容灾系统建设

1 引言

随着中国金融业数据集中的起步、发展和完善，很多银行建设了先进的集中式数据中心。数据大集中所体现出来的银行核心竞争力提升的优势，被业界广泛认同并随之引发了国内银行数据中心建设的浪潮，在过去的几年中，四大银行、全国性商业银行、城商、农商、农信联、邮储等都已建设或计划建设自己的集中式数据中心。

数据大集中在带来巨大好处的同时，也带来了风险大集中。如何应对和有效化解数据集中带来的风险？如何保证数据在各种灾难情况下的安全？如何保障业务连续性、维护银行声誉等等都是必须要面对的问题。伴随着数据大集中，容灾备份系统的建设成为一项紧迫的工作。

2 容灾建设相关标准

国际标准 SHARE78 定义了灾难恢复的多个级别，灾难恢复解决方案可根据多个方面制定，包括备份/恢复的范围、灾难恢复计划的状态、在应用中心与备份中心之间的距离、应用中心与备份中心之间是如何相互连接的、数据是怎样在两个中心之间传送的、有多少数据被丢失、怎样保证更新的数据在备份中心被更新、备份中心可以开始备份工作的能力。

为了规范数据数据中心和灾备中心的建设，人民银行在 2002 年制定了《关于加强银行数据集中安全工作的指导意见》，并在 2005 年进一步提出要求：全国各商业银行在 1～2 年内数据灾难备份标准达到 2～3 级，在各银行完成数据集中后的 2 年内数据中心的灾难备份标准必须达到 5～6 级。参考国际上相关组织在灾难恢复上的研究与实践，我国的国家标准 GB20988-2007-T《信息安全技术 信息系统灾难恢复规范》对容灾备份进行了标准化。表 1 为 SHARE78 与国标在各层次上的大致对应关系。

SHARE78		**《信息系统灾难恢复规范》GB/T 20988 - 2007**	
Tier-0	无异地备份数据	第 1 级	基本级。备份介质场外存，安全保管、定期验证
Tier-1	有数据备份，无备用系统 用卡车运送备份数据		
Tier-2	有数据备份，有备用系统 用卡车运送备份数据。	第 2 级	备份场地支持。网络和业务处理系统可在预定时间内调配到备份中心
Tier-3	电子链接，消除运送工具的需要，提高了灾难恢复速度	第 3 级	电子传输和部分设备支持。灾备中心配备部分业务处理和网络设备，具备部分通讯链路。
Tier-4	灾难恢复具有两个中心彼此备份数据，允许备份行动在任何一个方向发生。两个中心之间，彼此的关键数据的拷贝不停地相互传送着。在灾难发生时，需要的关键数据通过网络可迅速恢复，通过网络的切换，关键应用的恢复也可降低到小时级或分钟级。	第 4 级	电子传输和完整设备支持。数据定时批量传送，网络/系统始终就绪。温备中心模式。
Tier-5	保证交易的完整性 为关键应用使用了双重在线存储，在灾难发生时，仅传送中的数据被丢失，恢复时间被降低到分钟级。	第 5 级	实时数据传输及完整设备支持。采用远程复制技术，实现数据实时复制，网络具备自动或集中切换能力，业务处理系统就绪或运行中。
Tier-6/7	无数据丢失，同时保证数据立即自动地被传输到恢复中心。Tier6 被认为是灾难恢复的最高的级别，在本地和远程的所有数据被更新的同时，利用了双重在线存储和完全的网络切换能力。第 7 层实现能够提供一定程度的跨站点动态负载平衡和自动系统故障切换功能。	第 6 级	数据零丢失和远程集群支持。数据实时备份，零丢失，系统/应用远程集群，可自动切换，用户同时接入主备中心

表 1 SHARE78 与国标的对应

在设计容灾系统时，容灾要达到什么样的目标与层次，需要用一些定量的指标来衡量，这就是灾难恢复能力指标，如图 1 所示。

RTO(Recovery Time Object):恢复时间目标，指信息系统从灾难状态恢复到可运行状态所需要的时间，用来衡量容灾系统的业务恢复能力。

RPO(Recovery Point Time):恢复点目标，指业务系统所允许的在灾难过程中的最大数据量丢失，用来衡量容灾系统的数据冗余备份能力。

NRO(Network Recovery Object):网络恢复时间目标，指在灾难发生后网络恢复或切换到灾备中心的时间。通常，网络要先于应用恢复才有意义，但应用恢复后才能提供业务访问。

