一种超大容量自动光盘库的设计与实现

曹 强[1,2],严文瑞[1,2],姚 杰[2],谢长生[1]

(1. 武汉光电国家实验室,湖北 武汉 430074; 2. 华中科技大学 计算机学院,湖北 武汉 430074)

摘 要: 当前蓝光光盘的寿命已超过 50 年,光存储的可靠性远高于硬盘,寿命也远长于磁带,但单盘容量较小、存取性能较低的缺点限制了光盘在大规模归档系统中的应用。提出了一种新型超大容量机械手自动换盘的光盘库系统,该系统能够在标准尺寸的机柜中容纳 12 000 张蓝光光盘,数十个光驱可并行读写,对外的吞吐率达到 1 GB/s。除了高度并行之外,还使用了磁光电融合结构和虚拟化存储机制,通过磁电作为光存储的大容量缓存,提高存取性能,将大量的光盘存储空间虚拟成单个文件卷存储池。该系统的光盘调度、刻录和读取完全实现自动化,并提供给用户通用文件访问接口。综合这些技术,既发挥了光存储介质的大容量、长寿命、低成本、低能耗的优点,又克服了光存储系统速度慢、性能低的缺点,同时提供了用户友好的使用界面和环境,实现了与现有信息系统的无缝对接。

关键词: 光盘库; 数据归档; 文件系统; 蓝光; 存储系统; 自动化

中图分类号: TP334 **文献标志码:** A **DOI:** 10.3788/IRLA201645.0935003

Design and implementation of an ultra-large scale automatic optical disc library

Cao Qiang[1,2], Yan Wenrui[1,2], Yao Jie[2], Xie Changsheng[1]

(1. Wuhan National Laboratory for Optoelectronics, Wuhan 430074, China;
2. College of Computer Science, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: More and more digital information needs long term preservation, a cost-efficient storage system is needed to ensure long term data availability. Using tape and hard disk as storage medium can not meet the demands of long-term data preservation. Current blu-ray inorganic disc is able to store data more than 50 years, the fact that the drive and optical disc is separate makes it easy for optical disc to be stored. Optical disc is proved to be a choice for long term preservation. But the capacity of optical disc is too small compared to tape and hard disk, which limits its use in archival storage system. A novel large-scale Automation optical library system is introduced, a standard rack contains more than 10 000 discs, provides 1 GB/s throughput. The fundamental thought is tiered storage with HDD, SSD and optical disc and storage virtualization. HDD and SSD works as cache for optical storage level, thus improve performance; all the optical discs are set to a virtual volume pool. Optical library system implements automatic mechanical dispatch, file access and disc recording by the integration design controll and inner data structure. A common file system interface is provided to users. Experiments result shows that the system can automatically read and record files.

Key words: optical disc library; data archival; file system; blue ray; storage system; automation

收稿日期: 2016-08-01; **修订日期:** 2016-09-02

基金项目: 国家重点基础研究发展计划(2011CB302303);国家自然科学基金重点项目(61432007);
中央高校基本科研业务费专项资金(2013KXYQ003);信息存储系统教育部重点实验室专项资金

作者简介: 曹强(1975-),男,教授,博士,主要从事大规模存储系统、存储系统设计与优化、计算机性能评价方面的研究。
Email:caoqiang@hust.edu.cn

0 引 言

关键数字信息的长期保存是现代社会面临的一个重要问题。一些法律法规对于数据的长期保存也有强制性规定。比如金融交易信息至少应该保存 7 年甚至更长；关键设计文档至少应该保存 15 年；而医疗记录应该被保存至少 30 年[1]。互联网应用提供商往往希望能够把用户邮件或者上传照片保存超过 20 年。这些长期保存数据的一个特点，就是一旦被保存之后就很少被访问了，也称之为“冷数据”。随着大数据时代的到来，迫切需要低成本、长期可靠保存“冷数据”的技术和系统。

与主流磁盘(5 年寿命，能耗大[2])、固态盘(5 年寿命[3])和磁带(10 年寿命，但是保存条件要求较高[4-5])相比，光盘在数据长期保存的可靠性及能耗方面的优势比较明显。目前蓝光光盘的寿命已超过 50 年，同时，光存储由于具备存储介质与驱动器分离的特性，可以很好地满足归档存储要求。

在数据归档长期保存领域有一条公认的 3-2-1 原则：每份数据至少需要三个以上的副本，这些副本要被存储在至少两种不同的存储介质上，而这些存储介质中至少有一种可移动的离线存储介质。光存储介质正是满足这种要求的介质。

