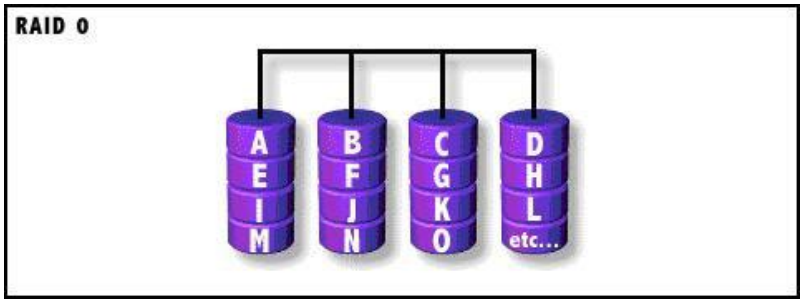
一种大规模存储系统能耗与性能的优化方法

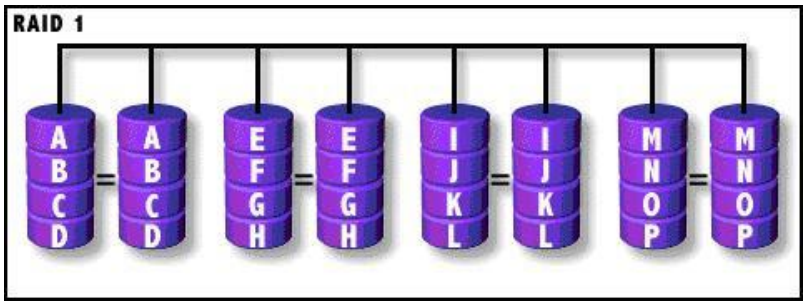
唐传慧　　2017202110055

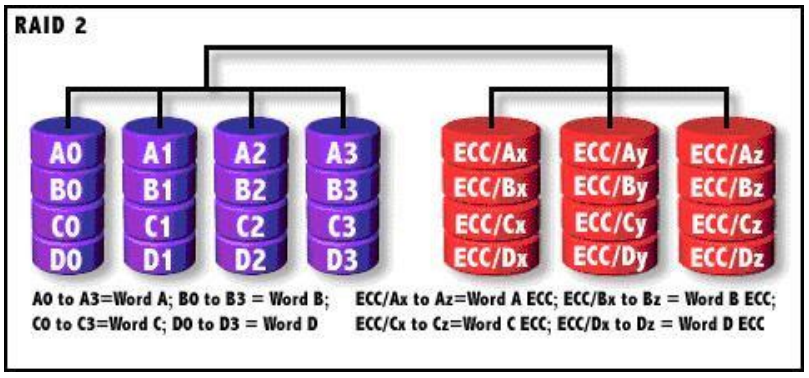
摘要　计算、物联网、移动互联、社交媒体等新兴信息技术和应用模式的快速发展,促使全球数据量急剧增加,推动人类社会迈入大数据时代.大数据应用背景下,用户对存储空间的需求越来越大,存储的数据类型也日益复杂化,这些现象的出现对传统的数据存储服务提出了极大的挑战.然而大数据的应用场景对大规模存储系统提出了进一步的要求,磁盘及其冷却系统是现代大规模存储系统中能耗的主体.传统的大规模存储系统的构建方式在提高服务规模及访问性能的同时也带来了巨大的能耗问题.数据中心的高能耗问题不仅仅是因为数据规模的急剧增大,系统资源的低效组织和无序管理所导致的能源利用率低下也是一个重要原因.此外在大数据环境下,数据价值的时效性往往表现为数据中所蕴含的知识价值随着时间的流失而衰减.大规模存储系统作为大数据的主要存储平台,必须满足大数据处理对数据存储平台的访问性能需求,避免成为大数据处理的性能瓶颈.大规模存储系统的能耗优化往往会对系统的访问性能产生一定的负面影响.已有的节能研究主要面向以随机访问为主的存储系统,并试图适应各种工作负载,导致难以取得进一步突破.因此,提出一种适于顺序数据访问的节能磁盘阵列 S-RAID5,采用局部并行策略:阵列中的存储区被分成若干组,组内采用并行访问模式,分组有利于调度部分磁盘运行而其余磁盘待机,组内并行用以提供性能保证.在模拟实验中,在满足性能需求、单盘容错的条件下,2 4 小时功耗测量实验表明:S-RAID5 的功耗为节能磁盘阵列 HiBernator 功耗的 59％, e RAID 功耗的 23％,PARAID、GRAID 功耗的 21％左右.

