负载均衡在服务器和网络的世界中并不是一个新的概念，许多产品都能够提供不同类型的负载均衡解决方案。比如，路由器能够在不同的路径之间分配流量到达相同的目的地，在不同的网络资源中平衡它们的负担。另一方面，一个服务器负载均衡设备，能在多台服务器之间分发流量。

最初，负载均衡设备只是满足简单的负载均衡需求，而如今这些产品已得到迅速的发展，能够提供更多更复杂的功能，如：负载均衡、流量管理、智能流量交换技术、内容交换技术甚至应用交换等。负载均衡设备能够对服务器、应用系统进行复杂而且准确的健康检查，从而提高系统的可用性和可管理性。因为负载均衡设备通常部署在服务器群的前端，它能够保护服务器免遭恶意用户的入侵从而提高安全性。根据 IP 包中的信息或应用请求的内容，负载均衡设备会作出智能的决策，并准确地将流量导向到正确的数据中心、服务器、防火墙、高速缓存或者应用系统。

负载均衡的需求

有两个重要的因素推动了负载均衡的出现：服务器和网络。随着国际互联网络(Internet)和企业网络(Intranet)的出现，连接着服务器和企业的雇员、客户、供应商以及合作伙伴的数字网络已经变得至关重要。整个信息服务系统如果不能提供服务或者性能差到不可接受的地步时，很可能会导致一些业务被迫停顿。举例来讲，如果要建立一个电子商务的网站，需要考虑许多部件：边界路由器、交换机、防火墙、高速缓存、Web 服务器、中间件服务器和数据库服务器，而针对不同应用需要不断增加服务器设备，最终数据中心填满了各种各样的服务器群。在这些服务器群中实现可扩展性、可管理性、可用性以及安全性变得非常复杂并具有挑战性，这也是智能高层交换技术出现的一个重要的推动力。从边界路由连接互联网络开始，一直到数据中心的后台数据库主机系统，负载均衡已经成为一个解决上述问题和挑战的新的强有力的工具。

服务器环境

至少有两个原因使得当今的企业和互联网络服务商(ISP)不断增加服务器的数量。第一，在互联网络时代需要配备许多不同功能的服务器或应用，包括：WEB 服务器、FTP 服务器、DNS、NFS、E-mail、ERP、中间件、数据库等等；第二，因为一台服务器无法提供足够的计算能力和容量，许多应用需要多台服务器同时提供服务。如果你与数据中心的任何一个人去交谈，他或者她就会告诉你，他们花了多少时间在不同的应用和服务器上解决可管理性问题、扩展性问题和高可用性问题。举例来说，如果电子邮件应用无法处理不断增加的客户访问的需求时，就必须配备另外一台电子邮件服务器，网络管理员还必须考虑如何将客户的请求分发到这两台电子邮件服务器上。如果一台服务器出现故障，网络管理员在维修这台服务器的同时，需要在另外一台服务器上运行这一应用，而一旦原来的服务器修好之后，它还必须回到原来的地方继续提供服务。对客户来说，这些工作都会影响到应用系统的性能和可用性。

扩展的挑战

如何扩展运算能力并不是一个新的问题。以前，一台服务器专门执行一个应用服务，如果这台服务器不能胜任这个工作时，就会购买一台更加强大的服务器来替换。因此提高系统的不同部件来增加服务器的运算能力就变得越来越有效。举例来讲，Intel 公司提出的著名的摩尔定律提到，处理器的处理速度正以每隔 18 个月翻一倍的速度增长着，但是对提高运算能力的需求的增长速度却更快。于是我们发明了集群技术，最初将它应用于主机系统。由于主机系统是专有的技术，所以生产主机的厂商都采用它们自己的技术，来配置一个主机集群以便分担计算任务，这样做是比较容易的。现在市场上出现了两种集群技术：松散耦合的系统和对称多处理技术。但是这两种集群技术都有局限性，而且从系统性能的曲线来看，性价比也没有足够的吸引力。

松耦合系统

松耦合系统是由许多相同功能的计算单元，通过系统总线互联在一起。每个计算单元包含处理器、内存、磁盘控制器、磁盘驱动器和网卡等部件，每个计算单元本身就是一台计算机。把这些计算单元组合在一起，即可组成一个松耦合的系统。象天腾(Tandem)公司的一个系统中就包含了 16 个计算单元。松耦合的系统，采用内部处理器间通讯将任务分担到多个计算单元来处理。松耦合处理系统只在计算任务可以很容易地被分割时才能够扩充。假设这样一个任务：在一个数据表中找出所有“Category”列的值等于 100 的记录，那么在操作时数据表就会被分割成四个相等的部分，而且每个部分都存放在被一个处理器控制的硬盘分区中。查询的任务也被分割成四个任务，而且每个处理器并行地执行查询的任务，最后再合并成一个完整的查询结果。但不是所有的计算任务都是可以分割的，比如需要更新剩余灯泡的库存数据，那么只有负责灯泡数据库的处理器才能够更新数据。如果灯泡的销售快速增长，使得更新库存的请求也快速增加，就可能使专门负责处理灯泡库存的处理器变成性能的瓶颈，而其他的处理器却很空闲。为了实现可扩展性，松耦合系统需要系统和应用软件的支持，即便是在任务可以被分割的情况的。松耦合系统在任务不可分割时，或者更新灯泡库存之类的突发情况下，就无法实现系统性能的扩展。

