Techtalk

Gluster in 7d

October 07, 2016

What's Gluster?

Figure 1:

“GlusterFS is a scalable network filesystem. Using common off-the-shelf hardware, you can create large, distributed storage solutions for media streaming, data analysis, and other data- and bandwidth-intensive tasks."

– https://www.gluster.org/

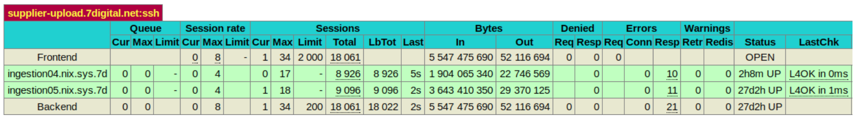
Where do we use it?

- Mediapool 777 & 888
- Supplier uploads
- Ingestion Share (new starfox)

Why do we use it?

- Redundancy on the cheap in the pre GCS world
- Less complex than legacy mediapools*
- **In Theory*

Simple use case

- Supplier uploads
- 

supplier-upload.7digital.net:ssh

	Queue			Session rate			Sessions						Bytes		Denied		Errors			Warnings			
	Cur	Max	Limit	Cur	Max	Limit	Cur	Max	Limit	Total	LbTot	Last	In	Out	Req	Resp	Req	Conn	Resp	Retr	Redis	Status	LastChk
Frontend				0	8	-	1	34	2 000	18 061			5 547 475 690	52 116 694	0	0	0					OPEN	
ingestion04.nix.sys.7d	0	0	-	0	4		0	17	-	8 926	8 926	5s	1 904 065 340	22 746 569		0		0	10	0	0	2h8m UP	L4OK in 0ms
ingestion05.nix.sys.7d	0	0	-	0	4		1	18	-	9 096	9 096	2s	3 643 410 350	29 370 125		0		0	11	0	0	27d2h UP	L4OK in 1ms
Backend	0	0		0	8		1	34	200	18 061	18 022	2s	5 547 475 690	52 116 694	0	0		0	21	0	0	27d2h UP	

-
```
root@ingestion05:~# gluster vol info
Volume Name: supplier-upload
Type: Replicate
Volume ID: 01cec521-7331-4b29-a90e-e23a823297e3
Status: Started
Number of Bricks: 1 x 2 = 2
Transport-type: tcp
Bricks:
Brick1: ingestion04:/export/brick00/supplier-upload
Brick2: ingestion05:/export/brick00/supplier-upload
```

What it looks like

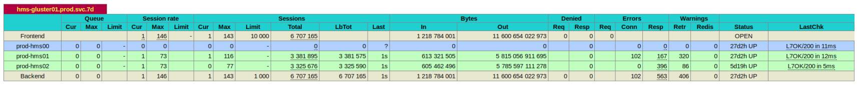
-
```
ops@ingestion05:~$ tail -2 /proc/mounts
/dev/mapper/debian-brick00 /export/brick00 ext4 \
rw,noatime,user_xattr,barrier=1,data=ordered 0 0
ingestion05:/supplier-upload /srv fuse.glusterfs \
rw,relatime,user_id=0,group_id=0,default_permissions,allow_other,max_read=131072 0 0
```

-
  ```
  [ftpadmin]
   comment = FTP Share Directory
   path = /srv/ftp
  ```

What it looks like (cont.)

-
  ```
  root@ingestion05:/export/brick00/supplier-upload/ftp# ls -t | head
  tunecore
  theorchard
  consolidated
  FUGA
  fuga_guvera
  kontor
  sonyset
  believe
  sonyhd
  labelworx
  root@ingestion05:/srv/ftp# ls -t | head
  theorchard
  kontor
  tunecore
  consolidated
  FUGA
  fuga_guvera
  sonyset
  believe
  sonyhd
  labelworx
  ```

