GlusterFS overview 1/3

- En stor parallel network filsystem over Ethernet
- Findes til Linux, OS X, NetBSD og OpenSolaris
- Udviklet af firmaet Gluster, der blev købt af RedHat i 2011

GlusterFS overview 2/3

GlusterFS består af:

glusterfsd
: En dæmon for at kunne export et lokalt filsystem som et gluster-volume

glusterd
: Bruges til volume management.
Dæmonen skal køre på alle export server.

gluster
: CLI frontend for GlusterFS kommandoer.

GlusterFS overview 3/3

GlusterFS kommandoer kan køres direkte med prefikset gluster

```
# gluster [command]
```

eller fra en “Gluster Console Manager” prompt:

```
# gluster
gluster> [command]
```

Alle CLI kommandoer kan ses med:

```
# gluster help
```

Volume Types 1/3

GlusterFS understøtter flg. *"Volumes Types":*

Distributed:
Distributes files throughout the bricks in the volume.
Useful where the requirement is to scale storage and the redundancy is either not important or is provided by other hardware/software layers.

Replicated:
Replicates files across bricks in the volume.
Useful where high-availability and high-reliability is critical.

Volume Types 2/3

Striped:
Stripes data across bricks in the volume.
Useful only in high concurrency environment accessing very large files.

Distributed Striped:
Stripe data across two or more nodes in the cluster.
Useful to scale storage in high concurrency environments where accessing very large files is critical.

Distributed Replicated:
Distributes files across replicated bricks in the volume.
Useful where the requirement is to scale storage and high-reliability is critical.
Also gives improved read performance in most environments.

Volume Types 3/3

Distributed Striped Replicated:
Configuration of this volume type is currently only supported for Map Reduce workloads.

Striped Replicated:
Configuration of this volume type is currently only supported for Map Reduce workloads.

GlusterFS eksempel 1/5

Først laves partitioner og filsystemer på alle noder.

Inode-størrelsen sættes op til 512 bytes for at lave plads til GlusterFS's xattr samt bedre sikker for diske > 1TB.

```
# fdisk -l /dev/sdb | tail -2
  Device Boot      Start         End      Blocks   Id  System
  /dev/sdb1            1      382819   976761560   83  Linux

# mkfs.xfs -i size=512 /dev/sdb1
```

Samt tilføjes til /etc/fstab og mountes.

```
# mkdir -p /data/brick
# echo "/dev/sdb1 /data/brick xfs inode64 1 2" >> /etc/fstab
# mount /data/brick
```

GlusterFS eksempel 2/5

Starter glusterd på alle noder, så de kan kommunikere.

```
# service glusterd start
# service glusterd status
glusterd (pid  2749) is running...
```

Og sætter glusterd til autostart efter boot.

```
# chkconfig glusterd on
# chkconfig --list glusterd
  glusterd    0:off  1:off  2:on  3:on  4:on  5:on  6:off
```

GlusterFS eksempel 3/5

På *node01* skal *node02*, *node03* og *node04* tilføjes.

```
# gluster
gluster> peer status
Number of Peers: 0
gluster> peer probe node02
peer probe: success.
gluster> peer probe node03
peer probe: success.
gluster> peer probe node04
peer probe: success.
gluster> peer status
Number of Peers: 3

Hostname: node02
Uuid: af3531e8-2b49-4c3d-b050-c1cafd2eac73
State: Peer in Cluster (Connected)

Hostname: node03
Uuid: c035ec55-c59e-4c3c-b104-66df7c40abe1
State: Peer in Cluster (Connected)

Hostname: node04
Uuid: 5b11a730-d3a7-45a1-8fd1-71314172d188
State: Peer in Cluster (Connected)
```

GlusterFS eksempel 4/5

Endeligt laves et område på alle noderne til et *distributed replicated* volume *gv01:*

```
# mkdir /data/brick/gv01
```

og volume *gv01* oprettes og startes på node01:

```
# gluster volume create gv01 replica 2 node0{1,2,3,4}:/data/brick/gv01
volume create: gv01: success: please start the volume to access data
# gluster volume  start gv01
volume start: gv01: success
# gluster volume  info gv01
Volume Name: gv01
Type: Distributed-Replicate
Volume ID: e25b8bf3-6f96-49c8-98e8-95155e54820f
Status: Started
Number of Bricks: 2 x 2 = 4
Transport-type: tcp
Bricks:
Brick1: node01:/data/brick/gv01
Brick2: node02:/data/brick/gv01
Brick3: node03:/data/brick/gv01
Brick4: node04:/data/brick/gv01
```

GlusterFS eksempel 5/5

Nu kan Gluster fileystemet mountes og bruges:

```
# mount -t glusterfs node01:/gv01 /mnt/gluster/
# df -h /mnt/gluster/
Filesystem      Size  Used Avail Use% Mounted on
node01:/gv01    1,9T   68M  1,9T   1% /mnt/gluster
```

GlusterFS NFS

Der er indbygget NFS v3-server support i GlusterFS, som kan enables eller disables nemt.

```
gluster> volume set gv01 nfs.disable on
volume set: success
gluster> volume set gv01 nfs.disable off
volume set: success
```

UDP er ikke understøttet.

