文章编号:1671-4598(2020)06-0038-04 DOI:10.16526/j.cnki.11—4762/tp.2020.06.008 中图分类号:TP333.5 文献标识码:A

企业级固态硬盘最新进展及测试技术研究

吴家隐[1],李先绪[2]
(1.广东邮电职业技术学院 计算机学院,广州 510630;
2.中国电信股份有限公司研究院,广州 510630)

摘要:企业级SSD(固态硬盘)的垃圾回收机制可能引发性能抖动,从而导致业务系统延迟甚至出错;目前,衡量固态硬盘的性能稳定性主要依赖于定性评价,尚未形成合理的定量评价指标;针对这一问题,研究了企业级SSD在生产工艺、闪存转换层及接口和协议等方面的最新进展;在总结现有SSD评价指标的基础上,采用变异系数作为性能稳定性的衡量指标,以FIO、Nmon和Memtester等为测试工具,提炼数据模型和QoS条件(服务质量),构建性能稳定性测试方案;该测试方案包括清空SSD、预热、压力测试以及数据读取等环节;实验结果表明,该方案可以科学合理地对SSD的性能稳定性进行定量评价;通过该方案,可以从规划设计环节开始对IT系统进行稳定性评估,保障核心业务系统的部署运营。

关键词:固态硬盘;性能稳定性;测试

Research on the Latest Development and Test Technology of Enterprise Solid State Drive

Wu Jiayin[1], Li Xianxu[2]
(1. School of Computing, Guangdong Vocational College of Post and Telecom, Guangzhou 510630, China;
2. Research Institute of China Telecom Corporation Limited, Guangzhou 510630, China)

Abstract: The garbage collection mechanism of enterprise solid-state drives may cause performance jitter, which may lead to business system delay or even error. At present, the measurement of SSD performance stability mainly depends on qualitative evaluation, owing to the lack of reasonable quantitative evaluation index. Aiming at this problem, the latest development of enterprise SSD in production technology, flash conversion layer, interface and protocol is studied. On the basis of summarizing the existing evaluation indexes of SSD performance stability, the coefficient of variation is used as the measurement index of performance stability. To build performance stability test scheme, FIO, Nmon and Memtester are used as test tools, when data model and QoS (Quality of Service) conditions are extract. The test scheme includes the steps of emptying SSD, preheating, pressure test and data reading. The experimental results show that the scheme can evaluate the performance stability of SSD scientifically and reasonably. Through this scheme, IT system stability can be evaluated from the planning and design stage, which is helpful for ensuring the deployment and operation of core business system.

Keywords: SSD (solid state drive); performance consistency; test

0 引言

固态硬盘(solid state drive, SSD)的出现,使单硬盘的性能相对机械硬盘提高了几个数量级,从而被广泛用于服务器、磁盘阵列等IT设备中。SSD硬盘主要由NAND闪存颗粒、NAND闪存控制器、接口芯片及接口等组成,这些零部件及相关的软件协议将影响到SSD盘的性能。

机械硬盘读写性能在生命周期内表现稳定,基本上保持一致。然而,固态硬盘在进行垃圾回收时,其写性能与垃圾回收机制产生冲突,用户读写请求的响应时间大大增加,从而出现性能抖动[1]。这种抖动对磁盘阵列的影响不能忽视,尤其是对于安装有多块硬盘的固态盘阵列。Kim等发现,在固态阵列中每块固态盘各自进行垃圾回收,将会放大垃圾回收导致的性能抖动,从而对固态阵列性能产生较大的影响[2]。此外,一些基于主机的固态硬盘在主机系统资源紧张时,也可能引起固态硬盘算法与主机上的业务系统的资源争用,从而产生性能稳定性问题。

