# 一种动态连接存储资源和计算资源装置的设计与实现*)

刘 宇[1,2] 施家元[3] 刘海洋[1,2]

(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]

(北京印刷学院教学技术与网络中心 北京 102600)[3]

**摘 要** 存储与计算的分离,重新定义了计算机的使用模式。计算资源和存储资源的动态重构,不仅提高了资源的利用率,而且简化了管理。一种实现这种新模式同时又兼容于传统机制的方案是在存储设备级截获数据流,并转接到网络,利用现在网络高带宽、可靠和灵活的特性提供高速、稳定和动态的服务。本文将介绍动态网络硬盘 nHD(network HardDisk)的设计和实现,给出并分析其运行性能。

**关键词** nHD,存储和计算分离,动态计算环境

## The Design and Implementation of Device which Dynamically Connects Storage and Computation

LIU Yu[1,2] SHI Jia-Yuan[3] LIU Hai-Yang[1,2]

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

(Graduate School of the Chinese Academy of Sciences, Beijing 100039)[2] (Beijing Institute of Graphic and Communication, Beijing 102600)[3]

**Abstract** The separation of storage and computation redefines the using model of computers. And the ability of reconnecting computation resources and storage resources not only improves the utilization of resources, but also simplifies the management. In this sense, we develop a method to implement this new model compatible with the traditional mechanism, which intercepts data stream onto networkat the storage device level, thereby taking advantage of high-bandwidth, reliability and flexibility of networking to provide high-speed, stable and dynamic accessing services. In this paper, we give a detailed introduction of the designand implementation of the dynamic network hard disks nHD, and present the actual testing and relative analysis.

**Keywords** nHD, Separation of storage and computation, Dynamic computing environment

## 1 引言

随着当今信息量爆炸性的增长[3],人们对存储的要求也在不断增长,然而传统的 DAS(直连式存储 Direct Access Storage)模型逐渐无法跟上信息时代存储发展的需要,主要体现在:①人们对数据服务的高性能、高扩展和高可用的要求大大提高,而存储和计算固定的模式缺乏灵活性和在线扩展能力;②日益庞大而又分散的信息造成资源利用的低效,同时也增加了管理的开销,人们急切需要更高效节省的存储管理模式;③如何实现高效的信息共享,是互联网时代最值得关注的问题,传统 DAS 模型只能通过副本拷贝进行信息共享,不仅低效而且难于维护一致性[1,2]。

造成这些问题的根本原因在于计算资源和存储资源本质上有较大差别。特别是计算资源作为人们日常使用的工具,其发展倾向于分散;而存储资源作为信息的载体,其发展则倾向于集中。传统计算机 DAS 模式是以计算资源和存储资源紧耦合方式构建而成,这种紧密的绑定关系,使得两种资源特性的相互冲突最终演变成了发展的相互束缚。因此,这些问题最根本的解决方式是实现计算资源和存储资源的分离和动态组合。

国家高性能计算机工程技术中心自主研发的 nHD(Network Hard Disk)系统,打破常理,在存储与计算分离的理论基础上,创造性地通过 NHD 系统的网络硬盘通道(nHD 卡)对存储资源和计算资源进行了彻底的分离,保证了按需动态构建计算资源和存储资源的能力的同时又兼容于传统使用机制,实现了集中存储而分散动态使用的存储模式,提高了资源利用效率,解决了信息分散带来的管理弊端。

本文第 2 节介绍相关研究的情况;第 3 节介绍 nHD 卡的总体框架、各模块的组成以及其各方面的优点和特性;第 4 节给出具体的性能参数并分析;最后总结全文,并介绍今后研究工作。

## 2 相关研究

网络软硬件平台技术的飞速发展和成熟,为存储和计算的分离的实现提供了合适的基础平台,对已有技术稍加整合即可构建高速稳定的资源动态构建系统。从存储市场的发展来看,目前已有多种不同类型的产品。

NC(网络计算机,Network Computer)并未分离存储资源和计算资源,但它利用网络动态绑定客户和资源。资源的使用方式较灵活,利用率也较高。但 NC 客户端需要专用设备的支持,中央服务器通常需要较高的成本,客户无法随意选择操作系统,而且 NC 很难提供给客户完全个性化的系统数据。

典型地分离出存储资源的系统是 NAS 和 SAN,它们也是通过高速网络动态重构存储资源和计算资源,由于分别基

*)本课题得到国家"973"基金项目(2004CB318205)资助。刘 宇 硕士研究生,主要研究方向为网络存储。施家元 在职硕士研究生,高级工程师,主要研究方向为网络技术、软件工程。刘海洋 硕士研究生,主要研究方向为为网络存储。

