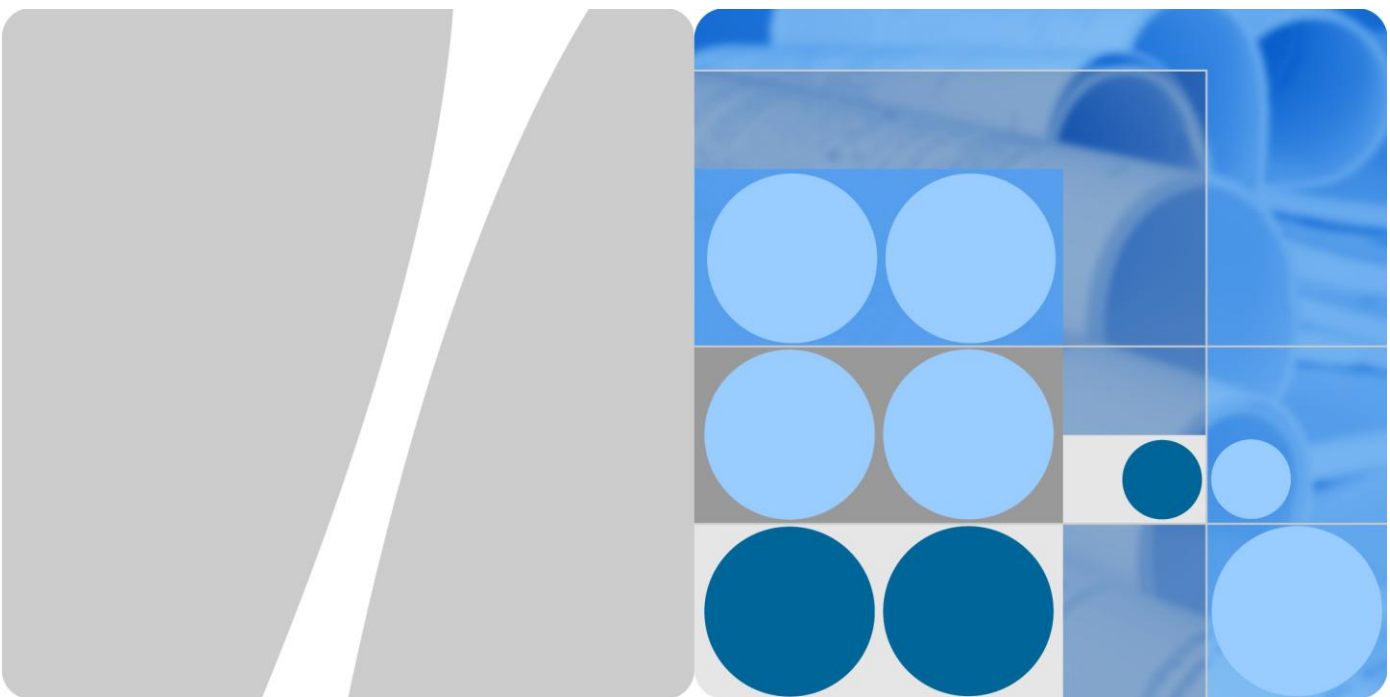




业务连续性容灾解决方案

双活数据中心解决方案技术白皮书 (HyperMetro)

文档版本 01
发布日期 2015-06-29

华为技术有限公司



版权所有 © 华为技术有限公司 2015。 保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明



和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： <http://e.huawei.com>

目 录

1 灾备建设的挑战与趋势	1
2 华为双活数据中心解决方案介绍	2
2.1 双活数据中心架构	2
2.2 双活数据中心部署	4
2.3 客户价值	6
3 双活数据中心关键技术	8
3.1 存储层双活	8
3.1.1 AA 双活架构	9
3.1.1.1 并行访问	9
3.1.1.2 免网关设计	9
3.1.1.3 I/O 访问路径	9
3.1.1.4 存储层组网	10
3.1.2 高可靠技术	11
3.1.2.1 跨站点集群	11
3.1.2.2 跨站点数据实时镜像	13
3.1.2.3 跨站点坏块修复	14
3.1.2.4 仲裁防脑裂	15
3.1.2.5 高可靠链路设计	19
3.1.2.6 分布式锁技术	20
3.1.3 高性能技术	23
3.1.3.1 数据零拷贝	23
3.1.3.2 FastWrite	23
3.1.3.3 地域优化访问	24
3.1.4 高可扩展性	26
3.1.4.1 两地三中心扩展	27
3.1.4.2 异构双活	27
3.1.4.3 本地保护	28
3.2 计算层双活	30
3.3 应用层双活	31
3.3.1 B/S 应用双活	31

3.3.2 C/S 应用双活.....	33
3.3.3 数据库双活.....	35
3.3.3.1 Oracle RAC 双活	37
3.3.3.2 DB2 双活	38
3.3.3.3 SQL Server 双活	39
3.4 网络层双活.....	40
3.4.1 网络架构	40
3.4.2 跨数据中心网络	40
3.4.3 业务访问网络架构	42
3.4.3.1 B/S 应用网络架构	42
3.4.3.2 C/S 应用网络架构	43
3.4.4 二层互联	47
3.4.5 负载均衡技术.....	49
3.4.5.1 站点间负载均衡	49
3.4.5.2 站点内负载均衡	52
3.5 传输层技术.....	54
3.6 安全层技术.....	54
4 可视化容灾管理.....	60
4.1 总体部署	60
4.1.1 部署方式	60
4.1.2 系统运行环境.....	61
4.2 应用支持矩阵.....	63
4.3 适用容灾场景.....	63
4.3.1 SAN 双活场景.....	63
4.3.1.1 技术特点.....	63
4.3.1.2 物理拓扑.....	64
4.3.1.3 逻辑拓扑.....	64
4.3.2 SAN 双活+快照场景	65
4.3.2.1 技术特点.....	65
4.3.2.2 物理拓扑.....	65
4.3.2.3 逻辑拓扑.....	65
4.3.3 SAN 双活+异步复制场景.....	66
4.3.3.1 技术特点.....	66
4.3.3.2 物理拓扑.....	66
4.3.3.3 逻辑拓扑.....	67
5 故障场景	68
5.1 GSLB 故障	68
5.2 SLB 故障	69
5.3 Web 服务器故障.....	71

5.4 应用服务器故障.....73

5.5 Oracle RAC 故障.....76

5.6 IBM DB2 故障.....77

5.7 阵列故障.....78

5.8 广域网链路故障.....80

5.9 站点间链路故障.....81

5.10 站点故障.....82

6 缩略语.....83

1 灾备建设的挑战与趋势

随着信息化技术的飞速发展，信息系统在各种行业的关键业务中扮演着越来越重要的角色。在通讯、金融、医疗、电子商务、物流、政府等领域，信息系统业务中断会导致巨大经济损失、影响品牌形象并可能导致重要数据丢失。因此，保证业务连续性是信息系统建设的关键。

目前，各行业普遍通过建设灾备中心来提高关键应用的业务连续性，在灾备中心保存生产数据副本。传统灾备部署方式为一个生产中心对应一个灾备中心，灾备中心平时不对外提供业务访问。一旦生产中心发生灾难，业务瘫痪，无法短时间恢复时，灾备中心才应需启动以提供业务访问。这种灾备系统面临以下挑战：

- 当生产中心遭遇供电故障、火灾、洪灾、地震等灾难时，需要手动将业务切换到灾备中心，有可能需要专业的恢复手段和长时间调试，业务中断时间长，无法保证业务连续运行。
- 灾备中心不能对外提供服务，常年处于闲置状态，资源利用率低。

华为针对客户提高资源利用率，在两个数据中心间实现负载均衡和灾难自动切换的需求，推出了端到端双活数据中心解决方案。

2 华为双活数据中心解决方案介绍

双活数据中心解决方案指两个数据中心均处于运行状态，可以同时承担生产业务，提高数据中心的整体服务能力和系统资源利用率。

业界目前有两种双活形态：AP 双活和 AA 双活。

- AP 双活通过将业务分类，部分业务以数据中心 A 为主，数据中心 B 为热备，而部分业务则以数据中心 B 为主，数据中心 B 为热备，以达到近似双活的效果。
- AA 双活则是真正的双活，同一个双活 LUN 的所有 I/O 路径均可同时访问，业务负载均衡，故障时可无缝切换。

华为双活数据中心解决方案为 AA 双活架构，以业界领先的 OceanStor V3 HyperMetro 功能为基础，与 Web、数据库集群、负载均衡、传输设备和网络等组件相结合，为客户提供 100km 内的端到端双活数据中心解决方案，确保业务系统发生设备故障、甚至单数据中心故障时，业务无感知，自动切换，实现 RPO（Recovery Point Objective）=0，RTO（Recovery Time Objective）=0（RTO 与应用系统及部署方式有关）。

2.1 双活数据中心架构

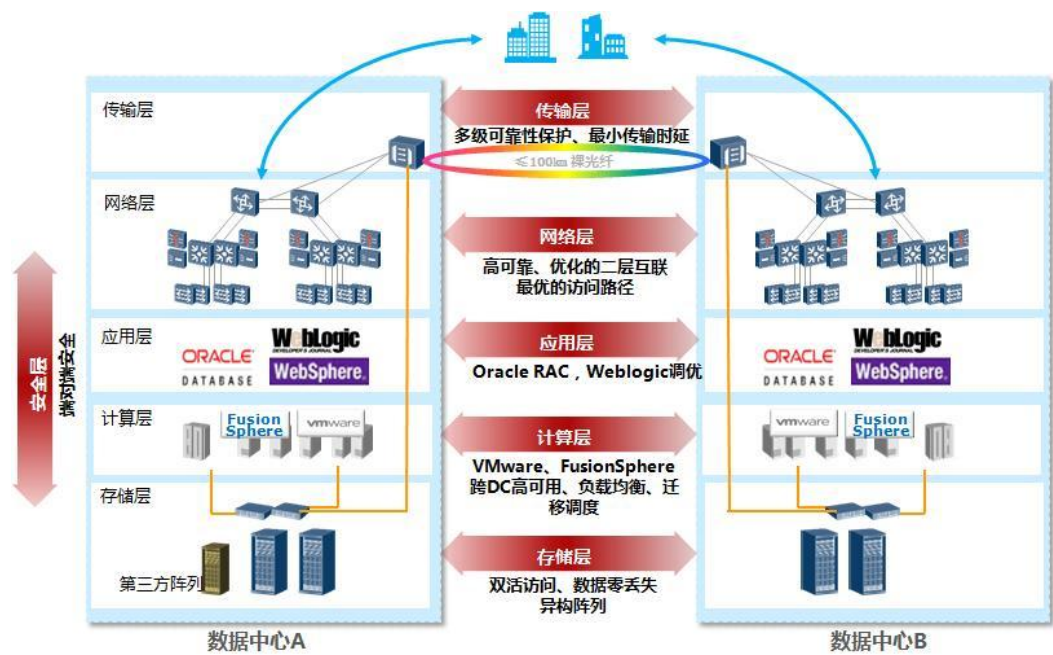
2.2 双活数据中心部署

2.3 客户价值

2.1 双活数据中心架构

端到端双活数据中心解决方案分为 6 层：存储层、计算层、应用层、网络层、传输层和安全层。逻辑架构图如图 2-1 所示。

图2-1 逻辑架构图



端到端双活数据中心解决方案，为了实现更好的可靠性、性能和负载均衡，针对每一层进行一些设计和优化，表 2-1 列举了一部分设计要点。

表2-1 6 层方案设计要点

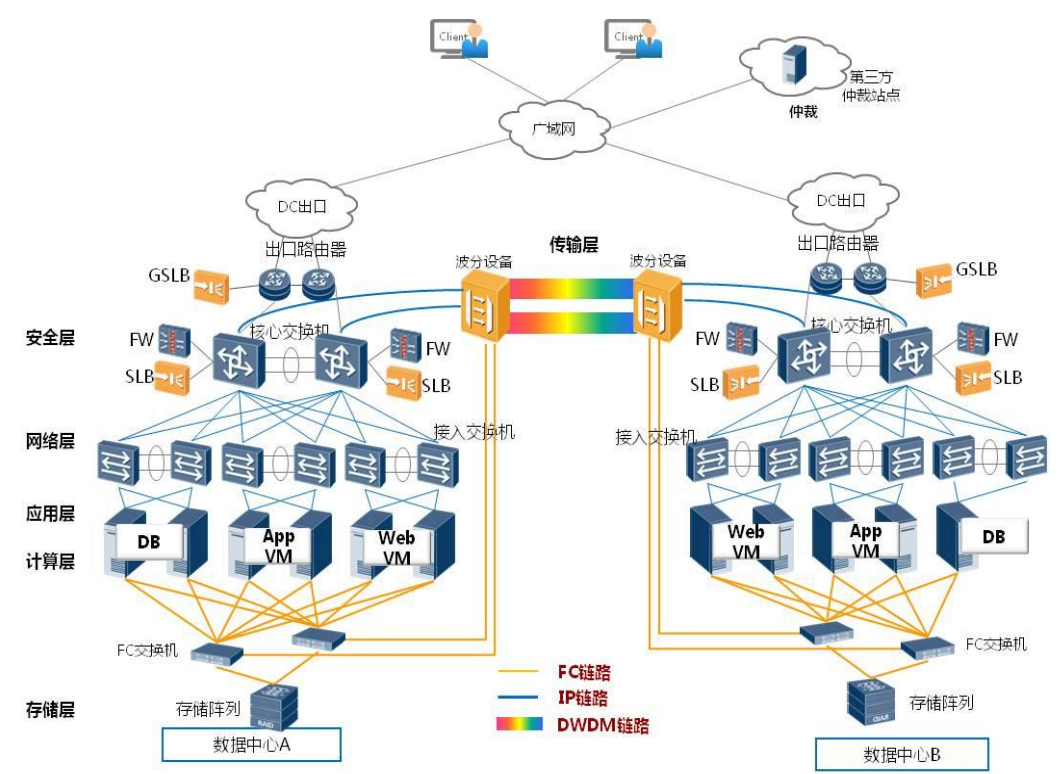
方案模块	设计要点
存储层	- 免网关双活架构。 - 通过 OceanStor V3 阵列的 HyperMetro 功能，实现存储层的双活，减少方案故障点，避免存储虚拟化网关的 I/O 性能瓶颈。 - 通过 OceanStor V3 阵列的 SmartVirtualization 功能，接管现网异构存储，将不同的存储资源的统一池化。 - 通过 FastWrite 功能，将标准的 1 个写 I/O 2 次往返优化为 1 次往返，提升写性能。 - 通过自适应网络功能，提升链路性能不均匀时的双活性能。
网络层	- 采用华为 CloudEngine 系列数据中心交换机的 EVN 技术。 - 通过 EVN 打通大二层网络，实现 2 层网络协议在 3 层网络上运行，保障了业务的跨数据中心互联互通。 - 通过多种互联优化技术，实现广播域隔离，达到数据中心间东西向流量优化。 - 通过双活网关和路由注入等访问优化功能，优化数据中心间南北向流量优化。
安全层	通过华为 USG 系列提供数据中心的 FW、IDS 等安全防护功能。

方案模块	设计要点
	通过数据中心间光传输加密技术，支持 AES-256 等多种加密算法，预防数据中心间数据窃听。
计算层	采用华为 FusionSphere、VMware 等虚拟化平台提供跨数据中心大集群技术，满足企业多种关键业务双活的需求。
应用层	- web 和 App 应用基于虚拟化集群提供更好的可靠性，通过负载均衡实现业务自动切换。 - 数据库基于双活 LUN 实现跨站点集群双活部署。
传输层	- 采用华为 OptiX OSN 系列作为双活数据中心的波分设备。 - 通过线路冗余，板卡冗余，设备冗余三种 1+1 保护方案，满足多级可靠性要求。 - 通过色散补偿等调优手段，确保传输层延时最小。

2.2 双活数据中心部署

整体的方案物理组网图如图 2-2 所示。

图2-2 物理组网图



各方案模块部署如表 2-2 所示：

表2-2 方案模块部署方式

模块	部署方式
存储层	- 跨 DC 的两套华为 OceanStor V3 系列存储阵列组成一个存储集群。 - 支持其中一台异构接管第三方存储，使用接管后的 LUN 与另一台 V3 上的 LUN 构建双活 LUN。
网络层	- 采用华为 CloudEngine 系列数据中心交换机作为核心交换机。 - 数据中心内部采用典型二层或三层物理架构组网，启用 EVN 形成二层通道，由核心交换机通过 CSS+链路聚合接入波分设备。 - 每个站点部署一台独立的 GSLB 实现站点间负载均衡。 - 每个站点部署 2 台 SLB，组成 HA 集群，实现应用层服务器的负载均衡。
应用层	- Web、App 层可以部署在虚拟机或者物理机上，DC 内的多台服务器组成集群，或者跨 DC 的多台服务器组成集群。 - 数据库建议物理机部署，跨数据中心组成一个集群。
计算层	使用华为 FusionSphere、VMware 等虚拟化平台，跨数据中心组成虚拟主机集群。

模块	部署方式
传输层	- 采用华为 OptiX OSN 系列 DWDM，每个站点部署 2 套波分设备。 - 如若不能设备级冗余，则需要至少每套波分设备配置 2 块传输板卡，实现板卡冗余。 - 将多路 FC 信号和 IP 信号复用到光纤链路上传输，每套波分设备通过两对裸光纤互联。
安全层	- 采用华为 USG 系列防火墙，每个站点部署 2 台防火墙，接入核心交换机。 - 在华为 OptiX OSN 系列 DWDM 启用传输加密功能。
仲裁	- 选择一个第三方站点部署仲裁设备和软件。 - 软件支持安装在物理服务器或虚拟机上。 - 仲裁服务器使用 IP 网络连接到双活数据中心的两套存储阵列。



说明

GSLB 是英文 Global Server Load Balance 的缩写，意思是全局负载均衡。作用：实现在广域网（包括互联网）上不同地域的服务器间的流量调配，保证使用最佳的服务器服务离自己最近的客户，从而确保访问质量。

SLB 是 Server Load Balancing，意思是服务器负载均衡。SLB 可以看作 HSRP（热备份路由器协议）的扩展，实现多个服务器之间的负载均衡。

2.3 客户价值

华为充分利用其宽产品线的优势，通过多产品的紧耦合，为客户提供端到端双活数据中心解决方案，也是业界唯一可提供端到端双活解决方案的厂家。

6 层 Active-Active 可靠性设计，业务负载均衡，应用零中断，数据零丢失，实现了业界最高等级的业务连续性保障。

华为双活数据中心解决方案有以下特点：

- AA 双活架构，数据零丢失，业务零中断（RPO=0，RTO =0）。
- 两个数据中心同时提供业务，充分利用灾备资源。
- 支持异构存储，保护已有设备投资。
- 方案扩展灵活，容灾可视化管理。

华为 6 层双活数据中心解决方案能给客户带来以下价值：

- 端到端双活设计：华为提供端到端双活方案设计，快速上线业务。
- 真 AA 精简双活，业务 7×24 小时持续运行：去存储网关精简双活架构，方案故障点少，系统可靠性更高，提供站点同时读写的 Active-Active 双活能力。
- 双活 I/O 优化，性能高：去存储网关的双活架构，I/O 处理路径最短。同时，通过锁预取优化，存储协议优化和地域优化访问等极大提升双活业务性能。

- 利旧现有设备，保护现有投资：存储支持异构接管第三方存储，计算层支持 FusionSphere 等虚拟化平台利旧已有服务器，保护现有投资。

3 双活数据中心关键技术

以下是华为双活数据中心解决方案的关键技术：

- 存储层：通过 HyperMetro 实现存储层的双活。
- 计算层：通过 FusionSphere、VMware 等虚拟化技术，提供虚拟机 HA 特性，故障时自动恢复。
- 应用层：通过应用集群和数据库集群技术实现双活。
- 网络层：通过 DWDM、EVN 等二层互联技术，实现低时延、高可靠的二层网络互联；通过网络设备的双活网关、RHI 等路径优化技术，以及全局负载均衡器、服务器负载均衡器实现双活就近接入或高可用网络切换。
- 传输层：通过设备冗余及板卡冗余构建可靠的双活传输网络。
- 安全层：通过防火墙和安全策略规划和设计保证访问安全，通过传输层加密特性，保证跨数据中心数据传输安全。

华为双活数据中心解决方案在以上 6 个层次上进行联动及联合，为客户提供端到端双活方案。

3.1 存储层双活

3.2 计算层双活

3.3 应用层双活

3.4 网络层双活

3.5 传输层技术

3.6 安全层技术

3.1 存储层双活

华为双活数据中心解决方案存储层基于 OceanStor V3 融合存储系统的 HyperMetro 特性实现。HyperMetro 采用 AA 双活架构将两套存储阵列组成跨站点集群，实现数据实时镜像，具有高可靠、高性能和高扩展的特点。本章节将对 HyperMetro 特性的主要技术原理和特点进行介绍。

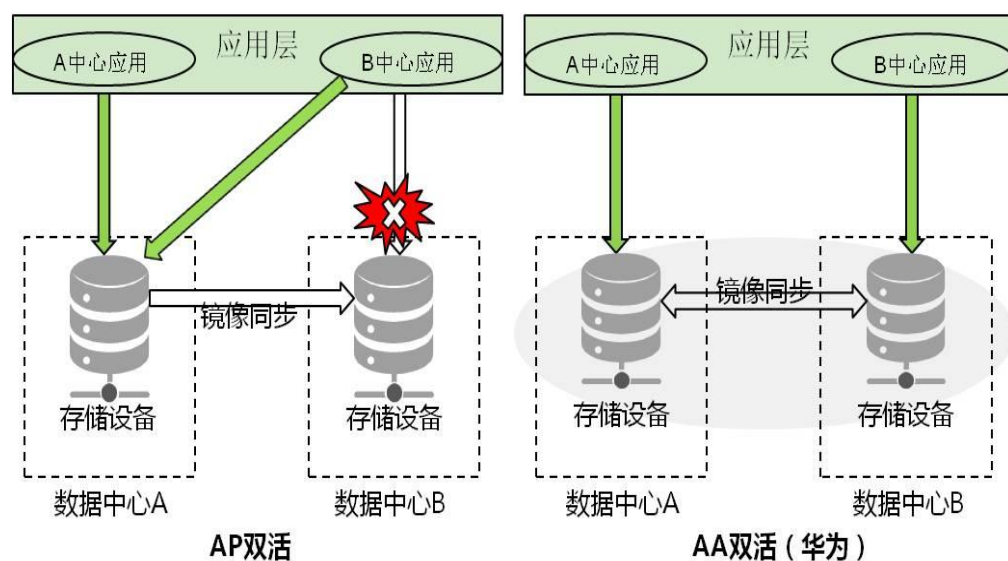
3.1.1 AA 双活架构

3.1.1.1 并行访问

HyperMetro 特性基于两套存储阵列实现 AA(Active-Active)双活，两端阵列的双活 LUN 数据实时同步，且双端能够同时处理应用服务器的 I/O 读写请求，面向应用服务器提供无差异的 AA 并行访问能力。当任何一台磁盘阵列故障时，业务自动无缝切换到对端存储访问，业务访问不中断。

相较于 AP 方案，AA 双活方案可充分利用计算资源，有效减少阵列间通信，缩短 I/O 路径，从而获得更高的访问性能和更快的故障切换速度。图 3-1 展示了几种双活方案的交互流程。

图3-1 存储双活架构



3.1.1.2 免网关设计

HyperMetro 双活架构无需额外部署虚拟化网关设备，直接使用两套存储阵列组成跨站点集群系统。最大支持 32 个存储控制器，即两套 16 控存储阵列组建双活关系。

本方案架构精简、与存储增值特性良好兼容，对客户价值如下：

- 减少网关故障点，提高方案可靠性。
- I/O 响应速度更快，无需经过存储网关转发，减少网关转发 I/O 时延。
- 双活可以兼容存储阵列已有特性，与 OceanStor V3 融合存储系统的其它 Smart 系列和 Hyper 系列特性组合，可为客户提供多种数据保护和灾备解决方案。
- 显著降低双活组网复杂度，便于维护。

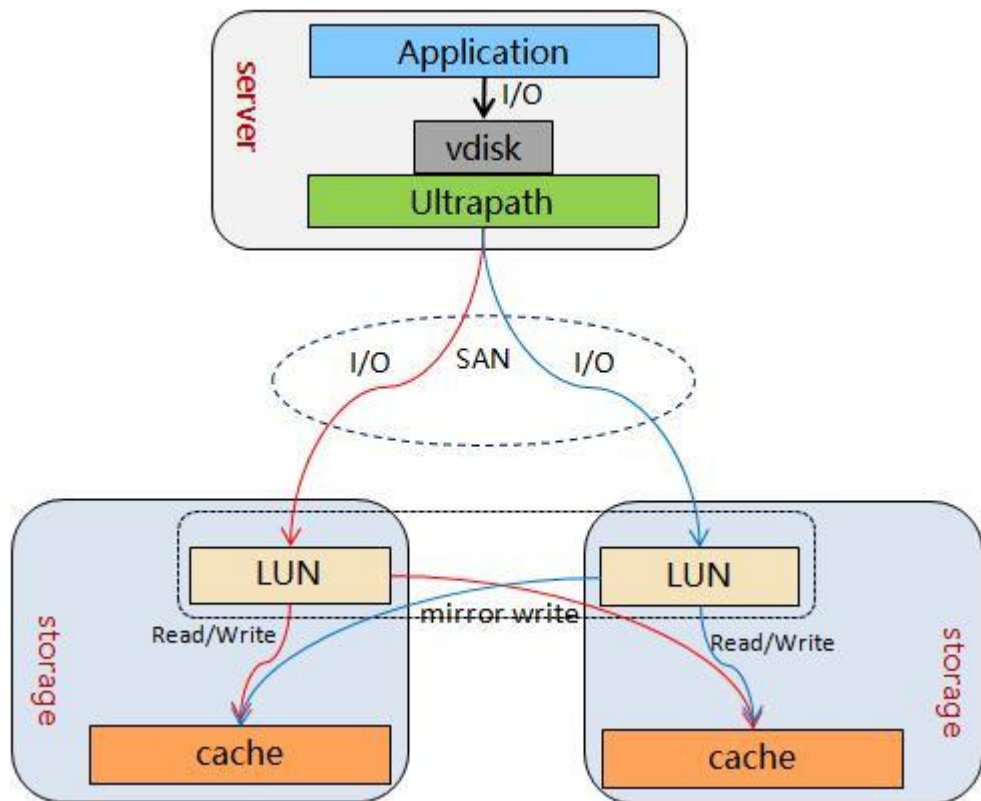
3.1.1.3 I/O 访问路径

HyperMetro 在应用主机侧，通过 UltraPath 主机多路径软件，将两台存储阵列上的双活成员 LUN 聚合为一个双活 LUN，以多路径 vdisk 方式对应用程序提供 I/O 读写能力。

应用程序访问 vdisk 时，Ultrapath 根据选路模式，选择最佳的访问路径，将 I/O 请求下发到存储阵列（参见 3.1.3.3 地域优化访问）。

存储阵列的 LUN 空间上接收到 I/O 请求后，对于读 I/O 请求，直接读本地 Cache 空间，将数据返回应用程序；对于写 I/O 请求，首先会进行并行访问互斥（参见 3.1.3.4 分布式锁技术），获取写权限后，将 I/O 请求数据同时写本地双活成员 LUN Cache 以及对端的双活成员 LUN Cache，双端写成功后返回应用程序写完成（参见 3.1.2.2 跨站点数据实时镜像）。

