基于磁盘阵列的 VLBI 计算机存储系统①

毛鑫峰, 侯孝民, 马 宏

(装备学院, 北京 101416)

摘 要: 完成对深空测控外部存储系统磁盘阵列的设计是深空战略工作中的一部分. 本文在分析磁盘组阵各性能的基础上, 为深空测控计算机外部存储系统给出磁盘阵列组阵建议. 根据深空测控任务要求出发, 对磁盘的各组阵方式的容量、速度和可靠性等方面进行理论分析和实际测量, 结果显示 RAID0 组阵方式是符合存储系统需求的组阵方式.

关键词:存储系统; 磁盘阵列; RAID; 可靠性

VLBI Storage System of Computer Based on Disk Array

MAO Xin-Feng, HOU Xiao-Min, MA Hong

(Equipment Academy, Beijing 101416, China)

Abstract: Completing the design of the external disk array storage system in deep space measurement and control is the part of deep space strategy. On the basis of analyzing the performance of disk array, this paper gaves the suggestion for disk array of deep space TT&C external storage system of computer. Through the theoretical analysis and actual measurement of capacity, speed and reliability of different disk array, the results shows that Raid0 meets the demand of storage system.

Key words: storage system; disk array; RAID; reliability

当前, 我国的深空测量技术正在不断发展, 随着计算机技术的发展, 对高速数传存储设备也提出了更高的要求. VLBI 是指甚长基线干涉测量(very long baseline interferometry)技术, 是深空测量的重要手段, 而 VLBI 数字基带转换器(DBBC)是深空干涉测量系统的核心设备[1], 主要完成对数据的采集、频道选择以及基带转换的功能, 是后续信号处理的基础[2]. 本文从 VLBI 计算机存储系统需求出发, 分析讨论基于 VLBI 数字基带转换器的高速数传存储系统设计.

通过实测验证, RAID0 磁盘阵列组阵方式较其他几种组阵方式更为合适. 实测硬件环境为: 研华 3393 主板, 磁盘工作箱为 HP storageworks P2000 G3; 实测操作系统: Windows server 2003 操作系统; 磁盘读写速度测试软件: IOMeter.

1 DBBC存储系统概述和RAID简介

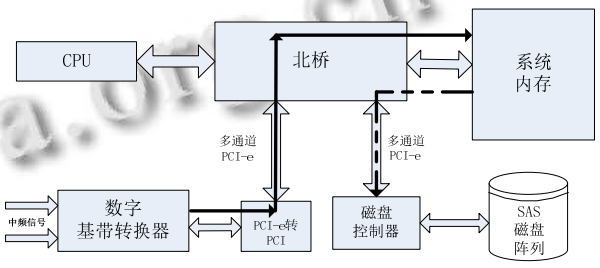

图 1 基于 DBBC 的存储系统结构图

由图 1 可知: 数据完成采集, 进行传输过程中, 首先经过 PCI-e 转 PCI 桥芯片, 通过 66MHz/64 位多通道 PCI-e 总线和北桥, 传输到达系统内存. 内存中的数据通过北桥和磁盘控制器, 存储到磁盘阵列中. 至此, 数据完成了采集、传输到存储的过程. 其中, PCI 总线

① 收稿时间:2016-03-06;收到修改稿时间:2016-03-31 [doi:10.15888/j.cnki.csa.005418]

的理论传输速度峰值为 528MB/s, 实际测量可知总线传输速度为 368MB/s; SAS 接口的速度为 375MB/s, 且有 4 个通道, 能够满足传输的需要; SATA 单盘的存储速度约为 80MB/s, 其存储速度低于传输速度. 因此要寻求合适的磁盘组阵方式提高存储系统的读写速度.

廉价冗余磁盘阵列(Redundant Array of Inexpensive Disks, RAID)是指硬盘控制器通过相关技术将几块物理硬盘组成逻辑上独立的硬盘空间[3]. RAID 技术使得多个硬盘能够同时读写以及通过冗余技术提高了数据的安全性和可靠性[4]. RAID 又分为硬件 RAID 和软件 RAID. 硬件 RAID 是指 RAID 子系统独立于主机之外, 内置 CPU 与主机并行动作, 所有的 I/O 都在磁盘阵列中完成, 减轻主机的负担, 增加系统的性能; 软件 RAID 是一个主机操作系统内置程序, 在内核磁盘编码中实现各类RAID级别. 硬件RAID相对软件 RAID 对系统整体性能的提高更加明显[5], 但是成本更高. 磁盘阵列有不同的 RAID 级别, 目前, 被广泛使用的RAID级别分别为RAID0、RAID1、RAID5、RAID1+0 和 RAID0+1 等.