对称多处理系统

对称多处理(SMP，Symmetric multiprocessing)系统采用多个处理器共享相同内存的技术，应用软件必须被改写成可适应多线程环境的系统，每个线程能够运行一个计算元语。这些线程共享同样的内存而且依赖于特别的通讯手段，如旗语或消息机制等。操作系统在多个处理器中分配任务，使多个处理器能够同时运行从而提供可扩展性。同样的，一个计算任务能否被清晰地分割并同时运行也是一个挑战。而且增加一个处理器，操作系统就需要在线程和处理器中通讯和协调，这样一来系统的可扩展性又会受到限制。

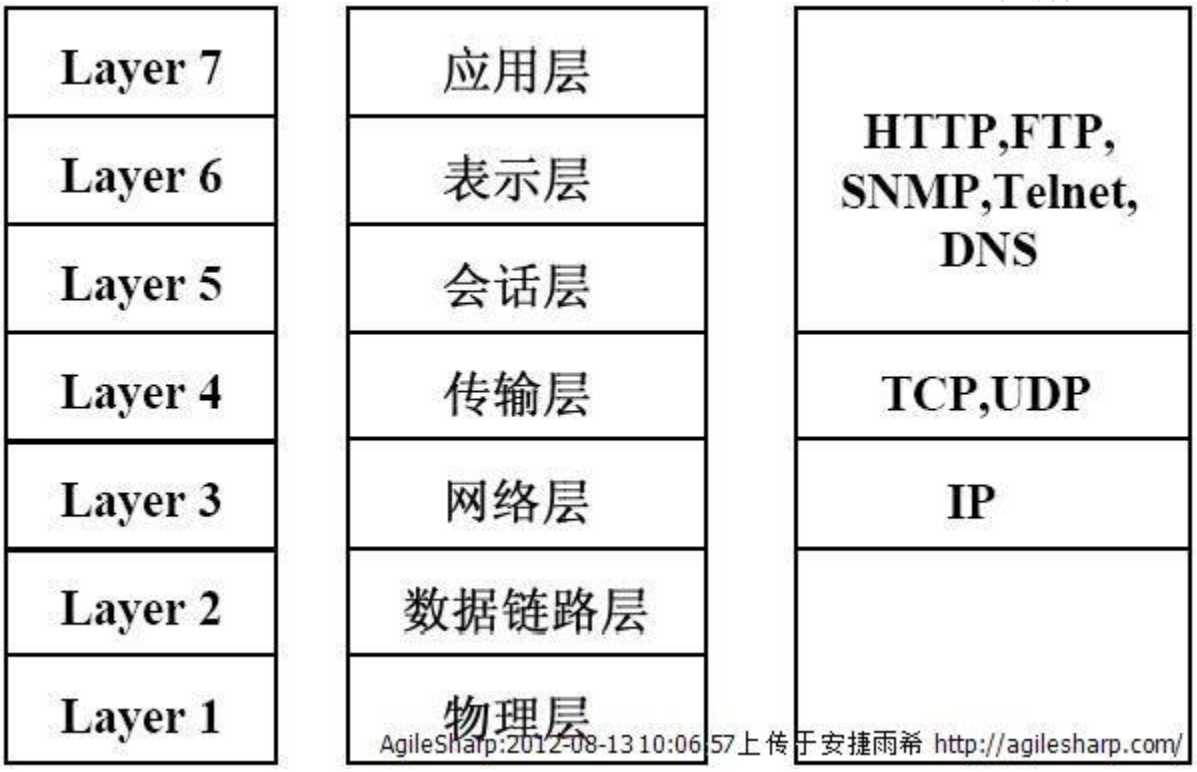
网络环境

传统的交换机和路由器根据数据包的 IP 地址或 MAC 地址转发数据包，但是它们不能满足当前复杂的服务器集群的需要。举例来说，传统的路由器或者交换机，无法智能地把请求数据包发送到特定的服务器或者高速缓存。如果目的服务器发生故障，不能提供服务时，传统的交换机仍会继续将客户请求发送到这个无效的目的地。为了了解传统的交换机和路

