并行内存文件系统

杨学军 王 磊 蒋艳凰 编著

国防科技大学出版社

并行内存文件系统

杨学军　王　磊　蒋艳凰　编著

国防科技大学出版社
·长沙·

图书在版编目（CIP）数据

并行内存文件系统/杨学军，王磊，蒋艳凰编著 .—长沙：国防科技大学出版社，2005.6
ISBN 7-81099-185-X

Ⅰ.并… Ⅱ.①杨…②王…③蒋… Ⅲ.并行内存贮器：内存贮器-文件系统 Ⅳ.TP333.1

中国版本图书馆 CIP 数据核字（2005）第 054681 号

国防科技大学出版社出版发行
电话：（0731）4572640 邮政编码：410073
E-mail：gfkdcbs@public.cs.hn.cn
责任编辑：何晋 责任校对：肖滨
新华书店总店北京发行所经销
国防科技大学印刷厂印装
*
开本：850×1168 1/32 印张：5.5 字数：138 千
2005 年 6 月第 1 版第 1 次印刷 印数：1-1500 册
ISBN 7-81099-185-X/TP·15
定价：12.00 元

本书的出版得到国家863重大软件专项服务器操作系统内核项目（2002AA1Z2101）资助。

目 录

第一章 绪 论

1.1 文件系统

在操作系统中，文件系统是指负责存取和管理文件信息的机构，即负责文件的建立、删除、组织、读写、修改、复制以及对文件管理所需的资源（如目录表、存储介质等）实施管理的软件部分。

有了文件系统，用户就可以用统一的文件观点去对待和处理各种存储介质中的信息，并通过文件系统去使用各种存储器。因此文件系统可以视为用户和文件存储器之间的接口，用户无需考虑存储器的信息组织，也无须记住信息在存储器上的存放和分布情况，借助文件名便可方面、灵活地对信息进行存取，所有访问事宜均由文件系统自动完成。由于磁盘所保存的信息不易丢失，而且磁盘文件系统的技术也很成熟，目前，绝大多数计算机系统均采用磁盘作为主要的文件存储器。

1.2 内存文件系统的发展契机

与磁盘文件系统不同，内存文件系统（Memory-based File

System）是指以半导体存储器而非传统的磁盘作为文件存储介质的文件系统。

半导体存储器可分为易失性（volatile）的和非易失性的（nonvolatile）两类。在不对存储器采取任何容错措施的情况下，当系统发生电源故障时，存储在易失性半导体存储器中的数据将丢失，而存储在非易失性半导体存储器中的数据则不会丢失。易失性半导体存储器的代表就是常用作计算机系统主存的 DRAM，非易失性半导体存储器有闪存（Flash Memory）、带后备电源的 SRAM 等。为了方便起见，在下文中，我们将所有半导体存储器统称为内存。目前的计算机系统一般以易失性的半导体存储器作为主存储器。不过，半导体存储器在计算机系统中的作用并不仅仅限于用作主存储器。非易失性的半导体存储器可以代替磁盘用作外部存储器，例如目前常用的 U 盘就是一例。

早在 20 世纪 80 年代末 90 年代初，人们就开始了内存文件系统[1~4]方面的研究。除了利用内存高带宽、低延迟的访问特性来提高系统的 I/O 性能外，早期的内存文件系统大多针对特定领域的需求进行研制。例如，无盘的移动设备、具备非易失性内存的设备、嵌入式系统、仅需存储临时文件数据的环境等。内存的容量和价格限制使得系统规模不可能很大，大多数系统基本上只面向串行应用，应用的领域和范围受到了很大的限制。

但是，受应用需求和技术发展的驱动，内存文件系统，尤其是面向并行计算系统的内存文件系统的研究获得了新的契机。下面我们分别从应用需求、技术发展和解决途径三方面引出并阐述内存文件系统的研究背景和动机。