More complex use case

- mp777
- 

hms-gluster01.prod.svc.7d

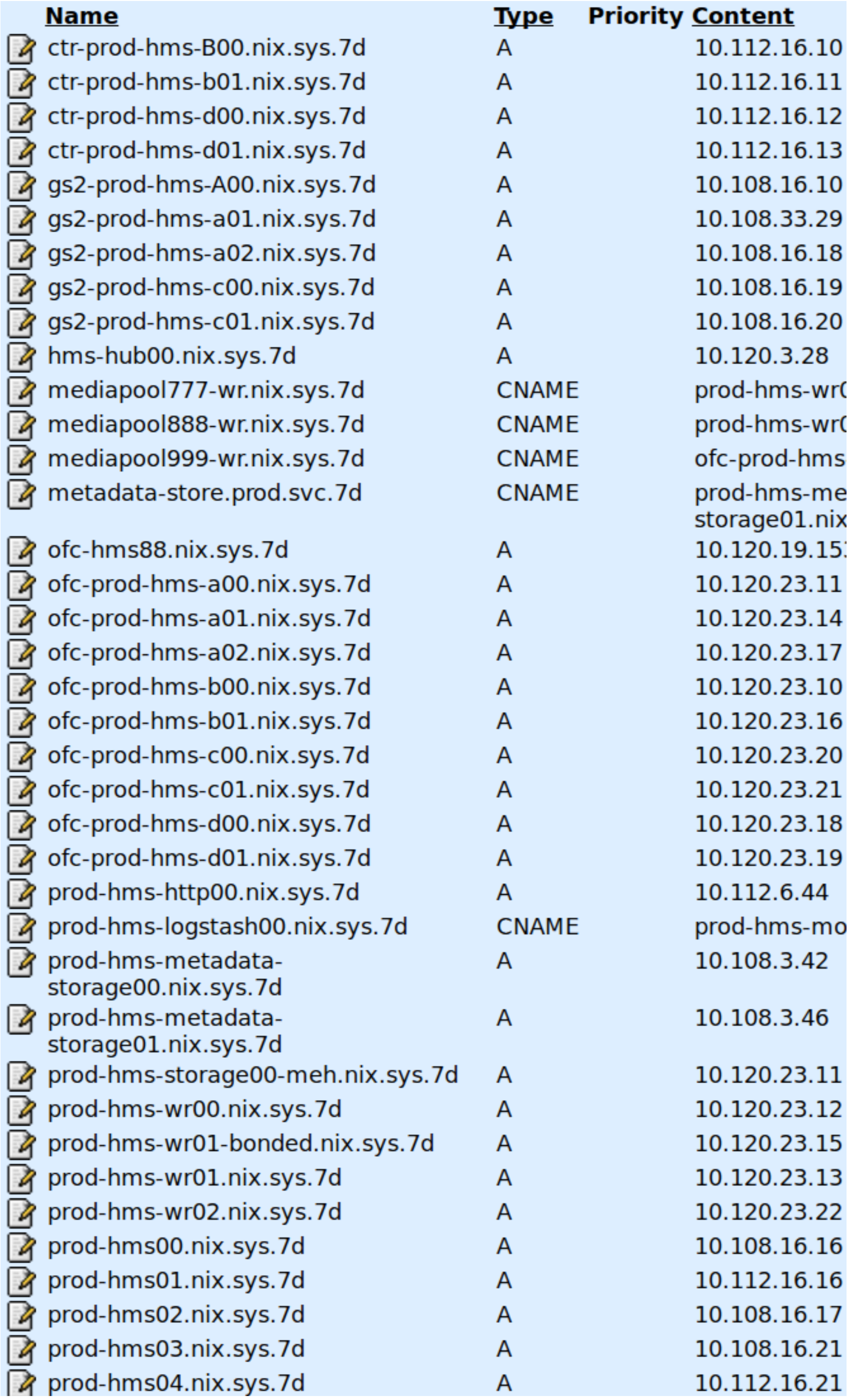
	Queue			Session rate			Sessions						Bytes		Denied		Errors			Warnings			
	Cur	Max	Limit	Cur	Max	Limit	Cur	Max	Limit	Total	LbTot	Last	In	Out	Req	Resp	Req	Conn	Resp	Retr	Redis	Status	LastChk
Frontend				1	146	-	1	143	10 000	6 707 165			1 218 784 001	11 600 654 022 973	0	0	0					OPEN	
prod-hms00	0	0	-	0	0		0	0	-	0	0	?	0	0		0		0	0	0	0	27d2h UP	L7OK/200 in 11ms
prod-hms01	0	0	-	1	73		1	116	-	3 381 895	3 381 575	1s	613 321 505	5 815 056 911 695		0		102	167	320	0	27d2h UP	L7OK/200 in 12ms
prod-hms02	0	0	-	1	73		0	77	-	3 325 676	3 325 590	1s	605 462 496	5 785 597 111 278		0		0	396	86	0	5d19h UP	L7OK/200 in 5ms
Backend	0	0		1	146		1	143	1 000	6 707 165	6 707 165	1s	1 218 784 001	11 600 654 022 973	0	0		102	563	406	0	27d2h UP	

-
  ```
  ops@prod-hms01:~$ df -h
  Filesystem              Size  Used Avail Use% Mounted on
  rootfs                  7.3G  1.5G  5.5G  21% /
  ```

```
udev                       10M     0   10M   0% /dev
tmpfs                     397M  2.2M  395M   1% /run
/dev/mapper/debian-root  7.3G  1.5G  5.5G  21% /
tmpfs                     5.0M     0  5.0M   0% /run/lock
tmpfs                     794M     0  794M   0% /run/shm
/dev/vda1                 228M   20M  197M  10% /boot
ctr-prod-hms-b00:/gv01   610T  606T  3.4T 100% /srv/gv01
```