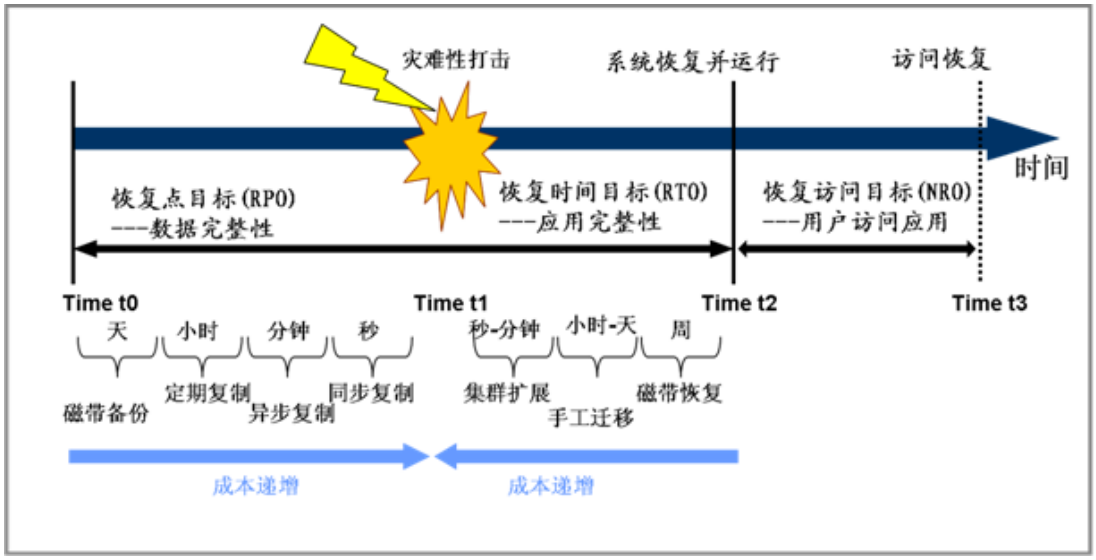

图 1 容灾恢复能力指标

在容灾设计和灾难恢复规划中，主要采用恢复能力指标（RPO 和 RTO）定量的分析灾难恢复目标，由此形成了灾难恢复的不同等级，如表 2 所示

阶段	灾备等级	RTO	RPO
初期建设目标	第 1 级	2 天以上	1 天至 7 天
	第 2 级	24 小时以上	1 天至 7 天
	第 3 级	12 小时以上	数小时至 1 天
中期	第 4 级	数小时至 2 天	数小时至 1 天
最终目标	第 5 级	数分钟至 2 天	0 至 30 分钟
	第 6 级	数分钟	0

表 2 不同等级的容灾能力要求

3 容灾建设模式讨论

3.1 容灾系统的组成

一个典型的容灾系统由灾备中心基础环境设施、数据备份系统、备份处理系统和网络通信系统、灾难恢复计划等组成，如图 2 所示。

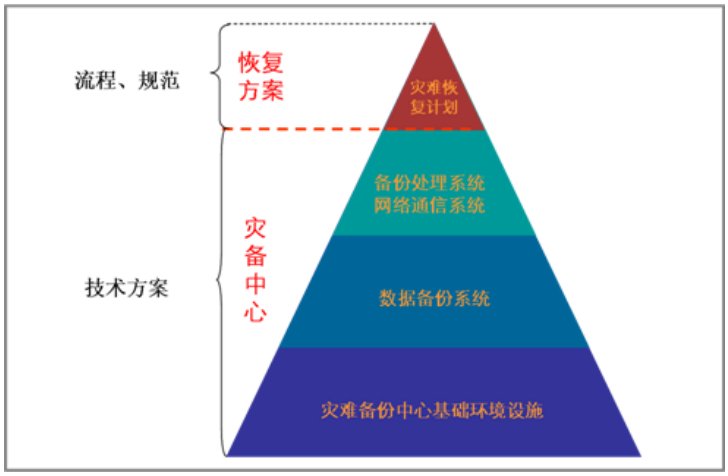

图 2 容灾系统组成

基础设施环境：要求能够保障数据备份系统和备份处理系统的工作，可提供完备的网络通讯设施，全面的供配电系统、综合布线、精密空调系统、消防及自动气体灭火系统、中心闭路监控、漏水检测等。

数据备份系统：是灾难备份系统的最基本要素。如何将数据（包含系统、应用和业务数据）完整、实时地复制到容灾中心，是银行在系统建设需要重点考虑的事项。

备份处理系统：是指在灾难恢复时需配备的主机系统、存储系统、应用软件等。容灾中心的网络通讯系统要求能够提供正常业务运行的数据备份通道和灾难发生时系统切换后的业务数据流转，保证关键备份业务峰值性能要求；备份处理系统所需要达到的处理能力和范围应基于恢复目标及成本效益等因素综合考虑。

灾难恢复计划：为保证灾备系统在故障发生时可按预期及时替代生产系统而制定的管理制度、规范和流程，如：安全管理、运维管理、恢复管理、变更管理、灾难恢复演练流程等。灾难恢复计划需定期维护、测试和演练及审核，以保持其持续可用性。

