【案例正文】

网络海量数据存储技术的应用

曾群 刘春年 屈文建 程晓

案例涉及的知识点：网络存储技术的基本概念、网络存储架构、容灾备份技术、常用设备及应用设计

案例真实性：真实

案例来源：来源于江西银行、南昌移动网络存储系统和海康威视视频存储系统的建设

摘要：网络存储技术(Network Storage Technologies)，就是以互联网为载体实现数据的传输与存储，数据可以在远程的专用存储设备上，也可以是通过服务器来进行存储。常用网络存储架构有直连附加存储、网络附加存储、存储区域网络等。本案例介绍了江西银行、南昌移动网络存储架构，海康威视视频云存储系统；对江西银行和南昌移动、海康威视视频云存储系统的实际网络存储架构进行了介绍。

Abstract：Network Storage Technologies is a way to realize the transmission and storage of data with the Internet as the carrier. The data can be stored on the remote dedicated storage device or through the server. Common network storage architectures include direct connection additional storage, network additional storage, storage area network, etc. This case introduces NanChang bank, NanChang mobile network storage structure, haikangwei visual video cloud storage system. This paper introduces the actual network storage structure of NanChang bank, NanChang mobile and hikvision video cloud storage system.

关键词：网络存储技术；DAS；NAS；IP-SAN；云存储

1引言

随着多媒体传输、数字图书馆、网络银行、网络电子交易平台等应用以大量集中式的数据实现，存储数据的容量已经开始从原来的B(字节)、KB、及MB急速增长成了GB，TB到乃至PB级的量级。企业以及互联网数据增长速率超过往年的50%，大数据时代已经到来。在此发展的基础上，用于存储大量用户信息的数据存储设备将不再作为服务器后端的辅助设备，它已经衍生为互联网时代中各个大中小企业用户作为最主要的成本支出。基于网络的数据存储已经发展成为继互联网和计算机产生之后的又一人类社会走向文明的标准。

国际数据公司（International Data Company，IDC）从四个方面定义了大数据，即海量的数据规模（Volume）、多样的数据类型（Variety）、快速的数据流转和动态的数据体系（Velocity）、巨大的数据价值（Value）。

作为一种信息资源，数据的重要性越来越突出。在如今高速发展的互联网时代，各国所有大中型企业及民营企业（如科研机构、电信部门、卫生医疗组织机构、金融行业等），其企业内部及企业间对其利用计算机和互连网络来实现数据传输处理业务数据将更加依赖。对于企业用户来讲，如果数据发生丢失或中断，由此产生的损失将是无法估量的，对小型的企业来说更将是毁灭性的打击，提高企业内部数据的可靠性对于企业来说是至关重要的。

在大数据时代，海量的数据所带来的一个巨大的挑战就是数据的存储问题，要想获得海量数据所带来的价值，首先要对数据进行存储，然后才能进行下一步的数据分析及价值提取。然而，数据的增长速度远远高于存储介质的扩容能力，要想单个存储服务器的存储能力满足数据的增长速度，所带来的人力消耗及资金消耗是难以承担的。此外，海量数据对现在的存储系统提出的诸如数据查询需求、容错纠错需求、安全性需求等都是大数据研究的重点难点。

2背景介绍

传统的分散式存储结构所存储数据分散，格式不统一，在传输中又会造成网

络拥挤等问题，渐渐难以满足实际使用的需求，也增加了企业运作成本。

随着数据存储系统规模越来越大，以及高清视频的大规模应用，存储系统中需要存储的数据和应用的复杂程度在不断提高，且视频数据需要长时间持续地保存到存储系统中，并要求随时可以调用，对存储系统的可靠性和性能等方面都提出了新的要求。在未来的复杂系统中，数据将呈现爆炸性的海量增长，提供对海量数据的快速存储及检索技术，显得尤为重要。

面对百PB级的海量存储需求，传统的SAN或NAS在容量和性能的扩展上会存在瓶颈。而云存储可以突破这些性能瓶颈，而且可以实现性能与容量的线性扩展，这对于追求高性能、高可用性的企业用户来说是一个新选择。

本案例主要即是讨论各种网络存储技术及相关协议在江西银行、南昌移动、海康威视视频存储系统上的应用，以及数据容灾备份技术等内容，通过对比不同存储技术在实际存储系统中的应用，让学生更直观地理解网络存储相关的知识点。

3理论基础及主要内容

3.1 网络存储相关概述

3.1.1网络存储技术概述

在传统的存储结构中，存储子系统是附属于主机的，它一般由磁盘阵列、光盘库、磁带库等组成大容量的存储空间，通过并行I/O通道与主机相连，如图1所示，然后经网络接口与LAN、WAN或MAN连接起来，这就像建立了一个数据传输的管道，数据从网络送到主机内存，接着送入到存储于系统。反之，则从存储子系统到主机内存，接着送入到网络。在这种方式下，能否正常访问存储的数据主要依靠服务器的可用性及网络的数据流量。这种基于网络的存储系统一船采用TCP/IP或SCSI协议。并且，远程用户在发送数据包时也不得不忍受LAN服务器和通过TCP/IP网络所带来的大的时间延迟。另外，网络和SCSI总线都不能实现容错功能，如果一个SCSI驱动器或一个LAN的连接导致服务器出错，或者SCSI服务器需要维护，这时，用户将不能进行数据存取访问，因此，服务器—磁盘这种存储方案正逐渐被网络存储所替代。

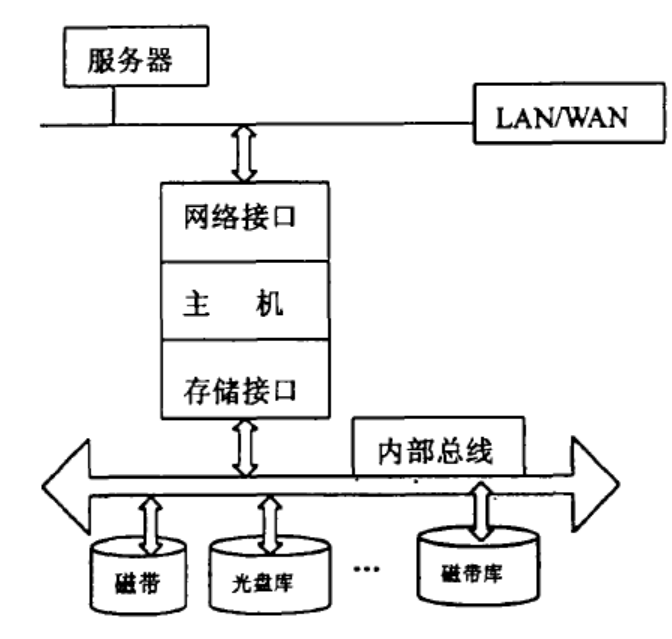


图1 传统的计算机存储结构

网络存储技术(Network Storage Technologies)，就是以互联网为载体实现数据的传输与存储，数据可以在远程的专用存储设备上，也可以是通过服务器来进行存储。网络存储技术是基于数据存储的一种通用网络术语。实际上，我们可以将存储技术分为三个阶段：①总线存储阶段；②存储网络阶段；③虚拟存储阶段。以存储网络为中心的存储是对数据存储新需求的回答。它采用面向网络的存储体系结构，使数据处理和数据存储分离；网络存储体系结构包括了网络和I/O的精华，将I/O能力扩展到网络上，特别是灵活的网络寻址能力，远距离数据传输能力，I/O高效的原性能；通过网络连接服务器和存储资源，消除了不同存储设备和服务器之间的连接障碍；提高了数据的共享性、可用性和可扩展性、管理性。

