MT²RAID:一种高可靠大规模磁盘阵列结构

王志坤 冯 丹

(华中科技大学计算机科学与技术学院 武汉 430074) (武汉光电国家实验室信息存储部 武汉 430074)

摘 要 传统的磁盘阵列一般采用集中式控制结构,其连接的底层磁盘数受系统总线的制约,容易出现性能瓶颈,且不能容两个以上磁盘出错。从模块化系统的组织方法出发,提出一种采用标准模块化存储单元组成的通过胖树结构互连的大规模磁盘阵列结构MT²RAID,分别就其各种数据分布的性能和可靠性进行了分析和讨论。原型系统测试结果表明,相比集中式磁盘阵列结构,MT²RAID也具有较高的性能。

关键词 磁盘阵列,扩展性,可靠性,模块化

MT²RAID:A High Reliable Architecture for Large Scale Disk Arrays

WANG Zhi-kun FENG Dan

(School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

(Division of Data Storage System, Wuhan National Laboratory for Optoelectronics, Wuhan 430074, China)

Abstract Traditional disk arrays have centralized control architecture. The number of connected disks is constrained by the system bus. Centralized RAID architecture easily generates performance bottleneck and can not tolerate more than two disks faults. This paper proposed MT²RAID, a modular tree-connected multi-tier RAID architecture for large scale disk arrays. MT²RAID is built from a collection of commodity components. Storage units are connected through fat-tree based interconnection channels. The performance and reliability of different MT²RAID level were also analyzed and discussed. Prototype experimental results show that MT²RAID also has performance advantages compared with centralized RAID architecture.

Keywords RAID, Scalability, Reliability, Modular

1 引言

随着磁盘容量的增加和价格的不断下降,磁盘已成为大规模存储系统中首选的存储设备。然而,由于单个磁盘本身性能和可靠性的限制,磁盘存储系统多是以磁盘阵列(Redundant Array of Independent Disk,RAID)[1]的形式出现的。RAID具有可靠性高、并行性好等优点,从提出之日起,就备受学术界和工业界的关注。然而,传统的RAID往往采用集中式控制结构,其连接的底层设备通道数受系统总线的制约,较大规模的阵列还会受到内部通道总线带宽和可靠性等问题的影响,且现有RAID级别不能容两个以上磁盘出错。高端企业级RAID往往采用专用、定制的硬件组件,并通过高度冗余的部件和通道设计来提高阵列的性能和可靠性,如EMC Symmetrix DMX[2],IBM ESS[3]和HDS USP[4]等,但这些系统都属于传统的单体式架构,价格昂贵,且只能纵向扩展(Scale-up),而不能横向扩展(Scale-out)。

随着计算机硬件和系统组织方法的发展,高性能存储系统已经从由专用硬件构成的集中式存储结构,演化到采用模块化存储节点组成的大规模存储系统,该方式具有更好的扩展性和灵活性。现在已经有了一些分布式RAID的研究[5-7],这些方法主要是在存储网络中的各存储节点之间按照软RAID或文件RAID方式进行数据分布,需要在客户端完成数据的分块与校验计算,对系统的性能影响较大。

存储系统从传统的直连存储(Direct Attached Storage,DAS)过渡到网络附加存储(Network Attached Storage,NAS)和存储区域网(Storage Area Network,SAN)等网络存储系统,都没有过多地关注拓扑结构。现有网络存储系统中一般采用总线或星型网络拓扑结构,虽然SAN可以互连很多存储设备,但其主要是为了解决数据共享和扩展性问题,因此不能提供很高的带宽。

网络互联拓扑结构是高性能并行计算领域的重要研究内容之一。通过高效地互联网络,可以大大提升高性能计算的性能。随着网络存储中高速互联网络的出现,如Infiniband、光纤通道和万兆以太网等,采用标准存储单元构建大规模、可扩展的磁盘阵列时,需要重新审视存储单元间的组织关系和互联拓扑结构,以增强系统性能和可靠性。