Complex (cont.)

-
  ```
  ops@prod-hms01:~$ grep gv01 /etc/fstab
  ctr-prod-hms-b00:/gv01 /srv/gv01 glusterfs \
  defaults,_netdev,backupvolfile-server=ctr-prod-hms-b01 0 0
  ```



	Name	Type	Priority	Content
✎	ctr-prod-hms-B00.nix.sys.7d	A		10.112.16.10
✎	ctr-prod-hms-b01.nix.sys.7d	A		10.112.16.11
✎	ctr-prod-hms-d00.nix.sys.7d	A		10.112.16.12
✎	ctr-prod-hms-d01.nix.sys.7d	A		10.112.16.13
✎	gs2-prod-hms-A00.nix.sys.7d	A		10.108.16.10
✎	gs2-prod-hms-a01.nix.sys.7d	A		10.108.33.29
✎	gs2-prod-hms-a02.nix.sys.7d	A		10.108.16.18
✎	gs2-prod-hms-c00.nix.sys.7d	A		10.108.16.19
✎	gs2-prod-hms-c01.nix.sys.7d	A		10.108.16.20
✎	hms-hub00.nix.sys.7d	A		10.120.3.28
✎	mediapool777-wr.nix.sys.7d	CNAME		prod-hms-wr
✎	mediapool888-wr.nix.sys.7d	CNAME		prod-hms-wr
✎	mediapool999-wr.nix.sys.7d	CNAME		ofc-prod-hms
✎	metadata-store.prod.svc.7d	CNAME		prod-hms-me storage01.nix
✎	ofc-hms88.nix.sys.7d	A		10.120.19.15
✎	ofc-prod-hms-a00.nix.sys.7d	A		10.120.23.11
✎	ofc-prod-hms-a01.nix.sys.7d	A		10.120.23.14
✎	ofc-prod-hms-a02.nix.sys.7d	A		10.120.23.17
✎	ofc-prod-hms-b00.nix.sys.7d	A		10.120.23.10
✎	ofc-prod-hms-b01.nix.sys.7d	A		10.120.23.16
✎	ofc-prod-hms-c00.nix.sys.7d	A		10.120.23.20
✎	ofc-prod-hms-c01.nix.sys.7d	A		10.120.23.21
✎	ofc-prod-hms-d00.nix.sys.7d	A		10.120.23.18
✎	ofc-prod-hms-d01.nix.sys.7d	A		10.120.23.19
✎	prod-hms-http00.nix.sys.7d	A		10.112.6.44
✎	prod-hms-logstash00.nix.sys.7d	CNAME		prod-hms-mo
✎	prod-hms-metadata-storage00.nix.sys.7d	A		10.108.3.42
✎	prod-hms-metadata-storage01.nix.sys.7d	A		10.108.3.46
✎	prod-hms-storage00-meh.nix.sys.7d	A		10.120.23.11
✎	prod-hms-wr00.nix.sys.7d	A		10.120.23.12
✎	prod-hms-wr01-bonded.nix.sys.7d	A		10.120.23.15
✎	prod-hms-wr01.nix.sys.7d	A		10.120.23.13
✎	prod-hms-wr02.nix.sys.7d	A		10.120.23.22
✎	prod-hms00.nix.sys.7d	A		10.108.16.16
✎	prod-hms01.nix.sys.7d	A		10.112.16.16
✎	prod-hms02.nix.sys.7d	A		10.108.16.17
✎	prod-hms03.nix.sys.7d	A		10.108.16.21
✎	prod-hms04.nix.sys.7d	A		10.112.16.21