而在企业级应用中,固态硬盘常用于读写操作频繁,性能要求高的系统中,如交易型性系统。交易型系统遇到高并发,大流量的访问请求时,固态硬盘的IO读写也随时增大,可能频繁引发垃圾回收机制,从而引发性能抖动,可能引起交易数据延迟甚至出错。因此,研究固态硬盘的性能稳定性对于IT系统具有重要的意义。

1 进展

1.1 闪存颗粒

目前,闪存存储颗粒的基本技术包括NAND和NOR,

收稿日期:2020-03-16; **修回日期**:2020-04-07。
基金项目:广东省普通高校青年创新人才类资助项目(2018GkQNCX140);广东邮电职业技术学院教研教改项目(201920)。
作者简介:吴家隐(1984-),男,广东雷州人,硕士,工程师,主要从事云计算、物联网与光电技术方向的研究。
李先绪(1971-),男,重庆人,硕士,高级工程师,主要从事IT基础设施及云计算方向的研究。

其中，SSD 主要采用 NAND 技术。单位体积闪存颗粒数量越多，相同体积的 SSD 盘容量越大。2D NAND 闪存采用平面型制造工艺制造，其容量的提高需要不断缩小半导体制程的线度，这就严重依赖于光刻技术的进步。在光刻技术进入 20 nm 后，NAND 闪存颗粒的制造成本将大幅度提高，而可靠性和良品率却不断降低。目前，商用光刻工艺已经达到 5 nm，但受限于半导体物理极限，再想提高的难度非常大，2D NAND 的容量已难以增长。

3D NAND 通过垂直方向上堆叠栅极，改变了 2D NAND 的平面扩展模式，突破了光刻精度的限度。在相同面积上可以堆积更多的存储单元，从而大幅度提高存储密度。3D NAND 的主要工艺流程包括交替沉积薄膜形成堆叠层、蚀刻高深宽比通道、填充字线钨金属以及蚀刻阶梯形成独立的接触面[3]。目前，企业级 SSD 主要采用 96 层 3D NAND 工艺。2019 年，美光宣布完成 128 层 3D NAND 芯片首次流片，海力士宣布量产 128 层 NAND SSD 并将在 2020 年普及[4]。我国长江存储也于 2019 年宣布已经量产 64 层 ED NAND 闪存[5]，并将于 2020 年开始研发 128 层堆叠技术。采用 3D NAND 技术进一步提高层叠层数时，还将面临着叠加层数的增加会引入更多的缺陷、高深宽比通道刻蚀缺陷容易导致短路和字符串干扰以及工艺复杂等问题[6]。

2015 年，美光与 Intel 联合推出 3D Xpoint 技术。与通过绝缘浮置栅极来捕获不同数的电子的 3D NAND 不同，3D Xpoint 使用阻变类的新型存储材料与双阈值选通器件耦合形成新存储单元结构。基于 3D Xpoint 的存储器其读写速度略低于内存，但远高于 NAND，随机写入速率是 NAND 的 1 000 倍。同时，3D Xpoint 也能通过堆叠增加容量，密度可以达到内存的 8～10 倍[7]。限于偏高的成本，3D Xpoint SSD 主要用于内存的拓展，或 NAND SSD 硬盘的缓冲。

1.2　闪存转换层

在机械硬盘中，操作系统读写数据的单位根据扇区的尺寸单元（512 字节）设置的。而在固态硬盘中，闪存的读写单位是 4KB 或 8KB 的页，而且闪存以块为单位进行擦除，在未完成擦除之前无法写入。这就导致操作系统的文件系统无法直接管理固态硬盘。为了解决这个问题，固态硬盘通过闪存转换层（flash translation layer，FTL）把对闪存写的读写操作虚拟成为磁盘的独立扇区，从而实现了逻辑地址到物理地址的转换。FTL 主要功能包括内存管理、垃圾回收、磨损平衡，对闪存的性能、使用寿命具有重要影响。除了上述的地址映射功能，FTL 的主要功能还包括垃圾回收和磨损均衡。

