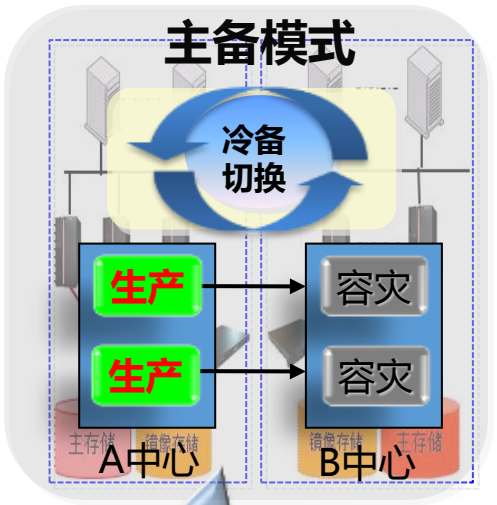
DBAplus

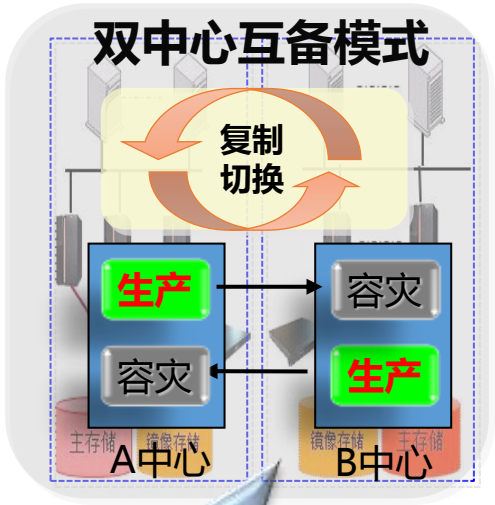
畅谈双活数据中心技术

朱祥磊

常见容灾模式

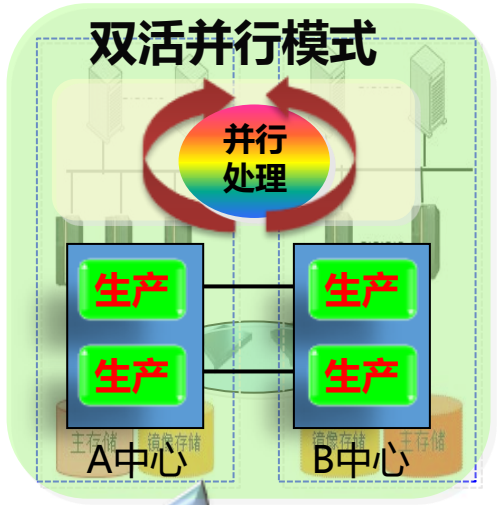


1、故障下需要定位+决策+切换流程，超过0.5小时
2、容灾侧资源闲置
3、SRDF、PPRC等技术



1、故障下需要定位+决策+切换流程，超过0.5小时
2、动态资源管理技术，容灾资源闲置
3、 SRDF、PPRC等技术

传统方案：“主备”模式或“互备”双中心模式

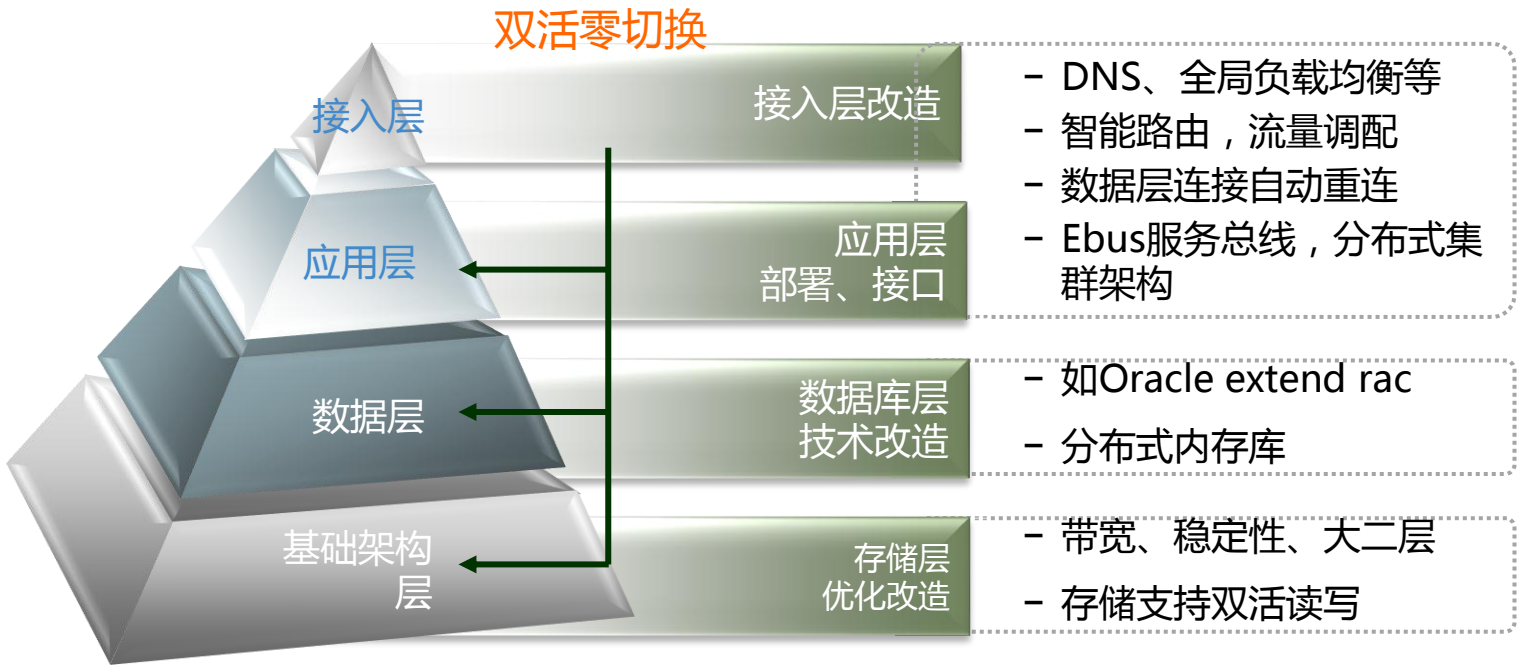


两个中心平时均可承担业务，同时对外服务，坏掉任何一方不影响。

“双活并行处理”模式，做到准0切换

双活应用架构

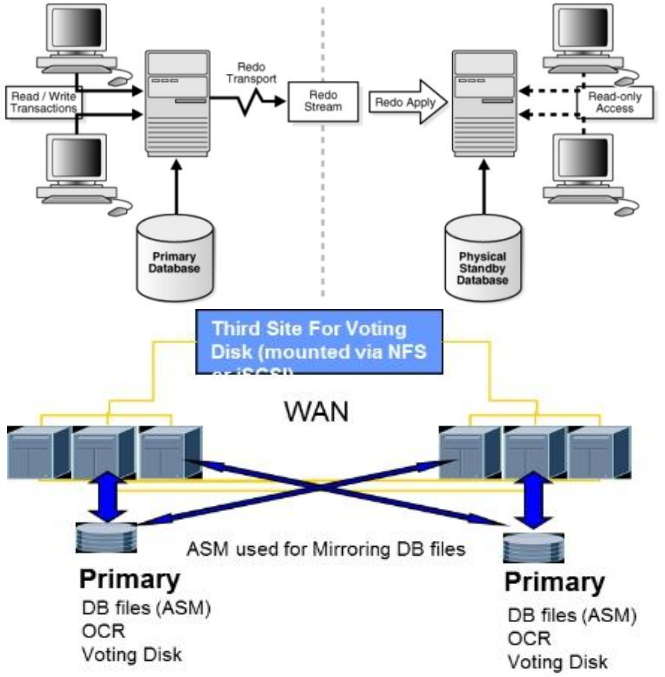
