计算机科学 2002Vol. 29№. 8

基于集群和分布式存储的大容量邮件系统研究

The Research of High-Capacity Email System based on Cluster and Distributed Storage

梁建民　祝明发　梁小萍

（中国科学院计算所　北京 100080）

Abstract　In this paper, firstly we discuss the characteristic of high-capacity email system and its key technologies, then the strategies of high-capacity email system such as load-balance, email-storage and user-authentication are in-depth analyzed. Finally, a hierarchical high-capacity email system architecture based on cluster and distributed storage is presented.

Keywords　Cluster, High-capacity email system, Load-balance, Distributed storage

1. 引言

电子邮件是 Internet 的传统服务，它基于可靠的顺序数据流传输，对网络连接和协议结构的要求较低，进行信息交流具有方便快捷可靠的特点，因此一直是 Internet 上使用最为频繁和广泛的服务。

随着 Internet 上的信息不断增长，信息类型、服务类型的不断扩展，传统的电子邮件技术已无法适应不断扩大的用户群和用户群对电子邮件不断增长的需求。多种信息的存储和管理、大用户量的并发访问、邮件系统的高可扩展性、高可靠性都要有成熟的大容量电子邮件技术作为支撑。

目前大型的邮件服务系统一般采用紧耦合结构的机群和邮件共享存储的方式实现，例如采用基于多处理器的高性能服务器机群和 NFS、SAN 等共享存储方式，这种架构的固有缺点是系统的扩展性、性能和可靠性完全依赖于一个或几个专用设备，这些专用设备本身很可能成为系统的瓶颈或单一故障点。共享式存储方式也影响了系统的 I/O 性能，使系统吞吐量降低，同时这些专用设备成本也相对较高。于是人们开始研究如何实现可扩展、高可用的邮件系统。

2. 大容量邮件系统的支撑平台——集群技术

邮件服务是一种 I/O 密集型业务，每次邮件访问无论发送还是读取都需要进行磁盘 I/O 操作，且用户访问间具有独立性，纯粹的分布式系统消息传递带来的开销很大，不适合于大型邮件系统；紧耦合结构的机群可以实现快速的 I/O 读写，但无法满足大容量邮件系统应具有的高可扩展性。集群技术把一组相互独立的计算机，用快速的内部网互联，组成一个松耦合结构的计算机系统，并以单一系统的模式加以管理，既具有快速的 I/O 读写能力，又能通过灵活的添加主机实现系统的可扩展性。对管理者来说，集群是高可靠的、不存在单一故障点的、可以统一管理的系统；对外部的用户来说，集群系统是一台高性能的服务器。和传统的高性能计算机相比，集群技术有以下优点：

1）高可扩展性　传统的高性能计算机在达到其自身的处理瓶颈之后，除了升级零部件可以获得很小的性能空间之外，无法再提高计算效率，唯一的解决方案就是购买更为强大的计算机。而集群的优点则在于其达到处理瓶颈时，可以通过增加计算机的数目来获得几乎是无止境的处理能力，系统可以根据需要而灵活地配置和变更。

2）高可靠性　传统高性能计算机的可靠依赖于单一计算机本身的软硬件的可靠性，一旦任何一点出现问题，都会造成系统瘫痪和服务中断。而集群模式中由多台计算机组成了一个低代价的冗余系统，单独一台机器的故障并不影响整个系统的正常运行，任何一台机器都可以随意更换和维修，而不会造成服务的中断。

3）较高的性能价格比　和传统的高性能计算机相比，集群技术具有较高的性能价格比，可以充分利用现有设备，节省投资。

3. 集群下的大容量邮件系统应具有的特征

1）单一登录点　整个邮件系统在网络中对用户只有唯一的网络名称，用户可以把整个系统看作一台高性能的服务器，不需要知道具体由哪一个节点提供服务。

2）高可靠性和高可扩展性　邮件系统应该具有可靠的容错机制，由于单个服务器的意外情况引起的系统崩溃在关键性应用中是不允许的，在集群系统中通过设备冗余避免系统的单一失效点。

邮件系统应该具有邮箱迁移能力，当迁移发生失误，或需要反向迁回时，系统应保留有原系统的备份，可以实现在线迁移。用户邮件应保证完整地发送给接收方，如果对方邮件服务器暂时不可用，应在一段时间内自动重试；如果发送失败，应回馈用户失败信息。