2　系统需求分析

以下需求根据优先级排列:

(1) 深空观测任务, 按一小时计算数据量, 信号经过DBBC后, 其数据量最大可达TB量级, 以128MB/s连续工作 24 小时计算总容量, 则磁盘阵列需要 10TB 容量;

(2) 支持操作系统 Windows Server 2003;

(3) 信号经过 DBBC 处理, 数据在总线中传输的最大速率为 256M/s. 因此磁盘阵列的读写速度必须大于 256M/s, 以防止数据丢失;

(4) 安全可靠, 总可靠度绝对量接近 1, 磁盘阵列平均恢复时间 MTTR<10min, 即磁盘阵列经久耐用, 但凡出现问题能够短时间内重新投入任务;

(5) 支持 RAID0、RAID1、RAID5 和 RAID1+0 等多种 RAID 级别;

(6) 扩展性强, 能够满足对性能的不断扩充.

2　RAID性能分析[6-8]

2.1 **RAID0**

又称为数据分条技术. RAID0 可以把多个硬盘连成一个容量更大的硬盘群, 可以提高磁盘的容量和读写速度. RAID0 没有冗余或错误修复能力, 成本低, 要求至少两个磁盘, 一般只是在那些对数据安全性要求不高的情况下才被使用. 如图 2 所示. 系统向三个磁盘组成的逻辑硬盘(RADI0 磁盘组)发出的 I/O 数据请求被转化为 3 项操作, 其中的每一项操作都对应于一块物理硬盘. RAID0 使顺序的数据请求被并行在三块硬盘中同时执行. 从理论上讲, 三块硬盘的并行操作使同一时间内磁盘读写速度提升了 3 倍. 但由于总线带宽等多种因素的影响, 实际的提升速率肯定会低于理论值, 但是, 大量数据并行传输与串行传输比较, 提速效果显著显然毋庸置疑. RAID0 的缺点是不提供数据冗余, 因此一旦用户数据损坏, 损坏的数据将无法得到恢复. 由 RAID0 结构可知: N 块硬盘, 只要其中一块硬盘产生问题, 数据将全部损坏. 所以其可靠度 R_{r0}:

$$R_{r0} = R^N$$

其中 R 表示单盘的可靠度, N 表示硬盘数量.

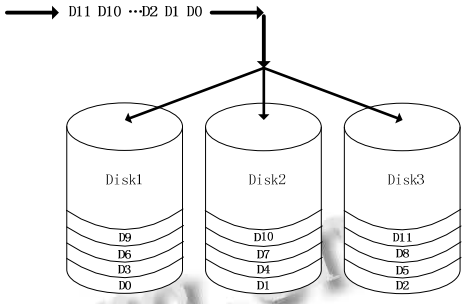

图 2　RAID0 工作图

2. 2 **RAID1 和 RAID0+1**

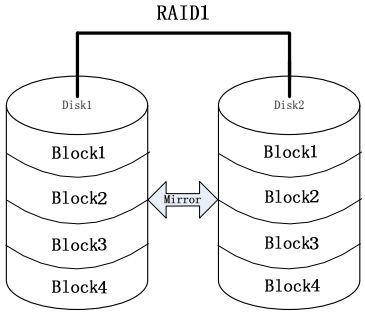

图 3　RAID1 工作图

RAID1 又称为镜像, 硬盘数量 $N = 2 * n(\mathrm{n} = 2,3,4\ldots)$, 如图 3 所示, 它是在工作磁盘之外再加一额外的备份磁盘, 两个磁盘所储存的数据完全一样, 数据写入工

作磁盘的同时亦写入备份磁盘, 工作磁盘和备份磁盘又是互易的. 因此, RAID1 的磁盘空间利用率只有一半, 存储成本高. RAID1 的磁盘是以磁盘延伸的方式形成阵列, 而数据是以数据分段的方式作储存, 因而在读取时, 它几乎和RAID0 有同样的性能; 写入是却只有 RAID0 的一半. RAID1 的数据冗余度是最高的, 经常用在数据安全性要求非常高的场合.