- 接入层：借助DNS、全局负载均衡等技术实现双活接入和智能路由，流量调配
- 应用层：基于开放分布式集群架构、或服务总线技术
- 数据层：需要构建双中心同时可读写的机制，如远程RAC
- 基础架构层：网络上对稳定性和带宽吞吐性能要求更高，甚至需要打通跨中心的大二层网络。存储方面，则需改变一主一备的读写机制，实现同时可读写。



- 数据层
- 存储层
- 接入/应用层
- 虚拟化/云平台
- 技术关键点

数据层双活三种方式

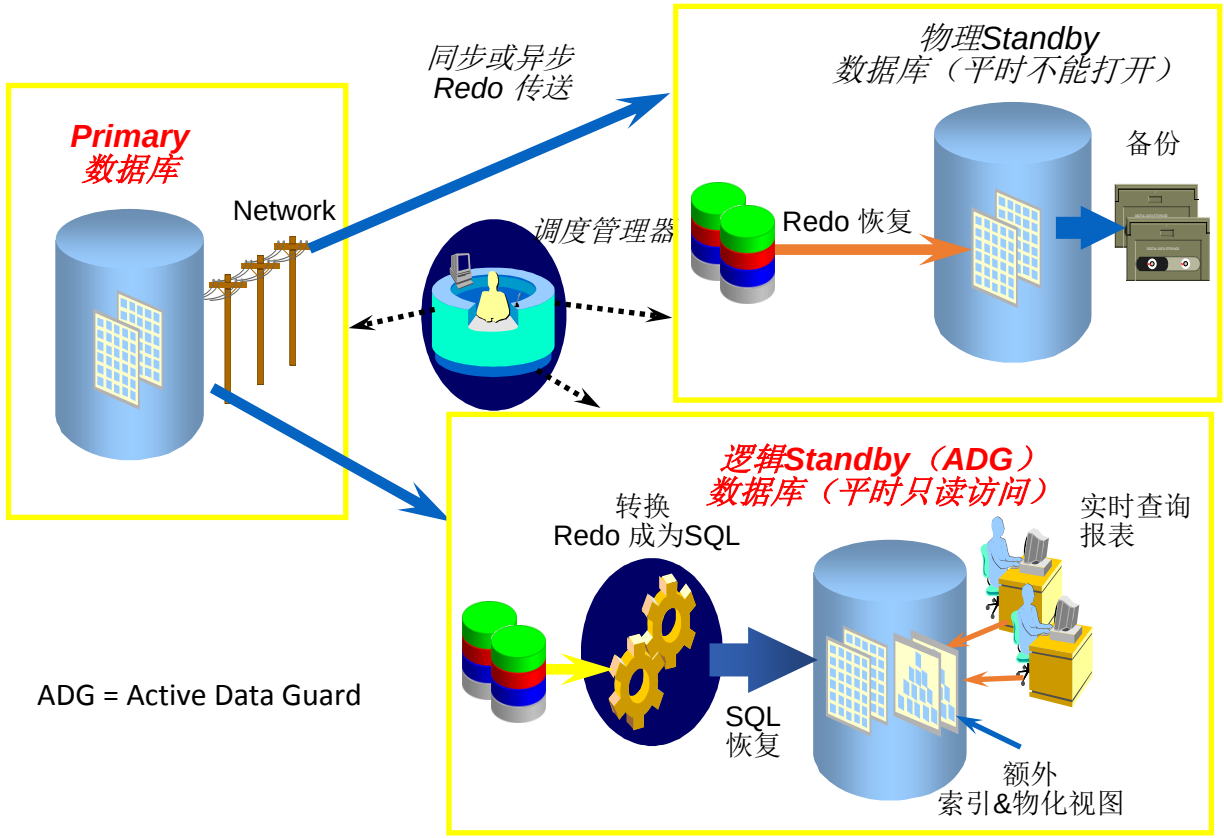
- **Active-Standby方式：**基于Oracle ADG技术，采用从主库向备库传输redo日志方式，备库恢复数据过程可以用只读方式打开进行查询操作，实现了部分双活功能，在主节点故障后可以将备节点切为
- **Active-Active方式：**通过Oracle Extend RAC实现多个集群节点同时对外提供业务访问。该方式做到故障无缝切换，提升应用系统整体性能。