2 MT²RAID体系结构

在分析、比较了高性能计算领域常用拓扑结构的基础上,

到稿日期:2009-12-08 返修日期:2010-03-04 本文受国家863计划项目(2009AA01A402),教育部创新团队(IRT0725),国家自然科学基金(60503059)资助。

王志坤(1980—),男,博士生,主要研究方向为磁盘阵列、网络存储系统等,E-mail:zkwang@smail.hust.edu.cn;冯 丹(1970—),女,教授,博士生导师,主要研究方向为计算机体系结构、并行存储系统和大规模网络存储系统等。

并考虑到存储系统 I/O 访问自身所具有的局部性和周期性的特点，提出了一种由标准模块化存储单元构成的并采用胖树结构互连的分层磁盘阵列组织结构 MT²RAID（Moduler Tree-connected Multi-Tier RAID）。它通过不同的通道及接口技术构成了一个模块化、可扩展、高可靠性的并行存储系统，其体系结构如图 1 所示。其中每个模块单元都包括 CPU、内存、磁盘、互联接口等部件。这些模块化的存储单元既可以是同构硬件，也可以是异构存储部件，并可以兼容不同接口、速度、性能的磁盘等。

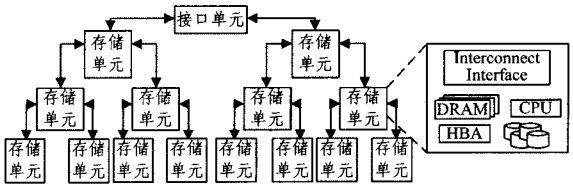


图 1　MT²RAID 体系结构

MT²RAID 是一种模块化分层存储结构，根据设定的树结构分支数，每层由个数不等的存储单元构成，各层存储单元之间由带宽不同的通道进行连接。由于 MT²RAID 结构中下层存储单元个数较多，为了充分利用各存储节点的并行性，采用胖树结构，以避免数据访问拥塞。胖树[8]是一种常用的拓扑结构，它用在多处理器系统中进行处理器之间的互联，比较有名的系统有 CM-5[9]等。由于其高带宽和灵活的寻址方式，最近它又被引入到高速互联网络设计中。胖树和传统树结构的主要区别在于胖树更加像真实的树结构。在一个传统的树结构中，无论树结构有多深，其链路带宽都是固定的，这将引起根节点的拥塞问题。而在胖树结构中，链路带宽从叶子到树根逐渐增大。树结构的另外一个好处是，其分支出现故障不会影响树的其他部分，而且基于树结构的存储系统可以逐步进行构建，通过增加树结构的深度或者分叉数目，可以很容易地扩展 MT²RAID 结构的规模。

在 MT²RAID 结构中，总共有两种类型的构建单元：接口单元和存储单元。其中接口单元位于树结构的根节点，对外提供访问通道，如 Infiniband、FC 或 iSCSI 通道等，负责接收命令请求并分发和返回数据，负责存储空间的管理与分配。为了避免接口单元节点成为单点失效点和性能访问瓶颈，可以在树根配置多个接口单元。而存储单元则提供磁盘存储空间并存放数据，其中每个存储单元中的磁盘既可以按照 RAID 方式进行组织，也可以仅提供磁盘组的形式。

与传统集中式控制结构的大规模磁盘阵列相比，所提出的 MT²RAID 大规模磁盘阵列结构具有如下优点：

1）模块化和成比例扩展。在 MT²RAID 结构中，每个构建单元都是一个独立的功能模块，包含自己的 CPU、内存、磁盘以及互联接口等。在增加或者删减存储模块的时候，可以成比例地增加或者减少处理能力、缓存、存储容量和带宽，以维持资源最优。

2）性价比高和升级方便。MT²RAID 中每个存储单元都是采用标准存储硬件组成，没有使用任何定制的硬件部件，可以很方便地利用当前最新的硬件部件进行扩展升级。这样既可以兼容已有存储单元，保护既有投资，也可使得整个系统性能随着硬件的发展而不断增强。