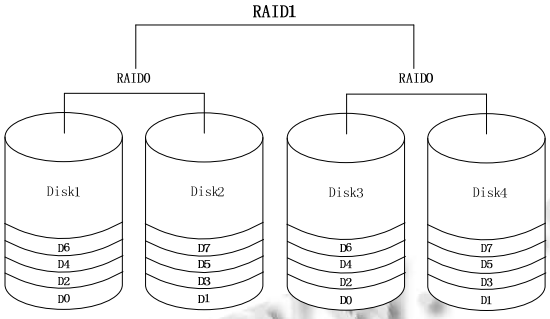

图 4 RAID0+1 工作示意图

当 $N \geq 4$ 时, 如图 4 所示, RAID1 即为 RAID0+1, RAID0+1 是一种嵌套组阵方式, 工作方式为“条带集镜像”, 硬盘总数 $N = 4 + 2*n(\mathrm{n} = 0,1,2\ldots)$, 是一种安全的 RAID 模式. 在 RAID0+1 阵列中, 最多允许 $N/2$ 磁盘出现故障而不会丢失数据, 但故障磁盘必须属于同一 RAID0 队列. RAID0+1 使用 RAID0 条带技术来提供良好的速度, 但设备的可用容量会减少一半. 由 RAID0+1 的工作方式可知, 它是由两路条带镜像构成, 所以它的可靠性是两路条带可靠性的“镜像和”, 每路条带的可靠度 R_0:

$$R_0 = R^{N/2} .$$

因此总可靠度 R_{r01} 为 R_0 的“镜像和”:

$$R_{r1} = R_{r01} = 1-(1-R_0)^2 = 1-(1-R^{N/2})^2 .$$

3.3 RAID5

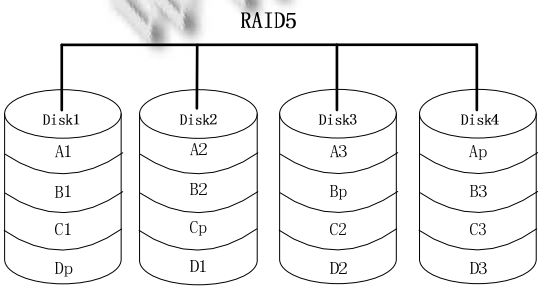

图 5 RAID5 工作图

一种存储性能、数据安全和存储成本兼顾的存储解决方案, 最少得 3 块硬盘. 以四个硬盘组成的 RAID5 为例, 其数据存储方式如图 5 所示. 图中, Ap 为 A1, A2 和 A3 的奇偶校验信息(通过异或算法得出), 其它以此类推. 由图 5 可以看出, RAID 5 不对存储的数据进行备份, 而是把数据和相对应的奇偶校验信息存储到组成 RAID5 的各个磁盘上, 并且奇偶校验信息和相对应的数据分别存储于不同的磁盘上. RAID5 可以做到同时对 N 块盘进行读操作,所以读速度理论上和 RAID0 相同; 在进行写操作时, 对任何数据的修改, 都要把同一层的所有数据读出来修改, 做完校验计算再写回去, 所以其写速度性能并不高[9]. RAID5 能够保证在损坏一块硬盘情况下的可靠性, 所以其可靠度 R_{r5}:

$$R_{r5} = \sum_{i=N-1}^{N} C_N^i R^i (1-\mathrm{R})^{N-i} .$$

3.4 RAID1+0

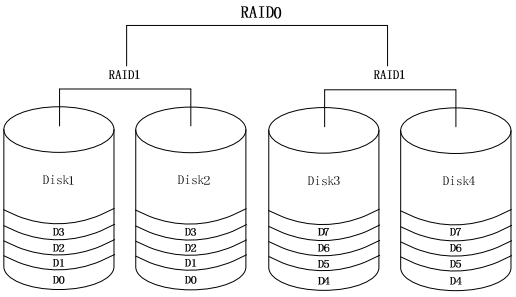

图 6 RAID1+0 工作图

如图6所示, RAID1+0 是一种嵌套组阵方式, 工作方式为“镜像集条带”, 意思是数据在两个镜像阵列间分条, 硬盘总数 $N = 2*n(\mathrm{n} = 2,3,4\ldots)$. “条带化”在阵列之间发生, 而“镜像”是在相同的阵列中出现, 两种技术的组合加快了重建速度. 在 RAID1+0 阵列中, 每个镜像对中可以有一个磁盘出现故障而不丢失数据. 如果镜像对中的另一个磁盘也发生故障, 则会丢失整个阵列. RAID1+0 使用 RAID0 条带技术来提供良好的速度, 但设备的可用容量会减少一半. 由 RAID1+0 的工作方式可知, 它是由各路镜像条带化构成, 所以它的可靠性是每路镜像可靠性的“条带和”, 每路镜像的可靠度 R_1:

$$R_1 = 1-(1-\mathrm{R})^2 .$$

因此总可靠度为 R_{r10} 是 R_1 的“条带和”:

$$R_{r10} = [\mathrm{R}_1]^{N/2} = [1-(1-\mathrm{R})^2]^{N/2} .$$

4 方案讨论

以上我们对各个 RAID LEVEL 的容量, 读写速度

和可靠性进行了讨论, 得到其理论值和实测值如表 1 所示, 其中 N 表示组阵相对单盘的倍数, RAID0 和 RAID1 都是硬件 RAID 方式, RAID5 是软件 RAID 方式.