大容量邮件系统应具有很好的可扩展性，在集群系统中我们既可以通过升级现有服务器增加服务能力，也可以动态地增加服务器数量。

3）完备的管理功能　提供对集群系统多个管理层面的管理模块，既包括对集群系统中各节点资源的统一管理、性能的监测，也包括对邮件系统的管理，如邮箱管理、群体配额等。

4）安全性和垃圾邮件的防范　建立统一的用户认证机制，实时邮件过滤，安全协议等。

5）多协议兼容　将简单邮件传输协议 SMTP 用于引入的邮件，POP3 和 IMAP 用于取回邮件。和 POP3 协议相比，

梁建民　博士研究生，主要研究方向：基于 Linux 的集群。祝明发　博士生导师，主要研究方向：高性能服务器。梁小萍　主要研究方向：计算机辅助设计。

IMAP 协议的好处在于可将用户个人的邮件分类地保存在邮件服务器的个人目录中，而不用下载到本地硬盘中，尤其适合于那些没有固定个人电脑或者经常出差在外的用户。邮件系统应同时提供这两种服务，以满足不同用户的需要。系统还应该支持 Web 方式的邮件收发。

4. 大容量邮件系统的关键技术分析

实现大容量邮件系统的关键包括如何解决邮箱的合理分布、大用户量并发访问的负载均衡、采用统一的认证模式既保证系统的安全性，又不使认证模块成为系统的瓶颈，以及邮件的高效存取等。

4.1 负载均衡策略

所谓负载均衡，就是对传输流进行调节和管理，把传输流动态地分布到使用相同应用程序的一组服务器上，负载均衡系统对这组服务器进行监控，并做出决策，选择性能最优、可用性最好的传输路由，从而为最终用户提供优质、可预测的服务质量。在大容量邮件系统中，负载均衡模块是整个系统和用户请求联接的纽带，直接影响整个系统的性能。

负载均衡可以用硬件实现，比如用交换机或路由器在硬件 MAC 层负载均衡，实现静态内容的快速均衡，但这种结构有不少局限性，例如，需要在应用层(Layer4)处理的每个数据包都必须被打开进行检查，以决定它们的目的端口。完成这种工作必须使用交换机上的中央处理器，因而降低了交换机的性能。另外，硬件实现方式常常缺少 SSL 会话、ID 跟踪、用户验证和应用安全检查等功能，且硬件方式的实现也比较昂贵。

相比之下用软件实现负载均衡更加灵活，网络管理员可以对服务器进行高度管理，如分析 CPU 内存使用情况，执行基于代理的内容管理，传输流也不必经过额外的设备，因此在理论上可以节省费用、提高性能。但是，软件方式应该考虑的问题是如何减少负载均衡对服务器速度和性能的影响，如何避免软件本身成为服务器上的瓶颈。软件实现负载均衡的方式一般来说有以下几种：

1)*基于应用层的负载均衡调度方法* 是将到达的用户连接通过应用层服务器转发到后台不同的服务器，取得回应后，再由应用层服务器返回给用户。该方法存在的问题是：从请求到达至处理结束，调度器需要进行四次从核心与用户空间的切换，从用户到转发服务器和转发服务器到后台服务器的两次 TCP 连接，系统处理开销特别大，致使系统的伸缩性有限，这种方式效率比较低。

2)*基于轮询的 DNS 方法* 这种方法是通过前端的 DNS 服务器把相同的域名轮流解析到不同 IP 地址，使负载分配到不同的服务器，从而提高整个系统的性能。但该方法存在的问题是：域名服务系统是按层次结构组织的，各级域名服务器都会缓冲已解析的名字到 IP 地址的映射，妨碍轮询方法在客户端生效，从而导致后台不同服务器间的负载不均衡。该方法适合于长期的负载均衡[5]。

3)*基于 IP 层的负载均衡调度方法* 在 IP 层进行负载均衡调度的优点是所有的操作都可以在操作系统核心空间中完成，所以调度开销很小，即使调度很多后台服务器，前端负载均衡模块也不会成为系统的瓶颈。例如 Linux 下的虚拟服务器系统(LVS)[2]。

LVS 在操作系统核心空间中将 IP 层上的 TCP/UDP 请求均衡地转移到不同的后台服务器上，前端负载均衡模块自动屏蔽掉后台服务器的故障，从而将一组服务器构成一个高性能的、高可用的虚拟服务器。LVS 有三种方式：VS-NAT (Virtual Server via Network Address Translation)、VS-DR (Virtual Server via Direct Routing)、VS-TUN(Virtual Server via IP Tunneling)。

VS-NAT 是通过网络地址翻译的方法转发用户数据包，用户的请求报文到达负载均衡模块时，负载均衡服务器根据各个后台服务器的负载情况，动态地从中选出一台，将报文的目标地址改写成选定后台服务器的地址，报文的目标端口也改写成选定后台服务器的相应端口，然后将报文发送给后台服务器，同时负载均衡服务器在维护的 Hash 表中记录用户和该服务器的连接，当这个连接的下一个报文到达时，从 Hash 表中可以得到原选定服务器的地址和端口，进行同样的改写操作，并将报文传给原选定的后台服务器。后台服务器的回应报文经过负载均衡服务器时，源地址和源端口被改为负载均衡服务器的地址和相应的端口，再把报文发给用户。当连接终止或超时，负载均衡服务器将这个连接的记录从 Hash 表中删除。这种方式的缺点是，请求和应答的数据包都需要通过负载均衡服务器重写，当后台服务器的数量增多时，负载调度器会成为集群系统新的瓶颈。