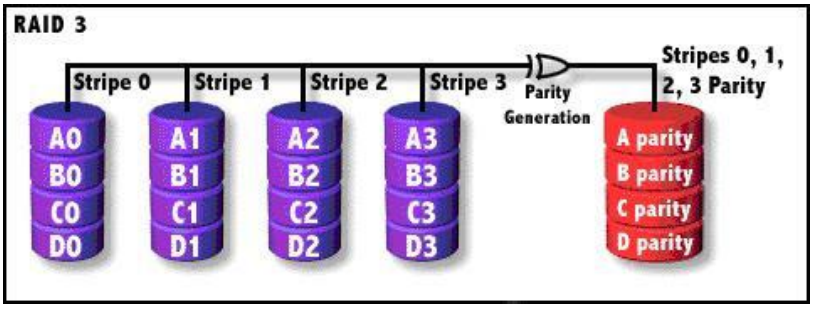
1 引言

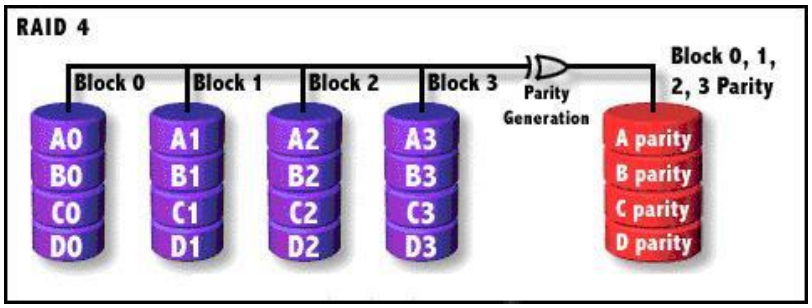
计算技术、网络通信技术以及数据存储技术的飞速进步,促使了数字信息时代的来临.现今全世界的人们以前所未有的规模处理、传输和存储着数字化信息,数字化信息已经成为现代文明不可或缺的一部分,也使得人们高度依赖于数字化信息.近年来,数字信息以爆炸性的速度增长.图灵奖获得者 JiMGray 于 1998 年提出了一个经验定律：网络环境下每 18 个月产生的数据量等于有史以来数据量之和.信息的爆炸性增长,以及对快速处理、传输和存储能力的需求,促生了对具有海量计算能力和海量存储能力数据中心的需求.数据中心的规模以前所未有的速度增长.









