vahidusefzadeh@gmail.com

Cluster File System

همانطور که می دانید در محیط RAC، همه database fileها(از قبیل datafileها، کنترل فایلها و redo log fileها) و فایلهای voting disk و OCR باید در یک فضای مشترک بین نودها قرار بگیرند به همین دلیل در زمان راه اندازی RAC نوع سیستم فایل فضای مورد استفاده بین نودها، باید طوری انتخاب شود تا بتواند امکان sharing را به وجود آورد در حال حاضر سیستم فایلهای مختلفی وجود دارند تا این امکان را فراهم کنند البته مزیت بعضی از آنها، به وضوح بیشتر است برای مثال سیستم فایلی که امروزه اوراکل در مستنداتش به استفاده از آن تاکید دارد، ASM و ACFS می باشد. بعضی از سیستم فایلهایی که این قابلیت را فراهم می کنند، به این اسم هستند:

1. Automatic Storage Manager (**ASM**)
2. ASM Cluster File System (**ACFS**)
3. Oracle Cluster File System (**OCFS2**)
4. Network File System(**NFS**)
5. Red Hat Global File System (**GFS**)

...

در ادامه مطالب مختصری را در مورد این سیستم فایلها ارائه خواهیم کرد.

NFS: برای راه اندازی RAC با استفاده از NFS، باید یک nfs server داشته باشیم که نودها بتوانند از طریق شبکه به آن فضا دسترسی داشته باشند.

برای استفاده از nfs، باید سرویس nfs در همه نودها استارت شده باشد و در فایل /etc/exports سرور nfs، فضا را به شکل زیر تنظیم کرد(البته تنظیم زیر به همه سرورهایی که این سرور را می بینند، اجازه دسترسی به فضا را می دهد که این کار از نظر امنیتی ممکن است مشکلی ایجاد کند):

```
/config   *(rw,sync,no_wdelay,insecure_locks,no_root_squash)
/grid     *(rw,sync,no_wdelay,insecure_locks,no_root_squash)
/home     *(rw,sync,no_wdelay,insecure_locks,no_root_squash)
/dbfile   *(rw,sync,no_wdelay,insecure_locks,no_root_squash)
```

برای اینکه همه نودهای مورد نظر این فضاها را ببینند، باید سرویس nfs را یکبار restart کرد:

```
service nfs restart
```

همچنین باید در فایل /etc/fstab هر کدام از نودها، مشخصات این فضاها را اضافه کرد تا بعد از reboot هر کدام از سرورها، نیازی به شناساندن مجدد فضاها نباشد:

```
nfsserver:/config   /u01 rw,bg,hard,nointr,tcp,vers=3,timeo=600,rsize=32768,wsize=32768,actimeo=0 0 0
nfsserver:/grid     /u02 rw,bg,hard,nointr,tcp,vers=3,timeo=600,rsize=32768,wsize=32768,actimeo=0 0 0
nfsserver:/home     /u03 rw,bg,hard,nointr,tcp,vers=3,timeo=600,rsize=32768,wsize=32768,actimeo=0 0 0
nfsserver:/ dbfile  /u04 rw,bg,hard,nointr,tcp,vers=3,timeo=600,rsize=32768,wsize=32768,actimeo=0 0 0
```

و در نهایت همه این فضاها با دستور زیر مونت می شوند:

```
mount –a
```

از دیگر ویژگی‌های مهم NFS می توان گفت:

۱. راه اندازی بسیار ساده ای دارد.

۲. به یک نقطه متکی است.

۳. کارایی بسیار مفتضحی دارد.

۴. برخلاف اکثر کلاستر فایل سیستم ها، NFS، علاوه بر database files و oracle home، گرید را هم پشتیبانی می کند و می توان گرید را روی آن نصب کرد.

GFS: با توجه به اینکه تجربه استفاده از این قابلیت را تا به حال نداشتم، تنها چند ویژگی از آن را در ادامه ذکر خواهم کرد.

ویژگی ها:

۱. تنها توسط Red Hat پشتیبانی می شود.

