

ENBD:扩展的网络磁盘设备^{*})

张玉国^{1,2} 张建刚¹ 许 鲁¹

(中国科学院计算技术研究所 北京 100080)¹ (中国科学院研究生院 北京 100039)²

摘 要 网络磁盘设备(Network Block Device, NBD)是一种稳定、高效的开源网络存储软件,可被用于构建基于 Linux 平台的网络存储系统。网络存储系统一般较为复杂,包含的模块通常要兼有多种功能,而 NBD 只能完成普通数据的网络传输,缺少网络存储系统要求的集群化、扩展性、元数据操作等功能。该文描述如何扩展 NBD,使之成为适合蓝鲸集群文件系统的网络存储软件。

关键词 网络存储, 集群文件系统, 网络磁盘设备

ENBD: Extended Network Block Device

ZHANG Yu-Guo^{1,2} ZHANG Jian-Gang¹ XU Lu¹

(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)¹

(Graduate School of the Chinese Academy of Sciences, Beijing 100039)²

Abstract Network block device(NBD) is a stable and open source network storage software, which can be used to construct a network storage system based on Linux. Normally network storage system is complex, including miscellaneous modules. But NBD just transmits data over network and lacks the functions of clustering, scalability, special operations required by a network storage system. This paper outlines how to extend NBD to apply for Blue Whale File System.

Keywords Network storage, Cluster filesystem, Network block device(NBD)

1 引言

NBD^[1]是 Pavel Machek 于 1997 年开发的一个开源网络存储软件,可以将位于远端机器上的存储空间映射为本地的虚拟磁盘,提供给本地用户使用。该软件设计新颖,在开发之初就被融入到 Linux 标准内核中,后经众多开源爱好者地共同开发,NBD 越发稳定,并且逐步支持 SMP、64 位体系结构等特性。NBD 可以和本地磁盘一样使用,也可被应用于构建简单的基于 Linux 平台的网络存储系统。

基于 NBD 的以上特点,我们选择 NBD 作为蓝鲸集群文件系统^[2](Blue Whale File System, BWFS)网络磁盘设备的开发原型。由于设计 NBD 的初衷只是为了提供网络磁盘,其核心——NBD 协议只能完成普通数据的传输。在 BWFS 环境下,我们不仅需要 NBD 能完成普通数据传输,还要服务于整个蓝鲸集群文件系统,它要支持多个存储服务器、资源路径查询、元数据访问,这些特性是标准 NBD(标准内核自带的 NBD 软件)不具备的。本文就是描述如何扩展标准 NBD 以满足上述特性,成为适合蓝鲸集群文件系统的网络存储软件。

本文组织如下:第 2 节对 BWFS 和标准 NBD 做了简单介绍;第 3 节对如何扩展标准 NBD 做了详细分析;最后总结全文,给出下一步工作展望。

2 背景介绍

2.1 BWFS

BWFS 是由中国科学院计算技术研究所国家高性能计算机工程技术研究中心自主研发的用于解决系统 I/O 瓶颈的海量网络存储系统的核心系统软件,主要面向机群结构环境提供高性能、并发、共享、高可用的分布式文件访问服务。与传统网络文件系统(NFS、CIFS)的体系结构相比,BWFS 融合了 NAS 和 SAN 的优点,采用元数据与数据相分离的带外传输模式(out-of-band),增强了系统的可扩展性,消除了传统带内传输模式的数据传输瓶颈,并有效解决了传统 SAN 设备数据文件级共享问题。

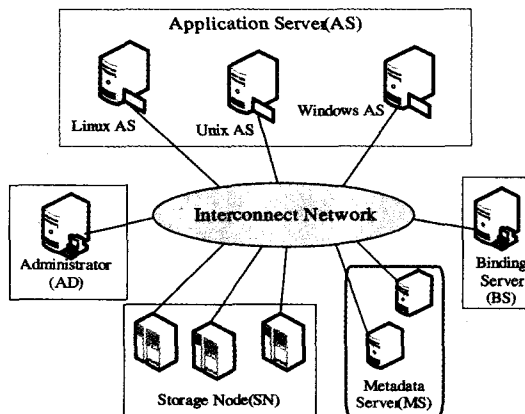


图 1 BWFS 系统架构

如图 1 所示, BWFS 的整体架构划分为应用服务器(Ap-

^{*})基金项目:国家“八六三”高技术研究发展计划基金项目(2004AA1Z2250)。张玉国 硕士研究生,主要研究方向为文件系统、网络存储;张建刚 博士,副研究员,硕士生导师,主要研究方向是分布式文件系统、海量存储等;许 鲁 博士,研究员,博士生导师,主要研究方向为操作系统、体系结构和网络存储。

