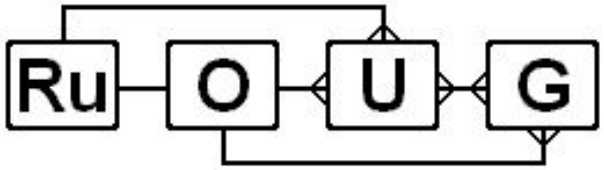


Деревянко Михаил

+ASM v.2009

Agenda

1. Сравнение ASM с аналогичными продуктами/технологиями.
2. Архитектура ASM.
3. Исследование внутренних структур ASM.
4. Опыт использования, проблемы, рекомендации.

Путь к ASM

- Standard I/O;
- Synchronous(Buffered) I/O - filesystem; biowait()
- Direct io filesystem; O_DIRECT
- Asynchronous I/O; aio

`filesystemio_options`

- filesystem;
- filesystem with mount flags;
- Raw device;
- Raw device over volume managers;
- ASM.

Преимущества ASM

- Облегчает администрирование (Oracle management files);
- Более низкая стоимость по сравнению с Volume manger's;
- Отсутствие ограничений, отличающих файловую систему;
- Mirror and Striping = Same ?;
- Отдал диски и забыл ☺;
  - Необходимость иметь одинаковые по размеру/производительности диски, чем больше диск по размеру тем больше данных туда разместит asm.
  - Возможность работе в RAC.
  - При выводе диска из строя нагрузка размазывает согласно PST до 10 дисков, аналогично при ребалансе скачок нагрузки "мягче".
  - Volume manager Drl vs asm mirror resilvering (asm учитывает особенности файлов)

Oracle RDBMS processes directly access the storage!

Недостатки ФС

Файловые системы:

-Следить за местом, авторасширение файлов* (?)

- Права на каталоги/тома…
- Имена файлов/каталогов;
- Выбор размера страйпа;

* Интересный баг Auto extend aud$.

Volume Manager

Проблемы есть везде;

- Возможность работы с большим кол-вом томов (high-end решения) или путей к дискам, все таки ASM не готов к этому…

-ASM=OSM = USM ?

- Отказ от raw в 12 ?

Набор утилит для работы с oracle, н-р, проверка db_block_cheksum - asm это делает на лету [scn…scn]

*интересный баг - ошибка в правах на redo logs - crash instance;

Надежность

Что надежней:

- LVM - ?
- VxVM - ?
- ASM - ?
- FS - ?

Рекомендовать какую-то одну из указанных технологий нельзя, так как любые проблемы стабильности выявляются только в результате длительной промышленной эксплуатации.

ASM

- ASM + linux = love? asmlib ☺;
- При обновлении ядра практически всегда есть в наличии актуальный asmlib (регулярно обновляется);

+<group>/<dbname>/<file_type>/<tag>.<file#>.<incarnation#>

* db_name берется с учетом db_unique_name

asmlib

- Device discovery

- I/O processing
  ASMLIB не kernel aio, а свой механизм.
  Reduce the nymber of calls to OS.
  A single call to asmlib multiple i/o.
  Открывает меньше file descriptors.
- Perfomance
  low cpu cost on high loaded system

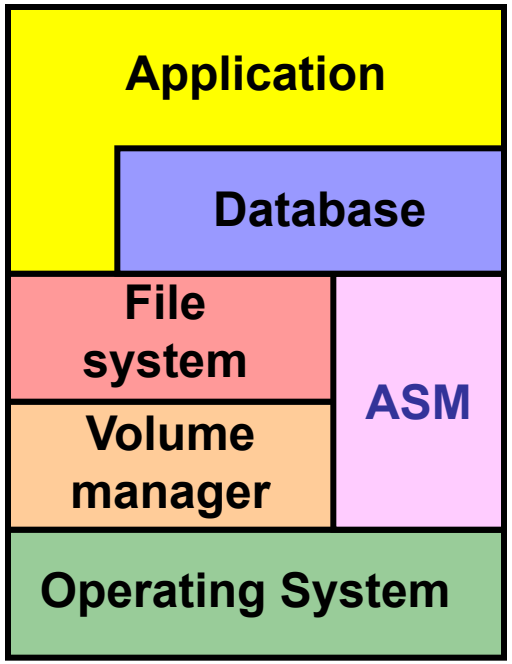
Архитектура ASM

instance_type = ASM

SQL > select instance_name from v$instance;