根据容灾效果，容灾系统分为数据级容灾和应用级容灾。

数据级容灾：异地容灾系统数据是本地关键应用数据的一个副本，当本地系统发生灾难时，系统至少在异地保存有一份可用的关键业务的数据。该数据可以是本地生产数据的完全复制(一般在同城实现)，也可以比本地数据略微落后，但必定是可用的(一般在异地实现)，而差异的数据通常可以通过一些工具(如操作记录、日志等)可以手工补回。

应用级容灾：是在数据级容灾基础上，在异地建立一套与本地生产系统相当的备份环境，包括主机、网络、应用、IP 等资源均有配套，当本地系统发生灾难时，异地系统可以提供完全可用的生产环境。

3.2 容灾模式

在灾备中心，IT 系统主要包括网络、计算、存储几个方面，分别对应着容灾系统的网络通信系统、备份处理系统、数据备份系统。灾备中心作为业务中心的备份，基于是否需要备用处理系统（服务器）以及专业人员支持，可以分为不同的灾备中心模式。

冷备模式（Cold Standby）：备份系统未安装或未配置成与主用系统相同或相似的运行环境，应用系统数据没有及时装入备份系统。

缺点：恢复时间长，一般要数天或者更长时间，数据的完整性与一致性差。

灾备等级：3 级，只适合于商业银行数据大集中初期的要求。

暖备模式（Warm Standby）：具备备份系统安装场地、备份主机、存储设备和通信设备，备份系统已经安装配置成与主用系统相同或相似的系统和网络运行环境，安装了应用系统业务定期备份数据。一旦发生灾难，直接使用定期备份数据，手工逐笔或自动批量追补孤立数据，恢复业务运行。

缺点：恢复时间较长，一般要十几小时至数天，数据完整性与一致性较差。

灾备等级：4 - 5 级 只适合于商业银行数据大集中初期的要求

热备模式（hot Standby）：备份系统处于联机状态，主用系统通过高速通信线路将数据实时传送到备份系统，保持备份系统与生产系统数据的同步。也可定时在备份系统上恢复主用系统的数据。一旦发生灾难，不用追补或只需追补很少的孤立数据，备份系统可快速接替主用系统运行，恢复生产。

优点：恢复时间短，一般几十分钟到数小时，数据完整性与一致性最好，数据丢失可能性最小。

灾备等级：5 - 6 级 当前金融行业主流容灾建设方向。

4 容灾技术与方案考虑

4.1 多中心的容灾方案

多中心容灾方案中，有同城灾备中心、异地灾备中心等多种结构，以及自建自用、共建或租用共享灾备等方式。目前比较主流的两地三中心结构(如图 3 所示)结合了“同城＋异地”的优点，在异地备份中心具有完整的灾难接管能力的情况下，建立同城备份站点，让同城灾备中心只是一个同步数据镜像站点的同时，使同城灾备中心具有应用接管能力。