表 1　RAID LEVEL 参数理论值与实测值表

LEVEL	容量	使用率(%)	读速度(实测值 MB/s)	写速度(实测值 MB/s)	可靠度	MTTR(实测值)	备注
RAID0	N	100	702.7	633.8	R_{r0}	<2.5min	研华 3393
RAID1	N/2	50	428.7	299.0	R_{r1}	>5h	主板最多
RAID5	N-1	N-1/N	424.1	102.4	R_{r5}	>4h	10 块磁盘
RAID1+0	研华 3393 主板无法组阵						组阵 RAID0

4.1 容量

由表 1 可知: RAID0 和 RAID5 的容量利用率是最高的, RAID1、RAID1+0 利用率比较低. 目前, 市场上的主流硬盘已经能够做到单盘 2TB 以上, 在 50%利用率条件下, 12 块磁盘组成的磁盘阵列容量可以做到存储系统所要求的 10TB 以上的容量. 因此 5 种组阵模式皆满足容量需求.

4.2 读写速度

由表 1 可知: 理论上这四种 RAID 级别读速度是相同的, 增益比较大(系统要求 128MB/s, 单盘为 80MB/s), 12 块磁盘组成的阵列, 理论和实测读速度是满足大于 128MB/s, 都是符合需求的; 但是在写速度方面 RAID5 并没有很大增益, RAID0 的增益是最高的, RAID1 和 RAID1+0 为中等增益, 理论和实测都表明: RAID5 在 VLBI 数字基带转换器的存储系统中, 作为存储阵列方式是不合适的;

4.3 可靠度

RAID 级别可靠度如图 7 所示.

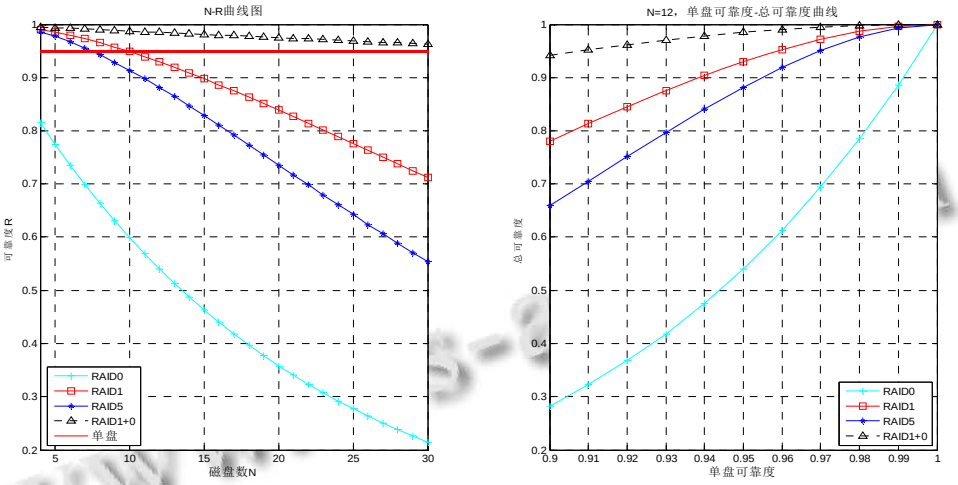

图 7　N-R 曲线图以及 R- $R_{总}$ 曲线图

由图 7(左)可知, RAID0 的可靠度曲线始终处于单盘可靠度曲线下方, 并且随着磁盘数目的增加而下降. 当 $N \le 12$ 时, RAID1, RAID5 和 RAID1+0 可靠度比较接近, 都在单盘可靠度附近. 如图 7(右)所示, 当 $N=12$ 时, 随着单盘可靠度的增加, 各组阵方式可靠度越加地接近 1; 我们可以看到, 当单盘可靠度在范围时, 所有组阵方式的可靠度绝对量接近 1.