3）全局缓存管理。每个组成单元都是一个独立的存储子系统，拥有自己独立的缓存。为了提高系统性能，通过全局缓存管理方法，可以充分利用其它存储单元中的空闲缓存资源，提高缓存命中率，减少磁盘 I/O。并且系统全局缓存大小和带宽随着存储单元个数的增加而线性增长。

4）模块处理能力。MT²RAID 中每个存储单元都具有自己独立的 CPU 部件。随着硬件的发展，尤其是多核处理器的出现，每个存储单元的处理能力显著增强。MT²RAID 除了提供存储服务外，还可以胜任计算密集型任务，如智能缓存预取、数据预处理、数据自动分级、数据加密等。并且随着系统规模的增加，MT²RAID 结构整体处理能力随之增强，特别适合于存储、计算混合型的应用环境。

3　MT²RAID 数据分布

3.1　MT²RAID 3 种数据分布方式

由于 MT²RAID 中每个存储单元只能提供有限的性能和可靠性，为了提高数据的访问性能和可靠性，在存储单元内部和存储单元之间采用双层 RAID 方式来提高数据的访问性能和可靠性。根据不同的数据分布方式和可靠性，设计并实现了 3 种不同的数据分布方式。为了便于描述，这里以二叉树结构来对 3 种 MT²RAID 结构进行说明，其中每个存储单元挂 4 个磁盘，存储单元内部使用 RAID0 的方式。

第一种数据分块方式是在各存储单元间采用数据分块方式 MT²RAID-S，其数据分布如图 2 所示。数据均匀分布到每个存储单元上，存储单元内部可以采用任意的 RAID 级别。MT²RAID-S 方式可以最大限度地利用底层 I/O 通道的并行性能，获得最大数据传输速率，但其缺点是不能对存储单元级别的故障提供保护。

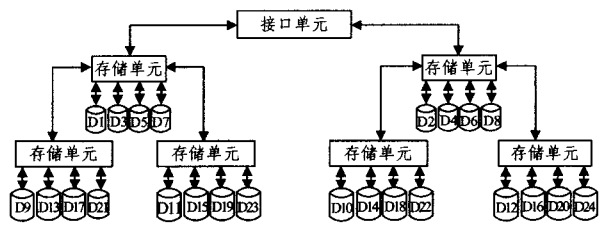


图 2　MT²RAID-S 数据分布

第二种数据分布方式采用镜像方式 MT²RAID-M，其数据分布如图 3 所示。以图中所用的二叉树结构为例，不同分支之间存储单元中的数据互为镜像。该结构具有很高的数据可用性，缺点是会损失一半的存储空间。在正常工作状态下，MT²RAID-M 的数据读操作只要从两个镜像存储单元中选择当前负载最轻的存储单元进行服务即可，能获得较好的系统读性能。但写操作需要等待数据分别写到镜像的存储单元中，才能向前端应用报告完成，从而具有一定的延迟。

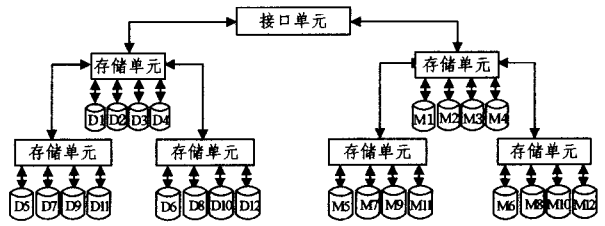


图 3　MT²RAID-M 数据分布

第三种数据分块方式在各存储单元之间采用奇偶校验方

式 MT²RAID-P，其数据分布如图 4 所示。其在每一层之中将数据块在存储单元之间按照块交叉校验方式进行存放。该方式具有比 MT²RAID-M 更高的空间利用率。读请求可以充分利用各存储单元之间的并行性，具有较高的读性能。但对于小写请求来说，其要在多个存储单元之间进行数据的“读-写-写”操作，对写性能影响很大。