- **数据逻辑复制软件方式：**通过实时抽取在线日志中的数据变化信息，然后通过网络将变化信息投递到目标端，最后在目标端还原数据，从而实现源和目标的数据同步。

方式1、oracle ADG

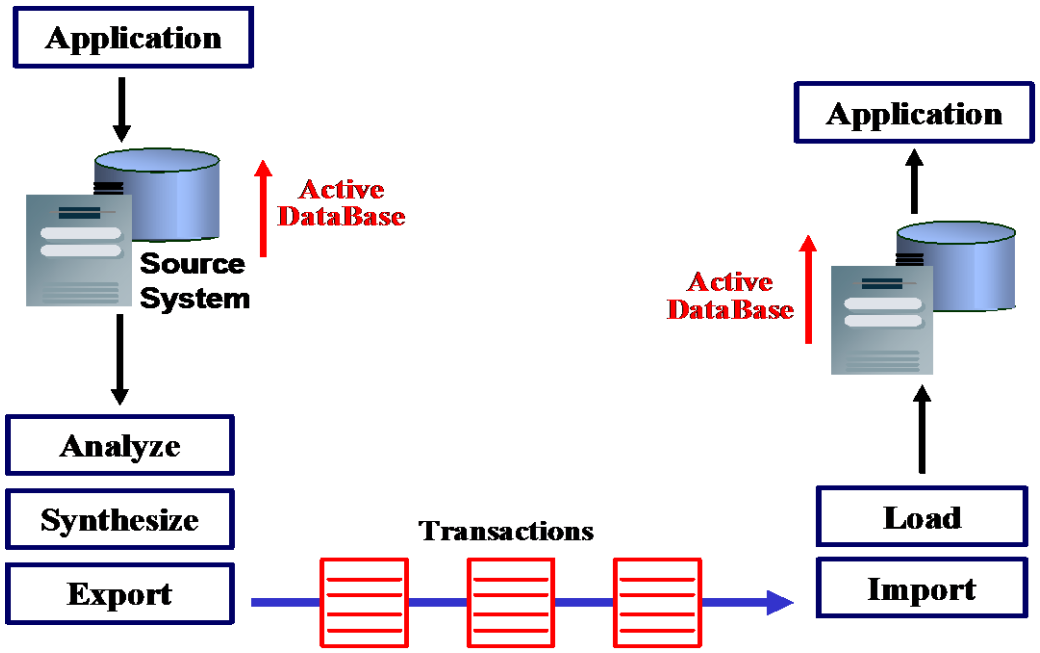
- 通过网络从生产向容灾传输归档或redo日志，容灾端恢复方式同步数据。
- Oracle 11g 以后容灾库可打开为只读模式，容灾切换时能快速alter为读写状态。
- 存储支持异构，OS需要同构
- 应用场景：
  - 作为应急或容灾
  - 作为读写分离
  - 作为数据保护手段（结合flash DB）



ADG = Active Data Guard

方式2、逻辑复制

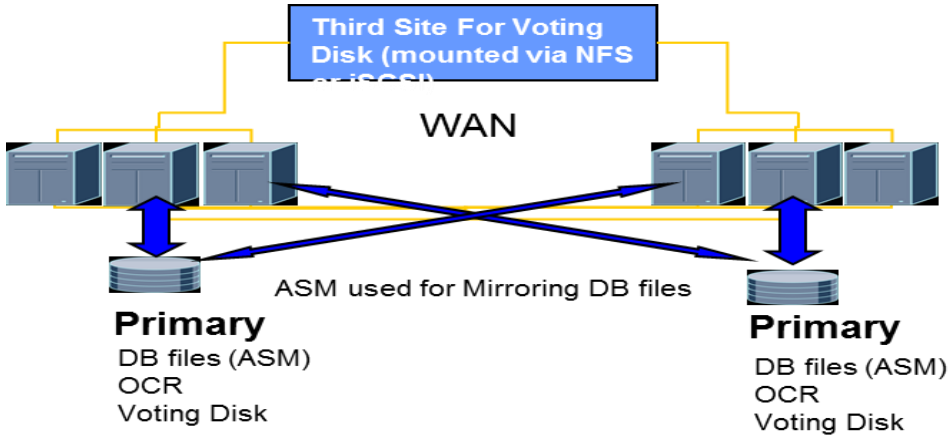
- 通过DSG、GoldenGate等逻辑复制技术实现跨中心数据库的相互复制
- 支持表级的复制
- 两个数据中心各建一套数据库，物理独立，同时能读写
- 基于数据库日志准实时复制数据
- 支持异构数据库、异构OS
- 可以实现一对一、一对多、多对一、双向复制等多种拓扑结构



