高能物理环境中混合存储系统的设计与优化

徐 琪[1,2] 程耀东[2] 陈 刚[2]

(中国科学院大学物理科学学院 北京 100049)[1] (中国科学院高能物理研究所 北京 100049)[2]

摘 要 高能物理是典型的数据密集型计算环境,数据处理包括模拟计算、重建计算以及物理分析。其中大文件计算占据较大比重,并且高能物理文件访问模式以跳读为主,因此大文件的高速访问成为整个系统性能的重要影响因素。首先剖析传统高能物理计算环境的典型架构及其文件访问模式的特点,介绍混合存储模式在高能物理计算环境中的优势,总结其数据访问方式的特点,对其各种读写方式进行数据测试;然后提出针对该环境的混合存储系统的部署设计和优化,使该环境下的数据读写性能得到明显提高;同时将成本因素考虑到系统设计中,实现了一个低成本高性能的存储系统。测试表明,混合存储系统在高能物理等大数据存储系统中具有高效的I/O性能。文中全面分析了影响其性能的各种因素,实现了最优化配置的低成本高性能混合存储系统,并对该系统的未来发展趋势进行了分析和展望。

关键词 海量存储系统,高能物理,混合存储系统,缓存,块设备,高性能计算,性价比

中图法分类号 TP399 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.10.014

Design and Optimization of Hybrid Storage System in HEP Environment

XU Qi[1,2] CHENG Yao-dong[2] CHEN Gang[2]

(School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China)[1]

(Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China)[2]

Abstract Computing in high energy physics (HEP) is a typical data-intensive application including simulation, reconstruction and physical analysis. Generally, the HEP experiment file is very big and the way of accessing to the files is usually skipping through large data blocks. Therefore, the performance of accessing to big files is one of decisive factors for the HEP computing system. Firstly, this paper analyzed the typical structure of the computing environment in high energy physics and the characters of accessing to files, introduced the advantages of hybrid storage system in high energy physics, summarized the characteristics of data access mode, evaluated the performance of different read/write mode, then proposed a new deployment model of hybrid storage system in high energy physics, which is proved to have higher I/O performance, at the same time the cost was considered to implement a high-performance system with low cost. The test result shows that the hybrid storage system has good performance in some fields such as HEP. Based on the analysis, it can help to get better I/O performance with lower price in High Energy Physics. At the last, the future of the hybrid storage system was analyzed.

Keywords Mass storage system, HEP, Hybrid storage system, Cache, Block device, HPC, Performance price ratio

伴随着人类对自然、物质组成、宇宙的不断探秘,高能物理的探索飞速前进,高能物理实验规模随之不断扩大,欧洲大型强子对撞机LHC、北京正负电子对撞机BEPCII、大亚湾中微子实验、羊八井国际宇宙线实验室等高能物理实验室每天都在提供海量的实验数据。经过长期的科学研究积累,越来越多的宝贵实验数据需要被高效存储。LHC每年会产生25PB的数据,并采用网格计算模型进行全球分布式的处理与分析[1]。BEPCII累计数据超过5PB,主要采用大规模计算集群进行数据分析。同时,正在建设的大亚湾中微子探测器二期实验等,在未来的高能物理实验中也将会产生更多的实验数据。

面对海量的实验数据,高效的存储设备成为了计算环境中不可缺失的部分,包括存储设备的性能、容量、可靠性和性价比等,都是存储设备必须考虑的,尤其是设备的读写性能与性价比更是重中之重。目前,主流的存储设备有硬盘、磁带、固态盘等,其中磁带多用于保存需持久性存储的冷数据,硬盘

到稿日期:2016-11-21 返修日期:2017-01-15 本文受国家自然科学基金项目:高能物理实验的大规模离线数据存储技术研究(11575223),国家重点研发计划项目:科学大数据管理系统(2016YFB1000605),国家自然科学基金项目:基于SDN的高能物理云数据中心弹性网络关键技术研究与应用(11605224)资助。

徐 琪(1990-),男,博士,主要研究方向为分布式存储系统、云计算,E-mail:xuq@ihep.ac.cn;程耀东(1977-),男,博士,副研究员,主要研究方向为云计算、海量存储、网格计算;陈 刚(1961-),男,博士,研究员,主要研究方向为大规模数据共享、网格技术。

多用于常见的计算设备，而固态盘虽具有高效的 I/O 性能但价格昂贵，因此多用于小数据存储[2]。将硬盘(HDD)和固态盘(SSD)有效地结合起来可以实现廉价、高效的大规模混合存储，为高能物理计算提供更优的存储环境。

