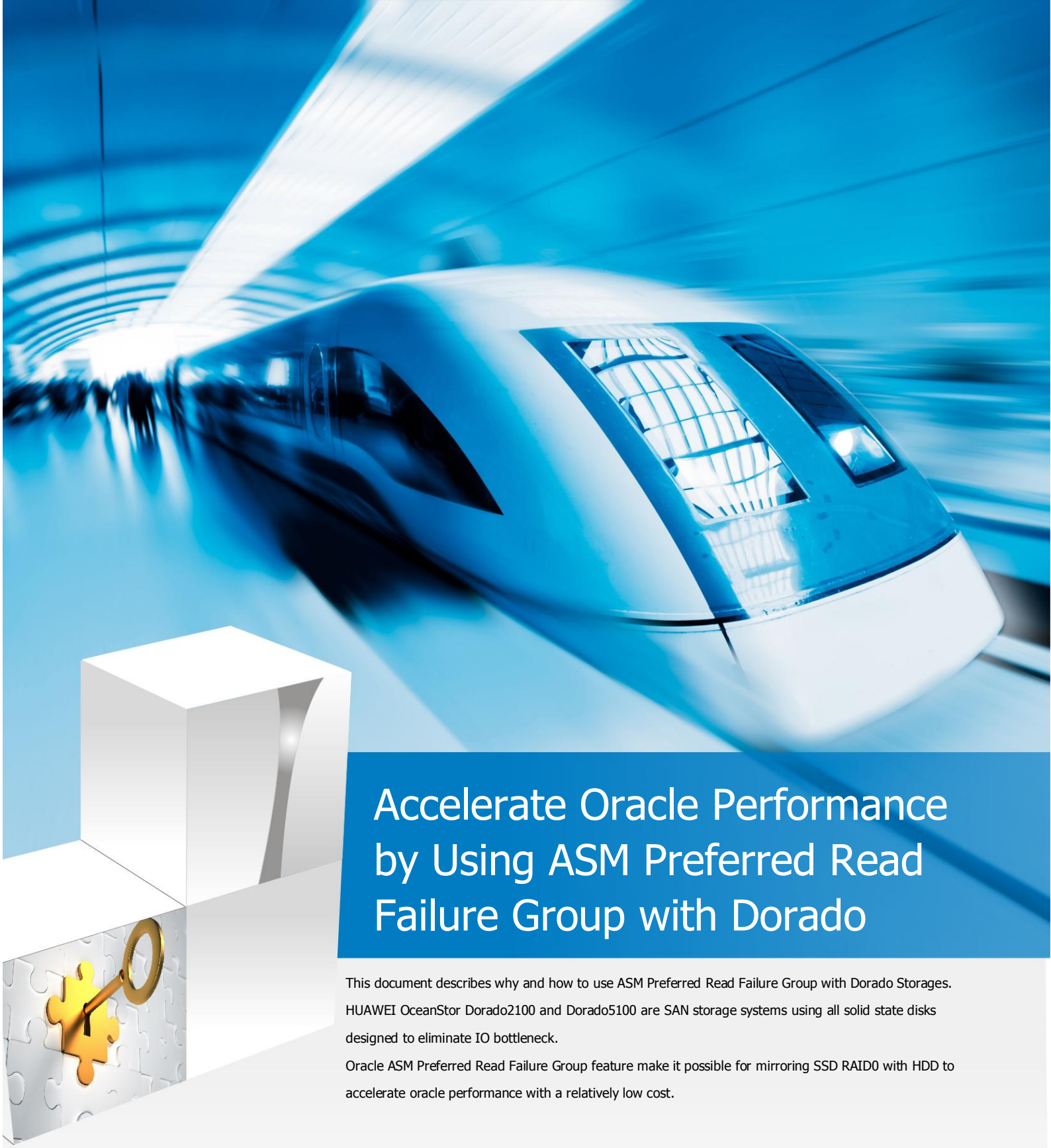


Accelerate Oracle Performance by Using ASM Preferred Read Failure Group with Dorado

This document describes why and how to use ASM Preferred Read Failure Group with Dorado Storages.

HUAWEI OceanStor Dorado2100 and Dorado5100 are SAN storage systems using all solid state disks designed to eliminate IO bottleneck.

Oracle ASM Preferred Read Failure Group feature make it possible for mirroring SSD RAID0 with HDD to accelerate oracle performance with a relatively low cost.

