全路径高可用 Web 应用系统的设计与实现①

Design and Implementation Entirely High Available Path through Web Application System

陈　鍊　黄　兵　（南京审计学院 信息科学学院 南京 210029）
李冠强　（南京审计学院图书馆 江苏 南京 210029）

摘　要: 为了构建数字图书馆的高可用 Web 应用系统，提出安全可用的四层物理子网体系结构，并且从零部件、系统数据到系统设备，特别是数据链路各个层次，综合运用 SAN、集群、VRRP、端口热备等各种容错技术，实现数字资源访问路径全程的高可用性。该系统已稳定运行一两年，可供各类 Web 应用系统的建设改造参考。

关键词: 网络规划　高可用性　容错技术

计算机和网络技术的应用改变了商务、政务、公共事务等许多社会行业、组织机构的运营方式，例如数字图书馆就以全天候全方位方式提供超越所有传统业务的文献信息和知识服务。新系统方便快捷不受时空限制的诸多优点导致传统运营方式、运营机构几乎废弃，不可逆转的现象迫使新系统的安全可用性变得至关重要，直接关系系统服务功能能否发挥效用和系统业务的连续不间断程度。本文以数字图书馆 Web 系统建设为例，通过系统四层物理子网体系结构的设计保障系统的安全性，然后从零部件、系统数据到系统设备，特别是数据链路各个层次，综合运用各种容错技术，全方位实现 Web 应用系统的数字资源访问路径全程高可用性。

1　高可用性及其技术方案

系统的可用性（Availability），是在规定条件下系统按规范成功运行、提供功能服务的能力。从数学视角看，可用（性程）度是时间函数 A（t），定义为在规定工作条件下系统在时刻 t 按规范正常提供功能服务的概率。用此术语，系统的高可用性 HA（High Availability）就是系统中断服务的时间期望值为无穷小，系统提供功能服务时间趋于无限，即处于随时可使用工作状态、能提供 365 ×24 小时不间断服务。通常用二元元件模型描述系统可用性 A[1]：

$$A = MTBF / (MTBF + MTTR)$$

其中可用性参数 MTBF 为平均无故障时间（可维修系统两次故障之间平均时间，Mean Time Between Failure），MTTR 为平均修复时间（Mean Time To Recover）。一般采用并联备份可能失效（Fault）元件的冗余设计方法提高系统的可用性，这样，根据马尔可夫（Markov）过程模型分析，可以推得 n 个节点组成的冗余多机系统的可用度为[2]：

$$A_n = 1 - (\lambda/(\mu + \lambda))^n$$

其中 λ 为节点的故障率，μ 为故障修复率（平均修复时间的倒数 1/MTTR）。由此可知，增加节点数 n、降低故障率 λ、缩短修复时间 MTTR，均可提高可用度 A_n；并且，当 n 取值 2 时效果最明显[1,2]，即冗余双机系统的（可用度/价格）比值可以达到最大。例如，节点的故障率 λ 为 1 次/24 小时，平均修复时间 MTTR 为 5 分钟，冗余双机系统的可用度 A2 可以达到 99.999%[2]。

2　安全可用 Web 应用系统的规划设计

安全性（Security）与可用性既有联系、又有所区别，安全性是可用性的前提、基础，为可用性奠定最基

① 基金项目 江苏省高校自然科学项目（02KJD120001）；南京审计学院教学研究项目（J2006B11）

本的使用环境,不安全的系统必然是不可用的;从安全角度看,可用性就是可被授权实体访问并按需求使用的特性[3],可用性与保密性(Confidentiality)、完整性(Integrality)是安全性的三大基本属性,是安全性通过保密、完整的质量属性实现的目标属性。简而言之,安全性是基础,可用性是目标。作为数字图书馆基础设施的图书馆 Web 应用系统,既要保证数字资源的安全可用,又要为用户提供一个开放友好的服务环境,为了实现似乎矛盾的规划目标,Web 应用系统设计成四层子网结构,如图 1 所示,由数据存储区域网——资源内网——用户外网——因特网(公网)等物理子网络组成。

