

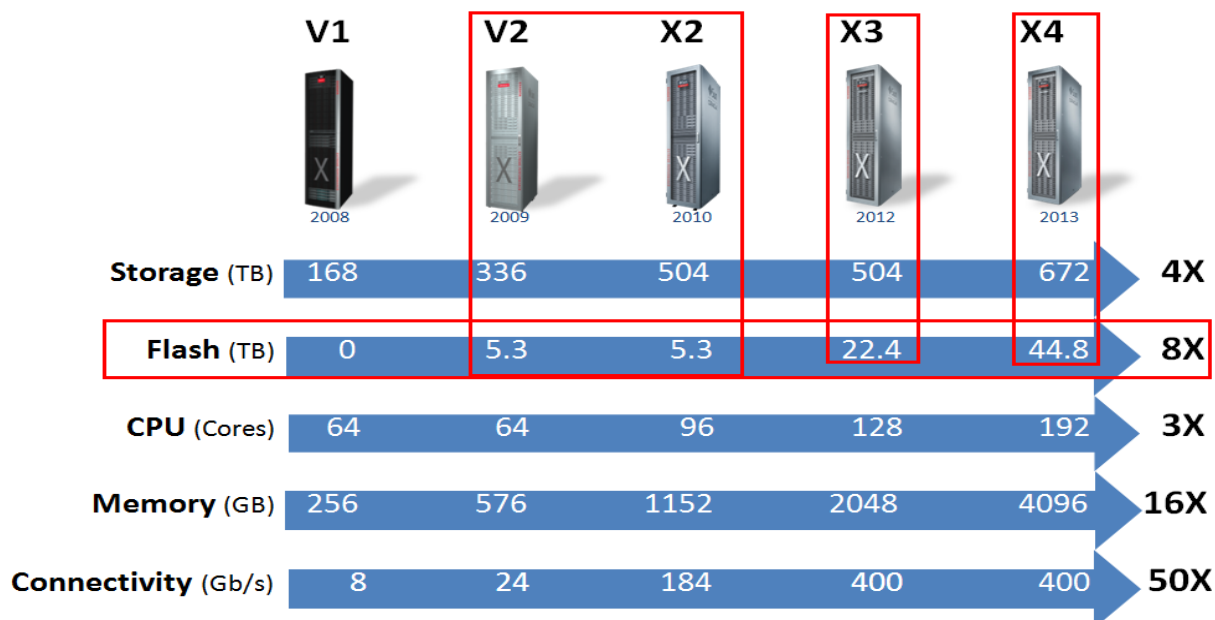
Exadata 上的 Writeback 和 Writethrough

第四篇 X3 和 X4 上的 Flash Card

(www.lunar2013.com)

从 X3 开始采用 Flash Accelerator F40 400GB Solid State Memory，每个 cell 上有 4 块 F40 卡，每个卡上有 4 个 100G 的 FDOM（每个 F40 的容量为 4*100GB=400GB），这样，每个 cell 的总 Flash 容量为 1.6T。可见从容量上已经突飞猛进了。

Hardware Generational Advances



另外，从相应于 X3 的 image 11.2.3.2.0 以后（常见版本是 11.2.3.2.1，最新的 11.2.3.2.x 版本是 11.2.3.2.2），可以支持 WriteBack 功能。该功能缺省是关闭的，可以手工启用这一功能。

对于 X2 或者 V2 的 F20 来说，可以升级 image 并打相应的 patch 来实现这一功能，具体请参见以前的 blog：

Smart Flash Cache on Exadata(1)—介绍
Smart Flash Cache on Exadata(2)—Write through
Smart Flash Cache on Exadata(3)—Write-back
Smart Flash Cache on Exadata(4)—使用 flash cache

After upgrade to 11.2.3.2.0,

- GDs on hard disks remain the same
- GDs on flash disks remain the same
- Flash Log (FL) is automatically created on flash disks and it has a default size of 512 MB
- WTFC is dropped and recreated on the remaining space
- If IORM's objective used to be "off", it will be changed from "off" to "basic" and will be enabled by default. If IORM was not "off" before, then it will remain the same.