图3-2 双活 IO 路径

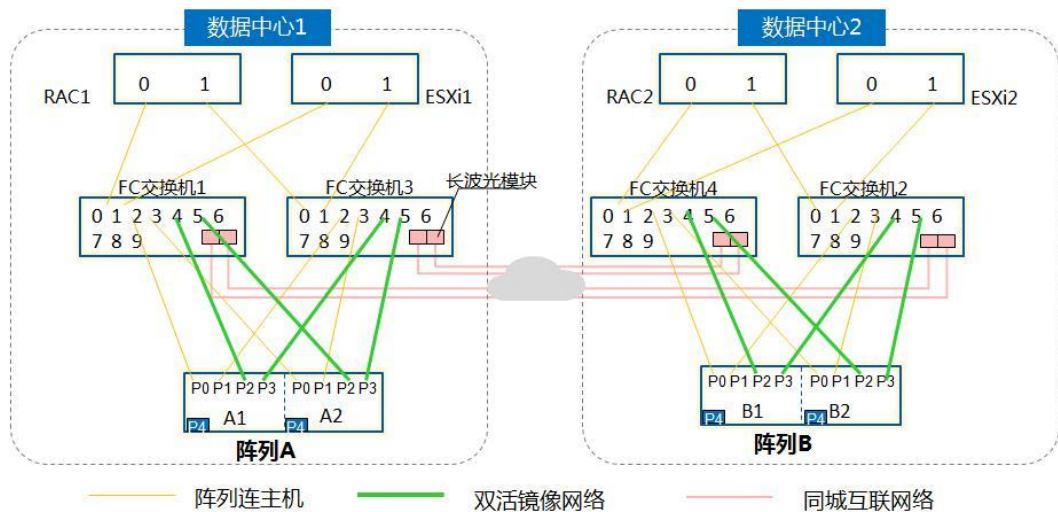


3.1.1.4 存储层组网

两套双活存储阵列间通信支持 FC 或 IP 链路，推荐使用 FC 链路。另外，存储阵列和仲裁服务器之间的链路采用更易于获取的 IP 链路。

图 3-3 为 4 控制器的双活组网情况下，FC 交换机端口占用情况。

图3-3 阵列双活方案组网



3.1.2 高可靠技术

HyperMetro 在继承 OceanStor 存储系统高可靠设计的基础上，全新设计了一些解决方案级高可靠技术，最大化提高了存储双活方案的可靠性。

本章节将从以几个方面介绍 HyperMetro 的高可靠技术：

- 跨站点集群
- 跨站点数据实时镜像
- 跨站点坏块修复
- 仲裁防脑裂
- 高可靠链路设计

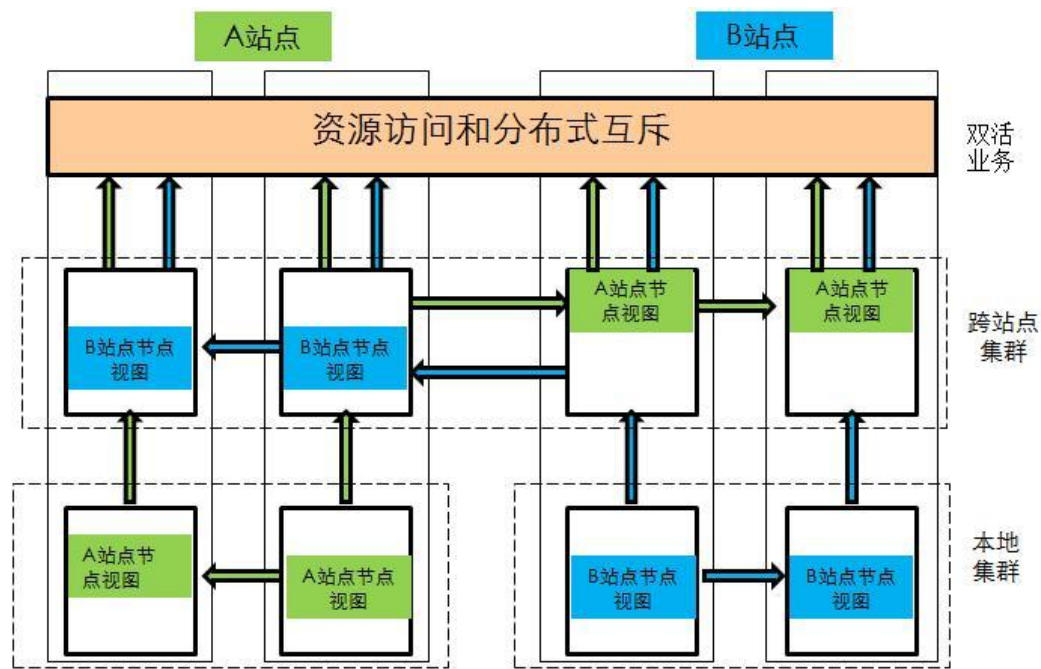
3.1.2.1 跨站点集群

两套独立的存储阵列组建成跨站点集群，并以跨站点集群为核心，提供双活存储架构，向应用服务器提供无差异的并行访问，处理应用服务器的 I/O 请求。

双活跨站点集群配置过程极为简单，只需要将两套存储阵列配置成双活域，即可完成跨站点集群配置。

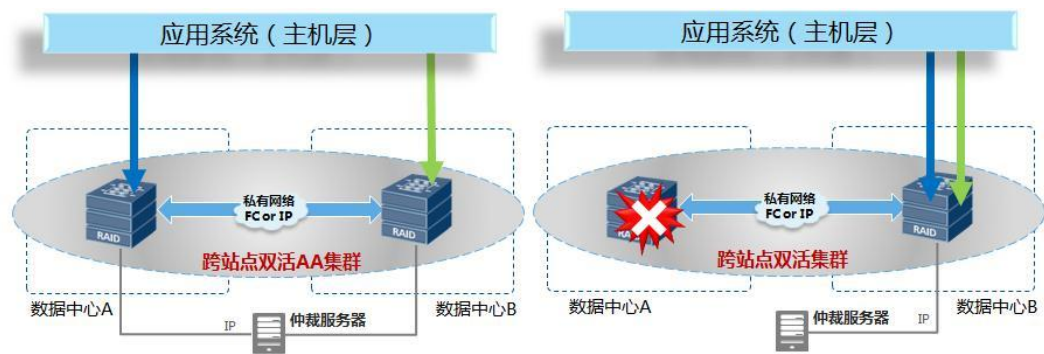
跨站点集群系统使用阵列间 FC 或 IP 链路作为通信链路，完成全局节点视图建立和状态监控。在全局节点视图基础上，跨站点集群系统提供分布式互斥等能力，支持 AA 双活架构。下图为双活跨站点集群示意图。

图3-4 双活跨站点集群



集群节点具有并发访问能力。当出现单个控制器故障时，其承接的业务将被切换到本地集群的其它工作控制器；本地集群工作控制器全故障时，则切换至跨站点集群另一个本地集群。

图3-5 双活访问与切换



在跨站点集群基础上，HyperMetro 以双活 Pair 或双活一致性组为单位提供服务 and 进行状态管理。

两套存储阵列上的双活成员 LUN 组成一个虚拟双活 LUN，通过实时镜像技术保持两个数据中心的成员 LUN 的数据实时一致。

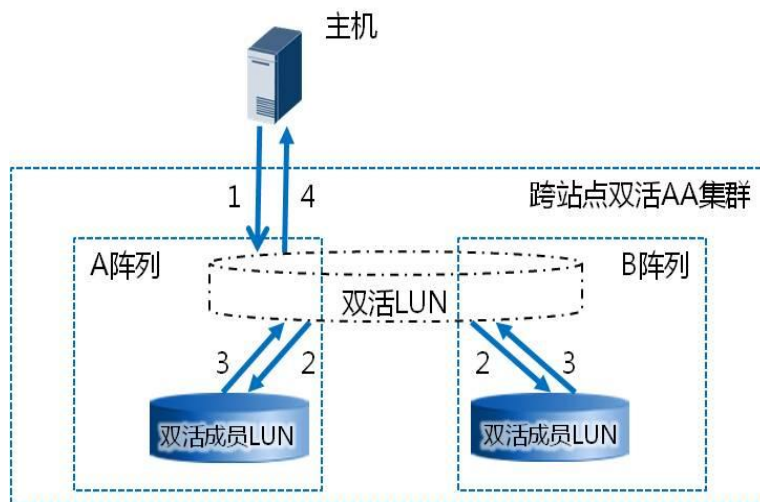
一致性组是多个双活 pair 的集合，可以确保单个存储系统内，主机在跨多个 LUN 进行写操作时数据的一致性。

一致性组进行分裂、同步等操作时，一致性组的所有双活 pair 保持步调一致。当遇到链路故障时，一致性组的所有成员对会一起进入异常断开状态。当故障排除后，所有成员同时进行数据的同步，从而保证从站点灾备阵列数据的可用性。

3.1.2.2 跨站点数据实时镜像

HyperMetro 通过实时镜像功能，保证两个站点存储阵列之间数据的实时同步。主机写操作通过实时镜像技术同时写入两个数据中心的双活成员 LUN，保持数据实时一致。具体的写 I/O 流程如图 3-6 所示。

图3-6 跨站点镜像



假如数据中心 A 阵列收到写 I/O，镜像处理流程如下：

1. 申请写权限和记录写日志：数据中心 A 阵列收到主机写请求，先申请双活 Pair 的写权限。获得写权限后，双活 Pair 将该请求记录写日志。日志中只记录地址信息，不记录具体的写数据内容。该日志采用具有掉电保护能力的内存空间记录以获得良好的性能。
2. 执行双写：将该请求拷贝两份分别写入本地 LUN 和远端 LUN 的 Cache。
3. 双写结果处理：等待两端 LUN 的写处理结果都返回。
4. 响应主机：双活 Pair 返回写 I/O 操作完成。

HyperMetro 支持断点续传功能。当某些故障场景（如单套存储故障）导致双活 Pair 关系异常断开时，HyperMetro 通过记录日志的方式，记录主机新产生的写 I/O。当故障恢复时，HyperMetro 将自动恢复双活 Pair 关系，并且将所记录的增量数据自动同步到远端，无需全量同步所有数据，整个过程对主机“透明”，不会影响主机业务。

双活 Pair 运行状态和主机访问状态关系见表 3-1。

表3-1 双活主机访问状态表

双活 Pair 运行状态	主机访问状态		状态描述
	主 LUN	从 LUN	

双活 Pair 运行状态	主机访问状态		状态描述
暂停	读写	不可读写	用户暂停双活镜像关系
待同步	读写	不可读写	阵列间链路故障或 I/O 错误导致双活镜像关系断开
同步中	读写	不可读写	恢复双活镜像关系时全量/增量同步双端差异数据
正常	读写	读写	两端 LUN 都进入双活 AA 实时镜像关系
强制启动	读写	不可读写	用户进行了强制将双活从 LUN 升级为主 LUN 的操作

双活 Pair 运行状态和镜像状态关系见表 3-2。

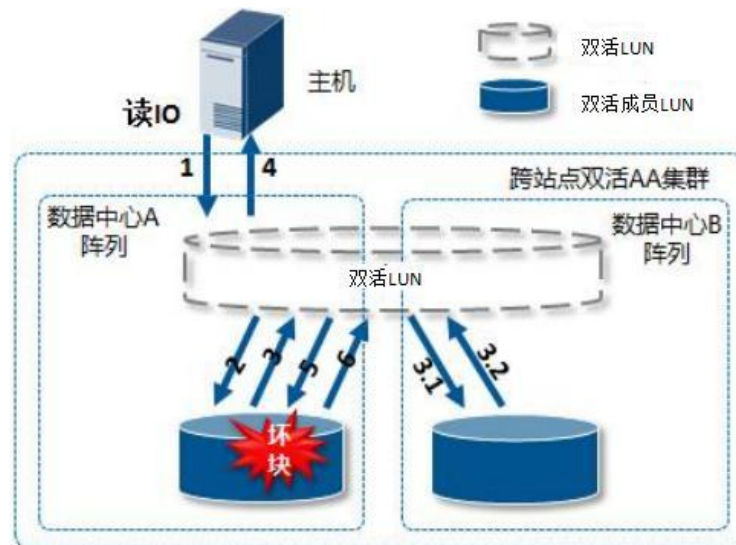
表3-2 双活镜像状态表

双活 Pair 运行状态	镜像状态	
	主 LUN	从 LUN
暂停/待同步/强制启动	不镜像，记录差异日志	不涉及
同步中	镜像写，后台复制差异	不涉及
正常	镜像写	镜像写

3.1.2.3 跨站点坏块修复

硬盘在使用过程中可能因为掉电等异常情况出现坏块，如果是可修复错误但是本端已经无法修复时，HyperMetro 将自动从远端阵列获取数据，修复本地数据盘的坏块，进一步提高系统的可靠性。

图3-7 跨站点数据修复



数据中心 A 阵列出现坏块时，从该阵列读 I/O 处理流程如下：

- 步骤 1 主机下发读 I/O。
 - 步骤 2 读本地 LUN。
 - 步骤 3 读取到坏块后，如果为可修复错误，执行步骤 4，否则执行 1、2 后流程结束。
 1. 重定向远端读。
 2. 远端读返回。
 - 步骤 4 将读数据返回主机，确保主机响应的快速返回。
 - 步骤 5 根据远端的读数据，进行本地写入修复。
 - 步骤 6 写修复结果返回。
- 结束

3.1.2.4 仲裁防脑裂

当提供双活 LUN 的两套阵列之间的链路故障时，阵列已经无法实时镜像同步，此时只能由其中一套阵列继续提供服务。为了保证数据一致性，HyperMetro 通过仲裁机制决定由哪套存储继续提供服务。

HyperMetro 支持按双活 Pair 或双活一致性组为单位进行仲裁。当多个双活 Pair 提供的业务相互依赖时，用户需要把这些双活 Pair 配置为一个双活一致性组。仲裁完成后，一个双活一致性组只会在其中一套存储阵列继续提供服务。例如，Oracle 数据库的数据文件、日志文件可能分别存放在不同的 LUN 上，访问 Oracle 数据库的应用系统存放在另一些 LUN 上，相互之间存在依赖关系。配置双活时，建议数据 LUN、日志 LUN 和应用 LUN 分别配置双活 pair，并且加入同一个一致性组。

HyperMetro 提供了两种仲裁模式：

- 静态优先级模式
- 仲裁服务器模式

配置双活 Pair 前，需要配置双活域，双活域为逻辑概念，包括需要创建双活关系的两套存储阵列和仲裁服务器。每个双活 Pair 创建时均要选择双活域，每个双活域只能同时应用一种仲裁模式。

仲裁服务器模式比静态优先级模式具备更高的可靠性，可保证在各种单点故障场景下，业务连续运行。因此，华为双活方案推荐采用仲裁服务器模式。

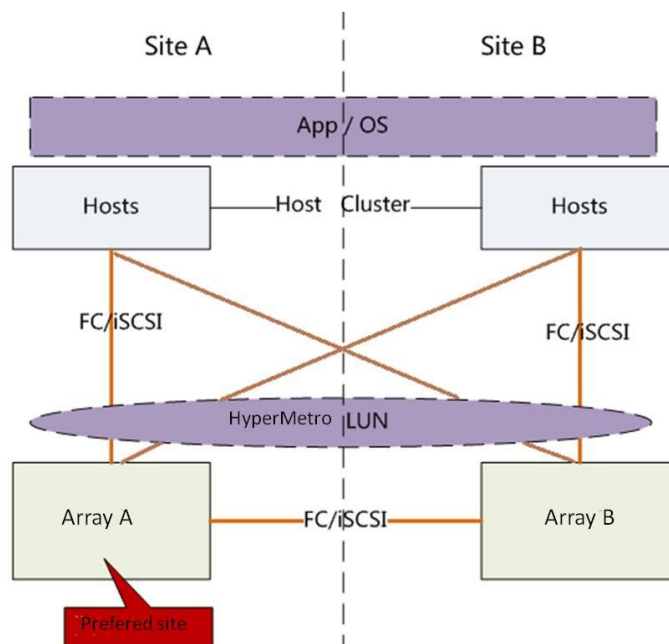
静态优先级模式

静态优先级模式主要应用在无第三方仲裁服务器的场景。用户可以按双活 Pair 或一致性组为单位，设置其中一端阵列为优先站点，另一端为非优先站点。如图 3-8 所示，不需要额外部署仲裁服务器。

该模式下，阵列间心跳中断时，优先站点仲裁胜利。

- 当发生阵列间链路故障，或者非优先站点故障时，优先站点上的 LUN 继续提供服务，非优先站点的 LUN 停止提供服务。
- 当优先站点阵列故障时，非优先站点不能自动接管双活业务，双活业务停止，需要人工强制启动非优先站点服务

图3-8 静态优先级部署





注意

有一种情况除外，当优先站点阵列主动下电维护时，非优先站点阵列立即接管所有双活业务，业务不会中断。

该模式的缺点是：两阵列之间的心跳丢失时，可能是站点间链路丢失或其中一个阵列故障，系统无法区分这两种情况。表 3-3 为静态优先级模式下的仲裁策略。

表3-3 静态优先级模式仲裁示意图

编号	示意图	仲裁结果
1		故障类型：链路故障 仲裁结果：H1 继续运行业务，H2 停止业务
2		故障类型：非优先故障 仲裁结果：H1 继续运行业务，H2 失效
3		故障类型：优先故障 仲裁结果：H1 失效；H2 停止业务，需要人工启动

仲裁服务器模式

使用独立的物理服务器或者虚拟机作为仲裁设备，仲裁服务器建议部署在第三方站点。这样可以避免单数据中心整体发生灾难时，仲裁设备也同时故障。如图 3-9 所示。

图3-9 仲裁服务器部署



仲裁服务器模式下，当存储阵列间心跳中断时，两端阵列向仲裁服务器发起仲裁请求，由仲裁服务器综合判断哪端获胜。仲裁获胜的一方继续提供服务，另一方停止服务。

仲裁服务器模式下如果有优先获得仲裁的要求，也可以配置站点优先级。优先阵列端具有仲裁获胜的优先权，心跳中断但其它正常时，优先阵列将获得仲裁胜利。

仲裁过程如图 3-10 所示：

图3-10 仲裁机制



- 1. 数据中心之间的链路断开时，跨站点阵列集群分裂为两个小集群。
- 2. 小集群分别抢占仲裁，优先阵列将优先抢占仲裁，抢占成功的小集群“获胜”，将继续对外提供服务，为应用提供存储访问空间；抢占失败的小集群则停止对外服务。
- 3. 中间链路恢复时，两个子集群检测到中间链路恢复正常，经过握手通信将两个小集群自动组成一个跨站点集群，双活关系恢复，以 Active-Active 模式提供服务。

表 3-4 列出了仲裁服务器模式下，各种故障场景下双活业务表现。

表3-4 各故障场景仲裁示意图

编号	示意图	仲裁结果
1		故障类型：仲裁失效 仲裁结果：H1、H2 继续运行业务
2		故障类型：一套阵列与仲裁之间链路故障 仲裁结果：H1、H2 继续运行业务
3		故障类型：一套阵列失效 仲裁结果：H1 失效，H2 继续运行业务
4		故障类型：阵列间链路中断 仲裁结果：H2 失效，H1 继续运行业务
5		故障类型：一套阵列与仲裁同时失效 仲裁结果：H1 失效，H2 停止业务

编号	示意图	仲裁结果
6		故障类型：一套阵列与对端、仲裁的链路同时中断 仲裁结果：H1 停止业务，H2 继续运行业务
7		故障类型：一套阵列失效，且对端与仲裁链路中断 仲裁结果：H1 失效，H2 停止业务
8		故障类型：仲裁失效，且阵列间链路中断 仲裁结果：H1 与 H2 均停止业务
9		故障类型：仲裁失效，且其与一套阵列链路中断 仲裁结果：H1、H2 继续运行业务

说明

H1 和 H2 表示组成双活 HyperMetro LUN 的两个阵列，C 表示对应的仲裁服务器。

强制启动

某些特定的多重故障情况下，仲裁机制优先保证数据的一致性，可能会将存活的双活成员 LUN 都停止主机访问。例如静态优先级模式下优先站点故障等场景，存活的双活成员 LUN 会被停止主机访问，用户或售后工程师可根据故障情况选择人工强制启动业务，快速恢复业务。

强制启动后，被强制启动端会升级为双活数据同步源端，强制启动端的双活成员 LUN 具有最新数据。链路恢复后，系统主动停止对端双活成员 LUN 主机访问。发起数据同步时，将以强制启动端的双活成员 LUN 数据覆盖对端。该过程中只会同步增量差异数据。



注意

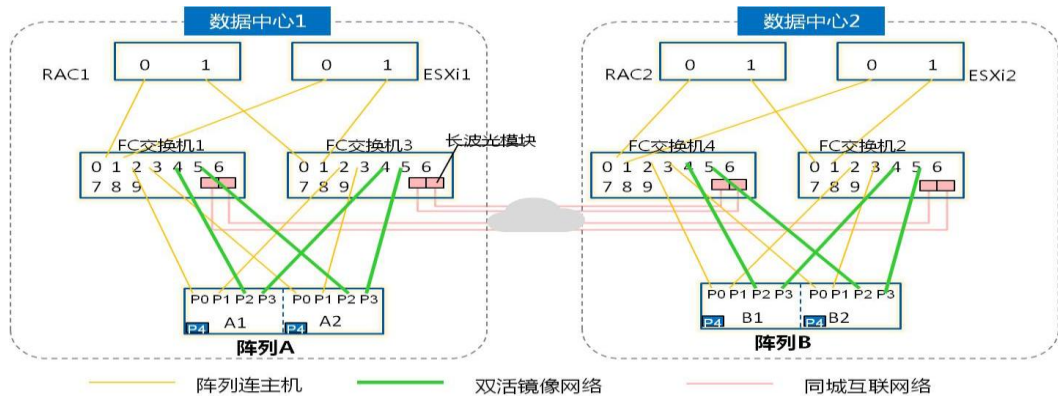
执行强制启动前，需要充分考虑双主风险，应在执行前确认两个数据中心的 LUN 状态和业务状态，确保对端存储已经停止工作。

3.1.2.5 高可靠链路设计

HyperMetro 阵列间组网支持 FC 或 IP 网络，可以根据用户的实际网络条件配置网络。支持使用阵列直连、通过 FC 光纤交换机连接、通过 IP 交换机连接等。为了更好的双活性能，推荐采用 FC 组网。

以‘跨数据中心 2 节点+2 节点双活组网’为例，建议每个控制器节点都与对端阵列建立 2 条阵列间镜像链路，并对两条链路的交换机进行隔离，获得最高的链路可靠性。

图3-11 以‘跨数据中心 4 控双活组网’为例



为保证双活性能，HyperMetro 对站点间双活链路的网络要求如下：

- 误码率 $\leq 10^{-12}$ 。
- 时延 $RTT \leq 2ms$ 。
- 无抖动，无丢包。
- 链路带宽>业务峰值带宽，最低要求 2Gbps。

除了高可靠的组网和链路质量要求外，双活阵列间镜像链路还应用了传输带宽自适应特性，该技术通过一系列自动检测机制，根据链路质量，动态调整各链路传输占比，减少数据重传率。

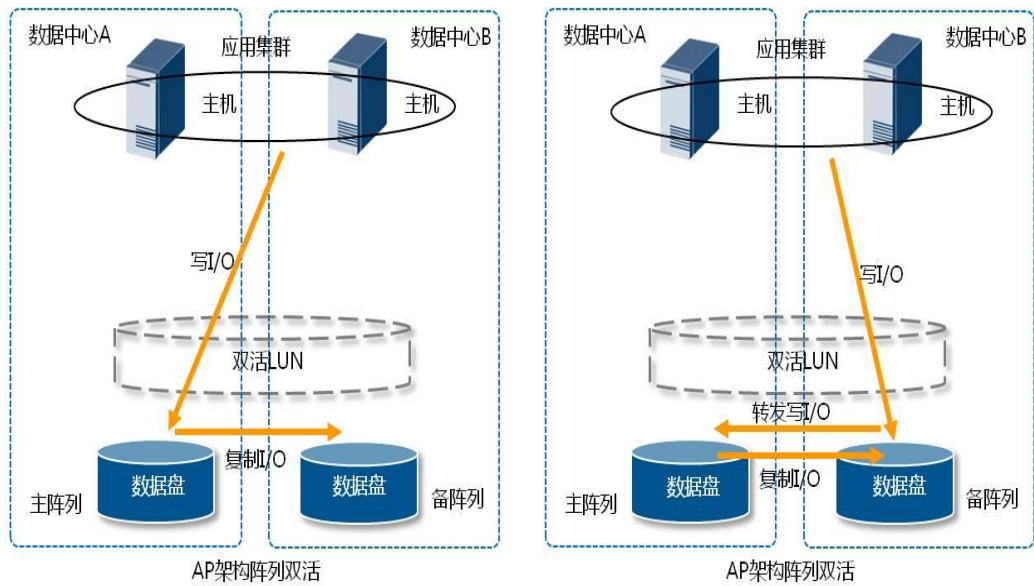
例如当前控制器与对端阵列控制器存在两条镜像链路 A 和 B，当系统检测到链路 A 因为误码等原因存在较大的传输时延时，会根据算法将链路 A 的流控带宽下降 20%，将更多的带宽压力导向链路 B，降低链路 A 的带宽压力，从而获得更稳定的传输时延。当链路 A 的传输时延较小时，系统又会根据算法将链路 A 的流控带宽上调 20%，从而恢复带宽吞吐量。

3.1.2.6 分布式锁技术

分布式互斥能力是实现 AA 双活的关键能力之一，双活分布式锁模块利用 Paxos 和 CHT（Consistent Hash Table）一致性算法，提供了分布式对象锁和分布式范围锁，从而满足 AA 双活的分布式互斥诉求。通过锁预取技术，可有效减少跨站点的数据传输量和通信交互次数，从而提升 I/O 读写性能。

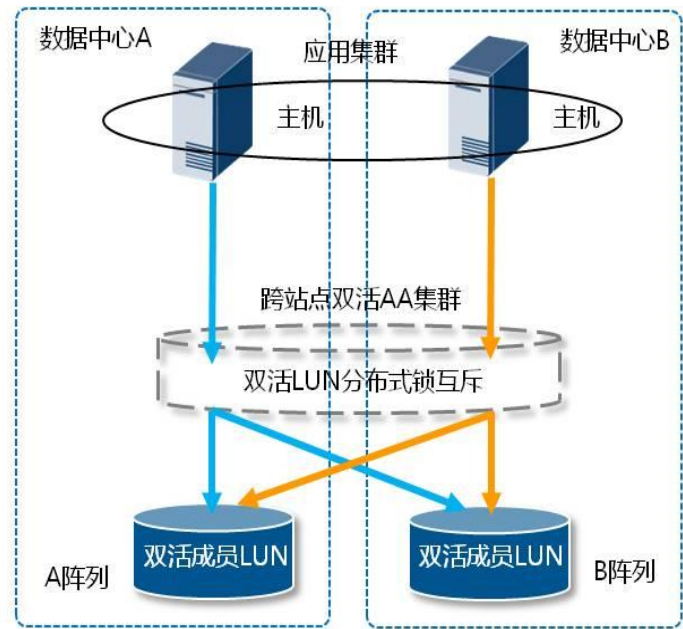
在 AP 双活架构中，由于主机无法通过从端直接访问双活 LUN，从端主机写数据时，必须将完整的写数据发送到主控端，再通过镜像链路把 I/O 从主控端同步到备控设备上，这样数据存在多次跨数据中心传输，严重影响写性能，如图 3-12 所示。