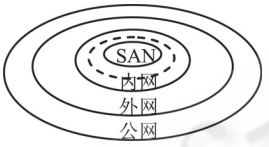

图 1　高可用 Web 应用系统体系结构

2.1 数据存储区域网 SAN

数据存储区域网是图书馆网络系统,更确切地说,是网络化图书管理信息 Web 应用系统的数据网,是数据存储核心区域。存储区域网 SAN(Storage Area Network)是一种由存储设备和基于 FC(Fiber Channel)协议的光纤设备组成的物理网,这种网络存储结构使主机和存储系统从物理连接和功能上都独立出来,分别组成自己的物理网络,可以获得更高效更安全的高可用、高可扩存储子系统[4]。

2.2 图书馆资源内网

资源内网是 Web 应用系统资源区、图书馆工作内网、系统安全区,除少数图书管理系统工作终端和用户资源查询无盘终端系统,主要由服务器主机组成,包括各种数据库系统后台数据库服务器及前台 Web 服务器的主机。事实上,后台数据库 DB 服务器与前台 Web 服务器形成未物理隔离的两个子网。资源内网的组成必须严格遵循两条组网原则:

(1)资源网的边界必须设置功能尽可能完善的多功能安全网关,形成物理隔离屏障,有条件开放个别端口,无条件关闭所有端口,阻挡 L2 ~ L7 层各种安全威胁,把守住整个图书馆资源的安全要塞;

(2)安全网关只允许从外向内的有限访问、禁止任何从内向外的访问。所有从外向内的访问都经过安全网关的地址映射 NAT(Network Address Translation)或端口映射 PAT(Port Address Translation)和访问控制 ACL(Access Control List),拦截所有非法访问、保证内网资源安全。

2.3 图书馆用户外网

用户外网由 DMZ 区和园区网两部分组成。个别不支持地址、端口转换的数据库应用系统和图书馆网站部分数据组成图书馆网络系统的非军事缓冲区 DMZ(Demilitarized Zone);安全度与之相当的园区网也看作 DMZ 区。这个区域的安全原则正好与资源内网相反:无条件开放所有用户权限,有条件限制极少数类别用户权限。对于单位的图书馆来说,这个子网区域就是本单位的园区网,为了保证这个子网区域的用户能正常使用所有软硬件资源,公共图书馆边界就须要再添置一台安全网关,阻挡公网上泛滥成灾的病毒、黑客和各种垃圾。

2.4 公网

教育科研网 CERNet、因特网等公网是不安全区域,用户可以访问无版权信息,并设有 VPN 网关的认证登录门户。

如上所述,系统设置两道安全网关,划分数据、资源和用户三个物理子网,分离用户与资源、资源管理系统与资源数据,不仅可显著提高系统资源的安全程度、简化网络安全管理,而且也使图书馆的软硬件资源使用的开放程度达到及至,并且为系统高可用性、高可扩性的实现奠定了基础。

系统拓扑如图 2 所示,为了便于综合布线,视具体情况,内外网可以设计成"接入层——核心层"的两层结构,或者"接入层——汇聚层——核心层"的三层结构。所有设备都以 VLAN 形式接入。资源内网与用户外网间的安全网关选用多功能安全网关 UTM,定位为网络边界多功能"一体化威胁管理"的 UTM(Unified Threat Management)是一种新型的多功能安全网关系统,采用先进的高可靠性嵌入式硬件平台,高度集成防病毒、防火墙、防入侵"老三防"安全技术于一身,实现单一设备对网络 L2 到 L7 全协议层的防护。基于相对陌生的光纤通道 FC(Fiber Channel)接口标准、64 位 RISC(Reduced Instruction Set Computer)处理器与 ASIC

(Application Specific Integrated Circuits)数据校验芯片架构的专用控制处理器、存储逻辑单元 LUN(Logic Unit Number)屏蔽和 FC 交换机通道分区(Zoning)的端口级屏蔽等安全技术,使得 FC SAN 相对于现行网络几乎可以说‘足够安全”,而基于 IP 的 iSCSI SAN 虽然具有易于融入现有网络应用的许多优点,但其安全程度令人担忧[5]。数据存储区域网 SAN 由基于 FC 协议的光纤交换机作为接口界面与外界隔离,形成数据存储区域网又一道安全网关。这些安全因素是设备选型不可忽略的关键。

