existe de nombreux systèmes de fichiers (certains ayant des avantages sur d'autres), dont entre autre l'ext (Extented FS), le NTFS (New Technology FileSystem), ZFS (Zettabyte FS), FAT (File Allocation Table).

Un système de fichiers distribué est un système de fichiers permettant le partage de données à plusieurs clients au travers du réseau. Contrairement à un système de fichier local, le client n'a pas accès au système de stockage sous-jacent, et interagit avec le système de fichier via un protocole adéquat.

Un système de fichier distribué est donc utilisé par plusieurs machines en même temps (les machines peuvent ainsi avoir accès a des fichiers distants, l'espace de noms est mis en commun). Un tel système permet donc de partager des données entre plusieurs client, et pour certains de répartir la charge entre plusieurs machines, et de gérer la sécurité des données (par réplication)

1.3 Le Grid'5000

Partie 2

NFS

2.1 Présentation

NFS (Network File System, système de fichiers en réseau en français) est un système de fichiers développé par Sun Microsystem, permettant de partager des données par le réseau.

Partie 3

GlusterFs

3.1 Présentation

Partie 4

MooseFS

4.1 Présentation

4.1.1 Présentation générale

MooseFS (Moose File System) est un système de fichiers répartis à tolérance de panne, développé par Gemius SA Le code préalablement propriétaire a été libéré et mis à disposition publiquement le 5 mai 2008. Il permet de déployer assez facilement un espace de stockage réseau, répartit sur plusieurs serveurs.

Cette répartition permet de gérer la disponibilité des données, lors des montées en charge ou lors d'incident technique sur un serveur. L'atout principal de MooseFS, au delà du fait qu'il s'agisse d'un logiciel libre, est sa simplicité de mise en œuvre.

En effet le tutoriel de prise en main, disponible sur le site du projet, explique de manière clair comment mettre en place une architecture distribuées en quelques heures. Concernant les utilisations, elles sont multiples et surtout, après la phase de configuration, l'évolution du système est très simple. L'ajout de serveurs, d'espace disque peuvent être gérés très facilement.

4.1.2 Aspect technique

Les montage du système de fichiers par les clients se fait à l'aide de FUSE.
MooseFS est constitué de trois types de serveurs :

1. Un serveur de métadonnées (MDS)
   Ce serveur gère la répartition des différents fichiers
2. Un serveur métajournal (Metalogger server)
   Ce serveur récupère régulièrement les métadonnées du MDS et les stocke en tant que sauvegarde.
3. Des serveurs Chunk (CSS)
   Ce sont ces serveurs qui stockent les données des utilisateurs.

Le point le plus important étant de bien dimensionner le serveur Master (qui stocke les métadonnées) afin de ne pas être limité par le suite. Donc pour ceux qui ne peuvent pas mettre en place des systèmes de stockage réseaux propriétaires assez coûteux, je vous conseille d'étudier cette possibilité. Elle vous permettra de partager des données sur plusieurs machines, de manière rapide, fiable, sécurisée et surtout peu coûteuse.

4.2 Mise en place

4.2.1 Environnement logiciel

Le système utilisé est une Debian Squeeze. MosseFs ne faisant pas partie des dépôts de la distribution, nous avons compilé le paquet à partir des sources dans leurs dernières version (1.6.20).

Partie 5

Ceph

5.1 Présentation

Partie 6

Comparaison

6.1 Test de performances

Afin de ne pas avoir de différence de matériel lors nos test, ceux-ci ont tous été réalisés sur un même cluster du Grid'5000 : Graphene.

Ce cluster est composé de 144 noeuds, avec pour caractéristique :
- 1 CPU Intel de quatre cœurs cadencé à 2.53 GHz
- 16 Go de RAM
- 278 Go d'espace disque

Nous avons réalisé un benchmark mesurant les performances (débit) de quatre type d'opérations sur le système de fichier distribué :

Écriture de petits fichiers : écriture des sources du noyau linux (décompressé).

Écriture de gros fichiers : écriture d'un fichier de 3 Go [1].

Lecture de petits fichiers : lecture des fichiers du noyau linux. Pour cela, nous avons compressé le dossier contenant le noyau (impliquant la lecture des fichiers), en redirigeant la sortie vers /dev/nul (afin que les performances du disque ne rentrent pas en jeux).