针对混合存储设备，本文重点剖析了其数据读写机制，对其各种读写方式进行了全面的数据测试，并根据测试结果分析其存储性能特点；然后根据其特点，针对高能物理计算环境提出优化措施并分析影响其性能的因素；最后介绍了混合存储设备的不足和发展趋势，以及未来混合存储系统在高能物理计算环境中的应用前景。

1 高能物理计算系统的混合体系架构

1.1 高能物理计算系统的典型架构及文件访问模式

高能物理计算是典型的数据密集型计算应用，并以大文件为主，具有数据高并发的特点。高能物理计算的本质为从海量数据中挖掘稀有事例，由于事例之间的无关性和独立性，导致高能物理多为一系列事件组成一个大文件，多个文件可以在多个机器上同时处理，而不需要相互通讯[3]。因此，高能物理计算的特点是高吞吐率的数据并发，不需要采用 MPI 等并行计算技术。基于这些特点，目前高能物理领域普遍采用集群计算系统，并将计算和存储分离，典型的系统结构如图 1 所示。

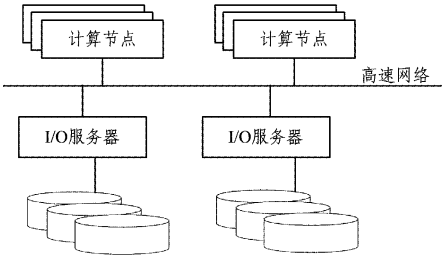

图 1 高能物理计算系统的典型结构

如图 1 所示，海量的实验数据存储在存储服务器即 I/O 服务器上，然后以高速网络为媒介将数据传输到各个计算节点，并在计算节点上进行数据分析。I/O 服务器及计算节点都以分布式节点的方式进行部署，然后通过作业管理系统和分布式存储系统分别对计算节点和 I/O 服务器进行管理。

高能物理数据处理包括模拟计算、重建计算以及物理分析 3 种类型。从探测器的建设到运行，一直到产生物理成果，都离不开模拟计算，该过程产生大量数据并消耗大量 CPU。事例重建用来处理探测器产生的原始数据与模拟计算产生的模拟数据，需要读入和写出大量数据，对 I/O 和 CPU 的需求都比较大。物理分析由物理学家发起，主要作用是读取重建数据进行数据处理和分析，这一步骤主要从海量数据中筛选出稀有事例，对系统的 I/O 要求很高。从上述可见，I/O 性能对高能物理计算环境有着十分重要的影响。高能物理计算节点进行任务分析时，大部分文件的连续请求大小分布在256k～4M 之间，每两个连续请求之间都有 offset，65%的 offset 绝对值分布在 1M～4M 之间，这说明文件的访问以较大数据块的跳读方式进行[4]。因此，采用缓存技术进行大数据块缓存，会极大地增强系统的 I/O 性能，提升高能物理计算环境的计算性能。

1.2 高能物理计算系统混合存储架构的设计

在整个高能物理计算数据流体系中，决定计算性能的主要因素是 I/O 服务器的吞吐率。在 I/O 块大小确定的情况下，I/O 服务器的 IOPS 成为影响高能物理计算系统性能提升的重要因素。现今主流的硬盘存储设备有 SATA 和 SAS 两种，SATA 读写慢，SAS 需要很多硬盘。不同类型的磁盘驱动运行特性由多个因素决定，其中包括硬盘转速、磁头速度以及每一转所读取的次数等[5]。同时，市场上闪存的价格又居高不下。针对这种情况，为了给高能物理计算提供良好的存储性能支撑，本文设计了高能物理计算系统中的混合存储架构，如图 2 所示。

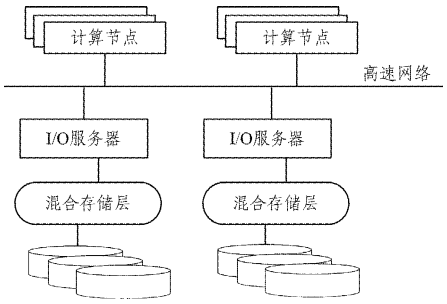

图 2 高能物理计算系统的混合存储结构

本混合存储系统，是在高能物理典型计算系统的 I/O 服务器与底层物理存储设备间添加了混合存储层，混合存储层使用了 Flashcache[6]技术。Flashcache 采用固态盘与机械盘混合的分层存储方式，它是一个 Linux 内核模块，使用时以内核模块的形式进行动态加载，不需要内核的修改，可以很好地保证磁盘上原有的数据布局[7]。Flashcache 作为一个通用的块设备缓存模块，有效地将固态盘读写性能好的特点和硬盘廉价且耐久性好的特点结合在一起，生成了特有的存储块设备，固态盘作为硬盘的缓存部分，加快了块设备的访存速度。虚拟块设备与磁盘逻辑地址一一对应，对于上层文件系统来说，访问该虚拟设备与访问磁盘无区别。