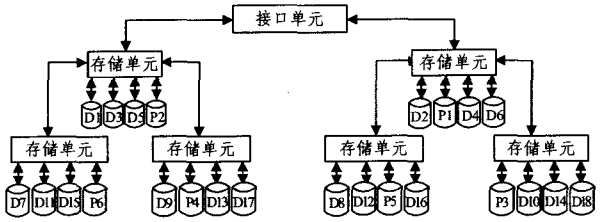


图 4　MT²RAID-P 数据分布

3.2　MT²RAID 性能和空间利用率分析

MT²RAID 每个存储单元内部可以使用 0～6 共 7 种 RAID 级别。选取其中最常用 RAID0，RAID1，RAID5，RAID6 4 种级别，与所提出的 3 种 MT²RAID 方式一起可以形成 12 种数据分布组合。基于基本 RAID 级别相对性能分析结果[10]，可以进一步分析所提出的 MT²RAID 各种数据分布的性能和空间利用率。其中，假定每个存储单元都采用同构节点组成，整个系统有 m 个存储单元，且每个单元中含有 n 个磁盘，所有存储单元内部 RAID 级别相同。由 $m\times n$ 个磁盘可构成 12 种 MT²RAID 数据分布方式。相对于 MT²-RAID-S0的性能和空间利用率如表 1 所列。

表 1　不同级别 MT²RAID 的相对性能和空间利用率

级别	小读	小写	大读	大写	空间利用率
MT²RAID-S0	1	1	1	1	1
MT²RAID-S1	1	1/2	1	1/2	1/2
MT²RAID-S5	1	Max(1/n,1/4)	1	(n−1)/n	(n−1)/n
MT²RAID-S6	1	Max(1/n,1/6)	1	(n−2)/n	(n−2)/n
MT²RAID-M0	1	1/2	1	1/2	1/2
MT²RAID-M1	1	1/4	1	1/4	1/4
MT²RAID-M5	1	1/2 * Max(1/n,1/4)	1	(n−1)/2n	(n−1)/2n
MT²RAID-M6	1	1/2 * Max(1/n,1/6)	1	(n−2)/2n	(n−2)/2n
MT²RAID-P0	1	Max(1/m,1/4)	1	(m−1)/m	(m−1)/m
MT²RAID-P1	1	1/2 * Max(1/m,1/4)	1	(m−1)/2m	(m−1)/2m
MT²RAID-P5	1	Max(1/m,1/4) * Max(1/n,1/4)	1	(m−1)(n−1)/mn	(m−1)(n−1)/mn
MT²RAID-P6	1	Max(1/m,1/4) * Max(1/n,1/6)	1	(m−1)(n−2)/mn	(m−1)(n−2)/mn

从表 1 中可以看出，不同 MT²RAID 级别的读性能不受磁盘和存储单元个数以及请求大小的影响，其性能都相同。而不同级别的小写相对性能在 $m>4$ 时只与存储单元内磁盘个数有关，这里取 $m=7$，即建立一个由 7 个节点组成的 3 层 2 叉 MT²RAID 结构，其结果如图 5 所示。从图中可以看出，MT²RAID-S0 性能最高，MT²RAID-S1 和 MT²RAID-M0 为它的 1/2，MT²RAID-S5，MT²RAID-M1 和 MT²RAID-P0 为它的 1/4，MT²RAID-M5 和 MT²RAID-P1 为它的 1/8，MT²-RAID-P5为它的 1/16，MT²RAID-S6 为 1/6，MT²RAID-M6 为 1/12，MT²RAID-P6 为 1/24。

从表 1 中可以看出，不同 MT²RAID 级别的大写请求相对性能与空间利用率的值相同，用图 6 来表示这两项的值。从图中可以看出，MT²RAID-S0 的性能和空间利用率最高，MT²RAID-P0 为它的 6/7，MT²RAID-S1 和 MT²RAID-M0 为它的 1/2，MT²RAID-P1 为它的 3/7，MT²RAID-M1 为它的 1/4。MT²RAID-S5，MT²RAID-S6，MT²RAID-M5，MT²RAID-M6，MT²RAID-P5，MT²RAID-P6 等级别随着存储单元中磁盘个数的增加，大写性能和空间利用率逐渐增大。