Det kan være nødvendigt at angive TCP til NFS-trafikken.

```
# mount -o proto=tcp,vers=3 nfs://node01/gv01 /mnt/s
```

Og i */etc/fstab*

```
node01:/gv01 /mnt/gfs nfs defaults,_netdev,mountproto=tcp 0 0
```

GlusterFS performance profiling 1/7

Man kan disable og enable performance profiling nemt.

```
gluster> volume profile gv01 start
Starting volume profile on gv01 has been successful
gluster> volume profile gv01 stop
Stopping volume profile on gv01 has been successful
```

Hvis der er enablet profiling, så kan det ses med *info*:

```
gluster> volume info
[...]
Options Reconfigured:
diagnostics.count-fop-hits: on
diagnostics.latency-measurement: on
[...]
```

GlusterFS performance profiling 2/7

```
gluster> volume profile gv01 info
Brick: node01:/data/brick/gv01
------------------------------
Cumulative Stats:
   Block Size:                16b+                  128b+                4096b+
 No. of Reads:                   0                      0                     0
No. of Writes:                 452                      1                     1

   Block Size:              8192b+              131072b+
 No. of Reads:                   0                      0
No. of Writes:                   1                 302585
 %-latency   Avg-latency   Min-Latency   Max-Latency   No. of calls         Fop
 ---------   -----------   -----------   -----------   ------------        ----
      0.00       0.00 us       0.00 us       0.00 us            467     RELEASE
[...]
      0.47    1508.15 us     131.00 us  170515.00 us            455      CREATE
     16.41      78.71 us      26.00 us  113718.00 us         303040       WRITE
     82.19  963528.60 us   33728.00 us 8473525.00 us            124       FSYNC

    Duration: 696380 seconds
   Data Read: 0 bytes
Data Written: 39660454664 bytes
```

GlusterFS performance profiling 3/7

Med *top* kan man få en masse informationer om mest åbne filer, antal åbne filer p.t., de meste læste/skrevne filer m.m.

```
gluster> volume top gv01 open
Brick: node01:/data/brick/gv01
Current open fds: 0, Max open fds: 3, Max openfd time: 2015-04-20 13:23:50.084800
Count                filename
=======================
6                    /allyears_10.csv.xz
6                    /allyears_10.csv.gz
[...]
Brick: node04:/data/brick/gv01
Current open fds: 0, Max open fds: 3, Max openfd time: 2015-04-20 12:41:41.001321
Count                filename
=======================
74                   /allyears_10.csv
10                   /allyears_10.csv.bz2
```

GlusterFS performance profiling 4/7

Top 3 over flest skrivninger:

```
gluster> volume top gv01 write list-cnt 3
Brick: node01:/data/brick/gv01
Count                filename
=======================
178632               /allyears_10.csv.gz
123955               /allyears_10.csv.xz
1                    /size-9998.done
Brick: node04:/data/brick/gv01
Count                filename
=======================
982908               /allyears_10.csv
128880               /allyears_10.csv.bz2
1                    /size-10000.done
[...]
```

GlusterFS performance profiling 5/7

Top 3 over flest læsninger:

```
gluster> volume top gv01 read list-cnt 3
Brick: node01:/data/brick/gv01
Brick: node04:/data/brick/gv01
Count                filename
=======================
3                    /size.txt
Brick: node02:/data/brick/gv01
Brick: node03:/data/brick/gv01
Count                filename
=======================
962299               /allyears_10.csv
```

GlusterFS performance profiling 6/7

Læse performance

```
gluster> volume top gv01 read-perf list-cnt 10
Brick: node01:/data/brick/gv01
MBps Filename                                                   Time
==== ========                                                   ====
2114 /allyears_10.csv.gz                                        2015-04-20 15:01:40.787796
Brick: node04:/data/brick/gv01
MBps Filename                                                   Time
==== ========                                                   ====
   0 /size.txt                                                  2015-04-20 14:41:03.386982
Brick: node02:/data/brick/gv01
Brick: node03:/data/brick/gv01
MBps Filename                                                   Time
==== ========                                                   ====
4096 /allyears_10.csv                                           2015-04-20 14:41:40.642246
2259 /allyears_10.csv.bz2                                       2015-04-20 15:00:49.1726
```