图3-12 AP 双活的数据访问



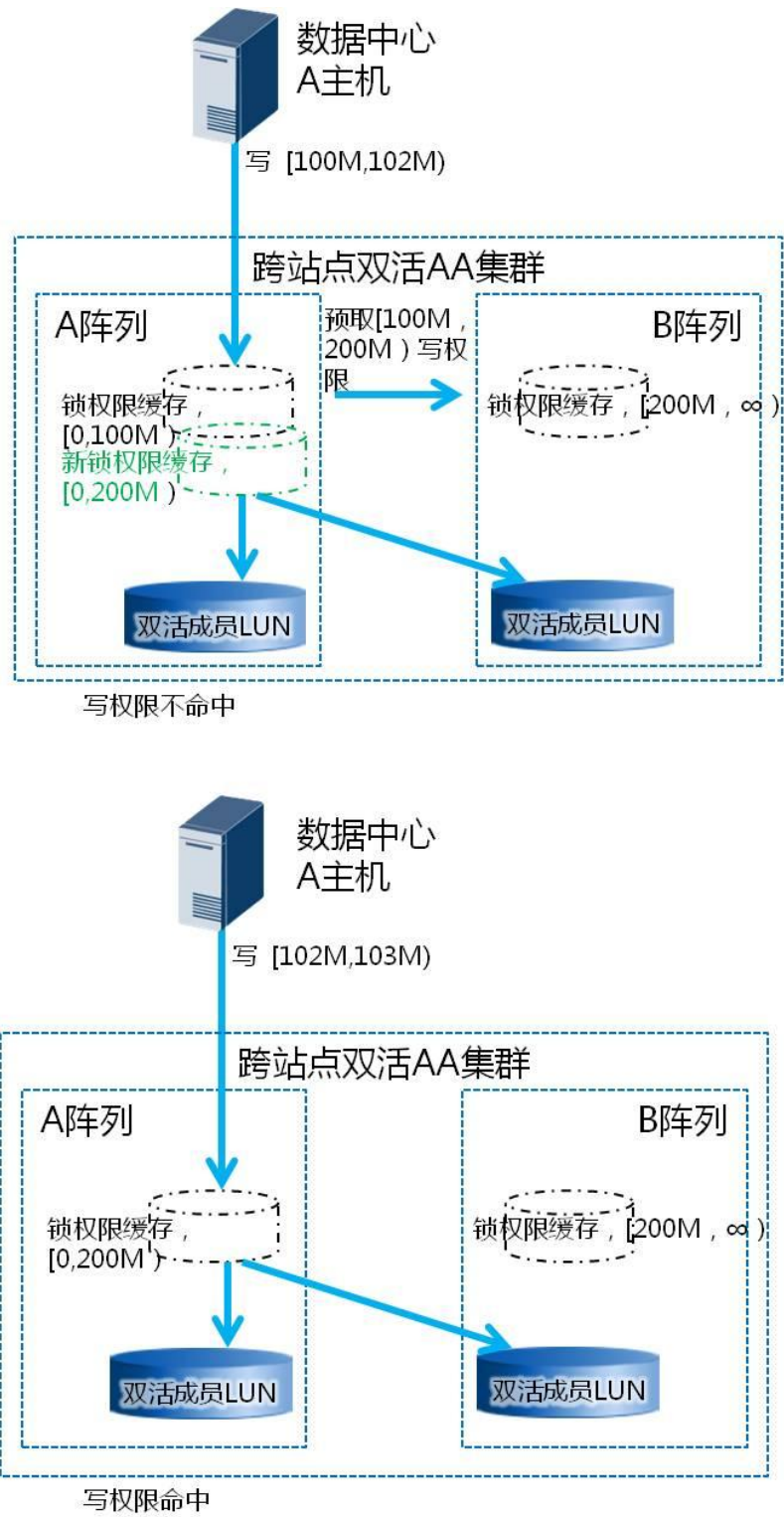
HyperMetro 以主机 I/O 粒度，对主机 I/O 访问的 LBA 区间加分布式范围锁进行并发互斥，从而达到双向实时同步的目的，该方案可省去不必要的阵列间数据传输带宽，并有效的减少数据传输次数。HyperMetro 双活数据访问如图 3-13 所示。

图3-13 应用分布式锁的双活数据访问



HyperMetro 分布式锁技术使用了智能的锁预取和缓存策略，在写权限本地无缓存的情况下，会通过较小的控制报文，向锁权限缓存节点申请写权限，并多预取部分区间的写权限缓存到本地。后续连续写 I/O 可快速在本地命中写权限，不需要再跨站点申请写权限。分布式锁实现原理如图 3-14 所示。

图3-14 分布式锁预取



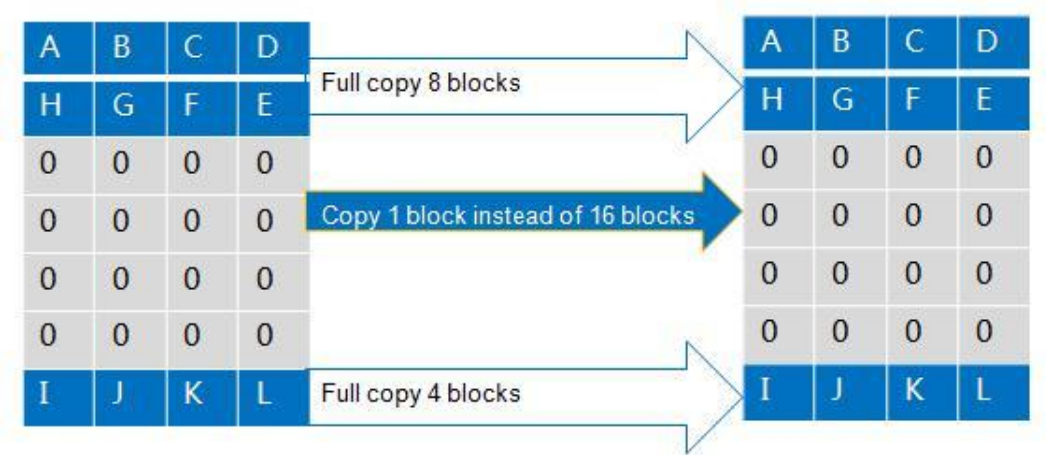
3.1.3 高性能技术

为了保证两个数据中心存储的数据实时一致，写操作都需要等待两端存储写成功之后再返回主机“写成功”。双活 I/O 性能因为实时双写导致一定的时延增加，华为 HyperMetro 设计了一系列 I/O 性能优化方案，减小对写时延的影响，提升整体双活的业务性能。

3.1.3.1 数据零拷贝

在双活镜像数据的初始同步或者恢复过程中的增量同步过程中，差异数据块通常有大量的零数据块，无需逐块复制，该功能叫数据零拷贝。例如，虚拟化场景下，新建虚拟机时会产生大量的零数据块，一个数十 GB 的操作系统盘，实际非零数据块仅 2-3GB。数据零拷贝原理图如图 3-15 所示。

图3-15 数据零拷贝



HyperMetro 零页面识别技术的实现方法如下：

通过硬件芯片，对数据拷贝源端进行快速识别，找出零数据，在拷贝过程中，对全零数据特殊标识，只传输一个较小的特殊页面到对端，不再全量传输。

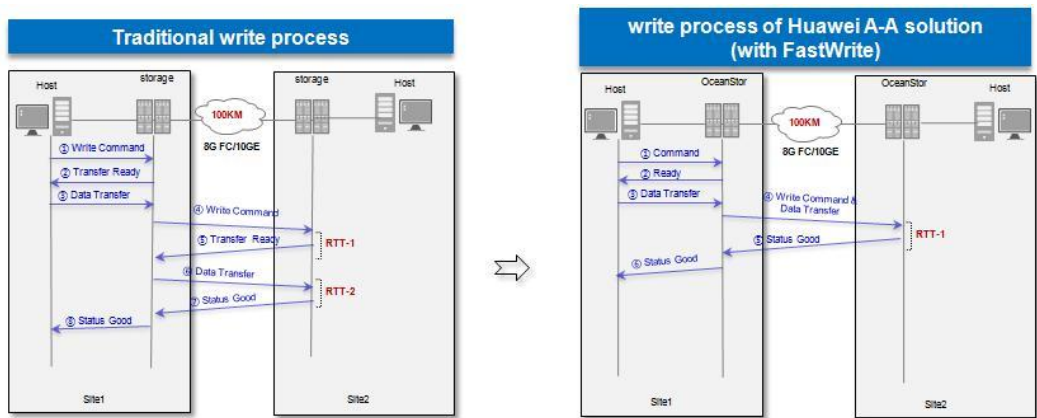
该技术可有效减少同步数据量，减少带宽消耗，缩短同步时间。

3.1.3.2 FastWrite

HyperMetro 通过 FastWrite 功能对阵列间数据传输进行了协议级优化，应用 SCSI 协议的 First Burst Enabled 功能，将写数据的链路传输交互次数减少一半。

正常的 SCSI 流程中，写 I/O 在传输的双端要经历“写命令”、“写分配完成”、“写数据”和“写执行状态”等多次交互。利用 FastWrite 功能，优化写 I/O 交互过程，将“写命令”和“写数据”合并为一次发送，并取消“写分配完成”交互过程，将跨站点写 I/O 交互次数减少一半。如图 3-16 所示。

图3-16 传输协议优化



3.1.3.3 地域优化访问

双活数据业务场景，两站点的距离远近，是影响 I/O 访问性能的关键因素。HyperMetro 特性通过与华为 OceanStor UltraPath 多路径配合，根据双活站点部署距离，提供了两种 I/O 访问策略供用户选择。

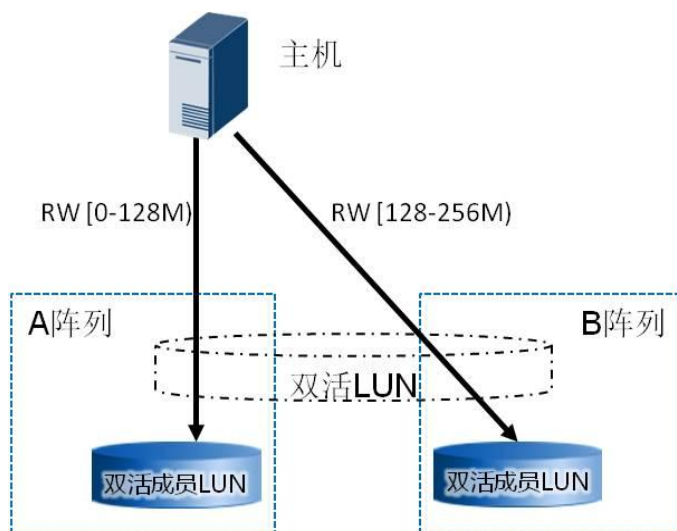
- 负载均衡模式
- 优选阵列模式

负载均衡模式

该模式下实现了 I/O 的跨阵列负载均衡，即 I/O 以分片的方式在两个阵列上下发。分片大小可配，例如分片大小为 128M，即起始地址为 0-128M 的 I/O 在 A 阵列下发，128M-256M 在 B 阵列下发，以此类推。

负载均衡模式主要应用于双活业务部署在同一数据中心的场景。在该场景下，主机业务访问两套双活存储设备的性能几乎相同，为最大化利用两套存储设备的资源，将主机 I/O 按分片方式下发到两套阵列上。

图3-17 负载均衡访问



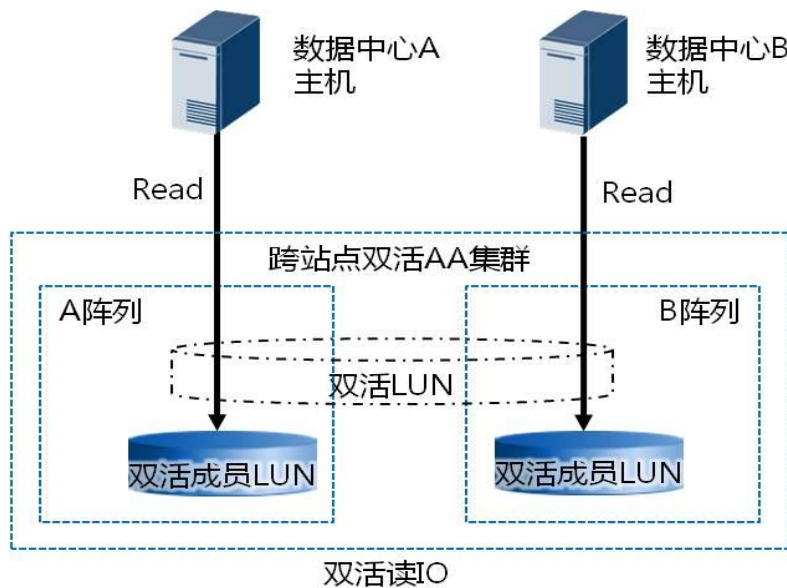
优选阵列模式

该模式下，由用户在 OceanStor UltraPath 上指定优选访问阵列，主机业务访问时，I/O 只会在用户设置的优选阵列路径上进行负载均衡下发，不产生跨阵列的 I/O 访问。只有当优选阵列出现故障时，才切换到非优选阵列下发 I/O。

优选阵列模式主要应用于双活业务部署在距离较远的双数据中心场景。在该场景下，双活数据中心的跨站点访问的代价较高，假如两个数据中心的链路距离为 100km，一次往返传输通常需要消耗约 1.3ms 时间。优选阵列模式可以减少跨站点交互次数，从而提升 I/O 性能。

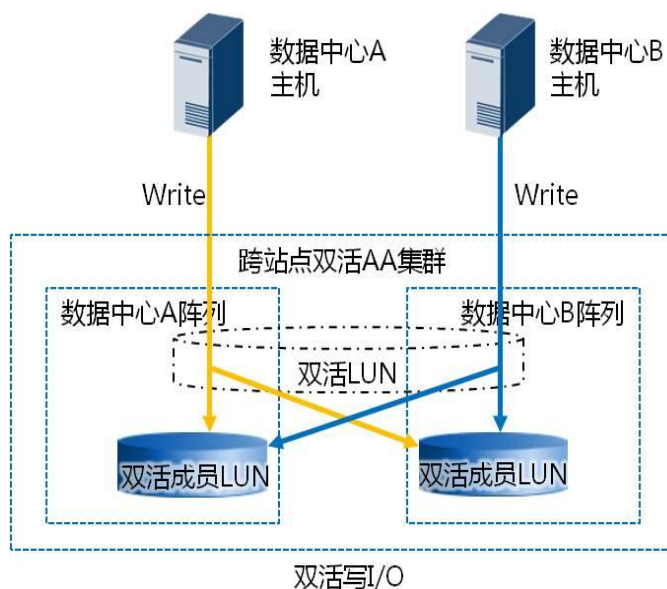
针对数据读场景，双活数据中心的业务主机只需要读本数据中心对应的双活存储阵列即可，避免主机跨数据中心读取数据，提升整体访问性能。

图3-18 优选阵列模式数据读



针对数据写场景，业务主机直接写本数据中心对应的双活存储阵列，避免主机跨数据中心转发数据，充分利用 HyperMetro AA 双活能力，AA 集群的每个控制器都能够接收写 I/O，由本地控制器处理本地主机的写 I/O 请求，减少跨数据中心的转发次数，提升方案整体性能。数据写 I/O 过程如图 3-19 所示。

图3-19 优选阵列模式数据写



3.1.4 高可扩展性

HyperMetro 双活特性通过与 OceanStor V3 融合存储系统的其它 Smart 系列和 Hyper 系列特性组合，可形成多种数据保护和灾备解决方案。

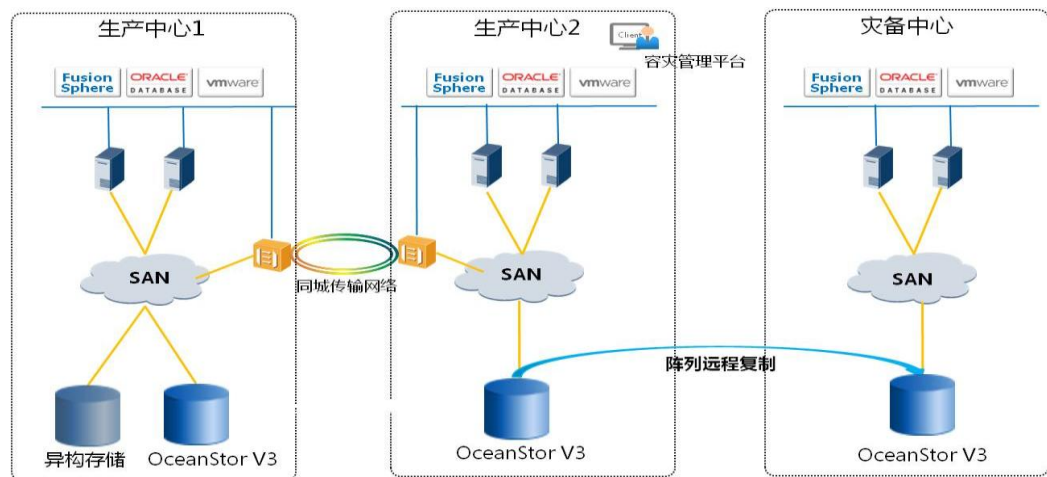
3.1.4.1 两地三中心扩展

近年来，大范围自然灾害时有发生，3DC（两地三中心）容灾解决方案越来越受到业界重视和认可。

所谓 3DC 即两地三中心，即一份数据有 3 份备份（包括自己）且分布在三个不同的地理位置即称之为三数据中心，通常是指生产中心，加上同城灾备中心以及异地灾备中心。

HyperMetro 特性与 OceanStor 统一存储系统的 HyperReplication 特性组合，并结合 BCManager 专业容灾管理软件实现 3DC 容灾。当生产中心发生灾难，可在同城灾备中心进行接管业务，并保持与异地灾备中心的容灾关系。若生产中心和同城灾备中心均发生灾难，可在异地灾备中心对远程复制进行主从切换，并拉起业务。

图3-20 3DC 容灾方案



相对于传统的同步复制+异步复制的 3DC 方案，双活+异步复制的方案具有更好的资源利用率和故障切换速度。双活数据中心实现同城容灾时，可将同一关键业务负载均衡到双数据中心，并且在单数据中心发生故障，业务零中断，数据零丢失。在部署层面，双活数据中心支持平滑扩展为两地三中心，可先期实现同城双活，待异地数据中心建设完成后，再添加异步复制，实现应用异地保护。

3DC 方案中，BCManager 能够提供简化管理的容灾拓扑展示与端到端监控功能，直观清晰的展示保护方案的状态与变化，实时监控相关设备部件，实现业务灾难切换前就识别问题与故障并协助用户排除，规避影响业务和增加成本的容灾切换发生。

3.1.4.2 异构双活

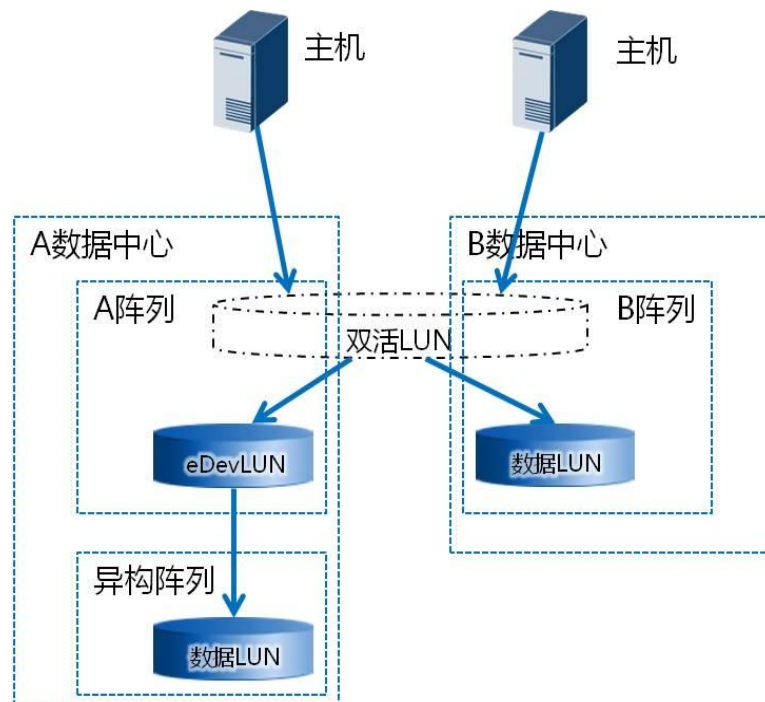
HyperMetro 与 OceanStor V3 融合存储系统的 SmartVirtualization 特性结合，利用 SmartVirtualization 特性的虚拟化功能，实现高可靠的存储资源整合和优化，提高资源利用率。

异构虚拟化就是把异构阵列映射到本端阵列（华为阵列）的 LUN，作为本端阵列能提供存储空间的逻辑盘，再在该逻辑盘上创建 eDevLun，逻辑盘为 eDevLun 提供了全部的数据存储空间，eDevLun 的元数据卷的存储空间由华为存储提供。异构虚拟化可保

证外部 LUN 数据完整性不被破坏，eDevLun 与原异构 LUN 具有不同的全球唯一名称 WWN。

由于 eDevLun 与本地 LUN 具有几乎相同的 LUN 属性，所以，异构虚拟化接管之后的 LUN 可以通过 HyperMetro 技术为异构 LUN 提供双活功能。

图3-21 异构阵列保护



3.1.4.3 本地保护

快照保护

当遭遇病毒攻击或人为误操作等异常时，数据中心的数据可能被破坏。采用虚拟快照技术，在操作前对现有数据卷激活快照，实现本地数据保护。

虚拟快照采用写前拷贝（COW）技术，仅将变化的数据拷贝至快照卷，系统资源占用少。当原卷数据被错误修改或删除，需要进行恢复时，采用快照对原卷进行回滚，实现数据恢复，并且，快照卷可分别映射给主机进行数据测试和挖掘，不影响生产业务。结合华为容灾管理软件，激活快照前触发数据库完全检查点，待数据写盘完成后，激活快照，保证快照数据完全满足数据库的一致性要求，快速拉起数据库。

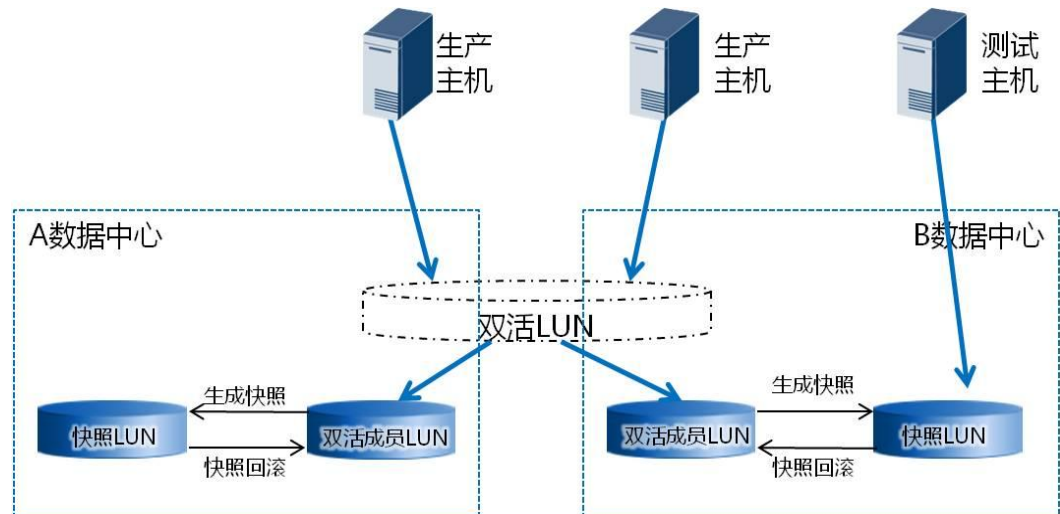


注意

如果将当前的双活数据进行快照回滚，数据会被覆盖，无法返回当前的状态。如果可能需要返回当前状态，需手动将当前状态打个快照。

HyperMetro 与 HyperSnap 结合，可分别在双活两则阵列上对双活成员 LUN 配置快照保护。

图3-22 有快照保护的双活

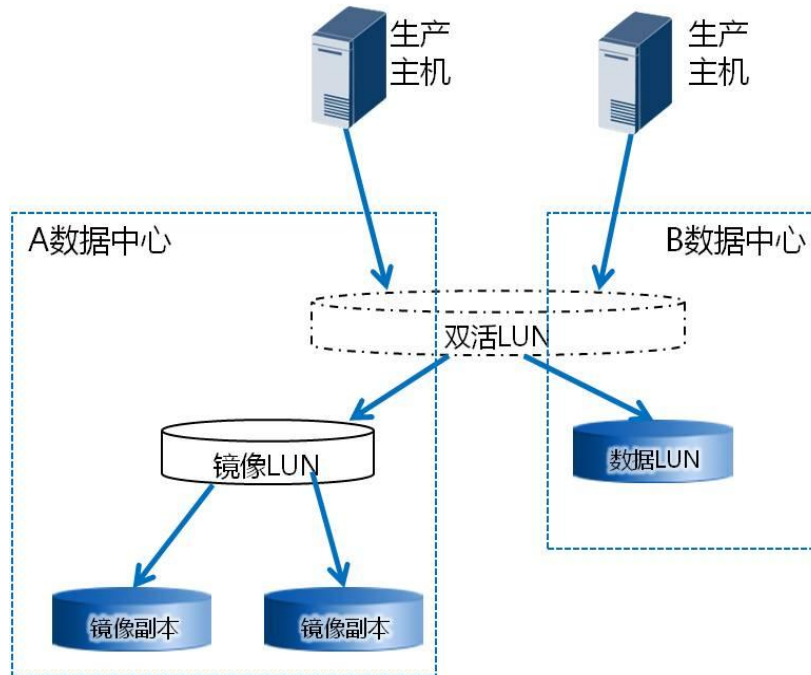


卷镜像保护

卷镜像（HyperMirror）是一种本地数据持续保护技术，可以为一个 LUN 创建两个镜像物理副本，提供持续的冗余保护而主机侧不受任何影响。当某个镜像副本故障时，存储系统可以利用卷镜像提供的另一个镜像副本，保证主机业务继续正常工作，并且可以保障数据不会丢失。华为 OceanStor V3 存储系统支持与第三方存储系统的 LUN 创建卷镜像，提高第三方存储系统 LUN 的可靠性。

HyperMetro 与卷镜像结合，可在镜像 LUN 基础上，再创建双活 Pair。这种组合可充分利用本地数据保护和远程数据保护的特点，提供更高级别的业务连续性保护。逻辑图如图 3-23 所示。

图3-23 有卷镜像保护的双活



3.2 计算层双活

通过物理机部署跨数据中心的集群，可以保证在各种故障场景下，快速完成业务切换，提供零中断零丢失的业务访问。

常见的虚拟化集群技术，有以下最突出的特点：

- HA 重启恢复虚拟机，业务有短暂中断
 - 发生宿主机故障时，运行其上的虚拟机在其它宿主机上自动重启恢复。该虚拟机业务会有短暂中断，虚拟机内存数据会丢失。
- 虚拟机集群资源利用率提升数十倍
 - 一台物理宿主机上能运行几台甚至几十台虚拟机，提供极高的资源利用率。

由此可见，解决虚拟机 HA 时业务中断是需要解决的首要问题，否则双活零业务中断无法满足。前述我们使用了负载均衡设备，可以实现在双活数据中心实现业务负载均衡，此时虚拟机上部署服务时，只需要在两个数据中心部署同样的服务，即可满足要求。因为当宿主机故障时，另一个数据中心的虚拟机上的服务可以实时接管业务。

当前，常见的几个虚拟化平台（VMware vSphere，FusionSphere）已经商用多年，稳定性和可靠性已经得到了验证。计算资源虚拟化的建议配置如下：

- 跨数据中心配置虚拟化集群，将计算资源虚拟化后，其上部署虚拟机。
- 配置 HA，使虚拟机受 HA 保护，故障时能自动恢复。
- 配置 DRS，使虚拟机按业务要求更好的分布在不同宿主机上。

- 网络大二层互通，使虚拟机能无障碍跨数据中心在线迁移，提供虚拟化平台更好的维护性，日常维护时业务不受影响。
- 双活存储平台提供的共享存储空间，映射给虚拟化集群的所有宿主机，以便虚拟机具备足够的灵活性。

计算层虚拟化双活改造后，虚拟机能更好的基于原计算资源负载均衡，资源利用率和运行效率能得到极大提高，新业务部署更为灵活和简单，所有虚拟机具备良好的高可靠性、在线迁移特性和易维护性。

虚拟机部署方式如下：

- 如果是 B/S 应用：Web 和 App 采用虚拟机部署，虚拟机不部署集群。SLB 可以检测服务器故障，将业务分发到正常运行的服务器上。
- 如果是 C/S 应用：App 如果采用虚拟机部署，虚拟机部署跨 DC 集群。