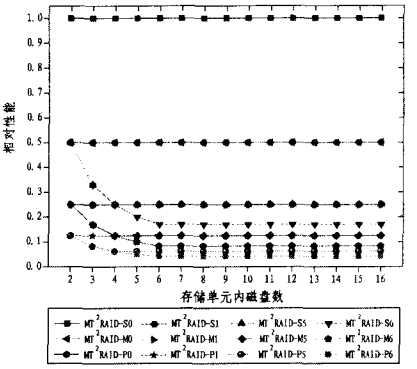


图 5　不同 MT²RAID 级别小写相对性能

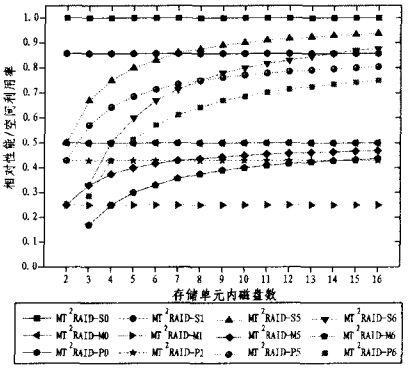


图 6　不同 MT²RAID 级别大写相对性能和空间利用率

3.3　MT²RAID 可靠性分析

存储系统一般采用平均数据丢失时间(Mean Time To Data Loss，MTTDL)来衡量其可靠性。该值可以采用数学公式进行分析[1,10]，也可采用马尔科夫链模型进行建模计算[11,12]。对于所提出的 MT²RAID 结构，其 3 种数据分布方式采用双层 RAID 方法来保护数据的可靠性，可采用嵌套马尔科夫链模型来分析其可靠性。其中各参数的含义如表 2 所列。为了简化分析，这里假设 MT²RAID 采用同构硬件单元组成，且每个存储单元中磁盘的个数相等。

表 2　可靠性模型参数及含义

参数	含义
n	存储单元中磁盘的个数
m	存储单元的个数
λ_d	磁盘失效率 $1/MTTF_d$
μ_d	磁盘修复率 $1/MTTR_d$
λ_A	阵列失效率 $1/MTTF_A$
μ_A	阵列修复率 $1/MTTR_A$

根据文献[11,12]中关于磁盘阵列的马尔科夫可靠性建模方法，考虑到磁盘平均修复时间远远小于磁盘平均失效时间，即 $\mu_d\gg\lambda_d$，可以首先得到各基本 RAID 级别 *MTTDL* 如下：

$$MTTDL_{RAID0}=\frac{1}{n\lambda_d}$$

$$MTTDL_{RAID1}\approx\frac{\mu_d}{n\lambda_d^2}$$

$$MTTDL_{RAID5}=\frac{(2n-1)\lambda_d+\mu_d}{n(n-1)\lambda_d^2}\approx\frac{\mu_d}{n(n-1)\lambda_d^2}$$

$$MTTDL_{RAID6}=\frac{(3n^2-6n+2)\lambda_d^2+(3n-2)\lambda_d\mu_d+\mu_d^2}{n(n-1)(n-2)\lambda_d^3}$$

$$\approx\frac{\mu_d^2}{n(n-1)(n-2)\lambda_d^3}$$

对于 MT²RAID-S0 而言，由于其没有采取任何数据冗余方式，任何磁盘故障都将导致数据丢失。对于 m 个存储单元，每个存储单元内部有 n 个磁盘，其可靠性是单个磁盘的 $\frac{1}{mn}$，因此 $MTTDL_{MT^2RAID\text{-}S0}=\frac{MTTF_{Disk}}{mn}=\frac{1}{mn\lambda_d}$

对于 MT²RAID-S1，将 $MTTDL_{RAID1}$ 带入 $MTTDL_{RAID0}$，即可得到

$$MTTDL_{MT^2RAID\text{-}S1}=\frac{MTTDL_{RAID1}}{m}=\frac{\mu_d}{mn\lambda_d^2}$$

