

磁盘存储测试技术研究

贺秦禄¹ 李战怀¹ 王乐晓¹ 段升强² 王惠峰¹ 孙 鉴¹

(西北工业大学计算机学院 西安 710129)¹ (西安建筑科技大学信息网络中心 西安 710055)²

摘 要 研究了测试海量存储系统中磁盘阵列 IOPS、数据传输率两项性能指标的技术与方法,提出了针对高性能磁盘阵列的并行测试技术,并使用此方法实现了对最新研制的高性能磁盘阵列的评测。通过研究影响单个磁盘性能指标的各项因素,并借助实验对其进行量化分析,最终实现了对整个磁盘阵列的性能评测。研究了影响磁盘阵列整体性能的关键瓶颈点,并给出了各项测试结果的合理性证明。

关键词 海量存储,磁盘阵列,性能测试

中图法分类号 TP393 **文献标识码** A

Disk Storage Testing Technology

HE Qin-lu¹ LI Zhan-huai¹ WANG Le-xiao¹ DUAN Sheng-qiang² WANG Hui-feng¹ SUN Jian¹

(Department of Computer Science, Northwestern Polytechnic University, Xi'an 710129, China)¹

(Department of Information Network Center, Xi'an University of Architecture and Technology, Xi'an 710055, China)²

Abstract We studied technologies about testing IOPS and data transfer speed of disk arrays in mass storage systems, and proposed a parallel testing technology towards the high performance disk array and realized the testing work on the newest developed disk array. By taking researches on many factors which effect disk performance seriously and quantifying them through experiments, we realized evaluations on the whole disk array. Finally, we studied bottlenecks of the disk arrays and proved that the results are correct.

Keywords Mass storage system, Disk array, Performance analysis

1 相关技术说明

磁盘是一种块设备,即数据分成一块块再存放在介质上,可以直接选择读写一块数据,定位这个块的速度比较快。随着技术的发展带来的成本下降,更有取代磁带机成为普及型数据存储的趋势。现在已有使用磁盘模拟磁带库的产品,这种产品在逻辑上与磁带库相同,但是由于磁盘的性能比流式数据存储的磁带有较大的优势,因此这种产品的性能得到了较大的提高,从而加快了备份存储系统前端数据的速度,降低了备份过程对上层应用的影响。

磁盘块大小一般为 512 字节。随着需求的变化,现在硬盘生产商“西捷”已经开始讨论研制块大小为 2k 字节的磁盘。测试过程中要明确块的大小。磁盘读取数据的过程也非常重要。磁盘读取数据时,总是整“块”进行读取,这就牵涉到一个读取位置对齐值的问题。

图 1 显示了两个连续的扇区。读取一个大小为 512 字节的磁盘数据时,将会有两种情况:一种是读取位置的开始正好是某个扇区的边界,在这种情况下,从读取位置开始读取该扇

区即可得到想要的结果;另一种情况是读取位置如图中所示的“开始”处,它位于一个扇区的中间,很明显,仅读取该扇区的数据不能满足数据读取的需要,这种情况下必须将下一扇区的内容读取出来。如图 1 所示,由于磁盘的读取是以扇区为单位进行的,扇区 1 与扇区 2 中的数据在第二种情况下都会被读取出来,然后取请求的数据返回给上层应用。这样,在测试过程中,如果不设定存取位置的对齐值,测试结果就可能会出现误差,或者测出来的结果不是最大性能指标,被测开发商会对测试结果提出异议。

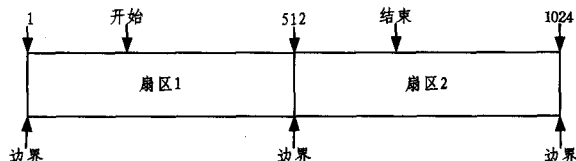


图 1 磁盘 IO 读取位置说明图

对磁盘来说,数据的读取速度很快,所以读取数据的时间并不是一次 IO 主要的时间耗费点。影响磁盘读取速度的关

到稿日期:2011-07-23 返修日期:2011-10-20 本文受国家“863”计划基金资助重大项目(2009AA01A404),国家自然科学基金(60970070)资助。