Dsg工作原理

方式3、Oracle 远程RAC

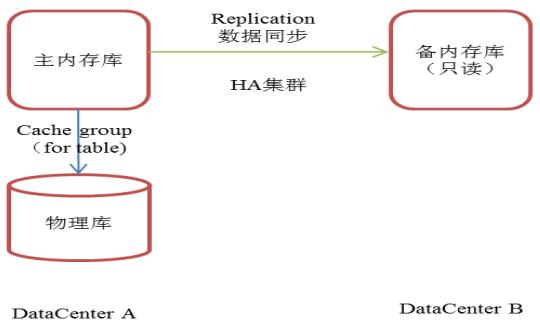
- Oracle Extended RAC以跨中心共享存储为基础，通过共享存储资源和Oracle Clusterware数据库集群管理，实现各个中心节点对数据库并行访问。
- 共享存储可以采用存储自身数据复制技术，存储虚拟网关或远程卷管理等技术，左图以Oracle ASM存储卷管理为例，实现数据的双向实时复制。
- ASM支持对本地磁盘的优先读取，避免跨数据中心的数据读取，提高I/O性能并减少网络流量；



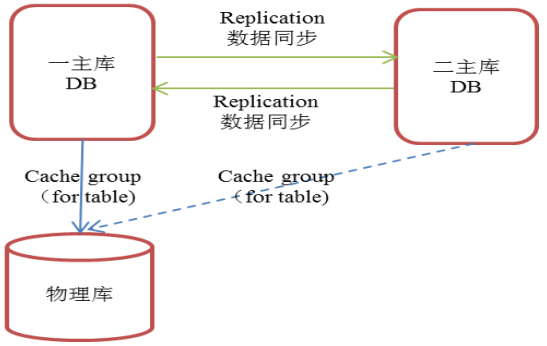
要点（ASM）：
•两个数据中心分别部署一套存储，各提供一套LUN设备给全部数据库主机。
•存储的SAN网络和RAC心跳网络需使用低延迟、高带宽的DWDM光纤链路。
•配置ASM磁盘组。每个磁盘组配置两个失效组，每个失效组对应来自一套存储的LUN设备。
•在第三个站点部署用于RAC的第3个投票盘，使用NFS的方式挂载到所有数据库主机。
•与管理普通的RAC系统类似，需要**重点加强对站点间光纤链路情况的监控与应急**。

内存库双活技术

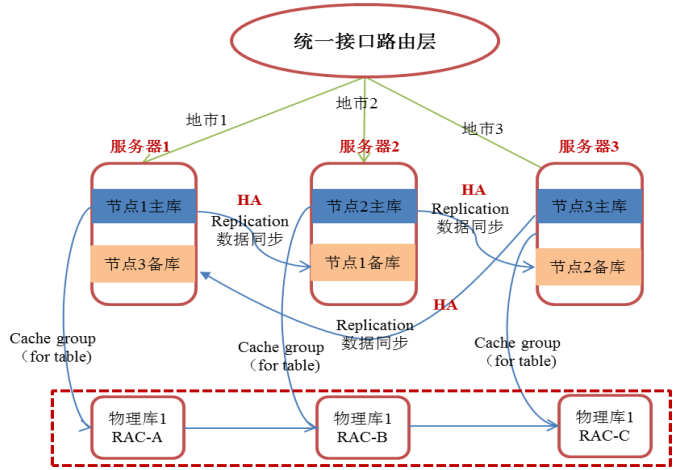
- 将数据常驻在内存中直接操作的数据库。
- 相对于磁盘，内存的数据读写速度要高出几个数量级，将数据保存在内存中相比从磁盘上访问能够极大地提高应用的性能
- 应用场景：用于实时计费、读写分离场景，主要有Oracle Times Ten，Altibase商用以及华为、亚信和斯特奇等自研产品。
- 内存库集群部署主要有HA模式，双活模式，线性拆分和分布式集群四种模式。



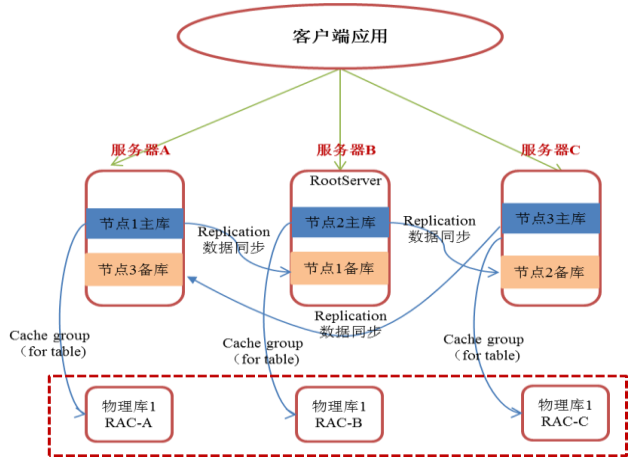
HA模式



双活模式



线性拆分模式

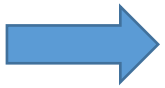


分布式集群模式

数据层双活技术比较

	技术种类	技术特征	数据一致性	双活读写	数据延迟	可维护性	可靠性	适用模式
物理库	基于数据逻辑复制软件	DSG、gg、shareplex等	逻辑错误会导致不一致，无法稽核	支持	存在延迟（和日志量有关）	较差，系统变更需要人工介入	较好，支持多线程，不影响生产，需定期重新同步	数据一致性要求较低或基于表的同步
	基于数据库自身	oracle active dataguard	一致（前提正常同步）	不支持	存在延迟（和日志量有关）	维护简单，支持线性扩展	较好，同步效率高，快速切换	读写分离场景
		Oracle Extended RAC	一致	支持	实时同步，没有延迟	较好	较好	核心系统对稳定性较高
内存库	HA模式	基于日志实时或异步同步	存在不一致风险	不支持	存在延迟	较好	一般	适合物理库较小
	双活模式	基于日志实时或异步同步	存在不一致风险	支持	存在延迟	较好	一般	
	线性拆分	基于日志实时或异步同步	存在不一、致风险	不支持	存在延迟	较差，复杂	一般	适合物理库较大
	分布式集群	基于日志实时或异步同步	一致	支持	存在延迟	较好	较好	适合核心系统