INSTANCE_NAME

+ASM



asm<->CSSD<-> rdbms

Cluster Synchronization Service Daemon

/opt/oracle/product/11.1.0/db_ruoug2/bin/localconfig reset

nohup /etc/init.d/init.cssd run >/dev/null 2>&1 </dev/null &

Архитектура ASM

Основные процессы:

ARBx – процессы ребаланса

CKPT – cross instance calls

DBWR,PMON,PSP0,PZ9x – аналогично RDBMS

SMON – общается с CSS

LGWR – пишет ACD(active change directory)

GMON (занимается drop/offline disk)

KATE (занимается online disk)

Архитектура ASM

Память:

- Memory_target – 256M;
- _db_cache_size (блоки по 4k);

+ SHARED_POOL_SIZE

  (DB_SPACE/100+2) External

  (DB_SPACE/50+4) Normal

  (DB_SPACE/33+6) High

+ 16 processes;

+ 1m large pool;

* Зависит от версии 11g. При открытии файла передает только direct extents, для закрытых файлов место в памяти не нужно

Архитектура ASM

- AU – allocation unit;
- Расположение au (зависит от размера дисков и никоим образом не зависит от i/o на диск!)

Datafile – 1 mb; много или мало ?

1m=8k*128 MBRC(8) по 16 блоков на несколько дисков*

Oracle “Read ahead” ?

Controlfile, redo 128kb; **

* Lewis блоки в памяти могут “помешать” идти fullscan и спровоцировать одноблочные чтения…

** controlfile при normal redur. имеет 3 копии

Основные представления ASM

v$asm_alias; (-)
v$asm_attribute(11g diskgroup properties); (-)
v$asm_client;
v$asm_disk (service oracleasm scandisk); -- читает заголовки дисков
v$asm_disk_iostat (rdbms only connected); (-)
v$asm_disk_stat;
v$asm_diskgroup; -- читает заголовки дисков
v$asm_diskgroup_stat;
v$asm_file;(-)
v$asm_operation; (-)
v$asm_template;

Asm reduradancy and failgroup

- External,normal(x2),high(x3)
- Failgroup (HBA,path,array,group)

Normal на нечетном количестве дисков.

Reduradancy не меняется в течение всего времени жизни Diskgroup.

Primary extent (+0-2 second extend) читаем всегда с primary extent, если он доступен; если нет, то с secondary до rebalance.

- 11g preffered reads (необходимо поднять версию dg)
- Partner Disks x$kfdpartner до 10, но всегда в других fg (PST)

ASM везде

- Asm на windows ☺ asmtool…
- Asm multipathing device mapper vs multipathd on linux ?

v$asm_diskgroup

Name	Type
GROUP_NUMBER	NUMBER
NAME	VARCHAR2(30)
SECTOR_SIZE	NUMBER
BLOCK_SIZE	NUMBER
ALLOCATION_UNIT_SIZE	NUMBER
STATE	VARCHAR2(11)
TYPE	VARCHAR2(6)
TOTAL_MB	NUMBER
FREE_MB	NUMBER
REQUIRED_MIRROR_FREE_MB	NUMBER
USABLE_FILE_MB	NUMBER
OFFLINE_DISKS	NUMBER
UNBALANCED	VARCHAR2(1)
COMPATIBILITY	VARCHAR2(60)
DATABASE_COMPATIBILITY	VARCHAR2(60)

V$asm_diskgroup

```
SQL> select
    TOTAL_MB,FREE_MB,REQUIRED_MIRROR_FREE_MB,
    USABLE_FILE_MB from v$asm_diskgroup;
```

- TOTAL_MB 3 200 000
- FREE_MB 1 800 000
- REQUIRED_MIRROR_FREE_MB 200 000
- USABLE_FILE_MB 800 000

USABLE_FILE_MB –сколько места еще свободно с учетом miiror. может быть меньше 0!