依次类推，MT²RAID 各种数据分布的 $MTTDL$ 值如下：

$$MTTDL_{MT^2RAID\text{-}S5}=\frac{1}{m\lambda_{A(RAID5)}}\approx\frac{\mu_d}{mn(n-1)\lambda_d^2}$$

$$MTTDL_{MT^2RAID\text{-}S6}=\frac{1}{m\lambda_{A(RAID6)}}\approx\frac{\mu_d^2}{mn(n-1)(n-2)\lambda_d^3}$$

$$MTTDL_{MT^2RAID\text{-}M0}\approx\frac{\mu_A}{mn^2\lambda_d^2}$$

$$MTTDL_{MT^2RAID\text{-}M1}\approx\frac{\mu_A\mu_d^2}{mn^2\lambda_d^4}$$

$$MTTDL_{MT^2RAID\text{-}M5}\approx\frac{\mu_A\mu_d^2}{mn^2(n-1)^2\lambda_d^4}$$

$$MTTDL_{MT^2RAID\text{-}M6}\approx\frac{\mu_A\mu_d^4}{mn^2(n-1)^2(n-2)^2\lambda_d^6}$$

$$MTTDL_{MT^2RAID\text{-}P0}\approx\frac{\mu_A}{m(m-1)n^2\lambda_d^2}$$

$$MTTDL_{MT^2RAID\text{-}P1}\approx\frac{\mu_A\mu_d^2}{m(m-1)n^2\lambda_d^4}$$

$$MTTDL_{MT^2RAID\text{-}P5}\approx\frac{\mu_A\mu_d^2}{m(m-1)n^2(n-1)^2\lambda_d^4}$$

$$MTTDL_{MT^2RAID\text{-}P6}\approx\frac{\mu_A\mu_d^4}{m(m-1)n^2(n-1)^2(n-2)^2\lambda_d^6}$$

假设在 RAID 中磁盘发生故障的概率是相互独立的，以实验中使用的希捷 ST3300831AS 300GB SATA 磁盘为例，其产品手册中标称的每年失效率（Annualized Failure Rate，AFR）为 0.34%，平均无故障时间大约为 250 万 h[13]。但 Gibson 等人对大约 10 万块各种接口的硬盘进行研究后的结论是：硬盘产品的实际年失效率 AFR 一般在 2%～4%之间，最高甚至可达 13%[14]。因此这里取磁盘的 $MTTF_d$ 值为 20 万 h，磁盘的平均修复时间 $MTTR_d$ 为 10h。存储单元控制器的 $MTTF_A$ 值为 40 万 h，其平均修复时间 $MTTR_A$ 为 20h。

分别分析了 MT²RAID 3 种数据分布可靠性与存储单元内部磁盘个数以及存储单元数的关系。MT²RAID 3 种数据分布可靠性随存储单元内部磁盘个数的变化如图 7 所示，MT²RAID 各级别可靠性随存储单元数的变化如图 8 所示。从两图可以看出，随着存储单元中磁盘个数的增加或者存储单元个数的增加，各 MT²RAID 数据布局的可靠性逐渐降低。MT²RAID-M6 和 MT²RAID-P6 由于最多可以容系统中任意 5 个磁盘故障，因而具有最高的可靠性。MT²RAID-M1，MT²RAID-P1，MT²RAID-M5 和 MT²RAID-P5 4 种方式，最多可以容系统中任意 3 个磁盘故障，可靠性也比较高。而 MT²RAID-S6 最多可以容系统中 2 个磁盘故障，具有较好的可靠性。MT²RAID-S1，MT²RAID-S5，MT²RAID-M0 和 MT²RAID-P0 4 种方式只能容单盘故障，具有一般的可靠性，而 MT²RAID-S0 则因为没有任何数据冗余，可靠性最差。