建议：在实际使用中应根据具体情况选择合适的方案，理论上只有Extended RAC为真正的双读双写

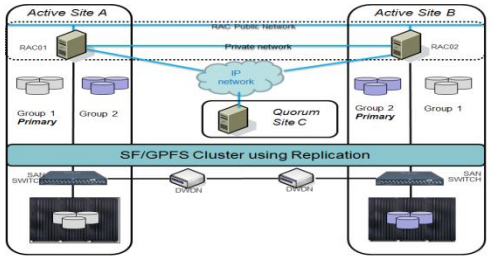


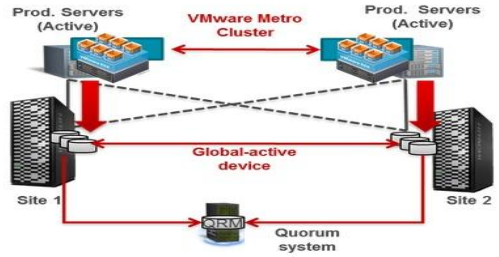
- 数据层
- 存储层
- 接入/应用层
- 虚拟化/云平台
- 技术关键点

存储双活流派

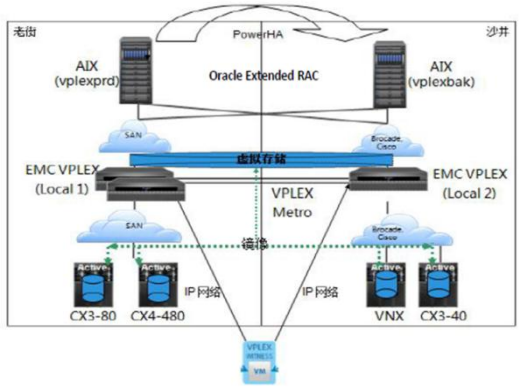
- 存储层作为双活系统核心基础架构平台，其双活技术在整个架构中起到关键作用，目前基于存储层双活方案主要有下面三种：
  - 基于远程卷管理软件的虚拟化，如：Symantec SF，IBM GPFS， Oracle ASM等
  - 基于存储网关虚拟化，如：EMC vplex、IBM SVC
  - 基于存储自身卷镜像技术，HDS GAD 、Huawei等

1. **卷管理软件虚拟化**：通过安装在主机上卷管理软件的逻辑卷镜像技术实现底层数据逻辑同步。





2. **存储网关虚拟化**：在每个站点新增存储虚拟化网关设备组成跨站点集群,并对存储卷进重新行封装，对外提供主机I/O访问。

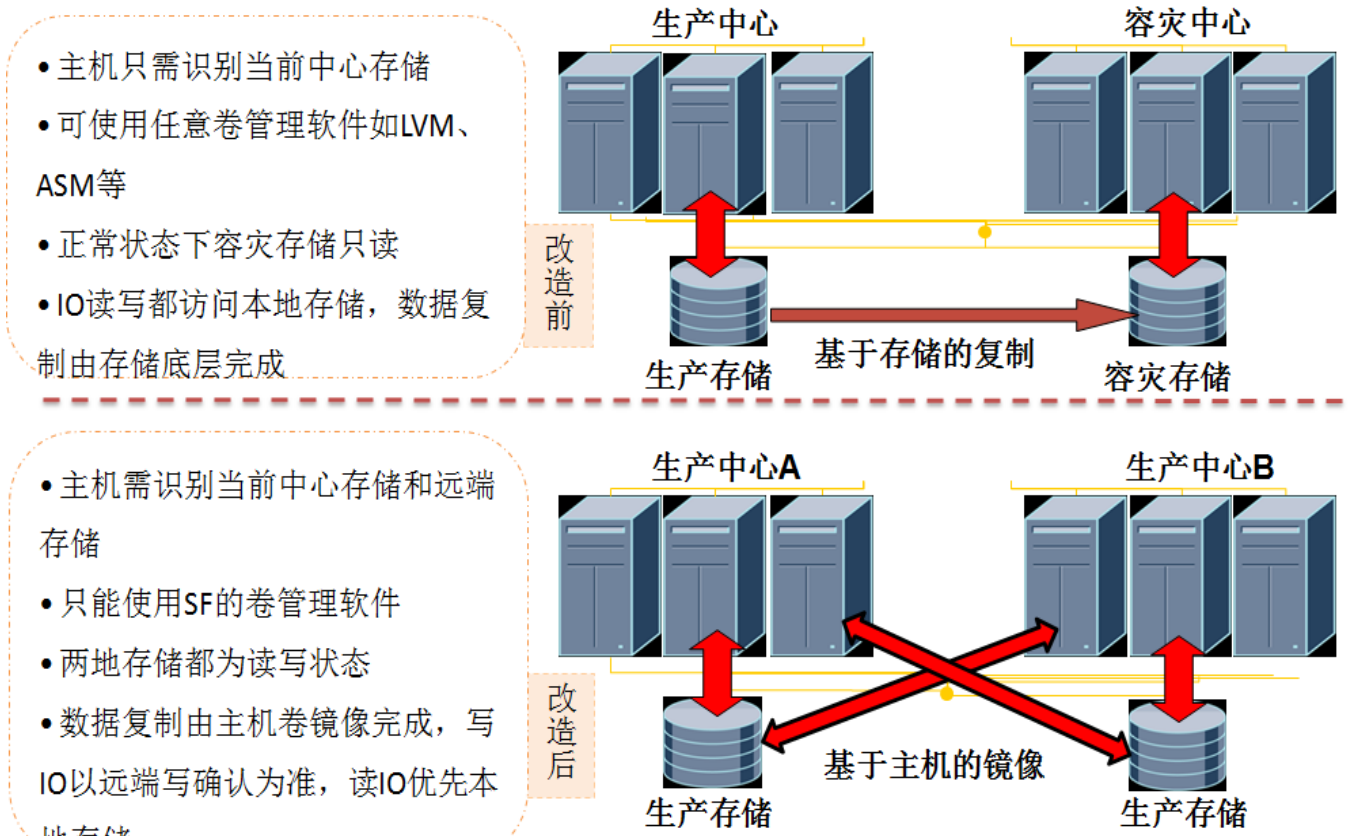


3. **存储卷镜像技术**：将两套磁盘阵列组成一个集群，两台存储上的LUN被虚拟化为一个虚拟卷，主机写操作通过卷虚拟化镜像技术同时写入两个数据中心的存储设备，保证站点之间数据实时同步。