于文件级的访问和基于块级的访问，因而互有优缺点却又正好相互补充。但目前 NAS 和 SAN 的主要使用方式是分离出用户数据，由于未分离系统数据，完成初始配置后基本不再改动映射关系，因而只是做到了物理分离而非逻辑分离。

无盘工作站理论上可以做到资源物理上和逻辑上的分离，目前的无盘工作站通常直接复用网络适配器，这种方式有几点不利因素：①过分依赖于宿主机的具体软硬件，这导致实现过程中不得不采取折中的手段，从而引入一些问题；②如果网络连接外网，则会存在潜在的安全问题；③数据的传输不得不占用客户端的处理器资源。种种限制尤其是第一点使得复用网络适配器的工作站至今没有较理想的产品，它们多数设计简单，且多采用静态映射方式，并未从逻辑上分离存储资源。

nHD 系统也是一种无盘工作站，但它选择在存储设备级别进行数据的截获，因而透明于磁盘控制器和上层操作系统，实现上灵活而不受限制；由 nHD 提供网络协议的卸载，无需占用主机 CPU 资源；而且用户可以选择使用专用网络，提高了安全性。nHD 作为一个通用的数据通道装置，巧妙无痕地分离了存储与计算。

## 3 系统设计

### 3.1 nHD 原型和结构

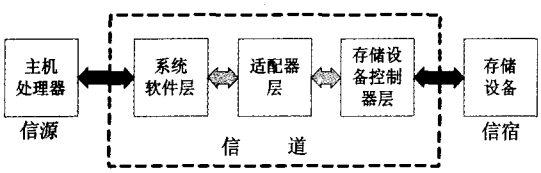


图 1 传统计算机系统数据存储通道的模型图

由于 X86 体系结构计算机使用相当广泛，在 PC 中占据很大比例，本文将以 X86 体系结构作为宿主机介绍 nHD。图 1 所示，是计算资源到存储资源的数据通路，主机处理器和存储设备分别处于存储通路的两端，相当于通信系统中的信源和信宿，而位于两者之间的系统软件层、适配器层和存储设备控制器层相当于通信系统中的信道。通过对该信道中某一层次的层间接口进行功能模拟，就可以获取对数据的控制权。

如 NFS、NBD 等技术都是从系统软件层截获数据的控制权，这些技术用于访问用户数据都已有较好的模块化实现，但将它们用于系统启动时访问系统数据则比较困难。这是因为 X86 体系结构发展过程中的各种历史遗留问题造成启动过程非常复杂，而且不同操作系统启动流程也差别很大，各种模块相互依赖性不同启动顺序也不一致，所以启动时在系统软件层截获数据不仅实现起来束缚很多，更严重的是不具备通用性，且对于操作系统升级也十分敏感。若在适配器层进行模拟，则需要芯片级的设计和控制，实现起来需要一定工作量，同时功能被放在平台相关的核心组件上，不利于兼容不同硬件平台。

因此 nHD 选择在存储设备控制层进行数据的截获数据，这样作为一个外设连接入主机，可以透明于上层操作系统和芯片组，屏蔽软硬件各种差异，从而获得最大的通用性和兼容性。无论软件硬件是何种架构，只要支持该类型外设，即能够完全透明地使用，无需任何改动。另一方面，在成熟的标准接口上开发外部设备也有相当多的经验可以借鉴，实现成本不高。

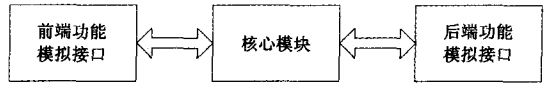


图 2 nHD 卡逻辑功能图

从功能的角度看，nHD 分为三个部分：前端的功能模拟接口，后端的功能模拟接口和动态连接支持模块，如图 2 所示。前端对主机模拟 IDE(一种磁盘控制接口 Integrated Device Electronics) 接口并遵循标准 ATAPI 协议，以能够将 nHD 伪装成一个标准 IDE 硬盘。IDE 是目前 PC 机使用最为广泛的存储接口，选择 IDE 可以获得最广泛的支持。我们用 FPGA 编程来构建硬件状态机，该状态机负责在主机的 IDE 控制器和 nHD 卡处理器之间进行控制信息和数据的传递。之所以引入 FPGA 这样一层，一是可以将整个 IDE 行为和 nHD 处理器外部设备的行为进行互相转换，烦琐的 IDE 命令应答控制和数据传输若全由软件实现则会相当烦琐，通过硬件转换成标准的中断和 DMA 操作，方便了程序控制。第二，硬件具有软件达不到的迅速响应能力。比如 FPGA 上电可立即投入工作，可以弥补软件启动慢的缺陷。X86 宿主机上电后会马上检测硬盘，这就需要硬件来提供迅速响应，进行通知并等待软件就位后处理。FPGA 的迅速加载复位能延缓主机检测，为 nHD 系统的启动以及建立网络连接争取了时间。第三，FPGA 可动态加载的优势使得与主机通讯的硬件部分具备可重构可升级的特性。