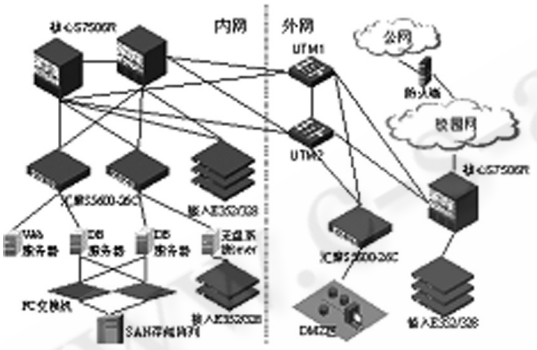

图 2　高可用 Web 应用系统拓扑

3　高可用性 Web 应用系统的实现

从容错技术角度来看,可用性就是通过冗余系统、解决系统单点故障 SPOF(single point of failure)的技术,不同的技术可以实现不同级别的可用性:

(1)级别 0:常规可用性,系统零部件冗余,实现零部件故障的容错能力,可用度 A 可以达到 99%,相当于年平均故障时间 87.6 小时(3.65 天);

(2)级别 1 数据可用性,系统数据冗余,实现中断系统运行、恢复系统数据的容错能力,可用度 A 可以达到 99.9%,相当于年平均故障时间 8.76 小时;

(3)级别 2:系统可用性,系统级设备冗余热备(Failover),实现中断系统服务、系统级设备的故障屏蔽切换容错能力,可用度 A 可达 99.99%,相当年平均故障时间 53 分钟;

(4)级别 3:系统高可用性,关键子系统集群(Cluster),实现几乎不中断系统服务、系统级设备及其相关数据链路(包括设备和线路)的故障自动切换和恢复容错能力,可用度 A 可达 99.999%,相当于年平均故障时间 5.3 分钟;

(5)级别 4 系统持续可用性,全系统集群,实现无任何切换或恢复造成系统瞬间的运行停顿和服务中断。

显然,低级别可用性是高级别可用性的实现基础,例如,数据可用性是系统可用性的基础[6]。根据‘以较低价格、通用设备、获得较高的可靠性”的 HA 系统设计原则[7],为实现 Web 应用系统资源访问全路径的高可用性(年均停机 5.3 分钟),主要选择影响系统业务、组成数字资源访问路径的关键零部件、关键设备、关键数据链路增加冗余热备份。

3.1　关键零部件的冗余容错

与用户数字资源相关链路的所有设备的电源、风扇、磁盘、关键引擎等易损关键部件都冗余备份、支持热插拔,中心机房通过高达 8h×40KVA 的不间断电源 UPS 和双制冷循环系统的精密空调提供稳定的运行环境,以此降低相关设备自身故障率、提高其 MTBF。

3.2　系统数据和关键设备的冗余容错

磁盘阵列系统的冗余控制器及其后端磁盘连接方式、前端冗余的光纤交换机和服务器冗余的适配器 HBA(Host Bus Adapter)卡等数据存储区域网 SAN 的硬件组成构造,本身就组成高可用性的存储系统[8,9]。RAID(Redundant Array of Inexpensive / Independent Disks)存储阵列支持 RAID0、1、5 多种存储方式,尤其是条带具有分布式校验块的 RAID5,即使整块物理盘损坏,换块新盘系统即可自动修复;还有 SAN 系统软件支持的动态扩容、本地镜像、远程复制等一系列高可用功能,为系统数字资源数据可用性提供了保障。

资源内网核心交换机冗余热备(Failover)由虚拟路由冗余协议 VRRP(Virtual Router Redundancy Protocol)实现[10]。如图 2 所示,配置相同的两台核心交换机通过 VRRP 组成具有虚拟 IP 地址的虚拟路由器,并以主控通告(Advertise)消息维持相互关系,所有服务器的缺省网关指向虚拟路由器地址配置为静态路由,当主控路由器失效时,VRRP 可以透明地切换优先级别最高的备份路由器升级为主控路由器,从而实现核心交换设备级故障屏蔽切换容错。由于交换机不支持负载均衡主主(Active / Active)模式,两台配置相同的 UTM 通过设备提供的 HA 功能,相互监测心跳(Heartbeat)工作在主从 A/P(Active / Passive)模式,实现安全网关系统设备级冗余热备。为保证监测可靠性配置