1.2.1 应用需求牵引

从过去十多年中计算机系统的发展情况来看，内、外存之间

的性能差距进一步扩大。表 1.1 给出了典型的计算机系统中存储层次的性能参数[20]。从表中的数据可以看出，在访问时间上，相邻存储层次中内存和磁盘之间的性能差距是最大的。

表 1.1 典型的计算机系统中存储层次的性能参数

	访问时间（ns）	带宽（MB/s）	容量（Byte）
寄存器	1 ~ 5	4000 ~ 32000	< 1K
cache	3 ~ 10	800 ~ 5000	< 8M
内存	70 ~ 400	800 ~ 3000	< 4G
磁盘	5000000	20 ~ 50	> 10G

Amdahl 法则[45, 110]表明，计算机系统的性能受限于系统中最慢的部件。内外存和微处理器间性能的不平衡性导致在大多数计算机系统中 I/O 子系统成为制约系统性能提高的瓶颈，在大规模科学计算和高端事务处理领域这一情况尤为明显。这使得人们，特别是那些急切想要打破 I/O 瓶颈的大规模科学计算和高端事务处理领域的科研工作者，在计算机系统的设备级、文件系统级、应用程序级等各个层次不断探索加速 I/O 操作的技术途径。

然而不幸的是，受限于磁盘固有的机械运动特性，磁盘的容量和数据访问带宽虽然逐年提高，对其访问延迟的改善却收效甚微。这使得基于磁盘的 RAID 技术、并行文件系统技术以及 I/O 库技术虽然对大块连续数据能够提供较高的访问带宽，但是对于访问延迟却难以控制，尤其对小粒度、随机分布的数据更是如此。

从应用需求来看，小粒度、随机 I/O 访问模式在现有的应用中并不少见。大量存在小粒度、随机 I/O 访问模式的应用主要

有：编译程序、web 服务、mail 服务、并行事务处理等。实际上，小粒度、随机 I/O 访问模式正是造成 I/O 瓶颈的一个重要原因，不同研究者对并行环境下的文件访问特性的研究[6, 13, 23, 29]都说明了这一点。列举如下：

• 以往的研究表明大文件大都是被顺序访问的。然而，2000 年伯克利分校 Roselli 等人对包括 NT 和 Unix、客户端和服务器、教育界和产业界的几种不同环境中的文件系统负载进行了分析[13]，他们的研究却表明目前大文件被随机访问的可能性比以往的研究结果大得多。

• 在科学计算领域中，卢凯博士对 Pablo 并行 I/O 踪迹进行动静态分析[23]，主要分析了并行景象生成应用（Render）、合成孔径雷达信息处理应用（SAR）、多通道电子散射应用（ESCAT）、流体力学模拟应用（PRISM）等四个典型的 I/O 受限的并行科学计算应用的 I/O 踪迹。卢凯博士的研究表明，造成某些并行计算 I/O 受限的重要原因是小粒度 I/O 访问模式而不是 I/O 访问的数据量。

• Nils Nieuwejaar 等人对并行科学计算应用的文件访问特性进行分析[29]，结果表明，在 MIMD 并行计算机系统中，存在大量小粒度的 I/O 请求。在某种程度上，这是数据在处理器间的逻辑划分模式与文件的物理分布模式不同所造成的。

• Galley 并行文件系统[6]设计者的经验也表明，许多应用向文件系统提出小粒度的、规则但非连续的请求。

• 我们对并行蒙特卡洛模拟程序进行分析，也得到了类似的结论，即小粒度、随机 I/O 访问模式是造成该模拟程序 I/O 受限的根本原因。并行蒙特卡洛模拟程序是大规模并行计算机系统中模拟核爆中光子、中子、电子的耦合、输运作用的重要程序之一。其中，被模拟的每个粒子在数据文件中大致只需以 0.2KBytes 的数据量表示，模拟程序的每个循环步骤仅随机访问