VS-TUN 利用 IP 隧道技术将请求报文封装转发给后台服务器，响应报文从后台服务器直接返回给客户。由于后台服务器有一组而不是一个，因此我们不可能静态地建立一一对应的隧道，而是动态地选择一台后台服务器，前端负载均衡模块将请求报文封装在另一个 IP 报文中，再将封装后的 IP 报文转发给选出的后台服务器；那台后台服务器收到报文后，先将报文解封获得原来目标地址，当发现报文的目标地址被配置在本地的 IP 隧道设备上，就处理这个请求，然后根据路由表将响应报文直接返回给客户。这样，我们可以利用 IP 隧道的原理将一组服务器上的网络服务组成在一个 IP 地址上的虚拟网络服务。

VS-DR 则将报文直接路由给目标服务器。负载均衡服务器根据各个服务器的负载情况，动态地选择一台后台服务器，不修改也不封装 IP 报文，而是将数据帧的 MAC 地址改为选出服务器的 MAC 地址，服务器收到该报文后，根据路由表将响应报文直接返回给客户。

三种方式中 VS-NAT 效率比较低，VS-TUN 技术受操作系统限制，要求所有的服务器必须支持"IP Tunneling"协议，相比之下，VS-DR 性能好，配置也比较简单。

4.2 大容量邮件系统的邮件存储策略

共享式集中存储 大容量邮件系统可以通过光纤通道和网络文件系统如 NFS、AFS、Coda 等实现全局的存储空间共享。存在的问题是：标准邮件存储格式为 mbox 形式。在这种格式下，多个邮件都保存在同一个文件中，因此进行邮件操作就必须加锁，以保证没有访问冲突，而 NFS 缺乏文件锁定机制，这就使得 NFS 存储方式不适合并发访问；另一个问题是，传统的方式使用一个单一的目录保存所有用户的邮件，在用户数量较多时文件系统的性能变差。

对第一个问题，我们采用 Qmail 的 Maildir[3]存储方式，每个邮件作为单独的一个文件保存在用户个人的邮件目录下面，这样就避免了加锁，具有安全、可靠的特点，缺点是小文件浪费磁盘空间。对第二个问题我们采用多级目录，每个目录下存储有限数量的文件，有效降低打开文件时的系统消耗。也可以采用数据库形式保存邮件，但由于邮件的操作大多为文件操作，且邮件大小变化较大，会造成性能和存储空间的浪费。

NFS 也可以用 Coda[1]文件系统代替，Coda 系统给集群中的计算机定义了三种角色：客户端，服务器端、系统控制服务器(SCM)。客户端即邮件服务器模块，服务器端可以看作给邮件服务器提供数据的后台存储模块，SCM 为 Coda 全局文件系统的管理提供单一的控制点。客户端运行有 venus 进程，提供服务器端数据的缓存，venus 和服务器端进程交互，保证缓存中数据的实时性。和 NFS 方式相比，Coda 的缓存机制可以使数据读取速度显著提高。

分布式存储 和共享式存储相比，分布式邮件存储具有更好的可扩展性和可靠性，每个邮件服务器只服务一部分用户，可以有效避免系统的单一故障点；I/O 操作只涉及到本地的存储设备，使服务的效率得到提高。

分布式邮件存储不通过网络访问其它服务器，因此可以采用任意的邮件存储格式，但同时也带来了新的问题：如何使用户的服务请求准确、可靠地转发到用户所在的邮件服务器，如何使用户能均匀地分布到不同的邮件服务器，保证系统的负载均衡。

对于上面的问题，在后面的具体设计中我们会详细介绍。对于第一个问题，我们在负载均衡模块和邮件服务模块之间增加一层路由模块完成用户认证和请求转发功能；第二个问题，通过在路由模块中的索引服务器上设置用户分布规则也可以解决。

4.3 大容量邮件系统的用户认证策略和系统维护策略

Linux 的标准邮件系统不适合大容量的邮件服务，有的系统用户标识只有 16 位，因此用户数量最多只有 64k，即使系统支持 32 位的用户标识，单台服务器用户数量也不能超过 10 万。Sendmail[4]邮件服务体系采用主机自身的认证系统，即邮件用户也是系统用户，每次认证都需要解析/etc/passwd 文件，在大量用户同时访问的情况下认证效率会大大降低，无法满足大容量的邮件服务的要求。