冗余的两条心跳线路。

3.3 服务器冗余容错问题

Web 应用系统的存储、服务器、核心交换机、安全网关等 4 个节点的系统设备都是关键设备，其中安全网关和核心交换冗余容错的目标侧重系统安全、访问通信可达，服务器冗余容错的目标侧重访问流量均衡、响应及时，存储冗余容错（SAN 就是存储集群）的目标侧重数据安全可用，全部节点数据链路冗余容错有相当难度和较高成本，因此，目前还未见从存储磁盘——服务器——核心交换机——安全网关全路径冗余集群的报道。不同的应用视具体需求对不同部分节点冗余容错（以核心交换机或带存储的服务器为系统做集群较多），这是符合 HA 系统设计目标和宗旨的。

为了避免由于每台服务器安装 4 块以上网卡造成的系统结构复杂和过多接触点而导致的系统故障率增加，避免图书馆数字资源多种数据库管理系统及其上层应用软件同时集群实现的难度、各家软件之间的协调兼容问题和集群经费问题，避免集群管理的监测模块形成软件自身 SPOF 而引入新失效源[7]，并且鉴于目前中档以上服务器主机硬件的平均无故障时间 MTBF 远远大于软件，鉴于图书馆资源数据安全可用、访问通信可达是系统建设的基本目标需求，因此舍弃过于复杂的服务器集群，以确保存储、核心交换和安全网关的冗余容错。

由于服务器是系统的关键设备，还是尽量采取零部件级别和系统数据级别冗余容错。数据库应用系统的数据库与 Web 分开后，应用数据存储的可靠性由数据存储区域网 SAN 得到保障，而 Web 服务器和数据库服务器的软件系统，则通过冗余服务器本地硬盘，以 RAID1 方式做系统软件镜象，提高服务器系统可用性。

3.4 关键数据链路的冗余容错

图书馆 Web 应用系统数字资源访问路径由存储——服务器——核心交换——安全网关等节点设备和相关数据链路组成，这些关键节点设备均按照 Failover 冗余热备（Standby）配置，但备份设备只要收到心跳通告就不会变为活动（Active）去屏蔽切换故障。所以，节点设备冗余热备并不能保证相关数据链路的冗余热备，还需解决三个问题。若冗余节点 A（A1、A2）与单节点 B 组成数据链路（图 3－1），首先，节点 B 必须提供分别与节点 A1、A2 连接的两个物理接口；其次，对物理链路 BA1、BA2 的监测，并能屏蔽故障物理链路、切换或恢复物理链路（目前主要采用可变 MAC 地址或 ARP 更新技术）[7]；若冗余节点 A（A1、A2）与冗余节点 B（B1、B2）组成数据链路（图 3－2），还要同时实现冗余节点 A（A1、A2）与节点 B1、冗余节点 A（A1、A2）与节点 B2、节点 A1 与冗余节点 B（B1、B2）、节点 A2 与冗余节点 B（B1、B2）的数据链路冗余容错，这就是第三个技术问题。

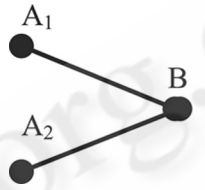

图 3－1　冗余－单节点

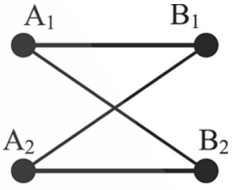

图 3－2　冗余－冗余节点

3.4.1 磁盘——服务器数据链路的冗余容错

数据存储区域网 SAN 本身的构造组成，已经形成从后端磁盘，经过控制器光纤通道、FC 交换机到服务器的物理链路冗余容错。从磁盘到控制器，磁盘阵列后端数据链路的冗余是通过阵列系统 Fabric 交换或 FC－AL（Arbitrated Loop）仲裁环实现负载均衡链路互备的[4]；数据库服务器通过冗余的两个 HBA 卡，经冗余 FC 交换机到阵列冗余控制器前端接口的前端数据链路，则是通过阵列配套的多路径负载均衡高可用软件（例如 HP StorageWorks Secure Path）的寻径技术[9]实现；驻留在服务器主机的软件，监测服务器与存储之间的数据链路，一旦发现路径故障，屏蔽故障的 HBA 卡，切换数据访问通道到备用路径，当故障修复软件自动恢复原来访问路径；同时软件通过动态负载均衡提高存储访问性能，并实现以 LUN 为单位的安全控制。