ASM extents

- Oracle extent транслируются в AU asm, asm не участвует в операциях ввода вывода, он участвует в ребалансе и поддержке актуальной карты экстентов.
- Можно попробовать настроить _asm_ausize и _asm_stripesize.*
  - Если диски одинаковые по размеру, то на них получится одинаковое число au.
- После каждых 20000 экстентов оракл увеличивает размер экстента AU x8 для каждого файла.
- Если получился перекос нагрузки то rebalance должен это поправить**

*Не встречал пока проблем, что данные были неадекватно отбалансированы, с большими объемами на ASM не пробовал.

**Для ребаланс идет обратный отсчет в v$asm_operation, но после окончания он еще доделывает накопившуюся во время ребаланса работу, поэтому долгое время стоит с 0.

asmcmd

asmcmd help (oracle11)

cd
cp
du
find
ls
lsct
lsdg
mkalias
mkdir
pwd
rm
rmalias

md_backup
md_restore

lsdsk
remap - пометить плохие сектора на диске

ASM metadata

В заголовке каждого диска:

- Disk Header

(disk name, disk number, diskgroup name, failure group name, disk size, AU size, creation time, mount time, compability ASM/RDBMS)

- Allocation Table (AT)

  (Allocation table blocks(ATB), Allocation table Entry(ATE) показывает file#,extent#->AU или free)

- Free Space Table (FST)

  (аналогично AT, только показывает пустые)

- Partnership Status Table (PST)

ASM metadata (Partnership Status Table)

- Version number;
- Timestamp;
- PST size (number of disks);
- Number of PST copies;
- Disk list for PST;
- Compability.

информация о дисках партнерах:

Disk status, number of partners, list of partners.

Последний блок PST heart beat необходим для того, чтобы диск нельзя было смонтировать с нескольких серверов.

Asm virtual Metadata

Находится в специальных asm файлах, доступных только для asm инстанса.

- File directory (file size, file block size, file type,redudancy,striping,creation time,modification,file layout – 60 directs extents, up 300 inderect);
- Disk directory(disk name, failure group name,disk size,disk free space,disk creation time);
- Active change directory(ACD);
- Continuing Operations Directory (COD);
- Template Directory;
- Alias directory;
- Attribute Directory;
- Staleness directory;
- Staleness Registry.

Asm virtual Metadata

File#0, AU=1: Partner Status Table (PST)

File#1: File Directory (files and their extent pointers)

File#2: Disk Directory

File#3: Active Change Directory (ACD)

- The ACD is analogous to a redo log, where changes to the metadata are logged.
- Size=42MB * number of instances

Asm virtual Metadata

File#4: Continuing Operation Directory (COD).

- The COD is analogous to an undo tablespace. It maintains the state of active ASM operations such as disk or datafile drop/add. The COD log record is either committed or rolled back based on the success of the operation.

File#5: Template directory

File#6: Alias directory

11g, File#9: Attribute Directory

11g, **File#12: Staleness registry**, created when needed to track offline disks

Ребаланс дисковой группы

Ребаланс не всегда зависит от скорости дисков:

Single instance:

-Buffer busy wait

RAC

- DFS lock handle (cross-instacnce lock)
- Buffer busy wait

На одной ноде ребаланс идет быстрее, и еще быстрее при startup restrict

При ребалансе происходит relocation экстентов, если экстент принадлежит файлу со статусом open, тогда на время переноса замораживается обращение со стороны rdbms к этому экстенту(операции, которые “ждали” экстент прочитают его со старого места)

