



“双活”数据中心的设计实现

樊昊

(中国电信股份有限公司泰州分公司, 江苏 泰州 225300)

摘要:随着传统数据中心向云数据中心的转变,政府、金融行业等的信息系统容灾备份成为确保业务连续性和数据可用性的重要手段。双活或多活的数据中心可以提升系统的可用性、提高设备利用率,成为未来数据中心建设的发展趋势。双活数据中心的实现涉及网络、虚拟化、存储等技术,从这些层面介绍了数据中心实现双活的相关技术,并通过在政务数据中心的实践,验证了这一技术的有效性。

关键词:数据中心;双活;虚拟化;云计算

中图分类号:TP311

文献标识码:A

doi: 10.11959/j.issn.1000-0801.2016029

An implementation of the “active-active” data center

FAN Hao

Taizhou Branch of China Telecom Co., Ltd., Taizhou 225300, China

Abstract: With the transition of the traditional data center to cloud data center, government and financial industry's information system disaster recovery and backup becomes an important means to ensure services continuity and data availability. Active-active data center can enhance system availability and improve the utilization of equipment, becoming the future trend of data center construction. Active-active data center involves networking, virtualization, storage technology. These related technologies to achieve active-active in the data center were described, and the effectiveness of the technology was verified by the implementation of the governmental data center.

Key words: data center, active-active, virtualization, cloud computing

1 引言

近年来,随着云计算和大数据的发展,数据中心成为政府、企业等处理其数据、信息的系统中心。信息系统作为政府、企业从事多项业务的基石,其数据完整性和可用性是业务持续运行的基础。尤其是涉及民生的政务系统、金融系统等,绝对的零数据丢失和数据完整性要求也是其最典型特征,依赖于常规的备份系统已经难以满足零数据丢失保障的要求。在高数据完整性的基础之上,一些业务具有非常高的业务连续性需求,基本保证 7×24 h 不间断运行。

针对大型数据中心所面临单点运行风险,需要设计一套支持零数据丢失和业务连续性保障的业务连续性计划来加以支撑。主、备数据中心间的存储复制、数据库复制、持续数据保护和多样化恢复等手段都应该成为数据中心业务连续性计划的必要组成部分来共同完成业务连续性计划目标。

灾难突发时如何保障政府、企业核心业务 7×24 h 的业务连续性,成为业界关注的问题。如何确保数据中心在重大灾难打击之下继续生存,提高业务连续性,是数据中心建设和持续发展的基础性工作。出于业务容灾备份的考

虑,一般都会建设2个及以上的数据中心;主数据中心用于承担用户的业务,备份数据中心用于备份主数据中心的数据、配置、业务等。主备数据中心之间主要包括冷备份、热备份以及双活3种备份方式。

冷备份方式:主数据中心承担业务,备用数据中心不对主数据中心进行实时备份,而是采用周期性或不定期的方式进行备份。一旦主数据中心发生阻断,业务随之中断,数据中心业务恢复的时间难以保证。由此可见,冷备份技术并不是一种真正意义上的灾备技术,无法适应数据中心高要求的发展。

热备份方式:主数据中心承担和响应用户的业务,各数据中心对主数据中心进行实时的备份。当主数据中心发生阻断时,业务从主数据中心切换至备数据中心,各数据中心自动接管主数据中心的业务,从而保证业务不会中断。热备份方式可以自动感知数据中心故障并且保证在规定的RTO(recover time objective,即灾难发生后,信息系统从停顿到恢复正常的时间要求)时间内,应用业务实现自动切换。

双活方式:由于热备份方式在通常情况下,各数据中心并不承担业务,这对数据中心的建设投资带来巨大的压力,且势必造成很大的资源浪费。采用双活的方式,多个数据中心不再有主备之分,同时承担业务并在线运行。此时,数据中心之间互为备份,并且进行实时备份。如果一个数据中心发生阻断,另外一个数据中心仍在正常运行,并将承担所有业务。对用户来说是不可感知的,业务几乎不受影响。双活