混合存储技术的设计基于 Linux 设备映射器框架(Device Mapper，DM)，位于虚拟文件系统层(VFS)与块设备层之间，如图 3 所示。在 Linux 内核中有很多基于 DM 框架的插件模块，如 LVM2(Linux Volume Manager Version 2)，EVMS(Enterprise Volume Manager System)，DMraid(Device Mapper Raid Tool)等[8]。

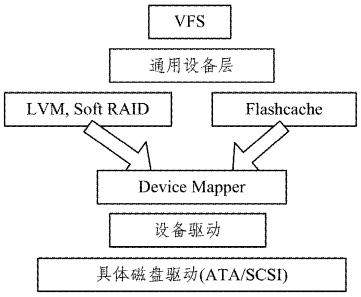

图 3 混合存储模块在内核栈中的层次示意图

混合存储将固态盘和硬盘上的数据以块(block)为单位进行划分，固态盘划分为元数据区和缓存数据区，之后默认以 512 个块为一个分组，通过组相连的缓存映射方式将硬盘块

映射到确定的缓存块中。硬盘块地址即磁盘块号以扇区为单位,利用直接映射的方式映射某个缓存组,然后通过线性哈希探测法映射到唯一的缓存块上,映射关系如式(1)和式(2)所示。

磁盘块所在分组＝磁盘块号/(块大小＊组大小)　　(1)

磁盘块所在数据块＝磁盘块号/(块大小＊组大小) mod 组数　　(2)

1.3　混合存储系统的工作机制

混合存储系统中,相比于容量较大的硬盘,作为其缓存部分的固态盘的容量是有限的,会有被耗尽的时候,因此混合存储系统使用了两种数据替换算法:FIFO(First In First Out)和 LRU(Least Recently Used)[9],FIFO 算法的原则是最早缓存的数据块被替换回硬盘,LRU 算法的原则是将最近没有调用的数据块写回硬盘。由于其元数据表是线性表,因此 FIFO 算法只需从分组元数据的入口位置顺序遍历,而 LRU 算法需要使用双向链表,链表头指向分组元数据,当有缓存块被访问时,链表尾指针指向该数据块,缓存替换时从 LRU 链表头指针指向的数据块开始查找[10]。

混合存储系统针对数据的读写请求采取了不同的缓存方式[11]。接收到 DM 层的请求后,首先判断请求是读缓存请求还是写缓存请求。读缓存中首先查找缓存区,查找结果分为读命中、读不中。读命中时直接从相应的缓存块中读取该数据,若读不中则先从磁盘中读取该数据,再将该数据写入缓存中,这样能避免将数据从磁盘复制到缓存中,但是增加了读不中的响应时间。若缓存块没有命中且找不到空闲或可替代的缓存块,则将数据访问重定向到磁盘。

写入模式有两种:写回模式(writeback)和写直达模式(writethrough)。写直达模式为同时写缓存和磁盘。写回模式中数据写到缓存中即是写入成功,等到脏缓存块的数量超过阈值或脏缓存块未经操作的时间过长时,再将脏块写到磁盘中,其固态盘作为缓存发挥了更好的性能。写请求中存在写命中和写不中两种情况,写命中时直接写数据到相应的缓存数据块中;写不中时则查找空闲或者可替换的缓存块,然后将数据写入该缓存块中,若找不到可以写入的数据块,则直接将数据写入磁盘。写入缓存操作完成后,需更新元数据区相应的元数据块。其中脏块写回是为了保证可用缓存块数量。脏块写回时会遍历整个分组,把相邻的脏块合并写回,以减少写操作的次数,进而减少磁盘的旋转操作。写回有如下两种条件策略。

1)阈值限制:分组内的脏块数量超过阈值时,进行脏块写回,直至脏块数量低于阈值。

2)时间限制:若分组内脏块存在的时间超过设定空闲时间没有被操作,则该块优先被写回。

2　混合存储系统的性能测试分析及优化

本文设计了适用于高能物理计算环境的混合存储系统,并针对该系统做了相应的比例读写压力测试,之后对测试结果进行了数学化分析,并根据分析结果对其进行了性能优化,验证了高能物理计算环境下该系统的优势性。

2.1　读写性能测试

2.1.1　测试环境介绍

(1)硬件测试环境

CPU:Intel(R) Xeon(R) E5-2407 v2 2.4GHz＊8

内存:8GB DDR3 1600MHz＊4

主板:Dell 0K7WRR

硬盘:HDD-Seagate ES.3 ST4000NM0033-9ZM170 磁盘＊1(转速 7200rpm,容量 4.0TB,缓存 128MB)