贺秦禄(1982—),男,博士生,主要研究方向为海量存储、数据管理,E-mail:luluhe8848@hotmail.com;李战怀(1961—),男,博士,博士生导师,主要研究方向为数据库、数据管理和海量存储;王乐晓(1987—),男,硕士,主要研究方向为网络存储、软件工程和网络软件;段升强(1981—),男,硕士,主要研究方向为网络数据存储、多媒体网络技术;王惠峰(1986—),男,博士生,主要研究方向为数据管理、海量存储;孙 鉴(1982—),男,博士生,主要研究方向为数据管理、海量存储。

键点是磁头的寻道过程。由于磁头寻道是一个机械过程,提高磁头寻道的速率需要设置合理的磁头调度策略,目前普遍使用的是电梯算法。在测试过程中,要测得被测设备的最大传输率,必须使磁头尽可能少地寻道,在确保顺序读写的同时,还要尽可能加快磁盘响应 IO 的速度。

2 单个磁盘测试

本节通过测试单个硬盘的多种指标研究测试硬盘的测试技术。整个部分包括测试环境、测试方法、测试结果与结论分析。

2.1 测试环境及配置

本节所有测试都是在图 2 的所示拓扑结构中进行的。

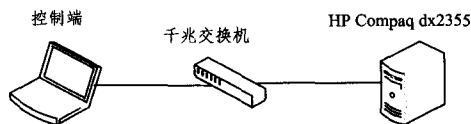


图 2 单硬盘测试拓扑结构图

软件信息:整个测试使用测试工具 iometer-2006.07.27;其总控端运行在安装有 Windows XP 系统的 PC 机上。测试端(dynamo)运行在安装有 RedHat Enterprise Linux 5.3 系统、内核为 2.6.18-128.el5xen 的 PC 机上。

硬件配置信息:型号为 D-Link DGS-1024T 的千兆以太网交换机;型号为 HP Compaq dx2355 的 PC 机,其配置信息为 CPU 型号 AMD Athlon(tm) Dual Core Processor 5000B、频率 2.6GHz、Cache 512kB、外频 200MHz;DDR2 2.5GB 内存,频率 800MHz;Western Digital 硬盘、SATA 接口、转速 7200 RPM;8MB 缓存;250GB 容量。

2.2 数据传输率及 IOPS 测试

2.2.1 测试方法

在服务节点上自带的大小为 250GB 的硬盘上分出一个大小为 30GB 的分区,设定 iometer 参数 outstanding ios 为 1、线程数为 1(保证所有对硬盘发起的 IO 操作为全顺序)、对齐值为默认的扇区大小(512)。然后分别进行全顺序、全随机读写,依次改变传输数据块大小,测得被测硬盘数据传输率与 IOPS 随传输数据块大小变化时的性能曲线。

2.2.2 测试结果

图 3 显示了使用不同存取策略得出的数据传输率性能指标。在整个测试过程中,分别进行了随机、顺序读写测试。

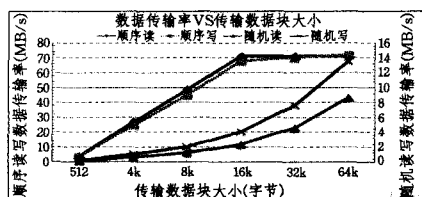


图 3 单硬盘数据传输率测试结果

从图 3 可以看出,在顺序读写时,硬盘的数据传输率性能指标会随着传输数据块的增大而呈线性增加趋势,并且当传输数据块达到 16kB 时会达到一个极限,基本上已达到了硬盘数据传输率的最大值。之后,若再增加传输数据块大小,硬盘的数据传输率并不会会有显著变化。对随机读写,数据传输

率则随着传输数据块的增加有显著的增加趋势。但总体来说,随机读写的数据传输率指标比顺序读写的数据传输性能要低很多。由此说明,对于上层应用,如果其要求较高的数据传输率性能,那么进行数据读写时应尽可能将数据进行顺序读写,以更好满足其性能需要。