plication Server, AS) 机群、元数据服务器 (Metadata Server, MS) 机群、存储节点 (Storage Node, SN) 机群, 绑定服务器 (Binding Server, BS) 和管理服务器 (Administrator, AD) 等几个部分, 其中 MS 机群负责处理 AS 机群的元数据请求, SN 机群负责处理 AS 机群的数据请求, BS 用于处理元数据的动态映射及负载均衡, AD 提供系统整体监控与管理。系统中各节点间彼此独立, 通过高速交换网络连接在一起。BWFS 支持系统的在线扩展, 用户可以根据系统的实际使用情况在线独立地扩展 AD 机群、MS 机群、SN 机群, 对系统进行整体均衡。

2.2 标准 NBD

NBD 的结构比较简单, 如图 2 所示。整个 NBD 采用 C/S 结构, 包括一个客户端驱动程序 (也叫 NBD 协议)、一个客户端控制软件和一个服务器端服务程序。客户端驱动程序虚拟出一个网络磁盘, 用户使用客户端控制软件连接服务器并且为网络磁盘设置参数 (比如和服务器的连接信息), 服务器端服务程序负责接受网络请求并把请求转发到挂载在 SN 上的本地存储设备, 客户端和服务端通过 TCP/IP 网络连接。

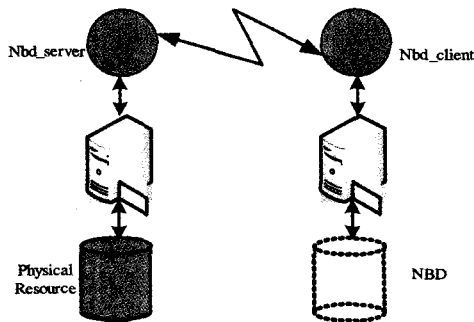


图 2 NBD 结构

NBD 可以完成很多应用。首先, NBD 可以和本地磁盘设备一样使用, 可以在上面创建并使用文件系统, 也可以当作原始设备 (Raw Device) 使用。其次, NBD 可被用于构建网络存储系统, 如远程磁盘镜像系统、远程数据备份系统等等, 文 [3] 描述了这类简单应用; 还可被用于一些较为复杂的网络磁盘阵列、分布式存储系统的研究工作, 如 NSS^[4] 系统, 中科院计算所的蓝鲸集群文件系统等等。

3 扩展 NBD 的设计与实现

根据 BWFS 的系统架构设计, 我们在 MS 端需要磁盘设备用于格式 BWFS MS 端本地的文件系统, 而我们又想使用 SN 提供的存储空间。标准 NBD 是一个较好的能实现我们设计意图的开发原型, 但是 BWFS 要求 NBD 能支持多个存储服务器、资源路径查询、元数据访问, 这些特性是标准 NBD 没有的, 我们需要就以上三个特性对其进行扩展。

在 BWFS 中, 我们需要扩展的三个特性均集中在 NBD 客户端驱动上, 因此标准 NBD 软件的其他两部分此处不再描述。

ENBD 扩展出的功能有如下几个部分:

3.1 支持多存储服务器

在 BWFS 中, AD 负责整个存储系统的逻辑资源管理, 包括逻辑资源的分配、回收和维护等功能。它具体负责两部分功能: 分配 SN 使用的逻辑资源和配置管理 MS、AS 的资源路由信息。这里的资源路由信息指的是资源 (block、inode、资源

组) 的位置信息, 包括对应的 SN 及其服务端口号等。

标准 NBD 支持一个 SN, 它只需要客户端控制程序把这个 SN 的 socket 链接传给内核态的 NBD 驱动即可开始传输。而在 BWFS 中, SN 是以集群方式来组织的, 当 ENBD 驱动需要传输数据时, 需要明确此次传输的目标 SN, 因此就需要做这部分扩展——查询 AD, 图 3 中的步骤 3、NO 和步骤 4 反应了这个查询过程。