由器的功能，以及智能内容交换机的先进性，我们必须先来研究一下 OSI 模型。

OSI 模型

OSI 模型是一个开放式的标准，它定义了不同的设备或计算机是如何互相通讯的。如图所示，它由 7 层构成，从物理层到应用层。网络协议如传输控制协议(TCP)、用户数据报协议(UDP)、互联网络协议(IP)和超文本传输协议(HTTP)，都对应 OSI 模型中不同的层。IP 是三层的协议，而 TCP 和 UDP 在第四层工作。每层的协议能够与另外一台设备上的相同层次的协议互相通讯，并与其紧挨着的上层或着下层的协议交换信息。



二、三层交换

传统的交换机和路由器工作在第二、三层中，也就是说它们根据网络第二、三层的包头信息来转发数据包，这也符合二、三层交换机的设计初衷，能够进行大量的数据传输工作。因为很多有用的信息，存放在更高层协议的数据段中，所以问题是，我们从一个可以观察更高层次协议的包头信息的交换机中能获得哪些益处呢？

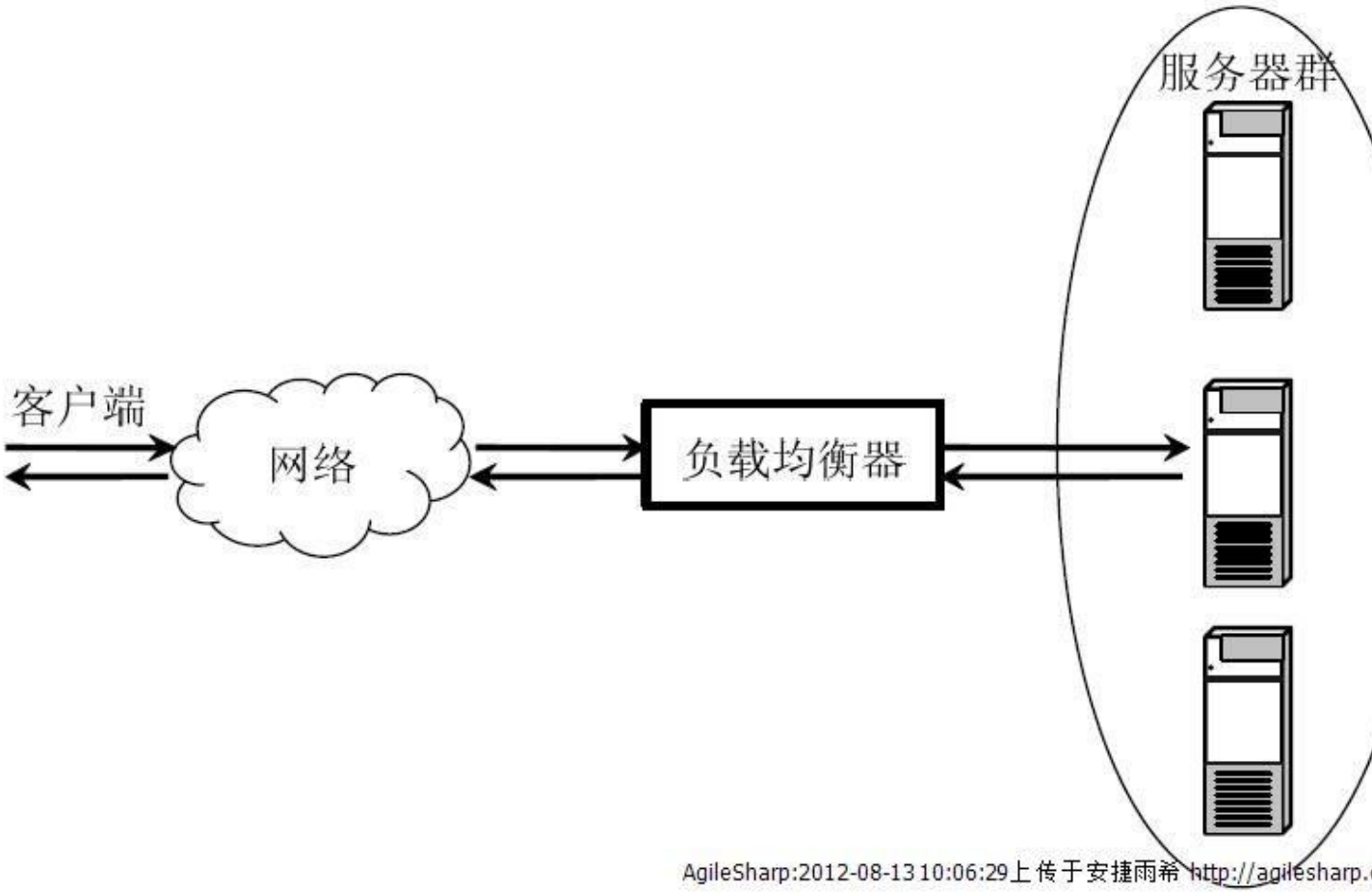
四到七层交换

网络四到七层交换技术，是指交换机根据数据包的第四到七层协议的包头信息来做数据包交换。TCP 和 UDP 是本书中最重要的第四层协议，TCP 和 UDP 包头中含有许多重要的信