Lecture de gros fichiers : lecture du fichier de 3 Go. Opération réalisé en faisant un "cat" du fichier, et en redirigeant la sortie vers /dev/nul afin de ne pas "polluer" le terminal.

1. Fichier créé avec la commande : dd if=/dev/zero of=/lieu/voulu bs=1G count=3

Partie 7

Conclusion

Partie A

Répartition des taches

A.1 Florent Lévigne

- Étude sur la mise en place de GlusterFS
- Réalisation d'un script de déploiement de GlusterFS
- Étude sur la mise en place de MooseFS
- Réalisation d'un script de déploiement de MooseFS
- Réalisation d'un script de benchmark pour système de fichiers distribué

A.2 Jean-François Garcia

- Étude sur la mise en place de GlusterFS
- Étude sur la mise en place de NFS
- Réalisation d'un script de déploiement de NFS
- Étude sur la mise en place de CephFS
- Réalisation d'un script de déploiement de CephFS

Partie B

Scripts

B.1 GlusterFs

Fichier deploimentGluster.rb :

```
#!/usr/bin/ruby -w
# encoding: utf-8

# r servation des noeuds (a lancer manuellement)
# oarsub -I -t deploy -l nodes=8,walltime=2
# oarsub -I -t deploy -l nodes=8,walltime=2 -p "cluster='graphene'"

# doit concorder avec la commande oarsub
numberOfClients = 5
numberOfServers = 3

infiniband = 1 # 1 : activ  , 0 : non activ  (ne change rien pour l'instant)

# cr ation d'un fichier contenant la liste des noeuds r serv s
`touch listOfNodes`
File.open("listOfNodes", 'w') do |file|
  file << `cat $OAR_FILE_NODES | sort -u`
end

# cr ation de deux fichiers contenant la liste des serveurs, et des clients
`touch listOfClients listOfServers`
serverWrited = 0
File.open("listOfNodes", 'r') do |node|
  File.open("listOfServers", 'w') do |server|
    File.open("listOfClients", 'w') do |client|
      while line = node.gets
        if serverWrited < numberOfServers
          server << line
          serverWrited += 1
        else
          client << line
        end
      end
    end
  end
end

# d ploiement des machines
puts "Machines en cours de d ploiement..."
```

```
#`kadeploy3 -k -a ../images/mysqueezegluster-x64-base.env -f listOfNodes` # image perso
`kadeploy3 -k -e squeeze-collective -u flevigne -f listOfNodes` # image collective


# Envoie d'un script de cr ation d'un r pertoire dans /tmp/sharedspace sur les serveurs
File.open("listOfServers", 'r') do |file|
  while line = file.gets
    machine = line.split.join("\n")
    `ssh root@#{machine} < createFolders.sh`
  end
end

# Envoie d'un script de cr ation d'un r pertoire dans /media/glusterfs sur les clients
File.open("listOfClients", 'r') do |file|
  while line = file.gets
    machine = line.split.join("\n")
    `ssh root@#{machine} < createMountDirectory.sh`
  end
end

masterServer = `head -n 1 listOfServers`.split.join("\n")

# g n ration des fichiers de conf, et envoie des fichiers de conf aux machines (serveurs et clients)
puts "Configuration des serveurs et des clients..."
`scp listOfServers root@#{masterServer}:`
`scp listOfClients root@#{masterServer}:`
`scp glusterfs-volgen.rb root@#{masterServer}:`
`ssh root@#{masterServer} ./glusterfs-volgen.rb`
#`ssh root@#{masterServer} < execScript/ex-glusterfs-volgen.sh`

# d marrage des serveurs
puts "D marrage des serveurs..."
File.open("listOfServers", 'r') do |file|
  while line = file.gets
    machine = line.split.join("\n")
    `ssh root@#{machine} < startGluster.sh`
  end
end

# montage du r pertoire par les clients
puts "Montage du r pertoire par les clients..."
File.open("listOfClients", 'r') do |file|
  while line = file.gets
    machine = line.split.join("\n")
    `ssh root@#{machine} < mountFs.sh`
  end
end

# r sum  des machines
puts "GlusterFS op rationnel"
puts "\nMachines clients :"
puts `cat listOfClients`
```

```
92
93 puts "\nMachines serveurs :"
94 puts `cat listOfServers`
95
96 puts "\nServeur maitre : #{masterServer}"
97
98 # nettoyage
99 #`rm listOfNodes listOfClients listOfServers`
```