** asm file names

Cern IT

View Name	X$ Table	Description
V$ASM_DISKGROUP	**X$KFGRP**	performs disk discovery and lists diskgroups
V$ASM_DISK	**X$KFDSK, X$KFKID**	performs disk discovery, lists disks and their usage metrics
V$ASM_FILE	**X$KFFIL**	lists ASM files, including metadata
V$ASM_ALIAS	**X$KFALS**	lists ASM aliases, files and directories
V$ASM_TEMPLATE	**X$KFTMTA**	ASM templates and their properties
V$ASM_CLIENT	**X$KFNCL**	lists DB instances connected to ASM
V$ASM_OPERATION	**X$KFGMG**	lists current rebalancing operations
N.A.	**X$KFKLIB**	available libraries, includes asmlib
N.A.	**X$KFDPARTNER**	lists disk-to-partner relationships
N.A.	**X$KFFXP**	extent map table for all ASM files
N.A.	**X$KFDAT**	allocation table for all ASM disks

Cern IT

Column Name	Description
NUMBER_KFFXP	ASM file number. Join with v$asm_file and v$asm_alias
COMPOUND_KFFXP	File identifier. Join with compound_index in v$asm_file
INCARN_KFFXP	File incarnation id. Join with incarnation in v$asm_file
XNUM_KFFXP	ASM file extent number (mirrored extent pairs have the same extent value)
PXN_KFFXP	Progressive file extent number
GROUP_KFFXP	ASM disk group number. Join with v$asm_disk and v$asm_diskgroup
DISK_KFFXP	ASM disk number. Join with v$asm_disk
AU_KFFXP	Relative position of the allocation unit from the beginning of the disk.
LXN_KFFXP	0->primary extent,1->mirror extent, 2->2nd mirror copy (high redundancy and metadata)

Reading ASM files with OS tools, using metadata information from X$ tables.

Reading strace -> asm files.

The tnsnames entry can be used to connect to ASM instances via Oracle*NET - the extra keyword (UR=A). UR=A allows to connect to 'blocked services'.

*applies to 10g, it is not needed in 11g.

11g Features

- 11g restricted in 11g – quicker rebalance;

preffered read;

- au-size (1|2|4|8|16|32|64MB);
  - FAST DATA RESYNC, в volume manager такого нет! (staleness registry)

Alter diskgroup data disk_repair_time='4 h';

- Alter diskgroup online disk all; -не падало
  - Настройка au-size для DG
  - Если asm сталкивается с невозможностью прочитать primary extent, то он пытается записать его в другое место на этом диске, а старое пометить как unusable; если возникает ошибка при записи, то сломавшийся диск asm переводит в offline (если и header не читается?) + пишет об ошибке в alert.log (если удалось записать хоть 1 копию, то для app ошибки нет) .

Лечение проблем

- Большая часть проблем решается обнулением заголовка asm дисков dd if=/dev/zer of=/dev/sdv3.
- Если диск “глючит”, то этого никак не видно, т.к. asm не ведет счетчик ошибок - только OS. Приходится менять его вручную.
- Alert.log для asm сложно читаем, хотя через некоторое время он становится понятней, если возникают ошибки невозможности прочитать блок - он пишет что пытается прочитать с зеркала (rdbms alert?)
- Авто-Offline диск только для для операций записи или чтения заголовка…
- Большую часть проблем и падений в процессе/post-процессе ребаланса 10,11 рекомендуется отложить до возможности разобраться с ними в спокойной обстановке.
- Добавлять диски с force в 11

Лечение проблем

- Висит vasm_disk, vasm-diskgroup-”приплыли”;
- ALTER DISKGROUP DATA CHECK;
- Редактор kfed

– `MOUNT FORCE`:
  - Переведет диски в offline
  - Не будет работать, если все диски в норме(10g);

Рекомендации

Иметь две DG, Data+Backups(FRA)

И если есть различные группы оборудования, то вынести в отдельную группу.

Up2date kernel

Список используемой литературы:

1. Oracle documentation
2. Cern docs: https://twiki.cern.ch/twiki/bin/view/PSSGroup/ASM_Internals

Вопросы и ответы