数据文件中的若干个粒子。也就是说，程序的 I/O 数据量并不大，对系统 I/O 带宽的需求也不高。但是，小粒度、随机 I/O 访问模式却导致其 I/O 受限。图 1.1 给出了并行蒙特卡洛模拟程序的大致 I/O 特性。

```
输入：数据文件 ssbk，处理器数 P，粒子数 N
输出：数据文件 ssbk，中间数据文件：ssbk_tmp（i），i=1，2，…，P

Begin

  ……

  DO
  For j=1，2，…，N/P DO
        每个处理器上的子任务在执行前根据本身的计算结果随机地
    从文件 ssbk 中读出若干个粒子的数据；
        中间计算：
        每个处理器上的子任务向本地的中间数据文件 ssbk_tmp（i）
    写入若干个粒子的数据；
    EndFor

    把各个子任务产生的 ssbk_tmp（i）合并成为新的 ssbk；
  Until 模拟结果收敛

  ……

End
```

图 1.1 并行蒙特卡洛模拟程序的大致 I/O 特性

从以上对小粒度、随机 I/O 访问模式的应用需求分析来看，我们相信，那仅仅是应用需求的冰山一角。实际上，为了匹配磁盘的 I/O 特性，不少未在上面列出的应用在代码化时也是经过程序员以增加工作量、牺牲可理解性等大量额外的软件或者智力上的开销才能具备大块、连续 I/O 访问特性的。如果持久存储层次本身能够对小粒度、随机 I/O 访问提供较好的支持，那些额外的优化工作本可以避免。

与磁盘对小粒度、随机 I/O 访问无能为力形成鲜明对比的是，内存固有的特性恰恰能够对应用的小粒度、随机访问模式提供高带宽、低延迟的支持。这使得合理利用整个系统的内存来加速随机、小粒度的 I/O 请求，成为目前乃至今后相当长一段时期内非常重要而有效的技术手段。

1.2.2 技术发展驱动

从计算机技术发展的现状及趋势来看，虽然磁盘的存储容量增长较快（大约每年 60% ~ 100%），但受到其固有的机械运动特性的制约，磁盘访问延迟和带宽的增长速度都大大落后于微处理器、内存以及网络性能的增长速度。

目前，虽然人们已经采取了利用 cache 技术减少机械操作[26]、通过 zone 技术[27]和 track _ aligned 技术[28]提高旋转效率等多种优化措施，但磁盘访问延迟与内存访问延迟仍大致相差 5 ~ 6 个数量级，与网络访问延迟大致相差 3 个数量级，而且该差距还有以每年 50% 的速率进一步增加的趋势[24]。这导致计算机系统尤其是大规模并行计算机系统中的 I/O 瓶颈问题越来越突出，成为制约系统性能提高的关键性问题之一。

1.2.2.1 曾经阻碍内存存储文件数据的障碍

显然，在层次存储系统中，当性价比在合适的范围内时，要访问的数据应该被尽量存储在离处理器较近的存储层次中才能提高系统的性能。特别地，对于 I/O 子系统而言，要访问的文件数据（包括文件的内容信息以及文件的元数据信息）应该尽量在主存而不是磁盘中才能提高 I/O 性能。然而，早期内存芯片的集成度有限兼之价格昂贵，计算机系统只能配置容量十分有限的内存作为主存，一般只以少量内存 cache 文件数据，而不可能用内存直接存储文件数据。

除了上述的容量和价格限制外，另一个阻碍内存存储文件数据的障碍来自于目前最常用的 DRAM 内存芯片固有的易失性。当系统发生电源故障时，如果内存中的数据在磁盘中没有相应的副本，那么就不能保证文件的持久性语义。否则，就必须像目前大部分磁盘文件系统所做的那样，真正的文件数据存放在磁盘上，仅将内存作为磁盘文件的 cache，内存中的数据以较短的时间间隔在适当的时机刷新到磁盘。在 1.2.3 小节我们将会给出以内存作为磁盘文件 cache 的不足。