3.1.2网络存储架构类型

目前网络存储架构大致分为三种：直连附加存储、网络附加存储、存储区域网络。这几种网络存储方式特点各异，应用在不同的领域。

3.1.2.1直连附加存储

直接网络附加存储(Direct Attached Storage，DAS)是指将存储设备通过SCSI接口或光纤通道直接连接到服务器上的方式。这种连接方式主要应用于单机或两台主机的集群环境中，主要优点是存储容量扩展的实施简单，投入成本少，见效快。网络上的任何客户端要访问某存储设备上的资源时必须经过服务器。由于连接在各个节点的服务器上的存储设备是独立的，因此整个网络上的存储设备其实是分散、独立而难以共享的。由于所有的数据流必须经过服务器转发，因此整个服务器的负担比较重，也将是整个系统的瓶颈。其具体架构如图2所示。

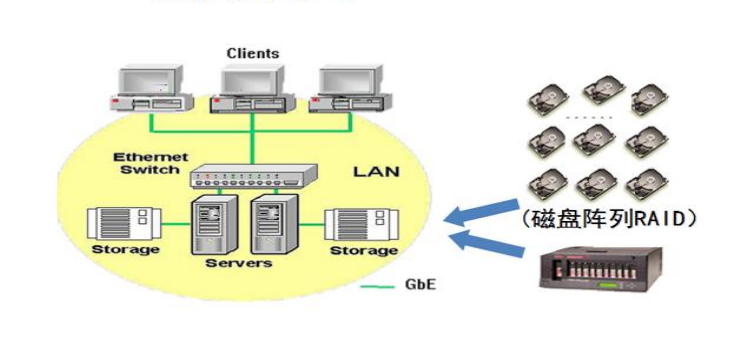


图2直接存储系统DAS架构

DAS是较早而且比较成熟的存储技术，可以使用与之连接的服务器上的操作系统来管理存储设备，操作系统具有独立性，存储设备和服务器可以分别购买，具有一定的价格优势。每个服务器或客户端连接自己的DAS设备，使得DAS设备在安装的时候不必考虑其他服务器或客户端对其的影响，网络架构有固定的模式可以套用，至今仍被很多用户继续使用。但是，DAS 构架下，数据的读写都依赖服务器，数据量增长后，响应性能下降；DAS 的架构决定了其很难实现集中管理，整体拥有成本较高，没有中央管理系统，数据的备份和恢复需要在每台服务器上单独做；不同的服务器连接不同的磁盘，相互之间无法共享存储资源，容量再分配很困难；DAS连接方式导致服务器和存储设备之间的连接距离有限制；因此，其对于小规模的运用比较合适。

3.1.2.2网络附加存储

网络附加存储(Network Attached Storage，NAS)是一种将分布、独立的数据整合为大型、集中化管理的数据中心，以便于对不同主机和应用服务器进行访问的技术。NAS中服务器与存储之间的通信使用TCP/IP协议，数据处理是“文件级”。NAS可附加大容量的存储内嵌操作系统，专门针对文件系统进行重新设计和优化以提供高效率的文件服务，降低了存储设备的成本。NAS被定义为一种特殊的专用数据存储服务器，包括存储设备和内嵌系统软件，可提供跨平台文件共享功能。NAS通常在一个LAN上占有自己的节点，无需应用服务器的干预，允许用户在网络上存取数据，在这种配置中，NAS集中管理和处理网络上的所有数据，将负载从应用或企业服务器上卸载下来。NAS架构中与传统的服务器为中心的存储系统相比，数据不再通过服务器内存转发，直接在客户机和存储设备间传送，服务器仅起控制管理作用，因而起到更快的响应速度和更高的数据带宽。具体架构如图3所示。

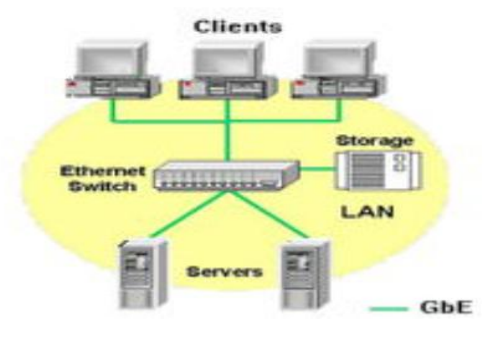


图3网络附加存储NAS架构

在 NAS 构架下，客户机对资源的请求直接发送给 NAS 设备，该请求返回的数据也直接发给请求者而无需服务器的中转，大大提高了响应速度和传输速率。NAS 采用标准网络接口，只要将NAS 设备连接到 TCP/IP 网络设备即可。服务器通过 FILE I/O 方式发送文件存储请求到存储设备 NAS，NAS 上一般安装有主机的操作系统，它将 FILE I/O 转换成 BLOCK I/O ，发送到内部磁盘。

3.1.2.3存储区域网络

存储区域网络（Storage Area Network，SAN）指存储设备相互连接且与一台服务器或一个服务器群相连的网络，其中服务器用 SAN 的接入点，通常由RAID阵列连接光纤通道组成，SAN和服务器以及客户机的数据通信通过SCSI命令而非TCP/IP,数据处理是“块级”。存储区域网络技术的出现，改变了传统的存储设备和计算机之间的联系方式。它将存储设备从网络分离出来，使它成为计算机网络的对等体。换言之，SAN技术将存储设备重新组合为网络的组成部分，而不是将它们视为特定的集成系统中的一部分。SAN是一种特殊的高速网络，连接网络服务器和诸如大磁盘阵列或备份磁带库的存储设备，SAN 置于 LAN 之下，而不涉及 LAN。利用 SAN，不仅可以提供大容量的存储数据，而且地域上可以分散，并缓解了大量的数据传输对局域网的影响。SAN 的结构允许任何服务器连接到任何存储阵列，不管数据置放在哪里，服务器都可以直接存取所需的数据。具体架构如图4所示。