B.2 MooseFs

B.3 Ceph

Fichier deploimentCeph.rb :

```
1  #!/usr/bin/ruby -w
2  # encoding: utf-8
3
4  ##############################################
5
6  # File Name : deploiementCeph.rb
7
8  # Purpose :
9
10 # Creation Date : 11-03-2011
11
12 # Last Modified : jeu. 17 mars 2011 14:54:51 CET
13
14 # Created By : Helldar
15
16 ##############################################
17
18 # doit concorder avec la commande oarsub
19
20 if ARGV[0] != nil
21   numberOfServers = ARGV[0].to_i
22   puts "Nb serveur : #{numberOfServers}\n"
23 else
24   puts "Veuillez relancer le script avec les bons param tres!\n"
25   exit
26 end
27
28 # cr ation d'un fichier contenant la liste des noeuds r serv s
29 `touch listOfNodes`
30 File.open("listOfNodes", 'w') do |file|
31         file << `cat $OAR_FILE_NODES | sort -u`
32 end
33 # cr ation de deux fichiers contenant la liste des serveurs, et des clients
34
35 `touch listOfClients listOfServers`
36 serverWrited = 0
37 File.open("listOfNodes", 'r') do |node|
38         File.open("listOfServers", 'w') do |server|
39                 File.open("listOfClients", 'w') do |client|
```

```
40                          while line = node.gets
41                                  if serverWrited < numberOfServers
42                                          server << line
43                                          serverWrited += 1
44                                  else
45                                          client << line
46                                  end
47                          end
48                  end
49          end
50  end
51
52  # d ploiement des machines
53  #puts "Machines en cour de d ploiement..."
54  #`kadeploy3 -k -e squeeze-collective -u flevigne -f listOfNodes` # image collective
55
56  # configuration du serveur
57  serveur_1 = `head -1 listOfServers | cut -d "." -f1`.strip
58  ip_serveur = `ssh root@#{serveur_1}  hostname -i`.strip
59
60  # g n ration du fichier de ceph.conf
61
62  `touch ceph.conf`
63  File.open("ceph.conf", 'w') do |file|
64       file << "[global]
65  ␣␣␣␣␣␣␣␣␣␣␣␣pid␣file␣=␣/var/run/ceph/$name.pid
66  ␣␣␣␣␣␣␣␣␣␣␣␣debug␣ms␣=␣1
67  ␣␣␣␣␣␣␣␣␣␣␣␣keyring␣=␣/etc/ceph/keyring.bin
68  [mon]
69  ␣␣␣␣␣␣␣␣␣␣␣␣mon␣data␣=␣/tmp/partage/mon$id
70  [mon0]
71  ␣␣␣␣␣␣␣␣␣␣␣␣host␣=␣#{serveur_1}
72  ␣␣␣␣␣␣␣␣␣␣␣␣mon␣addr␣=␣#{ip_serveur}:6789
73  [mds]
74  ␣␣␣␣␣␣␣␣␣␣␣␣debug␣mds␣=␣1
75  ␣␣␣␣␣␣␣␣␣␣␣␣keyring␣=␣/etc/ceph/keyring.$name"
76          if numberOfServers > 3
77           1.upto(3) { |i| file << "
78  [mds#{i␣-␣1}]"
79           host = `sed -n #{i + 1}p listOfServers | cut -d '.' -f1`.strip
80           file << "
81  ␣␣␣␣␣␣␣␣#{host}" }
82          else
83           file << "[mds0]"
84           host = `sed -n 2p listOfServers | cut -d '.' -f1`.strip
85           file << "
86  ␣␣␣␣␣␣␣␣#{host}"
87          end
88          file << "
89  [osd]
90  ␣␣␣␣␣␣␣␣sudo␣=␣true
91  ␣␣␣␣␣␣␣␣osd␣data␣=␣/tmp/partage/osd$id
92  ␣␣␣␣␣␣␣␣keyring␣=␣/etc/ceph/keyring.$name
93  ␣␣␣␣␣␣␣␣debug␣osd␣=␣1
94  ␣␣␣␣␣␣␣␣debug␣filstore␣=␣1
95  ␣␣␣␣␣␣␣␣osd␣journal␣=␣/tmp/partage/osd$id/journal
```

```
96  ␣␣␣␣␣␣␣osd␣journal␣size␣=␣1000"
97          1.upto(numberOfServers - 1) { |i| file << "
98  [osd#{i␣-␣1}]"
99          host = `sed -n #{i + 1}p listOfServers | cut -d '.' -f1`.strip
100         file << "
101 ␣␣␣␣␣␣␣#{host}" }
102 end
103 # copie du fichier ceph.conf vers le serveur
104 `scp ceph.conf root@#{serveur_1}:/etc/ceph`
105 puts "Envoy !"
106 # g n ration du fichier keyring.bin
107 `ssh root@#{serveur_1} cauthtool --create-keyring -n client.admin --gen-key keyring
        .bin`
108 `ssh root@#{serveur_1} cauthtool -n client.admin --cap mds 'allow' --cap osd 'allow
        *' --cap mon 'allow rwx' keyring.bin`
109 `ssh root@#{serveur_1} mv keyring.bin /etc/ceph/`
110 puts "Keyring␣g  n  r  !"
111 # montage
112 `ssh root@#{serveur_1} mount -o remount,user_xattr /tmp`
113 1.upto(numberOfServers - 1) { |i| serveurs = `sed -n #{i + 1}p listOfServers | cut
        -d "." -f1`.strip
114   `ssh root@#{serveurs} mount -o remount,user_xattr /tmp` }
115 puts "Montage␣fait!"
116 # d marrage du serveur
117 `ssh root@#{serveur_1} mkcephfs -c /etc/ceph/ceph.conf --allhosts -v -k /etc/ceph/
        keyring.bin`
118 `ssh root@#{serveur_1} /etc/init.d/ceph -a start`
119 puts "Serveur␣ceph␣d marr !"
120 # configuration des clients
121 0.upto(`wc -l listOfClients` - 1) { |i| clients = `sed -n #{i + 1}p listOfClients |
        cut -d "." -f1`.strip
122   `ssh root@#{clients} mkdir /ceph`
123   `ssh root@#{clients} cfuse -m #{ip_serveur} /ceph` }
124 puts "Clients␣mont s!"
```