在过去 10 年，由于网络音视频的飞速发展和软件分发的网络化，光存储在音视频和软件分发应用市场不断萎缩。但是随着以蓝光为代表的大容量光盘出现，光存储的主流应用已转向大规模数据存储和归档，特别是在数据中心的“冷数据”存储中呈现出良好的应用前景。目前市场上单张蓝光光盘容量可以达到 300 GB 以上[6]；随着纳米和材料技术的发展以及光衍射极限的突破，以全息、超分辨和多维光存储为代表的新型高密度光存储原理也取得重大进展，未来有望实现单张光盘容量超过 10 TB[7-9]，呈现出非常诱人的前景。此外，光存储还具有防电磁干扰和防水功能，在一些自然灾害中数据也可以被保存下来[10]。

目前光盘存储也有一些明显缺陷，主要是性能低、速度慢。如 100 GB 光盘标准规范仅定义 4倍速刻录，单张光盘读取和写入峰值速度低于 40 MB/s，远小于磁盘(150 MB/s)和磁带驱动器(200 MB/s)。目前单张光盘的容量和性能无法满足数据中心海量和高性能的存储需求。

为了使光盘在数据中心得到应用，需要使用系统级的技术来提高光盘存储系统的性能、可靠性以及可扩展性。

设计和开发了一种超大容量光盘库，把超过万张的光盘放置在一个库体内，通过机械装置实现光盘的自动存取，并使用大量光盘驱动器实现高并行的光盘读写操作，最后能够通过高速网络 Infiniband 或者光纤通道实现和前端系统的无缝高速互连[11]。在此大容量光盘库硬件之上，进一步设计了基于大容量光盘库的文件系统 OLFS (Optical disc Library File System)，可以将数千张光盘虚拟为一个大容量存储池，并为用户提供了标准的文件读写接口，对于上层应用隐藏底层光盘库机械调盘动作，以及光盘数据自动刻录和读取等物理操作。

1 相关工作

1.1 光盘库

传统光盘库主要来源于多年前的 DVD 点唱机，在机械方面，仅能容纳最多几百张光盘，使用专门盘匣存放 DVD 光盘；而且机械装置和电子控制装置是分离的，仅通过有限的光敏传感器定位机械臂位置，可靠性较低。在计算机硬件方面，使用单片机控制机械臂运动，并且采用老式 SCSI 接口和 SCSI 的 DVD 刻录机，必须通过前置主机和 SCSI 线缆及其接口存取光盘库。具体使用时需要管理员手工进行数据组织和刻录，刻录速度也低于 5 MB/s，而且没有纠错检错功能[12]。

目前新型大容量光盘库能够在单个机柜大小内容纳上万张大容量蓝光光盘，2014 年 1 月 Facebook 给出容纳万张光盘的原型物理样机；松下和索尼公司也推出 6U 机架 1 080 片光盘库，12 张光盘构成一个 RAID 组，设计专用前置机管理光盘库和 RAID 组。

1.2 光盘数据组织

目前最为流行的单张光盘数据组织格式是 UDF 文件系统[13]。当采用 UDF 文件系统写入时，光盘只要在光驱中被挂载即可直接读取文件数据。

UENO 等提出过基于光盘的防灾型存储系统[14]，将光盘作为磁盘的备份，并且采用分布式结构来实现容灾，数据通过网络传输，然后再用 SCSI 指令直接写入光盘。当需要读取数据时，数据先通过 SCSI

指令读取,然后通过网络传输到客户端。这种方案采用 SCSI 指令读写光盘,数据的读写都受到上层系统的制约,不具有普适性。

Thompson 介绍了一种基于 UDF 文件系统的光盘库存储系统[15]。该系统面向多媒体文件数据,通过 UDF 文件系统的特性将多张光盘虚拟为一个容量池,然而该存储系统依旧没能提供一个通用的文件访问接口。

上述工作都是考虑单张光盘的数据组织结构,并没有特别考虑如何有效管理大量光盘及其数据。对于大容量光盘库而言,人工管理成千上万张光盘及其上面的数据是不现实的[16],因此文中将针对光电机械一体化的大容量光盘库设计相适应的文件系统自动管理海量光盘及其数据。