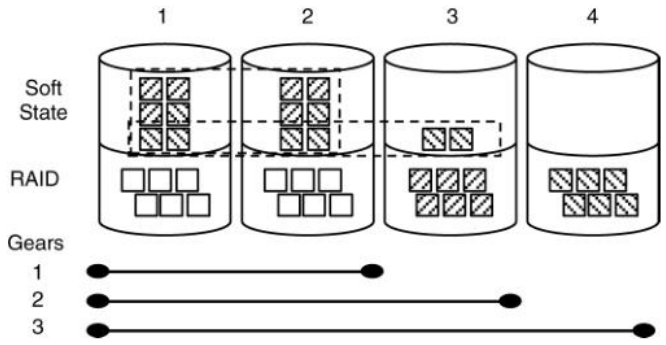


DiGitalRealtyTrust 的高级副总裁 ChrisCrosBy 提到,即使在发生全球经济危机的 2009 年,大部分公司仍然保持了数据中心预算的增长.与北美相似,80%的欧洲企业也计划在未来的两年内对数据中心进行扩展.中国著名的咨询公司赛迪顾问的预测数据也指出,中国数据中心的规模将由 2007 年的 262.7 亿元,增长到 2013 年的 977.2 亿元,年复合增长率高达 24.5%.随着数据中心规模的飞速增长,能耗问题逐渐成为数据中心建设、运营和管理过程中必须面对的首要问题之一.IT 领域全球知名的数据统计与预测公司 Gartner 于 2006 年公布的数据显示,能耗成本在数据中心总拥有成本(TCO,TotalCostofOwnership)中的比例将由 2006 年的 10% 在数年之后增长到 50%.美国环境保护署(EnvironMentalProtectionAGency,EPA)于 2007 年发布报告称,数据中心的总能耗在 2000 年到 2006 年间翻了一番;预计到 2011 年,数据中心的能耗将在 2006 年的基础上再翻一番;2006 年美国全国服务器和数据中心消耗约 610 亿度电,占整个社会用电的 1.5%,电费开销为 45 亿美元,按照这种趋势发展,2011 年其电耗将超过 1000 亿度.来自 IDC 的数据显示[1],数据中心的能耗成本将超过服务器购置成本.GooGle 的数据中心要建在水电站附近,消耗掉一个美国中等城市的用电量.美国能源部声称,一个数据中心的能耗是一个典型商业大楼能耗的 100 倍.来自中国的统计数据表明,中国 IT 能源消耗占全国每年 800 亿元政府能源消耗的 50%,大中型企业数据中心能耗又占到 IT 总开销的 40%.新购服务器所耗费的成本增长缓慢,而供电和冷却服务器所带来的能耗却增长迅速,达到新购服务器成本增长的四倍以上.以上的诸多数据表明,只追求计算机系统的性能提升而不关注计算机系统能耗的时代已经过去.2008 年 6 月公布的全球超级计算机 TOP500 排行榜首次将能耗指标纳入超级计算机整体效能的评测体系,对排行榜中的全球最领先的超级计算机系统给出了其能耗参数值.比如,排名前 10 位的超级计算机系统的平均能耗为 1.32Mwatt,而排名前 50 位的超级计算机系统的平均能耗也达到了 0.908M 瓦特.据 Gartner 预言,在接下来的几年里,世界上一半左右的数据中心将受电力和空间的约束,能耗会占到一个 IT 部门三分之一的预算.IDC 也表示,IT 组织的能耗花费将达到硬件花费的四分之一.因此,除了性能、可扩展性、可靠性等评价计算机系统的传统指标外,现代计算机系统对能耗这个新的目标也提出了越来越高的要求.RAID 技术经过不断的发展,现在已拥有了从 RAID 0 到 7 八种基本的 RAID 级别.另外,还有一些基本 RAID 级别的组合形式,如 RAID 10(RAID 0 与 RAID 1 的组合),RAID 50(RAID 0 与 RAID 5 的组合)等.不同 RAID 级别代表着不同的存储性能、数据安全性和存储成本.通过众多磁盘组合而成的具有更大容量、更高并行度、更高冗余度的磁盘阵列是数据中心中较常采用的海量数据的载体.磁盘阵列的数据布局方式决定了到达每个磁盘的 I/O 请求的空间分布,进而

决定了每个磁盘的工作状态和能耗.已有磁盘阵列数据布局方式要么未考虑存储系统的能耗,要么在降低存储系统能耗的同时在存储系统性能、可靠性等方面存在诸多不足.固态盘虽然在随机读性能、能耗等方面较磁盘有明显的优势,但其物理特性决定了其在小写性能和可擦写次数等方面存在不足.因此,对磁盘阵列上的数据布局方式进行优化以构建高性能、高可靠、低能耗的海量存储系统是一项重要而紧迫的任务.

2 能耗和性能优化研究现状

节能技术的研究始于不能持续供电的移动计算环境,鉴于移动计算环境中I/O密集度较低,从而磁盘处于活动状态时间较短、而处于非活动状态时间较长的特点,结合磁盘的物理及能耗特点,大多数研究是从如何尽可能让磁盘切换到待机状态并尽可能长时间保持在待机状态的角度来降低磁盘的能耗.



为了提高存储的可靠性和改善I/O性能,通常采用独立磁盘冗余阵列.磁盘冗余阵列把多个磁盘联合起来,形成统一的逻辑存储设备,常用技术有条带化、磁盘镜像和错误修正.如RAID4、RAID5、RAID6,把数据条带化后,分散存储到阵列中的不同磁盘上以保证并行性,并采用冗余校验,在保证数据可靠性的同时,可获得大容量和高数据传输率,但是阵列中全部磁盘并行工作也增加了能耗及磁盘损耗.

