Exadata 上的 Writeback 和 Writethrouth

第二篇　SSD 篇

(www.lunar2013.com)

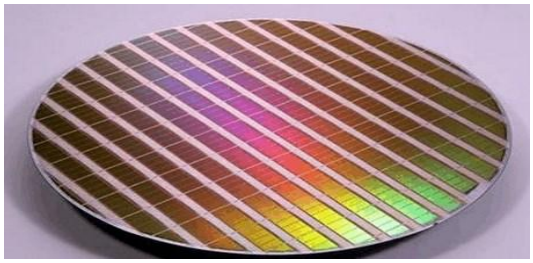
在了解 Exadata 上的 Flash Card 的使用方式之前，我也先给自己科普了一下 SSD 的相关知识，具体如下：

SSD，也就是我们常说的固态硬盘（Solid State Disk），他是用固态电子存储芯片阵列而制成的硬盘，由控制单元和存储单元（FLASH 芯片、DRAM 芯片）组成，好一点的还可以多一个缓存芯片。固态硬盘在接口的规范和定义、功能及使用方法上与普通硬盘的完全相同，在产品外形和尺寸上也完全与普通硬盘一致。

从发展实践来看，1970 年，StorageTek 公司(Sun StorageTek)开发了第一个固态硬盘驱动器。1989 年，世界上第一款固态硬盘出现。

固态硬盘（Solid State Drives），是用固态电子存储芯片阵列而制成的硬盘，其芯片的工作温度范围很宽,商规产品（0~70℃）工规产品（-40~85℃）。这些指标都的使用范围都远超过硬盘的工作规范，因此，我们常说固态盘适应的环境更多……

谈到固态盘（SSD），有一个众所周知的概念“NAND 存储器”，下图是一个 Nand Flash 的晶元：



Nand Flash 的晶元

实际上，由于固态硬盘技术与传统硬盘技术不同，所以产生了不少新兴的存储器厂商。厂商只需购买 NAND 存储器，再配合适当的控制芯片，就可以制造固态硬盘了。新一代的固态硬盘普遍采用 SATA-2 接口、SATA-3 接口、MSATA 接口、PCI-E 接口、NGFF 接口和 CFast 接口。

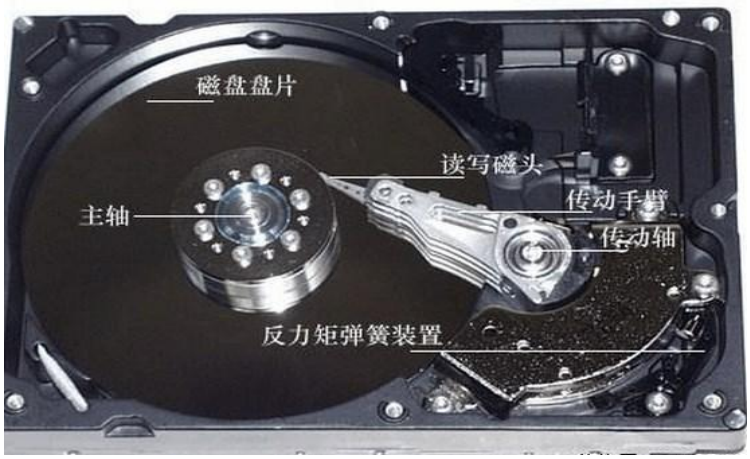
一、再详细了解为什么 SSD 访问速度远胜于 HDD（普通硬盘）之前，我们先了解下 HDD 的工作方式。

1，机械硬盘名字是 Hard Driver Disk，简称 HDD。

它不是液态或固态材质制造的，而是以铝合金材质的磁盘作为存储介质，马达来驱动盘片旋转，并由磁头来读写数据。这就是机械硬盘的基本构成，这与光盘的一些特性比较类似。

当机械硬盘需要读写数据时，将会接到指令，然后磁头会移动到相应位置，盘片也会转动以便让数据发生操作的区域到达指定位置。
这些动作所需要的时间就是寻道时间和潜伏周期，由于需要发生装置的移动，这些过程都需要几毫秒的时间。