2 光盘库整体结构

基于大容量光盘存储的需求,设计出以磁光电一体化的光盘库机柜架构,采用文件系统将海量光盘虚拟为一个可扩展的大容量存储池,并能提供标准的文件读写接口进行存取,屏蔽底层的光盘刻录和读取以及机械调度细节。

2.1 大容量光盘库结构

光盘库是由机械部分和软件部分组成的,其中机械控制部分包括可编程逻辑控制器 (PLC)、机械臂、传感器和机械转笼。硬件资源包括磁盘、光驱以及光盘。磁盘和光驱是通过 SATA 接口连接到位于机柜顶端的控制服务器上,光盘则摆放在转笼结构的托盘中。服务器与 PLC 通过 TCP/IP 协议通信来决定需要执行的机械操作。机械臂则负责准确地将光盘从转笼取到光驱或从光驱取回转笼。

2.2 光盘库机械结构

机械部分由转笼和自动机械臂组成。光盘库的关键机械物理结构如图 1 所示,海量的光盘摆放在转笼的光盘槽中,转笼截面形似 6 瓣莲花,也就是每层具有 6 个光盘槽。转笼分为多列,每列分为多组。目前一个转笼能够放置 6 120 张光盘。当转笼转动时,光盘槽能够弹出和收回,每个槽中摆放了光盘。自动机械臂仅需要在垂直方向上进行位移,光驱则放在机械臂的正前方,机械臂能将光盘槽中的光盘取出或放回。这些动作都由 PLC 进行控制,PLC 还控制诸多传感器来监控这些动作是否准确地完成。

PLC 提供了控制转笼的转动和机械臂的上下位移的一系列指令,服务器通过这些指令与 PLC 进行交互,实现自动化处理。

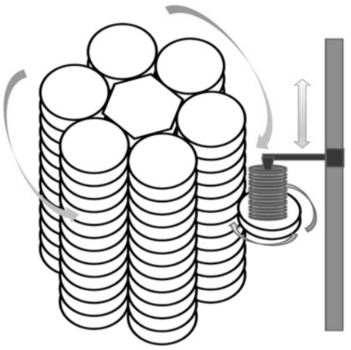

图 1 光盘库机械结构示意图

Fig.1 Schematic diagram of optical disc library mechanical structure

当执行取盘动作时,PLC 会接收到光盘所在的编号,根据编号可以得出光盘所在的列号以及在该列中的具体位置,转笼会转动该列到机械臂垂直方向将该光盘槽扇出,然后机械臂会移动到该光盘槽的位置将光盘抓出,再进行垂直移动到光驱位置,光驱弹出,这时机械臂松开,光盘会被放在光驱中,动作执行完成。这个过程中,机械臂抓盘、光驱弹出、机械臂放盘都有传感器进行监控。当出现故障导致动作未能执行完成时,这些传感器会返回错误,PLC 再将这些错误返回给服务器进行故障的分析和处理。执行退盘动作时流程和取盘动作相反。

2.3 光盘库系统结构

图 2 所示为大容量光盘库系统结构的逻辑结构。光盘库逻辑上由数据缓存区、光驱组和光盘组构成。由于常用的可记录光盘具有一次写的特点,并且单张光盘写性能有限,为了提高系统整体吞吐量并且适应光存储记录特性,光盘库使用硬盘或者固态盘构成数据缓存区,需要导入光盘库的数据首先高速存放在数据缓存区中,在满足条件后,一次性刻录整张光盘。此外,只有光盘装入光驱才能实现光盘数据的物理刻录和读取。光盘库通过机电一体化装置管理光驱组和光盘组,光驱组和光盘组之间通过机械臂实现光盘物理的装入和退出光驱。光盘库为每张光盘和光驱建立逻辑标识,光盘物理标识包括光盘全局编号、在光盘库中的物理位置、序列号和一些