图 4 显示了使用不同存取策略得出的硬盘 IOPS 性能指标。在整个测试过程中,分别进行了全随机、全顺序读写测试。

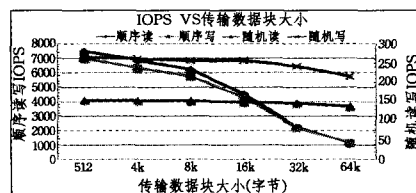


图 4 单硬盘 IOPS 测试结果

从图 4 可以看出,顺序读写的 IOPS 指标可达到随机读写 IOPS 的近 300 倍,可见读写方式对最终 IOPS 性能的影响之大。另一方面,全随机读写的过程可以看作是测试硬盘磁头寻道能力的有效方式。从结果可以看出,普通硬盘的寻道能力也只有 250 次/s 左右。可见,影响磁盘性能的最主要的因素还在于做机械运动的磁头,提高磁头的寻道能力对于提高整个硬盘的性能有着最为重要的意义。另外,从图 3、图 4 中可以看出,对于顺序 IO,读写操作的 IOPS 指标相差不大;而对随机 IO 来说,写操作的 IOPS 指标是读操作的近两倍。可能是因为硬盘的驱动程序对随机写操作有着更为有效的优化:对于读操作,驱动程序完全不知道下一次 IO 要读取的数据,每次 IO 都要重新寻道以读取数据,但其预读的策略在随机操作情况下起不到明显的作用(8MB 的硬盘缓存相对于 30GB 的硬盘分区太小);对于写操作,硬盘驱动可将最近几次的写数据操作缓存于硬盘缓存中,之后可以对最近几次随机写操作用电梯算法策略优化磁头寻道过程,以提高性能。

对于上层应用的开发者来说,只有对硬盘的性能损失点有足够的了解,并且能够掌握硬盘性能优化技巧,才能有效地提高应用程序的性能,以满足用户的需求,从而达到事半功倍的效果。而文件系统作为建立在硬盘之上的“上层应用”程序,只有合理优化硬盘的使用过程,才能达到相关指标的最大化。

2.3 低度随机测试

本小节进行随机 IO 比例不超过 10% 的低度随机读写硬盘的测试,以得到随机操作对硬盘数据传输率及 IOPS 两项性能指标的影响。

图 5、图 6 是低比例随机 IO 条件下测得的随机操作比例对不同大小的传输块的数据传输率与 IOPS 的影响。从图中可以看出,虽然随机操作的比例非常小,但是仅仅 6% 的随机操作比例就导致硬盘数据传输率与 IOPS 两项性能指标比全顺序操作时下降了一倍。这又一次验证了磁头寻道能力对整个硬盘性能的影响。而建立在硬盘之上的文件系统,必须最大程度地确保单个文件数据尽可能连续存储在硬盘上;并且在必要时要尽可能地减少产生硬盘随机操作的可能,最大化地优化硬盘性能,这对文件系统的性能有着至关重要的意义。

3.2 测试策略说明

测试目标:对磁盘阵列进行测试,测得其最大数据传输率、最大 IOPS。

测试方案:针对被测磁盘阵列只有 4 个盘的情况,为了测得最大数据传输率,采用 RAID0 形式,其它参数分别设置为条带宽大小为 64kB、Cache 预取策略为智能预取、Cache 写策略采用回写不镜像。

测试方法:每测试项测试时间设置为 5min,重复 3 次,之后求出平均值。

测试数据传输率性能最大值:首先使用 iometer 从 S1 发起对 LUN01-A-100G-S1 的测试。先设定 outstanding ios 参数为 1,worker(线程)数为 1 测得不同读写块大小(可分别设定传输数据块大小为 16kB、32kB、64kB、128kB、256kB、1MB、2MB、4MB,并为全顺序读、写)对数据传输率的影响,得到数据传输率最大值情况下的读写块大小。之后,逐步加大 outstanding ios 参数,测得使数据传输率性能最优的 outstanding ios 参数值。然后,在设定最优 outstanding ios 参数的同时,依次增加 worker 数,得到使数据传输率最优的 worker 数。到此,已经得到了使数据传输率性能最优的 outstanding ios 参数、worker 数、传输数据块大小。设定以上各值,同时发起 S1 (控制器 A)与 S3(控制器 B)两个节点对磁盘阵列的测试(测试项 1);测试项 1 是对处于两个不同控制器上的 LUN 进行的测试,还需要同时发起 S1 与 S2(其挂载了同一个控制器上的 LUN)对磁盘阵列的测试;测得数据之后,再同时发起 S1 至 S4 共 4 个节点的测试(测试项 2)。查看最后结果,如果测试项 2 的结果比测试项 1 的结果有明显的增加,那么可以继续从 RAID 组中分出相应的 LUN,再挂载到更多的机子上,最终测得磁盘阵列的数据传输率性能指标最大值。为了测得其最大性能指标,本节点所有的测试,如不特别说明,都是在全顺序读、全顺序写的条件下进行的。