也许对于人类的感觉来说，几毫秒的时间并不算很久，为什么我们经常会觉得机械硬盘经常会很慢，甚至用久了会更慢呢？
这是由于操作系统的读写机制造成的：
硬盘被分为若干个区域作为最基本的操作单位，这个单位被叫做“扇区”，当一个新数据写入时，会选择一个或几个扇区进行数据写入，这些扇区的位置都是挨着的，从逻辑上说它们是连续的，无论在读取还是写入的时候所需时间都比较短。



这里有一个关键的问题，即，所有数据都不是在建立之后就永远放在那里不会改变了，当原先写入的数据修改时，比如增加内容、数据量加大，而紧挨着原有扇区的位置已经有了其他数据，这些新数据就要写入到其他位置去。

那么我们在操作系统中看到的一个文件，在实际物理地址上并不是连续的，那么在再次读取该文件时，磁盘要进行的工作量就会加大，在最恶劣的情况，磁头和盘片会进行多次移动和转动，最后的工作时间也是成倍的增加。

这种情况在我们实际使用中并不少见，比如打开一个程序要很久，这是因为程序要加载很多系统文件、组件，这些东西都要从硬盘中逐一读取。比如游戏的加载时间，有大量的数据要读取，并且数据并不一定是连续的，甚至大部分都不可能是连续的。

以及，我们从使用经验上来看，都会觉的电脑会越用越慢，慢道受不了了，重装系统会让速度有所恢复，都是万恶的 HDD 工作原理造成的。

另外，大家在使用笔记本或者玩游戏的朋友在玩游戏时，估计都遇到过磁盘“咔咔”响或者系统很卡的情况，以前我们都说，这表示磁盘快坏了，那这个是为什么呢？

这里面有一个传统 HDD 硬盘的概念“磁头复位”，也就是说，硬盘的数据传输是通过磁头读写磁盘上的数据来完成的。在工作过程中，磁头并不与磁盘的盘面直接接触，两者之间有一层很薄的空气薄膜，这层空气薄膜是由于磁盘的高速旋转产生的。如果磁盘停止旋转，空气薄膜消失，磁头则会直接接触到盘片，这

无疑对盘片的寿命以及对存储在这块区域的数据造成不好的影响。因此在早期阶段，硬盘制造商一般会在对盘片的表面做特殊的处理。

这时 IBM 的工程师们提出了一种叫做 Load/Unload 的技术。简单来说，Load/Unload 技术有点像老式的点唱机，当盘片转速降低无法再产生空气薄膜的时候，就将磁臂以及磁头旋转一下，停靠到磁盘旁边的一个小斜坡上。这样就完全避免了磁头与盘片的直接接触。但是这些还是不能解决磁头复位后当程序请求读取硬盘数据时，磁头需要重新启动并寻址到指定位置,这一过程需要一定时间,而程序就会在这个间隙中出现假死现象。

其实，近 30 年来，磁盘存储技术的发展并不慢，不过仅限于存储密度方面，随着单位面积存储容量的提高，我们可以享用到更高容量的硬盘，但是读写数据的速度上并没有太大突破。

因为决定寻道时间、潜伏周期的关键因素：磁头移动速度和磁盘转动速度都已经接近了极限。

HDD 的最大弊端所在，它的物理移动：磁头移动和盘片转动造成了读写速度慢，越是不连续的文件，读写速度就越慢。这个对不连续的文件进行读写的操作，我们称之为随机读写。

磁盘的寻道瓶颈无法突破之时，SSD 出现了，闪存并不是一个新鲜事物，早在 1984 年，东芝就提出了快速闪存存储闪存的概念，但是由于磁盘已经发展的颇为成熟，而且影响传输速度的主要因素，所以市场上的闪存产品也都是些 U 盘或者加速卡之类，由于不再有寻道的延迟，自然速度更快也更加稳定。