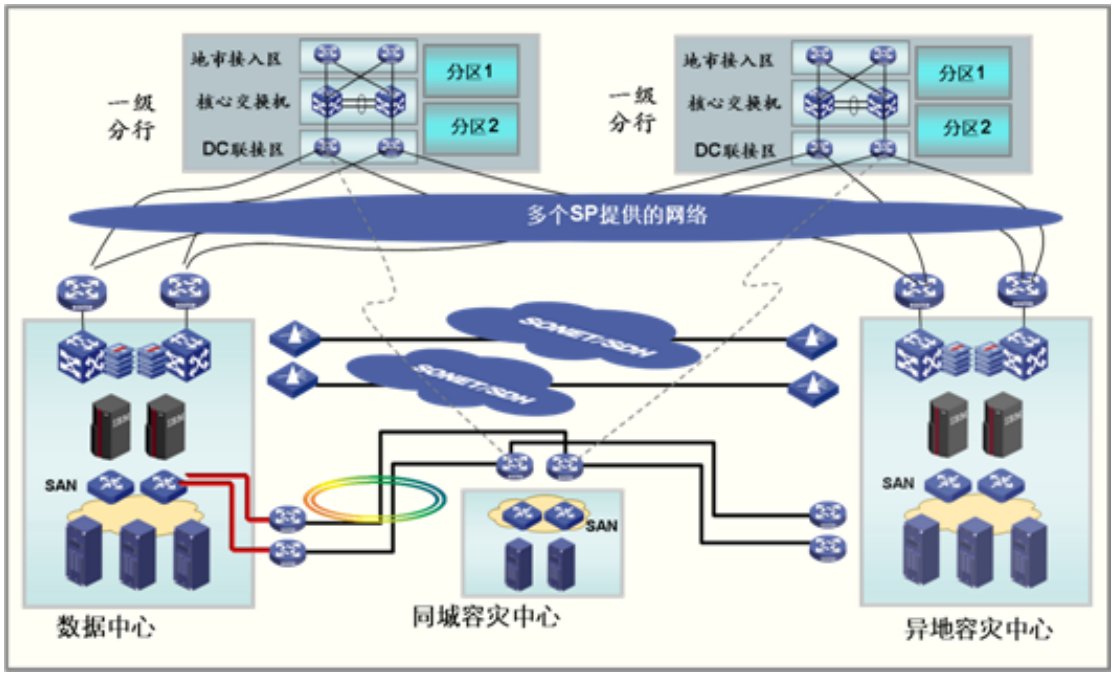

图 3 两地三中心容灾方案

银行灾备中心建设，以应用容灾为核心，以业务连续性为重点，实现安全生产与运营。在容灾能力上，两地三中心是当前最好的容灾模式，可以最大程度的保护数据和业务连续性，应对重大区域性灾难。

在容灾方式上，主数据中心的关键业务数据实施同步到同城灾备中心，即一笔数据写入会在两个数据中心同时写入后返回，保证同城两中心数据的完全一致性；本地数据写入完成后，再由主数据中心或同城灾备中心异步将数据复制到远程的异地灾备中心。

在运营方式上，多个中心之间可以在平时进行业务分担，也可以实现相互完全业务互备，而后端实现数据同步。

4.2 服务器集群

集群系统是一种提供高可用性、改善性能和增强应用软件可管理性的有效途径。随着基于多中心建设成为关键性业务和应用的容灾主流要求，集群技术的容灾应用也日益广泛。

集群可有效地提高系统的可用性，如果一个服务器或应用程序崩溃，集群系统中其它服务器在继续工作的同时，接管崩溃服务器的任务，最大限度地缩短用户服务器和应用程序宕机的时间。集群的另外一个优点是通过增加现有系统的的节点，提高了系统的延展性，使系统因故障中断的可能性降到最低。在这种架构中，多服务器的运行是针对相同的应用程序或数据库的。

集群技术已经发展为本地集群、同城集群、异地集群，可将关键应用系统在广阔的地理范围内进行扩展和备份。集群的运行一般要求服务器在同一局域环境内，因此要求在同一 VLAN 内，如图 4 所示，跨数据中心的同一个集群，需要将相同的 VLAN 从主中心延伸到灾备中心。