技术可采用多种故障检测的方式:ICMP monitor、TCP monitor、HTTP monitor、FTP monitor等。还可以实时检测服务器的运行状态、服务器负载均衡的情况,即使在没有故障的时候也可以根据应用业务量在多个数据中心之间调整。这样就充分利用了资源,从而避免一个数据中心常年处于闲置状态而造成浪费。且通过细致的资源整合,双活数据中心可以提供双倍的服务能力。

2 基础平台设计方案

2.1 设计目标

图1是双活数据中心整体方案的拓扑。

为了实现双活数据中心的设计,需考虑以下几个方面。

容灾半径:是指数据中心之间的直线距离,是衡量容灾方案所能承受的灾难影响范围的一个指标。不同灾难的影响范围是不同的,而距离也会影响到容灾技术的选择,如图2所示。

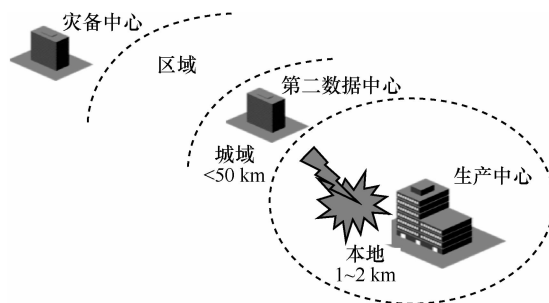


图2 数据中心容灾半径的选择

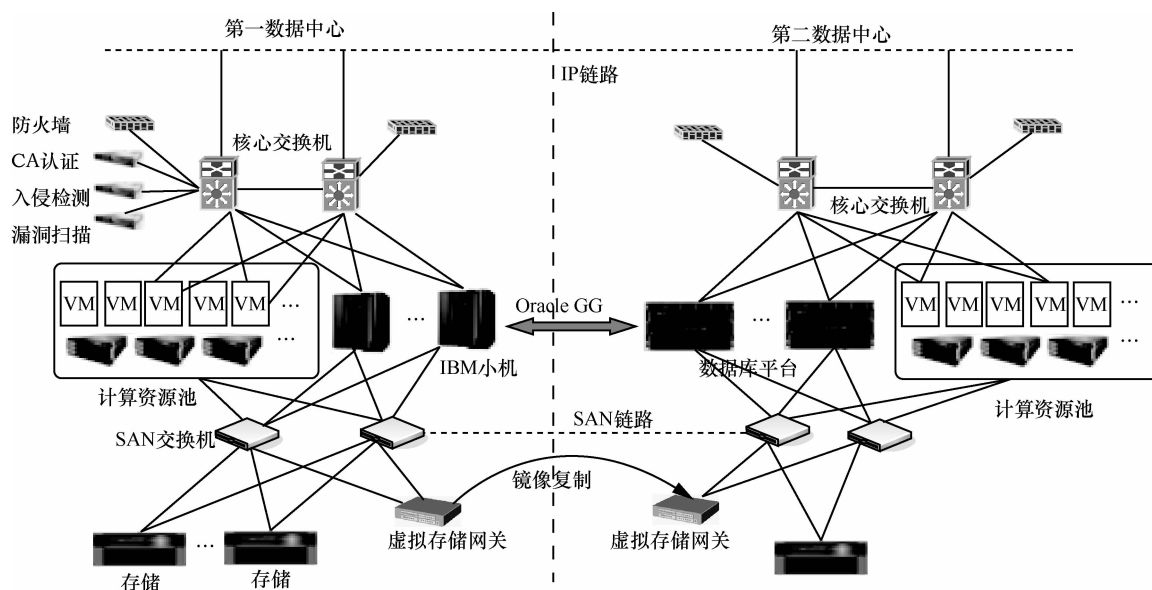


图1 双活数据中心拓扑



网络层: 将实现双活的数据中心之间的网络高速互联, 形成统一网络, 保障数据复制时效性, 同时保障数据库业务、中间件业务等系统的访问顺利切换。采用网络设备虚拟化、大二层网络等技术, 保障网络系统的冗余。双中心采用全局负载均衡, 可以同时提供业务访问能力。同时配置安全设备, 保障网络安全性。