实际上，我们在日常使用中绝大多数硬盘读写操作都是随机类型的，而 SSD 与 HDD 的最大差异就在于随机读写速度，这就是由 SSD 的基本构造决定的。

二、固态盘的存储介质分类

固态硬盘的简称是其英文缩写 SSD：Soild State Disk。

固态硬盘从存储介质分的话，有两种（一种是采用闪存（FLASH 芯片）作为存储介质，另外一种是采用 DRAM 作为存储介质。）

1，基于闪存的固态硬盘（IDE FLASH DISK、Serial ATA Flash Disk）：采用 FLASH 芯片作为存储介质，这也是通常所说的 SSD。

最大的优点就是可以移动，而且数据保护不受电源控制，能适应于各种环境，但是使用年限不高，适合于个人用户使用。例如：笔记本硬盘、微硬盘、存储卡、U 盘等样式。

我猜，基于闪存的固态盘，其用途可能很多时候是用于固定在各种 PCIe 卡上的一组存储数据的闪存芯片，比如典型的有两大阵营，一类是 Intel 的，一类是 Fusion IO 的，这两者之间的主要区别其实不在于这里讨论的采用何种 SSD 存储介质，而是在于他们的接口标准和具体实现工艺，具体的区别会在后面详细说说。

Intel SSD

Fusion-IO SSD

EMC XtremSF半高半长PCIe闪存卡

2，基于 DRAM 的固态硬盘采用 DRAM 作为存储介质

基于 DRAM 的固态硬盘采用 DRAM 作为存储介质，目前应用范围较窄。它仿效传统硬盘的设计、可被绝大部分操作系统的文件系统工具进行卷设置和管理，并提供工业标准的 PCI 和 FC 接口用于连接主机或者服务器。应用方式可分为 SSD 硬盘和 SSD 硬盘阵列两种。它是一种高性能的存储器，而且使用寿命很长，美中不足的是需要独立电源来保护数据安全。

我猜基于 DRAM 的固态盘，其用途可能很多时候是用于直接插在存储柜里面的 SSD 盘，例如这个图中的，当然这里的是插入到 PC 上的外置硬盘盒。

不论哪种 SSD，目前看来，固态存储技术已逐步渗透入服务器、混合存储阵列以及缓存设备的应用中。

不论是 Exadata、各家的一体机还是类似 IBM 推出的全闪存存储系列（FlashSystem），对于事务处理型数据库（OLTP）来说，IOPS 是关键指标之一。因为事务处理型数据库通常由 4~8KB 大小的记录构成，这些数据记录一般是被随机访问的，其性能很大程度上取决于磁盘存取时间（disk access time），即每秒钟多少次 IO 请求，也就是我们说的随机 IO 的 IOPS。

这里我在网上找到别人针对传统 HDD、基于闪存的 SSD 和基于 DRAM 的 SSD 的一个随机 IO 的测试比较，详细请参考： http://storage.it168.com/h/2008-07-16/200807161630445.shtml

测试结果如下：

传统硬盘：

Block Size	Read IO/s	Read MB/s	Avg Service Time - ms	Block Size	Write IO/s	Write MB/s	Avg Service Time – ms
512B	185	0.09	10.4	512B	290	0.14	6.7
1K	185	0.18	10.5	1K	290	0.29	6.5
2K	182	0.37	10.5	2K	283	0.57	6.9
4K	175	0.70	10.8	4K	280	1.12	6.3
8K	176	1.41	10.9	8K	284	2.27	6.2
16K	172	2.75	11.0	16K	264	4.23	6.3
32K	170	5.44	11.0	32K	237	7.58	6.6
64K	152	9.76	11.0	64K	211	13.51	6.5
128K	132	16.96	11.2	128K	183	23.48	8.0