基本属性;光驱物理标识也包括编号、在光盘库中的物理位置和关键参数等。

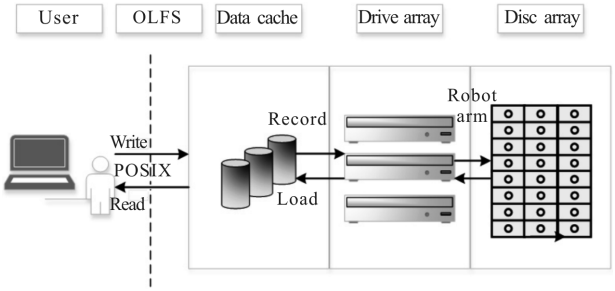

图 2 光盘库系统结构

Fig.2 Architecture of optical disc library

2.4 大容量光盘库文件系统 OLFS

在物理光盘库基础之上，需要设计和实现光盘库文件系统来管理海量光盘及其数据，全局逻辑结构主要设计全局文件系统索引和光盘索引结构。图2展示光盘库对外提供标准 POSIX 文件系统接口,把所有光盘的存储空间合并为一个具有统一文件目录树结构的逻辑存储卷，实现应用程序和用户对于光盘库的存取。在全局逻辑结构的基础之上,设计了大容量光盘库文件系统 OLFS,该系统为全局文件数据提供高效检索机制,方便全局数据的查找。

目前 OLFS 在单张光盘上数据存储采用 UDF 文件系统,并设计了一套 UDF 文件系统的底层光盘刻录机制。理论上 OLFS 可以设计一套专用的单张光盘数据组织格式，但是考虑到所有光盘能够和现有计算机系统中光盘管理系统兼容和可识别性，OLFS 采用标准 UDF 格式，这一点区别于现有松下光盘库基于 RAID 组的光盘库内部格式。但是为了提高光盘库整体数据可用性,OLFS 也在 UDF 格式基础之上设计盘内冗余机制和盘间冗余机制，通过生成冗余文件方式实现盘内和盘间数据冗余。

3 大容量光盘库实现机制

在介绍大容量光盘库的整体结构之后，该节讨论光盘库的主要设计考虑和具体实现机制。

3.1 文件分盘策略

OLFS 对外提供统一的文件卷逻辑视图，因此，应用程序无需感知物理光盘的存储边界。但是在实际数据存储过程中，需要将整个逻辑空间划分到各个物理光盘内，因此 OLFS 需要把统一全局文件卷逻辑空间分割到一张张物理光盘上。

OLFS 使用数据缓存区把导入的数据集首先保存在硬盘或者固态盘中;然后根据光盘分配算法,将整个数据集按照 UDF 格式划分为一定大小的光盘镜像,包括生成相应的冗余文件;之后一次性把镜像文件刻录到物理光盘中;最后在刻录成功后更新 MF 文件。数据分配的原则如下。

(1) 单张光盘的空间利用率

这里说的空间利用率指的是可用空间利用率。例如对一张容量为 25 GB 的光盘，能长时间稳定存放数据的区域不足 25 GB，用户可以设置其可用空间小于 25 GB 来保证数据安全，同理 50 GB 和 100 GB 的光盘也是如此。分配的单个数据集大小就是用户事先设定的可用空间大小，这样可以充分利用可用空间。

(2) 最大限度地保证关联性文件邻近存放

同一个目录下的文件可以视为关联性强的文件。当某个目录下的某文件被访问后,该目录下的其他文件被访问的可能性也很大。文件邻近存放就是将文件尽量存放在同一张物理光盘上，为了减少取盘调盘等耗时的机械动作，很有必要将同目录下的文件在光盘上邻近存放。

(3) 保证单张光盘可独立读取

单张光盘可独立读取是指每张光盘里的文件可以在光盘库以外的光驱上被读取，这就意味着光盘上的每个文件都是未分割的完整文件，会导致空间利用率下降。这样做的好处是当光盘库遇到一定自然灾害时，未损坏的光盘内容可以直接在光驱中被读取出来,不一定依赖于其他光盘数据的可存取,数据的可用性得到了提升。

然而同时满足上述三种原则的分割方法是不存在的。因为要充分利用可用空间就必须将数据集分割为指定大小,这样必然会导致部分文件的分割。目前 OLFS 采用的是遵循原则(1)、(2)的方法,数据集分配示意图如图 3 所示。从文件树的最底层开始遍历,保证文件的关联性最强,到了指定大小之后即划分为一个数据集,使光盘空间利用率最高。在这种分割方式下,每张光盘最多只会有两个文件被分割,也保证了较大的单张可读性。