测试 IOPS 性能最优值:设定传输数据块大小为 512B,其它测试方法和过程与数据传输率的测试方法相同。

3.3 数据传输率测试结果与分析

图 8 至图 11 显示了 iometer 不同参数对数据传输率的影响曲线。

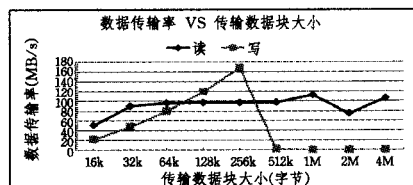


图 8 传输数据大小对数据传输率的影响



图 9 队列深度对数据传输率的影响

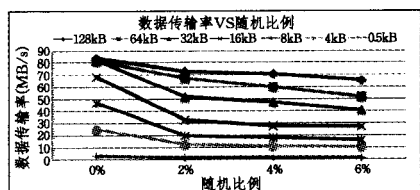


图 5 低度随机数据传输率测试结果

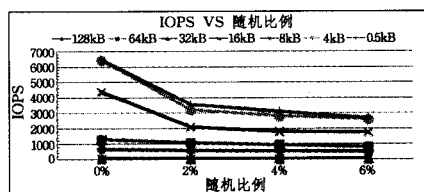


图 6 低度随机 IOPS 测试结果

3 磁盘阵列并行测试

本节通过使用 iometer 对单个磁盘阵列进行并行测试,验证磁盘阵列并行测试方法的可行性,得出测得磁盘阵列最大性能指标的测试方法与调优方法。

3.1 测试环境及配置

本节测试环境拓扑如图 7 所示。

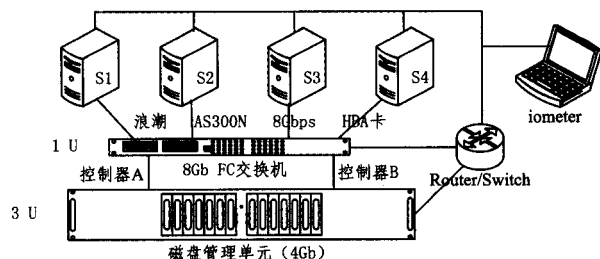
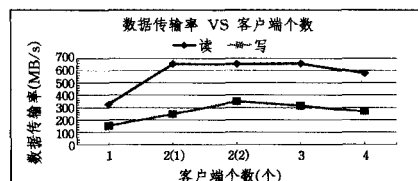
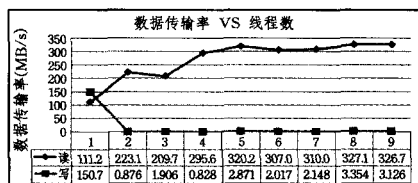


图 7 磁盘阵列测试拓扑结构图

被测磁盘阵列配置:共有 FC 接口、容量 300GB 的硬盘 4 块;共有两控制器,并设置成负载均衡模式;每个控制器有 4 个主机端口,每主机端口单向速率 4Gbps;每个控制器使用一根单模 FC 线通过一个主机端口连接到端口速率为 8Gbps 的 FC 交换机上。阵列中 4 个盘被分成两组,每组 2 个盘,组成一个 RAID 组(RAID 组 R1 与 RAID 组 R2)。RAID 组织形式都为 RAID0,条带大小为 64kB;从 R1 中划分出两个大小为 100GB 的 LUN(分别为 LUN01-A-100G-S1 与 LUN01-A-100G-S2),分配给控制器 A,并分别挂载到 S1 与 S2 上;从 R2 中划分出两个大小为 100GB 的 LUN(分别为 LUN01-B-100G-S3 与 LUN01-B-100G-S4),分配给控制器 B,并分别挂载到 S3 与 S4 上。