在固态硬盘中，数据不能直接覆盖写入闪存空间，而是需要在写前擦除无效数据。在 SSD 第一次使用时，由于盘内都还是已擦除状态，数据可以直接写入。在 SSD 的存储空间已经写满数据时，那就需要通过 FTL 将新的数据写到空闲的闪存空间中，再把逻辑地址指向新的物理地址。垃圾回收机制（garbage collection，GC）的具体过程包括：找到将要擦除的块，将要擦除的块内的有效页中的数据转移到空闲的块中，再擦除要擦除的块[8]。垃圾回收机制的触发时间和条件主要取决于所使用的算法，目前主要有定期触发、阈值触发、空闲时触发等方式。

闪存中存储单元的寿命主要取决于该存储单元的写操作次数，在达到寿命的写操作次数后，该单元失效形成坏块。磨损平衡是为了使写操作均匀分布在固态硬盘内不同的存储单元中，从而避免某些局部的写入过于频繁而形成坏点。

垃圾回收算法和磨损均衡算法的协作，有利于提高 SSD 性能，延长 SSD 的寿命。然而，垃圾回收算法启动时，如果处理不慎，容易引起性能抖动。

1.3　接口与协议

企业级 SSD 的主要接口类型和协议类型如表 1 所示。

表 1　SSD 接口与协议算法

物理接口	逻辑协议
SATA/SAS/PCIe	AHCI
PCIe /U.2	NVMe

SATA/SAS SSD 的逻辑协议就为 HDD 设计的高级主机控制器接口（advanced host controller interface，AHCI），通过 HBA 控制器与 CPU 连接。在固态硬盘的单元存储速度已经大幅度提高的情况下，SATA/SAS 接口已经成为进一步提高速率的瓶颈。

PCIe（peripheral component interconnect express）SSD 直接与 CPU 通信，路径更短，没有协议转换开销，因此具有比 SAS 更低的延时。PCIe SSD 又分基于设备端（Device Based）和基于主机端（Host Based）两种。Device Based 的 PCIe SSD 由主控芯片实现 FTL 算法，而 Host Base 的 PCIe SSD 由主机端安装的数据管理软件实现 FTL 处理，接入完成闪存的读写接口，固态硬盘的主控芯片只需要实现 ECC（error correcting code，错误检查和纠正）纠错、命令响应和闪存通道控制[9]。因此，Host Base 的 PCIe SSD 占用的主机资源更多，在主机端资源占用率较高时候，可能面临着业务系统和数据管理软件资源竞争，从而有可能造成性能的抖动的问题。此外，传统的 PCIe SSD 卡还存在着各厂商需要设计对应的驱动程序，难以形成统一的生态圈的问题。

NVMe 标准（NVM Express，全名非易失性存储主机控制器接口规范，即 Non－Volatile Memory Host Controller Interface Specification）为基于闪存的存储设备设计的，具备低时延和低系统开销的全新规范。通过 NVMe 协议，符合标准的盘都可以采用相同的驱动程序。NVMe SSD 是 Device Based PCIe SSD 的延伸，性能理论上可以获得和 PCIe 一样的性能。NVMe 的物理接口类型主要是 PCIe 和 U.2。目前，NVMe 已经得到了业界的认可，不仅得到了众多硬件厂商的认可，还获得了 Redhat、Oracle、微软等软件厂商的支持，已经全面进入商用。

2 测试方案

2.1 测试指标

主要指标包括：

1）IOPS（input/output operations per second，每秒输入输出操作数）。IOPS 是硬盘性能的重要指标，常用于衡量小数据块的随机写性能。在业务系统中，如数据库等应用，体现在存储端的压力通常是随机读和随机写。

2）延时/QoS（quality of service，服务质量）。延时是接收到服务请求到返回一个指令所消耗的时间，访问延迟增大时，业务系统也会有相应的延迟。QoS 可以分析一段时间内的延时表现，可以用一定的数据读写下延迟不大于指定时间的方式来表示[10]。