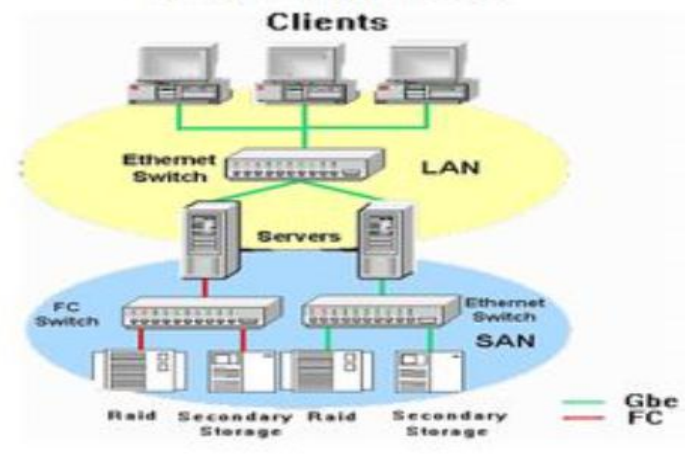


图4存储区域网络SAN架构

SAN是一种面向网络的、以数据存储为中心的存储结构，它采用可扩展的网络拓扑结构连接服务器和存储介质，因而把数据的存储、管理从服务器中独立出来，形成专有的存储网络。SAN面向服务器提供数据存储服务，服务器实现存储网和应用网间的连接与隔离。由于网络连接使服务器和存储设备之间具有多路、可选择的数据交换能力。使存储设备从服务器附属中分离出来，独立通过网络与服务器相连，消除了原来存储结构在可扩展性和数据共享方面的局限性。

3.1.3云存储

云存储是指通过集群应用、网格技术或分布式文件系统等功能，将网络中大

量各种不同类型的存储设备通过应用软件集合起来协同工作，共同对外提供数据存储和业务访问功能的一个系统，保证数据的安全性，并节约存储空间。简单来说，云存储就是将储存资源放到云上供人存取的一种新兴方案。使用者可以在任何时间、任何地方，透过任何可连网的装置连接到云上方便地存取数据。

与传统的存储设备相比，云存储不仅仅是一个硬件，而是一组由网络设备、存储设备、服务器、应用软件、公用访问接口、接入网和客户端程序等多个部分组成的复杂系统。各部分以存储设备为核心，通过应用软件来对外提供数据存储和业务访问服务。

3.1.4容灾备份技术

容灾备份是指对信息系统和数据进行备份，并能够在发生灾难后及时恢复业务的过程。灾备的目的是确保数据安全，保证关键业务持续运行。对灾备进行衡量的主要指标有恢复时间目标恢复点目标RPO（Recovery Point Objective）和RTO （Recovery Time Objective )。最完美的灾备方案是RTO和RPO均为0，表示当灾难发生时，业务完全不中断并且数据完全不丢失。RTO和RPO越小，业务系统的可用性越高，但是成本也越高。企业应针对不同类型的业务和数据设定不同的RTO和RPO目标，结合对现有生产环境改造的影响，选择相应的灾备方案。常用的容灾备份技术有以下几种：

1. 基于磁介质的灾备技术：这是传统的灾备技术之一，一般采用定时备

份的方式，把数据备份至磁介质(磁盘、磁带)，并把介质运输到异地保存。磁介质方式的缺点:如果系统数据量较大则需要大量磁介质、介质容易损坏、备份恢复操作复杂、需要大量人工操作；

1. 基于应用程序的灾备技术：在设计应用程序时就考虑对多份数据存储

的读写，灾备中心有一份与生产中心完全一致的数据，甚至双中心应用系统同时运行并对外服务。本技术的缺点：程序开发难度大，异地部署产生的延时对系统性能影响较大；

1. 基于数据库的灾备技术：通过数据库系统的复制功能实现数据备份，

例如Oracle Data Guard，Golden Gate等，该技术成本较低，应用较广；

1. 基于存储虚拟化技术：在存储层保证不同的存储设备中数据一致，与

上层操作系统、数据库、应用系统的架构无关，通用性好，不论是Windows还是Linux，Oracle或是MySQL都同样支持。

3.2 江西银行信息中心存储区域网络

3.2.1江西银行信息中心存储区域网络技术的选择

生活中常用的存储网络技术主要有NAS和SAN。NAS是通过IP网络访问的存储系统，采用NFS或者CIFS协议（NFS是Unix协议，CIFS是Windows协议）。这两种协议实际上都可以在文件层面对数据进行操作，所以在进行改动数据时，就是在改动整个文件。这有助于编辑改动不大频繁的电子表格的用户，不过并不适合大规模数据库，因为这种数据库可能会有容量达到数百兆的诸多文件。对银行集中存储这种大规模需求而言，NAS明显不适用。另外，NAS无法支持自动负载均衡，难以实现自动故障切换，这不能满足银行高可用性的要求。

SAN是一个由存储设备和系统部件构成的网络，所有的通信都在一个与应用网络隔离的单独的网络上完成，可以被用来集中和共享存储资源。SAN不但提供了对数据设备的高性能连接，提高了数据备份速度，还增加了对存储系统的冗余连接，提供了对高可用群集系统的支持。简单地说，SAN是关联存储设备和服务器的网络。它和以太网有类似的架构，以太网由服务器、以太网卡、以太网集线器/交换机及工作站所组成，SAN则由服务器、HBA卡、集线器/交换机和存储装置所组成。不过，SAN的配置在结构上往往更为严密，其设计考虑到了冗余性，并且尽量减少主机和磁盘之间交换机连接的数量。SAN有着极好的灵活性、安全性、扩展性等优势，而且技术已相对成熟，一些国内外大企业均有成熟应用案例。对于江西银行这样的商业银行，这些优点非常有利于促进存储基础设施建设。其中，光纤通道是个成熟的技术，被广为采用，具有性能好、稳定、安全的优点。在SAN中，光纤通道早已取得主导地位，并且由于它的高性能、高连通性以及支持基于块的存储协议成为SAN事实上的标准。首先，有效降低延迟，避免拥塞。光纤通道为存储网络提供了无可比拟的性能和效率。FC-SAN在全双工的传输能力下，2Gb/s的光纤通道在理论上能支持400MB/s的带宽。其次，协议简单。根据信道体系结构的特征，光纤通道标准把许多的协议处理分配到硬件上。故江西银行使用FC-SAN技术。

3.2.1.1 江西银行SAN拓扑结构

SAN网络拓扑结构有Cascade， Ring， Full Mash， Partial Mash，Core-Edge等多种不同的结构。江西银行采用Core-Edge拓扑结构，并采用双网结构，其拓扑结构如图5所示。