GlusterFS performance profiling 7/7

Skrive performance

```
gluster> volume top gv01 write-perf list-cnt 10
Brick: node01:/data/brick/gv01
MBps Filename                                                 Time
==== ========                                                 ====
4519 /allyears_10.csv.xz                                      2015-04-20 13:26:59.441095
4096 /allyears_10.csv.gz                                      2015-04-20 13:24:57.191450
   5 /allyears_10.csv.md5                                     2015-04-20 13:26:46.691494
   1 /size-9824.done                                          2015-04-20 14:44:52.224550
```

GlusterFS volume status 1/2

Med *volume status* kan man bl.a. se om de enkelte bricks er oppe:

```
gluster> volume status gv01
Status of volume: gv01
Gluster process                                 Port    Online  Pid
------------------------------------------------------------------------------
Brick node01:/data/brick/gv01                   49153   Y       8741
Brick node02:/data/brick/gv01                   49153   Y       2148
Brick node03:/data/brick/gv01                   49153   Y       2144
Brick node04:/data/brick/gv01                   49153   Y       8652
NFS Server on localhost                         2049    Y       8750
Self-heal Daemon on localhost                   N/A     Y       8751
NFS Server on node04                            2049    Y       8651
Self-heal Daemon on node04                      N/A     Y       8661
NFS Server on node02                            2049    Y       2157
Self-heal Daemon on node02                      N/A     Y       2172
NFS Server on node03                            2049    Y       2157
Self-heal Daemon on node03                      N/A     Y       2162

Task Status of Volume gv01
------------------------------------------------------------------------------
There are no active volume tasks
```

GlusterFS volume status 2/2

Og se de forbundne klienter:

```
gluster> volume status gv01 clients
Client connections for volume gv01
----------------------------------------------
Brick : node01:/data/brick/gv01
Clients connected : 8
Hostname                                               BytesRead    BytesWritten
--------                                               ---------    ------------
192.168.12.1:1008                                           4296            3984
192.168.12.2:1014                                        1793364         1012376
192.168.12.5:954                                            6144            4972
192.168.12.4:1011                                            828             428
192.168.12.4:1009                                            936             564
192.168.12.2:1010                                            828             428
192.168.12.1:1003                                         9474248         8831172
192.168.12.1:1001                                     39834482581     39858278358
----------------------------------------------
```

GlusterFS volume status 1/4

Med volume status kan man bl.a. se om de enkelte bricks er oppe:

```
gluster> volume status gv01
Status of volume: gv01
Gluster process                                 Port    Online  Pid
------------------------------------------------------------------------------
Brick node01:/data/brick/gv01                   49153   Y       8741
Brick node02:/data/brick/gv01                   49153   Y       2148
Brick node03:/data/brick/gv01                   49153   Y       2144
Brick node04:/data/brick/gv01                   49153   Y       8652
NFS Server on localhost                         2049    Y       8750
Self-heal Daemon on localhost                   N/A     Y       8751
NFS Server on node04                            2049    Y       8651
Self-heal Daemon on node04                      N/A     Y       8661
NFS Server on node02                            2049    Y       2157
Self-heal Daemon on node02                      N/A     Y       2172
NFS Server on node03                            2049    Y       2157
Self-heal Daemon on node03                      N/A     Y       2162

Task Status of Volume gv01
------------------------------------------------------------------------------
There are no active volume tasks
```

GlusterFS statedump 1/2

En statedump er en mekanisme til at få alle GlusterFS variabler samt tilstand af de enkelte GlusterFS processer dumpet.

```
gluster> volume statedump gv01
Volume statedump successful
```

Outputtet bliver default gemt i en fil med navnet /tmp/*brickname.PID.dump.timestamp*

Placeringen kan nemt ændres til f.eks. */var/log/glusterfs/statedump/*

```
gluster> volume set gv01 server.statedump-path /var/log/glusterfs/statedump
```

GlusterFS statedump 2/2

Indholdet af en statedump fil vil ligne dette:

```
DUMP-START-TIME: 2015-04-20 12:47:45.820438

[mallinfo]
mallinfo_arena=3043328
mallinfo_ordblks=71
mallinfo_smblks=3
mallinfo_hblks=12
mallinfo_hblkhd=16060416
mallinfo_usmblks=0
mallinfo_fsmblks=192
mallinfo_uordblks=779280
mallinfo_fordblks=2264048
mallinfo_keepcost=2225232

[global.glusterfs - Memory usage]
num_types=119

[global.glusterfs - usage-type gf_common_mt_dnscache6 memusage]
size=16
num_allocs=1
max_size=16
max_num_allocs=1
total_allocs=1

[global.glusterfs - usage-type gf_common_mt_event_pool memusage]
[...]
```