3）性能稳定性。性能稳定性，又称为性能一致性（performance consistancy），是固态硬盘质量的重要性能指标。在固态硬盘中，垃圾回收机制启动时，垃圾回收操作会与外部读写请求发生访问冲突。因此，需要在垃圾回收完成后才能完成外部读写请求的响应，这就引起了硬盘性能的抖动[1]，体现在 IOPS 值的波动以及延时的增加上。由于固态硬盘经常用于承载核心数据库或其他 IO 访问量大的应用。例如，性能抖动可能导致交易型存储系统中的交易数据延迟，甚至发生错误[11]。

目前，衡量性能稳定性主要靠人工查看 IOPS 的性能分布曲线，从定性角度来评价 IOPS 值的离散度。而在定量评价指标上，各值的定义也不大一致。Intel 定义性能稳定性的衡量标准为剔除最低的 0.01%的性能最小值以后，取余下数据中性能最小点与平均性能的百分比作为稳定性指标。如，最低 IOPS 的点为 10 000，而平均 IOPS 为 11 000，则性能稳定性指标为 10 000/11 000＝90.9%。然而这一定义仅能剔除个别性能值较低的数据，且只考虑性能较低值的波动，而并未考虑性能高值的波动。还有一种标准是统计一定 IOPS 以下的点的数量占总数量的占比，这种方式只考虑数量，并未考虑到每个点和点之间的波动幅度的影响。

鉴于现有性能稳定性指标的缺陷，本文提出能够表征概率分布离散程度的归一化量度的变异系数（CV，coefficientof variation）作为性能稳定性的衡量指标。在比较 SSD 性能的离散程度时，SSD 的性能相差较大，直接使用标准差来进行比较 SSD 的话，可能出现一些性能较差的 SSD 标准差偏小的问题。使用变异系数可以消除数值大小的影响，客观反映性能的抖动程度。SSD 的变异系数计算公式为：

变异系数＝标准偏差/平均值

变异系数越大，表示性能数据的离散程度越大，也就反映出其性能的稳定性越差。

Host Based 的 PCIe SSD 的 FTL 算法需要加载数据到主机内存资源中进行运算，因而其性能与 CPU 和内存状况有关。在内存资源紧张的情况下，有可能出现主机业务系统与 FTL 的资源竞争，从而产生性能波动。

2.2 测试工具

FIO 可以根据用户参数设定，产生特定类型 IO 操作的工具，可以用来模拟 I/O 负载匹配的作业文件，从而测试磁盘在不同 I/O 负载下的性能。

Nmon 可以采集 CPU 利用率、内存使用量、磁盘 IO 等信息。

Memtester 用于在主机产生指定强度的压力，消耗内存资源。

2.3 硬件环境

测试固态硬盘为 MLC PCIe SSD 卡，包括有 A 和 B 两个样本，分别插到测试主机的 PCIe SSD 插槽中。测试主机 CPU 为 Intel XEON E7－8870v3 系列 18 核 2.1 GHz 主频；内存为 128 GB，操作系统为 RedHat 6.5（X64）。

2.4 测试条件

1）数据模型：在承载核心数据库的应用时，数据块大小主要为 4 K。此外，通常情况下，大部分数据库的读写主要是 70%的随机读和 30%的随机写。为了贴近生产环境，设置数据模型为数据块大小为 4 K，读写比例为 70%的随机读和 30%的随机写。

2）QoS：QoS 条件的设定要考虑 PCIe SSD 本身的读写延时情况和应用对 PCIe SSD 的延时需求。

如果在 PCIe SSD 上部署数据库等应用，在实际使用中，大并发混合读写压力（比如 70%的随机读和 30%的随机写）下低写延时对数据库整体性能至关重要。

综上所述，本文将 QoS 设定为 70%随机读 30%随机写的 4 K，读延时不超过 1 000 μs，写延时不超过 200 μs。

2.5 性能稳定性测试方案

性能可靠性测试过程主要包括清空 SSD、预热、启动性能可靠性测试。具体测试过程如下：

1）清空 SSD：将 SSD 完整擦除一遍，即可清空 SSD 盘，从而使全新盘的状态，并静置 10 min，保证所有测试的可重复性，也即是从硬盘尚未写入数据开始，降低因为 SSD 上已有数据随机度不一致引起的测试误差。