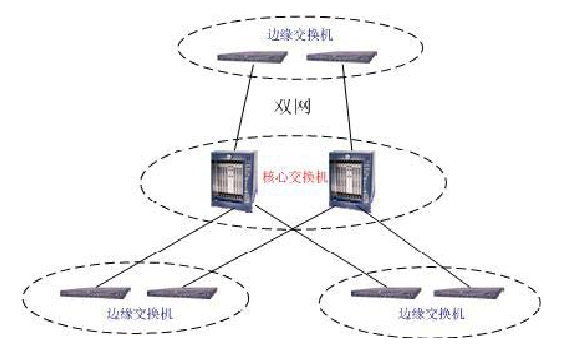


图5 江西银行双网核心-边缘架构

3.2.1.2 江西银行SAN设备的使用

搭建存储区域网络的设备主要有存储光纤交换机、服务器及存储的HBA卡、光纤线缆，而关键的设备就是存储光纤交换机。存储光纤交换机主要有三大品牌：Brocade， McData3， Cisco。江西银行使用Brocade的核心产品Brocade Silkworm 24000，如图6所示。



图6 Brocade Silkworm 24000光纤交换机

该产品具有卓越的性能与更高的可扩展性，其独特的功能可以满足银行的关键需求。Silkworm 24000建立在经过实践考验的Silkworm 12000导向器之上，可以在一个域中支持128个同时处于活动状态的2 Gbit/sec全双工端口。它提高了性能，降低了延迟，通过提高端口/域比率降低了管理成本，通过提高电源利用率降低了总拥有成本，并具备在将来升级到更快连接方式的能力。

Silkworm 24000还用于同包含Windows NT，Linux，Solaris，AIX等各种操作系统的开放异构环境集成，也能同时使用FICON的主机环境集成。这样，各机构就可以灵活地建立经济实用和易于管理的SAN Fabric。后者支持在企业中引入公用事业资源耗用计算模式和信息生命期管理(ILM)系统。Silkworm 24000具有超高可用性、简化管理工作、安全性高、互操作性、可扩展的模块化片式结构、先进的Fabric服务等特点。

3.2.2江西银行信息化中心SAN整体架构

江西银行采用了双冗余设备的存储区域网络架构，具体如图7所示。

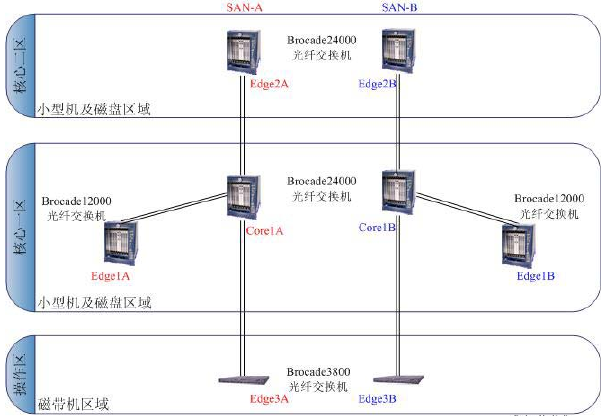


图7 江西银行信息中心网络存储架构图

3.2.2.1 江西银行信息化中心SAN架构设备分布

根据信息中心存储的规模、空间位置分布及端口实际需求，江西银行将两台64端口Brocade 24000交换机放置在核心一区(二层东区)，用作核心交换机(Corel A及Core1B；两台64端口Brocade 24000交换机放置在核心二区(二层西区)，用作边缘交换机(Edge2A及Edge2B；原有二台32端口Brocade12000交换机还放在核心一区，用作边缘交换机(Edge1A及Edge1B；原有两台16端口Brocade3800交换机改放在操作区(三层)，用作边缘交换机(Edge3A及Edge3B)。

两台24000交换机(Corel A和Core1B)是SAN存储网络的核心，也是核心一

区部分主机和存储集中的中心。将64端口的24000交换机作为SAN系统的核心，

它即可以连接其它交换机，也可以直接连接存储设备，将主要的数据访问本地化，

提高系统的性能。

两台24000交换机（Edge2A和Edge 2B）是SAN存储网络的边缘交换机，是核心二区主机和存储的集中的中心。它连接核心交换机，也直接连接本区域主机和存储设备，将主要的数据访问本地化，提高系统的性能。

原有两台12000交换机（Edge l A和Edge 1B）是SAN存储网络的边缘交换机，是核心一区原有连接开放平台服务器的存储集中的中心。它通过与核心交换机相连，可以访问其他区域的设备。

原有两台3800交换机（Edge 3A和Edge 3B）是SAN存储网络的边缘交换机，是操作区磁带设备集中的中心。由于该交换机为旧有设备，连接着一些开放平台服务器和存储设备，需要将这些设备连接迁移到新的SAN交换机，再将Edge 3A和Edge 3B用作操作区边缘交换机。所有服务器均需要通过它们访问磁带设备。

SAN网络中核心交换机与边缘交换机通过Trunk方式连接。开放平台服务器及存储设备互相访问时，可以采用Zoning（分区）功能区分开。

光纤交换机是1 G/2G自适应的，所以每个端口都可以接1 G/2G的设备（如1G端口的HDS 9960磁盘机、2G端口的HDS 9970磁盘机）、开放平台服务器设备1 G/2G的光纤卡（如IBM 6227和6228），磁带库的FC端口也可以直接连接到交换机上。

3.2.2.2 江西银行信息化中心SAN双网核心边缘结构

江西银行在网络设计时采用核心-边缘网络结构，整个网络以核心一区的Core 1 A它既是数据的中心，也是网络的中心。

作为数据的中心，大部分的开放平台服务器和存储设备都放在核心一区，所以在此区域的大部分应用都通过这台交换机来访问数据。作为网络的中心，其它区域的交换机都与它们相连。如果有必要，其它区域的开放平台服务器可以通过它们访问其它区域的存储。核心一区及核心二区的开放平台服务器都需要通过它们连接到操作区域来访问磁带设备。部分核心二区开放平台服务器通过核心交换机访问核心一区的存储设备。

为保证高可靠性和提高系统的运行速度，江西银行在每台主机和存储上都通过两个HBA连接到不同的Fabric网络中。在该架构中SAN交换机端口速率均为2Gb/s。但2G的设备可以向下兼容，包括交换机、HBA等都可自适应成1G的速度。