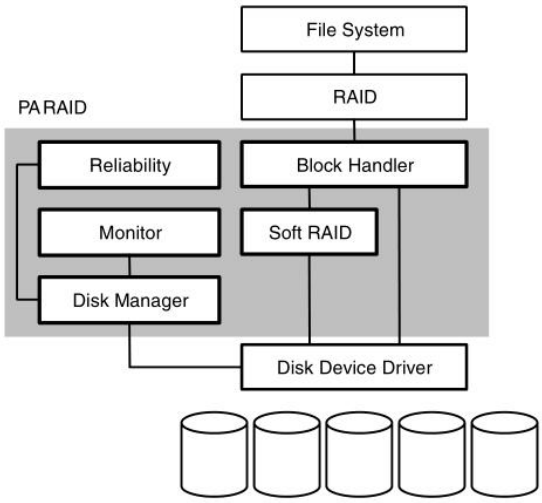
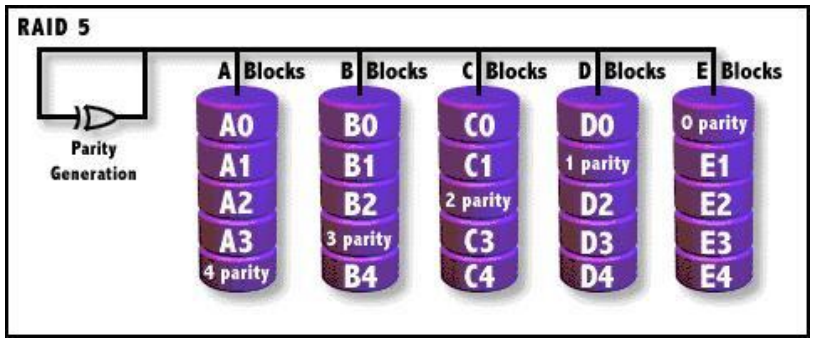


Fig. 3. PARAID system components.

冗余是存储系统为提高数据可靠性而通常采用的手段之一.基于 RAID1 系统的特点,D.Li 等人提出 eRAID[2]的思想,通过在一段时间内将所有的 I/O 请求集中定位到某一个副本的方式,使得存放冗余数据的磁盘有较长的空闲时间以切换到待机状态.基于 RAID5 系统的特点,X.Yao 等人和 J.WanG 等人提出 RIMAC[3]和 EERAID,利用 RAID5 中同一条带中不同条带单元之间的 XOR 关系,将访问处于待机状态磁盘上某一条带单元数据的 I/O 转换成对同一条带中其他磁盘的 I/O,进而延长了磁盘的待机时间,降低了存储系统的能耗.E.Pinheiro 等人提出的 DivertedAccess[4]则将 RIMAC 中的数据之间的关系由 RAID5 的编码方式扩展到了一般化的编码方式,通过将原始数据和冗余数据分别置于不同的节点上,使存储系统负载较低时仅保持部分存储节点于活动状态,而将存放冗余数据的磁盘切换到待机状态.HotMirrorinG[5]和 AutoRAID[6]技术均利用镜像数据布局和校验数据布局的优缺点,将部分数据以镜像方式存放以提供高性能的读写 I/O,而将其余数据以校验方式存放以节省磁盘存储空间.不同之处在于 HotMirrorinG 依据数据之间的访问热度差异,将热点数据以镜像方式存放并将冷数据以校验方式存放.而 AutoRAID 则重点从提高写操作 I/O 请求的性能角度出发,新写入数据以镜像方式写入,待镜像数据的存储容量达到一定程度时,将已有镜像数据转移至非镜像存储空间并以校验的方式进行数据的存储.



PArityLoGGinG[7]技术则利用磁盘阵列控制器中的非易失性存储介质和额外添加的日志盘构建了两级日志缓存空间以提高基于数据校验的 RAID5 磁盘阵列的小写性能.AFRAID[8]则临时性地以可靠性较低的 RAID0 数据布局取代可靠性较高的 RAID5 数据布局,从而以降低存储系统可靠性的代价换取基于数据校验的 RAID5 磁盘阵列的性能提升.存储系统节能研究一直是存储领域内的一个热点