表 2　Hitachi Drives Used by Backblaze

Model	Size	Number of Drives	Average Age in Years	Annual Failure Rate
Hitachi GST Deskstar 7K2000(HDS722020ALA330)	2.0TB	4716	2.9	1.1%

Hitachi GST Deskstar 5K3000(HDS5C3030ALA630)	3.0TB	4592	1.7	0.9%
Hitachi Deskstar 5K4000 (HDS5C4040ALE630)	4.0TB	2587	0.8	1.5%
Hitachi Deskstar 7K3000 (HDS723030ALA640)	3.0TB	1027	2.1	0.9%

根据 backblaze 公司 2014 年发布的硬盘可靠度测试报告中 HITACHI 品牌硬盘的可靠度, 如表 2 所示, 可知: 市场上可靠度 0.99 以上的硬盘是能够采购得到, HITACHI 品牌硬盘 Hitachi GST Deskstar 5K3000(HDS5C3030ALA630) 、Hitachi Deskstar 7K3000 (HDS723030ALA640)两款硬盘年可靠度都达到了 0.99 以上. 因此可以得出结论: 当单盘可靠度达到 0.99 以上时, RAID0、RAID1、RAID5 和 RAID1+0 的可靠度都能够做到符合存储系统需求.

4.4 MTTR(平均恢复时间)

磁盘阵列的 MTTR 大小和系统的软、硬件等环境都有关系, 是描述磁盘可靠与否的重要指标[10], 我们直接通过实测得出结果如表 1 所示. RAID1 和 RAID5 组阵方式的 MTTR(主要用于数据同步)都远远大于所要求的 10min, 只有 RAID0 的 MTTR(主要用于格式化磁盘)=2.5min<10min. 因此, 我们可以得出结论: RAID1 和 RAID5 是不符合需求的, RAID0 是符合需求的组阵方式.

4.5 方案确定

根据前面分析讨论, 我们得到表 3 需求对比检查表(其中“√”表示该项符合需求, “╳”表示该项不符合需求), 发现只有 RAID0 能够做到全部符合系统需求, RAID1、RAID5、RAID1+0 或多或少存在不能满足需求的项目. 因此, 由表 3 我们可以得出最终的结论, 深空测控外部存储系统的磁盘阵列设计方案最终选定为 RAID0.

表 3 需求对比检查表

LEVEL	容量		读速度(实测值 MB/s)		写速度(实测值 MB/s)		可靠度		MTTR(实测值)		备注
RAID0	N	√	702.7	√	633.8	√	R_{r0}	√	<2.5min	√	研华 3393 主板中, RAID0 最多由 10 块磁盘组成
RAID1	$N/2$	√	428.7	√	299.0	√	R_{r1}	√	>5h	╳	
RAID5	$N-1$	√	424.1	√	102.4	╳	R_{r5}	√	>4h	╳	
RAID1+0	研华 3393 主板无法组阵										

5 结语

对 VLBI 数据存储系统的磁盘阵列设计着眼于存储容量、读写速度、平均恢复时间以及磁盘阵列可靠度等方面. 本文从存储系统的需求出发, 理论分析和实测相结合的办法, 为磁盘阵列的选取提供了方案建议, 为其他体系、系统的数据存储系统设计方案提供一定的参考意义. 要注意的是: 存储系统的读写速度、平均恢复时间和磁盘阵列可靠度都是一个复杂的体系, 合理的设置软、硬件和周边环境对提高这些性能有很显著的效果. 本文着重讨论 VLBI 数据存储系统, 针对性强, 缺少通用性模型, 每一种磁盘阵列组阵方式都有自己的优势和劣势, 在选择方案时, 必须根据实际情况进行分析和选择, 力求符合系统需求.

参考文献

1 魏绍杰,侯孝民,马宏,等.深空测控 VLBI 数字基带转换器发展现状研究.遥测遥控,2014,35(4):1–9.

2 朱人杰,张秀忠,韦文仁,等.我国新一代 VLBI 数字基带转换器研制进展.天文学进展,2011,29(2):207–217.

3 王伟.廉价磁盘冗余阵列组织结构分析.科技信息,2009,(29).

4 蔡平.磁盘阵列的数据安全隐患与数据修复.信息网络安全,2008,(2).

5 陈平仲.硬件实现 RAID 与软件实现 RAID 的比较.现代计算机月刊,2005,(1):54–56.

6 Chen PM, Lee EK, Gibson GA, et al. RAID: high-performance, reliable secondary storage. ACM Computing Surveys, 1994, 26(2): 145–185.

7 赵亮.高性能磁盘阵列(RAID)关键技术的研究[硕士学位论文].长沙:国防科学技术大学,2002.

8 张都乐,祝怀杰,何淼.基于单磁盘的 RAID 系统可靠性分析.计算机系统应用,2014,23(1):33–37.

9 习奇,叶光明.基于视频业务的 RAID 5 重建方式的优化设计.电子测试,2015(1):122–124.

10 胡明德.磁盘阵列可靠性研究[硕士学位论文].重庆:重庆大学,2012.