3.3 南昌移动网络存储结构

NAS被定义为一种特殊的专用数据存储服务器，包括存储器件（例如磁盘阵列、CD/DVD驱动器、磁带驱动器或可移动的存储介质）和内嵌系统软件，可提供跨平台文件共享功能。NAS通常在一个LAN上占有自己的节点，无需应用服务器的干预，允许用户在网络上存取数据，在这种配置中，NAS集中管理和处理网络上的所有数据，将负载从应用或企业服务器上卸载下来。NAS通常在一个LAN上占有自己的节点，NAS技术对数据的存取则是基于文件方式（File）的，NAS的数据传输协议是建立在TCP/IP之上的，因此NAS非常适用于那些需要进行文件共享，对建设成本比较敏感的应用，有效降低总拥有成本，保护用户投资。通过NAS可以实现广泛的应用，支持NFS，CIFS，FTP，HTTP等，甚至支持iSCSI方式。

iSCSI协议分为initiator（发起端）和target（目标端）两个部分。通常发起端是具有网络数据操作请求的客户端，目标端是具有实际网络数据存储设备的服务器端。iSCSI是一个应用层协议。同时由操作系统的驱动程序实现的。一个完整的iSCSI体系结构由SCSI协议层，iSCSI数据封装层，TCP运输层，IP网络层，和物理链路层五个部分组成，具体如图8所示。

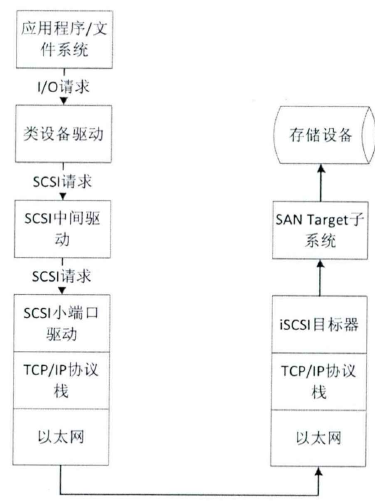


图8 iSCSI协议示意图

iSCSI协议作为一个应用层的协议是承载在TCP/IP协议之上的。由于iSCSI是块级别的协议，一次对网络数据的操作可能涉及到多个块，所以一次操作往往是由一次或多次TCP链接组成的。iSCSI具有如下优点：

（1）硬件成本低：构建iSCSI存储网络，除了存储设备外，交换机、线缆、接口卡都是标准的以太网配件，价格相对来说比较低廉。同时，iSCSI还可以在现有的网络上直接安装，并不需要更改企业的网络体系，这样可以最大程度地节约投入。

（2）操作简单，维护方便：对iSCSI存储网络的管理，实际上就是对以太网设备的管理，只需花费少量的资金去培训iSCSI存储网络管理员。当iSCSI存储网络出现故障时，问题定位及解决也会因为以太网的普及而变得容易。

（3）扩充性强：对于已经构建的iSCSI存储网络来说，增加iSCSI存储设备和服务器都将变得简单且无需改变网络的体系结构。

（4）带宽和性能：iSCSI存储网络的访问带宽依赖以太网带宽。随着千兆以太网的普及和万兆以太网的应用，iSCSI存储网络会达到甚至超过FC（Fiber Channel，光连通道）存储网络的带宽和性能。

（5）突破距离限制：iSCSI存储网络使用的是以太网，因而在服务器和存储设备的空间布局上的限制就会少了很多，甚至可以跨越地区和国家。

在南昌移动的实际应用中，正是采用NAS来实现iSCSI服务，将存储空间提供给各服务器使用达到降低成本、灵活部署、方便升级的目的。通过一台普通Pc server使用HBA卡与SAN交换机连接，SAN与磁盘阵列连接；PC server通过以太网与各服务器相连接，从而为NAS提供物理基础。在软件层面上，利用Linux系统的HBA驱动及本身的卷管理实现对磁盘存储的利用，使用SAN ZONE实现磁盘阵列的空间甄别；安装NAS管理软件，实现将容量按照规划分配给各服务器，并在安全的基础上进行管理。

3.3.1南昌移动NAS存储的总体架构

南昌移动数据存储系统是一个很复杂的大型数据存储中心，其总体结构十分复杂，为了简化及解释方便，图9中将大量的设备进行了简化和抽象。在需要提供存储服务的主机中拥有各种操作系统平台，为了简化，在这里只涉及到了UNIX操作系统，其它平台比较类似。

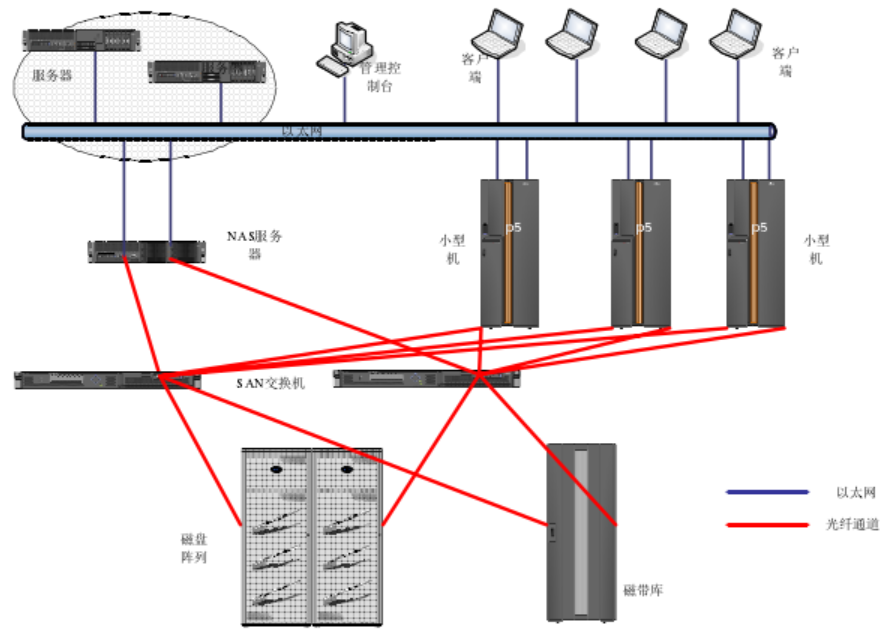


图9 南昌移动NAS存储架构示意图

如图9所示，圆圈中的服务器由NAS服务器通过IP网络提供存储容量，NAS服务器与两台SAN交换机通过光纤连接，SAN交换机与存储磁盘阵列通过光纤进行连接。而具备相应投资及软硬件到位的主机则通过HBA卡直接与SAN交换机相连，以提供延时极短、可靠性高的冗余解决方案。