SSD- Samsung SSD 850 PRO S25UNSAG401352F 磁盘＊1(容量 256GB)

(2)软件测试环境

操作系统:CentOS 7.0 x64 Linux

内核版本:Linux version 3.10.0-229.1.2.el7.x86_64

(3)测试软件

fio-2.2.8[12]

(4)测试环境说明

本文测试了高能物理计算环境下混合存储系统的读写性能。高能物理计算环境下有大量分布式节点,本文选取其中一个节点作为测试对象,对其进行全面的性能测试。

2.1.2　3 种存储设备的性能对比测试

在本文的测试实例中,首先在相同测试条件下针对混合存储设备、固态盘设备(SSD)和硬盘设备(HDD)进行了读写压力测试比较,测试结果如图 4－图 7 所示。

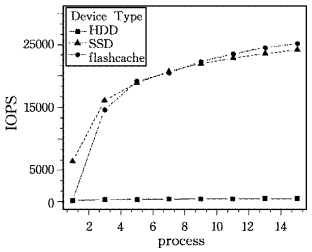

图 4　3 种存储设备随机读性能测试

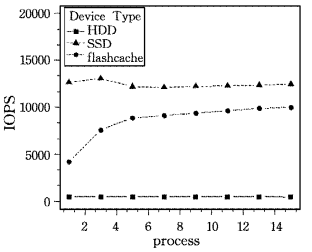

图 5　3 种存储设备随机写性能测试

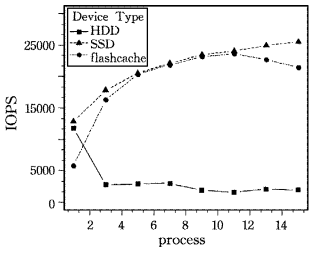

图 6　3 种存储设备顺序读性能测试

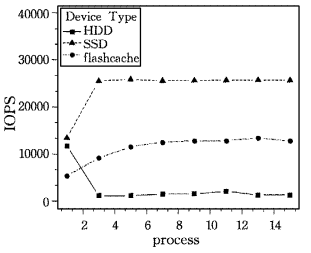

图 7　3 种存储设备顺序写性能测试

2.1.3　混合存储设备的混合比大小测试

混合存储设备是由固态盘作为硬盘的缓存层,因而提高了其访存性能。不同比例的混合构成对混合存储设备的性能有着很大的影响,可以说是混合存储设备性能的决定性因素。本文针对不同比例的混合存储构成进行了分析测试,如图 8－图 11 所示。在这些测试实例中,保持硬盘大小为 5GB,在 512MB～5GB 间不断调整固态盘的大小,不同的曲线表示固态盘含量的大小。由于随机读、写测试中,使用 5GB 固态盘的混合存储设备 IOPS 远远高于其他混合存储设备(原因详见 3.2 节),因此垂直轴使用 2 的对数模式进行标注。

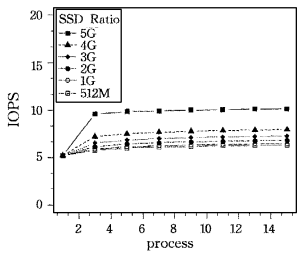

图 8 混合比随机读性能测试

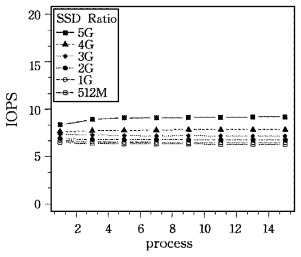

图 9 混合比随机写性能测试

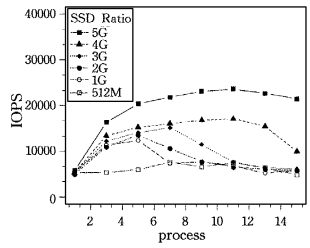

图 10 混合比顺序读性能测试

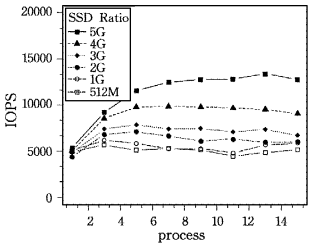

图 11 混合比顺序写性能测试

2.1.4 混合存储设备的压力文件大小测试

高能物理计算环境面对的主要是大文件读写的压力，数据文件大小也是影响系统表现的主要因素之一。为此，本文针对不同大小的文件进行了读写测试，测试结果如图 12－图 15 所示。