3.3 应用层双活

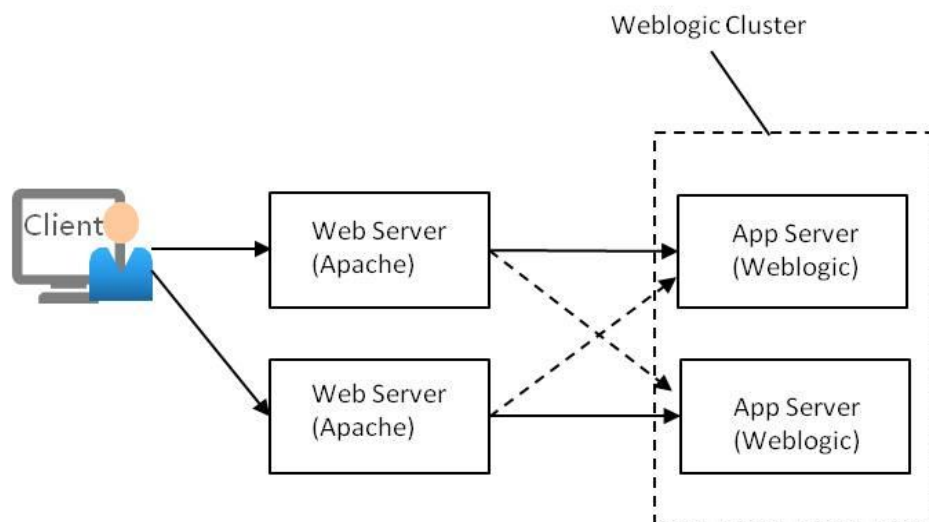
3.3.1 B/S 应用双活

工作原理

一个从浏览器发送的 HTTP 请求经过 Web 服务器（例如 Apache）之后会被重新定向给应用服务器（例如 Weblogic）。这个重定向的操作就是由 Web 服务器来完成的。

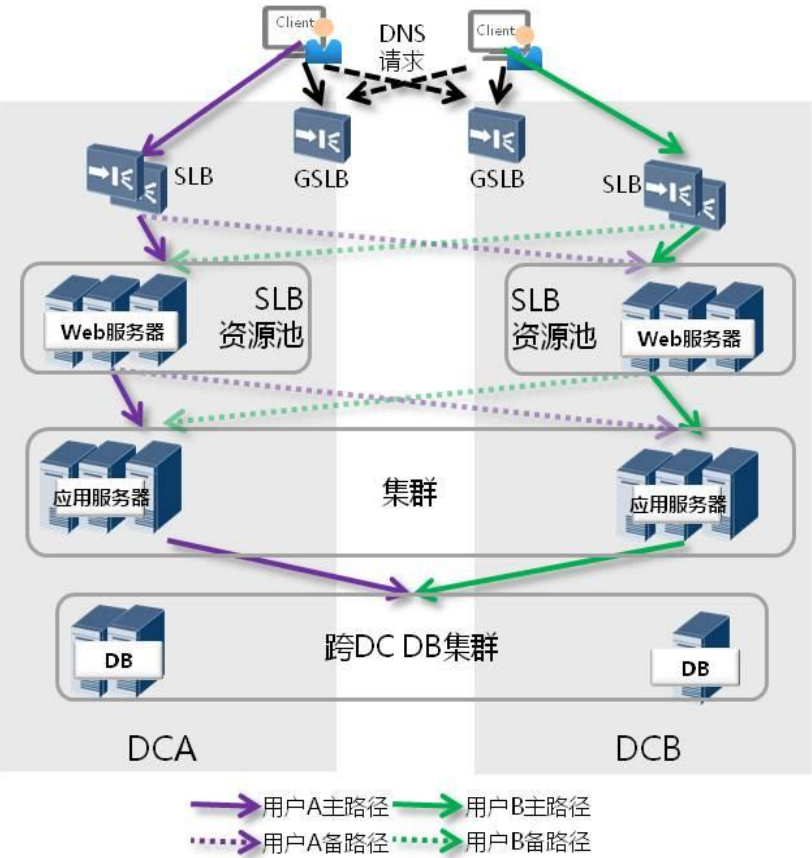
一般的，一台 Web 服务器对应一个应用服务器集群，从而达到负载均衡效果。

图3-24 Web 应用服务器集群部署



网络拓扑

图3-25 B/S 请求转发拓扑图



一个站点部署多台 Web 服务器，不组成集群，在 SLB（F5 LTM）上创建两个资源池，将一个站点内部的所有 Web 服务器组成一个资源池。双活数据中心方案分别创建 DC1 资源池和 DC2 资源池；

一个站点部署多台应用服务器，跨站点将同类应用服务器组成 AA 集群。

两个数据中心的应用服务器集群均连接跨 DC 的数据库集群。

HTTP Session 持久化管理

一个 HTTP Session 是指一个用户在同一个浏览器上对服务器发出的一系列请求。HTTP Session 使得运行在 Web 容器上的应用能够跟踪记录每个用户的一系列操作。

Session 持久化管理是指客户的所有后续请求都会被分发到同一台应用服务器上。通过 Session 持久化管理，系统的性能会得到提高，因为各个应用服务器不需要重新创建和维护 Session 信息，而且这样也避免了前面 Session 丢失的情况。

应用服务器集群跨数据中心创建，使用内存同步完成 Session 持久化管理，即使跨数据中心访问，也不会 Session 丢失。

负载均衡

B/S 应用的双活负载均衡，流程如下：

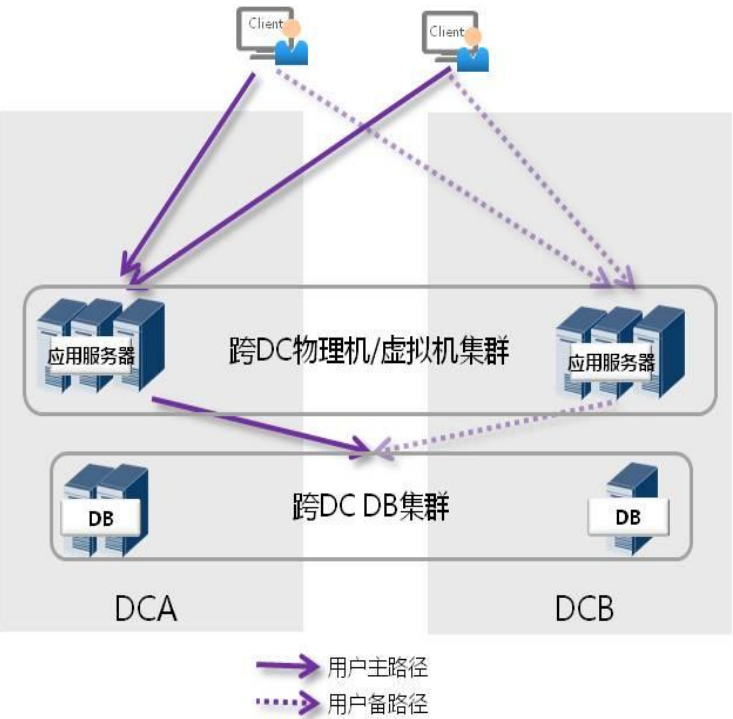
1. SLB 接受到新的 HTTP 请求，根据负载均衡算法分发到 Web 服务器；
2. Web 服务器接受到 SLB 分发下来的新的 HTTP 请求，然后在资源池中选择应用服务器节点进行转发。
 - a. HTTP Session 持久化检查，检查请求的 Cookie 或 URL 中的 jSession id 参数。如果匹配，那么插件就会将请求转给这台应用服务器。
 - b. 如果不能匹配。插件会根据既定规则寻找合适的应用服务器。
3. Web 服务器通过 TCP/IP 协议将浏览器请求转发给应用服务器。如果请求转发成功，应用服务器返回成功的 TCP/IP ACK 回复。如果超时无回复，Web 服务器会返回一个 500 错误。
4. 返回应用服务器处理请求结果。发送请求后，Web 服务器会进入等待状态。当应用服务器返回对应的结果后，Web 服务器结束其等待状态。如果请求处理失败，Web 服务器会把应用服务器 标为不可用的状态，并将请求转给另一台应用服务器。

3.3.2 C/S 应用双活

C/S 业务一般对外提供 IP 访问，例如使用一款 Client 软件，使用 IP 地址、用户名和密码登录某款应用。

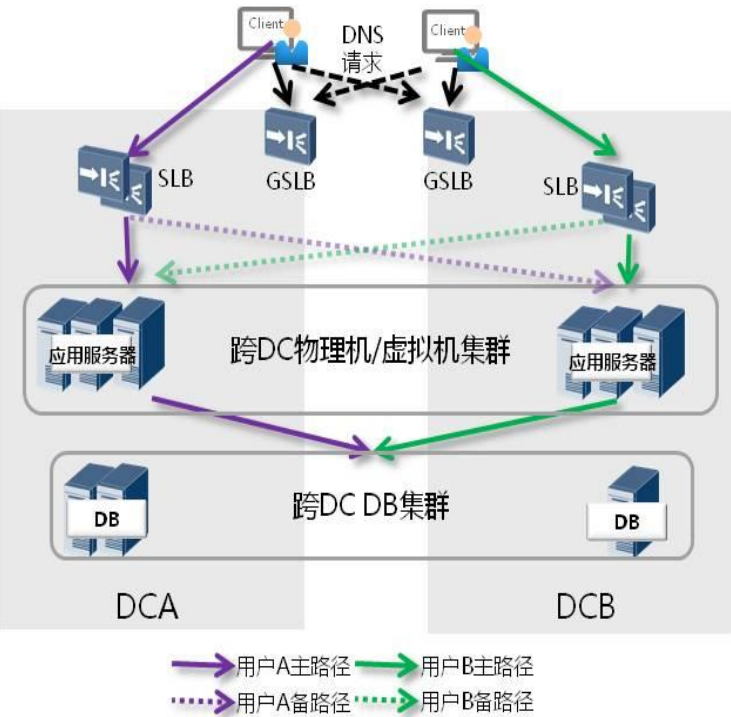
如果 C/S 应用不支持分布式部署。主路由只能在一个数据中心发布，所以应用只能在单数据中心运行，但支持故障自动切换至备站点。将不同应用分布到 2 个数据中心运行。

图3-26 C/S 应用双活原理（IP 访问，C/S 应用不支持分布式部署）



如果 C/S 应用支持分布式部署，在两数据中心运行（对外使用不同 IP 地址）。手动配置客户端对应的服务器端 IP，将不同客户负载均衡到两个 DC；

图3-27 C/S 应用双活原理（IP 访问，C/S 应用支持分布式部署）

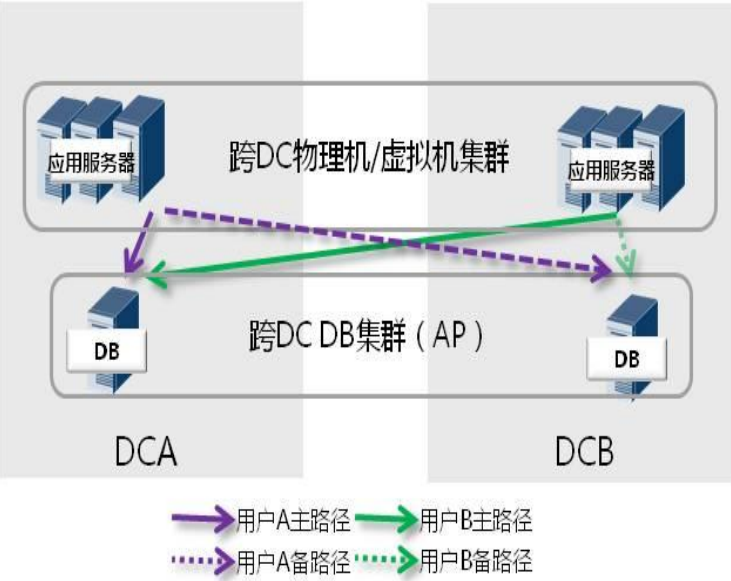


3.3.3 数据库双活

数据库层双活通过数据库集群实现，在业界主要分为两种方式：A/S（Active-Standby）集群与 A/A（Active-Active）集群。

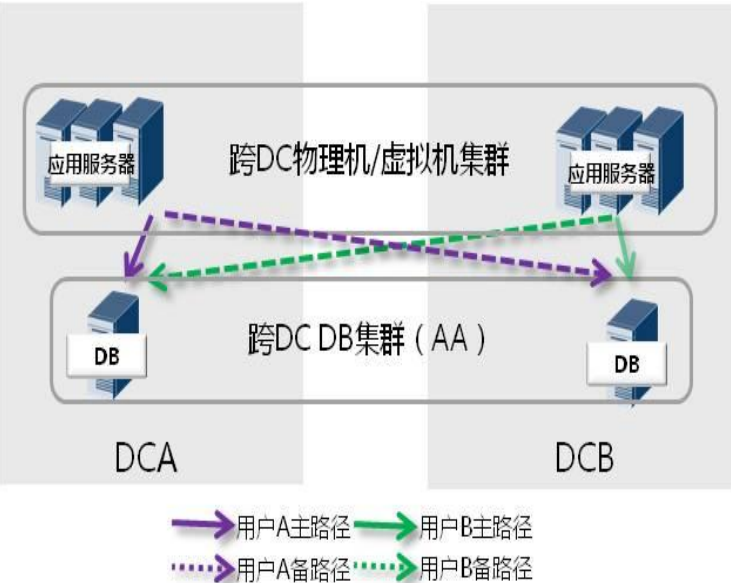
常见的 Active-Standby 集群系统有：IBM PowerHA、HP ServiceGuard、Microsoft WSFC、Veritas Cluster Server 等。在主节点故障后，集群备节点将自动重新启动应用系统，无需管理员干预，即所谓的“故障切换(Failover)”。管理员需要提前在所有集群主备节点上部署集群软件，由集群软件控制文件系统的挂载，应用系统服务的启动，以及公网 IP 地址的配置。主备集群架构如图 3-28 所示：

图3-28 A/S 集群



Active-Active 集群的多个集群节点可以同时提供同一项服务。采用 Active-Active 集群，可以做到故障的无缝切换，而且可以提升应用系统整体性能。当前典型、应用最广泛使用的是 Active-Active 集群，例如 Oracle RAC(Real Application Cluster)集群系统。Active-Active 集群架构如图 3-29 所示：

图3-29 A/A 集群



3.3.3.1 Oracle RAC 双活

Oracle RAC 以共享存储为基础，通过共享的存储资源，实现各节点对数据文件、重做日志文件、控制文件和参数文件的并行访问，且在单个节点出现故障时，业务能自动切换到正常节点，从而保证数据库系统正常可用。

存储双活提供的双活 LUN 作为共享卷，实现跨数据中心的 Oracle Extended RAC 集群构建。

Oracle Extended RAC 集群配合 Oracle 监听器技术，可实现客户端在数据中心间业务双活访问和负载均衡；配合 Oracle 透明应用程序故障转移（TAF）技术，当服务器或单数据中心故障时，使客户端能够在新的连接中继续工作，防止业务中断。

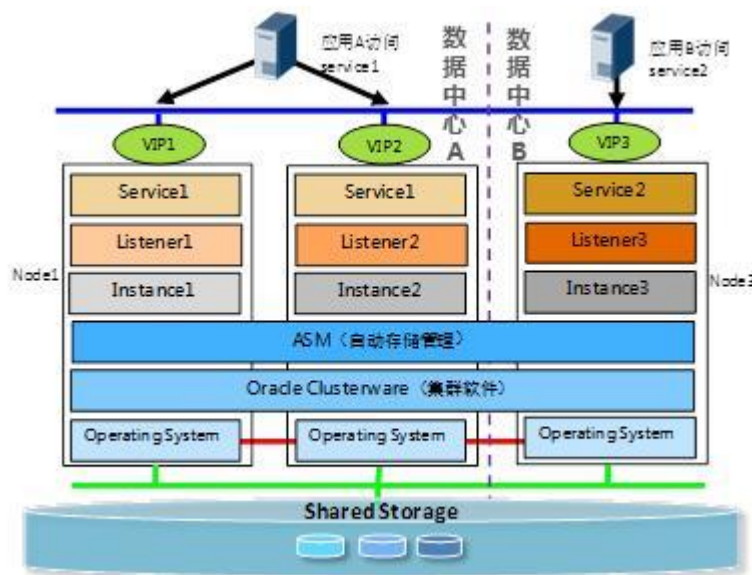
在出现中间网络心跳链路故障时，Oracle RAC 会采用如下的原则进行仲裁：

- 拥有最多节点数目的子集群（Sub-cluster with largest number of Nodes）获胜。
- 若子集群内数目相等则拥有最低节点号的子集群(Sub-cluster with lowest node number)获胜。

Oracle RAC 建议采用“2+1”的部署方式，数据中心 A 部署 2 台服务器，数据中心 B 部署 1 台服务器。保证在出现心跳链路故障时数据中心 A 的实例优先存活。若节点数一致，则建议部署时将节点号小的服务器部署在数据中心 A。

Oracle Extended Distance Cluster 节点部署如图 3-30 所示：

图3-30 RAC 网络拓扑图



为了避免跨数据中心进行数据的交互，建议在 Oracle RAC 层创建不同的 Service，实现业务分离。

通过 Oracle RAC 透明应用程序故障切换（TAF）的 PREFERRED 功能设置应用只访问本地实例；同时设置远端数据中心的实例为 AVAILABLE，只有本地实例都故障才切换到远端实例。

服务设置方式如表 3-5 所示：

表3-5 服务设置方式

	服务名	INSTANCE1	INSTANCE2	INSTANCE3
服务配置策略	SERVICE1	PREFERRED	PREFERRED	AVAILABLE
	SERVICE2	AVAILABLE	AVAILABLE	PREFERRED

3.3.3.2 DB2 双活

IBM DB2 数据库实现集群的方式主要有两种：第一种是通过 DB2 数据库的集群实现，如 pureScale 实现；另外一种是通过操作系统的集群实现，如 IBM PowerHA。

本方案采用 IBM PowerHA 集群实现 DB2 数据库的跨数据中心双活部署。

PowerHA 为主备集群模式，目标是消除单点故障，当一台服务器发生故障时，另一台服务器自动接管。PowerHA 集群技术通过提供冗余实现故障转移保护，同时通过并发/并行访问支持水平扩展。

PowerHA 采用“Shared Everything”的存储原则。在双活架构下，存储层需要提供可供多个数据库节点同时进行读写访问的共享卷。则 DB2 数据库的数据文件，日志文件都必须存放在共享磁盘上。DB2 支持以下的存储方式：

- 文件系统
- Raw Device（裸设备）

建议采用文件系统的方式为数据库提供存储空间。具有 Direct I/O 与 Concurrent I/O 特性的文件系统能提供与裸设备相似的性能，同时还简化了管理。

建议每个 LUN 上创建一个文件系统供数据库使用。创建文件系统的个数建议如下：

- 创建 1 个日志文件系统
- 创建多个数据文件系统

存储需要支持的预留方式：

- 对于 NON-Concurrent VG，需要支持 SCSI-3 Persistent Group Reservations (PGR)
- 对于 Concurrent vg 或 Enhanced Concurrent VG，则无需支持 SCSI-3 Persistent Group Reservations (PGR)

PowerHA 心跳配置建议：

IP 心跳与非 IP 心跳共存以达到冗余。采用 TCP/IP 网络与 Disk Heartbeating 作为 PowerHA 的心跳。在存储上创建一个 Enhanced Concurrent VG 用于 PowerHA 的 Disk Heartbeating。

PowerHA 有三种资源组接管方式：

- Cascading（级联）：Cascading 级联资源组定义了一个可以控制该资源组的所有节点的列表，以及每个节点接管资源组的优先级。当集群启动时，优先级最高的节

点拥有资源组，当该节点发生失败时，资源组被优先级较高的节点接管，当失败节点重新加入集群后，它将从接管节点夺回资源组的控制权。

- **Rotating(循环)**：一个 Rotating 资源组和一组节点相关联。资源组在所有被定义的节点之间轮转。当集群启动时，优先级最高的节点拥有资源，当控制资源的节点失败时，节点列表中对资源组具有较高优先级的节点将接管资源组。当失败节点重聚到集群后，不再取回资源组，将成为备用节点。
- **Concurrent（并发）**：Concurrent 资源组由多个节点共享，集群启动时，所有并发访问资源组的节点都获得该资源组，节点之间没有优先级的区别。当某节点失败，资源组只在该节点 offline，当故障节点重新回到集群后，资源组在该节点 online。

其中 Concurrent 资源组主要应用场景是配合 Oracle RAC 实现 A-A 的集群架构，在本方案中资源组不采用 Concurrent 方式。

3.3.3.3 SQL Server 双活

SQLServer 数据库实现集群的方式主要是通过操作系统的集群实现，如 Windows Server Failover Cluster（WSFC）。

Windows Server 故障转移集群提供了各种基础结构功能来支持所承载的服务器应用程序（如 Microsoft SQL Server）的高可用性和灾难恢复方案。如果一个集群节点或服务失败，则该节点上承载的服务可在一个称为“故障转移”的过程中自动或手动转移到另一个可用节点。

WSFC 集群中的节点协同工作，共同提供这些类型的功能：

- 分布式元数据和通知。集群中的每个节点上维护着 WSFC 服务和承载的应用程序元数据。除了承载的应用程序设置之外，此元数据还包括 WSFC 配置和状态。对一个节点的元数据或状态进行的更改会自动传播到集群中的其他节点。
- 资源管理。集群中的各节点可能提供物理资源，如直接连接存储、网络接口和对共享磁盘的访问。承载的应用程序将其本身注册为集群资源，并可配置启动和运行状况对于其他资源的依赖关系。
- 运行状况监视。节点间和主节点运行状况检测是通过结合使用信号样式的网络通信和资源监视来实现的。集群的总体运行状况是由集群中节点仲裁的投票决定。
- 故障转移协调。每个资源都配置为由主节点承载，并且每个资源均可自动或手动转移到一个或多个辅助节点。基于运行状况的故障转移策略控制节点之间资源所有权的自动转移。在发生故障转移时通知节点和承载的应用程序，以便其做出适当的响应。

WSFC 采用“shared everything”的存储原则，则 SQLServer 数据库的数据文件，日志文件都必须存放在共享磁盘上。SQLServer 建议的存储方式为文件系统，建议采用 NTFS 文件系统。

预留方式要求：需要支持 SCSI-3 Persistent Group Reservations (PGR)。

WSFC 仲裁模式：

推荐 2 节点部署，相对于更多节点能提供较高的资源利用率，在偶数个节点（如 2 节点）的集群架构下，采用“节点和磁盘的大多数”的仲裁模式。

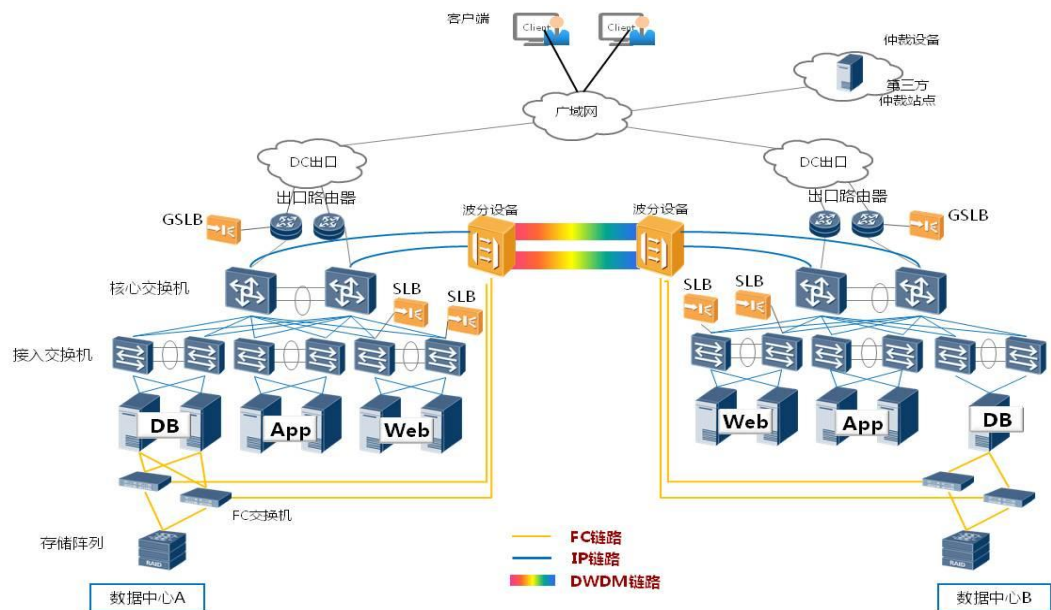
3.4 网络层双活

3.4.1 网络架构

总体网络架构涉及三个站点的部署规划，数据中心 A、数据中心 B、第三方仲裁站点。

- 建议两个数据中心距离不超过 100km，并具备裸光纤资源。
- 第三方仲裁站点同时连接至数据中心 A 和数据中心 B，无距离限制要求。

图3-31 总体网络架构图



说明

- 核心交换机中，公有网络应当与私有网络及仲裁网络采用逻辑隔离方式。
- 波分设备中，应采用不同波分通道承载这几类网络（尤其是大规模组网场景），最低要求是公有网络与私有网络分开，达到物理隔离的效果。根据带宽的利用率，公有网络可以与其他业务流量（如其他服务器间流量）共用波分通道。
- 数据库一般用于向上层应用服务器提供业务访问，不直接向广域网用户提供访问。

3.4.2 跨数据中心网络

为保障方案的可靠性，本方案采用数据传输链路与心跳链路分离设计的原则。通过 VLAN 或 VRF 隔离端到端流量，同时分配独立的物理互联链路，做到业务流量与集群心跳分离流量，互不影响。

涉及跨数据中心传输的业务有：

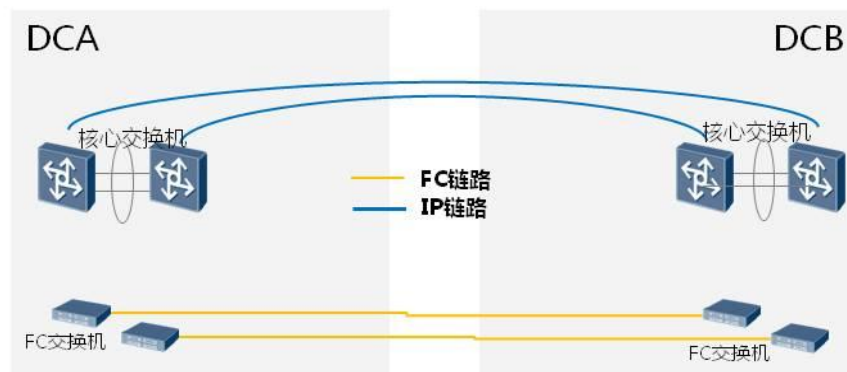
- 采用 FC 链路实现同城双数据中心间的数据实时同步；

- 采用二层以太网实现双数据中心间主机应用集群的心跳、同步互联链路通信。

如果两个数据中心间链路距离 $\leq 25\text{km}$ ，且裸光纤数量 ≥ 4 对。

- 4 台核心交换机建议 10GE 级联，需要 2 对裸光纤；
- 4 台 FC 交换机建议 FC 采用一对一级联，需要 2 对裸光纤。

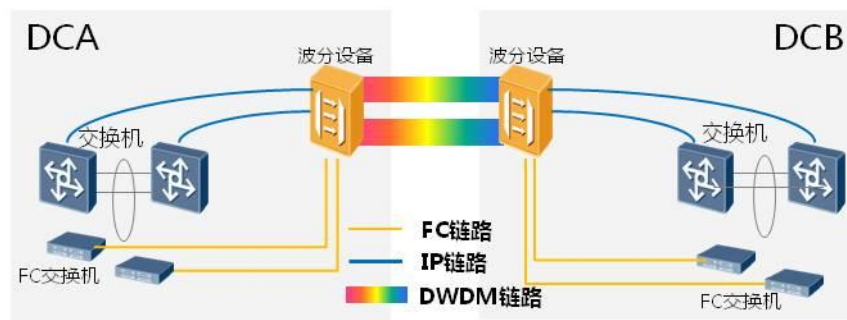
图3-32 链路距离 $\leq 25\text{km}$ & 裸光纤数量 ≥ 4 对



