From: Spitzer, Andras ([redacted])
Sent: Wednesday, July 08, 2009 11:18 PM
To: [redacted]
Cc: [redacted]
Subject: RE: Actimize QA DB Disk Performance

Matt,

I've been working on this issue in the last few days, and came to the following :

Analysis :

This cluster is an **Oracle 10R2 RAC with ASM**, the clusterware is Oracle also, **no VxVM/VCS installed/configured, volume manager and filesystem layer is handled by the Oracle ASM instance**. The two nodes are myukslulonadb01 and myukslulonadb02 running Linux, and each server is **connected to EMC CLARiiON** storage via 2 individual Emulex LP11002 HBAs. As we have dual path toward the storage frame and **Oracle ASM doesn't provide multipath functionality, EMC PowerPath 5.0 is installed** on both servers.

First, I ran some basic OS commands to check the health and throughput of the individual LUNs, and I found **unreasonably high service times and high utilization** (~100%) even during low I/O traffic. Further checking this weird behaviour, I found that the EMC PowerPath devices are not equally used as their underlying devs, it happened many times that the underlying device was way more loaded than the emcpower device. Normally this should never happen, because if you have EMC PowerPath installed, even though EMC PP allows you to access directly the devices beneath, it's an invalid use in a multipathing environment. Normally you have to access the LUNs via EMC PowerPath devices, which will provide you fault tolerant failover in case of a link/HBA failure, will manage the CLARiiON frame in case of a LUN trespass scenario.

In this **scenario it's even more important to use EMC PowerPath**, because the CLARiiON frame is an Active/Passive frame which means that the two HBA links are connected to CLARiiON's SP (Service Processors, namely SP A and SP B) and only one path can be active at a time (called LUN ownership, only one SP can be the owner of a specific LUN at a time), and in case of a LUN trespass (trespass is the process when SPs has to switch ownership over a LUN) only the Multipathing software can control the process (storage group's failovermode is set to 1, which means the multipathing software has to initiate the trespass, it won't happen automatically).

One additional reason for using EMC PP, because it's a policy to have the default ownerships of the LUNs shared among the SPs, for example half of the LUNs owned by SP A by default while the other half is owned by SP B, **sharing the load** on the two HBAs, links, and SPs.

By default in your ASM instance you have to refer to the device names in V$ASM_DISK, but in your ASM instance you have ORCL: symbolic names, which **indicated to me that you are using ASMLib for Linux**, which is an addon to Oracle ASM. ASMLib puts a small header on the disks which are put under ASMLib control, **creates a special filesystem area under /dev/oracleasm**, and during ASMLib startup it scans the /dev directory for valid ASMLib headers, when it finds one, it creates the device under /dev/oracleasm/disks with the ASMLib disk name (in your case it's ASMVOL01 - ASMVOL47 on your QA cluster, and ASMVOL01 - ASMVOL28 on your production cluster) and the major/minor numbers it found under /dev.

The problem here is that if **you have multiple paths to each CLARiiON LUNs** (in our case 3 different path to the same LUN, two paths for the two HBAs, and a 3rd pseudo path via EMC PowerPath) so you have to **configure ASMLib otherwise ASMLib will pick up the first which is listed under /dev which can be any of the 3**, and this can lead to serious problems. If you are using a multipathing software (especially if you use it with an Active/Passive SAN like CLARiiON) you want ASMLib to work with only the pseudo devices. I ran emcgrab, collected the /dev and /dev/oracleasm/disks contents, and wrote a small program which was able to resolve which paths ASMLib is currently using to access the CLARiiON LUNs, here is an example of the output (NODE 01 is myukslulonadb01 and NODE02 is myukslulonadb02 in this output) :