3.4.2 服务器——核心交换机数据链路的冗余容错

为避免接入过多网卡造成的故障和链路监测切换的虚拟网卡功能问题，在服务器与核心交换机之间引入二层交换机接入汇聚，既提供了物理接口、解决了链路监测切换问题，又符合网络层次化设计原则，有益于数据安全和系统可用性。此外，选用支持端口汇聚的交换机，不仅实现交换机上行级连数据链路负载均衡互为热备，而且提高了链路带宽。

3.4.3 核心交换机——安全网关数据链路的冗余容错

核心交换机——安全网关两节点设备均为冗余容

错配置,但其组成的数据链路实现冗余容错是个难点。相对于 Failover 冗余的安全网关,每台核心交换机提供接口不成问题,能够对链路监测、切换的 OSPF 动态路由实现每台核心交换机——冗余安全网关的数据链路 A/P 模式冗余容错。相对于 Failover 冗余的核心交换机,在采购设备时配置足够的端口,每台安全网关提供两个物理接口也不成问题;由于核心交换机和安全网关都是定型产品,不可能通过 MAC 地址、ARP 协议实现链路监测、切换,幸运的是有些安全网关提供二层端口监测功能:当端口无流量时,系统将流量从另一个与网络连接端口转发出去,并且可以设置这些端口优先级,这相当于安全网关端口的主从 A/P 热备。在端口监测功能支持下,安全网关与 VRRP 虚拟路由器组成每台安全网关——冗余核心交换机数据链路的 A/P 模式冗余容错。

这样,就基本实现存储——服务器——核心交换——安全网关各数据链路组成的数字资源访问路径全程的冗余容错,以存储集群、核心交换与安全网关集群的高可用性保障了图书馆 Web 应用系统服务业务的连续不间断运作。

4 结束语

图 2 是我院 2.7 万平米图文信息中心网络系统的拓扑图。冗余配置的 HP EVA 8000 及其 FC 交换机组成通道 4Gbps 的 SAN;两台 H3C S7500 构成资源内网交换核心、H3C S5600 作为服务器汇聚接入、H3C E352/E328 作为楼层接入;两台启明星辰的天青汉马 3000V – T 作为多功能安全网关隔离内、外网。系统集成后,从 UTM 到核心交换机、到服务器汇聚接入交换机,以及从服务器到存储,每个关键节点之间的 4 条线路,依次人为拔断、接插网络跳线,测试每条数据链路的切换和恢复,结果是在外网用户访问任何数据资源时,系统都能完全透明地在 2 秒钟内切换或恢复。进一步工作,研究如何对重要数据库应用的服务器节点解决 SPOF 以真正实现系统资源访问全路径的集群。

参考文献

1 梅启智,廖炯生,孙惠中. 系统可靠性工程基础. 北京:科学出版社,1987.

2 余胜生,周欣,周敬利,张俊. iSCSI 多级存储系统中的高可用性研究与实现. 计算机工程,2005,(3):207 – 209.

3 沈昌祥. 信息安全工程导论. 北京:电子工业出版社,2003.

4 蔡皖东. 基于 SAN 的高可用性网络存储解决方案. 沈阳:小型微型计算机系统,2001,(3):284 – 287.

5 朱立谷,赵青梅. SAN 的安全技术研究. 计算机应用研究,2003,(5):66 – 69.

6 郝永芳,舒云,王德军. Web 应用环境中的系统可用性设计. 计算机应用研究,2003,(6):123 – 126.

7 刘石丹,林晓东. 高可用性系统结构的研究与实现. 计算机工程与应用,1998,(3):20 – 40.

8 韩德志,耿红琴,李怀阳. 高可用性存储网络技术探析. 计算机应用研究,2004,(8):22 – 26.

9 Marc Farley. SAN 存储区域网络(第二版). 北京:机械工业出版社,2002.

10 H3C. S7500 系列以太网交换机操作手册. 杭州:华为三康技术有限公司,2007.