如果两个数据中心间链路距离 $> 25\text{km}$ 或裸光纤数量小于 4 对，建议使用 OTN 波分设备来构建两数据中心的同城网络。

以太网交换机和 FC 交换机同时连接到 OTN 波分设备，两个数据中心的 OTN 波分设备直接级联，需要 2 对裸光纤；

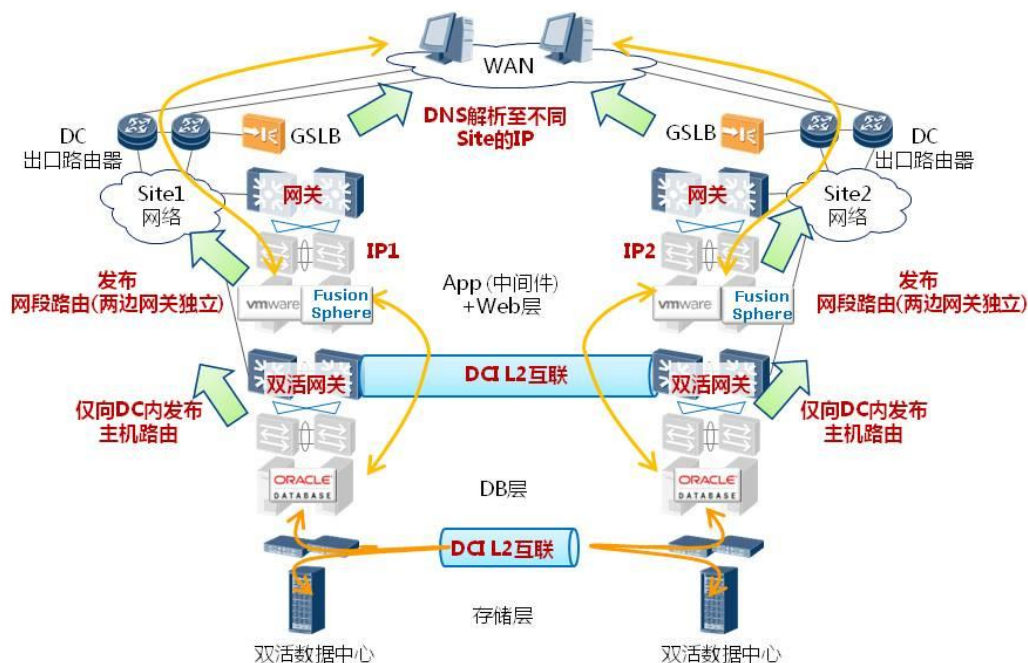
图3-33 链路距离 $> 25\text{km}$ 或裸光纤数量 < 4 对



3.4.3 业务访问网络架构

3.4.3.1 B/S 应用网络架构

图3-34 B/S 应用网络架构图



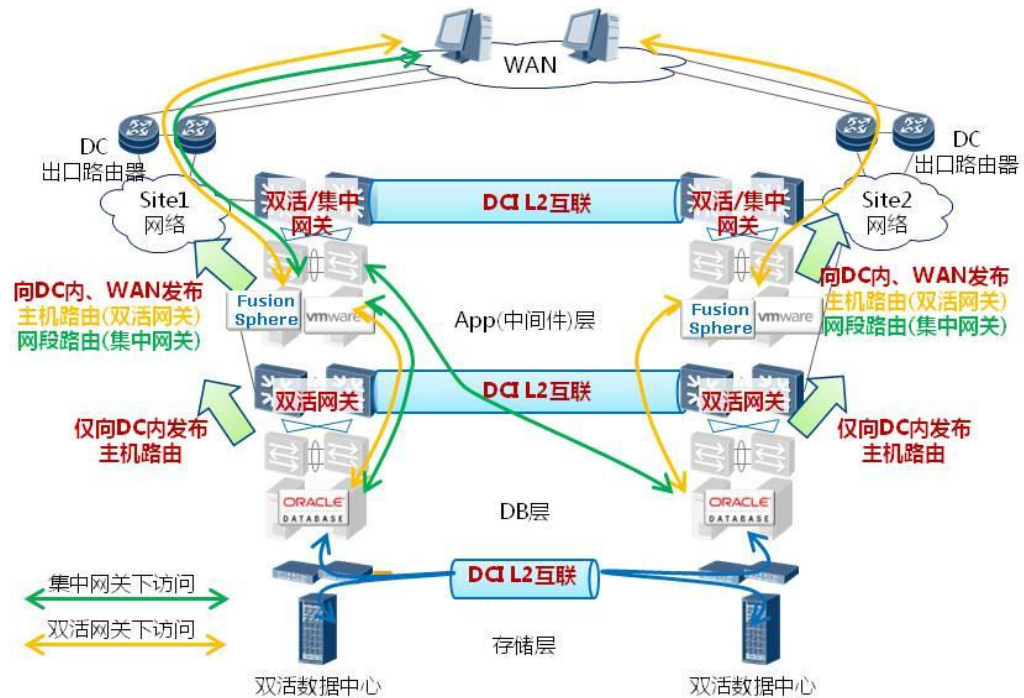
B/S 应用多为 Web 应用，对外提供域名访问。采用 Web-App-DB 三层结构，Web/App 采用虚拟机部署，虚拟机只在 DC 内部署集群，DB 采用物理机部署，跨 DC 部署 Oracle RAC 集群。Web/App 层提供双活访问，DB 仅对 App/Web 提供服务。

Web/App 层业务访问网络设计：采用 3 层物理网络架构的，在汇聚交换机配置 Web/App 层网关；采用 2 层物理网络架构的，在核心交换机配置 Web/App 层网关；两站点的网关相互独立，不同网段，对外发布网段路由。

DB 层业务访问网络设计：采用 3 层物理网络架构的，在汇聚交换机配置 DB 层网关；采用 2 层物理网络架构的，在核心交换机配置 DB 层网关；两站点之间需要二层互联，同一网段，每个站点部署双活网关，向 DC 内发布数据库的主机路由，本主机路由不会向广域网发布。

3.4.3.2 C/S 应用网络架构

图3-35 C/S 应用网络架构图



C/S 应用多为中间件应用，对外提供 IP 访问，无需 GSLB。采用 App-DB 两层结构，App 采用虚拟机部署、虚拟机采用集群部署方式，DB 采用物理机部署。App 层在单个数据中心运行，DB 仅对 App 提供服务。

DB 层业务访问网络设计：采用 3 层物理网络架构的，在汇聚交换机配置 DB 层网关；采用 2 层物理网络架构的，在核心交换机配置 DB 层网关；两站点之间需要二层互联，同一网段，每个站点部署双活网关，向 DC 内发布数据库的主机路由，本主机路由不会向广域网发布。

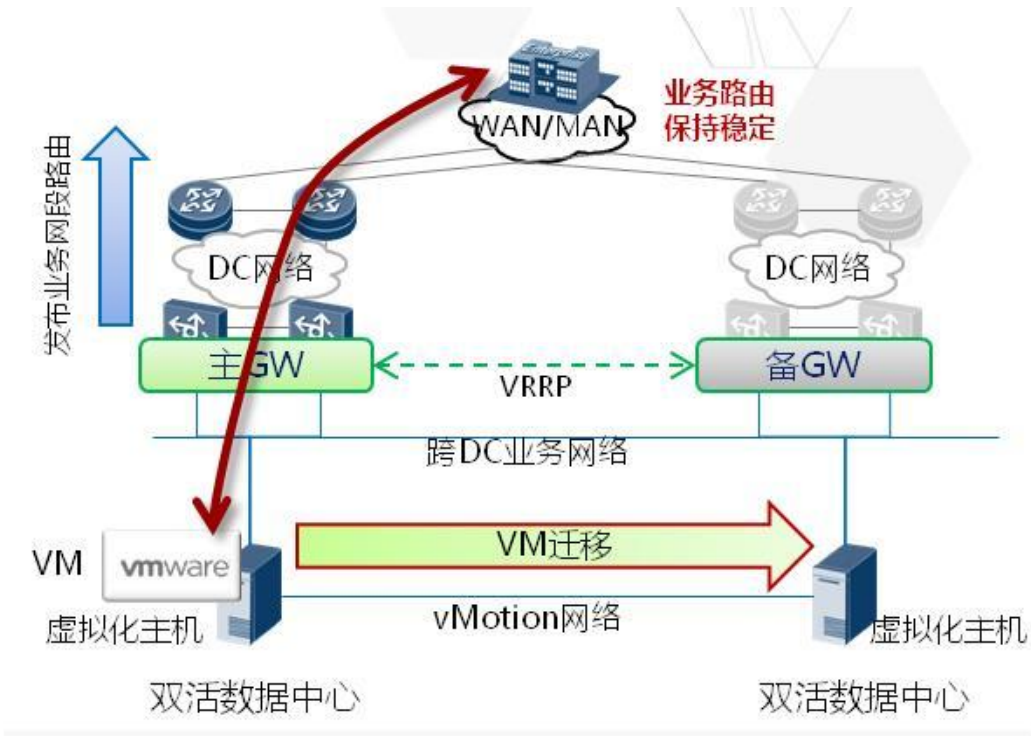
App 层业务访问网络设计：采用 3 层物理网络架构的，在汇聚交换机配置 App 层网关；采用 2 层物理网络架构的，在核心交换机配置 App 层网关；两站点之间需要二层互联，同一网段。网关设计有两种方式：

1. 集中网关，App 所在的 DC 的网关配置主 VRRP 网关，发布主路由，另一侧站点网关配置为备网关，发布备路由。如果 VM 跨数据中心迁移，主备网关不会跟随切换，存在跨数据中心访问的情况。
2. 双活网关，每个站点部署双活网关，网关动态感知 App 主机的位置，根据 App 主机所在的 DC 位置，由所在的双活网关向 DC 内、广域网发布主机路由，优化访问路径。

集中网关

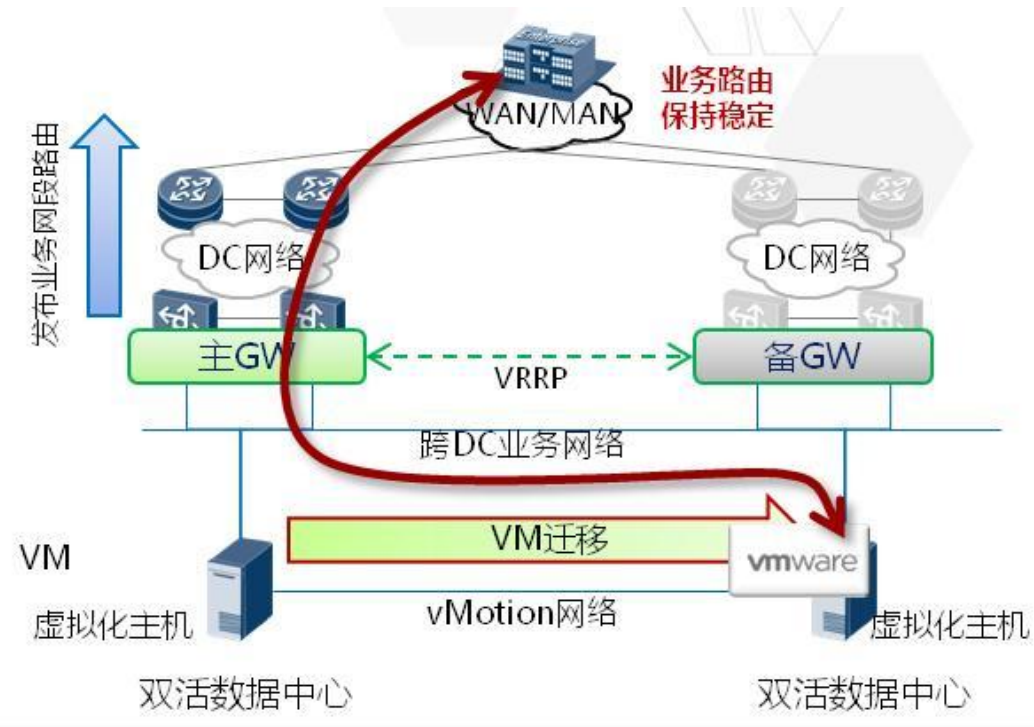
C/S 应用通常集中在单数据中心运行，在单数据中心发布网段路由，具备跨 DC 迁移的能力。

图3-36 C/S 应用路径-迁移前



虚拟机迁移后，集中网关不会跟随切换，存在跨数据中心访问的情况。但对外路由保持稳定，访问路径清晰，排障、运维较简单。

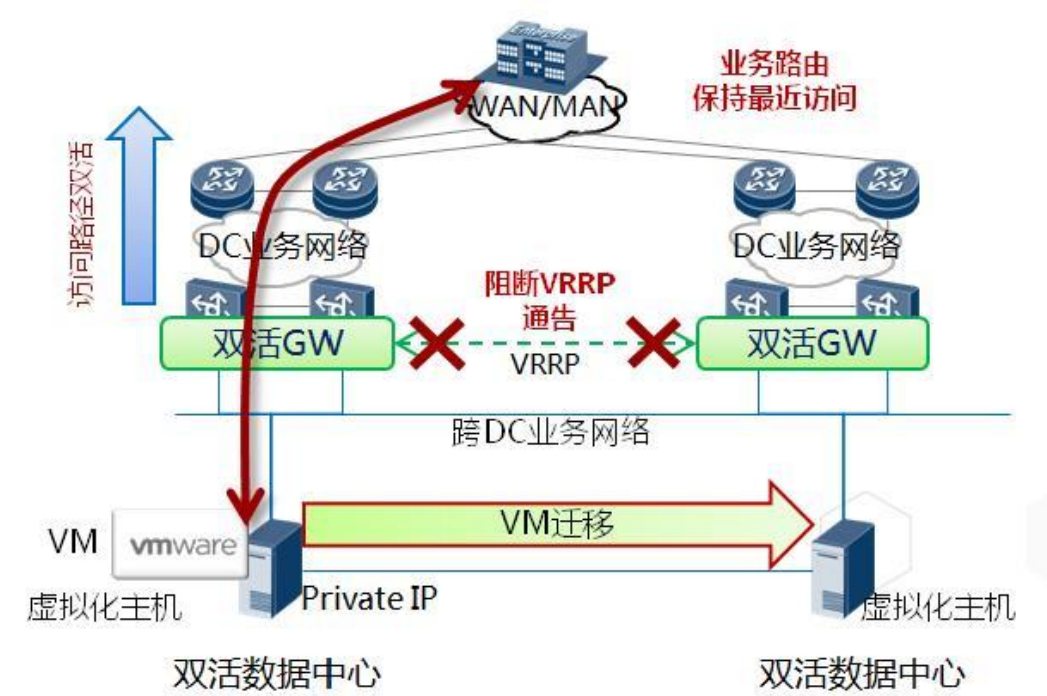
图3-37 C/S 应用路径-迁移后



双活网关

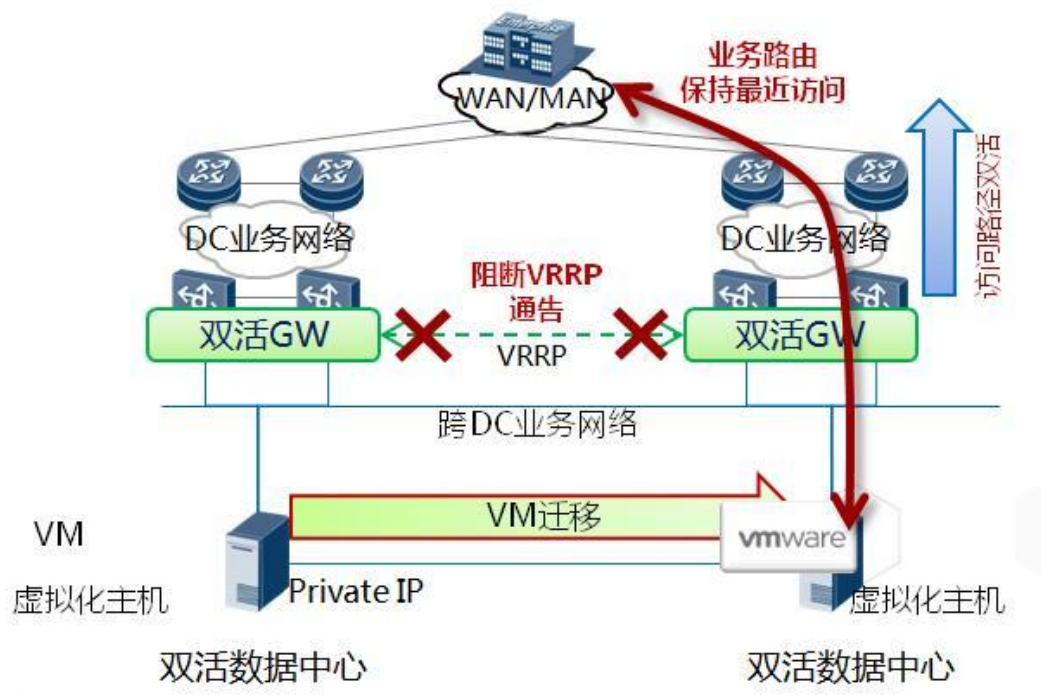
C/S 应用可随意在数据中心间迁移调度，根据 VM 所在位置，从数据中心就近的网关发布主机路由。

图3-38 C/S 应用路径优化-迁移前



虚拟机迁移后，发布 ARP 广播，双活网关发送 ARP 单播请求，探测、判断虚拟机所在位置。探测成功后，在新的网关位置对外发布虚拟机的主机路由，原有主机路由超时撤销。

图3-39 C/S 应用路径优化-迁移后



3.4.4 二层互联

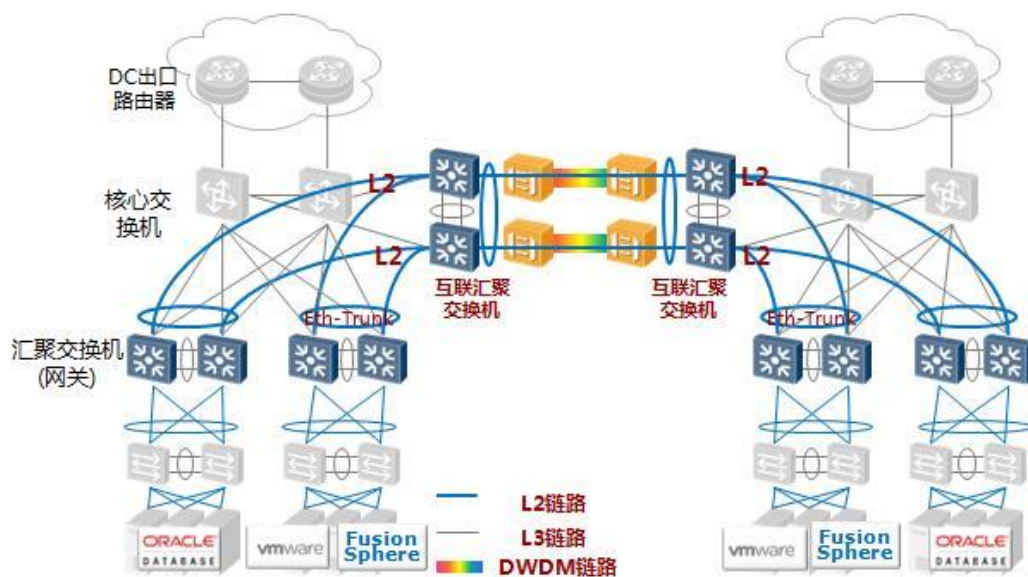
需要二层互联的网络有：

1. DB 层业务网络和心跳网络；
2. 虚拟机业务网络和业务管理网络；
3. C/S 应用的 App 层业务网络和心跳网络。

以下介绍典型的 2 种二层双活互联技术。

普通以太网互联设计

图3-40 中大型 DC 以太网网络架构



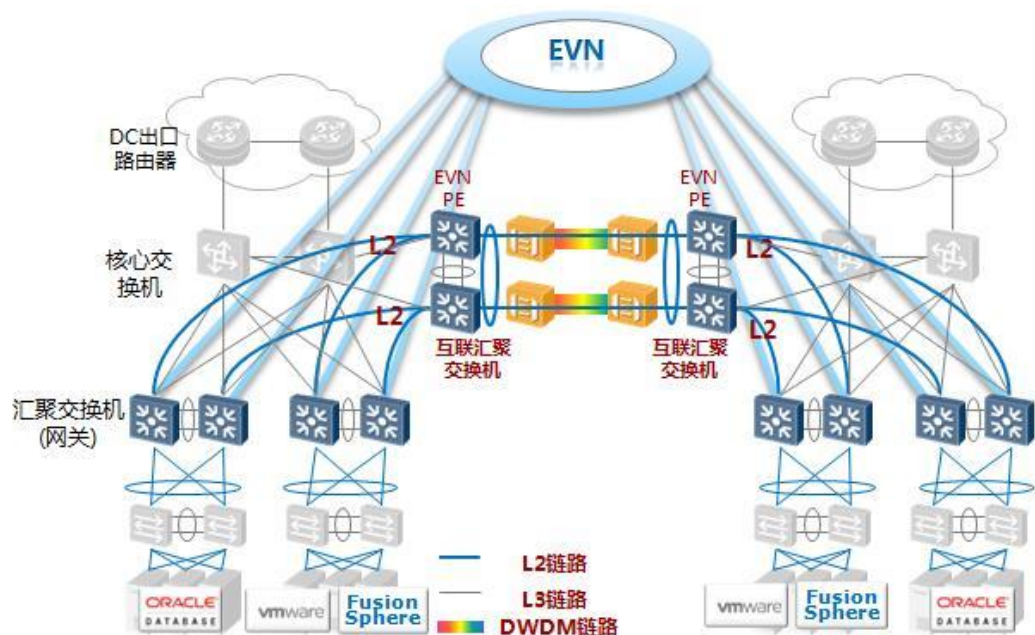
互联接入设计：每个站点部署互联汇聚交换机，站点内的网关交换机通过 CSS+链路聚合接入该互联汇聚交换机，互联汇聚交换机通过 CSS+链路聚合接入波分设备，CSS+链路聚合保证整网无二层环路。同时在汇聚互联交换机配置二层风暴抑制。

互联波分设计：每个站点部署波分设备，波分设备通过两条裸光纤互联；

小型数据中心互联，无需汇聚交换机，直接将接入交换机与核心交换机连接，再通过波分互联，在核心交换机配置二层风暴抑制。

华为 EVN 互联设计

图3-41 中大型 DC EVN 互联网络架构



互联接入设计：每个站点部署互联汇聚交换机，站点内的网关交换机通过 CSS+链路聚合接入该互联汇聚交换机，互联汇聚交换机通过 CSS+链路聚合接入波分设备，互联汇聚交换机运行 EVN PE，EVN PE 间形成 EVN 二层通道。DC 间三层互通，二层域完全隔离 ARP 广播、未知单播限制在本 DC；

互联波分设计：每个站点部署波分设备，波分设备通过两条裸光纤互联；

中小型数据中心的互联设计，无需汇聚交换机，直接将接入交换机与核心交换机相连，再通过波分互联，其它保持一致。

3.4.5 负载均衡技术

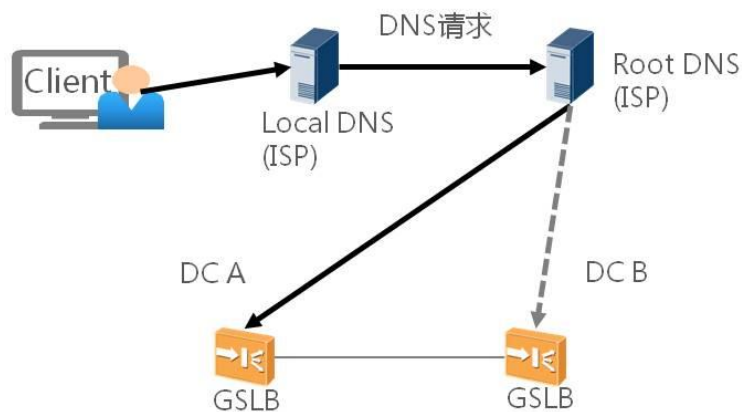
3.4.5.1 站点间负载均衡

域名访问

B/S 业务一般对外提供域名访问。

一个域名对应两个业务 IP 地址，分别部署在两个数据中心。通过 GSLB 或 DNS 实现站点间的访问负载均衡。

图3-42 站点间负载均衡原理（域名访问）



GSLB 可以采用专用的 F5 GTM。如果业务量小，也可以采用 windows 自带的 DNS 服务器，实现简单负载均衡（轮询）。

GSLB 跨站点负载均衡策略通常有两种：

1. 基于 Local DNS 请求所在的地理位置。
2. 基于 GSLB 与 Local DNS 的 RTT 最小。

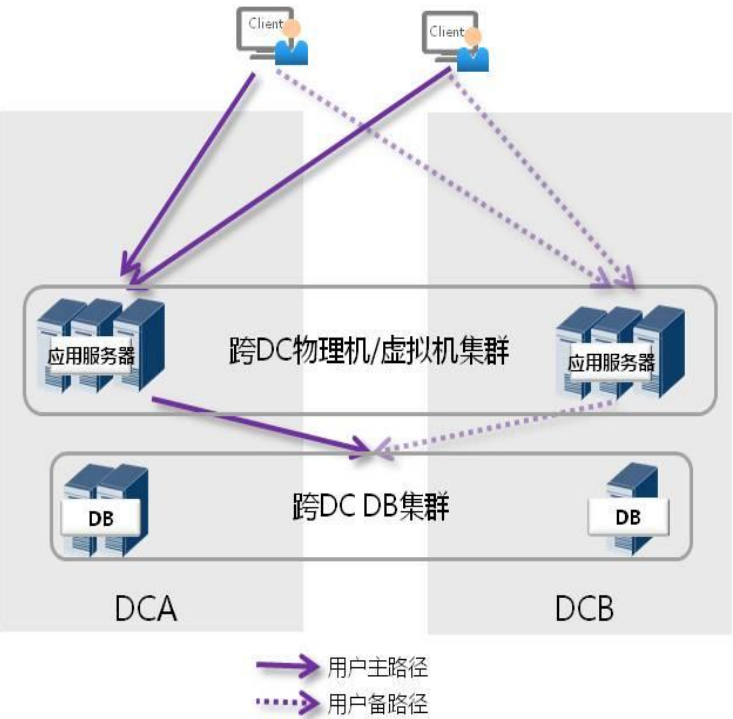
IP 地址访问

C/S 业务一般对外提供 IP 访问。

如果 C/S 应用不支持分布式部署。因为主路由只能在一个数据中心发布，所以一个应用的主机只能在单数据中心运行，但支持故障自动切换至备站点或跨 DC 虚拟机迁移。

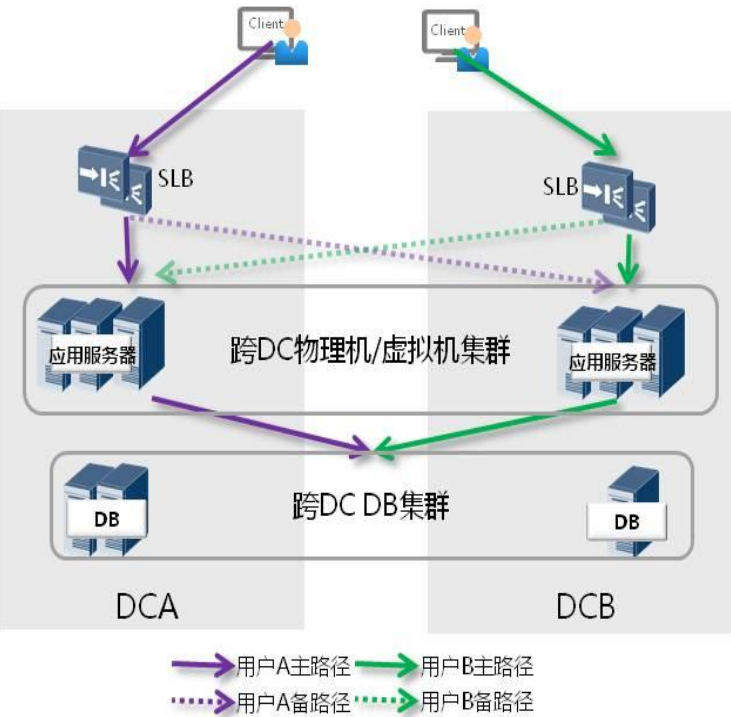
负载均衡策略：将不同应用分别部署到 2 个数据中心运行。

图3-43 站点间负载均衡原理（IP 访问，C/S 应用不支持分布式部署）



如果 C/S 应用支持分布式部署，在两数据中心运行（对外用不同 IP 地址）。负载均衡策略：手动配置客户端，每个客户端配置多个服务器端 IP 或只配置对应的服务器 IP，将不同客户负载均衡到两个 DC。