流派1、远程卷管理软件

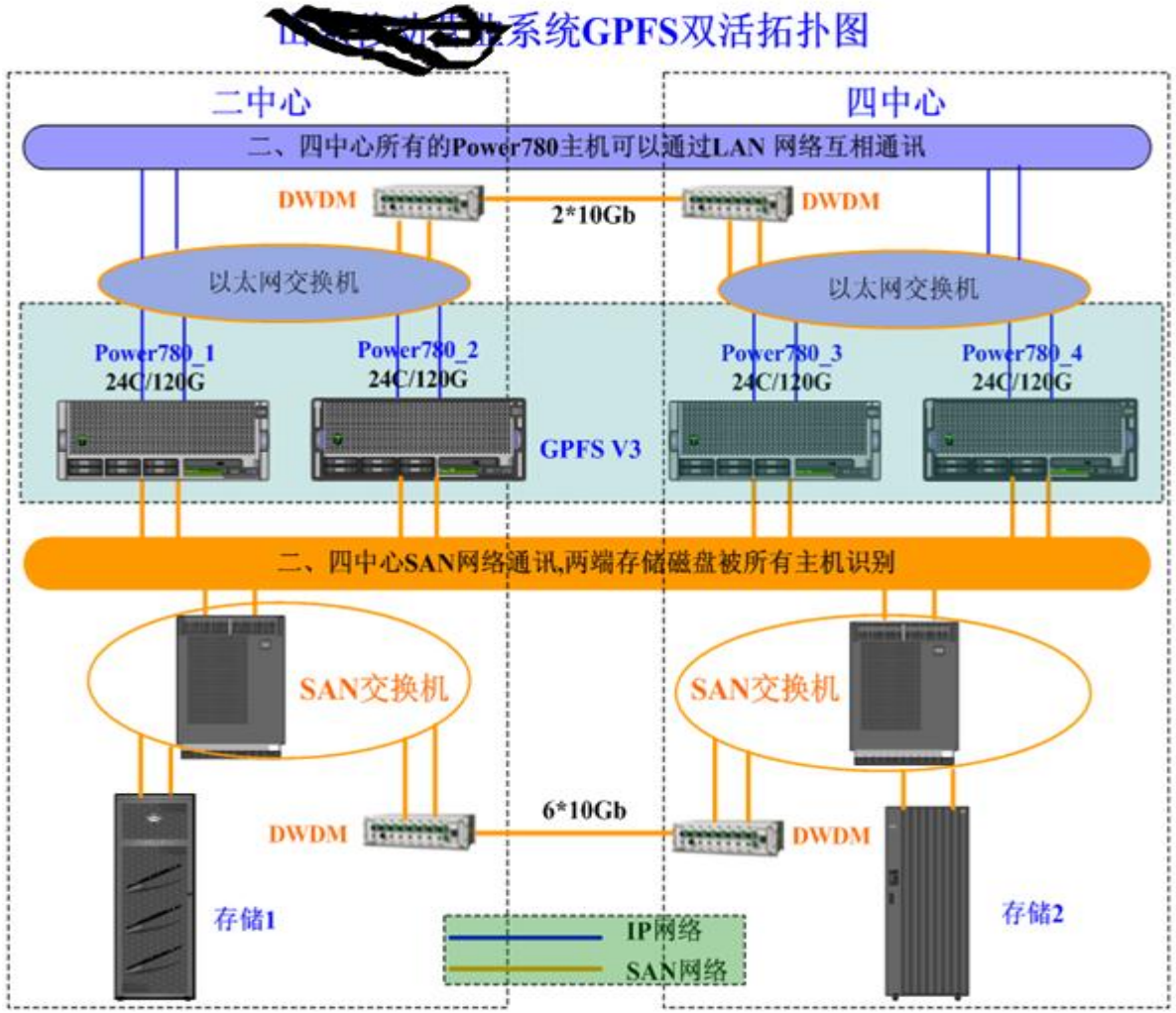
- **数据同步**：底层数据复制采用远程卷管理软件，如赛门铁克的storage Foundation（SF）、IBM的GPFS、Oracle的ASM等，通过逻辑卷镜像技术实现底层数据逻辑同步。上层应用采用Oracle Extended RAC方案实现远程多节点RAC,使生产和容灾节点都处于在线状态，应用逻辑访问的是同一个数据库。
- **数据读写**：支持双读写。
- **数据一致性**：完全一致。