۲. زمانی که اندازه دیتافایلها بسیار بزرگ باشد، خوب جواب می دهد به طوری که سرعت I/O آن برای یک نود، از EXT3 بیشتر است.

۳. زمانی که اندازه فایل بسیار کوچک باشد، سرعت بسیار کند خواهد شد.

۴. . قابل shrink کردن نیست.

۵. نسخه GFS تنها 9i تا 10.2 را پشتیبانی میکند و نسخه 11g به بعد را بطور کلی پشتیبانی نمی کند.

۶. برای GFS در نسخه 6.1 اگر همه نودها ۳۲بیتی باشند، حداکثر تا اندازه 16TB و اگر همه نودها ۶۴ بیتی باشند، تا اندازه 8EB پشتیبانی می شود.

OCFS2: امکان پشتیبانی از اوراکل با نسخه های مختلف را دارد البته باید نسخه OCFS2 با نسخه بانک سازگار باشد که جدول زیر شرایط حداقلی را مشخص می کند.

OCFS2_version	DB_version
1.2 , 1.4	9i, 10gR1, 10gR2, 11gR1
1.4.1	11.2.0.2 , 12c

برای تعیین نسخه OCFS2 نصب شده در OS، می توانیم از دستور زیر استفاده کنیم:

```
cat /proc/fs/ocfs2/version
```

```
OCFS2 1.2.9 Tue May 19 07:43:29 EDT 2009 (build a896806cb852dd7f225004092b675ede)
```

که البته برای لینوکس 6.0 و 5.6 نسخه OCFS2 برابر با 1.6.3 می باشد.

vahidusefzadeh@gmail.com

بر روی این نوع از کلاستر سیستم فایل می توان اوراکل را نصب کرد(البته در حالت RAC منظور است) همچنین در 10g می توانستیم بر روی این فایل سیستم نصب کلاستر داشته باشیم ولی در 11gR2 به بعد، با آمدن گرید، امکان نصب آن بر روی ocfs2 ممکن نیست هر چند که استفاده از این کلاستر سیستم فایل از 11gR2، کمتر متداول است.

برای استفاده از این ویژگی، ابتدا باید packageهای مربوط به OCFS2 را نصب کرده باشیم و بعد ازآن باید مراحل زیر را روی همه نودها انجام داد:

```
[root@rac1 ~]# service o2cb configure
Load O2CB driver on boot (y/n) [y]: y
Cluster stack backing O2CB [o2cb]:
Cluster to start on boot (Enter "none" to clear) [ocfs2]: ocfs2
Specify heartbeat dead threshold (>=7) [31]:
Specify network idle timeout in ms (>=5000) [30000]:
Specify network keepalive delay in ms (>=1000) [2000]:
Specify network reconnect delay in ms (>=2000) [2000]:
Writing O2CB configuration: OK
Setting cluster stack "o2cb": OK
Cluster ocfs2 already online
[root@rac1 ~]# /etc/init.d/o2cb load
[root@rac1 ~]# /etc/init.d/o2cb start
Setting cluster stack "o2cb": OK
Cluster ocfs2 already online
```

بعد از اجرای دستورات ذکر شده، باید مشخصات نودها را در فایل /etc/ocfs2/cluster.conf قرار داد که این کار هم از طریق کنسول و هم از طریق دستور قابل انجام است برای استفاده از کنسول OCFS2 نیاز است تا package اضافه نصب شود(برای مثال ocfs2console-1.2.2-1.i386.rpm). در صورت استفاده از کنسول، می توان به راحتی مشخصات را اضافه کرد و در صورتی که از کنسول استفاده نشود، باید فایل مذکور را به صورت زیر اصلاح کرد:

```
node:
        ip_port = 7777
        ip_address = 192.168.20.50
        number = 0
        name = rac1
        cluster = ocfs2
node:
```

```
    ip_port = 7777
    ip_address = 192.168.20.52
    number = 1
    name = rac2
    cluster = ocfs2
cluster:
    node_count = 2
    heartbeat_mode = local
    name = ocfs2
```

بعد از این مرحله، می توانیم با دستور زیر فضا را با این کلاستر سیستم فایل، فرمت کنیم:

```
mkfs.ocfs2  -b 4K  -C 32K  -N  2  -L  oracle  /dev/sda1
```

منظور از N– تعداد نودها است و L– هم برچسب را مشخص می کند.