图3-44 站点间负载均衡原理（IP 访问，C/S 应用支持分布式部署）



3.4.5.2 站点内负载均衡

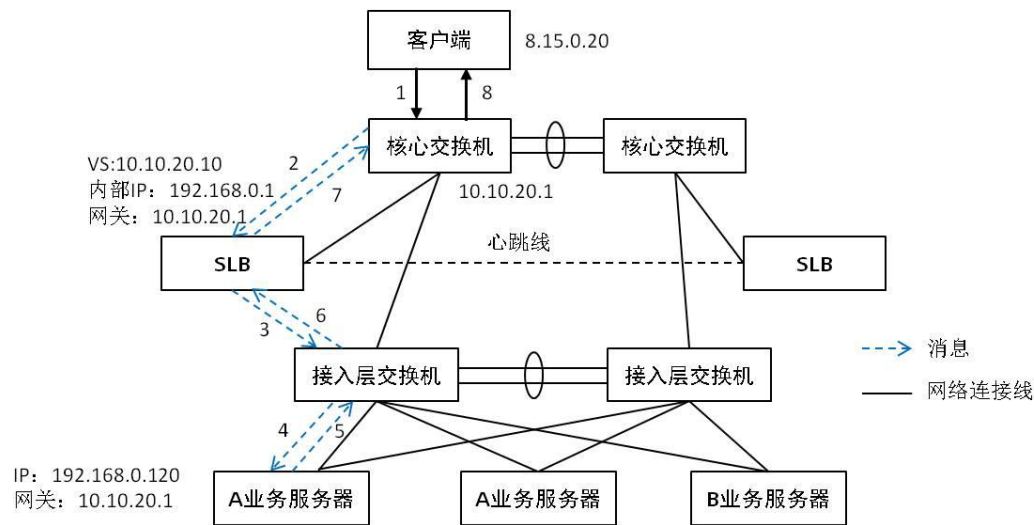
站点内的负载均衡使用 SLB 实现，本方案中 SLB 使用 F5 LTM 部署，支持 http/https、FTP、Diameter、SMPP、SIP 协议等大量应用层协议的负载均衡健康检测。SLB 对外提供的业务接口（包括 IP 地址和端口），接收来自客户端的业务请求。

网络拓扑

双活数据中心方案中，SLB 推荐使用三层旁挂方式组网。

三层旁挂方式，采用将 SLB 旁挂在核心交换机，将业务服务器接到接入交换机上的方式，适合大规模组网的应用场景。组网如下图所示：

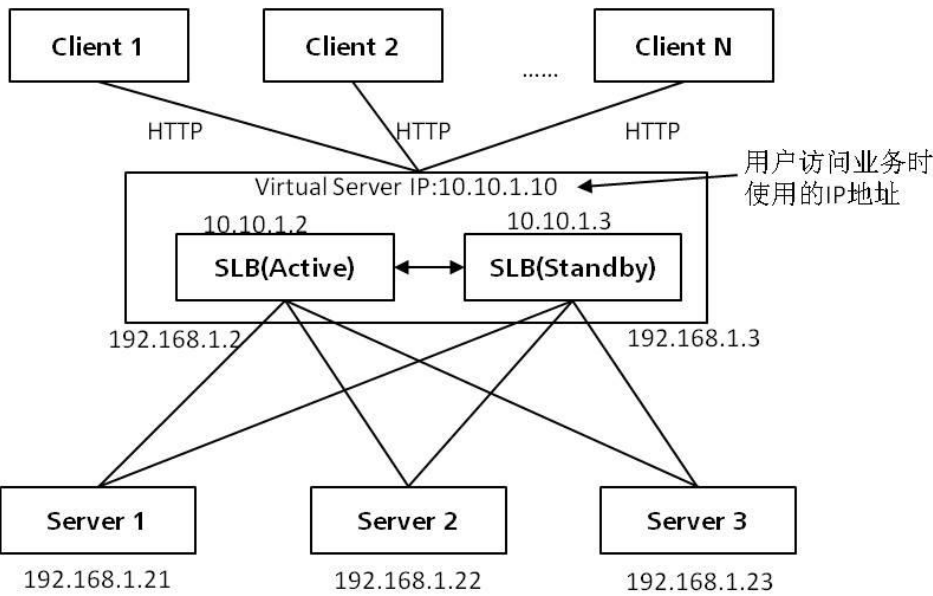
图3-45 SLB 三层旁挂



一个站点部署两台 SLB，组成双机热备集群，两站点各部署一个 SLB 集群，当主用 SLB 故障时会自动切换到备用 SLB 上。

HTTP 协议负载均衡

图3-46 HTTP 负载均衡原理图



SLB 实现 HTTP 负载均衡原理如上图所示。SLB 对外提供业务访问的 IP 地址为浮动 IP，所有 Web 客户端访问浮动 IP。

SLB 基于 HTTP 业务，提供多种调度算法：

1. 最少连接数：最少连接数调度算法将客户端请求分发至拥有最少连接数的真实服务器。
2. 轮询：调度算法在资源池中的真实服务器之间进行轮询分发调度。
3. 加权最少连接数：基于最少连接数调度算法的基础上考虑性能权重，将更多的连接请求分发到连接数少的真实服务器。
4. 加权轮询：基于轮询调度算法，但是将会给权重大的服务器分发更多的连接请求。

SLB 基于 HTTP 业务，提供多种健康检测策略，健康检查用于确保远程服务器的可用性，SLB 支持基于 Ping、UDP、TCP、HTTP 和 SHELL 脚本的健康检查。健康检查失败的节点将不再接受新的请求，直至它重新被健康检查为可用状态。

3.5 传输层技术

华为 OSN 系列波分产品采用了先进的低时延处理技术，针对时延问题进行了优化，非常符合对对延时要求较高的数据中心互联应用。

- 华为采用了先进的 GMP 封装技术，具有超低的电层处理时延。
- 华为 OTN 产品可以根据需要采用不同的 FEC 方式，以减少 FEC 引入的延时。在满足传输需求的情况下，可以关闭 FEC 或者采用延时较小的 FEC 方式，最大化的缩小系统延时。
- 针对系统色散补偿引入的延时，华为通过一下方式来减小时延：
 - 华为可以提供基于 FBG（Fiber Bragg Grating）的色散补偿方案，与传统的色散补偿光纤相比，FBG 本身基本不引入延时（小于 0.1us），具有非常大的低时延优势。
 - 针对需要使用传统色散光纤补偿的场合，华为提供一系列模块化的色散补偿模块，可以针对需要进行精确的色散补偿，尽量减少色散补偿光纤的长度。
 - 华为 40G/100G 系统，采用先进的 DSP 算法，无需色散补偿，彻底消除色散补偿光纤对延时的影响。同时华为率先支持 100G 二代软判技术，延时得到大幅下降。
- 针对光纤路径延时，华为提供 OTN 延时测量功能，当存在多个光纤路径可以选择时，可以选择最优的传输路径。

3.6 安全层技术

双活数据中心除传统的数据中心内的安全威胁外，还有跨数据中心数据安全的考虑。华为针对数据中心和跨数据中心传输有全面的安全解决方案。

例如，华为针对数据中心网络分区，提供了以下多种安全方案：

- 互联网接入区安全方案
- 外联网接入安全方案
- 内网接入区安全方案
- 数据中心互联区安全方案

- 数据中心服务区安全方案
- 网管维护区安全方案

本文中，针对双活数据中心的业务特点，安全问题主要关注两个双活数据中心解决方案组件：防火墙和波分传输设备。华为波分传输设备可通过加解密传输板卡解决数据传输中的安全问题，该加解密为硬件加解密，性能高，几乎对传输时延没有影响。下面我们重点介绍以下几种重要的访问安全方案设计。

传输层低时延加密方案

加密网络的构成如下图所示。客户业务在本端客户侧通过配置了加密功能的设备后，以加密后的业务形式在运营商网络传输，并在远端客户侧解密，确保业务的安全传输。

图3-47 加密网络构成

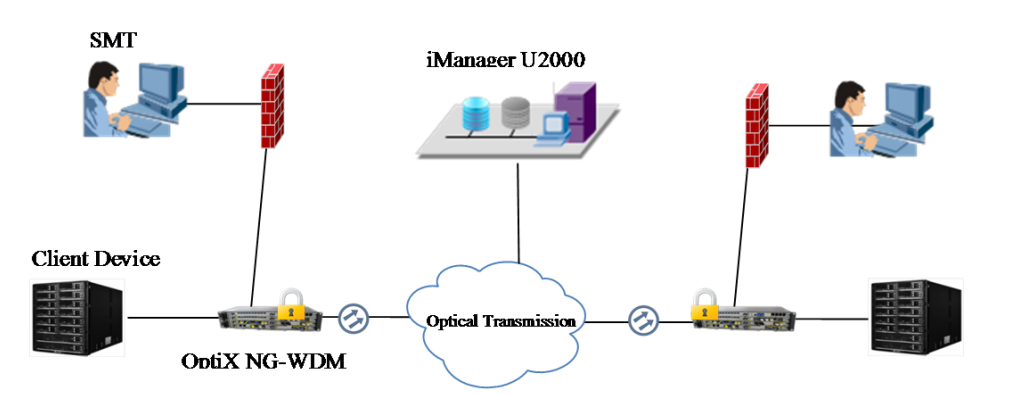


表3-6 安全管理单元部署及通信方式

网络管理系统分类	部署位置	与网元通信方式	远程管理网元的功能差异
安全管理终端（SMT，Security Management Terminal）	DC 内业务分区	通过基于 SSHv2.0 协议的安全通信方式，与传送加密设备通信。	网元上运行安全管理系统（SMS，Security Management System），实现加密参数的设定和启用，以及加密功能运行情况的监控。
iManager U2000	DC 内管理分区	通过基于 SSL3.0/TLS1.0 协议的安全通信方式与网关网元通信，通过网关网元基于	网元上运行配置管理、性能管理、告警管理等子系统，通过 U2000 来访问这些管理功能实现

网络管理系统分类	部署位置	与网元通信方式	远程管理网元的功能差异
		ECC 方式转发，实现与加密传送设备通信。	业务配置和网络运维。

加密网络的管理和运营需具备以下条件：

- 传输设备管理员完成网络的建设和部署。
- 传输设备管理员完成基本的网元管理和业务参数的配置。
- 传输设备管理员完成业务的发放。
- 传输设备管理员为最终业务用户创建用于加密管理的网元账户。
- 最终业务用户完成 SSH 客户端软件的安装。

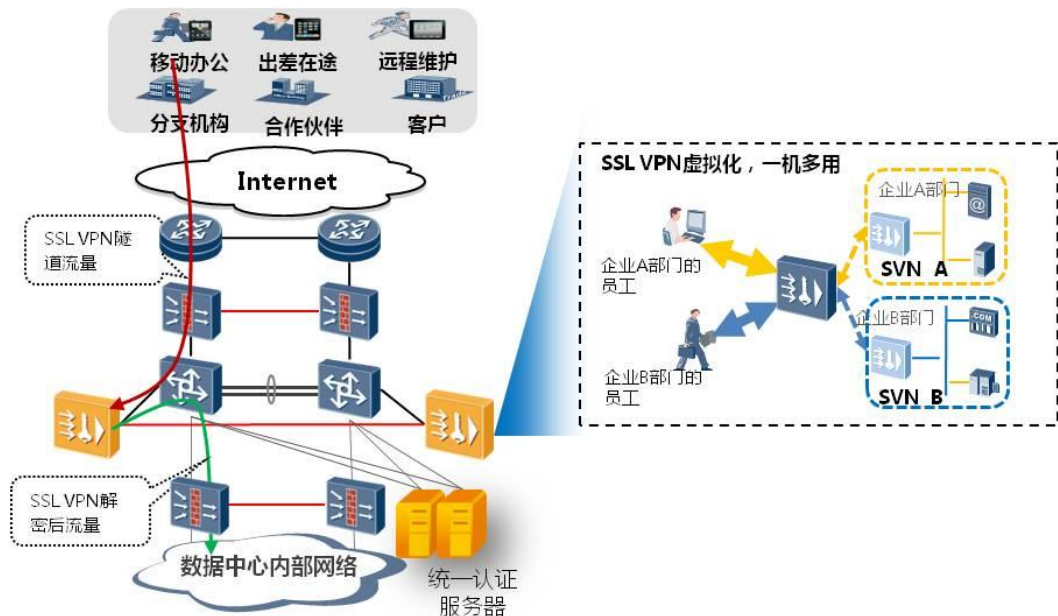
加密算法的选择：

考虑到一层加密解决方案需要应用在数据中心、SAN 等场景，需要对大带宽的数据进行加解密，因此，对加密算法有加密强度高、高效、简单、随机性等要求，而且要便于硬件实现。综合来看，当前安全业界较为流行的 AES（Advanced Encryption Standard）算法，能满足上述要求，是比较合适的选择。

AES 算法的密钥长度有 128、256、512bit 几种。同时 AES 支持 5 种算法工作模式：CBC、ECB、CTR、OFB、CFB。其中，CTR 模式被广泛用于 ATM 网络安全和 IPSec 应用中。综合来看，采用 256bit 的密钥程度并工作于 CTR 模式的 AES 算法无疑从各方面来讲，是传送产品业务加密解决方案中加密算法的最佳选择。

SSLVPN 安全接入方案

图3-48 SSLVPN 安全接入方案

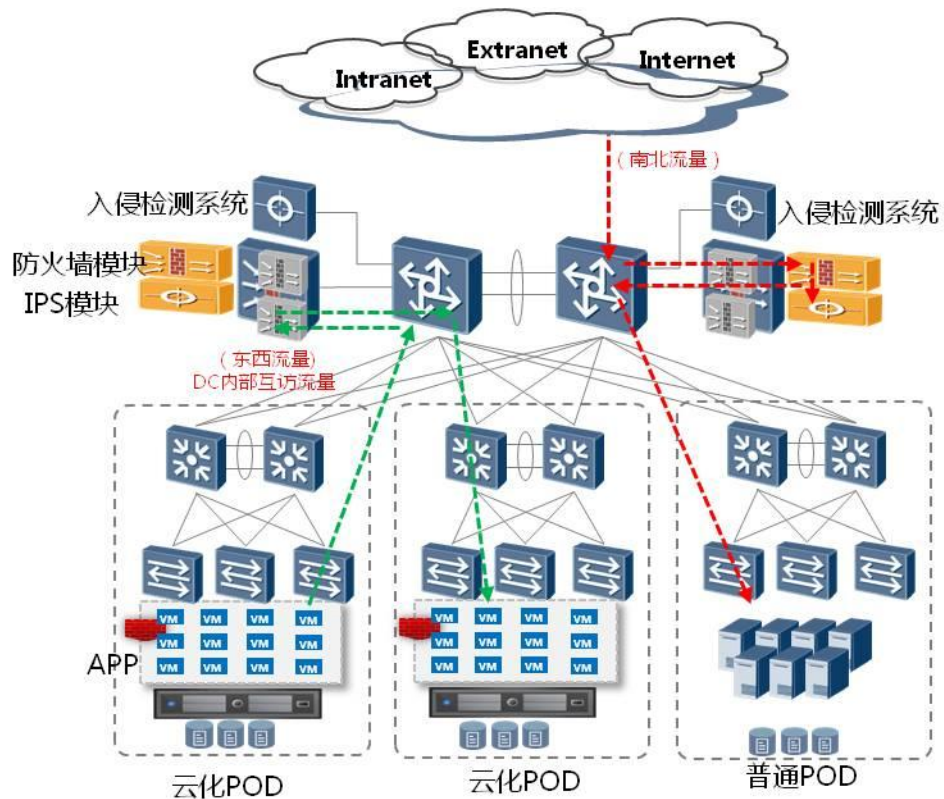


部署方案

- SVN 网关一般旁挂在网络出入口防火墙之后，应用服务器之前，防火墙开启 SSL 默认端口 443 。
- SVN 网关上配置一个私网 IP，与内网服务器路由可达；对外公布一个公网 IP 地址供用户访问，该地址可以配置在 SSL 网关上，或由防火墙做 NAT。
- 远程用户通过 Web 浏览器或者客户端软件访问，输入网关 IP 地址，报文先到达防火墙，防火墙发现目的地址是 SVN 网关，则转发给 SVN 处理，SVN 解密后再发送到目的服务器。

数据中心内部网络典型安全设计

图3-49 DC 内部网络典型安全架构图



- 不同业务区互访控制 & 安全隔离（东西流量管控）

实现设备：USG 统一安全网关

实现方式：统一安全网关旁挂

对于需要互访的区域或服务器，将流量牵引至防火墙通过安全策略进行访问控制，严格实现最小访问授权；

对于互相独立、业务上无互访需要的两个业务区，在防火墙上进行严格的访问隔离策略，或者通过防火墙虚拟化技术为不同业务区分配独立的v 防火墙配合交换机的虚拟化技术实现流量的完全隔离。

- 数据中心外部与内部的互访控制（南北流量管控）

实现设备：统一安全网关

实现方式：统一安全网关旁挂（防火墙业务板+入侵防御系统业务板）

使用防火墙严格控制各地市的访问来源及访问目的；

对于进入数据中心业务区的流量启用入侵防御系统进行清洗，尤其是对互联网业务进入的流量。做到严格访问授权、攻击防护及恶意代码过滤等。

- 入侵攻击\恶意渗透的监测取证

实现设备：入侵检测系统

实现方式：核心交换机镜像流量给入侵检测系统，对入侵攻击进行检测、预警及取证。

- 虚拟化服务器内部流量管控

实现设备：DeepSecurity

实现方式：通过在 ESXi 平台中部署 DeepSecurity 安全组件，将需要管控的虚拟机流量导入虚拟防火墙中按规则进行过滤。

4 可视化容灾管理

OceanStor BCManager 是一款基于华为存储双活、快照、远程复制等数据保护特性的专业容灾管理软件，提供应用保护、管理自动化和管理可视化的服务能力。BCManager 可实现生产中心到灾备中心的容灾站点管理，实现业务主机应用、业务存储、灾备存储端到端的容灾资源管理与监控，实现双活容灾方案的拓扑展示，并且在快照和异步远程复制保护时，能保证应用数据的一致性，将人工单点操作转化为一键式自动执行。本章节将对 OceanStor BCManager 软件的主要技术特点进行介绍。

4.1 总体部署

4.2 应用支持矩阵

4.3 适用容灾场景

4.1 总体部署

4.1.1 部署方式

OceanStor BCManager 基于 B/S 架构，通过浏览器即可进行容灾的管理。它包含 2 个子系统：BCManager Agent、BCManager Server。

- BCManager Agent 安装在业务主机上，提供主机、应用的发现功能。
- BCManager Server 安装在独立服务器上，提供整个容灾管理系统的配置、调度等业务功能。

针对双活解决方案的不同场景，BCManager 有两种部署方式。在双活容灾场景，以及双活+快照保护的容灾场景下，BCManager Server 部署在生产中心 1 或者生产中心 2，与业务主机、业务存储连接在同一个管理平面，见图 1；在双活+异步复制保护的容灾场景下，BCManager Server 部署在灾备中心，与生产中心的业务主机、业务存储，以及灾备中心的灾备主机、灾备存储分别连接在同一个管理平面，见图 2。

图4-1 华为存储双活保护管理部署图

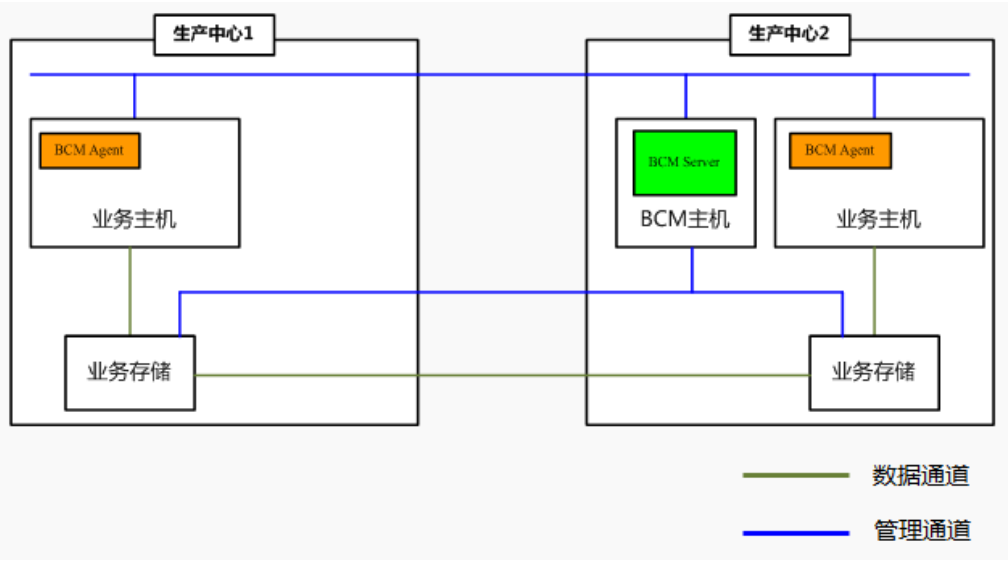
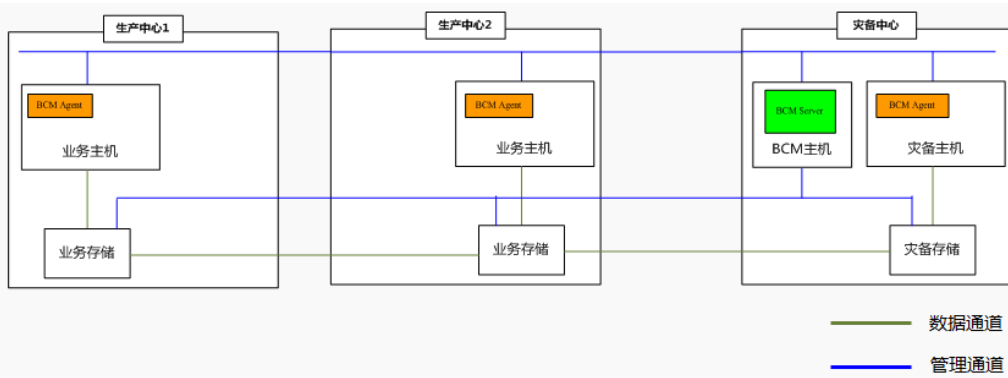


图4-2 华为存储双活+异步复制容灾管理部署图



在两种部署方式下，系统组件间的通信协议和方式是一样的，如下所示。

- BCManager Server 与存储间：REST over Https
- BCManager Server 与 BCManager Agent： SOAP over Https
- BCManager Agent 与存储：iSCSI 或 FC

4.1.2 系统运行环境

OceanStor BCManager 容灾管理软件的部署环境要求如下表所示：

规格项		规格要求	备注
Server	操作系统	SUSE Linux Enterprise Server 11 SP1 64 位	

		Windows Server 2008 R2 企业版 64 位	
	CPU	最低配置：2Core 1.6GHz CPU 标准配置：4Core 1.6GHz CPU 及以上	
	内存	最低配置：DDR 4GB 内存 标准配置：DDR 8GB 内存及以上	1、如果系统的配置为 4G 物理内存，则要求可用物理内存至少 2G。如果系统的配置为 8G 物理内存，则要求可用物理内存至少 4G。 2、如果安装环境是虚拟机，虚拟机不能采用内存动态分配的方式。 3、如果安装环境是 Linux 操作系统，swap 要求如下：1）当内存为 4GB~16GB 时，swap 的取值与内存大小相同。2）当内存大于 16GB 时，swap 的取值为 16GB。
	硬盘 空闲 空间	最低配置：大于等 于 10GB 标准配置：空闲空 间大于等于 100GB	
	数据 库	GaussDB V100R003C10	
	管理 网络 带宽	最低带宽配置： 2Mb/s 标准带宽配置： 10Mb/s	数据中心间管理网络
	部署 方式	集中式部署	仅部署一台 OceanStor ReplicationDirector Server 容灾管理服务器，要求该管理服务器必须与各数据中心间管理网络互通。
分布式部署		主备容灾、两地三中心等解决方案场景中最多部署 2 台 OceanStor ReplicationDirector Server 容灾管理服务器，要求管理服务器间的管理网络互通。	
Agent	内存	标准配置：DDR 1GB 内存	对业务主机的最低要求。
	硬盘 空闲 空间	标准配置：100MB	对业务主机的最低要求。
	部署	部署在应用服务器	在生产站点和灾备站点的应用服务器上均

	方式	上	需部署。
--	----	---	------

4.2 应用支持矩阵

OceanStor BCManager 容灾管理软件既能对物理机上的数据库应用，也能对虚拟机进行保护，支持的应用矩阵如下表所示。数据库应用包括 Oracle RAC 集群、SQL Server MSFC 集群，以及 IBM AIX 高可用集群 PowerHA 上的 Oracle 应用或 DB2 应用。虚拟机的保护则支持华为的 FusionSphere，以及 VMware 虚拟机。

容灾方案	容灾技术	数据库				虚拟化环境	
		Oracle		SQL Server	IBM DB2	FusionSphere	VMware
		RAC	PowerHA	MSFC	PowerHA	5.0、5.1	5.0、5.1、5.5
双活数据中心容灾解决方案	双活（SAN）	√	√	√	√	√	√
	双活（SAN）+ 快照（SAN）	√	√	√	√	√	√
两地三中心容灾解决方案	双活（SAN）+ 异步复制（SAN）	√	√	√	√	√	√

4.3 适用容灾场景

4.3.1 SAN 双活场景

4.3.1.1 技术特点

在双活的容灾环境中，OceanStor BCManager 容灾管理软件能端到端的直观的展示从业务主机应用到应用所使用的存储，以及业务存储到灾备站点存储之间的容灾关系，对容灾环境进行监控，帮助用户掌握双活数据中心的容灾状态。

OceanStor BCManager 提供物理拓扑和逻辑拓扑两种可视化管理方式。物理拓扑主要是监控容灾环境中的所有网元和链路，不仅能监控业务主机、业务存储和仲裁服务器的故障，还能监控主机到存储、存储到仲裁服务器之间链路的故障，以及存储双活的状态。

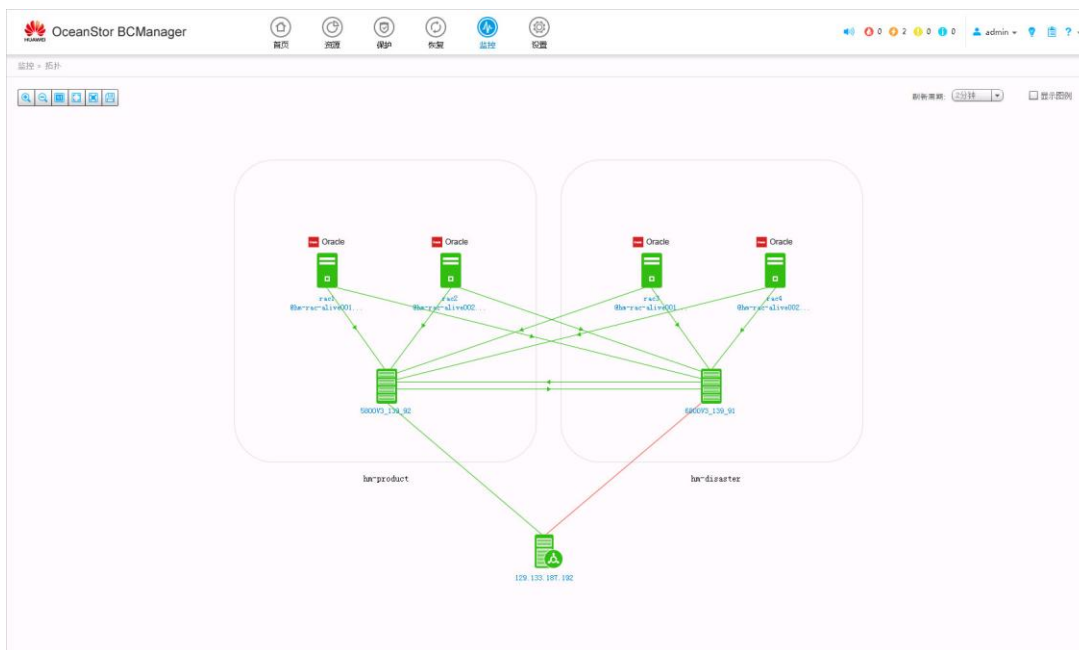
态。同时，在物理拓扑上还能呈现出主机、存储、仲裁服务器的名称、IP 地址、状态等基本信息，以及存储间所有的 FC 或者 iSCSI 链路信息。逻辑拓扑则是从应用的视角出发，展现应用与主机磁盘、主机磁盘与存储双活子 LUN 之间的逻辑关系。通过逻辑拓扑，可以看到应用（或实例）的信息，以及双活 LUN 的名称、ID、WWN、容量和状态。