	NODE 01			NODE 02		
ASM Volume	**DEVICE**	**CLAR LUN**	**SP**	**DEVICE**	**LUN**	**SP**
ASMVOL01	**/dev/sdh**	0x01ae	A	**/dev/sdh**	0x01ae	A
ASMVOL02	**/dev/sdg**	0x01aa	A	**/dev/sdg**	0x01aa	A
ASMVOL03	**/dev/sdf**	0x01ac	A	**/dev/sdf**	0x01ac	A

```
ASMVOL04        /dev/sde            0x01a8   A  /dev/sde            0x01a8   A
...
ASMVOL43        /dev/sdaw           0x00c2   A  /dev/emcpoweraw     0x00c2   A
ASMVOL44        /dev/sdax           0x00c4   A  /dev/emcpowerax     0x00c4   A
ASMVOL45        /dev/sday           0x00c6   A  /dev/emcpoweray     0x00c6   A
ASMVOL46        /dev/sdaz           0x00c8   A  /dev/emcpoweraz     0x00c8   A
ASMVOL47        /dev/sdba           0x00ca   A  /dev/emcpowerba     0x00ca   A
```

As you can see, **Oracle ASMLib is not configured properly**, as it's using the devices beneath EMC PowerPath, and only some cases it's using the proper /dev/emcpower* pseudo devices . On myukslulonadb01 all the devices are referring to the devices beneath, on myukslulonadb02 33 out of 47 is referring to use the devs directly, as a summary **80 devices out of 94 is using the devices directly, which means if it would occur a LUN trespass** (which can happen anytime, a glitch in the SCSI subsystem, a maintenance on the CLARiiON frame, this is absolutely normal and not considered an issue) **now, your Oracle instance would take those LUNs immediately offline, most likely bringing down both nodes on your Oracle RAC.**

The real scary fact is that **even your production server is configured** (or misconfigured if you like) **the same way**, which means the possibilities I wrote above all true for your production servers as well. Also, such configuration when you don't use the EMC PowerPath devices even though EMC PowerPath is installed (and failovermode is set to 1 on CLARiiON) can be the **root cause of a major performance degradation accessing those LUNs**. Frankly speaking with failovermode 1 set on your storagegroup (which is the correct setting if you have EMC PowerPath installed) you shouldn't even be able to communicate via the non-owner SP, so we can say you are lucky that it even works today (as I can see you communicate via the non-owner SP to many of the LUNs).

Every possible documentation states that in case of failovermode set to 1 (called the Passive Not Ready mode), you can communicate only via the owner SP, the non-owner SP should reject any I/O requests, so I don't really understand why it works here, but this is the beauty of the CLARiiON internals. Anyway, in short, it's essential to use EMC PowerPath pseudo devices (starts with /dev/emcpower*), and not the standard /dev/sd* devices.

During my investigation I also found that **on myukslulonadb02 you had a missing device name for /dev/sdbm**. You use /dev/sdbm1 (1 at the end of the device referes to the partition number) as ASMVOL28, but because you missed the whole-disk device name /dev/sdbm, EMC PowerPath couldn't identify the device so it was missing from it's configuration :

```
Pseudo name=emcpowerah
CLARiiON ID=CK200071500501 [myukslulonadb_cluster]
Logical device ID=60060160C02B1D00E0AF659FD8C6DD11 [LUN 1142]
state=alive; policy=CLAROpt; priority=0; queued-IOs=0
Owner: default=SP B, current=SP A
==============================================================================
---------------- Host ---------------   - Stor -   -- I/O Path -  -- Stats ---
###  HW Path                I/O Paths    Interf.   Mode    State  Q-IOs Errors
==============================================================================
   2 lpfc                     sdbq      SP B0     active  alive      0      0
   0 UNKNOWN                  unknown   SP A0     active  alive      0      0
```

I **fixed this,** simply created the appropriate device under /dev, and after a few hours **EMC PowerPath** already **picked up, now correctly** :

```
Pseudo name=emcpowerah
CLARiiON ID=CK200071500501 [myukslulonadb_cluster]
Logical device ID=60060160C02B1D00E0AF659FD8C6DD11 [LUN 1142]
state=alive; policy=CLAROpt; priority=0; queued-IOs=0
Owner: default=SP B, current=SP A
==============================================================================
---------------- Host ---------------   - Stor -   -- I/O Path -  -- Stats ---
###  HW Path                I/O Paths    Interf.   Mode    State  Q-IOs Errors
==============================================================================
   0 lpfc                     sdbm      SP A0     active  alive      0      0
   2 lpfc                     sdbq      SP B0     active  alive      0      0
```