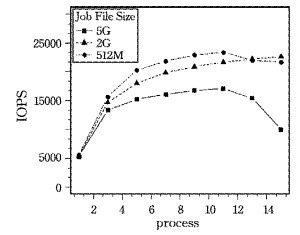

图 12 压力文件顺序读测试

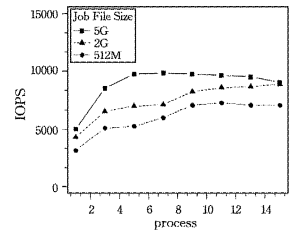

图 13 压力文件顺序写测试

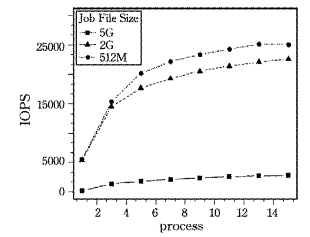

图 14 压力文件随机读测试

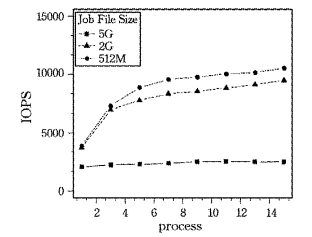

图 15 压力文件随机写测试

2.1.5 混合存储设备的缓存块大小测试

前文已提及混合存储系统是针对块设备进行调度的，所有缓存操作的基本单元都是缓存块，缓存块的大小(block size)对混合存储系统的性能有着很重要的影响。因此本文进行了多次的缓存块划分，并对每种缓存块进行读写性能测试，测试结果如图 16－图 19 所示。

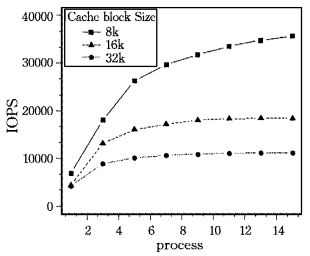

图 16 缓存块顺序读测试

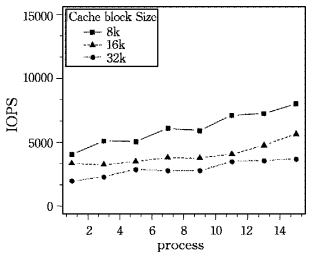

图 17 缓存块顺序写测试

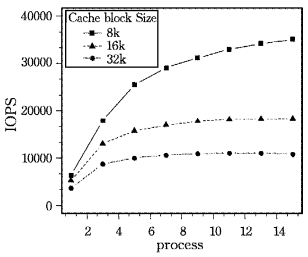

图 18 缓存块随机读测试

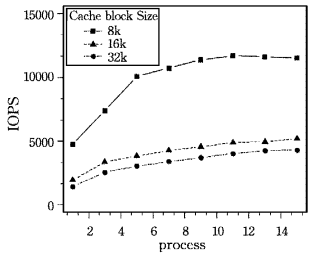

图 19 缓存块随机写测试

2.2 测试结果分析

从 2.1.2 节的测试结果可以看出，多进程并发情况下，随着进程的增多，硬盘的读写性能较稳定且较低；反之，固态盘表现出了较高的读写性能，与硬盘形成了鲜明的对比；混合存储的块设备表现出了较优异的性能，在读方面基本可以与固态盘保持相同量级，在写方面也可以达到令人满意的性能。本文发现，在典型的高能物理存储环境下，硬盘能提供的存储性能远不及混合存储设备，同时混合存储设备的价格远低于性能略高的固态盘设备。从性能和成本角度来看，混合存储系统都表现出了优势性。

本文在 2.1.3 节、2.1.4 节以及 2.1.5 节分别对影响混合存储系统性能的 3 个重要因素进行了测试。前文已提及混合存储设备是固态盘和硬盘的混合设备，这种混合设备的混合比例直接影响了其性能。2.1.3 节的测试中始终保持硬盘的大小为 5GB，将固态盘的大小从 512MB 不断调整至 5GB，并测试混合存储设备的 I/O 性能变化。2.1.3 节中提到固态盘大小为 5GB 时，混合存储设备的 IOPS 远高于其他数据点，这是因为此时缓存的大小等于数据文件的大小，混合存储设备上的文件读写全部位于缓存区，相当于直接读写固态盘设备，因此 IOPS 极高。当固态盘所占比例下降时，其性能开始较大幅度地下降，因为此时数据出现了缓存不命中的情况，导致出现数据块替换，开始读写硬盘。从测试结果可以看出，固态盘比例越高，混合存储设备性能就越高，并发进程数达到 11 时其性能基本可以达到最高值。