数据存储层: 采用虚拟化存储网关对异构存储平台进行存储虚拟化整合, 建设统一的虚拟化存储平台, 实现两中心的数据底层存储复制。

数据库层: 采用高端的 x86 服务器, 利用主机虚拟化技术, 建设虚拟化服务器集群, 打造高可靠的核心业务系统的数据库。两中心数据库可采用 GoldenGate 等数据库复制技术, 提供业务级数据安全复制, 复制级别达到秒级、亚秒级。建议采用数据库读写分离技术进行应用部署, 提高业务保障级别。

应用层: 采用服务器虚拟化技术, 建设虚拟化应用服务器集群, 部署中间件和业务系统, 提供应用业务支撑。采用负载均衡设备或者应用中间件集群进行负载均衡设计。

2.2 网络资源池

数据中心网络核心设备作为整个网络的核心节点, 要求设备可以长时间不间断运行, 考虑足够的冗余性和安全性。且单板可以热插拔, 便于对设备进行维护时保证业务不间断地运行。同时该设备还能够提供高密度的接入, 便于接入交换机、汇聚交换机的接入。两个数据中心分别采用 2 台高端的数据中心交换机作为核心交换机, 双机通过万兆以太网链路进行互联, 并采用网络虚拟化技术对同一层面的设备进行横向整合, 将两台设备虚拟为一台设备, 统一转发、统一管理, 并实现跨设备的链路捆绑。该方式不会引入环路, 无需部署 STP 和 VRRP 等协议, 极大地简化网络协议的部署, 缩短设备和链路收敛时间(毫秒级)。链路采用负载分担方式工作, 资源利用率得到极大的提升。

网络双活是数据中心双活的基础。双活数据中心需要将同一个网络扩展到多个数据中心, 在不同数据中心之间可以通过大二层网络连接, 实现服务器和应用的虚拟化数据中心互联, 如图 3 所示。目前, 常见的大二层网络技术有 IRF、TRILL、SPB、EVI 等。另外, VMware、微软等厂商提出了包括 VxLAN、NvGRE 在内的 overlay 技术解决方案, 在虚拟化层的 vSwitch 中将二层数据封装在 UDP、GRE 报文中, 在物理网络拓扑上构建一层虚拟化网络层, 从而摆脱对网络设备层的二层、三层限制。

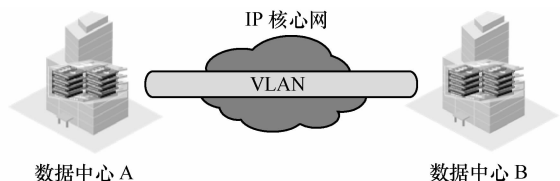


图 3 数据中心通过二层互联

2.3 存储资源池

存储虚拟化技术是将底层存储设备进行抽象化统一管理, 向服务器层屏蔽存储设备硬件的特殊性, 而只保留其统一的逻辑特性, 从而实现了存储系统的集中、统一、方便的管理。

与传统存储相比, 虚拟化存储的优点主要体现在: 磁盘利用率高, 传统存储技术的磁盘利用率一般只有 30%~70%, 而采用虚拟化技术后的磁盘利用率高达 70%~90%; 存储灵活, 可以适应不同厂商、不同类别的异构存储平台, 为存储资源管理提供了更好的灵活性; 管理方便, 提供了一个大容量存储系统集中管理的手段, 避免了由于存储设备扩充所带来的管理方面的麻烦; 性能更好, 虚拟化存储系统可以很好地进行负载均衡, 把每一次数据访问所需的带宽合理地分配到各个存储模块上, 提高了系统的整体访问带宽。