Jarvis WANG

wangyaohui@huawei.com

IT Storage Solution & Verification, Enterprise BG

2012-9-28 Version 1.1

HUAWEI

Why ASM PRFG

As we know, whether in OLTP (online transactional processing) or OLAP (online analytical processing) systems, Oracle is I/O intensive. Using SSDs (solid state disks, also known as flash disk) to store oracle files, user can benefit from the low latency, high random IOPS, high sequential throughput and low power consumption features. SSD is expensive, so you may pay a lot of money for the redundancy data on SSDs to guarantee the reliability of your IT system. Mirroring SSD with HDD and only reading from SSD is a good idea to balance performance and cost. ASM Preferred Read Failure Group (PRFG) provides such a right feature.

OLTP Bottleneck

The following table shows the top 5 wait events in one oracle OLTP database running on traditional magnets disks (10K RPM SAS disk). 96.87% of DB time is spent on waits of "db file sequential read", and each wait costs 15 milliseconds. The high latency of the slow magnets disks is the bottleneck of the OLTP system.

Event	Waits	Time(s)	Avg wait (ms)	% DB time	Wait Class
db file sequential read	426,694	6,492	15	96.87	User I/O
DB CPU		355		5.30	
db file parallel read	729	58	79	0.86	User I/O
log file sync	39,938	30	1	0.45	Commit
gc cr grant 2-way	56,407	18	0	0.27	Cluster

OLAP Bottleneck

The following table shows the top 5 wait events in one oracle OLAP database running on traditional magnets disks (10K RPM SAS disk). More than 80% of DB time is spent on User I/O. The User I/O is definitely the bottleneck of the OLAP system.

Event	Waits	Time(s)	Avg wait (ms)	% DB time	Wait Class
direct path read	4,604,339	567,141	123	63.67	User I/O
direct path read temp	1,955,162	147,298	75	16.54	User I/O
DB CPU		38,874		4.36	
db file sequential read	117,944	16,399	139	1.84	User I/O
direct path write temp	597,138	13,507	23	1.52	User I/O

Capacity and Performance of Hard Disk

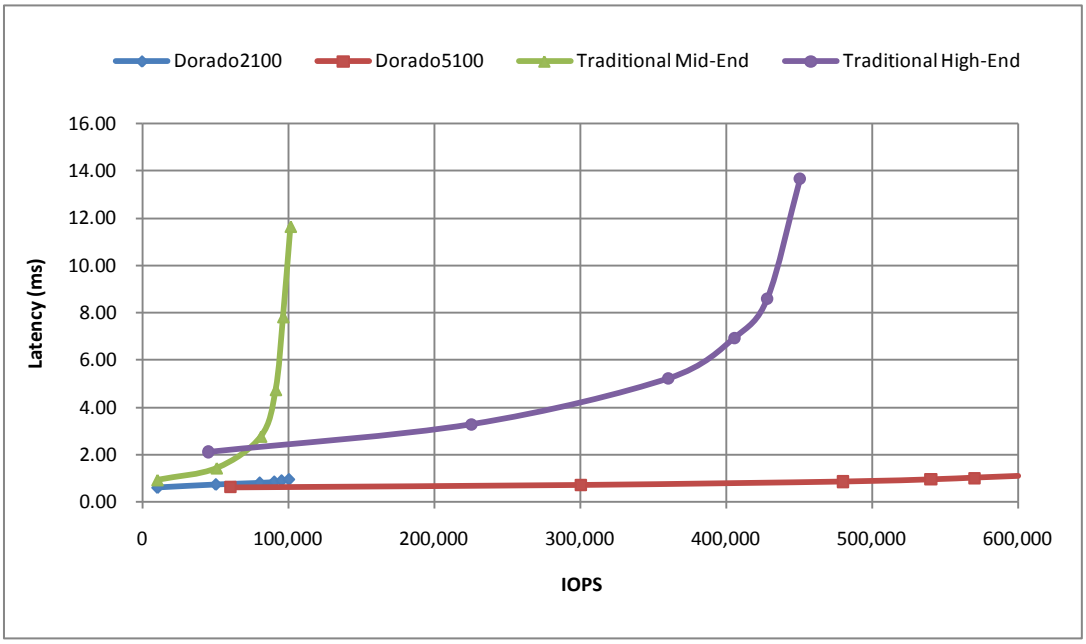
The following table shows the capacity and performance change of enterprise hard disk in the last 5 years, the capacity is increased year by year, but the performance stays the same level. When deploying Oracle OLTP database on hard disk, users will buy much more spindles than the capacity needs to suport the high random I/O requirements, while the free capacity can't be used to store other things because that will make the performance of database bad. Tiering storage or pure SSD solution are now strongly recommended in Oracle OLTP database, and the idea that mirroring SSDs with HDDs is a better solution fit into the requirements.