也就是说，从 Image 11.2.3.2.0 开始，Flash Card 功能更加强大了：

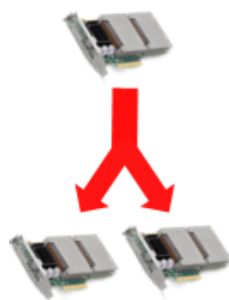
1, Grid disks (GDs) : Flash disk group

2, flash log (FL) : 从 Exadata Storage Software 11.2.2.4 开始引入了 Exadata Smart Flash Logging 的新特性,该特性允许 LGWR 进程将 redo 同时并行写入 flash cache 和 disk controller 中，只要在 flash cache 和 disk controller 中有任意一个率先写完成就会通知 RDBMS 数据库继续工作，该特性着眼于改善了 Exadata 对 Redo 写入的响应时间和吞吐量。

3, flash cache (either WBFC or WTFC) on flash disks.

从 X4 开始采用 800GB PCI Express Flash Accelerator F80 SAS HBA，也就是说支持基于 Flash 的硬件压缩功能，且每个 cell 上有 4 块 F80 卡，每个卡上有 4 个 200G 的 FDOM(每个 F80 的容量为 4*200GB=800GB)，这样，每个 cell 的总 Flash 容量为 3.2T，比 X3 的又增加了一倍的容量。

Exadata Flash Cache Compression



- Exadata uses compression to expand smart flash cache
 - Data automatically compressed as it is written to flash cache
 - Automatically decompressed when it is read out of flash cache
 - Up to 2X more data fits in smart flash cache, so flash hit rates will improve and performance will improve for large data sets
- Flash cache compress/decompress implemented in hardware
 - Performance is same as uncompressed – millions of I/Os per second
 - ZERO performance overhead
 - Supported on X3 or X4 storage servers (requires F40 or F80 cards)

开启 X3 的 FlashCache Compress 功能: alter cell flashCacheCompress=TRUE

开启 X4 的 FlashCache Compress 功能: alter cell flashCacheCompX3Support= TRUE

FlashCache Compress 功能需要在数据库中开启 Advanced Compression Option

这里，我们再回来思考上一篇《[Exadata 上的 Writeback 和 Writethrough-4-V2 和 X2 上的 Flash Card](#)》中提到的一个问题：是否可以将 Exadata 上的 Flash Card 作为 ASM disk 使用？

当时我猜测大家这么用的原因有两点：

- 1，在 Image 11.2.3.2.0 以前的版本，对于 Flash disk 没有 WriteBack 的功能，因此对于需要大量写入的场景，Flash disk 优势不大
- 2，考虑到 ASM 中，同一个磁盘组中的 Failure Group 的大小最好是一致的（官方推荐），那么可以想象，此时的 F20 的容量太小，难以发挥很大作用。

如今，这两点显然已经不成问题了：

- 1，X4 上，每个 cell 上可以建立 4 个最大为 800G 的 ASM Disk（X4 的每个 HDD 是 1T）
- 2，X3 开始已经支持 WriteBack 模式了

理论上，如果对于写要求很大负载的系统来说，我们可以考虑使用 Flash Card 直接做 ASM Disk 了，那么你知道如何设计和使用 Exadata 上的 Flash Card 作为 ASM Disk 来达到性能最大化和最大的可靠性么？

关于 SSD 的使用寿命，我查找了一下资料，计算公式如下：

$$\text{寿命 (年)} = \frac{\text{实际容量 (GB)} \times \text{P/E 次数}}{\text{实际写入 (GB/天)} \times 365 \text{天}}$$

实际容量 = 硬盘容量 × 使用率（一般不低于 95%）

实际写入 = 写入文件大小 × 写入放大率

这里以 128G 的 SSD 为例，若 7x24，Flash Card 每秒写入 100MB，每个存储服务器的 Flash Card 大概能够用多久？

$$128\text{GB} \times 4 \times 1\text{M 次} / 100\text{MB} / (365 \times 86400\text{s}) = 162 \text{ Year}$$

写入次数	SLC	SLC: 10万次
	MLC	MLC: 1万次

上述算法来自于互联网，目前没有找到官方对于 Flash Card 寿命的说法。