4.3.1.2 物理拓扑

OceanStor BCManager 容灾管理软件通过分析各网元之间的关系，构造全局物理拓扑。通过查询主机 HBA 卡与存储的启动器信息，识别出主机与存储间的连接关系，通过查询双活存储与存储间的链路信息，构造双活存储与存储的链路关系。根据保护对象的双活 PairID 获取该双活 Pair 关系的高可用域，通过该域获取仲裁服务器信息，以及仲裁链路，从而构造生产存储与灾备存储至仲裁服务器的链路关系。

图 1 展示的是 Oracle RAC 应用在存储双活场景下的物理拓扑示例。

图4-3 Oracle RAC 应用在存储双活场景下的物理拓扑图

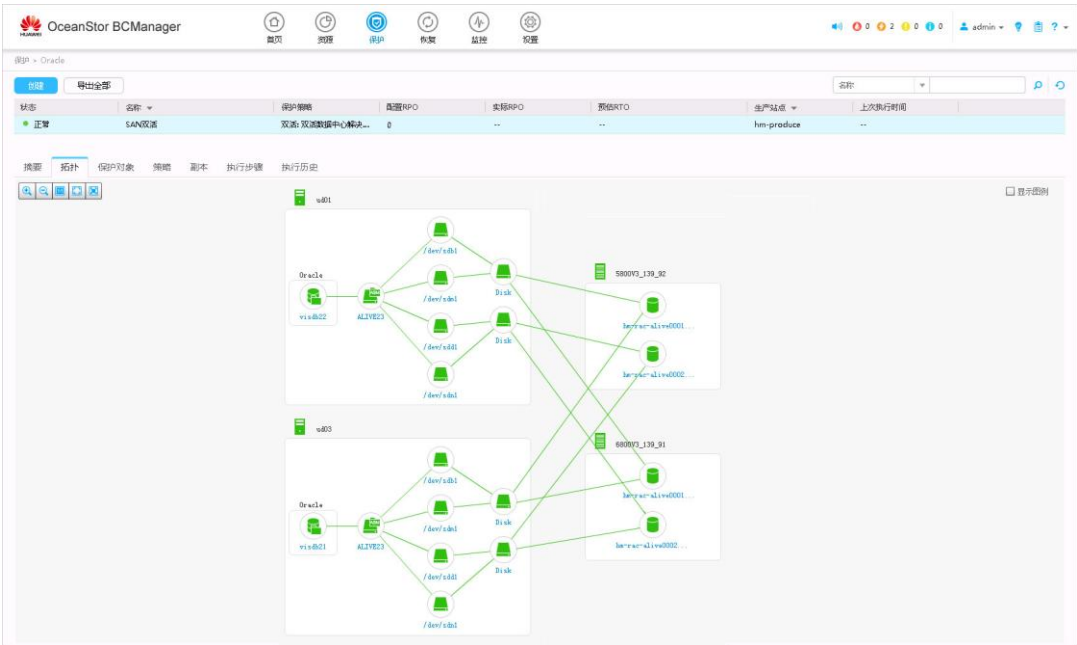


4.3.1.3 逻辑拓扑

OceanStor BCManager 容灾管理软件通过分析应用数据与业务存储的关系，构造逻辑拓扑。针对物理机上的数据库应用，通过业务主机上的 agent 获取保护的数据库信息，以及数据库文件存放的主机磁盘信息，通过主机和存储的映射关系，查询主机磁盘与存储双活子 LUN 之间的逻辑关系，从而构造应用的逻辑拓扑。对于虚拟机的保护，则是通过虚拟化平台提供的接口获取虚拟机与数据存储的关系，从而构造虚拟机的逻辑拓扑。

图 1 展示的是 Oracle RAC 应用在存储双活场景下的逻辑拓扑示例。

图4-4 Oracle RAC 应用在存储双活场景下的逻辑拓扑图



4.3.2 SAN 双活+快照场景

4.3.2.1 技术特点

在双活+快照的容灾场景中，OceanStor BCManager 容灾管理软件不仅能对双活容灾环境进行可视化监控，还能提供应用一致性、双站点打快照、基于快照副本的测试，以及快照回滚同时恢复双活关系的功能，实现一键式自动化的管理。

OceanStor BCManage 可以对应用使用的双活子 LUN 分别配置时间策略，通过自动或者手动的方式给双站点打快照，打快照的时候能通过 Oracle 的热备模式或者数据库悬挂 IO 的方式将缓存中的数据刷到磁盘上，以保证应用的一致性。

OceanStor BCManage 提供基于快照副本的测试功能，测试时不会影响生产业务。执行测试演练的过程中，BCManage 软件会自动创建快照的副本、映射快照的副本给测试主机、在测试主机上配置存储、下发数据库启动命令，以及测试数据库的连接。

OceanStor BCManage 可以对存储快照执行一致性回滚，在回滚过程中自动配置主机与存储的映射关系、在业务主机上扫描和挂载数据存储、在业务主机上启动和测试数据库，并且在回滚成功后重新启动存储双活，自动恢复双活关系。

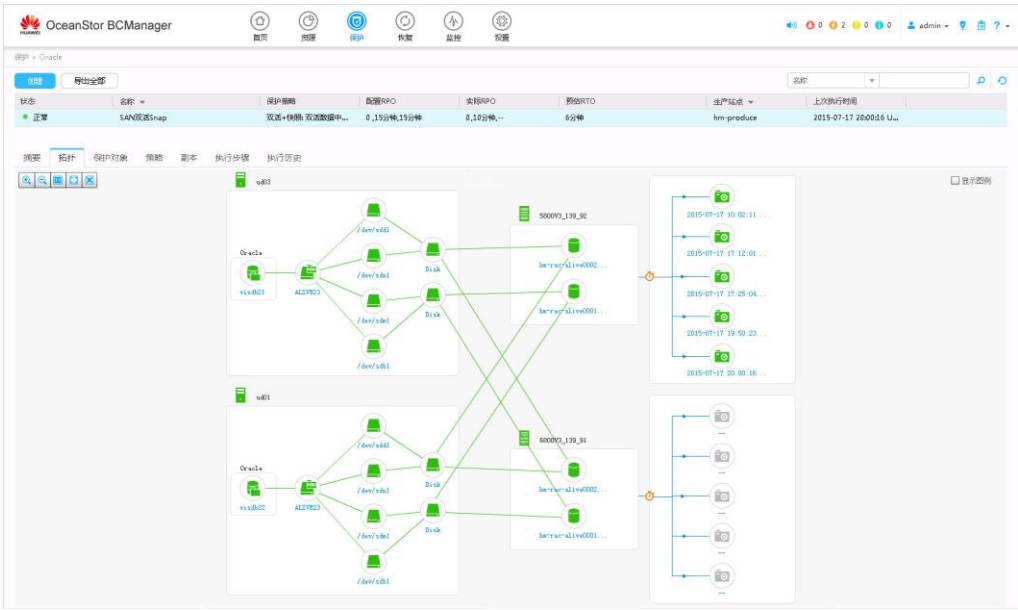
4.3.2.2 物理拓扑

SAN 双活+快照场景的物理拓扑同 SAN 双活场景，参考 4.3.1.2 章节。

4.3.2.3 逻辑拓扑

SAN 双活+快照场景中，应用的逻辑拓扑除了展现出应用与业务存储的逻辑关系外，还会展示双活子 LUN 的快照副本情况，以及快照策略，如图 1Oracle RAC 应用示例所示。

图4-5 Oracle RAC 应用在存储双活+快照场景下的逻辑拓扑图



4.3.3 SAN 双活+异步复制场景

4.3.3.1 技术特点

在双活+异步远程复制的容灾场景中，OceanStor BCManager 容灾管理软件不仅能对两地三中心的容灾环境进行可视化监控，还能支持并联和级联场景下的应用一致性、容灾测试演练、计划性迁移、故障恢复、重保护，以及回切等一系列的自动化容灾管理功能。

在这个场景中，容灾测试和恢复主要是在异地灾备站点上执行。若生产中心和同城灾备中心均发生灾难，可通过 BCManager 在异地灾备中心对远程复制进行主从切换，并拉起灾备中心业务。在日常维护中，可通过 BCManager 在异地灾备中心执行容灾测试演练和计划性迁移。

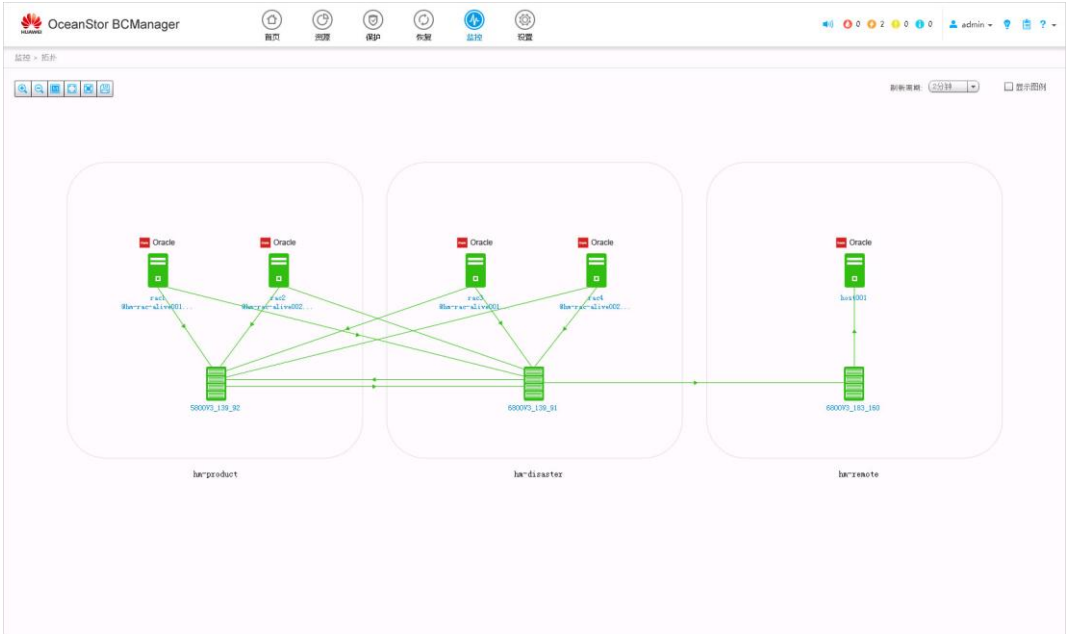
OceanStor BCManager 将人工单点的复杂过程转化为一键式自动执行的简单快捷操作，提供可视化和自动化的容灾管理。在执行计划性迁移过程中，BCManager 会自动暂停业务存储的双活关系，分裂业务存储和异地灾备存储的远程复制关系，设置从端 LUN 可写，将灾备端的存储映射灾备主机，自动拉起灾备端业务。重保护后，BCManager 会自动重建保护关系，将异地灾备存储的数据同步到生产中心。BCManager 进行生产回切的时候，还能自动恢复存储的双活关系，操作简单便捷。

4.3.3.2 物理拓扑

SAN 双活+异步复制场景的物理拓扑中会根据双活子 LUN 与异地灾备存储间的远程复制关系，构造异步复制组网，同时根据灾备主机上的 HBA 信息以及灾备存储上的启动器信息，构造出灾备主机与灾备存储的关联关系。

最终的全局物理拓扑如图 1Oracle RAC 应用示例所示。

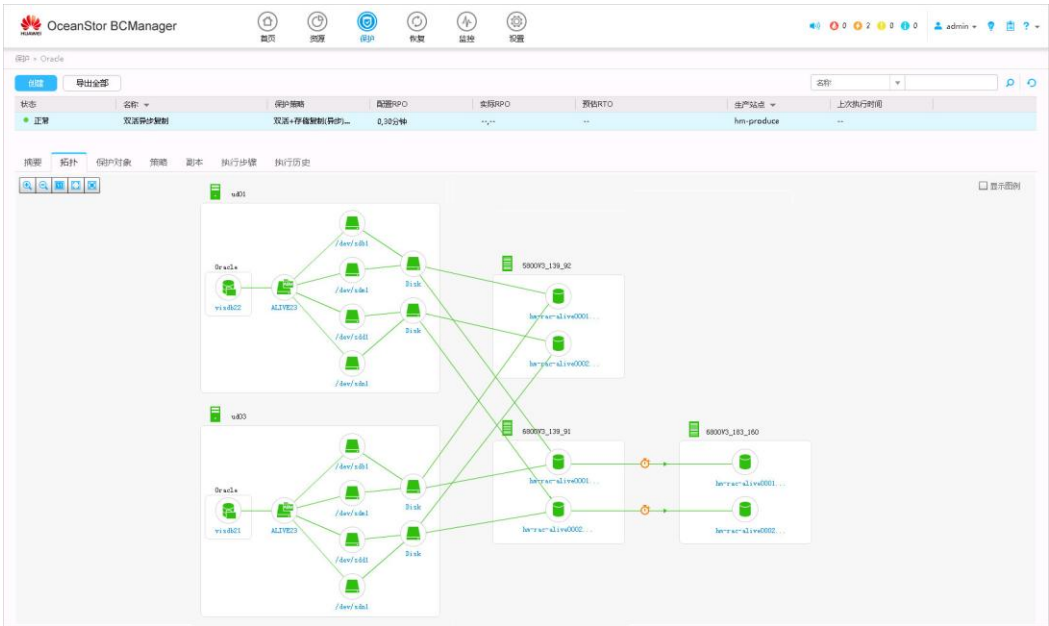
图4-6 Oracle RAC 应用在存储双活+异步远程复制场景下的物理拓扑图



4.3.3.3 逻辑拓扑

SAN 双活+异步复制场景的逻辑拓扑中将展示双活子 LUN 与异地灾备存储 LUN 之间的远程复制关系，同时在远程复制关系链路上呈现保护组配置的策略信息，如图 1Oracle RAC 应用示例所示。

图4-7 Oracle RAC 应用在存储双活+异步远程复制场景下的逻辑拓扑图



5 故障场景

本章主要介绍两个数据中心相距 100km 场景下，方案可靠性验证结果。

在各单部件和链路故障，甚至整个数据中心故障，都可以实现业务自动无缝切换。

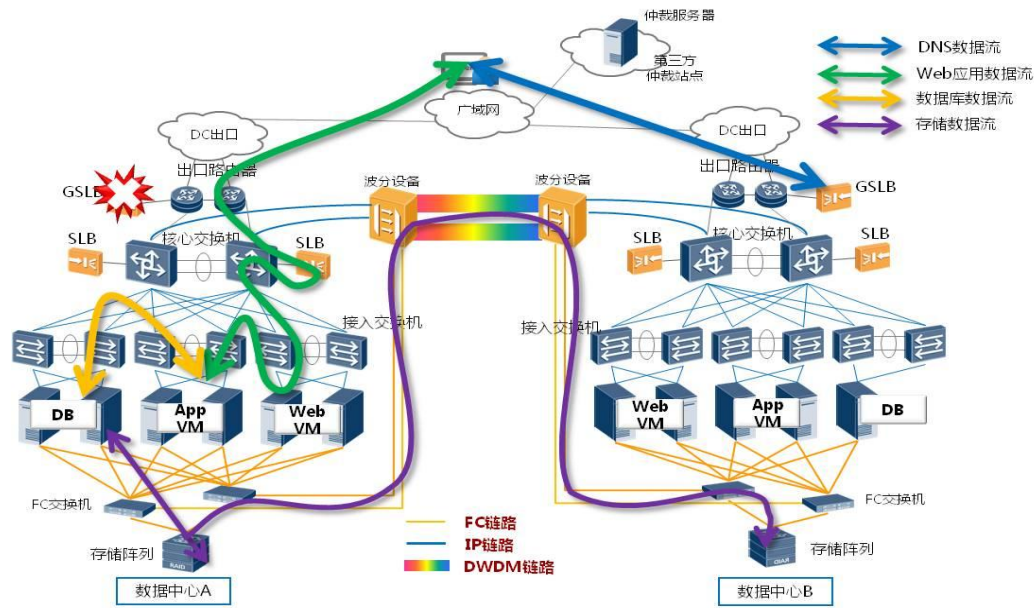
以下列举各种故障场景的连续性分析：

- 5.1 GSLB 故障
- 5.2 SLB 故障
- 5.3 Web 服务器故障
- 5.4 应用服务器故障
- 5.5 Oracle RAC 故障
- 5.6 IBM DB2 故障
- 5.7 阵列故障
- 5.8 广域网链路故障
- 5.9 站点间链路故障
- 5.10 站点故障

5.1 GSLB 故障

两站点各部署一台 GSLB（Global Server Load Balancing），一个 DC 中只需配置一台。上级 DNS 通过轮询方式保证第一台 GSLB 故障时其他的 GSLB 可以接替工作。

图5-1 GSLB 故障-数据流向示意



数据中心 A 的 GSLB 故障，原来用户继续使用数据中心 A 的 Web 服务器访问，业务正常运行，如下描述新用户的访问情况，处理过程如下：

1. 新用户发起访问前，查询该域名的 DNS。
2. 上级 DNS 通过轮询方式，发现数据中心 A 的 GSLB 故障，访问数据中心 B 的 GSLB，数据中心 B 的 GSLB 根据负载均衡策略返回 IP 地址给用户。
3. 新用户根据 IP 地址发起访问。

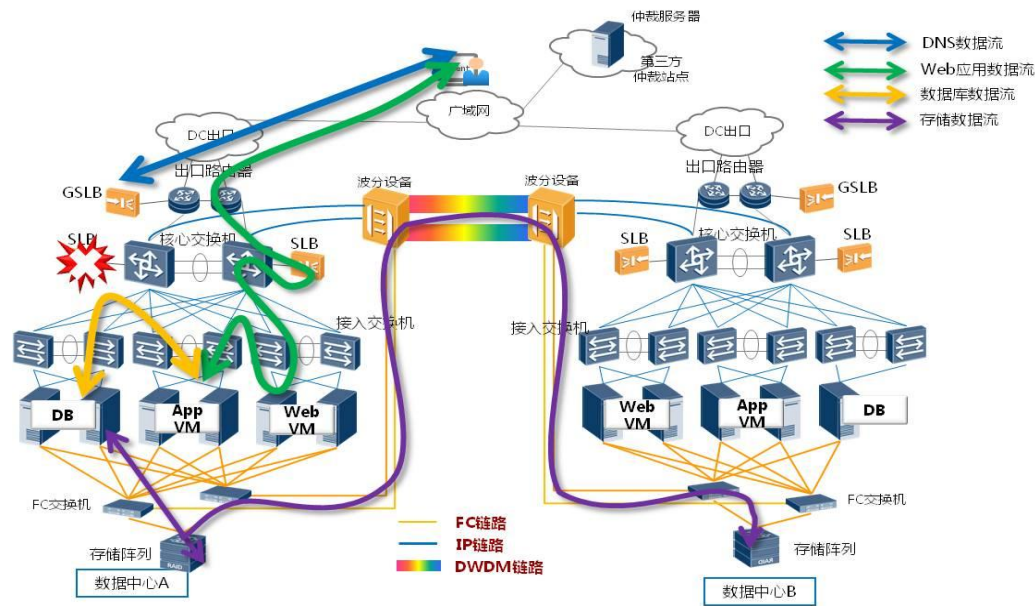
数据中心 A 的 GSLB 故障恢复后，上级 DNS 将访问数据中心 B 的 GSLB，上层业务无影响，负载均衡策略不变。

5.2 SLB 故障

每个站点部署两台 SLB（Server Load Balancing），组成双机 HA 集群，自动同步集群各节点的配置信息，确保所有节点配置文件的一致性；两站点各部署一个集群，各自独立。

站点内一台 SLB 故障

图5-2 站点内一台 SLB 故障-数据流向示意图



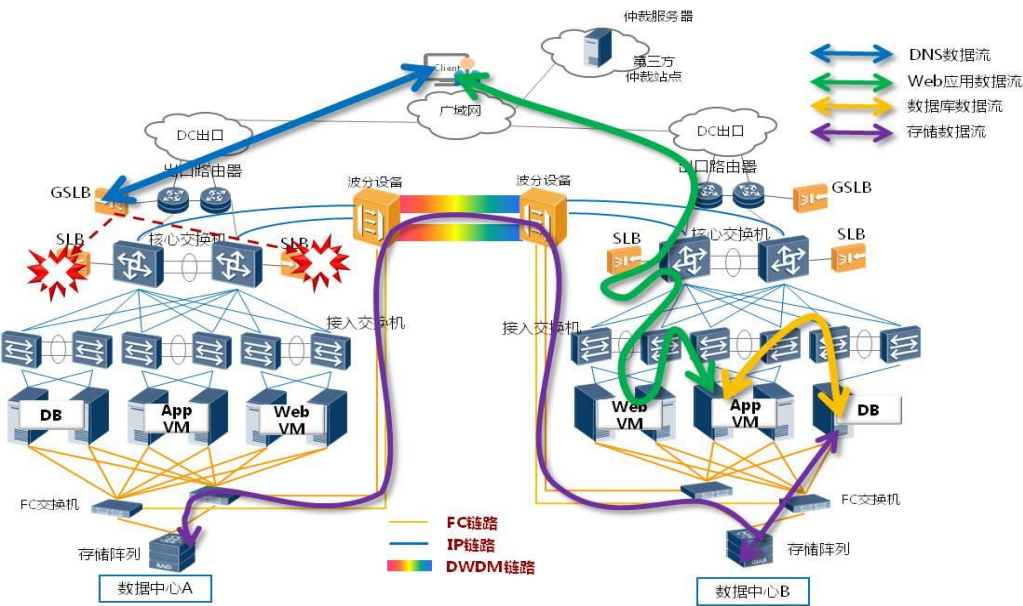
站点内 SLB 主节点故障，处理过程如下：

1. 站点内部 SLB 备节点接管 Virtual Server IP。
2. 客户端将 HTTP 请求依然发至 SLB 的 Virtual Server IP。
3. Web 服务器和应用服务器仍然照常工作，不受影响。

数据中心 A 的 SLB 故障恢复后，SLB 自动重组 HA 集群，客户端仍然访问 SLB 的 Virtual Server IP，上层业务无影响，单个客户的 HTTP 会话能够保持。

站点内两台 SLB 故障

图5-3 站点内两台 SLB 故障-数据流向示意图



数据中心 A 内两台 SLB 故障，处理过程如下：

1. SLB 对外的 Virtual Server IP 将无法访问。
2. 数据中心 A 的 GSLB 探测到本中心 SLB 的 Virtual Server IP 无法访问，修改负载均衡策略，不再往本中心 SLB 发送。
3. 客户端侧的 Local DNS 缓存超时，重新学习到指向数据中心 B 的 DNS。
4. 客户端将 HTTP 请求发至数据中心 B 的 SLB 的 Virtual Server IP。

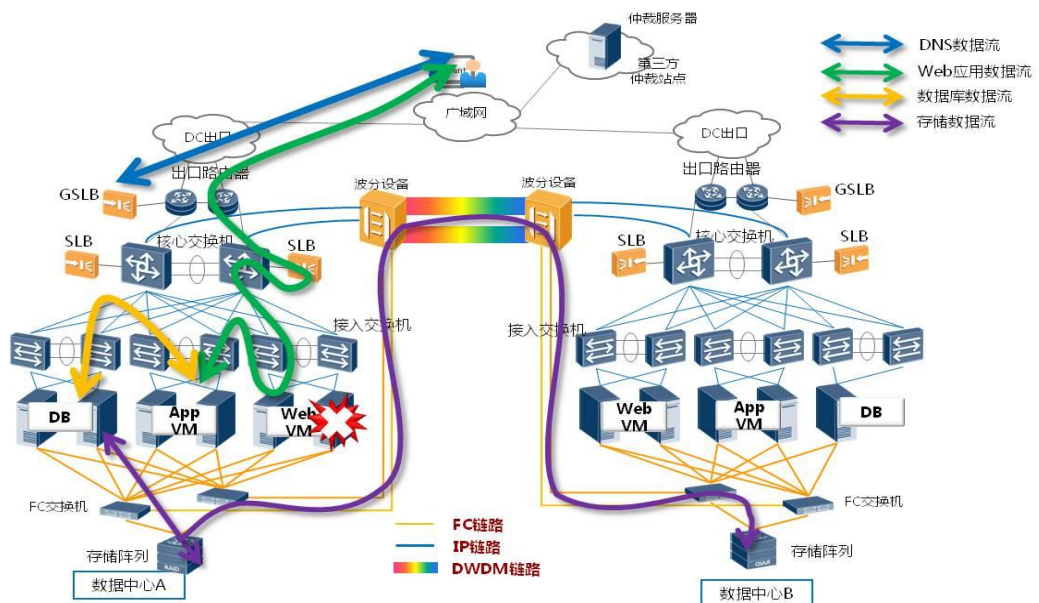
数据中心 A 的 SLB 故障恢复后，GSLB 探测到 SLB 的虚拟服务 IP 恢复正常，修改负载均衡策略，分流新业务到数据中心 A。

5.3 Web 服务器故障

一个站点部署多台 Web 服务器，创建一个集群。由 SLB 创建一个 pool（负载均衡池），将站点内的所有 Web 服务器组成一个资源池。这样分别创建 DC1 资源池和 DC2 资源池。

站点内一台 Web 服务器故障

图5-4 站点内一台 Web 服务器故障-数据流向示意图



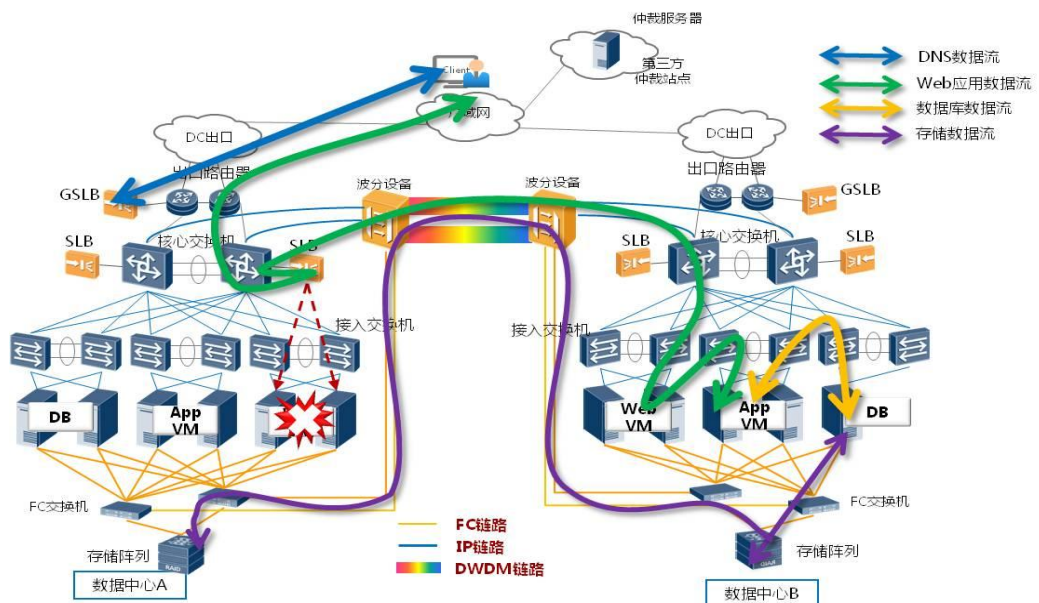
假如站点内一台 Web 服务器故障，处理过程如下：

1. 一台 Web 服务器故障，无法提供服务。
2. 客户端将请求发至 SLB 的 Virtual Server IP。
3. SLB 根据查询到上次处理该会话的 Web 服务器节点，但检测到故障，分发至本站点的另外一台 Web 服务器。
4. 新 Web 服务器指定应用服务器节点进行访问，单个客户的 HTTP 会话能够保持。