该系统的工作流程：client端是使用NAS空间的主机，当client需要访问数据时，操作系统调用卷管理，卷管理再通过ioctl到ISCSI驱动函数（iSCSI initiator） , ISCSI驱动函数生成socket链接，链接到NAS服务器（iSCSI target）后完成相应的IO操作。而 NAS服务器在接收到client的访问需求后，将对磁盘阵列进行访问；磁盘阵列上的各个逻辑盘是通过光纤网络映射为UNIX系统的硬盘设备。逻辑盘（logical disk）是逻辑上的概念，它可以由多个物理disk（各物理disk的一部分或全部）按照一定的RAID方式构成，可以进行并行访问或存储冗余数据，与普通物理硬盘一样。因为UNIX上的设备都被看作特殊文件，访问设备类似于文件访问，设备驱动程序都提供打开和读写接口来操作真正的硬件，另外还提供IO Control接口来进行特殊的设备控制或从设备获取信息。与disk相关的specialfile都在/dev/dsk或/dev/rdsk目录下，dsk目录与rdsk目录的差别是打开rdsk disk设备后，read/write/ioctl等调用将直接发送给硬件，而dsk目录的设备则还经过操作系统的中间处理（一般是文件系统）。整个数据访问过程如图10所示。

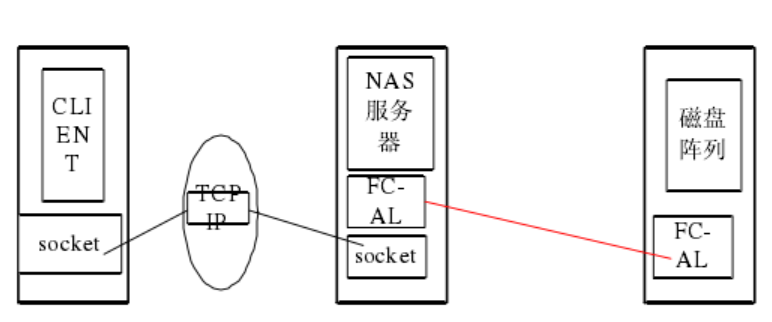


图10数据访问示意图

3.3.1.1 南昌移动iSCSI的实现

iSCSI的实现有多种方法，不同的架构具有不同的实现方式。

（1）控制器架构

iSCSI的核心处理单元采用与FC光纤存储设备相同的结构。即采用专用的数据传输芯片、专用的RAID数据校验芯片、专用的高性能cache缓存和专用的嵌入式系统平台。打开设备机箱时可以看到iSCSI设备内部采用无线缆的背板结构，所有部件与背板之间通过标准或非标准的插槽链接在一起，而不是普通PC中的多种不同型号和规格的线缆链接。

这种类型的iSCSI存储设备核心处理单元采用高性能的硬件处理芯片，每个芯片功能单一，因此处理效率较高。操作系统是嵌入式设计，与其他类型的操作系统相比，嵌入式操作系统具有体积小、高稳定性、强实时性、固化代码以及操作方便简单等特点。因此控制器架构的iSCSI存储设备具有较高的安全性和和稳定性。

控制器架构iSCSI存储内部基于无线缆的背板链接方式，完全消除了链接上的单点故障，因此系统更安全，性能更稳定。一般可用于对性能的稳定性和高可用性具有较高要求的在线存储系统，比如：中小型数据库系统，大型数据的库备份系统，远程容灾系统，网站、电力或非线性编辑制作网等。

（2）连接桥架构

整个iSCSI存储分为两个部分，一个部分是前端协议转换设备，另一部分是后端存储。结构上类似NAS网关及其后端存储设备。前端协议转换部分一般为硬件设备，主机接口为千兆以太网接口，磁盘接口一般为SCSI接口或FC接口，可连接SCSI磁盘阵列和FC存储设备。通过千兆以太网主机接口对外提供iSCSI数据传输协议。后端存储一般采用SCSI磁盘阵列和FC存储设备，将SCSI磁盘阵列和FC存储设备的主机接口直接连接到iSCSI桥的磁盘接口上。

iSCSI连接桥设备本身只有协议转换功能，没有RAID校验和快照、卷复制等功能。创建RAID组、创建LUN等操作必须在存储设备上完成，存储设备有什么功能，整个iSCSI设备就具有什么样的功能。

（3）PC架构

PC架构可以理解为存储设备建立在PC服务器的基础上。即选择一个普通的、性能优良的、可支持多块磁盘的PC（一般为PC服务器和工控服务器），选择一款相对成熟稳定的iSCSI target软件，将iSCSI Target软件安装在PC服务器上，使普通的PC服务器转变成一台iSCSI存储设备，并通过PC服务器的以太网卡对外提供iSCSI数据传输协议。

在PC架构的iSCSI存储设备上，所有的RAID组校验、逻辑卷管理、iSCSI

运算、TCP/IP运算等都是以纯软件方式实现，因此对PC的CPU和内存的性能要求较高。另外iSCSI存储设备的性能极容易受PC服务器运行状态的影响。

当前由于PC架构iSCSI存储设备的研发、生产、安装使用相对简单，硬件和软件成本相对较低，因此市场上常见的基于PC架构的iSCSI设备的价格都比较低，在一些对性能稳定性要求较低的系统中具有较大的价格优势。

（4）PC+NIC架构

PC+iSCSI Target软件方式是一种低价低效比的解决方案，另外还有一种基于PC+NIC的高阶高效性ISCSI方案。即通过使用带iSCSI功能的以太网卡来提供iSCSI访问。

南昌移动的存储解决方案中使用了上述二与三两种方式的混合体。既使用了类似NAS及其后端的存储设备，又使用软件方式实现iSCSI功能。其采用了完全免费的开源软件openfiler操作系统及其所提供的iSCSI target功能，在客户端则直接使用操作系统自带的iSCSI initiator利用普通网卡即可实现对存储的访问。

客户端的数据流程如图11所示，当client或是应用需要访问数据时，操作系统调用卷管理，卷管理再通过ioctl到iSCSI驱动函数（iSCSI initiator），iSCSI驱动函数生成socket链接，链接到NAS服务器（iSCSI target）后完成相应的IO操作。而 NAS服务器在接收到client的访问需求后，将对磁盘阵列进行访问；磁盘阵列上的各个逻辑盘是通过光纤网络映射为UNIX系统的硬盘设备。