برای mount این فضا، باید بسته به محتوای دیسک، گزینه هایی را مشخص کرد برای مثال اگر بخواهیم این فضا database file را در خود نگه دارد، باید از دو خصیصه datavolume,nointr استفاده کرد در صورتیکه اوراکل مورد استفاده، 10g باشد و قرار باشد clusterware بر روی این فضا قرار بگیرد، در صورت عدم از استفاده از این خصیصه، سرویسهای clusterware استارت نخواهند شد.

```
mount –o  datavolume,nointr, _netdev  -t ocfs2  /dev/sda1  /db
```

در صورتی که قصد داشته باشیم oracle home را در این مسیر قرار دهیم، استفاده از خصیصه های ذکر شده لازم نیست:

```
mount -t  ocfs2  /dev/sdb1  /orahome
```

از دیگر ویژگی‌های مهم OCFS2 می توان گفت:

۱. در سیستم عاملهای Oracle Linux, RHEL پشتیبانی می شود.

۲. سیستم فایل OCFS2 مشکلات سیستم فایل GFS را ندارد و از نظر کارایی از GFS بهتر می باشد.

۳. مونت فضا توسط سیستم عامل انجام می شود(برخلاف ACFS).

۴. direct I/O و asynchronous I/O را پشتیبانی می کند.

۵. block size می تواند در اندازه های 512 byte ، 1K ، 2K، 4K تعیین شود که اندازه 4K توصیه می شود و Cluster Size که مشخص کننده کوچکترین واحد تخصیص فضا به یک فایل می باشد می تواند یکی از مقادیر 4K، 8K، 16K، 32K، 64K ، 128K، 256K ، 512K و 1M را به خود بگیرد.

۶. اگر cluster size برابر با 1M باشد، حداکثر اندازه این سیستم فایل می تواند برابر با 4PB باشد. اگر cluster size برابر با 4KB باشد، اندازه فایل سیستم حداکثر برابر با 16TB خواهد بود.

۷.حداکثر اندازه یک فایل می تواند 16TB باشد و حداکثر تعداد subdirectory هم برابر با 32000 خواهد بود.

ACFS : همانطور که می دانیم هنگام استفاده از ASM بعضی از قابلیتهای سیستم فایلهای دیگر به مثل OCFS2، ext3 و ... برای کاربر سخت تر بدست می آید و با استفاده از دستورهای رایج سیستم عامل، تحصیل آن ممکن نخواهد بود و همچنین تنها می توان چند نوع محدود از فایلها را در آن قرار داد از اوراکل 11gR2 این امکان به وجود آمد که در لایه ای بالاتر از ASM، فضایی ایجاد شود تا بتوان با آن، این کمبودهای ASM را برطرف کرد. این قابلیت جدید، ACFS نام دارد.

همانطور که گفته شد، این کلاستر سیستم فایل بر روی ASM قرار می گیرد و منابع آن توسط گرید مدیریت می شود و بسیاری از فایلها از قبیل فایلهای application، فایلهای اجرایی، trace file، alert log و حتی نرم افزار اوراکل را می توان در ACFS ذخیره کرد ولی ویژگی مهمی که از اوراکل 12c به آن اضافه شد، پشتیبانی از database fileها می باشد البته کماکان نمی توان گرید را بر روی آن نصب کرد.