Complex (cont. 2)

-
  ```
  root@ctr-prod-hms-b00:~# gluster vol info
  Volume Name: gv01
  Type: Distributed-Replicate
  Volume ID: 94cff867-4a2c-4cc3-9a9b-8033623d49b6
  Status: Started
  Number of Bricks: 25 x 2 = 50
  Transport-type: tcp
  Bricks:
  Brick1: ofc-prod-hms-b00:/export/brick000
  Brick2: ctr-prod-hms-b00:/export/brick000
  Brick3: ofc-prod-hms-b00:/export/brick001
  Brick4: ctr-prod-hms-b00:/export/brick001
  Brick5: ofc-prod-hms-b00:/export/brick002
  Brick6: ctr-prod-hms-b00:/export/brick002
  Brick7: ofc-prod-hms-b00:/export/brick003
  Brick8: ctr-prod-hms-b00:/export/brick003
  Brick9: ofc-prod-hms-b00:/export/brick004
  Brick10: ctr-prod-hms-b00:/export/brick004
  Brick11: gs2-prod-hms-a00:/export/brick000
  Brick12: ofc-prod-hms-a00:/export/brick000
  Brick13: gs2-prod-hms-a00:/export/brick001
  Brick14: ofc-prod-hms-a00:/export/brick001
  Brick15: gs2-prod-hms-a00:/export/brick002
  ```

Complex (cont. 3)

```
Brick38: ofc-prod-hms-b01:/export/brick003
Brick39: ctr-prod-hms-b01:/export/brick004
Brick40: ofc-prod-hms-b01:/export/brick004
Brick41: gs2-prod-hms-a02:/export/brick000
Brick42: ofc-prod-hms-a02:/export/brick000
Brick43: gs2-prod-hms-a02:/export/brick001
Brick44: ofc-prod-hms-a02:/export/brick001
Brick45: gs2-prod-hms-a02:/export/brick002
Brick46: ofc-prod-hms-a02:/export/brick002
Brick47: gs2-prod-hms-a02:/export/brick003
Brick48: ofc-prod-hms-a02:/export/brick003
Brick49: gs2-prod-hms-a02:/export/brick004
Brick50: ofc-prod-hms-a02:/export/brick004
Options Reconfigured:
cluster.self-heal-daemon: off
cluster.data-self-heal: off
```

```
cluster.metadata-self-heal: off
cluster.entry-self-heal: off
```

Complex (cont. 4)

-
  ```
  root@prod-hms01:~# netstat -natp| grep glusterfs | head
  tcp        0      0 10.112.16.16:935        10.120.23.11:49182      ESTABLISHED 1917/gl
  tcp        0      0 10.112.16.16:928        10.120.23.11:49186      ESTABLISHED 1917/gl
  tcp        0      0 10.112.16.16:956        10.120.23.10:49155      ESTABLISHED 1917/gl
  tcp        0      0 10.112.16.16:969        10.108.16.10:49185      ESTABLISHED 1917/gl
  tcp        0      0 10.112.16.16:926        10.112.16.11:49153      ESTABLISHED 1917/gl
  tcp        0      0 10.112.16.16:944        10.120.23.17:49155      ESTABLISHED 1917/gl
  tcp        0      0 10.112.16.16:958        10.108.33.29:49154      ESTABLISHED 1917/gl
  tcp        0      0 10.112.16.16:947        10.120.23.17:49152      ESTABLISHED 1917/gl
  tcp        0      0 10.112.16.16:929        10.120.23.11:49185      ESTABLISHED 1917/gl
  tcp        0      0 10.112.16.16:961        10.120.23.10:49152      ESTABLISHED 1917/gl
  root@prod-hms01:~# netstat -natp| grep glusterfs | wc -l
  51
  ```