存储资源池采用 SAN 架构技术构建, 按照功能分为虚拟化平台存储资源池、数据库平台存储资源池。通过部署存储虚拟化网关, 对不同厂商的存储设备进行整合, 将数据中心存储虚拟化, 实现异构存储资源池化。存储虚拟化网关部署充分考虑到数据中心双活需求, 业务系统数据在几个数据中心各配置一套, 并通过光纤直连、DWDM 等设备直连或 WAN 等方式实现连接。采用分布式存储方式, 建立统一的存储虚拟化资源池, 实现存储双活, 保障系统数据不丢失, 如图 4 所示。

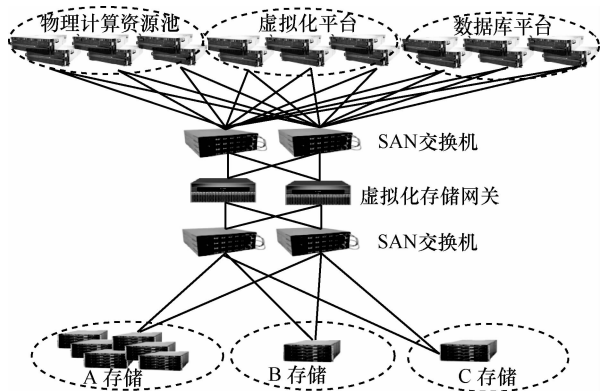


图 4 存储虚拟化网关部署拓扑

2.4 计算资源池

通过服务器虚拟化技术将服务器物理资源抽象成逻辑资源,让一台服务器变成几台甚至上百台相互隔离的虚拟服务器,不再受限于物理上的界限,而是让 CPU、内存、磁盘、I/O 等硬件变成可以动态管理的“资源池”,从而提高资源的利用率,简化系统管理,实现服务器资源池化的整合,提供了更好的可迁移性,让计算资源对业务的变化更具适应力,如图 5 所示。每个用户都可以在他们的虚拟机上运行程序、存储数据,甚至虚拟机崩溃也不会影响系统本身和其他的系统用户。所以,虚拟机模型不仅允许资源共享,而且实现了系统资源的保护。目前,可选的虚拟化主要产品包括 VMware vSphere、Citrix XenServer、Oracle VM Server for x86 等。

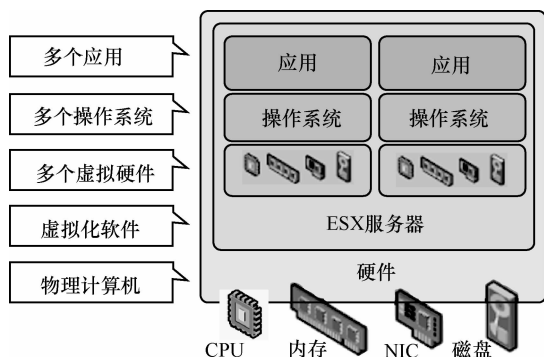


图 5 虚拟硬件技术结构

不同数据中心的虚拟化主机按业务分组组成负载模式,共同承担保证业务访问,实现两中心虚拟机迁移及 HA 切换等。此外,计算资源池还要预留一部分资源,供计划内维护业务系统时的资源迁移以及非计划内意外宕机等情况,建议每个虚拟化集群预留 30% 的资源池容量。

3 业务容灾方案

3.1 业务容灾目标

应用级容灾从流程上实现了全业务的连续性需求。在容灾切换时,除了切换核心的数据库数据外,还包含 IP 地址、中间件服务、用户级业务的切换。

在进行容灾设计时,应针对数据中心业务承接情况给出容灾的合理化建议,分类实现容灾备份系统。以某数据中心为例:对于核心数据库业务,提供实时数据保护,实现实时应用级容灾,数据库读写分离,分别部署于两中心;对于非核心业务,实现应用级容灾,数据近似零丢失,应用业务可在两中心切换,切换时间取决于业务大小,基本在 30 min 内实现切换。