数据中心 A 的 Web 服务器故障恢复后，SLB 自动重新将其加入资源池，上层业务无影响，单个客户的 HTTP 会话能够保持。

站点内所有 Web 服务器故障

图5-5 站点内所有 Web 服务器故障-数据流向示意图



假如站点内所有 Web 服务器故障，处理过程如下：

1. 数据中心 A 的 Web 服务器对外的 IP 将无法访问，无法提供服务。
2. 数据中心 A 的 SLB 检测到所有 IP 均无法访问，SLB 的 http 检查失败。
3. GSLB 检测到 SLB 的 http 故障，修改负载均衡策略，分流业务到数据中心 B 的 SLB。
4. 客户端将 HTTP 请求发至 SLB 的 Virtual Server IP。
5. SLB 查询不到上次处理该会话的 Web 服务器，分发至数据中心 B 的一个 Web 服务器。
6. 新 Web 服务器查询不到故障节点的会话信息，客户 Session 将丢失，要求客户重新登陆。
7. 连接数据中心 B 的客户业务正常运行；连接数据中心 A 的客户 HTTP 会话不能保持，客户重新登陆。

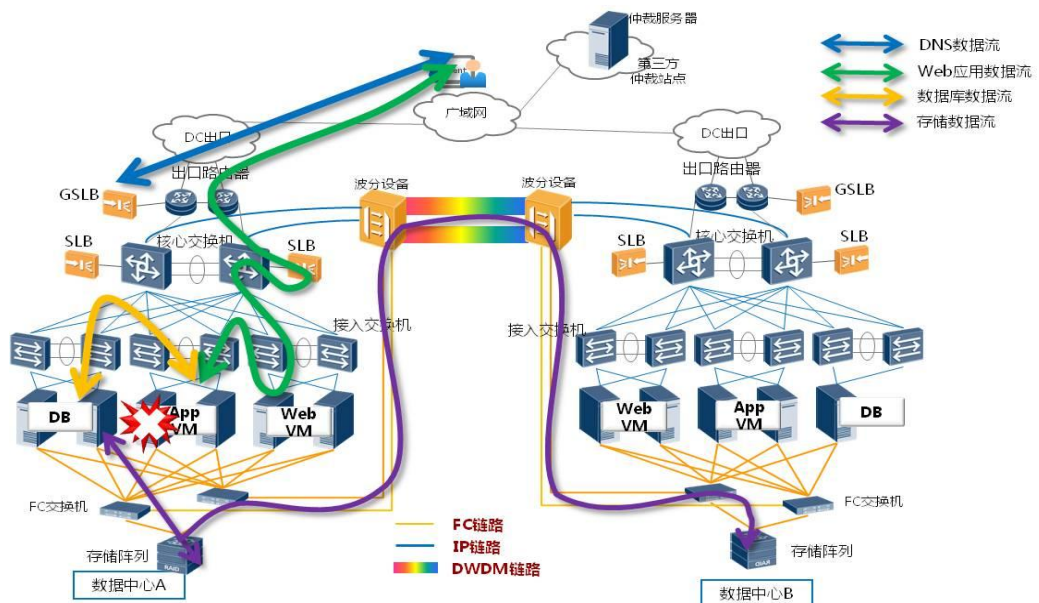
数据中心 A 的 Web 服务器故障恢复后，SLB 检测到后自动加入资源池，上层业务无影响，单个客户的 HTTP 会话能够保持。

5.4 应用服务器故障

一个站点内部署多台应用服务器，可以部署在物理机或者虚拟机上，站点内多台服务器配置为一个集群。

站点内一台应用服务器故障

图5-6 站点内一台应用服务器故障-数据流向示意图



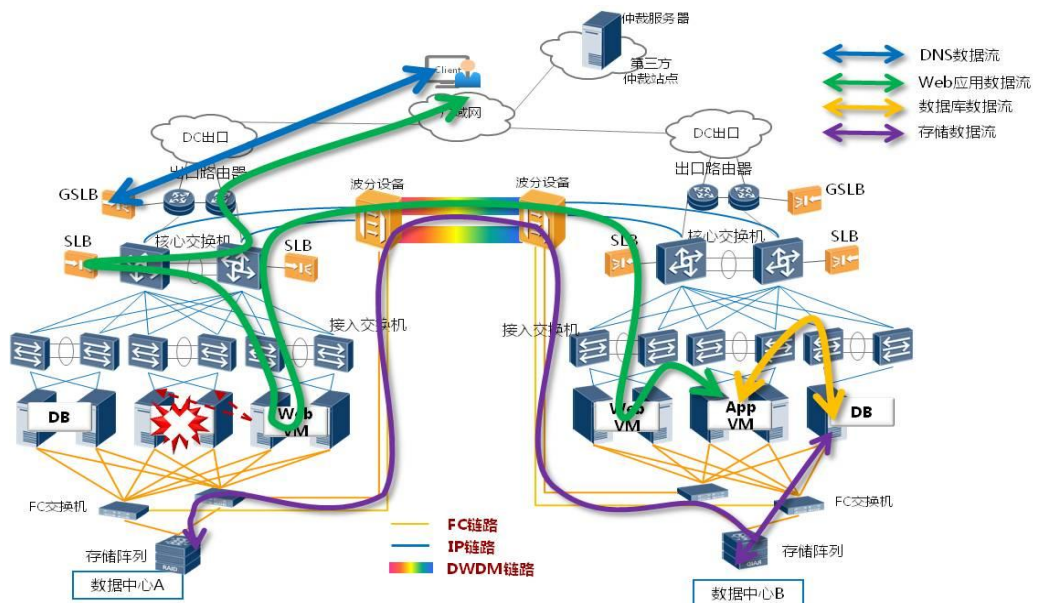
假如站点内一台应用服务器故障，处理过程如下：

1. 一台应用服务器故障，无法提供服务。
2. 客户端将 HTTP 请求发至 SLB 的浮动 IP。
3. SLB 根据查询到上次处理该会话的 Web 服务器节点，分发给原来的服务器。
4. Web 服务器检测到应用服务器节点故障，将后续 HTTP 请求发送到正常状态的节点；单个客户的 HTTP 会话能够保持。

数据中心 A 的应用服务器故障恢复后，上层业务无影响，单个客户的 HTTP 会话能够保持。

站点内所有应用服务器故障

图5-7 站点内所有应用服务器故障-数据流向示意



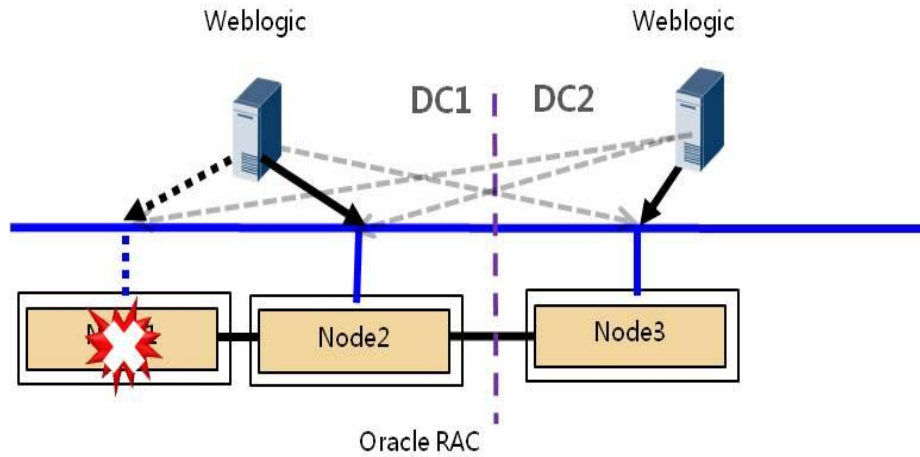
假如数据中心 A 的应用服务器全故障，处理过程如下：

1. 数据中心 A 的应用服务器无法提供服务。
2. 数据中心 A 的 Web 服务器探测到应用服务器无法访问，修改负载均衡策略，分流业务到数据中心 B 的应用集群。
3. 客户端将 HTTP 请求发至 SLB 的浮动 IP。
4. SLB 根据查询上次处理该会话的 Web 服务器，分发至该 Web 服务器。
5. 数据中心 A 的 Web 服务器探测到本中心的应用服务器均无法访问，修改负载均衡策略，分流业务到数据中心 B 的应用集群。

数据中心 A 的应用服务器故障恢复后，上层业务无影响，单个客户的 HTTP 会话能够保持。

5.5 Oracle RAC 故障

图5-8 Oracle RAC 故障示意



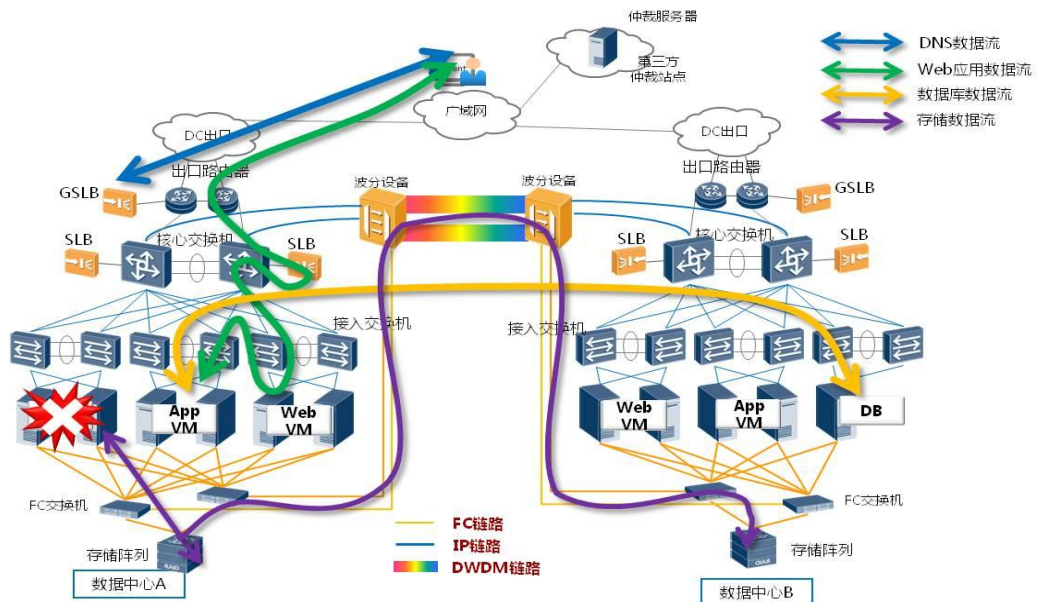
Oracle RAC 建议采用“2+1”的部署方式，数据中心 A 部署 2 台服务器，数据中心 B 部署 1 台服务器：

- 如果数据中心 A 的 1 台 Oracle 服务器故障，数据中心 A 内 Oracle RAC 单个数据库服务器故障，业务不切换到数据中心 B，业务在 DC A 内的另外一个数据库服务器上运行。
- 如果数据中心 A 的 2 台 Oracle 服务器故障，数据中心 A 内所有 Oracle RAC 数据库服务均短暂悬挂，随后，业务自动切换到数据中心 B 的数据库服务器上继续运行。

故障恢复后，数据库服务器故障修复后，自动加入集群，无需人工干预。

5.6 IBM DB2 故障

图5-9 IBM DB2 故障示意-数据流向图



IBM DB2 (PowerHA)建议采用“1+1”的部署方式，每个数据中心各部署 1 台服务器，故障及恢复过程：

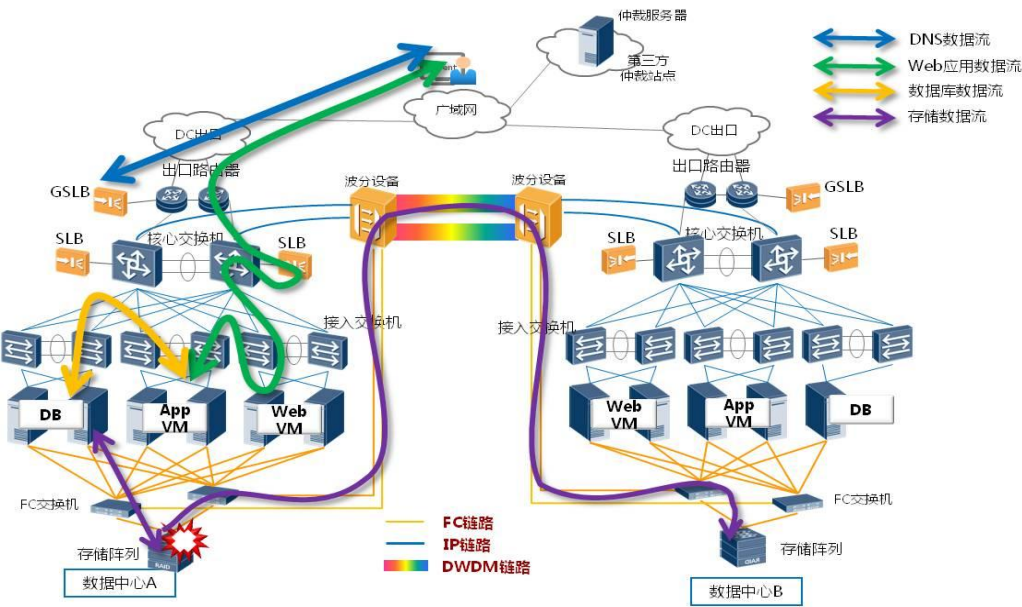
1. 数据中心 A 内 PowerHA 节点、网络或网卡发生故障，资源组自动切换到数据中心 B，DB2 数据库在数据中心 B 内运行。
2. PowerHA 为主备集群，当故障发生时，无法做到业务不中断，中断时间根据需要接管的资源组的个数相关。
3. 数据库服务器故障修复后，如果采用 rotating 的方式业务继续运行在数据中心 B。
4. 如果采用 Cascading 接管方式，资源组自动切换到数据中心 A，DB2 数据库在数据中心 A 内运行，业务中断，中断时间根据需要接管的资源组的个数相关。

本文介绍 IBM DB2 故障切换原理，MS SQLServer 原理类似。

5.7 阵列故障

阵列单控故障

图5-10 阵列故障示意-数据流向图



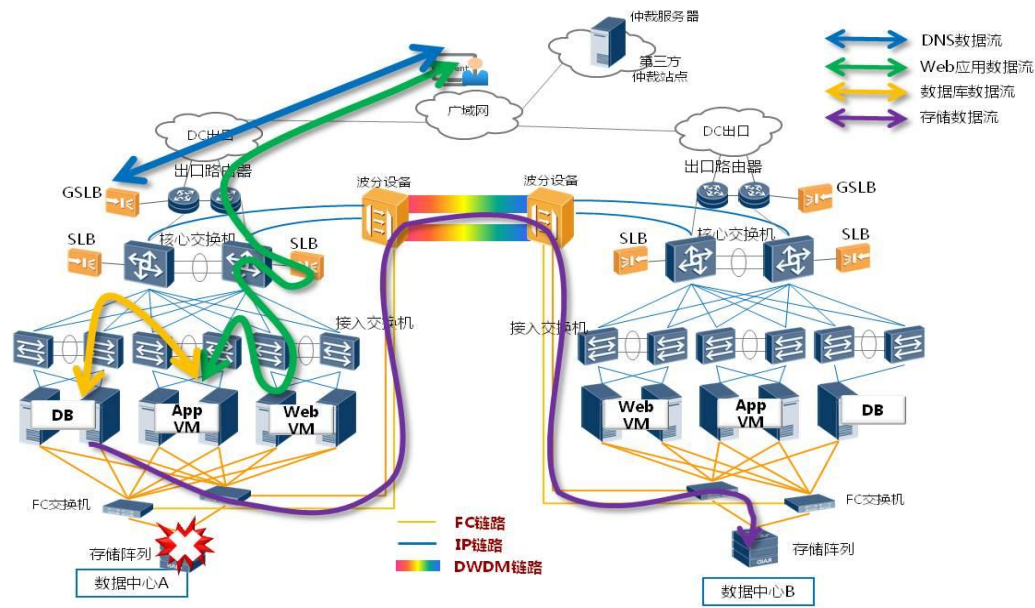
当一个数据中心阵列单控故障时，业务 I/O 自动切换到其它正常的阵列控制器处理，业务无中断。假设数据中心 A 的阵列单控故障，处理过程如下：

1. 数据中心 A 阵列单控掉电。
2. 多路径检测到单控路径中断，业务 I/O 自动切换至其它控制器继续运行。

数据中心 A 的阵列单控恢复上电后，业务自动重新负载均衡，上层业务无影响。

阵列双控故障

图5-11 阵列故障示意-数据流向图



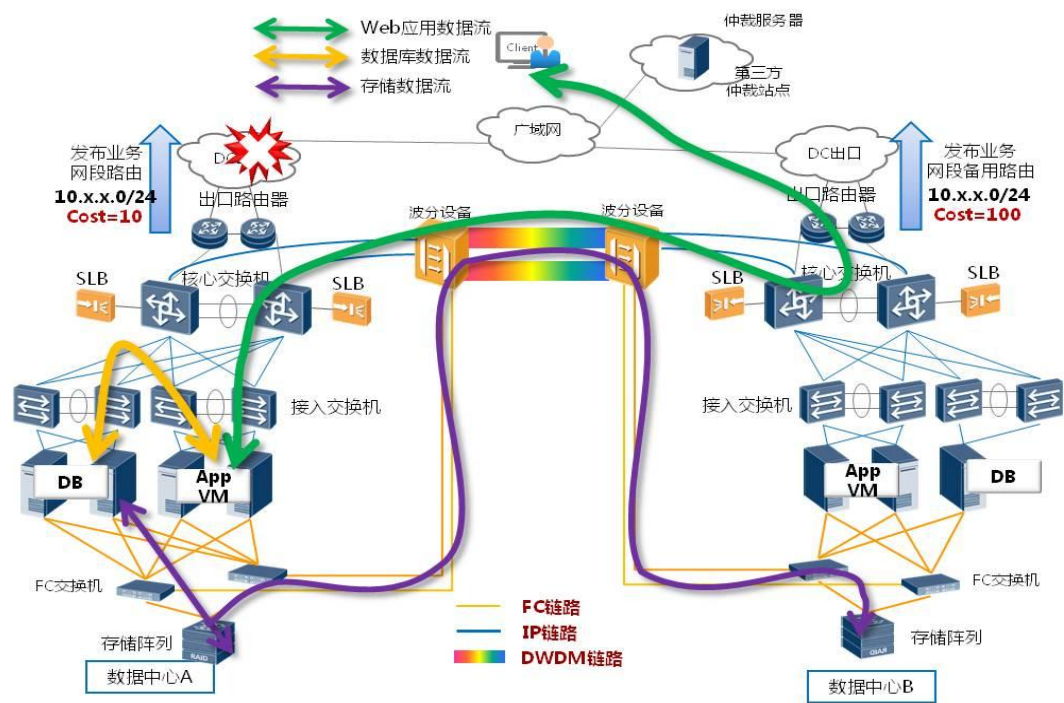
当一个数据中心阵列双控故障时，阵列心跳中断，发生仲裁，另一个数据中心获得仲裁，继续提供读写访问，业务无中断。假设数据中心 A 的阵列双控故障，处理过程如下：

- 1. 如果是数据中心 A 阵列主动通过管理软件下电，自身发送命令至数据中心 B 阵列，告知其接管业务。如果不是主动掉电，则会发生仲裁。
- 2. 数据中心 A 阵列同时停工所有双活 LUN。
- 3. 多路径置所有至数据中心 A 阵列的路径为不可用，所有 I/O 直接转发送数据中心 B 阵列。
- 4. 数据中心 B 阵列对新接收 I/O 记录差异位图。

数据中心 A 的阵列恢复上电后，双活关系自动恢复，根据差异位图盘的记录自动同步新增数据，上层业务无影响。

5.8 广域网链路故障

图5-12 广域网故障示意



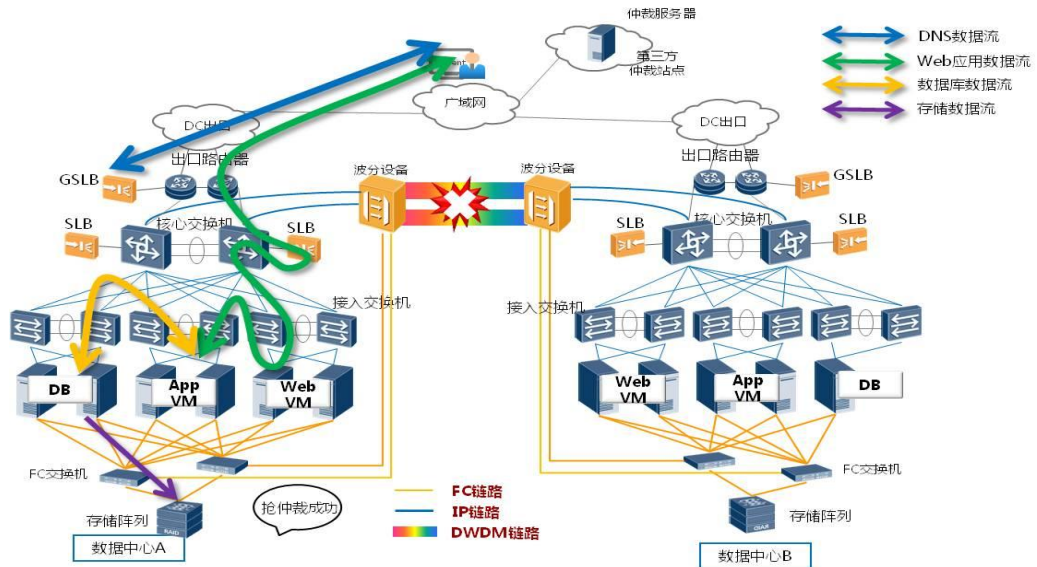
如果 DC 网关采用集中网关。正常情况下，数据中心 A 发布 cost 为 10 的网段路由，数据中心 B 发布 cost 为 100 的网段路由。故障发生，处理过程如下：

1. 数据中心 A 的主网段路由撤销，数据中心 B 发布 cost 为 100 的网段路由，经收敛后变为首选路由。
2. 客户端根据路由，访问数据中心 B，通过数据中心 B 二层到达数据中心 A 的 Web 应用。
3. 两个数据中心业务不发生切换，正常运行。

数据中心出口故障切换与广域网链路故障相同。

5.9 站点间链路故障

图5-13 站点链路故障示意-数据流向图

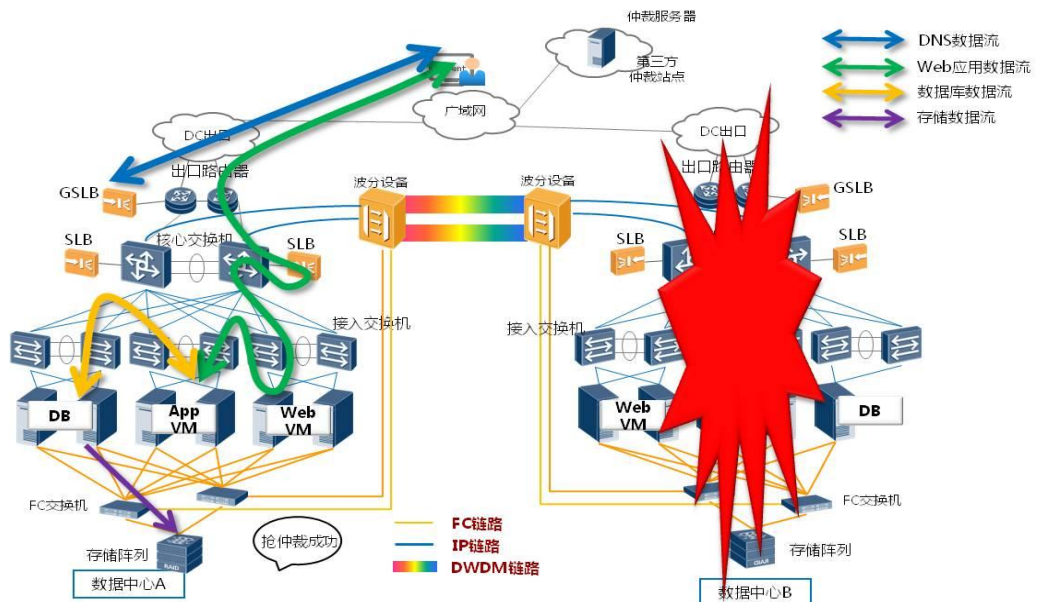


同城网络包括业务数据同步网络、阵列心跳网络和 Oracle RAC 私有网络，当同城网络故障时，两个数据中心的阵列均发现心跳中断，设置为优先的阵列抢占仲裁成功，接管所有的业务，另一套阵列停止提供业务，数据库同时发生仲裁，由于只有一个数据中心能提供 LUN 供读写，业务自动切换至该中心。详细处理过程如下：

1. 同城网络链路故障，阵列检测到心跳网络链路故障，阵列开始抢占仲裁。
2. 如果数据中心 A 的阵列仲裁抢占胜利，数据中心 B 的阵列停工所有双活 LUN。
3. 阵列将该数据中心 B 的 LUN 状态置为不可用，阵列双活关系故障。
4. 对于 Oracle 来说，数据中心 B 的服务器到数据中心 A 的阵列链路故障，业务 I/O 不能正常访问，数据中心 B 的服务自动切换到数据中心 A。
5. 数据中心 B 的 Web 和应用服务器无法访问数据库。
6. GSLB 的健康检测发现 Web 和应用服务器出现异常，修改负载均衡策略，不分发到数据中心 B。

5.10 站点故障

图5-14 站点故障示意



设备全冗余架构部署，当一个数据中心发生停电或火灾等灾难时，另一个数据中心阵列抢占仲裁胜利，接管所有的业务，业务自动切换。详细处理过程如下：

1. 同城网络链路故障，数据中心 A 的阵列检测到心跳网络链路故障，阵列开始抢占仲裁。
2. 数据中心 A 的阵列无法访问到数据中心 B 阵列，阵列将该数据中心 B 的 LUN 状态置为不可用，阵列双活关系故障。
3. 对于 Oracle 来说，数据中心 B 的服务自动切换到数据中心 A。
4. GSLB 的健康检测发现 Web 和应用服务器出现异常，修改负载均衡策略，不分发到数据中心 B。

6

缩略语

缩略语	英文解释	中文解释
VIS	Virtual Intelligent Storage	虚拟化智能存储
COW	Copy on Write	写前拷贝
FC	Fibre Channel	光纤通道
IP	Internet Protocol	Internet 协议
iSCSI	Internet SCSI	互联网小型计算机系统接口
LUN	Logical Unit Number	逻辑单元号
NAS	Network Attached Storage	网络附加存储
DWDM	Dense Wavelength Division Multiplexing	密集型波分复用
PCI	Peripheral Component Interconnect	外围设备互联
RAID	Redundant Array of Independent Disks	独立磁盘冗余阵列
SAS	Serial Attached SCSI	串行 SCSI
SATA	Serial Advanced Technology Attachment	串行 ATA
SCSI	Small Computer System Interface	小型计算机系统接口
TCP	Transmission Control Protocol	传输控制协议
UPS	Uninterruptible Power Source	不间断电源
RAC	Real Application Clusters	实时应用集群

缩略语	英文解释	中文解释
ASM	Automatic Storage Management	自动存储管理
FEC	forward error correction	前向纠错
DCM	dispersion compensation module	色散补偿模块
FBG	fiber Bragg grating	布拉格光栅
DCF	dispersion compensation fiber	色散补偿光纤
RPO	Recovery Point Objective	恢复点目标
RTO	Recovery Time Objective	恢复时间目标
TAF	Transparent Application Failover	透明应用程序故障转移