برای استفاده از این ویژگی می توان از asmca یا command line استفاده کرد که در این قسمت دستوراتی را که برای ایجاد فضایی با سیستم فایل ACFS در اوراکل ۱۲.۱ استفاده شده را می بینید:

۱. فرض کنید که یک DISKGROUP با نام usef داریم که در حال حاضر مونت شده است:

```
ASMCMD [+] > lsdg
State    Type    Rebal  Sector  Block       AU  Total_MB  Free_MB  Req_mir_free_MB  Usable_file_MB  Offline_disks  Voting_files  Name
MOUNTED  EXTERN  N         512   4096  1048576   7152     3608                0            3608              0             Y  USEF/
```

```
[root@rac1 ~]# oracleasm createdisk usef  /dev/sdg1
Writing disk header: done
Instantiating disk: done
```

```
CREATE DISKGROUP usef  EXTERNAL REDUNDANCY  DISK '/dev/oracleasm/disks/USEF' SIZE 1019M
ATTRIBUTE 'compatible.asm'='12.1.0.0.0','au_size'='1M';
```

مقداری که compatible.asm می گیرد باید بزرگتر از 12.1 باشد البته برای 11gR2 باید بزرگتر از 11.2 باشد.

۲. باید بر روی این DISKGROUP یک volume بسازیم و سپس آن را فعال کنیم:

```
ALTER DISKGROUP USEF  ADD VOLUME volume11   SIZE 819200K;
alter diskgroup USEF  enable volume 'volume11';
```

با دستور زیر، مشخصات این volume را خواهیم دید :

vahidusefzadeh@gmail.com

ASMCMD [+] > volinfo -G USEF volume11

Diskgroup Name: USEF

Volume Name: VOLUME11

Volume Device: /dev/asm/volume11-151

State: DISABLED

Size (MB): 832

Resize Unit (MB): 64

Redundancy: UNPROT

Stripe Columns: 8

Stripe Width (K): 1024

Usage:

Mountpath:

۳. با دستور زیر volume را به فرمت ACFS در می آوریم:

[root@rac1 ~]# /sbin/mkfs -t acfs /dev/asm/volume11-151

mkfs.acfs: version   = 12.1.0.2.0

mkfs.acfs: on-disk version   = 39.0

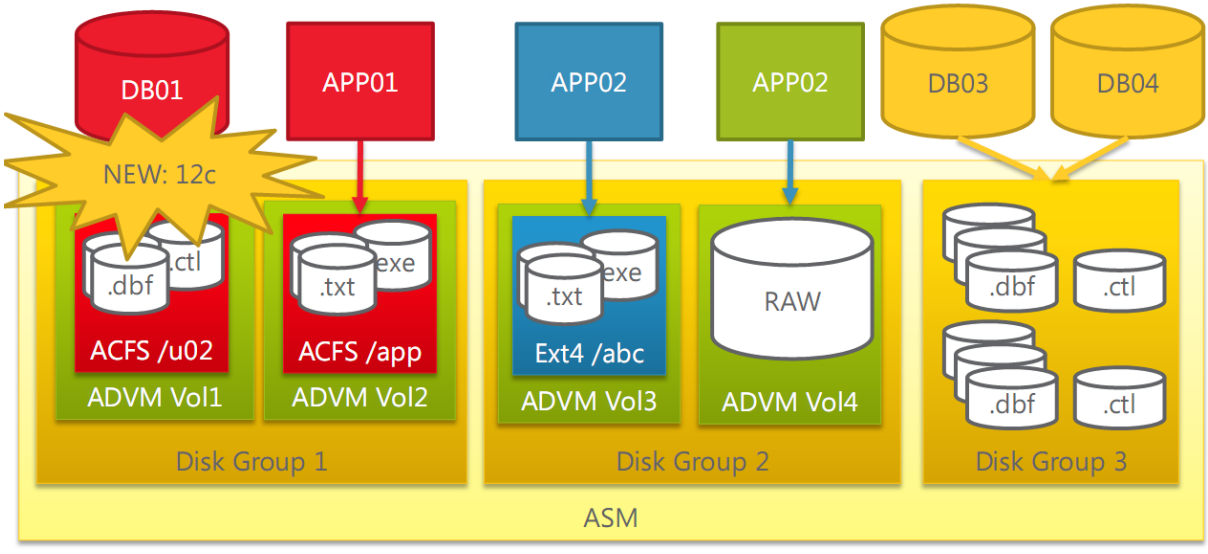
mkfs.acfs: volume   = /dev/asm/volume11-151

mkfs.acfs: volume size   = 872415232 (832.00 MB)

mkfs.acfs: Format complete.