Year	Capacity	RPM	OLTP Throughput	OLAP Throughput
2008	73 GB	15 K	200 IO/s	30 MB/s
2009	146 GB	15 K	200 IO/s	30 MB/s
2010	300 GB	15 K	200 IO/s	30 MB/s
2011	600 GB	15 K	200 IO/s	30 MB/s
2012	900 GB	15 K	200 IO/s	30MB/s

OceanStor Dorado

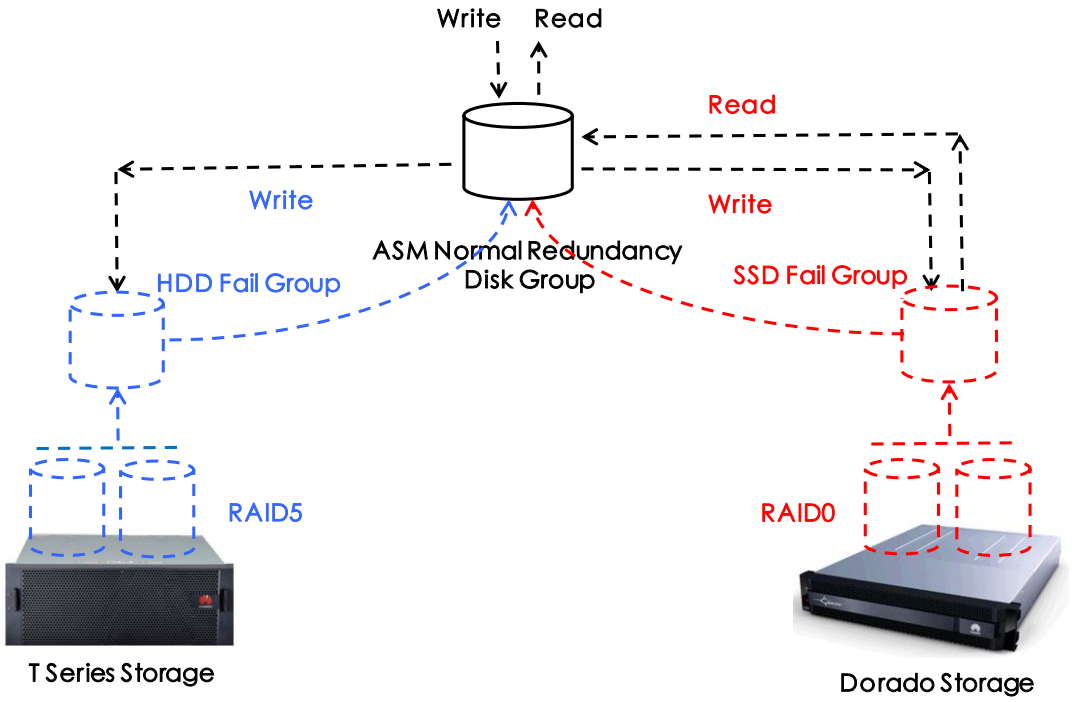
HUAWEI OceanStor Dorado2100 and Dorado5100 are SAN storage systems using all solid state disks, they are designed to eliminate IO bottleneck, and accelerate mission-critical applications by reducing latency. Using Dorado2100 or Dorado5100 to accelerate oracle performance is the best choice.

The following chart shows the SPC1-LIKE (OLTP workload) benchmark results of OceanStor Dorado and traditional storage systems. The latency of Dorado is lower than 1ms, but the latency of traditional storage is higher than 10ms. Dorado could significantly improve the response time and throughput of OLTP systems.