以上障碍以及技术的惯性使得磁盘文件系统仍然是目前串行或者并行文件系统中的主流。

1.2.2.2 技术发展对内存存储文件数据障碍的弱化

从 20 世纪 90 年代中后期开始，随着计算机技术的发展，上述的一些情况已经发生了变化，阻碍内存存储文件数据的障碍与以前相比已经大大减小了，这也就使得我们必须重新考虑内存在缓解 I/O 瓶颈中所应该扮演的角色。

(1) 大容量和低价格的内存

从容量和价格上来看，微电子和半导体技术的发展使得内存

芯片在集成度、容量和价格上有了很大的改善。内存的价格已经相对较低，在计算机系统的整体价格中占的比重正逐步减小。

由于内存芯片的价格低廉以及集成度的提高，单台计算机或者并行计算机的单个结点配备 512M 乃至更多的内存已是很常见的现象。对于目前主流的并行计算机系统而言，系统的每个结点一般都配备了不少内存，随着系统规模越来越大，整个系统中的全部物理内存从容量上来说已经很大，表 1.2 给出了若干大规模并行计算机系统的物理内存配置情况。

表 1.2 若干大规模并行计算机系统的物理内存配置情况

系统名称	处理数目	物理内存总容量（TB）
Earth Simulator	5120	10
Thunder	4096	8.2
ASCI White	8192	6.2
ASCI Blue-Pacific SST	3904	1.9
ASC Linux Cluster	1920	3.8

实际上，并行计算机系统的物理内存不仅总容量巨大，而且在系统正常运行过程中的平均空闲容量也相当可观。1999 年加州大学 Acharya 等人的研究[7]表明：在非专用的 Cluster 环境中，对单个结点的内存而言，256M 内存平均有 180M ~ 200M 空闲，32M 有 12M ~ 19M 空闲，随着内存容量增加，空闲的比例也相应增加；对整个 Cluster 的内存而言，如果考虑所有结点，平均有 60% ~ 68%的内存空闲，如果仅考虑空闲结点，平均有 53%的内存空闲。

Acharya 等人虽然只针对非专用 Cluster 系统进行了研究，但

据此可以判断，对于配置更高的专用 Cluster 系统或者大规模并行计算机系统而言，系统中平均的空闲物理内存容量应该更大。将这些内存组织起来存储文件数据、加速 I/O 操作应该是合理的选择。

(2) 新的存储层次——网络内存

随着网络技术的进步，与磁盘的相应参数相比，无论是 MPP 中的互连网络还是 Cluster 中的专用网络[117~119]或者通用局域网络的带宽和延迟都获得了较大的改善。磁盘延迟每年仅降低 10%，带宽仅提高 20%，而网络延迟则每年降低 20%，带宽提高更达到 45%。图 1.2 给出了几种典型快速网络的单向传输时间[116]。

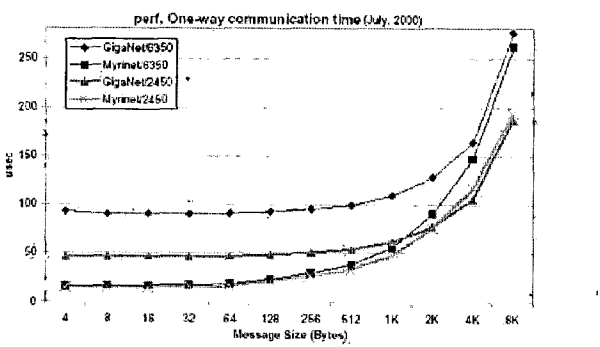

图 1.2 几种高速网络的 one-way 传输时间

网络技术的飞速发展使得网络的带宽和延迟都优于目前的磁盘设备，尽量使用系统中其它结点的内存而不是本地磁盘来存储数据将使 I/O 操作更加高效。这也使得分布并行计算机系统中出现了一个新的存储层次——网络内存[24, 116]。网络内存的访问延迟和带宽参数虽然比本地内存稍差，但大大优于磁盘的相应参数。通过一定的软硬件机制，它可以成为比本地内存更大、更可靠的存储介质。用它来存储本应存储在磁盘中的数据可以大大改