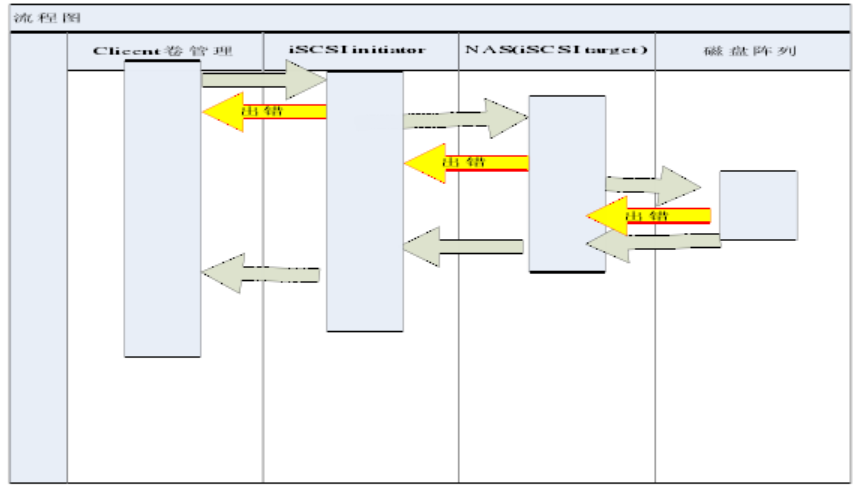


图11 数据流程图

3.3.1.2南昌移动卷管理的实现

南昌移动在卷管理的实现上采用了LVM的功能。卷组在创建时将在卷组中的每一块盘的头部写入VGDA（volume group data area ），VGDA的主要内容是VGID，LVM记录位置、LVM记录长度、物理卷信息、PP\_SIZE，VGSA长度、VG\_TYPE及各LV的MAP信息、物理卷的PVID等，如果只有两块盘则在第一块盘上写入两份VGDA。根据卷组的类型，这个VGDA区的大小将不同。卷组的构成如图12所示。

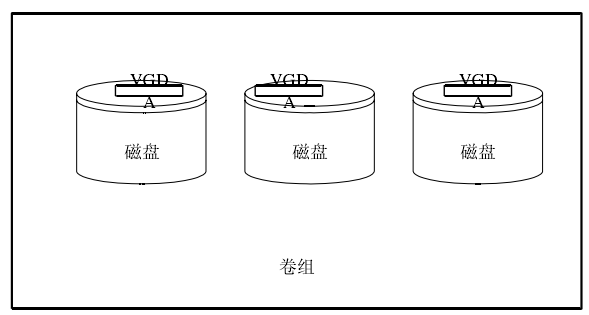


图12 卷组示意图

物理卷在第一个扇区写有PVID信息，这是硬盘在主机对其进行初始化时生成，注意在正常情况下与VGDA中的PVID应一致，如果不同将导致卷组出错无法使用。物理卷被分成若干的PP （physical partion），PP作为分配给逻辑卷使用基本单位。在经典的卷管理中每个物理卷最多1016个PP，经过改进现在已经可以通过技术手段消除这个限制，同一个卷组中的PP大小是一致的。物理卷的构成如图13所示。

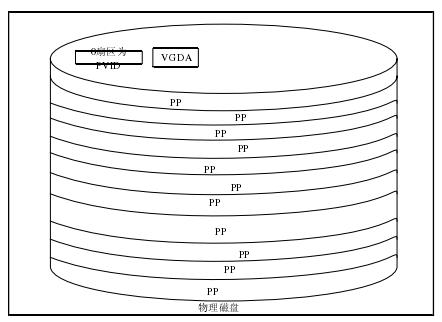


图13 物理卷示意图

逻辑卷是信息的逻辑组织，它以卷组为母体，与物理卷无关。逻辑卷是文件系统的数据存储底层，文件系统将IO交给逻辑卷来完成。逻辑卷由物理卷的PP构成，不受物理卷物理限制。其构成如图14所示。

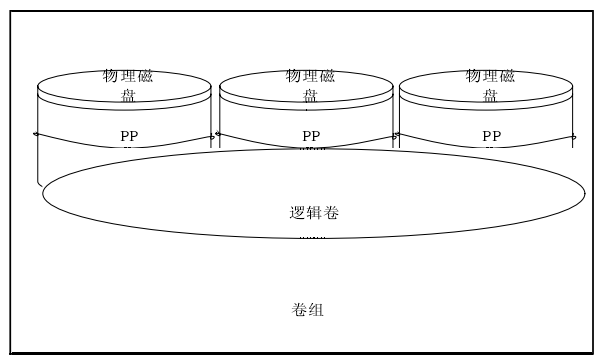


图14 逻辑卷示意图

3.3.1.3南昌移动iSCSI initiator的实现

为了访问提供服务的NAS，使用了iSCSI协议，目前的操作系统通常内置了iSCSI initiator，实现对iSCSI target的访问。它的访问方式是由驱动生成虚拟设备，虚拟设备通过丁CPIP协议与target产生socket联接，然后进行相应的配置后提供服务。如图15所示。

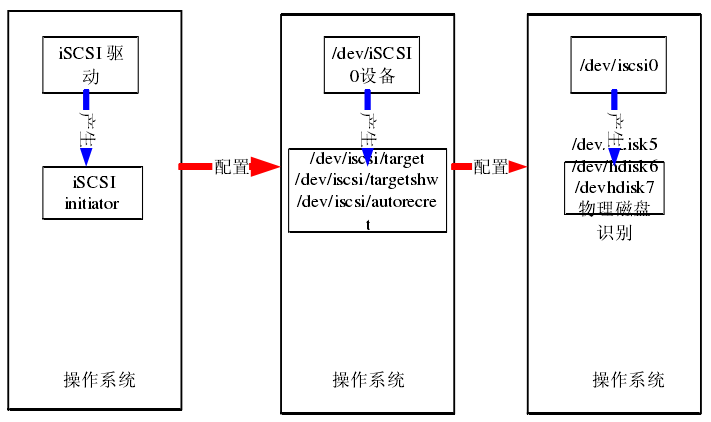


图15 iSCSI initiator示意图

3.3.1.4南昌移动NAS（iSCSI target）的实现

为了实现NAS功能，南昌移动使用了一台闲置的PC server，安装运行openfiler系统，运行NAS服务功能。在openfiler上，安装配置HBA卡，使其可以直接访问磁盘阵列，通过光纤交换机或直接连接到磁盘阵列的光纤口，实现对磁盘阵列的访问；安装运行NAS软件，提供通过浏览器进行配置NAS，实现将磁盘阵列的容量进行排列组合，提供不同的容量给不同的主机。如图16所示。

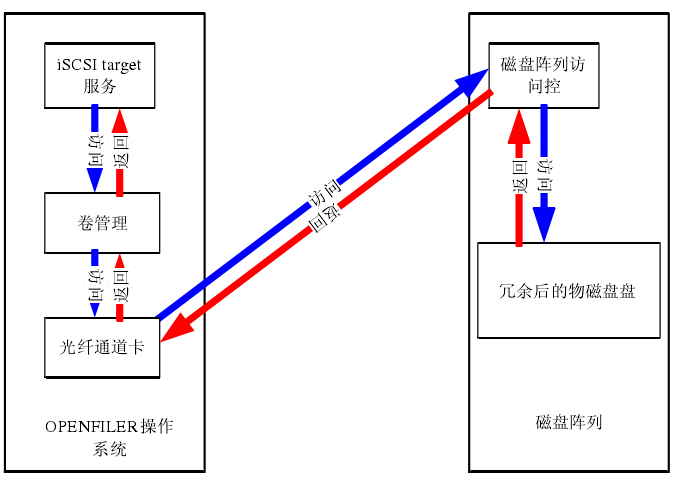


图16 NAS实现示意图

3.4海康威视视频云存储系统架构

海康威视视频云存储系统是由海康威视自主研发的一套针对安防监控行业应用的专业级云存储系统，采用软硬件一体化设计，融合了集群应用、负载均衡、虚拟化、云结构化、离散存储等技术，可将网络中大量各种不同类型的存储设备，通过专业应用软件集合起来协同工作，共同对外提供高性能、高可靠、不间断的视频、图片数据存储和业务访问服务。海康威视视频云存储系统与通用的文件云存储系统在底层的云架构设计上有很多类似之处，但是海康威视视频云存储系统结合了视频和图片的数据特点进行深度的架构调整和应用融合。海康威视视频云存储拓扑结构如图17所示。