交换节点配置信息:图中 FC 交换机为 40 口 Brocade 5100,其单端口单向速率为 8Gbps;千兆以太网交换机型号为 D-Link DGS-1024T。

测试节点 S1 至 S4 配置信息:Inspur AS300N 存储服务器;数据宽度 64 位、频率 1333MHz、DDR3 内存 16GB;Dual Intel Nehalem 5502 双核存储专用处理器(LGA 1366)两颗,频率 1.87GHz,4MB Cache,对齐值 64 字节,133MHz 外频;安装有型号为 Emulex LPe1250、单向速率 8Gbps 的 Red Hat Enterprise Linux 5.3 系统,内核为 2.6.18-128.el5xen;测试



从图 8 可以看出,传输数据块对磁盘阵列数据传输率的大小有着非常显著的影响。在一定范围内,传输率会随着数据块的增大而增大。并且写带宽比读带宽最大值要大。对于写带宽在传输块为 512kB 时的急速下降,究其原因可能是 RAID 组只由两块盘组成,数据块大小可能使 LUN 在响应写操作时写操作顺序不当,造成随机写的比例大大增加。

从图 9 可以看出,队列深度(oustanding ios 参数)的影响非常小,很可能是使用的 Linux 内核对异步 IO 的实现不完全造成的。而且,写带宽仍然比读带宽要大。

从图 10 可以看出,线程数的影响也非常显著,特别是读带宽会随线程数的增加而增大,而写带宽会有很大的下降。本文没有对性能下降的原因进行分析与证明。

图 11 中的“2(1)”指被测的两个 LUN 是同一个控制器上的两个 LUN；而“2(2)”指被测的两个 LUN 分配在了两个不同的控制器上。从图 11 可以看出，客户端个数的影响也较为显著，但并不是客户端个数越多越好，而是在增加到一定程度之后，反而会有所下降。

3.4 IOPS 测试结果与分析

图 12 至图 14 显示了 iometer 相关参数对磁盘阵列的 IOPS 性能指标的影响。

outstanding io数(个)	读 (IOPS)	写 (IOPS)
1	3200	1400
3	3300	1400
5	3300	1600
6	3300	1700
7	3200	1400
8	3300	1500
9	3100	1500
11	3300	1500

图 12 队列深度对 IOPS 的影响

IOPS VS 线程数

线程数	读 (IOPS)	写 (IOPS)	混合 (IOPS)
1	3367.37571	1683.79594	3367.37571
2	1187.95402	5.88978286	1187.95402
3	851.467413	4.45577133	851.467413
4	330.64242	5.10496566	330.64242
5	244.812753	5.69785	244.812753

图 13 线程数对 IOPS 的影响

客户端个数 (控制器个数 X)	读 (IOPS)	写 (IOPS)
1	~3000	~1500
2(1)	~2500	~2500
2(2)	~6500	~2500
3	~6000	~2500
4	~4800	~2500

图 14 客户端个数对 IOPS 的影响

从图 12 可以看出,队列深度对阵列 IOPS 几乎没有影响。经过调查发现,Linux 并没有实现真正的异步 IO,使用 Windows 系统进行 iometer 测试将会得到更好的结果。图 13 则说明,随着线程数的增加,IOPS 性能会出现显著的下降。从图 14 可以看出,同时对处在同一控制器上的 LUN 进行测试,其 IOPS 性能参数会比测试单个 LUN 的性能要低;当同时对处在不同控制器上的 LUN 进行测试时,IOPS 性能指标可以达到最大。

从上面的测试结果可以看出,使用 iometer 测试磁盘阵列 IOPS 与数据传输率性能时,要设定相关最优参数值才能测得其性能最优值。在测试最终研制的磁盘阵列时,需要寻找其性能最优点。而对上层的并行文件系统而言,要达到较好的效果,其设计机制要着重考虑这些因素。

3.5 磁盘阵列满负载测试

上一小节证明了依次确定 iometer 相关参数测度方法的可行性,本小节将会对最新研制的满负载磁盘阵列进行测试。本测试的性能寻优过程由于篇幅较大,将不会给出。本小节的目的是通过 iometer,使用 3.3.4 节中的测试方法实现对满负载磁盘阵列的测试,验证测试方法的有效性。