■ 远程卷管理软件改造前后变化



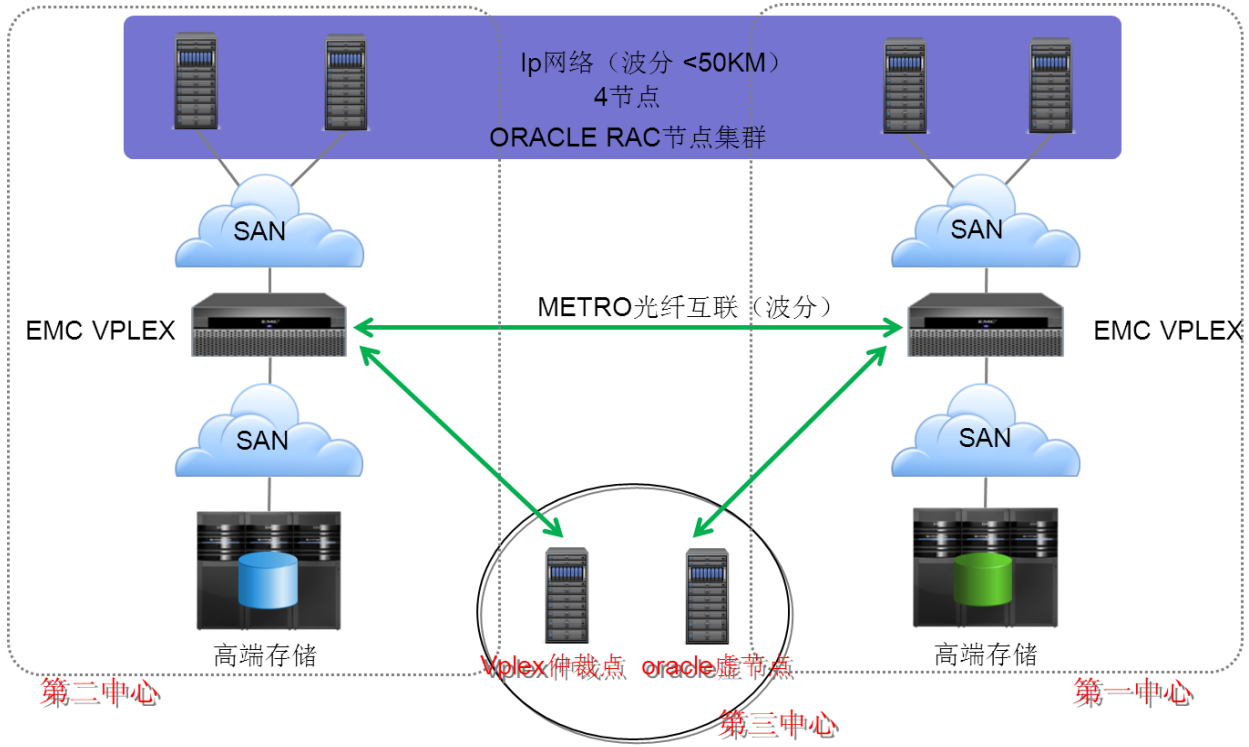
流派1、远程卷管理软件

- 案例： IBM GPFS+oracle 11g Extended RAC
- 应用场景：一边承担生产，另一端承载统计分析和查询
- 实施要点：
  - 网络改造：需要打通两个中心间大二层网络。
  - 底层存储链路改造：需要认到对端机房存储，带宽要求高。
  - 提供可靠性较高的二层网络（心跳网络）
  - 提供可靠性较高的共享存储（投票盘）
  - 对底层链路和距离要求高：距离太远会导致响应变慢，官方建议50KM之内。



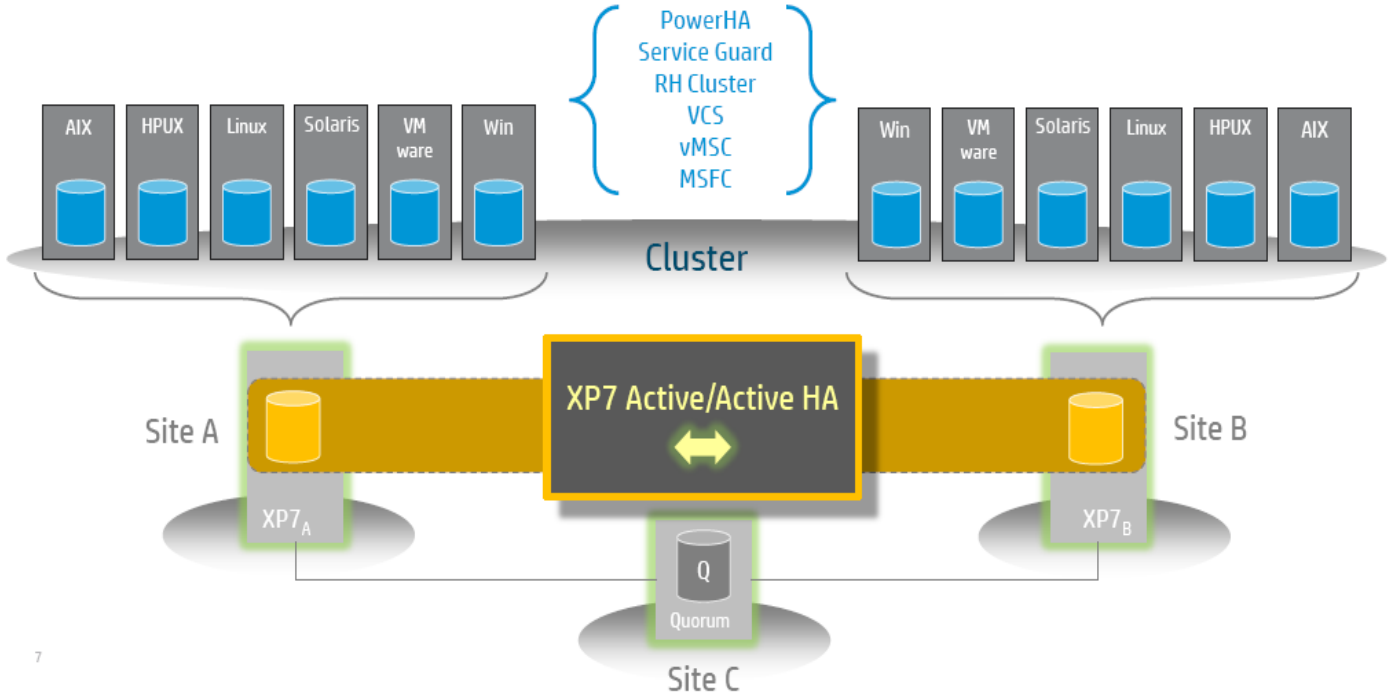
流派2、存储网关虚拟化

- 实现原理：将存储虚拟化技术(EMC的vplex)和Oracle的远程RAC技术结合，实现跨中心的数据双活访问。
- 跨中心的两个存储通过网关设备虚拟成一个对外访问，内部实时同步，保持数据的一致性，平时两边主机分别访问本地存储，故障情况下可跨中心访问对方存储。
- 对于同一个数据块的读写冲突机制，是由Oracle RAC来保证的。
- 具备脑裂预防服务器“witness”：witness是VPLEX的仲裁装置；