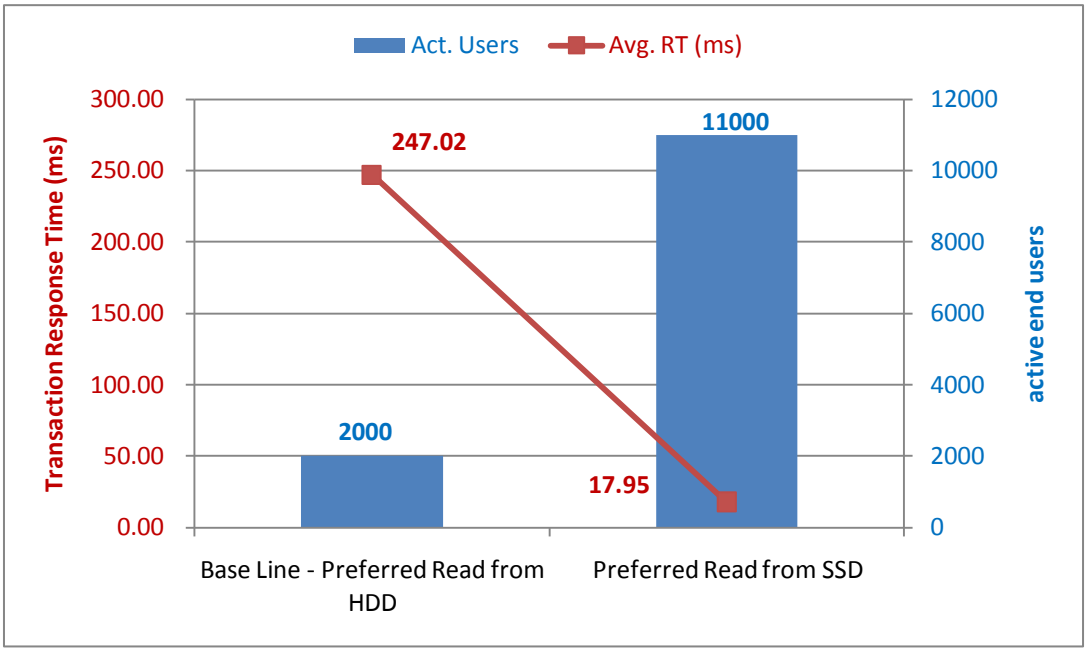
ASM PRFG

ASM normal redundancy disk group is created on two or more fail groups; each block on one fail group has a copy on one of the other fail groups. The following figure describes a normal redundancy disk group created on two fail groups. One fail group is created on two HDD RAID5 arrays from OceanStor T series storage, and the other is created on two SSD RAID0 arrays from OceanStor Dorado storage. The SSD fail group is set as preferred read group. DB blocks are only read from SSD fail group except the group fails, and "dirty pages" are both written to HDD and SSD fail groups. The low latency of Dorado will make the transaction response time much lower.



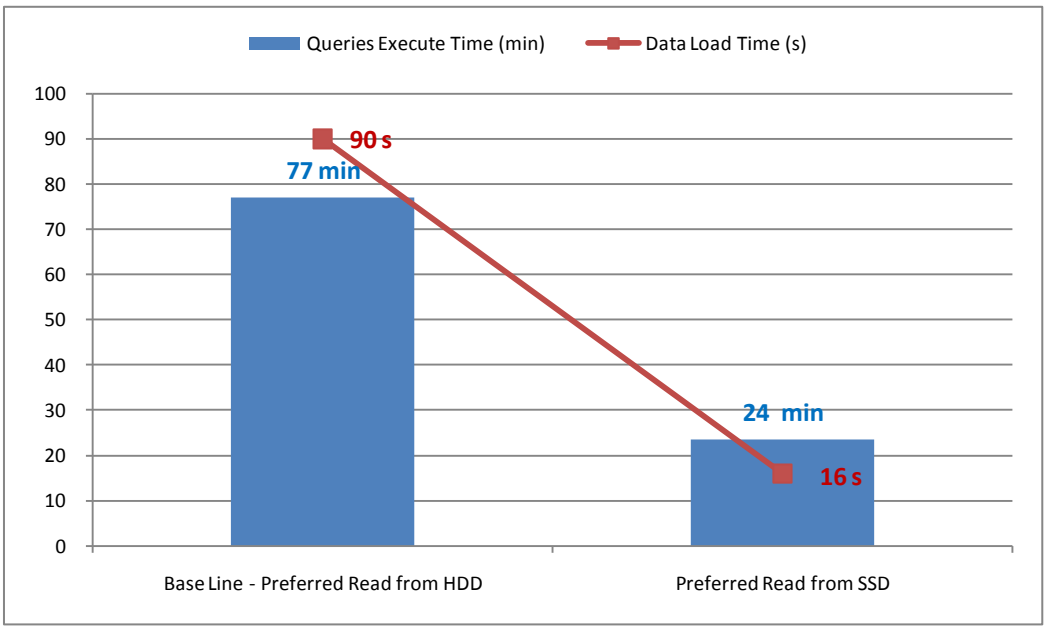
OLTP Acceleration

The following chart shows the improvement of maximum number of active end users (Act. Users) and transaction response time (Avg. RT) when setting SSD as preferred read failure group. The **maximum number of active end users** is increased from 2000 to 11000, **increased to 550%**. The **transaction response time** is reduced from 247.02 ms to 17.95 ms, **reduced to 7.3%**. Users are significantly benefited from the low latency and high random IOPS of Dorado.