3.2 数据中心入口选择

数据中心可选择多运营商接入,通过部署链路负载均衡设备,用来实现多广域网线路选路和冗余备份,使用户可以通过最快线路访问数据中心内部业务系统,加快了数据中心访问速度。同时,当任何一条线路出现故障,用户依然可以通过另一条广域网线路访问数据中心服务器,提高了数据中心的可靠性。

通过全局负载均衡设备的智能 DNS 功能,可以实现两数据中心的灾备功能。当第一生产中心的所有服务器出现故障或受到攻击而不能提供服务时,全局负载均衡设备会引导用户(自动或人工)去第二生产中心访问业务。

3.3 核心数据库业务容灾

核心数据库业务中业务复杂,访问量大,包含结构化和非结构化数据。数据重要程度高,要求当生产中心发生严重灾难时,能实时接管并提供应用服务,保障业务连续性。

(1) 非结构化数据

非结构化数据一般为大量的不同格式的图片、文档、音视频文件等,存放于光纤磁盘阵列上。对于非结构化数据,只要数据能完整的复制到灾备中心,通过相应的文件系统格式,就可以进行访问。因此,将非结构化数据通过虚拟化存储网关进行存储整合后,实现统一存储虚拟化,两中心数据完全一致,为上层应用提供透明访问。

(2) 结构化数据

Oracle RAC 集群或使用专业的数据库复制软件进行异构的数据库复制,是常用的业务连续性解决方案,可实现数据库层的双活。

Oracle RAC 可以实现多节点多活,并可在节点之间负载均衡,从而保证数据的唯一性;而支持异构的数据库复制技术,可以在异构的 IT 基础结构之间实现大量数据的数据捕捉、转换和投递,打破了原来硬件及操作系统等方面的限制,避免了引入过多种类的软硬件产品,降低了运营维护的复杂度和投入,有利于灾备系统的恢复和切换。

以数据库复制软件 GoldenGate 为例,由于数据库文件在不同的业务平台下,数据文件和参数会有一些区别,只有实现两端数据一致性保障,才能确保核心数据的安全。GoldenGate 可以从数据库的日志中解析数据的变化(数据量只有日志的 1/4 左右),然后将数据转化为自己的格式,直接通过 TCP/IP 网络传输,无需依赖于数据库自身的传递方式,而且可以通过很高的压缩比率对数据进行压缩,



大大降低带宽需求。在目标端,可以通过交易重组、分批加载等技术手段,大大加快数据投递的速度和效率,降低目标系统的资源占用,可以在秒级实现大量数据的复制。

此外,采用数据库读写分离技术,将读写数据流分别分担于两中心,进行业务负载均衡。当数据写操作发生时,不影响数据读操作,且可以用分离的数据库接管主业务系统。

3.4 非核心业务容灾

非核心业务完全承载于数据中心的云平台,全部进行业务虚拟化。同时采用基于存储网关的统一存储虚拟化进行设计。数据中心之间通过大二层网络互联,使虚拟机可以在数据中心之间进行快速迁移和 HA 切换。

4 政务数据中心的“双活”实践

随着政府信息化建设的不断推进,尤其是国家提出建设智慧城市后,智慧交通、智慧医疗、智慧教育等平台纷纷上马。各地政府为解决过去电子政务基础资源分散、资源难以共享、重复投资建设、资源浪费等问题,陆续开始了政务云数据中心的建设,以实现政务基础资源的整合和共享。整合后的政务数据中心具有“大集中”、“虚拟化”、“云计算”等特征,通过灾备能够提供业务连续性,保证各类政务业务的持续运行。

4.1 方案概述

某市政府在对其数据中心进行资源整合的同时,在两个数据中心引入了灾备能力。灾备选用“同城双中心”方案,根据数据中心的业务情况,将应用的备份分为主备应

用以及双活应用。拓扑示意如图 6 所示。

两个数据中心不是同期建设,因此规划和部署了不同网段的 IP 地址:第一数据中心规划了 10.0.0.0/16 网段的 IP 地址,第二数据中心规划了 20.0.0.0/16 网段的 IP 地址。