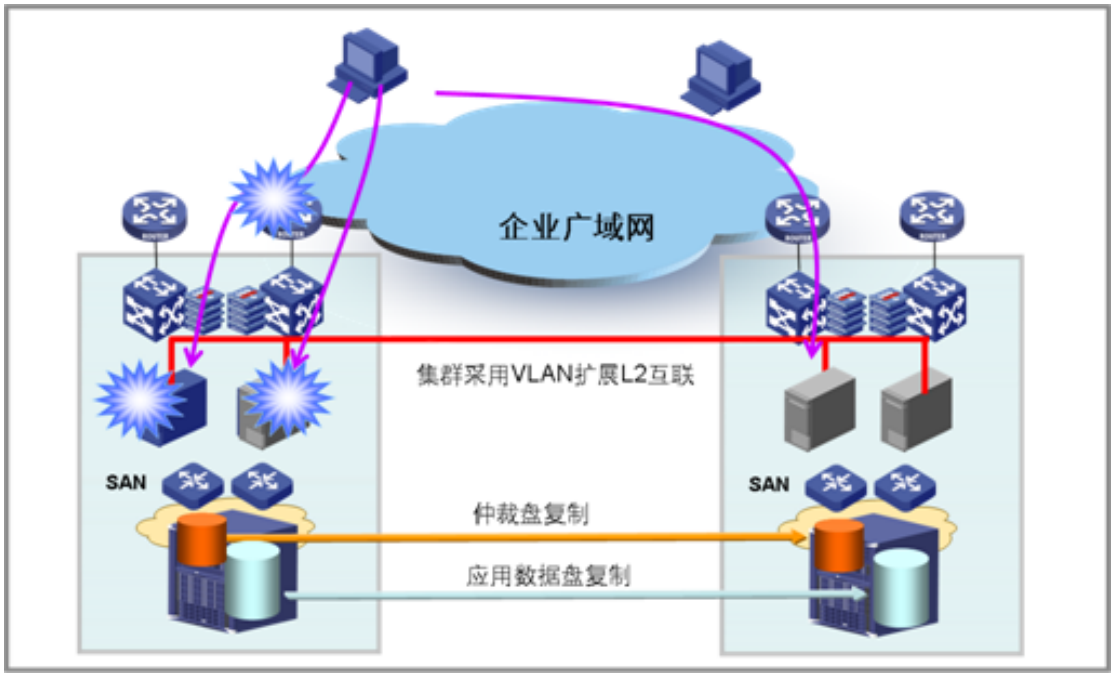

图 4 跨中心的服务器集群

集群容灾过程如图 4 所示，主数据中心和灾备中心均有存储系统，正常情况下主中心 Master 服务器获取对存储的控制(即取得对 Quorum 盘——仲裁盘的访问，Quorum 盘用来保存集群的信息，这些信息用来维护集群的完整性以及使节点保持同步，Quorum 盘在某一时刻只能被一个节点所拥有，并用来决定由哪个节点来拥有集群的所有资源)，Quorum 信息与应用数据会被复制到灾备中心，当主中心 Master 故障，主中心本地备用服务器接管对 Quorum 盘的控制继续提供服务，如果主中心发生灾难，灾备中心集群内的服务器将挂接本地 Quorum 盘并提供服务(如果距离和 IO 允许，容灾设计也可允许灾备中心的服务器获取主数据中心存储的控制)。

4.3 数据的复制

数据复制是容灾中的关键技术，一般分为同步复制和异步复制。

同步复制是把生产数据以完全相同的方式复制到异地，每一个 I/O 操作都需要等待远程复制完成后才释放。实时性强，远端数据与本地数据完全同步，可以达到数据的零丢失。同步复制通常有基于主机逻辑卷和基于磁盘系统 I/O 的两种方式，如图 5 所示。

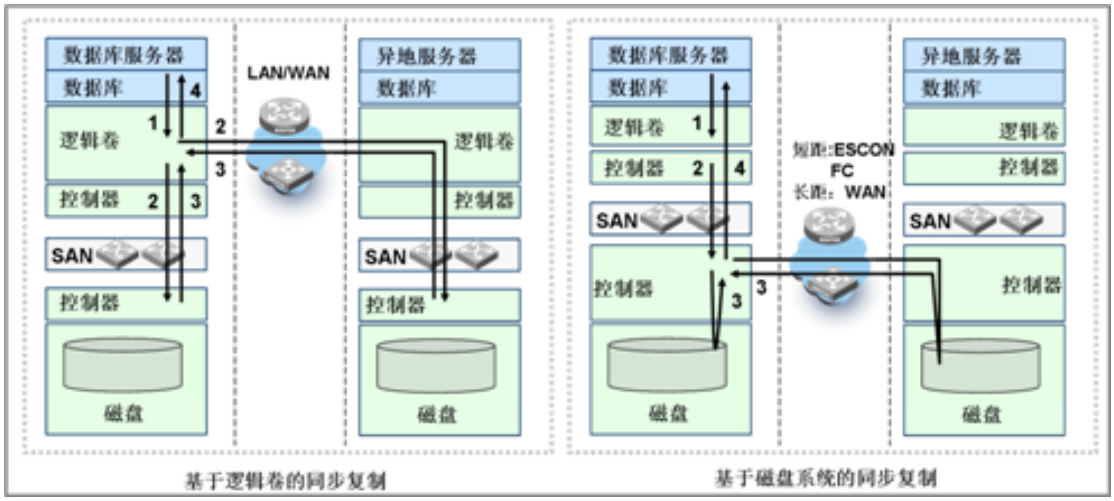

图 5 两种同步复制

基于逻辑卷的复制：当主机发起一个 I/O 请求到达逻辑卷层后，逻辑卷管理层在向本地磁盘系统发出 I/O 请求的同时，通过 TCP/IP 网络向异地系统发出 I/O 请求，本地卷管理层等待本地磁盘和异地系统的 I/O 全部返回后再进行本次 I/O 返回操作。

基于磁盘系统的同步复制：主机发起的 I/O 请求到达磁盘控制器，磁盘控制器一方面写入本地磁盘，同时将操作复制到异地磁盘系统，本地磁盘系统需要等待两个 I/O 操作均返回后才向主机返回本次 I/O。

同步复制可以保证两地数据的完全一致性，但同步容灾过程中本地系统必须等到数据成功写入异地系统，才能进行下一个 I/O 操作，同步容灾数据复制一般只在较短距离或同城范围部署(10km~100km)，超过 100 公里以上一般采用异步复制方式进行容灾。

异步复制中，本地 I/O 操作完成后直接返回，而不需等待异地 I/O 的返回，甚至，异步复制并非针对每个 I/O 进行复制，而是根据数据的增量或时间等方式进行复制。

4.4 容灾网络

容灾网络包含容灾路由规划、主机互联、链路选择等几个方面。

路由规划

路由协议选择，一般有 BGP/OSPF/IS-IS。BGP 具备较强的路由控制与策略功能，适合大型网络互联，适用于银行骨干网；OSPF/IS-IS 路由控制能力一般，适用于数据中心内部、局域网内部、分行内部。如图 6 所示为一种两地三中心容灾模式的规划方式，每个数据中心/灾备中心/分支机构为一个 BGP 自治域，通过 EBGP 控制域间路由，对于链路故障、灾难引起的故障，都需要对 BGP 的路由引入、分发进行规范性设计，以达到故障恢复、灾难恢复后路由收敛的可预见、可控制。