基于闪存的固态硬盘（flash-based SSD）：

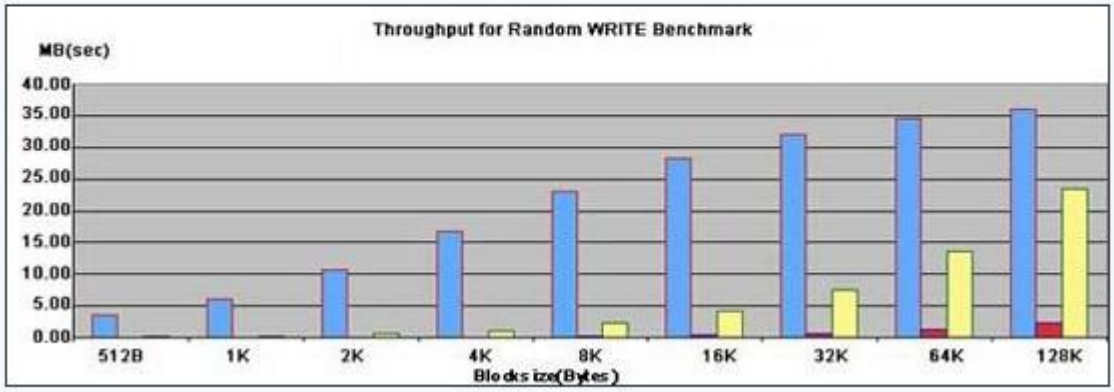
Block Size	Read IO/s	Read MB/s	Avg Service Time - ms	Block Size	Write IO/s	Write MB/s	Avg Service Time – ms
512B	1315	0.66	1.4	512B	22	0.01	92.5
1K	1217	1.22	1.5	1K	22	0.02	91.7
2K	1206	2.41	1.5	2K	21	0.04	92.3
4K	1075	4.30	1.7	4K	21	0.09	94.5
8K	906	7.26	2.0	8K	21	0.17	92.5
16K	666	10.66	2.8	16K	21	0.34	93.7
32K	447	14.33	4.2	32K	21	0.68	102.1
64K	322	20.62	5.9	64K	19	1.23	106.7
128K	204	26.16	9.5	128K	18	2.37	113.2

基于DRAM的固态硬盘：

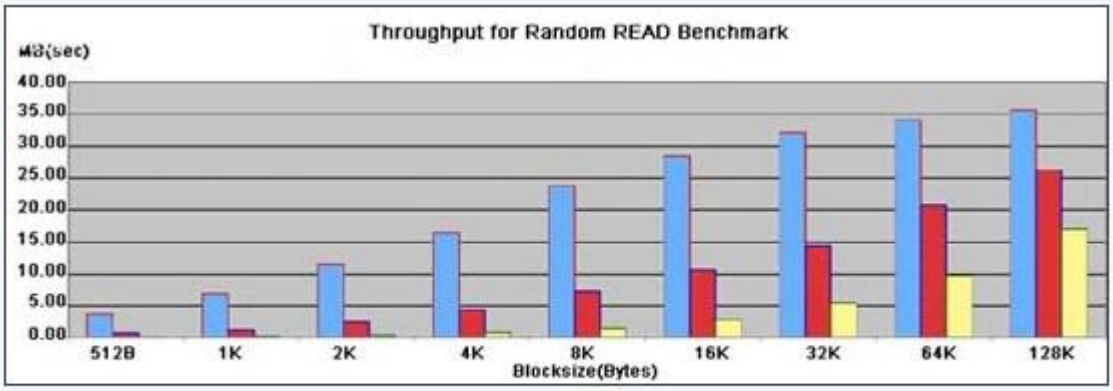
Block Size	Read IO/s	Read MB/s	Avg Service Time - ms	Block Size	Write IO/s	Write MB/s	Avg Service Time – ms
512B	7388	3.69	0.2	512B	7238	3.62	0.2
1K	6794	6.79	0.2	1K	6015	6.01	0.2
2K	5752	11.50	0.2	2K	5353	10.71	0.3
4K	4091	16.36	0.4	4K	4184	16.74	0.4
8K	2959	23.68	0.6	8K	2875	23.00	0.6
16K	1771	28.35	1.0	16K	1773	28.37	1.0
32K	999	31.98	1.8	32K	1004	32.14	1.8
64K	531	34.01	3.5	64K	540	34.57	3.5
128K	277	35.47	6.9	128K	281	36.08	6.8