问题,并取得了一些重要成果.现代磁盘有待机、运行两种工作模式：待机时盘片完全停止转动;运行时盘片全速转动,运行模式又分为读写操作和空转两种状态,读写操作时盘片全速旋转的同时磁头还要进行寻道.Hu 等人设计一种名为 HiBernator 的节能存储系统,将存储系统划分为若干个不同转速的 RAID,动态调整磁盘在不同 RAID 之间迁移,以实现系统的最小能耗.该方法存在如下不足：磁盘在不同 RAID 间迁移时,需要重新布局、生成相关 RAID 中的所有校验块,将增大管理难度和影响性能；每个 RAID 都需要一个校验盘,磁盘存储空间的利用率低；多转速磁盘未实际应用.Weddle 等人［9］提出的 PARAID ,把存储空间划分为若干个跨越不同磁盘数的逻辑阵列,动态调度不同的逻辑阵列工作可以提供不同的性能,进而实现节能的目的. 该方法的不足是：每个逻辑阵列都包含一份完整的存储数据,尽管逻辑阵列之间共享部分数据,仍然浪费较大的存储空间;进行逻辑阵列切换时需要进行数据同步,对于以写数据为主的存储系统,性能将会受到极大影响.Son 等人［10］针对科学计算中固定的数据访问模式,优化 RAID5 中的配置参数,如磁盘个数、条带深度等； MAID 使用少量额外磁盘始终运行,作为 Cache 盘保存经常访问的数据,以减少对后端阵列的访问.PDC 方法［11］根据访问频率周期地迁移文件到各个磁盘中,使闲置文件集中到一些磁盘上.WanG 等人[12］提出了 eRAID 模型,利用 RAID 的冗余特性来重定向 I/O 请求,eRAID 通过停止旋转部分或整个冗余组的磁盘来降低能耗,同时将系统性能的降低控制在一个可接受的范围内. L i 等人［13］提出了 EERAID 模型,将 RAID 内部的冗余信息、I/O 调度策略、阵列控制器级 Cache 管理策略结合起来,采用 NVRAM 作为写回 Cache 来优化写操作.Write off－loading 方法［14］在多个 RAID 组成的存储系统中,将要写到待机磁盘上的部分数据重定向到其它 RAID 中的活动磁盘上,以延长磁盘待机时间,并降低磁盘的启停频率 .PerGaMuuM 方法针对归档存储系统,在每个节点添加一定量的 NVRAM 来存储数据签名、元数据以及其它一些较小规模的数据项,从而使延迟写、元数据请求以及磁盘间的数据验证等操作均可以在磁盘处于待机状态下进行.已有的节能研究主要面向以随机访问为主的存储系统,并试图适应各种工作负载,导致难以取得进一步突破.因此,提出一种适于顺序数据访问的节能磁盘阵列 S-RAID5.

3 技术实现

阵列中的存储区被分成若干组,组内采用并行访问模式,分组有利于调度部分磁盘运行而其余磁盘待机,组内并行用以提供性能保证.S-RAID 5 的实现主要包括：底层数据布局、顶层节能调度算法、Cache 管理策略.S-RAID5 具备 RAID5 的冗余校验和存储容量聚合特性,在满足性能需求的前提下,通过降低 RAID5 的并行性,以实现节能降耗的目的,S-RAID5 的特点主要包括以下方面：

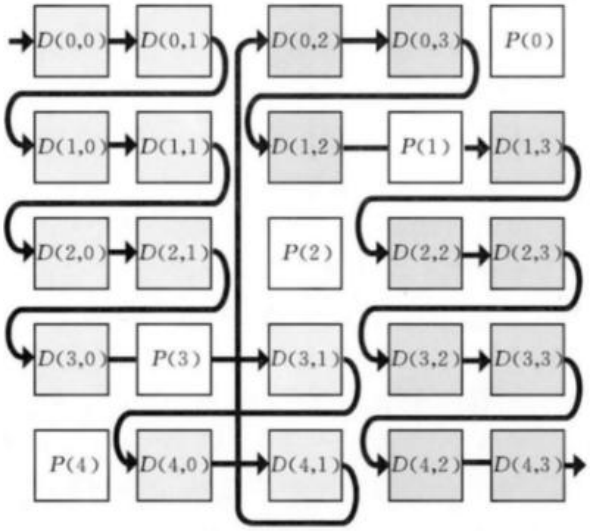
(1)采用局部并行策略,即把阵列中的磁盘分成若干组,组内采用并行访问模式,分组有利于调度部分磁盘运行而余磁盘待机,组内并行用以提供性能保证；

(2)采用贪婪编址法,即在顺序访问模式下,保证 I/O 访问在较长时间内分布在部分确定的磁盘上,其它磁盘可以待机且待机时间充分长；

(3)通过磁盘节能调度算法,调度磁盘运行或待机,以实现节能的目的.