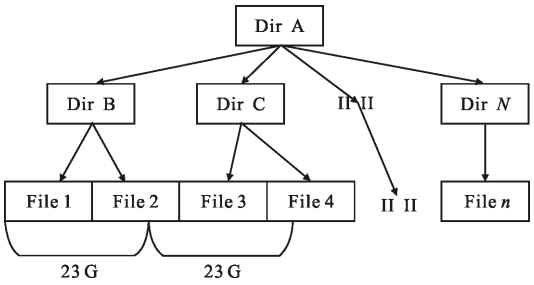

图 3 文件分割示意图

Fig.3 Schematic diagram of file partitioning

3.2 光盘管理

由于可记录光盘具有一次写特性，对于已经刻录过的光盘不能够再次刻录，也就是不能实现原地更新。当数据写入光盘时，需要取空白光盘用于刻录。因此 OLFS 对于所有物理光盘建立两个索引来保证准确的光盘查找。索引结构如图 4 所示，一个是未使用光盘的索引，它记录了每张光盘的物理位置和可用容量信息；另一个是已刻录光盘的索引，它记录的是光盘的唯一标记和物理位置。系统启动时会建立这两个索引队列，这两个索引队列会一直存放在内存中。当数据集刻录到某一张光盘时，从未刻录队列中取出该光盘标识加入已刻录光盘索引，然后将该光盘的唯一标记赋值为数据集的唯一标记，读取时即可通过对应关系取出指定光盘。

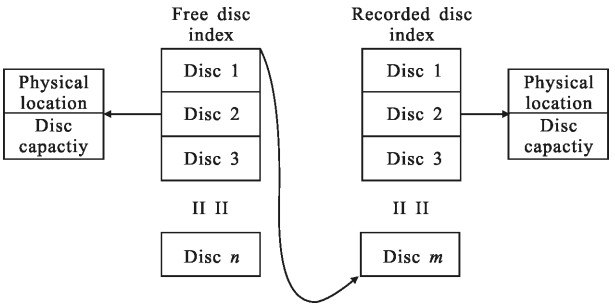

图 4 光盘索引结构

Fig.4 Disc index structure

光盘库中的文件会记录到一张或者多张物理光盘中，索引文件会记录每个文件及其相应的物理存储位置。物理存储位置包含光盘的物理编号和光盘内部地址，光盘内部地址包括该文件在光盘中的首地址和尾地址。由于是非实时刻录和一次写，因此文件在光盘中是连续存放的，从而只需要记录首尾地址。

此外，一个大文件可能跨越单个物理光盘的边界。例如一个大文件分布在 $M(M>=1)$ 个物理光盘上，则其索引项这需要记录 M 条记录，每条记录为光盘标记和文件在该光盘中的首地址和尾地址。

3.3 文件自动刻录

文件首先要传送到缓存盘上分割为数据集，针对每一个数据集发起一个光盘刻录任务，所有的刻录任务存放在一个先进先出队列中。由于光驱数量是有限的，队列中的等待任务仅在光驱空闲时被下发到刻录端。

不同的缓存策略添加队列的时机是不一样的。根据数据集分割原则，可以采用的缓存策略有两种：(1) 流式缓存刻录策略；(2) 全局缓存再刻录。

在流式缓存刻录策略中，文件传输到缓存盘的过程中是按底层目录开始遍历的顺序传输，在传输中即根据到达的次序实时分割成了光盘大小的数据集，每有一个数据集传输完成就将该数据集加入刻录队列。当刻录开始之后，刻录程序会从缓存盘上读取数据并刻录到光盘，此时前端还在保持数据的写入，这种方法要求缓存盘能够同时支持读和写的数据量，对于缓存盘读写吞吐率要求很高，考虑带磁盘在多个读写任务下性能较差，目前 OLFS 使用固态盘作为缓存盘。

在全局缓存再刻录策略中，采取预先缓存多张光盘数据，甚至全部缓存所有导入的整个数据集，然后事后按分割原则进行全局分割，分割完成之后再将多个数据集一次性加入刻录任务队列。这种方案需要缓存盘的容量较大，对同时读写的吞吐率要求不高，因此可以采用磁盘作为缓存盘。此外，在这种策略之下可以对于待刻录数据集进行全局离线分析，因此能够得到最优光盘数据分配策略。这部分工作将在未来集成到 OLFS 之中。

3.4 光盘调度策略

在光盘库中存在大量光盘和若干个光驱，由于光存储介质和读取驱动能够物理分离，在实际运行中，需要考虑光盘到光驱的优化调度，一方面能够实现光盘取盘、退盘的自动化的设计，另一方面实现调盘的延迟优化。

光盘存取动作需要通过机械臂完成，具体过程可以分为两类：从逻辑光驱 X 把光盘库放置到逻辑光盘槽 Y；从逻辑光盘槽 Y 中的光盘取到逻辑光驱 X (X、Y 为光盘、光驱逻辑号)。由于机械臂的个数小于