2）预热：使用 FIO 对 SSD 全盘顺序写满两遍，然后 4K 随机写 3 小时进行预热；

开启读写测试，全盘顺序填充数据两次。执行两遍全盘随机写，将数据全部打乱，保证后面的所有随机读写测试都是全盘随机。

预处理过程 1：顺序写两遍

```
fio －－filename＝/dev/nvmexx －－ioengine＝libaio －－direct＝1 －－thread －－norandommap －－name＝init_seq －－output＝init_seq_1.6T.log －－rw＝write －－bs＝128k －－numjobs＝1 －－iodepth＝64 －－loops＝2
```

预处理过程 2：4K 随机写 3 小时

```
fio －－filename＝/dev/nvmexx －－ioengine＝libaio －－direct＝1 －－thread －－name＝init_rand －－output＝init_rand_1.6T.log －－rw＝randwrite －－bs＝4k －－numjobs＝8 －－iodepth＝32 －－ramp_time＝60 －－runtime＝10800 －－time_based
```

3）性能稳定性测试：使用 FIO，以 70：30 的随机读/随机写比例的 4K 数据块，对 SSD 盘进行压力测试，持续 12 小时。代码如下：

用 nmon 记录 12 个小时数据

nmon －s 20 －c 2161 －f &.

4 K 混合随机读写 12 个小时，前面 fio ramp_time 300 秒的数据点不统计在内

fio －－filename＝/dev/nvmexx －－ramp_time＝300 －－runtime＝43200s －－time_based －－ioengine＝libaio －－direct＝1 －－thread －－norandommap －－name＝randrw_4K －－rw＝randrw －－rwmixread＝70 －－bs＝4k －－numjobs＝8 －－iodepth＝30 －－log_avg_msec＝2000 －－write_iops_log＝iops_4K_Steady_1.6 －－output＝randrw_4K_Sum_Steady_1.6.log －－group_reporting

4）数据读取：通过 nmon analysis 解析生成的 nmon 文件，在解析得到的 excel 中 DISKXFER 页找到对应的测试对象，不统计前面 FIO _ ramp _ time 大于或等于 300 s 的数据点，取 23 个小时数据计算 IOPS 平均值和变异系数升（标准偏差/平均值）。

5）从 FIO 输出文件中获取读写 clat 平均延时，要求在限定延时范围内。

2.6　内存加压下的性能稳定性测试方案

内存加压条件下的性能可靠性的测试过程与性能稳定性测试的区别在于，在性能稳定性测试步骤中，使用 Memtester 对内存持续加压。加压时间以 4 小时为一个周期，每周期内 Memtester 分别占用主机 0、50%、70%到 90%的内存的情况下，对 PCIe SSD 的性能稳定性进行测试。通过内存加压条件下的性能稳定性测试，可以测试出影响 SSD 性能的内存使用率，以及 SSD 的性能稳定性随内存的变化量。Memtester 代码示例如下：

nohup memtester 64G ＞ memtest_50_1.6.log &.