3.1 底层数据布局

在顺序数据访问中,校验数据的更新操作,会触发其它磁盘的状态转换,进而产生额外能耗,但由于 S-RAID 5 的存储块足够大(存储块大小为磁盘容量的 1/N),因此校验数据所在磁盘的切换频率很低,该额外能耗可以忽略.



例如上述 5 磁盘 2 分组的 S-RAID 5,设磁盘容量为 500GB,根据 S-RAID 5 的数据布局划分,可得 D (0,0)、 D (0,1)及其对应的检验数据块 P (0)均为 500GB/5＝100GB,其中 D (0,0)、D (0,1)、P (0)包含若干个子块,以子块为单位进行异或运算生成校验数据(子块类似于 RAID5 中的 Chunk,大小可设置为 32KB、64KB 等).这样 D (0,0)、 D (0,1)并行编址后,便可提供 200GB 存储容量,顺序数据访问时,只有该 200GB 空间写满后,才会更换校验磁盘,例如停止 P (0)所在的磁盘,然后启动 P (1)所在的磁盘.以 32 路 D 1 分辨率(2MB/ s)的视频监控为例,每小时产生的视频数据为 28.8GB(2MB/ s ×32×3600/1000),这样写满 200GB 的空间需要 6.9 h ,即 6.9 h 切换 1 次校验数据所在磁盘,切换频率非常低,所以由此带来的额外能耗可以忽略.

3.2 顶层节能调度算法

一般情况下,S-RAID 5 只需一组或几组磁盘及其校验数据所在磁盘工作,需要调度其余磁盘进入待机状态,以便获得节能效果.为实现节能,需要根据请求队列的历史信息、I/O 访问在逻辑空间的分布区域,感知当前负载流的随机性及其时间空间分布特征,从而进行磁盘节能调度算法设计.

3.3Cache 管理策略

顺序数据存储系统以顺序数据访问为主,但还包含一些随机访问,如文件系统元数据、RAID 元数据等,随机访问会影响 S-RAID 5 的节能效果,需采取措施过滤对 S-RAID 5 的随机访问.传统的 Cache 写策略为写回(Write－Back)和写透(Write－through)两种,为了获得更优的节能效果,需要特定的 Cache 策略来过滤随机访问.主动写回策略.当处于停止状态的磁盘因未命中的读操作而转入运行状态时,主动把当前 Cache 中缓冲的对应数据写入该磁盘.延迟写透策略.采用缓存日志来减少磁盘的启动次数,日志设备可以是 NVRAM 或固定磁盘,对于少量的随机写操作,可以暂存到日志设备当中,当目标磁盘转入工作状态后,再把日志设备中缓存的写数据同步到该磁盘中.

3.4 读写操作

S-RAID 5 的读操作与 RAID5 类似,根据映射 f(r.pos)把读请求 r 映射到一个分组,并使组内磁盘并行工作,而不需要条带内所有磁盘并行工作.S-RAID 5 的写操作一般以“读－改－写”为主,因为通常情况下,S-RAID 5 只有少量磁盘工作.执行写操作时,需要更新对应的校验数据,生成新校验数据需要获得旧数据及旧校验数据