光驱数(目前光盘库使用两个机械臂),更为主要的是上述物理取盘退盘动作可能会共同使用统一物理通道(例如滑道),因此 OLFS 需要把物理调盘操作进行串行化。在实际实现中,调度程序会维护一个先进先出的动作队列,将需要执行的动作依次加入队列末尾,然后再由动作处理部分取出动作队列前面的任务进行处理,这样在密集读写时也不会造成机械手操作顺序的紊乱。

4 OLFS 具体实现

文件系统需要提供标准的文件访问接口来监控文件的读写操作,把对文件的读写转换为对光盘的读写操作,建立文件索引和光盘索引,保证数据能刻录到光盘,在需要读取时能读取到正确的数据。

4.1 OLFS 实现架构

实现结构如图 5 所示。OLFS 是基于用户态文件系统 FUSE[17]实现的,使用 FUSE 可以在文件系统处理函数中调用用户态的函数,使实现文件系统变得很方便。FUSE 在 VFS 层注册好 OLFS 的文件系统操作函数,当用户层进行文件操作时,VFS 会通过 FUSE 将操作转到用户态的 OLFS,然后 OLFS 再根据用户的具体请求来实现刻录或者读取数据操作。

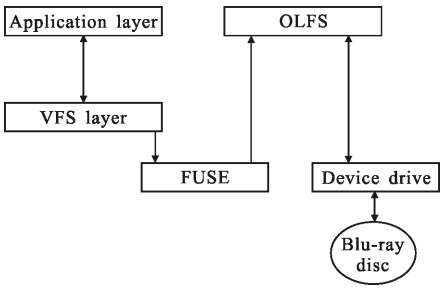

图 5 文件系统实现机制

Fig.5 Implementation of OLFS

4.2 实现流程

光盘库整机充当存储服务器,客户机可以通过 samba[18]协议访问光盘库,通过网络来进行数据的交互。samba 是基于 CIFS 实现的。

文件通过 samba 协议传输到服务器端,在将文件流分成数据集时,每个数据集内部的文件路径是与整个文件集局部类似的。文件流在传输时是以文件夹顺序和文件顺序传输的,所以流式缓存策略的实现只需要在传输时计数分割即可,文件的相关性自动得到保证。

5 实验结果与分析

5.1 测试环境

测试平台使用华中科技大学和广州紫晶光电有限公司共同研发的 ZL 12240S 光盘库硬件,该光盘库能够容纳 12240 片光盘,实现一体化光盘归档应用。ZL 12240S 的具体参数如表 1 所示,目前安装 24 个先锋专业光驱。ZL 12240S 提供两个万兆以太网接口,控制器中包含 5 块 240 GB 的 SSD 组成的 RAID0 阵列和 12 块 4 TB 磁盘组成的 RAID0 阵列作为缓存。

测试平台包含两台客户机。客户机配置为:i3 双核 CPU,500 GB 的 5 400 转磁盘,4 GB 内存,万兆网卡。客户机与光盘库机器同在一个万兆局域网内。

表 1 光盘库部分参数

Tab.1 Parameters of optical disc library

Parameter	Value
Max loaded disc count	12 240
Max storage capacity	1.2 PB
Supported drive count	1–24
Average disc load time	<5 s

测试中需要借助工具对拷贝速度进行采样,计算得出一系列性能参数,对实验结果进行分析。

5.2 光盘刻录性能

图 6 给出单个光驱的峰值刻录速度。刻录是从光盘内圈开始刻录,可以看到光盘峰值刻录吞吐量从内圈到外圈逐渐加速,这是由于目前光盘恒定角速度旋转,外圈线速度逐渐增加。基本上达到 60%光盘容量时,峰值性能达到近 50 MB/s。

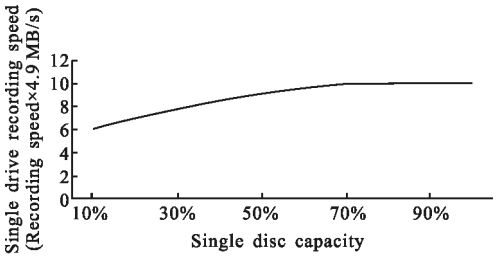

图 6 单个光驱实时峰值刻录速度

Fig.6 Peak burning speed of single optical drive

但是目前蓝光标准规定的最大刻录为 6 倍速，光驱固件中可以对光驱的刻录速度进行设置。当使用更高倍数刻录时，可能导致刻录失效率上升。因此出于对光盘数据可靠性的保证，目前光驱中光盘的刻录和读取速度都是恒定的，将刻录和读取速度设定为 6 倍速，约等于 24 MB/s。