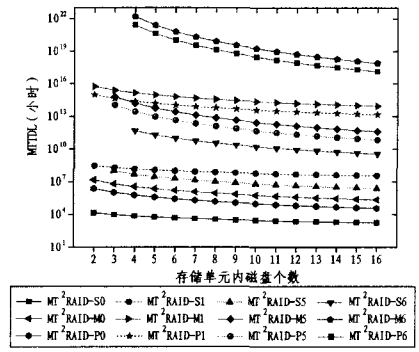


图 7　MT²RAID 可靠性与存储单元中磁盘个数的关系（m=7）

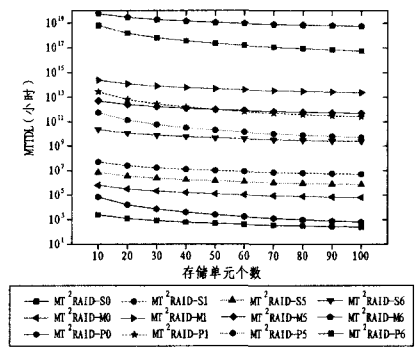


图 8　MT²RAID 可靠性与存储单元个数的关系（n=8）

4　性能测试

4.1　实验环境

基于前述设计思想和实现方法，建立了一个由 7 个模块化存储单元组成的、深度为 3、分叉为 2 的 MT²RAID 原型系统，其拓扑结构如图 9 所示。其中每个存储单元都是由标准硬件构成，其硬件配置如表 3 所列。每个存储单元中运行 Linux RedHat FC4 操作系统，内核版本 2.6.11。为了匹配胖树结构层间带宽的要求，接口单元和中间存储单元之间通过两对 2Gbps 的光纤链路点对点连接，中间存储单元和叶子存储单元之间通过两个独立的 1Gbps 以太网交换机相连，并使用 UNH iSCSI[15] 软件进行数据传递。

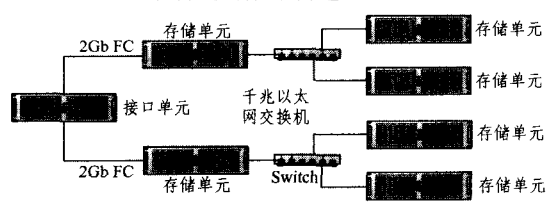


图 9　MT²RAID 原型系统拓扑结构图

表 3　存储单元硬件配置

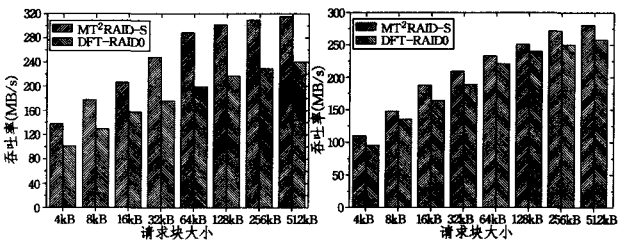
部件	描述
中央处理器	Intel Pentium4 Xeon 3GHz
内存	512MB DDR
主板	Supermicro X6DH8-XB
SATA 主机适配器	HighPoint RocketRAID 2240
SATA 磁盘	希捷 300GB(ST3300831AS)×8

FC 主机适配器	Qlogic Qla2310F 2Gb 光纤卡
网卡	板载 Broadcom BCM5721 千兆网卡

4.2 测试结果及分析

为了反映 MT²RAID 结构性能上的优势，采用流行的基准测试程序 Iometer[16] 对 MT²RAID-S0 与 DFT-RAID0[17] 双通道光纤磁盘阵列进行了对比测试。Iometer 具有一个合成负载发生器，可以指定负载的访问模式、读写操作比例、并发请求数和请求块大小等。为了对系统性能进行压力测试，采用两种极端访问负载：100%的顺序读和 100%的顺序写，块请求大小从 4kB 到 512kB，每次增加 1 倍。

图 10 所示的对比测试结果表明，MT²RAID-S0 的最大持续读性能有 315MB/s，而双通道的 DFT-RAID0 最大持续读性能只有 240MB/s。这正是由于 MT²RAID-S0 将读请求分散到各个存储单元上，充分利用了存储单元之间的并行能力。而 DFT-RAID0 虽然有两个对外的光纤通道，但是所有数据请求都将竞争内部的 CPU、内存和磁盘等资源，其服务请求队列较长，性能有限。对于写请求而言，由于两种阵列结构均采用“写回”策略，DFT-RAID0 作为商业磁盘阵列，具备较大的写缓存设计，写性能与 MT²RAID-S0 比较接近。