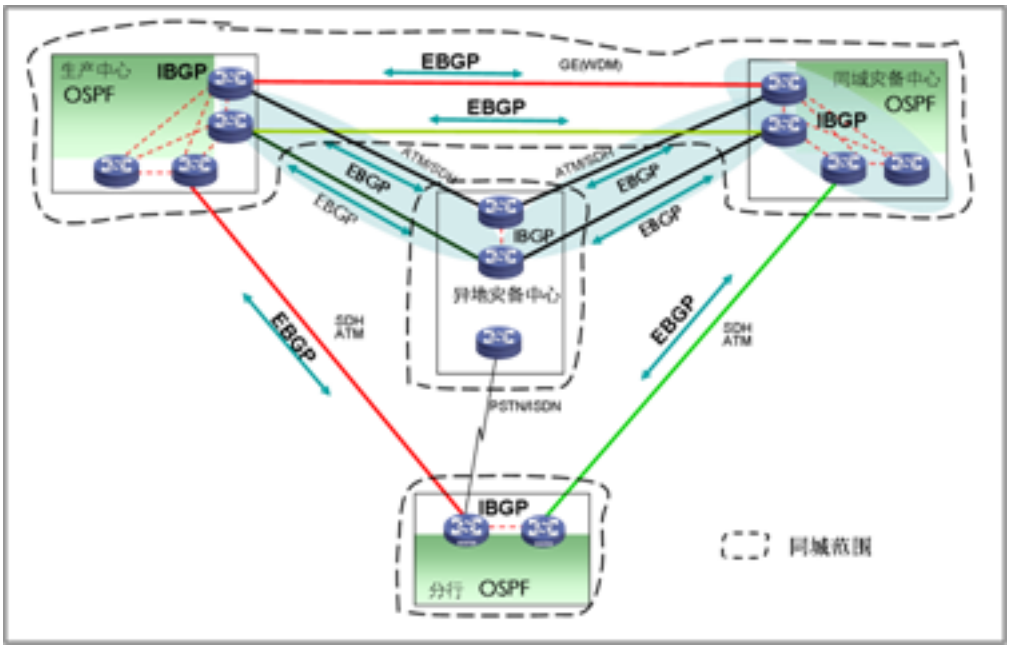

图 6 容灾路由协议基本规划

容灾路由基本策略：

- 当备中心服务器与主中心服务器 IP 地址不在相同网段时，发布两个网段的路由不会影响业务；
- 当两中心服务器采用相同地址或相同网段时，会出现同一子网 IP 地址由不同数据中心或容灾中心对外发布的情况，处理不当则会影响应用系统的访问，此时需要设置主中心路由具有高优先级发布或备份中心路由不对外发布；
- 路由设计使访问能够在分行与主中心链路故障时绕行容灾中心到主中心，防止各分行至主中心链路故障意外情况造成的业务中断(路由不可达)；
- 灾难发生时，确保容灾中心基础网络路由可达，容灾中心网络设备地址必须与主中心网络设备地址不同。

主机互联

多中心之间的主机如果运行集群协议，一般要求在同一个 VLAN 内，因此需要构建一个跨数据中心的集群网络。如图 7 所示，这个网络要求保证主机集群二层可达，同时要有冗余能力，当前可提供的建网技术包括二层 VPN(VLL、VPLS)、传输冗余直连/光纤冗余直连(需要生成树)、虚拟化网络技术。