目前的光驱虽然有 24 个，但是分为两组，一组用来刻录光盘，一组用来读取光盘。每组光驱的刻录读取聚合带宽为 280 MB/s。

5.3 机械臂取盘性能

数据存取在数据缓存区或者光驱缺失时，需要进行机械手取盘动作。在 ZL 12240S 机器中，光盘是存放在转笼中的，转笼有 6 列，每列有 5 层，每层有 17 个槽匣，每个槽匣里面放了 12 张光盘，总共有 12 240 张光盘。机械手执行一次移动距离最远的取盘操作的时间数量级是十秒级的。具体的取盘位置和所消耗时间如表 2 所示。

表 2 机械手取盘时间

Tab.2 Robot fetch time

Location	Load time/s
Row 5 line 1	73.2
Row 5 line 2	74.3
Row 5 line 3	74.4
Row 5 line 4	74.8
Row 5 line 5	74.7
Row 5 line 6	73.4

机械手取盘的时间=转笼转到指定位置时间+机械手上下移动抓盘时间+开光驱门放盘时间。第 5 层是在最下面的一层，取列 6 中光盘槽中的盘转笼需要转动一整圈，此时机械手需要移动的距离最远，转笼转动的角度最大，需要的时间也最大。

此外，光盘被装入光驱中，光驱需要识别光盘内容后才能开始读取数据，这会额外产生大约 60 s 的延迟。

5.4 数据集写入测试

根据文件系统的实现模型，文件是先通过远程网络存取方式写入到光盘库的缓存盘，然后再刻录到光盘，所以写入速度与网络和缓存盘的性能相关，由于元数据的写入影响，写入数据的速度和数据集的平均大小有关，数据集总大小为 20 GB。该节主要测试数据写入速度(均取平均值)。

文件写入性能测试是将总大小一定的数据集拷入远程指定目录中，每次拷入的数据集的平均大小分别为 1、2、4 MB，然后记录拷贝时间，最后计算出平均性能，具体性能如表 3 所示。

表 3 写入速度和刻录速度

Tab.3 Write speed and record speed

Average file size/kB	Write speed/MB·s^{-1}
4	87.3
16	109.5
64	117.8
128	118.5

文件写入时间中包括写入 4 kB 大小的索引文件时间和写入实际数据的时间。当平均文件大小为 4 kB 时，索引文件的写入时间占用较大，导致写入性能低于平均文件大小为 16 kB 的数据集的写入。在数据集平均大小远大于 4 kB 时，索引文件的写入时间影响变小，导致数据集的写入速度基本无变化。

5.5 文件读取测试

将刻录到光盘的文件读取出来，测量其读取速度。当读取单个文件时，FUSE 采用的策略是顺序读取，所以无法实现单文件双光驱并行读取。测试包括从光盘读取和从缓存盘和光盘混合读取。读取速度如表 4 所示。

表 4 读取速度

Tab.4 Read speed test

File size/kB	Read speed/MB·s^{-1}
1	72.3
20	81.5
45	86.2
150	84.4

在读取大小为 1 GB 的文件时，光驱处于初始工作状态，转速较慢，读取速度在一定时间后到达稳定值。在 45 GB 文件和 150 GB 文件读取测试中文件是混合存放的。45 GB 文件有 90%在光盘上，10%在缓存盘中；而 150 GB 文件则是有 75%在光盘上，其余部分在缓存盘中，这个延迟时间也包含调盘延迟时间，所以其读取的速度比起光驱读取速度要快。

5.6 使用 FUSE 文件系统和直接使用 samba 的写性能对比

在使用 FUSE 进行文件系统的 mount 时，可以通过调节参数来实现控制单次 write 操作写入的数据块大小，FUSE 默认的是内存中每有 4 kB 缓冲数据就写入到缓存盘，可以通过修改 mount 参数将 4 kB 改为 128 kB。缓存盘的写入和内存操作相比是很耗时间的，所以该参数的设置与否对写入速度有较大的影响。表 5 为使用 FUSE 以及 samba 的一些写入性能测试。