B.4 NFS

Fichier deploimentNFS.rb :

```
1  #!/usr/bin/ruby -w
2  # encoding: utf-8
3
4  ###############################################
5
6  # File Name : deploiementNFS.rb
7
8  # Purpose :
9
10 # Creation Date : 17-03-2011
11
12 # Last Modified : jeu. 17 mars 2011 16:29:07 CET
13
14 # Created By : Helldar
15
16 ###############################################
```

```
17
18 `cat $OAR_FILE_NODES | sort -u > listOfNodes`
19
20 # D ploiement des machines
21 #puts "Machines en cour de d ploiement...\n"
22 #`kadeploy3 -k -e squeeze-collective -u flevigne -f listOfNodes # image collective`
23
24 serveur = `head -1 listOfNodes`.strip
25 puts "Le serveur : #{serveur}!\n"
26 # Suppression du serveur de la liste
27 `sed -i 1d listOfNodes`
28
29 puts "Configuration du serveur...\n"
30 `scp exports root@#{serveur}:/etc/`
31 `ssh root@#{serveur} /etc/init.d/nfs-kernel-server restart`
32
33 puts "Configuration des clients...\n"
34 line = `wc -l listOfNodes | cut -d ' ' -f1`.strip.to_i
35 puts "Il y a #{line} nodes"
36 1.upto(line) { |i| clients = `sed -n #{i}p listOfNodes | cut -d "." -f1`.strip
37   `ssh root@#{clients} mkdir /tmp/partage`
38   `ssh root@#{clients} mount #{serveur}:/tmp -t nfs /tmp/partage` }
```

