邮件系统高可用性系统切换的设计与实现

The Design and Supplement of Email Dual System's Manual HA

春增军　（深圳市广东核电集团信息技术中心　518124）

摘要：标准的双机系统都是在磁盘阵列（Disk Array）和集群（Cluster）软件的支持下实现的。而邮件系统的双机一般更是需要他们的支持。本文详细论述了如何在没有集群（cluster）软件和共享磁盘整列的条件下实现两台 SUN 邮件服务器通过手动切换实现双机高可用性（HA）的解决方案。此方案的特点是结构简单、易管理、性价比较高，而且相当于双主机、双阵列，并具有良好的可靠性，对中小企业来说也是一个很好的解决方案。

关键词：集群（cluster）　高可用性（HA）　邮件服务器　目录服务器（LDAP Server）　镜像　手工 HA

1　引言

为了给用户提供持续稳定的电子邮件服务，一般 ISP 及大型企业的电子邮件系统均采用基于集群（Cluster）软件及磁盘阵列的双机甚至多机系统。但对于中小企业来说，集群（Cluster）软件及磁盘阵列的费用过于昂贵，同时集群（Cluster）软件管理起来也比较复杂。如何在没有集群（Cluster）软件及磁盘阵列的情况下实现双机服务，是一个很有意义的技术问题，对中小企业来说也是一个性价比很好的解决方案。

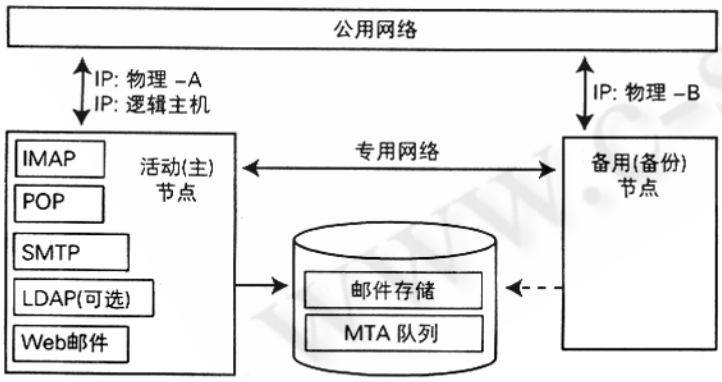

图 1　基本不对称或"备用"高可用性模型

本文以最新的 Sun Messaging Server 6.0 为例，说明如何实现邮件系统的手工高可用性（High Availability，HA）系统切换。

2　SUN 高可用性解决方案

SUN 邮件系统推荐的双机方式有三种：不对称（备用）、对称、N +1（N Over 1）。

2.1　不对称（备用）系统

基本不对称或"备用"高可用性模型（图 1）由两个群集主机或"节点"构成。这两个节点被指定了一个逻辑 IP 地址和关联的主机名。

在这种模型中，只有一个节点在给定时间内处于活动状态，备份或备用节点大部分时间内处于空闲状态。这两个节点之间的单一共享磁盘阵列由活动节点或"主"节点配置和管理。邮件存储分区和邮件传输代理（MTA）队列就驻留在这个共享的卷上。

故障切换之前，活动节点为 Physical – A。故障切换期间，Physical – B 成为活动节点，并且会切换共享的卷，以便由 Physical – B 管理该卷。Physical – A 上的所有服务都将停止，并在 Physical – B 上启动。

此模型的优点在于，备份节点是专用的并且是完全为主节点保留；故障切换发生时，备份节点上不存在资源争用。然而，此模型也就意味着备份节点大部分时间内处于空闲状态，因此资源利用率很低。

2.2　对称系统

基本对称或"双重服务"高可用性模型由两个主机构成，每个主机都有自己的逻辑 IP 地址。每个逻辑

节点都与一个物理节点相关联，并且每个物理节点都控制一个具有两个存储卷的磁盘阵列。一个卷用作其本地邮件存储分区和 MTA 队列，另一个卷是其同伴的邮件存储分区和 MTA 队列的镜像。