三种硬盘的评测数据比较

蓝色部分数据：基于DRAM的固态硬盘
红色部分数据：基于闪存的固态硬盘
黄色部分数据：传统硬盘



随机写入性能比较



随机读性能比较

三、基于闪存的 SSD 的结构

基于闪存的固态硬盘是固态硬盘的主要类别，其内部构造十分简单。

固态硬盘内主体其实就是一块 PCB 板，而这块 PCB 板上最基本的配件就是控制芯片、缓存芯片和用于存储数据的闪存芯片。

1、主控的功能：　类似于 SSD 的大脑。控制数据写入，纠错，擦除等，可实现性能优化，数据加密和写保护功能，数据安全擦除模式，自毁功能等。目前市面上比较常见的固态硬盘有 Intel、SandForce、Indilinx、JMicron、Marvell 以及 Samsung 等多种主控芯片。不同的主控之间能力相差非常大，在数据处理能力、算法、对闪存芯片的读取写入控制上会有非常大的不同，直接会导致固态硬盘产品在性能上差距高达数十倍。

2、FLASH 存储芯片

右边图中，红色的分别为主控芯片（上面的）和缓存芯片（下面的），后面连续的 8 个是闪存芯片。

SSD 最基本的单位就是闪存芯片英文名字叫做 Nand Flash（也就是前面提到的 Nand Flash 晶元），这是一种非易失性内存芯片，通过充电、放点的方式写入和擦除数据，速度相当快。
由于在读写操作中完全通过电路来传输信号，因此不会存在类似 HDD 那样移动磁头、旋转盘片等动作，因此大大减少了处理时间。

然而，Nand Flash 也分为几种：

（1）SLC 全称是单层式储存 （Single Level Cell），因为结构简单，在写入数据时电压变化的区间小，所以寿命较长，传统的 SLC NAND 闪存可以经受 10 万次的读写。而且因为一组电压即可驱动，所以其速度表现更好，目前很多高端固态硬盘都是都采用该类型的 Flash 闪存芯片。

（2）MLC 全称是多层式储存（Multi Leveled Cell），它采用较高的电压驱动，通过不同级别的电压在一个块中记录两组位信息，这样就可以将原本 SLC 的记录密度理论提升一倍。作为目前在固态硬盘中应用最为广泛的 MLC NAND 闪存，其最大的特点就是以更高的存储密度换取更低的存储成本，从而可以获得进入更多终端领域的契机。不过，MLC 的缺点也很明显，其写入寿命较短，读写方面的能力也比 SLC 低，官方给出的可擦写次数仅为 1 万次。

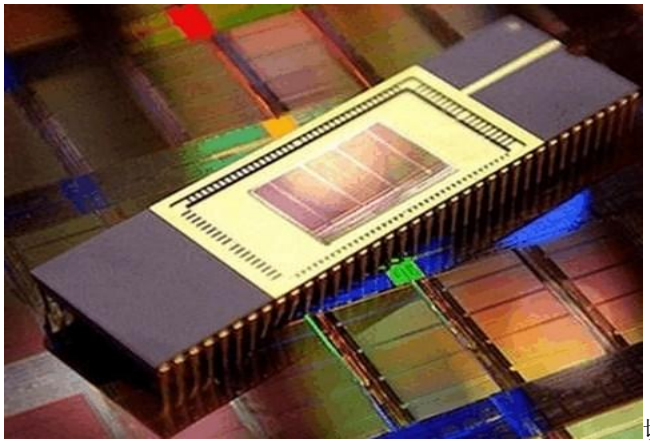
目前消费级 SSD 甚至不少企业级 SSD 都是用 MLC（多层单元）闪存，这种闪存的写入性能不如 SLC（单层单元）闪存，寿命也较之短很多，但是价格要低很多。就算这样，目前 SSD 的成本也没有降低到人人都能接受的程度，价格仍然是影响 SSD 进一步普及的障碍。

3、缓存芯片（这个不是必须的，有些 SSD 上有，有些没有）
主控芯片旁边是缓存芯片，固态硬盘和传统硬盘一样需要高速的缓存芯片辅助主控芯片进行数据处理。这里需要注意的是，有一些廉价固态硬盘方案为了节省成本，省去了这块缓存芯片，这样对于使用时的性能会有一定的影响