OLAP Acceleration

The following chart shows the improvement of analytical queries execute time and text data load time when setting SSD as preferred read failure group. The **queries execute time** is reduced from 77 minutes to 24 minutes, **reduced to 31%**. The **data load time** is reduced from 90 seconds to 16 seconds, **reduced to 17%**. Users are significantly benefited from the low latency and high random throughput of Dorado.



How to Use ASM PRFG

It's very simple to use ASM Preferred Read Failure Group (PRFG). The following steps introduce an example of using ASM PRFG with OceanStor Dorado and T Series Storage.

Create RAID Groups and LUNs

1. Create two RAID5 group on T Series Storage, each with seven 300GB SAS disks.

```
ssh admin@S5500T
admin@S5500T's password: Admin@storage
admin:/>createrg -n HDD1 -l 5 -list 1,0:1,1:1,2:1,3:1,4:1,5:1,6:
admin:/>createrg -n HDD1 -l 5 -list 1,7:1,8:1,9:1,10:1,11:1,12:1,13:
```

2. Create one LUN on each HDD RAID5 groups and map to the oracle host group

```
admin:/> createlun -rg 0 -susize 128 -n HDD1 -c a
admin:/> createlun -rg 1 -susize 128 -n HDD2 -c b
admin:/> addhostmap -group 1 -devlun 0
admin:/> addhostmap -group 1 -devlun 1
```

3. Create two RAID0 groups on Dorado storage, each with nine 200GB SLC SSDs.

```
ssh admin@Dorado5100
admin@Dorado5100's password: Admin@storage
admin:/>createrg -n SSD1 -l 0 -list
1,4:1,5:1,6:1,7:1,8:1,9:1,10:1,11:1,12:
admin:/>createrg -n SSD2 -l 0 -list
1,13:1,14:1,15:1,16:1,17:1,18:1,19:1,20:1,21:
```

4. Create one LUN on each SSD RAID0 group and map to the host

```
admin:/> createlun -rg 0 -susize 128 -n SSD1 -c a
admin:/> createlun -rg 1 -susize 128 -n SSD2 -c b
admin:/> addhostmap -group 1 -devlun 0
admin:/> addhostmap -group 1 -devlun 1
```

5. Scan devices on Oracle database node (Linux) as OS user "root"

```
upadm start hotscan
upadm show array
```

```
[root@HOST4 ~]# upadm show array
Hostname    = HOST4
Domainname  = (none)
Time        = 08/20/2012 10:08:53

------------------------------------------------------------
Array managed by UltraPath.
------------------------------------------------------------
Array ID          WWN                          Module Name
 0          21000022a1050c50                   Dorado5100
 1          21000022a1046452                   S5500T
------------------------------------------------------------
```

```
upadm show lun array=0 | grep LUN
upadm show lun array=1 | grep LUN
```

6. Change the I/O scheduler of HDD LUNs to "deadline" and SSD LUNs to "noop"

```
for sd in `ls -l /dev/disk/by-id/scsi-360022a110004* | awk -F '/' '{print $NF}'`; do echo deadline > /sys/block/$sd/queue/scheduler; done
for sd in `ls -l /dev/disk/by-id/scsi-360022a110005* | awk -F '/' '{print $NF}'`; do echo noop > /sys/block/$sd/queue/scheduler; done
```

7. Change the owner of the device to Oracle OS User and make soft link

```
chown oracle:dba /dev/disk/by-id/scsi-360022a11000*
su – oracle
cd /opt/oracle/data
ln -s /dev/disk/by-id/scsi-360022a1100046452037bab7100000000 hdd1
ln -s /dev/disk/by-id/scsi-360022a11000464521884063200000001 hdd2
ln -s /dev/disk/by-id/scsi-360022a1100050c500188610100000000 ssd1
ln -s /dev/disk/by-id/scsi-360022a1100050c5001886cf500000001 ssd2
```