表 5 samba & FUSE 写入性能测试

Tab.5 samba & FUSE write performance

samba/MB·s^{-1}	samba+FUSE/MB·s^{-1}	samba+FUSE configured/MB·s^{-1}
320.6	87.4	133.5

通过 samba 网络共享直接拷入数据的性能和 samba 与 FUSE 修改参数的性能差别较小，说明用户层文件系统 FUSE 带来的性能损失处于可以接受的范围，但依然有优化的空间。

6 结 论

文中利用光盘的长期低成本保存的特性，提出了一种基于光盘的超大容量光盘库系统，解决了单光盘容量和性能的限制，实现海量光盘虚拟化和性能优化，解决海量光盘及其数据的自动化管理问题，而且为光盘库提供了通用的文件系统访问接口，读取写入测试中，光盘库文件系统能自动化进行取盘调度来实现文件的刻录和读取。

未来需要做的工作有：(1) 增加光盘库的缓存盘容量，实现冷温热数据分层机制；(2) 分析在用户态文件系统 FUSE 性能消耗原因，减少性能消耗。

参考文献：

[1] Wang Bin, Pan Xinhua, Tan Ke. Application of information lifecycle management in medical data preservation [J]. ***Value Engineering***, 2012, 31(12): 157-158. (in Chinese)

[2] Kumar S, McCaffrey T R. Engineering economics at a hard disk drive manufacturer[J]. ***Technovation***, 2003, 23(2): 749-755.

[3] Boyd S, Horvath A, Dornfeld D. Life-Cycle Assessment of NAND Flash Memory [J]. ***Semiconductor Manufacturing IEEE Transactions on***, 2011, 24(1): 117-124.

[4] Okazaki Y, Hara K, Kawashima T, et al. Estimating the archival life of metal particulate tape [J]. ***Magnetics IEEE Transactions on***, 1992, 28(5): 2365-2367.

[5] Feeney R, Science and Technology Council, Academy of Motion Picture Arts and Sciences. The digital dilemma: strategic issues in archiving and accessing digital motion picture materials [R]. US: Academy Imprints, 2008.

[6] Nikoobakht B, El-Sayed M A. Preparation and growth mechanism of gold nanorods (Nrs) using seed-mediated growth method [J]. ***Chemistry of Materials***, 2003, 15(10): 1957-1962.

[7] Gu M, Li X. The road to multi-dimensional bit-by-bit optical data storage [J]. ***Optics and Photonics News***, 2010, 21(7): 28-33.

[8] Mikami T, Mochizuki H, Sasaki T, et al. Twenty-layer optical disc fabricated by web coating and lamination [J]. ***Japanese Journal of Applied Physics***, 2013, 52(9S2): 09LC01.

[9] Zijlstra P, Chon J W M, Gu M. Five-dimensional optical recording mediated by surface plasmons in gold nanorods[J]. ***Nature***, 2009, 459(7245): 410-413.

[10] Iraci, Joe. Disaster recovery of modern information carriers: compact discs, magnetic tapes, and magnetic disks [J]. ***Canadian Conservation Institute Technical Bulletin***, 2002, 25: 1-15.

[11] Balakrishnan S, Black R, Donnelly A, et al. Pelican: a building block for exascale cold data storage [C]// Proceedings of the 11th USENIX Conference on Operating Systems Design and Implementation, 2014: 351-365.

[12] Lijding M E, Jansen P, Mullender S. Real-time scheduling of a tertiary-storage juke-box [J]. ***Stw Technology Foundation***, 2001: 135-140.

[13] Kamada J, Kokachi Y. Security enhancements for UDF (Universal Disk Format)[J]. ***Ipsj Sig Notes***, 2001, 53: 19-24.

[14] Ueno M, Murata S, Iwatsu S, et al. A disaster-tolerant widely distributed file system using optical disk libraries [J]. ***Japanese Journal of Applied Physics***, 1999, 38(3S): 1795.

[15] Thompson C. Optical disc system for long term archiving of multi-media content [C]//Systems, Signals and Image Processing (IWSSIP) International Conference on, IEEE, 2014: 11-14.

[16] Gu M, Li X, Cao Y. Optical storage arrays: a perspective for future big data storage [J]. ***Light: Science & Applications***, 2014, 3(5): e177.

[17] Rajgarhia A, Gehani A. Performance and extension of user space file systems [C]//ACM Symposium on Applied Computing. 2010: 206-213.

[18] Hertel C R. Implementing CIFS: The Common Internet File System[M]. New York: Prentice Hall, 2004.