✓ نکته ای خارج از موضوع:

البته این امکان وجود دارد که این volume را به فرمت های دیگر به مثل ext3 هم در آوریم:



SQL> ALTER DISKGROUP USEF ADD VOLUME volume12 size 1M;

Diskgroup altered.

[oracle@rac2 ~]$ /sbin/mkfs –t ext3 /dev/asm/volume12-151

mke2fs 1.43-WIP (20-Jun-2013)

Filesystem label=

OS type: Linux

Block size=1024 (log=0)

Fragment size=1024 (log=0)

Stride=0 blocks, Stripe width=0 blocks

16384 inodes, 65536 blocks

3276 blocks (5.00%) reserved for the super user

First data block=1

Maximum filesystem blocks=67108864

8 block groups

8192 blocks per group, 8192 fragments per group

2048 inodes per group

Superblock backups stored on blocks:

8193, 24577, 40961, 57345

Allocating group tables: done

Writing inode tables: done

Creating journal (4096 blocks): done

Writing superblocks and filesystem accounting information: done

[root@rac2 ~]# mount -t ext3 /dev/asm/volume12-151 /ext_dir

[root@rac2 ~]# df -h /ext_dir/

Filesystem Size Used Avail Use% Mounted on

/dev/asm/volume12-151 62M 5.4M 54M 10% /ext_dir

۴. پوشه ای که می خواهیم فضا به آن مونت شود را می سازیم و با دستور acfsutil، مشخصات فضای acfs را ثبت می کنیم این دستور به نوعی کار /etc/fstab را انجام می دهد.

[root@rac1 ~]# mkdir /acfs_usef11

[root@rac1 ~]# chown -R oracle.oinstall /acfs_usef11/

[root@rac1 ~]# /sbin/acfsutil registry -a /dev/asm/volume11-151 /acfs_usef

acfsutil registry: ACFS-03132: mount point /acfs_usef already exists in the Oracle Registry as:

vahidusefzadeh@gmail.com

Mount Object:

Device: /dev/asm/volume1-194

Mount Point: /acfs_usef

Disk Group: USEF_GRP1

Volume: VOLUME1

Options: none

Nodes: all

همچنین برای از رجیستر درآوردن acfs_usef می توان از دستور زیر استفاده کرد:

```
[root@rac1 ~]# /sbin/acfsutil registry -d /acfs_usef
```

برای مونت کردن این فضا و نیز از مونت درآوردن آن، از دو دستور زیر استفاده می کنیم:

```
[root@rac1 ~]# /bin/mount   -t   acfs   /dev/asm/volume1-194     /acfs_usef

[root@rac1 ~]# umount    /acfs_usef
```

همانطور که در این قسمت می بینید، فضا به هر دو نود اضافه شده است:

```
[root@rac1 ~]# df -h /acfs_usef/

Filesystem          Size  Used Avail Use% Mounted on

/dev/asm/volume1-194  832M   79M  754M  10% /acfs_usef

[root@rac2 ~]# df -h /acfs_usef/

Filesystem          Size  Used Avail Use% Mounted on

/dev/asm/volume1-194  832M   79M  754M  10% /acfs_usef
```

البته راه دیگری هم برای انجام مرحله آخر وجود دارد:

```
[root@rac2 bin]# ./srvctl add   filesystem     -d /dev/asm/volume1-194      -m  /acfs_usef     -u oracle
-fstype ACFS    -autostart     ALWAYS

 [root@rac2 bin]#  ./srvctl  start  filesystem  –d  /dev/asm/volume1-194
```

از دیگر ویژگی‌های مهم ACFS می توان گفت:

۱. می تواند برای کلاستر یا single استفاده شود(برخلاف OCFS2).