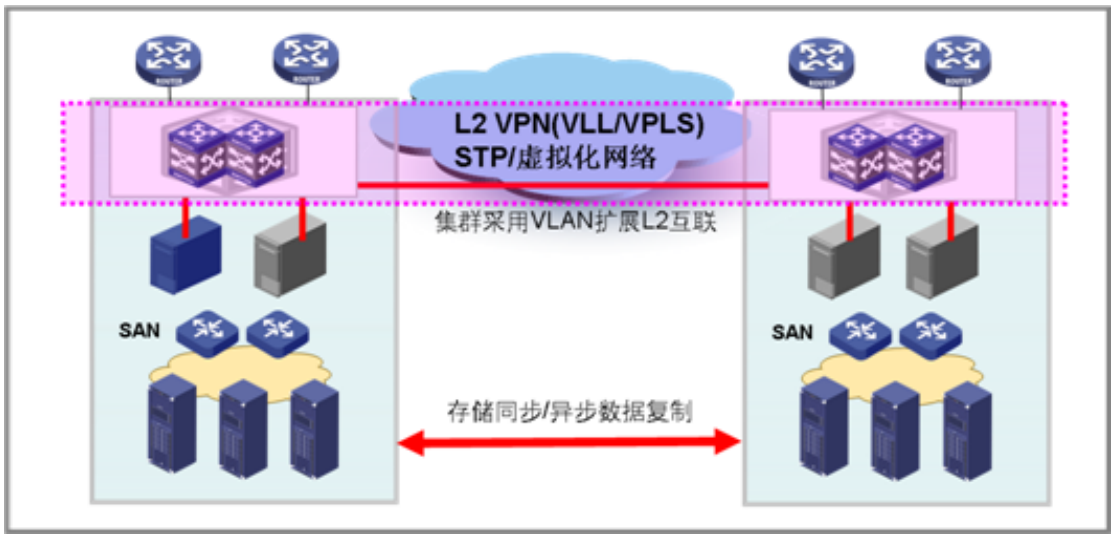

图 7 跨数据中心的主机集群

H3C 虚拟化网络技术 IRF2 可以基于以太网底层介质(L2 VPN/传输/光纤等)构筑多数据中心间的二层无环网络，既提供了大范围主机集群的二层连接性，又可通过冗余捆绑链路提供高可靠性，如图 8 所示。

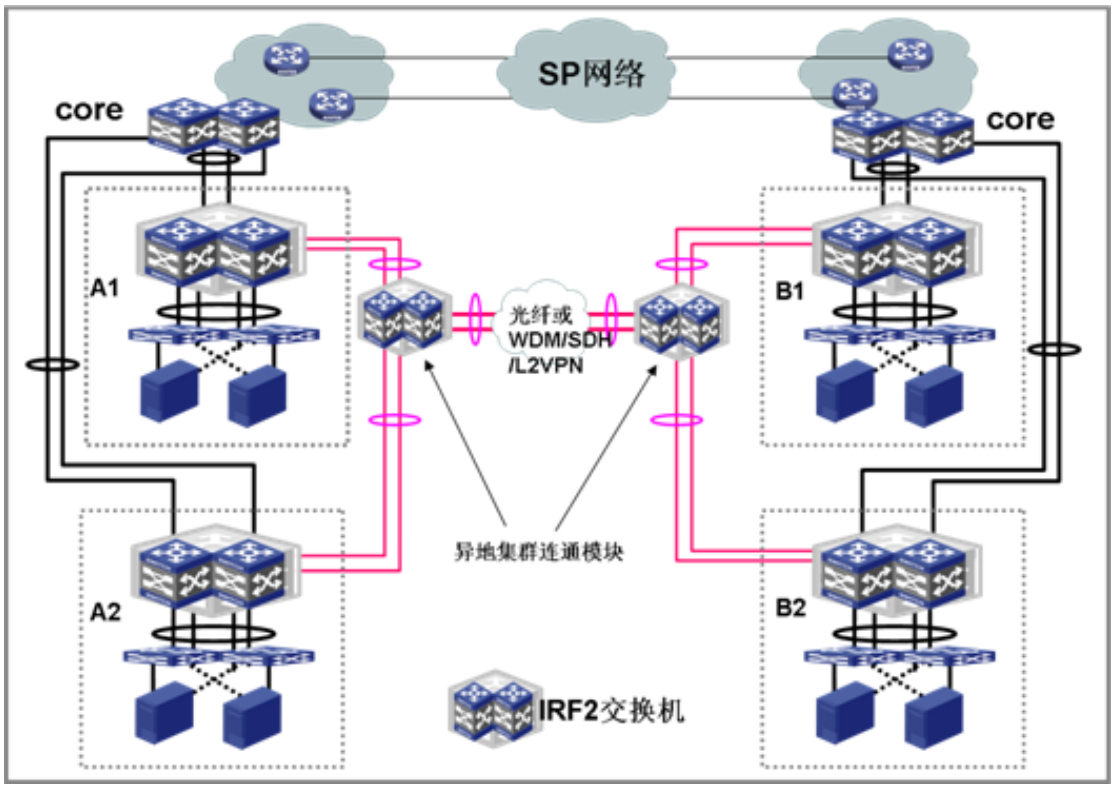

图 8　IRF2.0 支持多中心二层主机集群互联

链路选择

- 分行-主中心、分行-容灾中心：当前国内银行多选择 ATM，技术发展方向为 CPOS-SDH。
- 主中心-同城容灾中心：WDM 支持 100km 同城互联，一般的同城中心距离在 40-60km。WDM 可承载多种接口链路类型 FC/FICON/ESCON/GE/10GE.
- 主中心-异地容灾中心：POS/GE-SDH

5 灾难恢复规划与流程

建设容灾系统必须有明确的灾难恢复规划，有明确定义的灾难恢复流程、演练制度等措施来确保业务的长期稳定运行。容灾建设涉及的基本规划与流程内容建议考虑以下方面：

灾难恢复目标及范围：定义灾难恢复的范围、恢复的方式、RTO、RPO、NRO；确定恢复业务品种范围、需要有限恢复的网点和渠道。

灾难宣告流程：灾难告警，由生产运行负责人向灾备中心预警，主备切换；灾难评估，负责人召集专家团进行评估，确定是否做灾难切换；灾难宣告，负责人向上级、灾备中心、各业务部门宣告灾难切换，启动灾难恢复计划。