从 2.1.4 节的测试结果可以看出，测试文件小于 5GB 时 IOPS 较高，且测试文件大小变化时 IOPS 并没有较大波动，较小文件的 IOPS 略高于大文件；而测试文件超过 5GB 时会出现缓存不命中的情况，IOPS 急剧下降。可见，面对不同压力的访存，混合存储设备的性能会受到一定的影响。

在 2.1.5 节的测试结果中，缓存块大小(bs)为 8kB 时 IOPS 值较大，16kB 时 IOPS 略高于 32kB 时的 IOPS，但是相差不大。缓存块是硬盘和固态盘上数据存储的基本单元，因此当缓存块较大时每个存储单元存储的数据量变大，数据划分粒度变粗，I/O 调度次数就会减少，因此 IOPS 会更高。另外，当缓存块小于 16kB 时，IOPS 会有很大的提高，这是因为 Flashcache 是以 16kB 的数据页与内存层数据缓存池进行交互的，不足 16kB 的数据块会被合并至 16kB 后进行调度[13]。但缓存块过小会导致 CPU 过于繁忙，使得系统 I/O 速度下降。综上，缓存块大小为 4～8kB 较好。

2.3 混合存储系统的优化

从 2.2 节的测试结果中得到了混合存储设备性能的总体

变化趋势。为获取高能物理计算环境下的高性价比存储设备,本文定义变量 VOF(Value of Flashcache)来表示其性价比高低。前文已介绍在进程数为 11 时,各种读写方式的性能基本达到最高。截取进程数为 11 的相应数据点,以硬盘设备的读写性能为基准数值,计算其差值,以随机写为例,结果如表 1 所列,纵向代表混合存储设备中固态盘所占大小。

表 1　随机写性能测试差异值

混合存储设备中固态盘大小	IOPS 测试值	与硬盘性能差异值
0(硬盘)	471	0
512MB	545	74
1GB	632	161
2GB	836	563
3GB	1297	826
4GB	2513	2042
5GB	9619	9148

除性能外,本文还考虑到设备成本。据市场调查发现,固态盘容量增大至 2 倍,价格大致增加至 1.5～3.0 倍,不同品牌略有差异。文中定义了设备价格增长率 Price_x,用于表示相同容量固态盘与机械硬盘的价格差值。将价格增长率考虑到 VOF 变化中,使用性能差(IOPS 差值,其中 IOPS_x 表示含有 xGB 固态盘的混合存储设备 IOPS,IOPS_0 表示相同容量机械硬盘设备的 IOPS)与价格增长差的比值表示系统性价比(如式(3)所示)进行系统性价比评估,得出其性价比增长率,如图 20 所示(图为散点图,曲线为其多项式拟合图)。使用多项式将散点进行拟合,得到式(4)所示的多项式。

$$VOF=(IOPS\_x-IOPS\_0)/Price\_x \tag{3}$$

$$y=-3.681x^6+45.276x^5-245.26x^4+613.26x^3-722.49x^2+383.86x+1E^{-06} \tag{4}$$

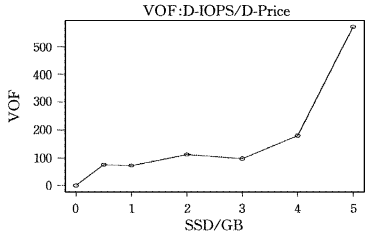

图 20　随机写实例的性价比增长

从图 20 可以推解出,当固态硬盘为 3.5GB 左右时,混合存储设备的性价增长比达到 100 量级,即达到高量级,此时系统对于固态硬盘的消耗相对最小,混合存储设备具有较高的 VOF 及性价比。

影响混合存储设备性能的因素主要有 4 点:硬盘和固态盘的混合比例、并发访问进程数、缓存块大小以及被访问数据文件的大小。其中,并发进程数和被访问数据文件大小主要取决于用户需求,并且影响力相对较小。实验主要针对存储设备混合比和缓存块大小的配置对系统进行改进和性能优化。混合存储的混合比决定了缓存数据区中文件的读写命中率,提高混合存储比可以提升缓存命中率,进而加速文件的访存速率。缓存块是逻辑设备操作的基本单位,相当于一块固态硬盘与一块机械硬盘的映射结合体,数据在缓存块上读写时,首先在固态盘上进行查询,缓存块的大小直接决定了单次访存的命中率大小。

通过测试结果发现,本环境下的最优配置为:硬盘与固态盘比例为 1.43:1,并发进程数为 11,缓存块大小为 8kB,数据文件大小为 2GB。

上述测试为高能物理实际存储环境的缩小比例测试。高能物理存储环境为 PB 级存储,在高量级的数据环境下,存储系统的性能也适用本文中相应的比例参数。使用该配置参数对本地高能物理计算环境进行优化配置,并对实际高能物理作业的运行情况进行测试,测试结果如图 21、图 22 所示。