Issues

-
  ```
  ops@prod-hms03:~$ ls -l /srv/gv02/hms/track/000/049/009/850/17
  ---------T 1 root root 0 Nov  4  2015 /srv/gv02/hms/track/000/049/009/850/17
  ```

Issues (cont.)

-
  ```
  Track: 45086687, 17
  gv02: gs2-prod-hms-c00.nix.sys.7d
  On 2 brick(s)
  Result 0: -rw-rw-r-- 2 www-data www-data 5080970 May  5  2015 /export/brick002/gv02/hms
  Brick brick002 owner: www-data size: 5080970
  Result 1: -rw-rw-r-- 2 www-data www-data 5080970 May  5  2015 /export/brick003/gv02/hms
  Brick brick003 owner: www-data size: 5080970
  gv02: ctr-prod-hms-d00.nix.sys.7d
  On 1 brick(s)
  Result 0: ---------T 2 root root 0 Jul 23 06:43 /export/brick001/gv02/hms/track/000/045
  Brick brick001 owner: root size: 0
  gv02: gs2-prod-hms-c01.nix.sys.7d
  On 1 brick(s)
  Result 0: ---------T 2 root root 0 Dec 30  2015 /export/brick001/gv02/hms/track/000/045
  Brick brick001 owner: root size: 0
  ```

```
gv02: ofc-prod-hms-d00.nix.sys.7d
On 1 brick(s)
Result 0: ---------T 2 root root 0 Jul 23 06:43 /export/brick001/gv02/hms/track/000/045
Brick brick001 owner: root size: 0
gv02: ofc-prod-hms-c00.nix.sys.7d
On 2 brick(s)
Result 0: -rw-rw-r-- 2 www-data www-data 5080970 May  5  2015 /export/brick002/gv02/hms
Brick brick002 owner: www-data size: 5080970
Result 1: -rw-rw-r-- 2 www-data www-data 5080970 May  5  2015 /export/brick003/gv02/hms
Brick brick003 owner: www-data size: 5080970
gv02: ofc-prod-hms-c01.nix.sys.7d
On 1 brick(s)
Result 0: ---------T 2 root root 0 Dec 30  2015 /export/brick001/gv02/hms/track/000/045
Brick brick001 owner: root size: 0
On 6 servers and 8 bricks, suggested actions:
ofc-prod-hms-c00.nix.sys.7d: cp /export/brick003/gv02/hms/track/000/045/086/687/17 /tmp
gs2-prod-hms-c00.nix.sys.7d: rm /export/brick002/gv02/hms/track/000/045/086/687/17
gs2-prod-hms-c00.nix.sys.7d: rm /export/brick003/gv02/hms/track/000/045/086/687/17
ctr-prod-hms-d00.nix.sys.7d: rm /export/brick001/gv02/hms/track/000/045/086/687/17
gs2-prod-hms-c01.nix.sys.7d: rm /export/brick001/gv02/hms/track/000/045/086/687/17
ofc-prod-hms-d00.nix.sys.7d: rm /export/brick001/gv02/hms/track/000/045/086/687/17
ofc-prod-hms-c00.nix.sys.7d: rm /export/brick002/gv02/hms/track/000/045/086/687/17
ofc-prod-hms-c00.nix.sys.7d: rm /export/brick003/gv02/hms/track/000/045/086/687/17
ofc-prod-hms-c01.nix.sys.7d: rm /export/brick001/gv02/hms/track/000/045/086/687/17
ofc-prod-hms-c01.nix.sys.7d: rm /srv/gv02/hms/track/000/045/086/687/17
ofc-prod-hms-c00.nix.sys.7d: install -g www-data -o www-data -m 664  /tmp/tmp.JD2j8fnkH
ofc-prod-hms-c00.nix.sys.7d: rm /tmp/tmp.JD2j8fnkHM
```

Fixes

-
  ```
  12:53:54-ccoffey@ciaranc:~/hms/code (master)$ python ./gluster_heal.py
  58 broken file(s) reported
  15 actually broken
  15 file(s) fixed, 0 not fixed
  ```

Questions?