(a) 顺序读性能　　(b) 顺序写性能

图 10　性能对比测试结果

结束语　构建大规模磁盘阵列时，扩展性和可靠性成为首要考虑的目标。与现有高端磁盘阵列采用的专用定制部件不同，本文提出一种由标准存储模块构成、采用胖树结构互连的大规模磁盘阵列结构 MT²RAID。MT²RAID 通过存储单元内部和存储单元之间的双层数据冗余方法来提高 MT²RAID 结构中数据的可靠性。理论分析表明，各种不同 MT²RAID 级别在性能、存储空间利用率和可靠性上各有优势。其中 MT²RAID-M6 和 MT²RAID-P6 级别可以容任意 5 个磁盘故障，具有很高的可靠性。原型系统测试结果表明，与传统的集中式控制结构磁盘阵列相比，采用标准存储单元组成的 MT²RAID 结构也具有较高的系统性能。

参 考 文 献

[1] Patterson D, Gibson G, Katz R. A Case for Redundant Arrays of Inexpensive Disks (RAID) [C]// Proceedings of the 1988 ACM SIGMOD International Conference on Management of Data. Chicago, Illinois, USA, June 1988: 109-116

[2] EMC Corp. EMC Symmetrix DMX Family [EB/OL]. http://www.emc.com/ products/family/symmetrix-family.htm

[3] Hartung M. IBM TotalStorage Enterprise Storage Server: A Designer's View[J]. IBM Systems Journal, 2003, 42(2): 383-396

[4] Hitachi Data Systems. Universial Storage Platform [EB/OL]. http://www.hds.com/products/storage-systems/universal-storage-platform-vm.html

[5] Stonebraker M, Schloss G A. Distributed RAID — A New Multiple Copy Algorithm [C]// Proceedings of the 6th International Conference on Data Engineering (ICDE'90). Feb. 1990: 430-437

[6] Edward K, Chandramohan A. Petal: Distributed Virtual Disks [C]// Proceedings of the 7th International Conference on Architectural Support for Programming Languages and Operating Systems. Cambridge, MA, Oct. 1996: 84-92

[7] Long D D E, Montague B R, Cabrera L F. Swift/RAID: A Distributed RAID System [J]. Computing Systems, 1994, 7(3): 333-359

[8] Leiserson C E. Fat-trees: Universal Networks for Hardware-Efficient Supercomputing[J]. IEEE Transactions on Computers, 1985, 34(10): 892-901

[9] Leiserson C E, Abuhamdeh Z S, Douglas D C, et al. The Network Architecture of The Connection Machine CM-5 [C]// Proceedings of the 4th ACM Symposium on Parallel Algorithms and Architectures. San Diego, CA, USA, 1992: 272-285

[10] Chen P, Lee E, Gibson G A, et al. RAID: High-performance, Reliable Secondary Storage[J]. ACM Computing Surveys, 1994, 26(2): 145-185

[11] Baek S H, Kim B W, Jeung E, et al. Reliability and Performance of Hierarchical RAID with Multiple Controllers [C]// Proceedings of the 20th Annual ACM Symposium on Principles of Distributed Computing (PODC'01). Newport, RI, USA, Aug. 2001: 246-254

[12] Rao K, Hafner J L, Golding R A. Reliability for Networked Storage Nodes [C]// Proceedings of the International Conference on Dependable Systems and Networks (DSN'06). Philadelphia, PA, USA, June 2006: 237-248

[13] Seagate Technology LLC. Barracuda 7200.8 Serial ATA Product Manual[S]. Publication number: 100325576, Rev. F, Aug. 2007

[14] Schroeder B, Gibson G A. Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You? [C]// Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST'07). San Jose, CA, USA, Feb. 2007: 1-16

[15] UNH-iSCSI Initiator and Target for Linux [EB/OL]. http://unh-iscsi.sourceforge.net

[16] Iometer Benchmark (version 2006.07.27) [EB/OL]. http://sourceforge.net/projects/Iometer/

[17] DFT ES1600 RAID[EB/OL]. http://www.dft.com.cn/product/ productdetail.aspx? id=596