息可以作为智能交换的依据。举例来说，HTTP 协议通常在 TCP 80 端口提供 Web 服务。如果一个交换机能够识别 TCP 端口号码，对一个请求它就可以优先处理，或者拒绝访问，或者将其重定向到一个特定的服务器。只需要检查 TCP 和 UDP 的端口号码，交换机就能够识别许多常规的应用，包括 HTTP、FTP、DNS、SSL 和流媒体协议等。利用 TCP 和 UDP 的相关信息，四层交换机能够通过分发 TCP 或者 UDP 请求，进而达到多台服务器的负载均衡。四七层交换机这一术语，既是真实存在的，也是市场推广的噱头。大部分四七层交换机至少工作在网络第四层，而且许多交换机确实能够识别四层协议以上的信息。至于能识别网络四层以上多少信息或者哪一层的信息，不同的产品就大相径庭了。

负载均衡：定义和应用

随着 Internet 的出现，网络变成了世界的中心。连接真实世界和 Intranet 的 Internet 逐渐成为商业运营主体，IT 基础架构可以分成两类设备：计算机，包括客户端和服务器，和连接这些计算机的交换机和路由器。从概念上讲，负载均衡器是网络和服务器之间的桥梁，如图所示，一方面，负载均衡器能够识别许多高层的协议，能够智能地与服务器进行沟通；另一方面，负载均衡器能够识别网络协议，可以跟网络设备高效率地整合在一起。



图中 负载均衡与服务器群

负载均衡至少有四种主要的应用：
1)：服务器负载均衡；
2)：广域网服务器负载均衡；
3)：防火墙负载均衡；
4)：透明高速缓存负载均衡。服务器负载均衡负责将客户的请求分发到多台服务器，用以扩展服务能力，并且能够使应用系统具有容错能力。广域网服务器负载均衡负责将客户的请求导向到不同数据中心内的服务器群中，以便为客户提供更快的响应时间，并且在某一数据中心出现灾难性事故时提供智能的冗灾处理。防火墙负载均衡将请求负载分发到多台防火墙，用来提高安全性能，获得更高的处理能力。透明高速缓存的交换能够把流量透明的导向多台高速缓存，把静态的内容交给高速缓存处理，以此卸载网站服务器的压力，从而提高网站服务器的性能、加快客户端的响应时间。

负载均衡产品

负载均衡产品有很多种类型，大致可以分为三类：软件产品、功能服务器和交换机，下面分别予以介绍。软件负载均衡产品运行在负载均衡服务器上，这类产品包括 Resonate、Rainfinity 和 Stonebeat。功能服务器产品是包含必要软件和硬件的黑盒子，可以实现 Web 交换。这些设备可能是一台简单的服务器或着个人电脑，配备有专用的操作系统和软件，或者是一个专有的软件和硬件的设备。F5 Networks 和 Radware 都提供这种产品。交换机产品通过利用某些硬件和软件，使传统的二三层交换机扩展到更高的层次。

而许多供应商只能够将二三层交换功能集成在 ASIC 芯片中，却没有一个产品能够将所有的四七层交换功能集成在 ASIC 中，尽管些厂商宣称能够实现。多数情况下，硬件只能辅助提升产品性能，而主要的工作还是由软件来实现。比较著名的厂商如 F5 Networks、Cisco Systems、Foundry Networks 和 Nortel Networks 都提供高层交换机产品。负载均衡是属于服务器的功能呢还是属于交换机的功能呢？答案并不重要，重要的是，哪一种类型的负载均衡产品能够满足您对性价比、功能模块、稳定性、可扩展性、可管理性和安全性等方面的需要呢？本书不准备推荐任何类型的产品，但是会涉及到这几种类型的负载均衡产品的功能和概念。

命名的难题

负载均衡有很多名称：二到七层交换机、四到七层交换机、Web 交换机、内容交换机、网络流量管理交换机、主动式流量管理设备。它们都完成相同的工作，只是在功能上稍微有些差异。虽然负载均衡器是一个描述性的词汇，但它引入了更多更深层次的功能，所以有些厂商使用 Web 交换机这种称呼。我们使用负载均衡器这个词汇，简单明了。不管是什么样的产品，负载均衡都是其根本。

相关链接：

负载均衡详解第二篇：服务器负载均衡的基本概念-网络基础

负载均衡详解第三篇：服务器负载均衡的基本概念-使用负载均衡器的服务器群

负载均衡详解第四篇：服务器负载均衡的基本概念-负载均衡时数据包流程

负载均衡详解第五篇：服务器负载均衡的基本概念-健康检查