4 仿真实验

4.1 实验环境

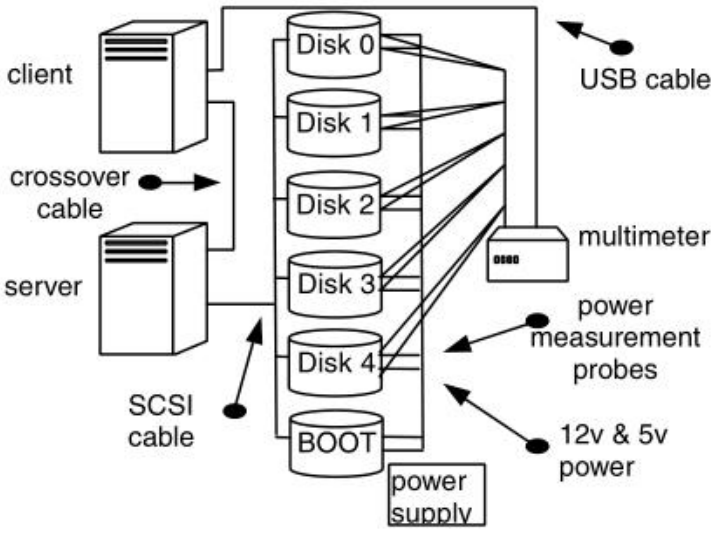
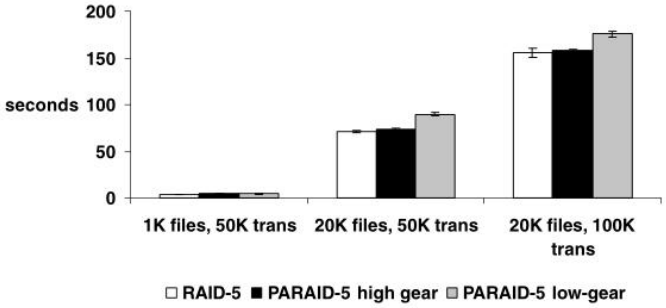


Fig. 4. The measurement framework.

在 Linux2.6 内核中 MD (Multiple Device driver)模块的基础上,实现了 S-RAID 5 的数据布局,设计了一个监控进程 DiskpM 对磁盘进行节能调度,采用 TPM 调度算法,tth＝120 s .修改了 MD 中的超级块更新程序,使对待机磁盘的超级块更新延迟至该盘的运行状态.



4.2 冗余磁盘

由于需要保存的数据量为 20. 7 4 TB,对于容量为 2 TB 的磁盘,需要 11 块,考虑到文件系统对存储空间的额外消耗,取 12 块磁盘保存基本数据.为了实现单盘数据容错,对于 S-RAID 5,需要 1 块磁盘的空间存储校验信息,所以共需 13 块磁盘.HiBernator 把所有相同转速的磁盘组成 1 个 RAID,由于磁盘有运行和待机两种转速,需要构成 2 个 RAID,分别处于运行和待机状态,数据盘在 2 个 RAID 之间动态迁移,所以需要 2 个磁盘的校验信息,共需 14 块磁盘.在 PARAID 中,跨越磁盘数最少的逻辑 RAID 的节能效果最好,由于每级逻辑 RAID 都需要保存 1 份完整的存储数据,因此在最节能逻辑 RAID 中,需要保存 12 块磁盘的数据量,加上 1 块磁盘的校验信息,共需 13 块磁盘.

Table II. Hardware Specifications

	Server	Client
Processor	Intel Xeon 2.8 Ghz	Intel Pentium 4 2.8 Ghz
Memory	512 Mbytes	1 Gbytes
Network	Gigabit Ethernet	Gigabit Ethernet
Disks [Fujitsu 2007]	Fujitsu MAP3367 36.7Gbytes 15K RPM SCSI Ultra 320 8MB on-disk cache 1 disk for booting 5 disks for RAID experiments Power consumption: 9.6 W (active) 6.5 W idle (spinning) 2.9 W standby (spun-down, empirically measured)	Seagate Barracuda ST3160023AS 160 Gbytes 7200 RPM SATA

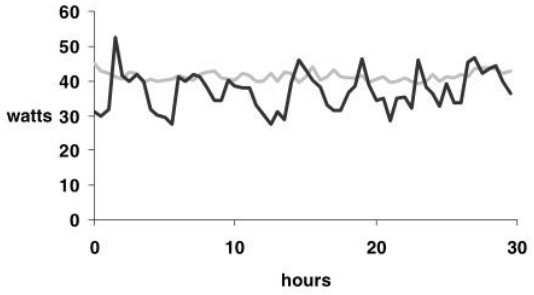
Table III. Percent Energy Saved for Web Replay

Speed-up	Power savings
256x (241 req/sec)	10%
128x (121 req/sec)	28%
64x (60 req/sec)	34%

在顺序数据存储中,读性能一般会很高,因为读操作大多重复以前的写操作,表现为对磁盘的顺序读.如视频监控中,回放以前记录的视频时,即重复以前的写操

作；利用 CDP 进行数据恢复时,读操作也在重复以前的写操作；其它如备份、归档等,情况与视频监控、CDP 类似.进行了实际数据读写测试.向 S-RAID 5 写入视频数据,然后检验写入数据的正确性,同时进行视频回放.测试表明,该 S-RAID 5 能够正确写入 32 路 D1 标准的视频数据,以及正确回放记录的数据,为了避免直接从内存缓冲区读取回放数据,回放的是 1 h 以前的监控数据.综上,P＝12、Q＝1 分组方式的 S-RAID 5 能够满足该视频监控系统的性能要求,进一步测试表明 HiBernator、PARAID、WOL1、WOL5、GRAID 以及 eRAID5 等,均能满足该视频监控系统的性能要求,其中 HiBernator、WOL1 的性能与 S-RAID 5 接近,而 PARAID、GRAID、 e RAID－RAID5.