在对称高可用性模式（图 2）下，两个节点同时都是活动的，并且每个节点都是彼此的备份节点。正常情况下，每个节点只运行邮件传送服务器的一个实例。

故障切换期间，会关闭有故障的节点上的服务，并在备份节点上重新启动这些服务。此时，备份节点将同时从这两个节点运行 Messaging Server 并同时管理两个单独的卷。

此模型的优点在于，两个节点同时处于活动状态，因此能充分利用计算机资源。但是，在故障期间，备份节点中会存在较多的资源争用，因为它要同时从两个节点运行 Messaging Server 的服务。因此，您应该尽快修复有故障的节点并将服务器切换回其双重服务状态。

此模型还提供了一个备份存储数组；如果磁盘阵列发生故障，备份节点上的服务可以拾取其镜像。

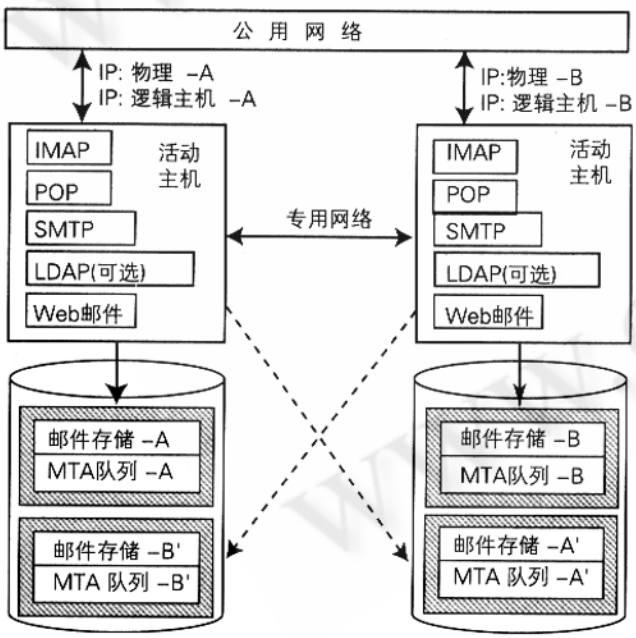

图 2　对称高可用性模式

2.3 N+1（N Over 1）系统

N + 1 或"N over 1"模型在多节点不对称配置下运行。需要 N 个逻辑主机名和 N 个共享磁盘阵列。单一备份节点被保留为所有其它节点的备用节点。备份节点能够从 N 个节点同时运行 Messaging Server。

在一个或多个活动节点的故障切换期间，备份节点将承担有故障的节点的工作。

N + 1 模型的优点在于，可以将服务器负载分发到多个节点上，并且只需一个备份节点承担所有可能的节点故障。因此，计算机空闲比率为 1/N，而单一不对称模型情况下为 1/1。

2.4 高可用性模型的优点和缺点

SUN 公司三种高可用性模型也属于标准的集群系统模型，其优缺点如下表 1。

表 1　三种高可用性模型的优点和缺点

模型	优点	缺点	建议采用的用户
不对称	·配置简单 ·备份节点 100% 保留	·不能充分利用计算机资源	计划将来扩大规模的小型服务提供商。
对称	·系统资源使用率较高 ·可用性较高	·备份节点上存在资源争用 ·镜像的磁盘会降低磁盘写入性能	近期内不计划扩展备份系统的中等规模的服务提供商。
N + 1	·负载分布 ·易于扩展	·配置复杂	需要不受限制地分布资源的大型服务提供商。

3　手工 HA 系统设计

通过以上分析，我们已经了解了标准集群系统的原理。要在没有集群（Cluster）软件及磁盘整列的情况下实现类似双机的工作模型，关键是按双机的工作原理设计系统。以下案例我们主要按主备（不对称）系统架构设计。

建设我们两台服务器的主机名为 jt－mail1 和 jt－mail2，虚拟主机名为 jt－mail，则系统架构如图 3。

系统结构说明如下：

（1）两台机器（jt－mail1、jt－mail2）的整体结构设计为主备方式，始终以虚拟机器名 jt－mail 的方式给用户提供 Internet 邮件服务，当主机故障时，通过简单的手工命令切换到备机提供有奖服务，即实现手工 HA 动态切换。其架构相当于 SUN 的标准方案一（不对称），但没有共享磁盘阵列及集群（Cluster）软件。