B.5 Benchmark

Fichier benchmark.rb :

```
1  #!/usr/bin/ruby -w
2  # encoding: utf-8
3
4  if "#{ARGV[0]}" == ""
5    puts "Doit prendre en param tre le nombre de clients participant au bench."
6    exit(1)
7  end
8
9  $clientsOfBench = "#{ARGV[0]}"
10
11 # chemin du fichier contenant la liste des clients
12 listOfClients = "/home/flevigne/glusterFs/listOfClients"
13
14 # chemin ou   crire   les donn es du benchmark
15 whereToWrite = "/media/glusterfs"
16
17 # chemin du fichier contenant les r sultats
18 $outputRes = "./resOfBench"
19
20 # le client doit avoir dans /home/flevigne :
21 # - linux-2.6.37.tar.bz2 : noyau linux compress
22 # - bigFile : un fichier de 3 Go
23
24 # fichier contenant la liste des clients participant au benchmark
25 `touch clientOfBench`
26 `head -#{$clientsOfBench} #{listOfClients} > clientOfBench`
27
28 # si le fichier $outputRes n'existe pas, on le cr e.
```

```ruby
if !File.exist?($outputRes)
  `touch #{$outputRes}`
end

`echo "\nBenchmark sur #{$clientsOfBench} clients" >> #{$outputRes}`

$numberOfClients = open("clientOfBench").read.count("\n").to_i

puts "Lancement du benchmarck sur #{$numberOfClients} clients."


# lance un travail
# parametres :
# - name : nom du travail (str)
# - work : chemin du script de travail (str)
# - whereToWrite : chemin ou   crire   les donn es du benchmark (str)
# - size : taille (en Mo) du/des fichier(s) a ecrire/lire (float)
def startBench(name, work, whereToWrite, size)
  puts "bench : #{name} en cours..."

  totalSize = size.to_i * $clientsOfBench.to_i
  workFinished = 0
  startOfBench = Time.now

  # execution du sript pour tous les clients
  File.open("clientOfBench", 'r') do |file|
    while line = file.gets
      fork do
        machine = line.split.join("\n")
        `scp #{work} root@#{machine}:/root`
        `ssh root@#{machine} ./#{work} #{whereToWrite}`
        exit(0)
      end
    end
  end

  # on attend que tous les clients aient fini leur travail
  1.upto($numberOfClients) do
    pid = Process.wait
    workFinished += 1
    puts "Machine(s) ayant termin  leur travail : #{workFinished}"
  end

  endOfBench = Time.now
  duration = endOfBench - startOfBench

  puts "Toute les machines ont termin  leur travail."

  puts "  --> Le benchmark \"#{name}\" a dur  #{duration} secondes.  (debit : #{totalSize / duration} Mo/s)"

  `echo "#{name} : #{duration} sec : #{totalSize / duration} Mo/s" >> #{$outputRes}`
end


```

```
# lancement du benchmark
startBench("ecriture de petits fichiers", "writingSmallFiles.sh", whereToWrite, 479)
startBench("ecriture de gros fichiers", "writingBigFiles.sh", whereToWrite, 3076)
startBench("lecture de petits fichiers", "readingSmallFiles.sh", whereToWrite, 479)
startBench("lecture de gros fichiers", "readingBigFile.sh", whereToWrite, 3076)

# nettoyage du syst me de fichier distribue (necessaire pour enchainer les benchmark)
puts "Nettoyage de l'espace de travail..."
oneClient = `head -1 clientOfBench`.strip
`ssh root@#{oneClient} rm -r #{whereToWrite}/*`

puts "\nBenchmark termine"
```

Fichier writingSmallFiles.sh :

```
#!/bin/bash

whereToWrite=$1

nameOfMachine=`uname -n`

# creation du repertoire de travail de la machine
mkdir "$whereToWrite/$nameOfMachine"

# decompression dans ce repertoire
cd "$whereToWrite/$nameOfMachine"
tar -xf /home/flevigne/linux-2.6.37.tar.bz2

# copie dans le rep partage
#cp -r /home/flevigne/linux-2.6.37 "$whereToWrite/$nameOfMachine"
```

Fichier writingBigFiles.sh :

```
#!/bin/bash

whereToWrite=$1

nameOfMachine=`uname -n`

# on copie le gros fichier au lieu voulu
cp /home/flevigne/bigFile "$whereToWrite/$nameOfMachine"
```

Fichier readingSmallFiles.sh :

```
#!/bin/bash

whereToWrite=$1

nameOfMachine=`uname -n`

cd "$whereToWrite/$nameOfMachine"

# lecture des fichiers du noyau linux (compression (donc lecture) redirig vers /dev/nul)
tar -cf /dev/null linux-2.6.37
```

Fichier readingBigFile.sh :

```
1  #!/bin/bash
2
3  whereToWrite=$1
4
5  nameOfMachine=`uname -n`
6
7  cd "$whereToWrite/$nameOfMachine"
8
9  # lecture du gros fichier
10 cat bigFile > /dev/nul
```