由于对磁盘阵列性能的要求越来越高,传统磁盘阵列的架构已不能满足需求。同时为满足更高的可扩展性、可靠性、可用性、超高容量,现在高端磁盘阵列框架已得到了全面革新,单一阵列已演变成一小型的存储区域网络,通过多路径软件、分级存储软件实现对高端阵列的虚拟化,向用户提供简便的阵列操作与使用接口。

3.5.1 高端阵列测试拓扑

高端阵列测试拓扑如图 15 所示。

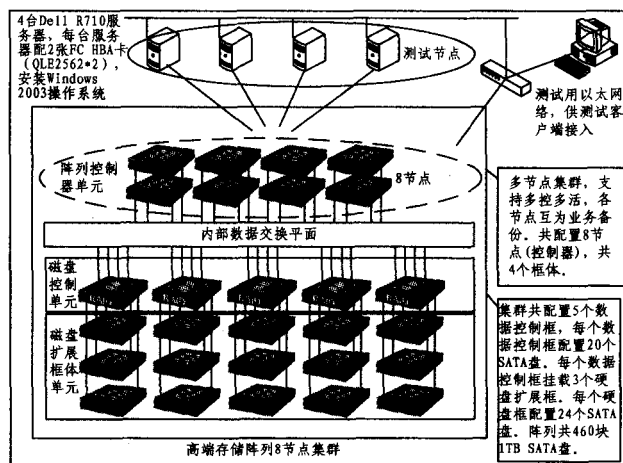


图 15 高端阵列拓扑

4个测试节点配置信息:型号 Dell R710;4颗 Intel Xeon E5620 CPU(主频 2400MHz,四核),共 16 核;6 根内存条(1066MHz,2GB),共 12GB;Windows Server 2003 Enterprise Edition Service Pack 2 系统;2 张 QLE2562(Qlogic 8Gb Dual-port FC HBA),每个测试节点有 4 根 FC 线与控制单元连接;测试工具 iometer-2006. 07. 27. win32. i386,采用联机测试。

8个控制器节点配置信息。2根内存条(667MHz,4GB),共8GB,仅用于阵列增值业务,非数据Cache;4个主机端口,

8Gbps FC 接口卡,测试时使用其中 2 个端口连接测试主机;4 个主机端口 8Gbps FC 接口卡(总速率为 24Gbps),使用其中 4 个端口连接内部数据交换平面;控制器之间关系为负载均衡。

阵列内部交换平面配置信息:4 个 20 口的 FC 交换机,单端口单向速率为 8Gbps。

磁盘控制单元信息:5 个磁盘控制单元;每个控制单元中有双控制器(Standby 模式);每个控制单元使用 4 个 4Gbps 速率的 FC 端口连向内部交换平面;内存大小 32GB,其中 28GB 用于数据 Cache;自身带有 20 个 1TB 大小的 7200RPM SATA 盘;使用 6 个 4Gbps FC 链路连接后端硬盘扩展框;每个磁盘控制器后面连接有 3 个硬盘扩展框;每个硬盘扩展框中有 24 块 1TB 大小的 7200RPM SATA 盘。

RAID 与 LUN 配置信息:共配置 50 个 RAID 组,每个 RAID 组 9 个盘;每个 LUN 大小为 500GB,共 80 个 LUN。

测试工具参数设定信息:队列深度(oustanding ios)为 90。

3.5.2 测试结果

IOPS 测试块大小:512B。

表 1 磁盘阵列 IOPS 测试结果

方式	IOPS			
	1	2	3	平均
全顺序读	1307508	1306649	1308867	1307675
全顺序写	885388.9	883478.1	883615.5	884160
全随机读	861279.2	860314.3	860385.6	860659.7
全随机写	711172.4	702073.4	701805.6	705017

数据传输率传输块大小:512k。

表 2 磁盘阵列数据传输率测试结果

方式	数据传输率(MB/s)			
	1	2	3	平均
全顺序读	9879.506	9879.766	9877.059	9878.777
全顺序写	10618.41	10612.54	10611.13	10614.03
全随机读	9910.732	9910.059	9912.918	9911.236
全随机写	10736.77	10733.85	10726.68	10732.43
全顺序 60%读	14067.15	14074.43	14067.52	14069.7