（2）两台机器都安装了标准的 LDAP 服务器：Sun ONE Directory Server 5.2，配置成双主（Dual Master）模

式,可互相复制目录服务器数据(如帐号信息)。保证了用户帐号信息在两台机器上的实时同步。

(3)两台服务器均装安装同样的邮件系统:Sun Messaging Server 6.0。邮件系统的配置信息及邮件数据在两台机器上都存放在单独的文件系统下(类似于标准 SUN 标准集群系统 2 中对称模式中两个服务器共享两个阵列),以镜像的方式实现两台机器数据同步。

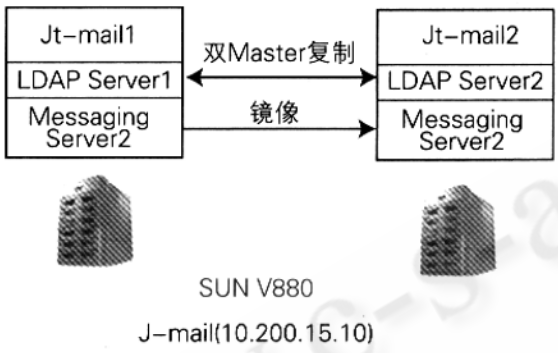

图 3　手工 HA 双机系统结构

(4) 主机 jt - mail1 的 IP 为 10.200.15.1,备机 jt - mail2 的 IP 为 10.200.15.2,用户始终通过虚拟机器名 jt - mail(IP 为 10.200.15.10)访问邮件系统。

4 系统实现

4.1 软件安装

(1) 系统安装。两台 SUN V880 安装 Solaris 9,用于邮件系统的 10 个 10000 转 73GB 光纤盘以 Raid 1 +0 方式实现。提供最大可能的冗余。(SUN 的手册上只写了支持 Raid 1 +0,未写如何实现。SUN 工程师一般只作过 Raid 0 +1,通过理论分析和反复实践,我们在 Solaris 9 上利用系统自带的 SUN Volume Manager 软件实现了 Raid 1 +0)。

(2) 邮件系统(Sun Messaging Server 6.0)安装。在两台机器上同时安装邮件系统软件,同时注意单独设立/home1/msgstore 目录用于存放邮件系统配置信息和邮件数据,相当于在主机上模拟磁盘阵列的存储方式。

4.2 配置目录服务的复制

两台机器上的目录服务器(LDAP Server)配置成双主(Dual Master)方式,可互相实时复制数据(用户帐号信息),保证 jt - mail1 故障时,jt - mail2 上有完整的用户帐号数据信息。

实现方式:

经过反复试验,以下项目在目录服务中属于公共信息,需定义数据复制:

- o = internet
- o = pab
- o = usergroup

而目录服务中以下项目属两个目录服务专有信息,不能定义复制:

- dc = dnmc,dc = com,dc = cn,dc = cn(缺省域信息)
- o = comms - config
- o = NetscapeRoot

4.3 启用逻辑 IP(两台机器上完成)

使用安装软件自带的 admin_ip. pl 实用程序将物理主机的 IP 地址更改为 Administration Server 逻辑主机的 IP 地址。需在两台主机上同时运行此实用程序。

- #perl admin_ip. pl ' cn = Directory Manager' xxx 10.200.15.1 10.200.15.10
- #perl admin_ip. pl ' cn = Directory Manager' xxx 10.200.15.2 10.200.15.10

注:启用逻辑 IP 以后,使用户通过逻辑 IP(10.200.15.10)访问邮件系统成为可能,为手工 HA 双机方式实现提供可能。

4.4 启用逻辑主机名(两台机器上完成)

最终用户是通过主机名而非虚拟 IP 访问邮件系统,所以需配置邮件服务器绑定虚拟主机名。

- #./configutil - o local. hostname - v jt - mail. dnmc. com. cn

以上命令执行后,邮件系统配置文件需重新编译,系统才能正常工作。

4.5 完成手工 HA

至此,已基本完成邮件系统本身的手工 HA 配置,这时可通过 http://jt - mail 或 http://10.200.15.10 以 WEB 方式收发邮件。