待查询的信息有 block 号、inode 号、资源组号与 SN 的对应关系, ENBD 向 AD 发出一个大小为 12 字节的请求, 请求格式为:

type	num
------	-----

AD 会返回一个大小为 36 字节的响应:

type	num	ip	port	start	end
------	-----	----	------	-------	-----

其中 type 为查询类型, 为 0 时表示查询 block 或者资源组, 为 1 时是 inode; num 为待查询的 block 号或者 inode 号或者资源组号; ip 为待查询资源所在 SN 的 ip 地址; port 是运行在 SN 上 enbd_server 服务进程的服务端口; start 和 end 分别为该 SN 含有的资源空间。type 和 num 这两个域在请求和响应中完全相同, 可用于传输正确性校验, 响应结构的后四个域才是 ENBD 需要的传输目标信息。

3.2 资源路径查询

Cache 技术是计算机领域的一项重要技术, 能大幅度提高计算机系统的性能, 随着计算机技术的发展, 其应用领域大大扩展。在 BWFS 中, 资源路由信息也要 Cache。

3.1 节提到 ENBD 每次向 AD 发出请求, 都会收到一条响应描述对应 SN 的信息。这些资源路由信息是海量的, 而且响应也是大量重复的。因此我们设计了全局逻辑卷管理器 (Global Logical Volume management, GLVM) 模块并把它放置在 MS 端, 功能是维护 ENBD 从 AD 上查询得来的 SN 的相关信息。这样以后 ENBD 再次发出资源路由查询请求时, 首先查询 GLVM, 如果 GLVM 已缓存该信息, 就直接返回给 ENBD, 否则 ENBD 就需要访问 AD, 进而更新 GLVM 的缓存信息。

为了实现上述设计意图, 我们也需要扩展标准的 NBD。我们以图 3 中的若干步骤描述这些扩展, 该过程是 BWFS MS 端本地的文件系统需要写块号为 bno 的块时, 系统的资源使用流程。

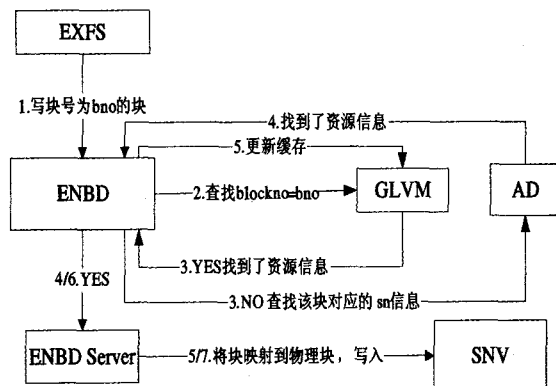


图 3 系统资源使用模型

3.3 增加元数据访问协议

文件系统需要管理磁盘资源,为用户提供一致的访问接口。在 BWFS 下,MS 管理系统的元数据,而这些元数据和普通数据一样存放在 SN 上;标准 NBD 已具备了普通数据的访问协议,还不具备元数据访问协议,因此我们需要为 ENBD 增加元数据访问协议。

我们称标准 NBD 提供的元数据访问协议为普通协议,普通协议与特殊协议可以通过 I/O 请求(request)的块号域加以区分,定义如表 1。

表 1 Enbd driver 协议块号定义表

块号(block)	用途
0—3	读写超级块
4	申请空闲 block 资源
5	申请空闲 inode 资源
6	读指定的 inode
7—17	保留块号
18	释放空闲 block 资源
19	释放空闲 inode 资源
20	写指定的 inode
21—25	保留块号
26	获取指定资源组的资源统计信息
27	获取系统所有资源组的信息
28	获取指定资源组的属性信息
29—127	保留块号
128—	普通数据的读写块号

ENBD 的普通协议与标准 NBD 协议一致,包括所有请求块号不小于 128 的读写请求,以及请求块号在 0-3 的超级块请求,其他位于 4-127 的协议号均为特殊协议。

ENBD 驱动和 enbd_server 之间网络请求 enbd_request 的结构如表 2 所示。

表 2 enbd_request 结构

成员	用途
magic	enbd 请求幻数标识, ENBD_REQUEST_MAGIC = 0x25609513
type	enbd 读写请求标识, 0 = READ 1 = WRITE
handle ^[8]	enbd 请求的序号标识,通常为请求的 request 内存地址
from	读写偏移,以字节计算
len	读写长度,以字节计算