在两个数据中心增加 VxLAN 网关设备,通过 VxLAN 技术在两个数据中心实现二层互通。双中心完成大二层互通后,形成跨越两个数据中心的统一资源池,虚拟机可以在双中心间进行迁移。此外,通过 IP 地址和 VLAN 的规划,将资源池划分为两级:一级资源池的应用采用双活方式进行备份,二级资源池的应用采用主备方式进行备份。

进行整合的初期,为了保证数据中心访问流量流向在改造实施后仍与改造前相同,暂时不在双中心对业务进行负载均衡。在双中心的资源池网关设备上针对不同的 IP 地址段启用 VRRP。通过调整 VRRP 优先级以及发布路由的优先级,使外网与 10.0.0.0/16 网段的交互流量均经过第一数据中心,外网与 20.0.0.0/16 网段的交互流量均经过第二数据中心。即正常情况下,两个数据中心的二层互通链路不承载外网访问业务流量。

4.2 资源池故障测试

当第一数据中心资源池中的虚拟机出现故障时,通过高可用性、容错等备份技术,虚拟机被自动移动至第二数据中心,其 IP 地址不发生变化。用户访问该主机时,仍从第一数据中心的外网入口进入,并经由双中心之间的二层互联,访问到位于第二数据中心的备用主机。

在两个数据中心现有的资源池搭建测试环境,具体硬件配置见表 1。

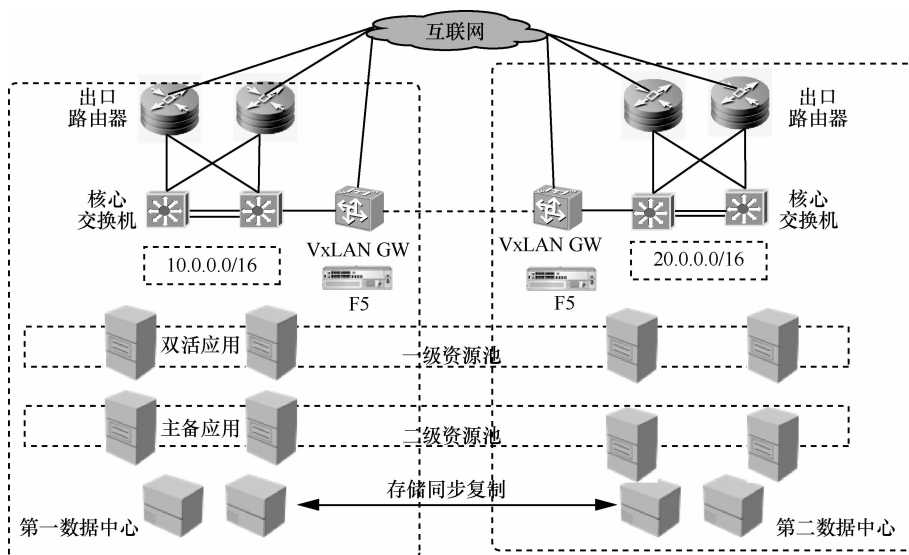


图 6 政务数据中心整合示意

表 1 资源池配置

资源名称	型号	硬件配置
计算资源池	浪潮 NF8560M2	4×Intel Xeon E7-4820 处理器 128 GB 内存 2×300 GB 15K 硬盘/独立 RAID 卡 (512 MB) 标配 2 个吉比特以太网口, 2 块单端口万兆光纤网卡, 2 个 8 GB HBA 卡 1+1 冗余电源
存储资源池	浪潮 AS1000G6	冗余控制器 48 GB 缓存, 支持缓存断电保护 8 个 8 GB FC 36 块 600 GB 15 K SAS 硬盘, 共 21.6 TB 冗余电源、风扇等