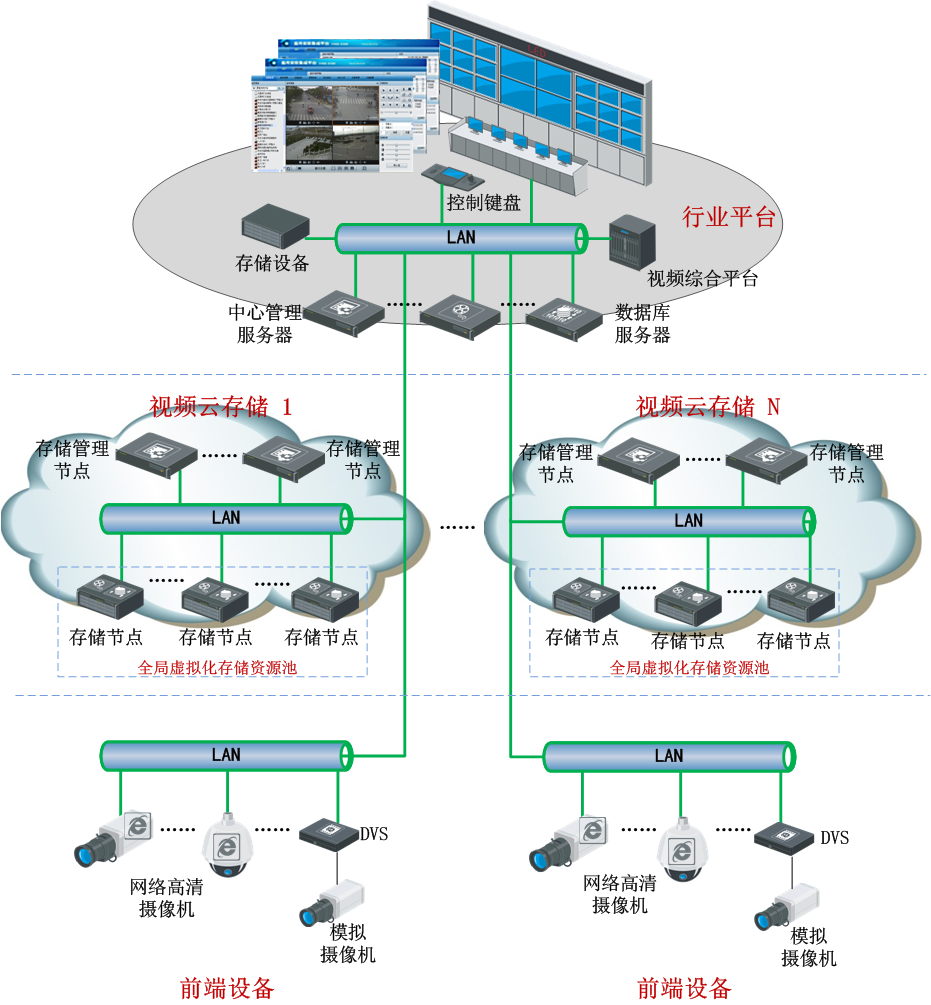


图17 海康威视视频云存储拓扑结构

在海康威视视频云存储系统设计中主要采用了如下技术：

1. 采用存储全域虚拟化技术对具有海量存储需求的用户提供透明存储

构架，可持续扩容避免瓶颈限制，可以更有效的进行资源管理，灵活增减空间，达到最大程度上合理利用空间的效果；

1. 采用集群技术，解决单/多节点失效造成的业务中断问题，并利用负

载均衡技术充分利用各存储节点的性能，提升系统的可靠性和安全性；

1. 采用离散存储技术，保障了用户高效读写和业务的持续性；
2. 采用统一完善的接口，降低对接成本、平台维护成本和用户管理的复杂度；
3. 采用开放的集成构架，可兼容业界各类标准的iSCSI/FC存储设备，保护用户现有存储投资资源；
4. 采用数据备份和容灾技术，保证云存储服务的安全稳定。

3.4.1海康威视视频云存储系统存储主机介绍

海康威视视频云存储系统存储主机是海康威视针对高性能、高可靠性需求而推出的千兆网络存储系统，如图18所示。提供4-8Gbps的传输带宽和单模块高达144TB的海量存储空间，且支持扩展。采用强大的64位四核处理器、稳定的体系架构、RAID 6技术、热插拔硬盘、多千兆网口及冗余热备电源、智能控制风扇，既满足了高性能的要求又彻底保护了用户的数据安全。系统具有如下管理功能：中文Web可视化管理，多设备集中管理和全面状态监测报警，批量升级；绿色节能，磁盘及RAID预警容错，带宽预留等；访问控制、身份识别、MAC地址绑定等安全机制。



图18 存储主机

4小结

本案例基于课程的理论讲述，结合江西银行、南昌移动、海康威视实际运行的网络存储系统作了详细介绍和说明，主要讨论了各种网络存储技术的应用和常用网络设备，具体包括对网络存储技术相关概念和发展、网络存储架构的类型、备份与容灾相关概念的介绍；讨论了网络存储常用的接入技术和江西银行、南昌移动网络存储架构，海康威视视频云存储系统；对江西银行和南昌移动、海康威视视频云存储系统的实际网络存储架构以图的形式进行介绍，并对其使用的设备及相关协议进行了具体的介绍。学生通过本案例的学习能够更好地理解网络存储技术抽象概念和实际应用，同时本案例的设计能极大地促进课程教学并提高学生的学习兴趣。

5附件：

本案例讲义的附件为课程的PPT及案例使用说明（教师指导书）。

6思考题

（1）SCSI协议模型可分为哪三层；

（2）SCSI的异步传输与同步传输有什么不同；

（3）数据备份的类型有哪几种；

（4）存储虚拟化的方法有哪些；

（5）磁盘子系统的I/O通道的配置有哪几种方式。