考虑到安全性、性能和可管理性，我们采用非系统用户作为邮件用户。Qmail[3]提供了虚拟域和虚拟用户的机制，但它缺乏良好的管理工具，通过安装配置专门配合 Qmail 的 Vmailmsg 软件包，能增强 Qmail 的口令验证功能，使用户可以通过 POP3 及 IMAP 访问自己的虚拟邮箱。

大容量邮件系统的系统维护除了对邮件系统的有效管理之外，还包括安全的邮件迁移。邮箱迁移一般发生在集群中邮件服务器出现故障或某个服务器负载过重，它的一部分负载应该迁移到其它服务器上；或者一个新的服务器加入也有必要在服务器之间重新分布用户，因此，每个邮件服务器的使用率和邮件服务器的每个用户的使用率需要有一个度量，一个简单的办法是通过各自的日志文件统计。可以通过一个脚本收集统计信息，内容应该包括：通过 SMTP 传送的信息数，这里是考虑到接收 SMTP 信息需要的系统资源；引入 SMTP 邮件的大小，这里是考虑到处理较大邮件需要的附加系统资源；POP3 请求的数目；邮箱的大小，这里是考虑到在用户邮箱中扫描所需的资源。

5. 一种基于集群和分布式存储的层次型大容量邮件系统结构

通过以上的讨论，我们给出一种基于集群和分布式存储的层次型大容量邮件系统结构如图 1，这里我们把邮件系统划分为四个功能模块：

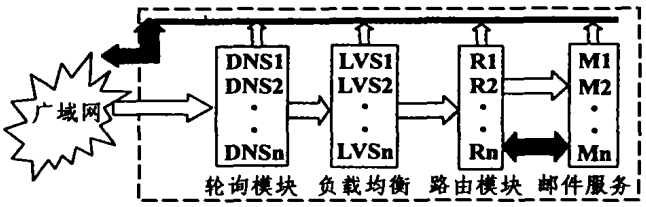

图 1 基于集群和分布式存储的层次型大容量邮件系统结构

轮询模块包含一组 DNS 服务器，通过为单一域名设定多个不同的 IP 地址把用户的请求定向到负载均衡模块的不同主机上。

负载均衡模块由一组运行 LVS 的转发服务器组成，转发服务器上运行有监视路由模块中服务器负载、结点的健康状况的进程，当路由模块中服务器对 ICMP ping 不可达或网络服务在指定的时间没有响应时，监视进程通知操作系统内核，将该服务器从调度列表中删除或者宣告失效。这样，新的服务请求就不会被调度到坏的结点。转发服务器根据路由模块中服务器的实时负载情况，决定用户请求的转发目标。

路由模块由一组索引服务器组成，索引服务器保存了用户的基本信息和用户邮箱所在服务器的信息。这部分的工作首先是完成用户身份认证、安全检查，然后把用户的请求转发给邮件服务模块，并使用户和邮件服务器建立直接的连接。由于用户基本信息只占很小的空间，保存用户信息可以采用 LDAP 目录服务，系统不需要保存多个口令的拷贝，只在 LDAP 里保存一份就可以了，使邮件系统具有良好的可扩展性。另外，在索引服务器上设置用户分布规则可以使用户均匀地分配到邮件服务模块的不同邮件服务器，比如可以把邮件地址的首字母为 a-k 的邮件发到一个服务器，而其它邮件地址发到其它服务器，在 sendmail 中可以通过简单的配置实现邮件分割。

邮件服务模块由一组邮件服务器组成，每台服务器为一组不同的用户服务。路由模块和邮件服务模块之间有快速的专用通道，当邮件服务模块的用户信息改变时，路由模块的 LDAP 服务器应得到及时的更新。当用户通过路由模块的身份认证、安全检查，用户就和特定的邮件服务器建立起直接的连接。

结束语 集群技术和分布式存储技术避免了系统的单一故障点，提高了大容量电子邮件系统的可用性和可扩展性；层次结构的数据分流，使系统的流量合理地分布到不同邮件服务器上，加快了系统对用户请求的响应速度；基于内部网的路由模块完成用户认证工作，既减轻邮件服务器的负担，又进一步加强了系统的安全性。

邮件用户划分到不同的邮件服务器的规则和方式还有待于进一步探讨，系统虽然是针对邮件服务的设计，同样的结构也可适用于其它大用户量、I/O 密集型的服务中。

参 考 文 献

1 http://www.coda.cs.cmu.edu
2 http://www.Linuxvirtualserver.org
3 http://www.Qmail.org
4 http://www.sendmail.org
5 Cardellini V, Colajanni M, Yu P S. Dynamic Load Balancing on Web-server Systems. IEEE Internet Computing 1999,3(3):28～39