4.6 故障处理

正常情况下公共地址 10.200.15.10(jt - mail)是 10.200.15.1(jt - mail1)的虚拟地址,而这时也可以通过两台机器的物理地址 http://10.200.15.1 和 http://10.200.15.2 收发邮件,所以故障时只需在备机(jt -

mail2)上启动公共地址 10.200.15.10 即可实现手工 HA 切换系统:

ifconfig ge1:2 up

注:只需上面这一个命令即完成主机故障时手工 HA 切换系统。

手工 HA 方式与标准双机模式的差别见下表 2。

表 2　手工 HA 方式与标准双机模式的差别

	优点	缺点
手工 HA	·费用低 ·管理简单实用 ·相当于双主机、双阵列,两份数据	·故障时不能自动切换
标准双机	·故障时自动切换	·需 Cluster 软件和磁盘阵列,费用高 ·管理复杂,可能某些原因引起 Cluster 软件本身出错

4.7 数据镜像的实现

(1) 镜像软件的选择。以上方式保证了在主机故障时,备机在手动 HA 切换后可继续为用户提供邮件服务。但毕竟我们没有共享磁盘证列,用户看不到自己以前的邮件数据。如何在主机故障时保证用户依然能看到自己的邮件数据?

此问题通过镜像技术实现。

对于选择 Unix 作为应用平台的的中小型企业或网站来说,往往面临如何实现数据远程备份或者网站镜象的问题,虽然有商业化的备份和镜象产品可供选择,但这些产品的价格往往过于昂贵。

通过网络进行远程数据备份或者网站镜象的最简单的方法就是使用 wget,但是这种方式每次都需要将所有数据都重新在网络上传输一遍,而不考虑哪些文件是经过更新的,因此效率非常低下。尤其在需要备份的数据量很大的时候,往往需要花费数个小时来在网络上进行数据传输。

我们采用一种高效的网络远程备份和镜象工具 – rsync,它可以满足绝大多数要求不是特别严格的备份需求。

(2) 客户机 jt – mail2 配置

测试:

/usr/local/bin/rsync – vzrtopg —delete—progress mail@10.200.15.12::msgstore /home1/msgstore/store/partition —password – file =/etc/rsync.pass

上面这个命令行中 – vzrtopg 里的 v 是 verbose,z 是压缩,r 是 recursive,topg 都是保持文件原有属性如属主、时间的参数。——progress 是指显示出详细的进度情况,——delete 是指如果服务器端删除了这一文件,客户端也相应把文件删除,保持真正的一致。

mail@10.200.15.12::msgstore 表示对该命令是对服务器 10.200.15.12(jt – mail1)中的 msgstore 模块进行备份,mail 表示使用 mail 来对该模块进行备份。

——password – file =/etc/rsync.pass 来指定密码文件,这样就可以在脚本中使用而无需交互式地输入验证密码了,这里需要注意的是这份密码文件权限属性要设得只有 root 可读。

(3) 配置说明。目前 jt – mail1 和 jt – mail2 均有 2 个千兆以太网卡(ge1、ce0),用户通过 ge1 千兆以太光卡访问邮件服务器,所以以上配置采用另一个 ge1 的热备(IPMP 技术)千兆以太网卡 ce0(10.200.15.12、10.200.15.14)来实现数据镜像。

(4) 规则

针对 rsync 软件的特点及应用需求,制定如下规则:

· 复制用户更新数据:即当主机上收到一份邮件时,复制到备机上;当主机上删除一份邮件时,备机上的邮件也删除。

· 复制频率:10 分钟/次

5 小结

通过以上设计和实现步骤,即可以实现在没有集群软件和共享磁盘阵列的条件下两台邮件服务器的手工 HA 切换。

参考文献

1 Sun Java Enterprise System 2003Q4 Installation Guide, Sun Microsystems, Inc.

2 Sun ONE Messaging Server 6.0 Patch 1 Update 1 Release Notes, Sun Microsystems, Inc.

3 Sun ONE Messaging Server 安装指南, Sun Microsystems, Inc.