通过部署 VMware HA 进行资源池故障恢复测试, 测试环境虚拟机配置见表 2。

表 2 测试虚拟机配置

名称	IP 地址	操作系统	应用
HA1	10.100.15.1	VMware EXR5.1	ESX 集群节点 1
HA2	10.100.15.2	VMware EXR5.1	ESX 集群节点 2
Windows	10.100.15.3	Windows 2008 R2	VMware vCenter
Wintest	10.100.15.101	Windows 2003 Server	HA 测试

在两个数据中心的统一资源池上创建两个测试集群节点, 并启用 HA。在集群上创建一台配置为 2.4 GHz 处理器、2 GB 内存的 Windows 2003 Server 企业版虚拟机, 用于故障切换测试。测试时关停一个集群节点, 通过 ping 工具测试验证主机的可用性, 业务整体恢复时间约为 1 min 45 s。

通常, VMware HA 的业务中断时间主要由 HA 心跳检测时间、虚拟机系统启动时间以及应用启动时间组成, 这个时间一般在 3 min 以内。因此, 测试结果是正常的。

4.3 网络出口故障测试

当第一数据中心网络出口设备发生灾害(停电、出口阻断等)时, 10.0.0.0/16 网段在第一数据中心的网关失效, 路由也随即撤销。第二数据中心检测到第一数据中心网关不可用后, 将其网关切换成 VRRP 主用状态, 且路由同时生效。此时, 访问 10.0.0.0/16 网段的流量从第二数据中心网关进入, 并经由双中心之间的二层互联, 访问位于第一数据中心的资源池。

由于资源池并未受到故障影响, 因此, 该场景下的业务恢复时间主要受制于 VRRP 设置的发送通告报文定时器时间。测试中, 将第一数据中心网关子接口关闭来模拟故障, 使用 ping 工具测试验证主机的可用性, 业务恢复时间可达到秒级。通过部署快速检测协议, 加快 VRRP 触发

故障倒换, 可以进一步缩短业务恢复时间。

4.4 Web 应用故障测试

用户通过域名访问位于第一数据中心 Web 应用时, 部署在两个数据中心的任意一个 F5 解析出主用地址为 10.X.100.100, 备用地址为 20.X.100.100。当第一数据中心发生单点故障导致主用 Web 服务器不可用时, 主用 IP 地址不可用, 并由 F5 调度到 20.X.100.100 提供服务。测试中, 通过关停 Web 测试服务器虚拟机模拟故障, 由于备用主机已经处于“存活”状态, 因此用户几乎不会感知到 Web 应用故障。

5 结束语

数据中心是支撑云计算、大数据发展的基础, 本文介绍了一种“双活”数据中心的实现方案。通过双活技术, 可以搭建一个高可靠性的信息化平台。数据中心中信息系统双活的实现方式还有很多, 结合数据中心各类信息系统的具体需求来进行方案设计尤为重要。此外, 真正的实现数据中心各类信息系统的连续运行, 一套完善的运维管理制度和流程、维护管理人才的培养和储备也是必不可少的。

参考文献:

[1] 史晨阳, 寿弘宇. 面向未来的同城双活中心网络架构建设[J]. 金融电子化, 2014(3): 52-54.
SHI C Y, SHOU H Y. The construction of future oriented network architecture for double-live data center [J]. Financial Computerizing, 2014(3): 52-54.

[2] 李宏伟, 肖伟. 存储虚拟化技术在双活数据中心中的应用[J]. 邮电设计技术, 2013(9): 9-13.
LI H W, XIAO W. The application of storage virtualization technology to double-live data center[J]. Designing Techniques of Posts and Telecommunications, 2013(9):9-13.

[3] 朱明明, 夏寅贵, 徐小飞. 基于 SDN 的数据中心网络研究[J]. 邮电设计技术, 2014(3): 23-29.
ZHU M M, XIA Y B, XU X F. Research on data center network based on SDN [J]. Designing Techniques of Posts and Telecommunications, 2014(3): 23-29.

[作者简介]



樊昊(1979-), 男, 中国电信股份有限公司泰州分公司网络操作维护中心工程师, 主要研究方向为大型 IP 城域网组织、优化以及数据中心的设计规划。