四、SSD 的接口和工作方式

SSD 的系统接口、供电部分，以及驱动方式都与 HDD 没有差别，其主要改变是构成单元（也就是前面说的基于哪种存储介质）和物理工作方式（可以理解为采用哪种接口等等）。

SSD 的内部构造包括 PCB 板、主控制器芯片和闪存芯片，有些产品还会有缓存。


切割后的 Nand Flash 芯片

一块 SSD 是由多个 Nand flash 闪存颗粒组成的，我们可以将每一个闪存颗粒看作是一个独立的存储单位，然后由主控制器将他们做了一个 RAID 并联。

也就是说 SSD 的读写是“多线程”的，每次的工作并不会只局限于一个颗粒之上，主控可以让数据分解并同时在不同颗粒上进行写入，这样以来速度自然会更快了。

这也是 SSD 速度快的原因之一。当然，主控要做的事情远非这么简单。

SSD 闪存也是有最小操作单元的，和机械硬盘相比，Nand Flash 的一个比较特殊的区别是写入与擦出操作最小单位不同，写入最小单位为 4KB,这个 4KB 大小的单元称之为“页”（Page），而擦除则为 512KB，叫做“块”（Block）。

也就是说，在空白单元上写入，可以以页为单位来进行，但是若要删除这个数据，就需要将整个块进行擦除操作。

并且当有一个块中的数据需要删除时，会先对需要删除的数据进行标记而非真正物理擦出，然后当再次需要在同一物理位置写入之时，会将有效数据保留，复制到新的块上，然后擦写原来的块。

听起来似乎很复杂，简单的说，SSD 的写入机制就是原本需要写入 1MB 大小的数据，实际操作量是会大于这个数值的，具体是多少，就要看主控制器的算法是否具备高效率，而实际随机写入速度则取决于运算速度是否够快。

这就是为什么我们经常听到客户说，“SSD 运行不稳定，有时会出现 HANG 死的状态”，同时也诠释了为什么 Oracle 数据库运行的环境上，不建议 Redo 放在 SSD 上的原因。

和 HDD 的相同之处是，SSD 也需要逻辑地址来管理，然而操作系统的逻辑地址最小单位是 512B，SSD 的最小写入单位则是 4KB，这其中就需要 CPU、芯片组和主控制器依次工作。

除此之外，主控制器还要负责分配每个闪存芯片的任务量，全盘闪存状态的监控，各个块的管理，数据校验等等，工作相当多而繁杂，这也是为什么在一些新主控上会使用到 ARM 双核心处理器，因为主控的性能会直接影响到 SSD 的速度。

因此，闪存是基本存储单元，而主控制器则是 SSD 的心脏，负责运算和任务分配，两者的结合才是一款 SSD 性能的真正体现。

在网上找到了下面的 SSD 和 HDD 的对比，我作为一个硬件盲，没有辨别的能力，但是作为长时间玩数据库的人来说，我比较敏感的是“数据恢复”那一行，我理解这里的数据恢复不是说数据库在无奈下的使用 DUL 去挖掘数据的过程，而是从存储层面操作的数据恢复。

即便如此，我们分析了上面的一些原理，相比也了解了，基于 SSD 的数据库，Oracle Active Dataguard 是必须的，O(∩_∩)O 哈哈~

固态硬盘和传统硬盘特性的比较

项目	固态硬盘		传统硬盘
容量	256MB~1TB		320GB~4TB
价格	高		低
连续读写速度	SATAII	R/W:250MB/220MB/S	W:40MB~50MB/S
	SATAIII	R/W:550MB/500MB/S	
写入次数	SLC	SLC：10万次	无限制
	MLC	MLC：1万次	
盘内阵列	可		极难
工作噪音	无		有
工作温度	极低		较明显
防震	很好		较差
数据恢复	难		可以
重量	轻		重

江湖传说，SSD 也是可以通过专业设备进行数据找回的，如下图，可见做 SSD 存储层恢复将是未来一个很有前景的行业，O(∩_∩)O 哈哈~