۲. بصورت آنلاین می توان فضا را کاهش یا افزایش داد:

```
[root@rac1 ~]# /sbin/acfsutil size +450M –d /dev/asm/vvolume1-194   /acfs_usef
```

acfsutil size: new file system size: 805306368 (768MB)

[root@rac1 ~]# /sbin/acfsutil size -150M –d /dev/asm/vvolume1-194   /acfs_usef

acfsutil size: new file system size: 671088640 (640MB)

۳. می توان از اطلاعات آن snapshot تهیه کرد:

[root@rac1 ~]# /sbin/acfsutil snap create snap1 /acfs_usef

از مزایای snapshot، می توان به بازگردانی فایل به یک نسخه قدیمی موجود در snapshot(برای نمونه می توان از این ویژگی برای برگشت به قبل از اعمال patch استفاده کرد) و بازیابی فایل حذف شده اشاره داشت.

البته برای بکاپ گیری از دیتا با این ویژگی، شاید راه بهتر این باشد که از این ویژگی در data guard استفاده کنیم تا بار اضافه ای بر روی بانک اصلی تحمیل نشود در صورت انجام این کار(استفاده از data guard) توصیه می شود برای نگاستن کارایی بانک از پروسس arch به جای LGWR استفاده کنیم یعنی data guard در مود maximum performance اجرا شود.

۴. Oracle Cluster Registry (OCR) و voting files را پشتیبانی نمی کند.

۵. مونت فضا برای ACFS بر عهده گرید می باشد.

۶. از اوراکل 12c می توان دیتافایل را در ACFS قرار داد:

SQL> create tablespace usef_tbs datafile '/acfs/data01.dbf' size 5m;

Tablespace created.

SQL> select name from v$datafile;

+DATA01/RAC/DATAFILE/system.279.891703291

+DATA01/RAC/DATAFILE/undotbs2.291.891706449

+DATA01/RAC/DATAFILE/sysaux.278.891703149

+DATA01/RAC/DATAFILE/undotbs1.281.891703587

/acfs/data01.dbf

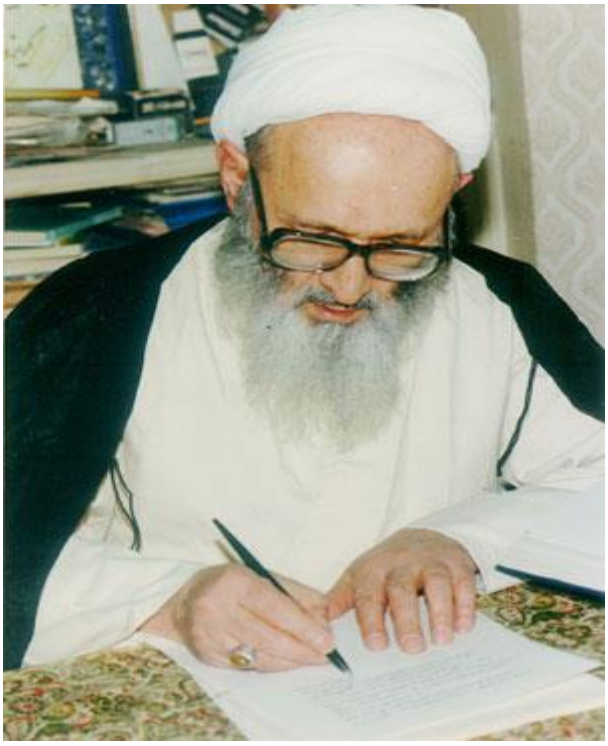
+DATA01/RAC/DATAFILE/users.280.891703585

6 rows selected.

۷. ACFS از نظر کارایی تقریبا شبیه به ASM است و ویژگی های ASM از قبیل striping، mirroring، rebalancing و .. را در خود دارد. البته در بعضی از منابع گفته شده که ACFS به خاطر لایه اضافه تری که ایجاد می کند، نمی تواند کارایی ASM را داشته باشد.

۸. از دیگر ویژگی های ACFS می توان به replication، tagging، security و encryption اشاره کرد.

vahidusefzadeh@gmail.com



علامه حسن زاده آملی:

الهی از من برهان توحید خواهند و من دلیل تکثیر.