nHD 后端在标准以太网接口上采用 NBD 协议来动态连接访问存储设备，将 nHD 伪装成一个以太网客户端。以太网是当前最流行的网络架构，性能优异而且价格低廉。NBD 是基于 TCP/IP 上的网络块级设备访问协议，特点是简单且高效。我们用 PHY 芯片搭配集成在 SOC 处理器中的 MAC 控制器来处理网络物理层和 MAC 层，而上层协议的卸载则交给软件完成。为了具有较强的实用性，在后端增添一个网口和一块交换芯片，使得 nHD 的 2 个网口具备交换功能，主机的网卡接入 nHD 另一个网口就能和 nHD 卡同时连入网络。这样提供了网络的复用使用方式，而不需要增加任何网络设备。

nHD 核心部分由处理器和系统软件构成，主要包括 2 个模块：动态映射模块和数据传输模块。Linux 系统多方面优势促使我们选择它作为系统软件平台：①功能强大且可配置，实现了协议栈，相比专用系统更具灵活性和可扩展性，利于今后的功能扩充；②开发的软件有较强的移植性，利于今后移植到不同硬件平台；③性能优异、运行稳定，且版本以固定周期升级。软件以固件形式保存在 FLASH 中，运行时载入内存，因而整套软件具备可升级能力。

### 3.2 动态映射模块

将网络纳入数据通路后可以利用网络的动态性构建动态映射的存储系统，通过建立不同的 NBD 连接可将 nHD 映射到不同的存储空间。动态映射使得存储资源和计算资源可以进行随意组合，它是存储与计算分离的基础。动态映射机制模块，为整个存储计算提供了分离机制。从此不再是使用者同具体硬件绑定的那种传统使用模式，取而代之的是使用者与自己的数据绑定。这样使用者可以用自己的硬盘启动而又不局限于使用固定的机器，不仅提高了硬件资源的利用率，同时为后端对存储的灵活管理提供了基础。

不同于在客户机系统软件层截获数据，在存储设备层截

获数据，需要整个动态映射过程对客户机完全透明。我们采用公共启动盘的方法实现，整个动态连接机制如下：①客户机上电后，nHD通广播方式找到管理服务器；②nHD先映射到指定的公用虚拟硬盘，客户机于是以正常启动流程从该公用虚拟盘读入并运行一段启动程序，该程序同服务器交互，获取服务列表并提供给用户。例如用户可通过帐号登录并获取自己私有可使用的虚拟硬盘列表，在不同的虚拟硬盘之间进行选择等等；③选择具体硬盘后，管理服务器会根据用户的选择通知具体SN导出该虚拟硬盘，然后通知nHD重新进行映射；④重启宿主机，系统就会从重新映射后的网络硬盘启动。

### 3.3 数据传输模块

nHD卡本质上是一个数据通路，通道的两端是网络和IDE通道，它们的带宽决定了整个nHD通道的带宽，因此我们需要做的就是保证nHD通道的低延时。

整个数据路径由硬件软件协同控制。FPGA和MAC控制器都通过DMA方式操作内存进行数据传输，以减少CPU负担。软件主要进行协议转换和对硬件的调度。整个数据通路如图3所示。图中淡灰色双向箭头显示了前后端硬件到内存的DMA通路。深灰色dma箭头是我们进行的优化措施。由于网络是将数据分割打包进行传输，直接调用Linux内核的协议栈，所有数据会被拷贝到指定的缓冲区(buffer)中，这次拷贝完全由处理器进行，同时顺便完成对整个TCP报文的奇偶校验。所有经过的数据都需要进行这种不必要的拷贝，不仅浪费了处理器资源，而且增加了数据传输的延时，降低了服务质量[4]。