For the **ASMLib with EMC PowerPath issue, you have to configure /etc/sysconfig/oracleasm**, there are two variables you can set in order to correct this issue. ORACLEASM_SCANORDER, and ORACLEASM_SCANEXCLUDE. With ORACLEASM_SCANORDER we can specify to process the EMC PowerPath pseudo devices first, with the ORACLEASM_SCANEXCLUDE.we can exclude all the direct paths (/dev/sd*) to the

CLARiiON LUNs.

For example :

```
ORACLEASM_SCANEXCLUDE=sd
ORACLEASM_SCANORDER=emcpower
```

This should correct this, it would filter out the devices /dev/sd*, and would prefer the devices /dev/emcpower*. As you can see, these variables are not set currently on your ASMLib :

```
[sysas2t@myukslulonadb02 sysconfig]$ egrep "
(ORACLEASM_SCANORDER|ORACLEASM_SCANEXCLUDE)" /etc/sysconfig/oracleasm
# ORACLEASM_SCANORDER: Matching patterns to order disk scanning
ORACLEASM_SCANORDER=
# ORACLEASM_SCANEXCLUDE: Matching patterns to exclude disks from scan
ORACLEASM_SCANEXCLUDE=
```

Also I found that **on your production hosts the EMC PowerPath licenses are expired** :

```
[root@myukslplonadb02 cmds]# /etc/opt/emcpower/EMCpower.LINUX-5.0.0/bin/cmds/powermt
check_registration

Key BZP6-MB4K-TFDM-QPBK-M69P-GASB
  Product: PowerPath
  Capabilities: All
  Expired: Mar, 2008

[root@myukslplonadb01 ~]# /etc/opt/emcpower/EMCpower.LINUX-5.0.0/bin/cmds/powermt
check_registration

Key BZP6-MB4K-TFDM-QPBK-M69P-GASB
  Product: PowerPath
  Capabilities: All
  Expired: Mar, 2008
```

This means your **EMC PowePath is running in a demo/basic/degraded mode on your production hosts**.

Summary :

I found three major problems :

1. ASMLib is not configured properly to filter the devices to scan under /dev so only the **EMC PowerPath** pseudo devices are used. As a result 90% of the selected and currently in use devices are direct paths (/dev/sd*), instead of EMC PowerPath paths (/dev/emcpower*), **causing major availability and performance issues**. **This is not just an issue on your QA cluster, but on your production cluster as well.**

Fix : amend the ASMLib configuration /etc/sysconfig/oracleasm (you can use my example), also I would advise a full system boot on both nodes to make sure the ASM instance picks up all the LUNs via the pseudo devices. After you succesfully applied this filter to your Oracle RAC, I would advise you to run 'powermt restore' which would force EMC PowerPath to restore the LUN's current ownership to the default ownerships, resulting some level of load sharing on the dual links. This command should be transparent to the Oracle RAC, as now you would already using only EMC PowerPath pseudo devices, no direct devices. (if you would run powermt restore today, most likely it would bring down both of your Oracle RAC nodes) Btw, if you have more Oracle RAC & ASM installation like this, this seems to me a typical issue so you should check those as well.

Futher info about ASMLib & EMC PowerPath : Metalink doc 394956.1

2. on myukslulonadb02, you had **a missing device /dev/sdbm**, so EMC PowerPath couldn't add to it's configuration

Fix : already applied, now EMC PowerPath can see the path and it became part of it's configuration

3. expired EMC PowerPath licenses on your production hosts (your QAs have valid licenses) Your EMC PowerPath is running in a basic/degraded mode on your production host. This should be fixed even before 1.), otherwise you'll force

ASM to use a multipathing software without a proper license installed.

Fix : get valid EMC PowerPath licenses, and apply them

Matt, fixing these issues most likely will make most of your performance related problems go away, but of course if you keep experiencing performance issues after these fixes, I continue my investigation.

Regards,
sendai