该代码表示，memtester 消耗的内存大小为 64 G，即占用 50%的内存。

3　实验结果与分析

3.1　性能稳定性

使用上述性能稳定性测试方案，对 1.6 T 容量的 PCIe SSD 进行测试，并整理测试结果如下：

从时延来看，A 产品的读时延为 514 μs，写时延 91 s，B 产品的读时延为 899 μs，写时延为 20 μs，均满足 QoS 要求。

如图 1 所示，A 产品进入稳态的时间较晚，在 2 小时 03 分时进入稳态，在 2 小时 03 分到 3 小时 43 分间产品的曲线比较平滑，整体离散度较小。在 3 小时 43 分后，A 产品的 IOPS 曲线产生了持续抖动，整条曲线有毛刺感，从定性分析即可得出，该产品的性能稳定性较差。B 产品在测试开始 1 小时 31 分钟后进入稳态，随后 IOPS 的性能曲线比较平滑直到测试结束。

根据变异系数计算公式，选取进入稳态后到结束测试这段时间内的 IOPS 性能值，计算 A 产品和 B 产品的离散度得到，A 的变异系数为 8.46%，而 B 的变异系数为 0.75%，A 的变异系数比 B 高一个数量级。至此，本方案

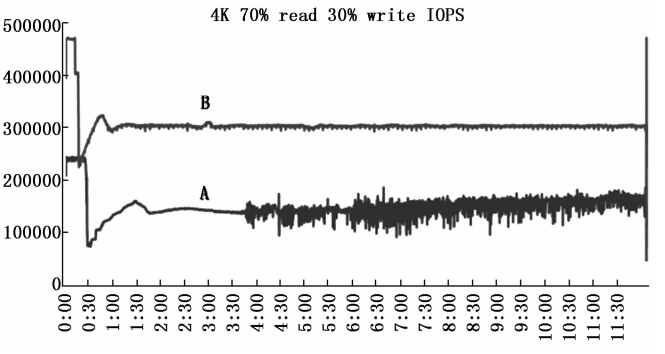

图 1　性能可靠性测试

可以定量地评价固态硬盘的离散程度，其数值可以用于 IT 系统的存储规划、

从上述测试结果分析可以看出，本文所提出的测试方案，可以定量地评测出 SSD 的性能稳定性。用户可以根据业务系统需求，在使用成本、SSD 性能稳定性和业务系统需求的之间综合考虑，选取适合的 SSD。

3.2　内存加压下的性能稳定性

图 2 是 B 产品在不同内存压力下的性能测试结果。在启动测试后的第一个 4 小时周期内，压力测试工具设置加压为 0，在进入稳态后，B 产品的变异系数值（*CV*1）为 0.72%。第二个 4 小时周期内，压力测试工具在主机上产生 50%的内存压力，此时 B 产品的变异系数值（*CV*2）增加到 0.93%。在第三、第四个小时周期内，压力测试工具分别产生 70%、90%的内存压力，B 产品对应的变异系数值（*CV*3、*CV*4）分别为 1.03%、1.02%。由此可见，B 产品在内存资源消耗加剧时，SSD 的变异系数略有增加，但变化不大，具有良好的性能稳定性。

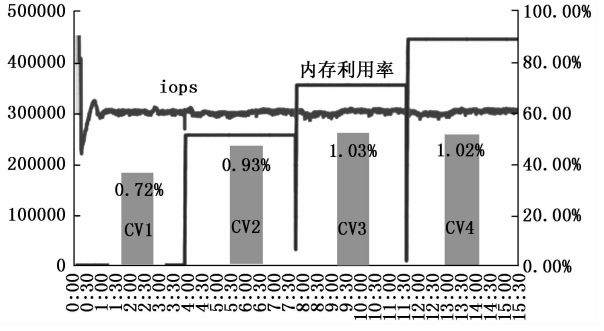

图 2　内存压力测试的性能可靠性

4　结束语

本文介绍了企业级固态硬盘在生产工艺、闪存转换层及接口和协议方面的最新进展。针对目前业界定量衡量性能稳定性指标缺失的现状，提出以变异系数作为衡量指标，并指内存资源竞争可能引起性能波动。本文以 FIO、Nmon 和 Memtester 等为测试工具，设置数据模型和 QoS，并设计了性能稳定性测试方案。实验证明，本方案能够定量地评

（下转第 46 页）