灾难恢复团队及人员构成：责任人通常是主管信息科技的行长；技术方案团队、业务方案团队、对外联络团队等。

联络清单：提供灾难发生情况下的紧急联系信息，包括银行信息科技、业务、业务管理、后勤支持保障等各部门，以及供应商、维保公司、客户、政府部门。

灾难切换流程：灾难恢复技术团队按预先制定的规程恢复交易系统后，技术解决方案团队与业务解决方案团队人员针对恢复的业务完整性、数据及时完整性、网点和服务渠道范围进行审核和案例验证。

灾难恢复规划还需包含有计划的和无计划的容灾切换。

有计划的切换：应对紧急情况，该种情况允许主中心的所有应用系统正常关闭，允许网络在主中心完全 down 前仍然可用，可基本确保主中心与灾备中心的数据完全同步和一致，整个切换过程可控。

无计划的切换：灾难不可预知突然发生，主中心所有应用系统都是非正常结束，数据的完整性和一致性需要额外手段进行修复，网络在容灾中心数据库、应用启动前须切换完成。

灾难恢复规划与制定流程虽然重要，但是基于恢复流程的演练本身更为关键。有计划的、重复的演练，能够完善规划，熟练的全员演练可以优化恢复过程，使得灾难发生时快速、反射式应对，有效构建新的生产环境。图 9 显示了一个简化的容灾演练过程。

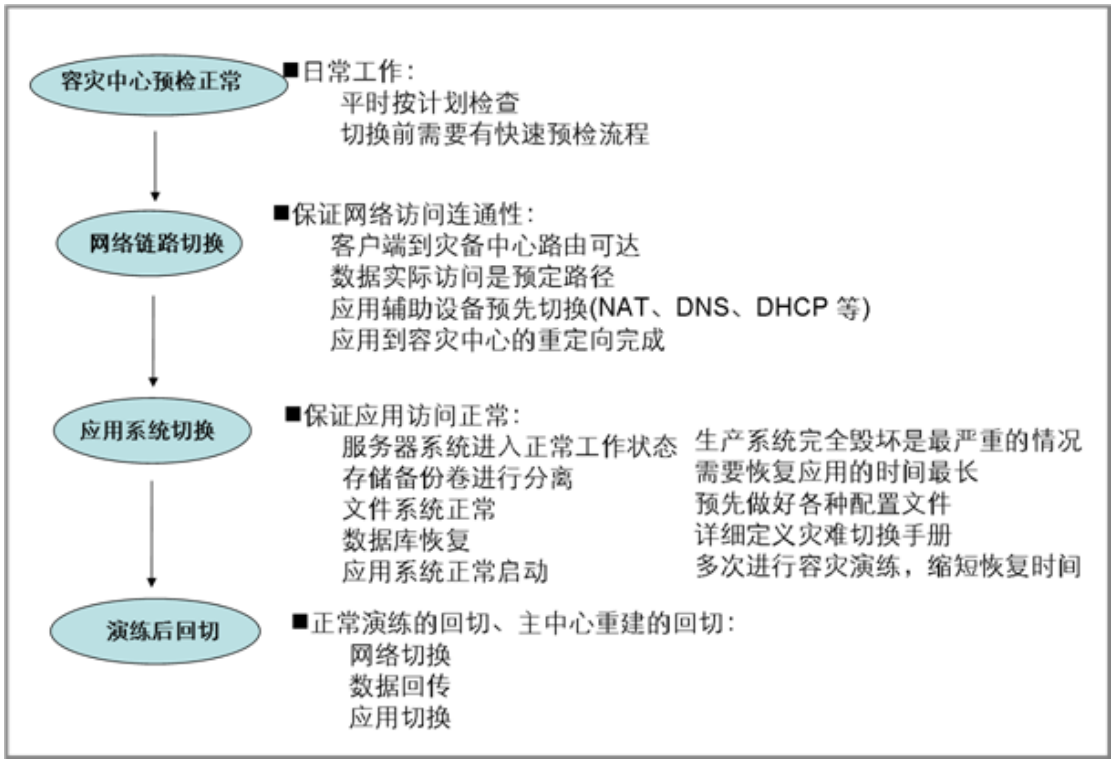

图 9 一个简化的灾难恢复演练流程

6 结束语

容灾建设的要素有很多，全面考虑业务连续性体系的每一个方面，抓住最为重要的环节，进行深入而细致的研究，就能够使有限的资源发挥更大的能量。容灾建设关乎着金融企业持续生存发展要义，容灾需要打通技术、环境、人等各个环节，形成一个完整流畅的灾难应对共同体，才能确保灾备系统成为抵御不可抗灾难的有力工具。