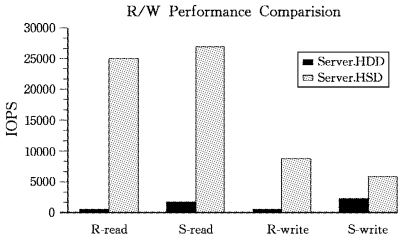

图 21　优化前后读写性能的对比

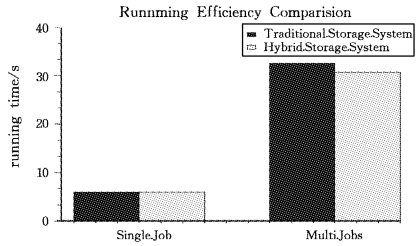

图 22　BOSS 作业运行时间的对比

图 21、图 22 分别显示了传统存储系统和优化后的混合存储系统的读写性能及高能物理 BOSS 作业的运行情况。结果表明,优化后的读写性能,尤其是读性能,得到了很大提升,并且高能物理作业,尤其是多作业时,其完成时间大幅减少。

3　混合存储系统的不足及发展前景

本文中高能物理混合存储系统引用了 Flashcache 技术,该技术主要有两点不足:1)元数据的管理,其采用同步更新和批量更新方式,当不断有元数据改变时,会进行更新等待,增加了更新时间,这时需要将其优化为异步更新来加快元数据处理速度;2)没有进行 I/O 并行优化,不能提供很好的并发性能,所有 I/O 都为串行执行,这是由于整个缓存使用一个互斥锁,锁的粒度过大,可以考虑引用数据集(set)的概念,每个数据集使用一个互斥锁,这样可以提供更好的 I/O 并行化,从而优化系统性能。

随着 CPU 计算速度的不断提升,存储设备的 I/O 性能成为了系统性能的主要瓶颈,尤其是在高能物理这种数据密集型计算环境下,突发的大数据访问对存储设备的 I/O 性能造成了巨大的挑战,单纯的硬盘设备已经渐渐不能满足这种高 I/O 需求。2006 年,三星公司发布了第一款固态硬盘笔记本,2009 年固态盘呈现井喷式发展,并逐渐占据了存储市场。各种混合存储设备也在此时踏入了存储设备的舞台,Flashcache 就是在 2010 年被 Facebook 推出的。

固态盘具有读写速度快、防震抗摔、低功耗、轻便等优点,但是也有容量小、寿命短、售价高等缺点,尤其是售价高、容量

(下转第 90 页)

[6] LITZKOW M,LIVNY M,MUTKA M. Condor-A Hunter of Idle Workstations [C]// Proceedings of the 8th International Conference of Distributed Computing Systems. IEEE,1988:104-111.

[7] XU R S,LANG P F,CHEN Y Q,et al. BES Offline Data Processing [J]. High Energy Physics and Nuclear Physics,1991,15(7):577-583. (in Chinese)
许榕生,郎鹏飞,陈雅青,等. 北京谱仪数据的离线处理[J]. 高能物理与核物理,1991,15(7):577-583.

[8] WANG Y F. A Neutrino Experiment Using the Daya Bay Reactor [J]. Physics,2007,36(3):207-214. (in Chinese)
王贻芳. 大亚湾反应堆中微子实验[J]. 物理,2007,36(3):207-214.

[9] NIE S M,ZHANG J L,TAN Y H,et al. Real Time Transmission and Analysis of the Yangbajing Cosmic Rays Observation Data [J]. Nuclear Electronics and Detection Technology,2007,27(1):14-17. (in Chinese)
聂思敏,张吉龙,谭有恒,等. 羊八井宇宙线观测数据实时传输及处理系统[J]. 核电子学与探测技术,2007,27(1):14-17.

[10] 江门中微子实验[EB/OL]. http://www.ihep.cas.cn/dkxzz/juno.

[11] 高海拔宇宙线观测站[EB/OL]. http://www.ihep.cas.cn/dkxzz/lhaaso.

[12] TORQUE Resource Manager-Adaptive Computing[EB/OL]. http://www.adaptivecomputing.com/products/open-source/torque.

[13] Maui-Adaptive Computing [EB/OL]. http://www.adaptivecomputing.com/products/open-source/maui.

[14] RAMAN R,LIVNY M,SOLOMON M. Matchmaking:Distributed Resource Management for High Throughput Computing [C]//Proceedings of the Seventh IEEE International Symposium on High Performance Distributed Computing. Chicago,1998.