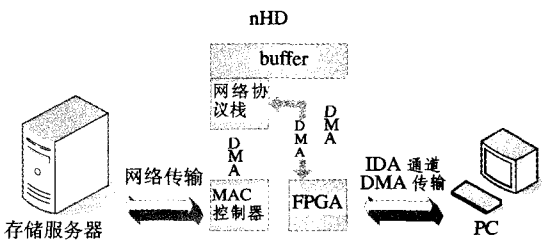


图3 nHD卡数据传输

我们降低延时的方法是在不影响协议栈正常工作的前提下，为数据传输开辟一条捷径，直接从协议栈进行DMA的传输，并将发送和接受的数据的校验工作分别交给FPGA和MAC硬件执行。这个方法不仅将处理器解放出来，更巧妙的是，由于网络的数据分割成包传输，这使得每一个包完成前一操作后就能启动后续操作，于是很自然地形成了数据传输的流水线，硬件上使处理器、2个DMA得到了充分的并行，极大地降低了延时。限于篇幅，这部分的详细工作将在其他文章中介绍。

### 3.4 其它特性

nHD独立于宿主机的代理身份还有利于进行资源监控：①可以作为主机资源管理代理，管理服务器能通过nHD对宿主机进行资源监控；②提供对自己的管理、配置以及重构的接口，nHD自身具备的升级特性，可随时方便地调整和扩展功能；③能方便地获取I/O访问统计信息，具有一定科研价值。

## 4 性能评价

nHD卡目前已经在两个不同的硬件平台上实现。一个采用ARM Samsung S3C2500芯片，处理能力较弱，数据cache关闭，运行经由Linux内核代码裁减而成的uClinux-2.4.x。另一个使用MIPS gs32i芯片，处理能力强，数据cache打开，内核使用Linux-2.4.18。应用程序和驱动程序则是各个平台共用一份。我们在接近理想情况的环境中进行了测试：服务器配置为pentium 4 1.5GHz、512MB内存、80G硬盘，通过百兆环境和单个nHD客户端直连。由于读具有较强的实时要求，在实际使用中对应用程序性能影响较大，因而我们重点测试和分析了读的性能。处理能力相对较强的gs32i表现出更优异的性能，各接口和总体读性能如下：

| 平台 | TCP/IP | NBD(64kB) | IDE接口性能 | 总体通道性能 |
|---|---|---|---|---|
| gs32i (MB/s) | 11.71 | 9.37 | 14.55 | 8.45 |
| s3c2500 (MB/s) | 5.14 | 4.66 | 14.78 | 4.03 |

同时我们使用Agilent 16903A逻辑分析仪对影响性能的各个因素进行了更微观的分析。使用2MHz的采样频率(500ns采样一次)跟踪了数据在协议栈、dma等各个环节的延时。

测试结果表明，s3c2500平台性能落后于gs32i的主要原因是处理器性能过低。我们发现，处理器性能过低不单只是线性增加流程的处理时间，当网络包的到达速率大于处理器的处理能力时，还会造成原有流水处理操作串行化，导致性能进一步下降。

总的来说只有采用处理能力合适的芯片才能充分发挥和利用网络带宽。虽然目前nHD的性能不到10M，但这已达到20世纪90年代初高端硬盘的性能，对于I/O负载不大的普通桌面应用而言几乎没有差别。随着网络性能和带宽的进一步发展，在千兆、万兆的网络环境中实现无疑可以超越目前硬盘的性能。

**总结和未来工作** 全文介绍了网络硬盘通道nHD卡的灵活且广泛兼容的软硬件接口，可移植性强的软件模块，低延时的数据传输机制，以及动态连接机制和高度可控的资源管理机制。随后给出了在2个不同平台上nHD实现的各项性能参数，并对数据进行了对比和分析。

nHD使得主机的存储模式发生了巨大的变革，摆脱了原有使用模式的许多缺陷，但是目前nHD的实现在性能方面做得还不够。未来的工作中，我们会考虑在千兆甚至更快的网络环境中存储动态映射的实现方式。在后端SN上，一方面，我们将研究由主机内存、nHD卡和SN构成的多级cache存储系统的行为，降低数据在SN上的延时，增强数据通路整体的流水性；另一方面寻找解决在多用户并发情况下SN磁盘I/O成为瓶颈问题的合适方法。

## 参 考 文 献

1 马一力，傅湘林，韩小明，等. 存储和计算的分离[J]. 计算机研究与发展，2005，42(3)：520～530

2 刘振军，许鲁，尹洋. 蓝鲸SonD动态服务部署系统[J]. 计算机学报，2005，28(7)：1110～1117

3 SIMS of UC Berkeley. How Much Information. http://www.sims.berkeley.edu/how-much-info/,2000-11-10

4 Wang S, Su J S. A survey of technology for TCP acceleration [J]. Journal of Software, 2004,15(11):1689～1699