响应 enbd_reply 的结构如表 3 所示。

表 3 enbd_reply 结构

成员	用途
magic	enbd 应答幻数标识, NBD_REPLY_MAGIC = 0x67446698
error	enbd 请求结果正确标识, 0 = ok, else error
handle ^[8]	enbd 请求的序号标识,来自对应请求的 handle 部分

有了网络请求和响应的结构,我们就可以具体地看一个特殊协议是如何使用它们来完成元数据访问的。块号为 4 的请求,目的是为了向 SN 申请空闲的 block 资源。其 enbd_request 请求的格式为:

```
magic = NBD_REQUEST_MAGIC;
type = READ;
handle = &req;
from = 4 * 512;
len = gid(资源组标识);
```

由于 request 的块号域是以扇区为单位的,因此上面格式中的 from 的大小应为块号域的值再乘上扇区的大小。SN 收到上述请求后,会返回一个响应包,如果握手正确,SN 会再向 ENBD 返回空闲的块资源。

ENBD 就是通过以上方式来实现对存储系统的元数据操作的,其他元数据访问协议实现方法类同。

总结 本文简单介绍了标准 NBD,阐明了由于其功能不完全满足蓝鲸集群文件系统的需求,需要对其进行扩展,随后重点分析了如何扩展该软件以支持蓝鲸集群文件系统对多存储服务器支持、资源路径查询、元数据访问三个特性的需求。上述工作对于利用开源软件进行复杂系统集成或开发有指导意义。

下一步的工作是分析在现有 NBD 扩展基础上蓝鲸集群文件系统的整体性能。

参考文献

- 1 Machek P, Network Block Device. <http://atrey.karlin.mff.cuni.cz/~machek/nbd/nbd.html>
- 2 杨德志,黄华,张建刚,许鲁.大容量、高性能、高扩展能力的蓝鲸分布式文件系统.计算机研究与发展,2005(6):1028~1033
- 3 Ares P T, The Network Block Device. <http://www.linuxjournal.com/article/3778>
- 4 Novell Storage Services File System Administration Guide for NetWare 6.5. <http://www.novell.com/documentation/nw65/index.html?page=/documentation/nw65/nss-enu/data/hn0r5fzo.html>
- 5 plexity[J]. 电子科学学报(英文版),2006,23(6):873~876
- 6 魏芳,李学明.基于 MMX 技术的 H.264 解码器优化[J].计算机工程与设计,2004,25(12):2218~2224
- 7 List P, Joch A, Lainema J, et al. Adaptive Deblocking Filter[J]. IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, 2003, 13(7):614~619
- 8 Lappalainen V, Hallapuro A, Hamalainen T D. Complexity of Optimized H.264 Video Decoder Implementation[J]. IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, 2003, 13(7):717~725
- 9 Bryant R E, O'Hallaron D. Computer Systems A Programmer's Perspective[M]. America: Pearson Prentice Hall, 2003
- 10 Intel Wireless MMX Technology Developer Guide[M]. America: Intel Corporation, 2002
- 11 Lindroth T, Avessta N, Teuhola J, Seceleanu T. Complexity Analysis of H.264 Decoder for FPGA Design[J]. ICME, 2006: 1253~1256
- 12 Lee J, Moon S, Sung W. H.264 Decoder Optimization Exploiting SIMD Instructions[C]. IEEE Asia-Pacific Conference on Circuits and Systems, 2004

(上接第 61 页)

结束语 本文介绍的解码器实现方案适用于运行 Windows Mobile 5.0 平台的 PPC 和 Smartphone 设备, WMMX 优化方案仅适用于支持 Intel® WMMX 指令集的 PLA27x 处理器系列。该方案针对嵌入式移动视频监控系统的要求而研制,同样适于手机视频点播和手机电视应用的视频解码器优化。

参考文献

- 1 Richardson I E G. H.264 and MPEG-4 Video Compression: Video Coding for Next-generation Multimedia [M]. America: John Wiley & Sons, Ltd, 2003
- 2 刘鹏飞,刘志,安平,等. H.264/AVC 解码器优化的研究[J]. 中国图象图形学报,2006,11(11):1627~1630
- 3 Xi Yinglai, Hao Chongyang, Lai Changcai. Fast Sub-pixel Motion Estimation Techniques Having Lower Computational Com-