[15] LAHIFF A,DEWHURST A,KELLY J,et al. HTCondor at the RAL Tier-1 [OL]. https://indico.cern.ch/event/272785/contributions/1612799.

[16] Center for High Throughput Computing,University of Wisconsin-Madison. HTCondor Manual [EB/OL]. http://research.cs.wisc.edu/htcondor/manual/v8.5/3_1Introduction.html.

(上接第 79 页)

小,已成为其不能替代硬盘的主要问题[14]。Flashcache 等分层混合存储技术应运而生,并且表现出较好的性能。目前随着工艺的逐渐精进,纯固态盘设备已经逐渐普及,纯硬盘设备开始慢慢被取代。相信在不久之后,纯固态盘设备会拥有较大的存储空间和较低的价格,进而可能完全替代硬盘设备,但就目前而言,Flashcache 等分层混合存储技术仍起着十分重要的作用。

结束语 海量的高能物理实验数据、高并发的大文件访问对文件存储系统有着极高的要求,传统的硬盘设备已经不能很好地满足系统需求。混合存储系统有着较高的性价比,可以为高能物理计算环境提供廉价的高性能存储设备,虽然它有着一定的缺陷,但是针对高能物理计算环境,部署混合存储系统可以带来很高的性能收益。本文对混合存储技术进行了详细的原理分析和数据测试,并对其性能影响因素进行了细致化分析,总结出了混合存储的优化配置公式,对高能物理以外的其他大数据系统也有着借鉴意义。随着存储设备制造工艺的不断改进,纯固态盘设备逐渐普及,未来的存储系统架构、存储技术也会不断革新进步,但是目前分层混合存储技术仍有着重要作用。

参 考 文 献

[1] WLCG-Worldwide LHC Computing Grid[OL]. http://lcg.web.cern.ch/LCG.

[2] CABRERA L,LONG D D E. Swift:Using Distributed Disk Striping to Provide High I/O Data Rates[J]. Computing Systems,1991,4(4):402-441.

[3] CHENG Y D,SHI J Y,CHEN G. A survey of High Energy Physics Computing System[J]. e-Science Technology & Application,2014,5(3):3-10. (in Chinese)
程耀东,石京燕,陈刚. 高能物理计算环境概述[J]. 科研信息化技术与应用,2014,5(3):3-10.

[4] CHENG Y D,WANG L,HUANG Q L,et al. Design and Optimization of Storage System in HEP Computing Environment [J]. Computer Science,2015,42(1):54-58. (in Chinese)
程耀东,汪璐,黄秋兰,等. 高能物理计算环境中存储系统的设计与优化[J]. 计算机科学,2015,42(1):54-58.

[5] MITUZAS D. Flashcache at Facebook From 2010 to 2013 and beyond[EB/OL]. [2013-10-9]. https://www.facebook.com/notes/facebook-engineering.

[6] Facebook/Flashcache[OL]. https://github.com/facebook/flashcache.

[7] MORE A,GANJEWAR P. Dynamic Cache Resizing in Flashcache[J]. Advances in Intelligent Systems and Computing,2015,327:537-544.

[8] 敖青云. 存储技术原理分析:基于 Linux 2.6 内核源代码[M]. 北京:电子工业出版社,2011:363-476.

[9] LEE D,CHOI J,KIM J H,et al. On the Existence of a Spectrum of Policies That Subsumes the Least Recently Used (LRU) and Least Frequently Used (LFU) Policies[J]. SIGMETRICS Performance Evaluation Review,1999,27(1):134-143.

[10] RAMOS L E,GORBATOV E,BIANCHINI R. Page placement in hybrid memory systems. [C]// Proceedings of the international conference on Supercomputing. ACM,2011:85-95.

[11] YANG Z Y. Design and Implementation of Hybrid Storage Scheme Based on Flashcache[D]. Wuhan:Huazhong University of Science and Technology,2013. (in Chinese)
杨昭宇. 基于 Flashcache 的混合存储方案设计与实现[D]. 武汉:华中科技大学,2013.

[12] Freecode/fio[OL]. http://freecode.com/projects/fio.

[13] FRÜHWIRT P,HUBER M,MULAZZANI M,et al. InnoDB Database Forensics[C]//24th IEEE International Conference on Advanced Information Networking and Applications (AINA). IEEE,2010:1028-1036.

[14] LIN T J. Principle,Composition and Application of Several Typical Solid State Drives[J]. Computer and External Equipment,1999(1):16-21. (in Chinese)
林天静. 几种典型固态盘的原理、组成形式及应用[J]. 电子计算机与外部设备,1999(1):16-21.