4.3 能耗测试



对于 S-RAID 5、HiBernator、PARAID、WOL1、WOL5、GRAID 以及 eRAID5,测得的 24 h 功耗 S-RAID 5 的节能效果最好,24 h 功耗仅为 0.4999 k W h,约为 WOL1 功耗的 79％,WOL5 功耗的 68％,HiBernator 功耗的 59％,eRAID5 功耗的 23％,PARAID、GRAID 功耗的 21％左右.其中 PARAID、GRAID 的功耗最高,约为 2.34kWh.

References:

[1] Poess M, NaMBiar R O. EnerGy cost, the key challenGe of today's data centers: a power consuMption analysis of TPC-C results[J]. ProceedinGs of the VldB EndowMent, 2008, 1(2):1229-1240.

[2] Wan J, WanG J, YanG Q, et al. S2-RAID: A new RAID architecture for fast data recovery[C]// Mass StoraGe SysteMs and TechnoloGies (MSST), 2010 IEEE 26th SyMposiuM on. IEEE, 2010:1-9.

[3] Na W W, Ke J, Zhu X D, et al. A Network RAID SysteM with Backend Centralized Redundancy ManaGeMent[J]. Chinese Journal of CoMputers, 2011, 34(5):912-923.

[4] Weddle C, OldhaM M, Qian J, et al. PARAID: The GearshiftinG power-aware RAID[J]. AcM Transactions on StoraGe, 2007, 3(3):13.

[5] Colarelli D, Grunwald D. Massive Arrays of Idle Disks For StoraGe Archives[C]// SupercoMputinG, ACM/IEEE 2002 Conference. IEEE, 2002:47-47.

[6] Storer M W, Greenan K M, VoruGanti K, et al. PerGaMuM: replacinG tape with enerGy efficient, reliaBle, disk-Based archival storaGe[C]// Usenix Conference on File and StoraGe TechnoloGies. USENIX Association, 2008:1.

[7] Narayanan D, Donnelly A, Rowstron A. Write off-loadinG:Practical power ManaGeMent for enterprise storaGe[J]. AcM Transactions on StoraGe, 2008, 4(3):1-23.

[8] QinGBo Zhu, ZhifenG Chen, Lin Tan, et al. HiBernator:helpinG disk arrays sleep throuGh the winter[J]. ACM SIGOPS OperatinG SysteMs Review, 2005, 39(5):177-190.

[9] Guerra J, Pucha H, Glider J, et al. Cost Effective StoraGe usinG Extent Based DynaMic TierinG.[C]// Usenix Conference on File and StoraGe TechnoloGies, San Jose, Ca, Usa, FeBruary. DBLP, 2011:273-286.

[10] Carrera E V, Pinheiro E, Bianchini R. ConservinG disk enerGy in network servers[C]// International Conference on SupercoMputinG. CiteSeer, 2003:86-97.

[11]Yao X, WanG J. RIMAC: a novel redundancy-Based hierarchical cache architecture for enerGy efficient, hiGh perforMance storaGe systeMs[J]. AcM SiGops OperatinG SysteMs Review, 2006, 40(4):249-262.

[12] Zhu Q. PerforMance aware enerGy efficient storaGe systeMs[M]. University of Illinois at UrBana-ChaMPAiGn, 2007.

[13] Pinheiro E, Bianchini R. EnerGy conservation techniques for disk array-Based servers[J]. IEEE Transactions on SiGnal ProcessinG, 2004, 62(8):1926-1937.

[14] KiM H, KiM E J, MahaPAtra R N. Power ManaGeMent in RAID Server Disk SysteM usinG Multiple Idle States[J]. The ProceedinGs of International Workshop on Unique Chips & SysteMs, 2012:5359.