3.5.3 瓶颈分析

阵列中有 5 个磁盘控制单元,每单元与内部交换平面使用 4 个 4Gbps 端口连接,其总带宽为:5(磁盘控制单元数) * 4(连接的主机端口数) * 4(主机端口单向速率) = 80Gbps = 10GB/s;8 个阵列控制单元,每单元与内部交换平面使用 4 个 8Gbps 端口连接,但 4 个端口总带宽为 24Gbps(受到其端口后端卡的限制),故 8 控制器与内部交换平面总带宽为 24Gbps * 8 = 192Gbps = 24GB/s;4 个测试节点,每节点上有两张双口 8Gbps FC 卡连接向控制单元,故测试节点的总带宽为:8Gbps * 4(单节点端口数) * 4(测试节点个数) = 128Gbps = 16GB/s。以上分析的为单向瓶颈,由于其使用的都为 HBA 卡,它是一种全双工网络接口卡,因此以上分析所有组件的双向瓶颈应为单向瓶颈的 2 倍。从表 2 中结果可以看出对于全读或全写操作的测试结果,能达到大约 10GB 左右,此结果正是以上分析的 5 个磁盘控制单元与内部交换平面之间的瓶颈;在之后的测试中,为了测出更优性能,使用了读、写混

合的测度方法(60%读、40%写),测得的结果可达到 14GB/s,此值与其理论值 20GB/s 相差较远,故将不被视为瓶颈。对于 IOPS 的测试,因为测得 IOPS 值情况下的传输带宽没有达到各组件性能值的上限,磁盘控制单元在此种情况下没有成为瓶颈,所以并不进行读写混合的测试项。另外,从表 2 中的结果可以看出,随机读写测试结果并非比顺序读写的测试结果低,所以它可以大大地提高上层文件系统的性能。

结束语 本文分析了测试单个硬盘以及磁盘阵列的测试方法,并对影响磁盘性能的瓶颈进行了分析与验证。使用 iometer 测得了数据传输率与 IOPS 两项性能指标的最优值,找出了测试工具 iometer 影响测试结果的各项参数。最后,通过对高端磁盘阵列的测试验证了使用 iometer 测试磁盘阵列的可行性,并对测试结果进行了分析,得出了被测高端阵列的性能瓶颈点。

参考文献

[1] Katcher J. PostMark; A new File System Benchmark[R]. TR-3022, Network Appliance Inc., October 1997

[2] Agrawal N, Arpaci-Dusseau A C, Arpaci-Dusseau R H. Generating Realistic Impressions for File-System Benchmarking[C]//FAST 2009

[3] Agrawal N, Bolosky W J, Douceur J R, et al. A Five-Year Study of File-System Metadata[C]//FAST'07. San Jose, CA, February 2007

[4] Leung A W, Pasupathy S, Goodson G, et al. Measurement and Analysis of Large-Scale Network File System Workloads[C]//Proceeding of the USENIX Annual Technical Conference. Boston, MA, June 2008

[5] Muthitacharoen A, Chen B, Mazieres D. A Low-Bandwidth Network File System[C]//Proceedings of the 18th ACM Symposium on Operating Systems Principles (SOSP-01). Banff, Canada, October 2001:174-187

[6] Cha M H, Lee S M, Jin K S, et al. Design and implementation of a consistent update method for multiple file replicas in a distributed file system[C]//10th International Conference on Advanced Communication Technology. 2008;2063-2066

[7] Chao H C, Liu T J, Chen K H, et al. A Seamless and Reliable Distributed Network File System Utilizing Web space[C]//10th IEEE International Symposium on Web Site Evolution. Beijing, China, 2008;65-68

[8] Long L H, Choi E, Kim S, et al. Data Structures for Storing File Namespace in Distributed File System[C]//4th International Conference on Networked Computing and Advanced Information Management. Gyeongju South Korea, 2008;250-255

[9] Pletka R, Cachin C. Cryptographic security for a high-performance distributed file system[C]//24th IEEE Conference on Mass Storage Systems and Technologies. 2007;227-232

[10] Xiong J, Li J Y, Tang R F, et al. Improving data availability for a cluster file system through replication[C]//22nd IEEE International Parallel and Distributed Processing Symposium (IPDPS 2008). 2008;306-313