8. Done.

Create ASM Disk Group

1. Create a normal redundancy disk group in ASM instance.

```
export ORACLE_SID=+ASM
export ORACLE_HOME=/opt/oracle/grid_home
sqlplus / as sysasm
CREATE DISKGROUP DG1 NORMAL REDUNDANCY
 FAILGROUP HDD DISK
  '/opt/oracle/data/hdd1' NAME HDD1,
  '/opt/oracle/data/hdd2' NAME HDD2
 FAILGROUP SSD DISK
  '/opt/oracle/data/ssd1' NAME SSD1,
  '/opt/oracle/data/ssd2' NAME SSD2
ATTRIBUTE 'compatible.asm' = '11.2';
```

2. Done.

Set Preferred Read Failure Group

1. Set ASM instance parameter "**asm_preferred_read_failure_groups**".

```
export ORACLE_SID=+ASM
export ORACLE_HOME=/opt/oracle/grid_home
sqlplus / as sysasm
alter system set asm_preferred_read_failure_groups='DG1.SSD' scope=both;
```

2. Done.

ASM PRFG Best Practices

This chapter introduces best practices when using ASM Preferred Read Failure Group (PRFG) with Dorado Storage System.

RAID Level

Dorado provides 4 kinds of RAID level: RAID10, RAID5, RAID0, and RAID1. When using ASM PRFG, data is redundant between fail groups, so it's unnecessary to redundant the data in RAID level. Level **RAID0** provides maximum random performance and is the **best RAID level** for ASM PRFG. In despite of the high reliability of SSDs, RAID0 has chances to fail. When that happens, all blocks will be read from HDD fail group, the **performance will degrade**, if that can't be tolerated, you could **chose RAID5 or RAID10 to avoid**.

Write Policy

Dorado provides 3 kinds of write policies for LUN: write through, write back with cache mirroring, and write back without cache mirroring. Write through is the default policy, in which mode "dirty pages" evicted out of Oracle Buffer Cache are directly written to the backend SSDs on Dorado, which policy can be used for LUNs store **user tables and indexes**. For LUNs store **redo log files and archive logs**, you could change the policy to "**write back with cache mirroring**" with the help of HUAWEI technical support engineers.

In "write back" mode, "dirty pages" are written to Dorado Cache Pool, and later synced to SSDs in the background with LRU-LIKE algorithm. The latency is very low because blocks are only written to Dorado memory. With "cache mirroring", each write I/O from Oracle is first written to the cache of LUN's owner controller, and at the same time transferred to another controller through the mirror channel between the two controllers, making the latency of write I/O higher than "no cache mirroring".

Linux I/O Scheduler

Oracle database is widely deployed on Linux operating system. There're 4 kinds of I/O scheduler on block devices in Linux kernel 2.6: "noop", "anticipatory", "deadline", and

"cfq". The default I/O scheduler is "cfq", which is not suitable for Dorado. For Oracle database, "deadline" is recommended for OceanStor T series storage systems and "**noop**" is recommended for OceanStor Dorado storage system.

NOTE

Using the following command, you can change the I/O scheduler of "/dev/sdb" to "noop" and "/dev/sdc" to "deadline":

```
echo noop > /sys/block/sdb/queue/scheduler
echo deadline > /sys/block/sdc/queue/scheduler
```

SLC or eMLC

Two types of SSD are supported on Dorado, SLC (Single-Layer Chip) and eMLC (Enterprise Multi-Layer Chip). SLC has much better random write performance and more number of block erase count, but more expensive than eMLC. SLC and eMLC almost have the same random read performance.

The write ratio of OLTP workload is typically 20% - 60%, but there're scenarios lower than 20%. For write intensive OLTP workload, SLC is a better choice considering performance and erase count. For read-mostly or read-only OLTP workload, eMLC is a better choice.

In OLAP database, the data is written once and read many times, and the data is periodically loaded into database. Choosing eMLC is a better idea for OLAP workload.

Copyright © Huawei Technologies Co., Ltd. 2012. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Technologies Co., Ltd.

Trademark Notice

HUAWEI,  and  are trademarks or registered trademarks of Huawei Technologies Co., Ltd.

Other trademarks, product, service and company names mentioned are the property of their respective owners.

General Disclaimer

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

HUAWEI TECHNOLOGIES CO., LTD.

Huawei Industrial Base

Bantian Longgang

Shenzhen 518129, P.R. China

Tel: +86-755-28780808

www.huawei.com

PROVIDED BY HUAWEI STORAGE PERFORMANCE LAB