流派3、基于存储自身卷镜像

- 不需要额外软硬件，需要采用特定高端存储设备，如VSP、XP7以上才可以。
- 存储网络架构没有改变，易于实行。
- 两边存储可以同时读写。
- 上层需要结合Oracle远程RAC实现双活



存储层双活技术对比

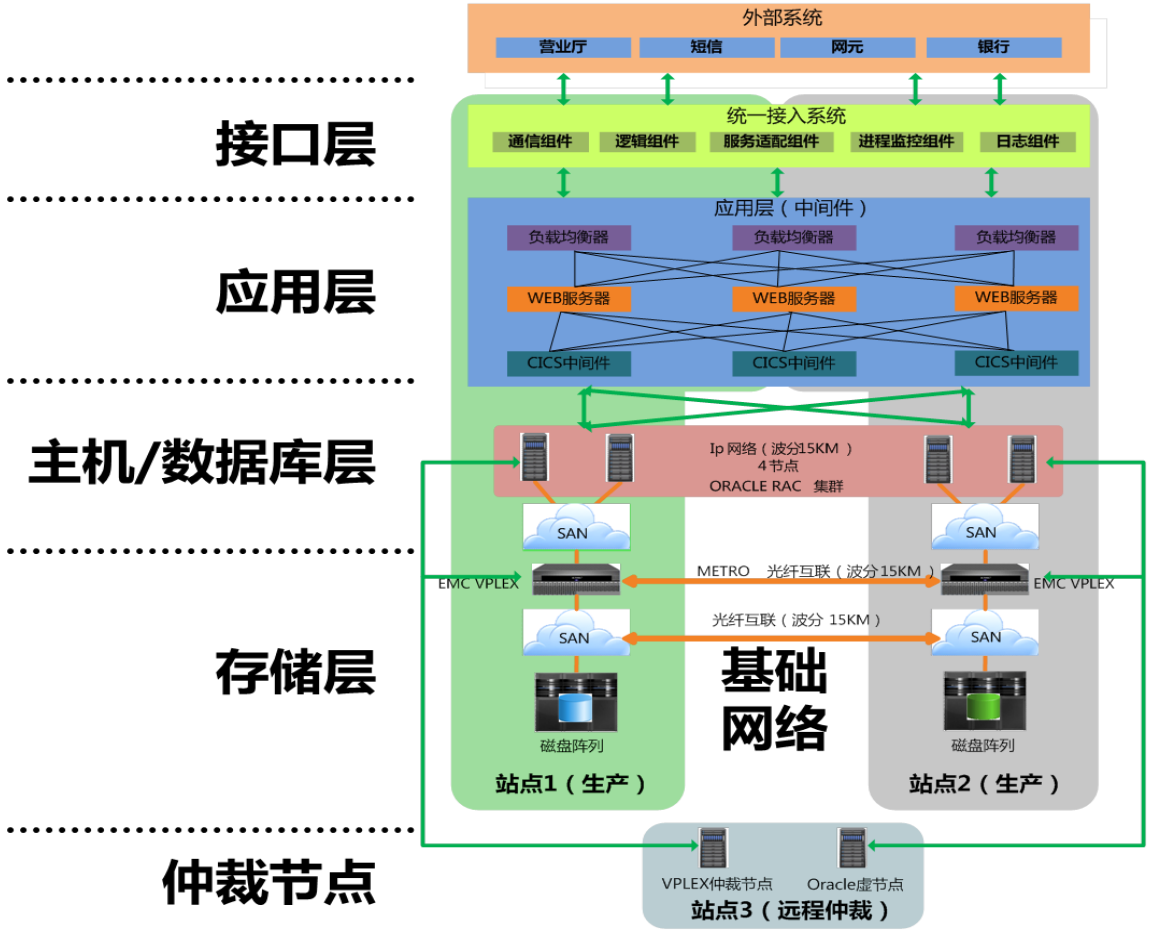
技术特征	技术特征	数据一致性	双活读写	可靠性	异构性	投资成本	优缺点
基于远程卷管理（软件虚拟化）	Symantec SF AIX LVM IBM GPFS Oracle ASM	RPO=0	支持	较差	支持异构	成本较低	
基于存储网关虚拟化	EMC Vplex IBM SVC 华为 VIS 飞康 NSS	RPO=0	支持	较差	支持异构	较高	组网复杂，可靠性差，数据同步性能差
基于存储卷镜像（存储自身虚拟化	**HDS GAD Huawei OceanStor V3**	**RPO=0**	**支持**	**较好 RTO=0**	**不支持**	**较低**	**组网简单，维护方便，但技术较新，实用经验少**
基于存储HA机制	IBM powerHA HyperSwap、日立的HAM	RPO=0	支持	较差	不支持	投资较高	采用高端存储虚拟化软件，有一端远程读写效率低

整体看红色为最优方案，但要根据实际情况选择，上述方案均需要Extend RAC支持。

- 数据层
- 存储层
- 接入/应用层
- 虚拟化/云平台
- 技术关键点

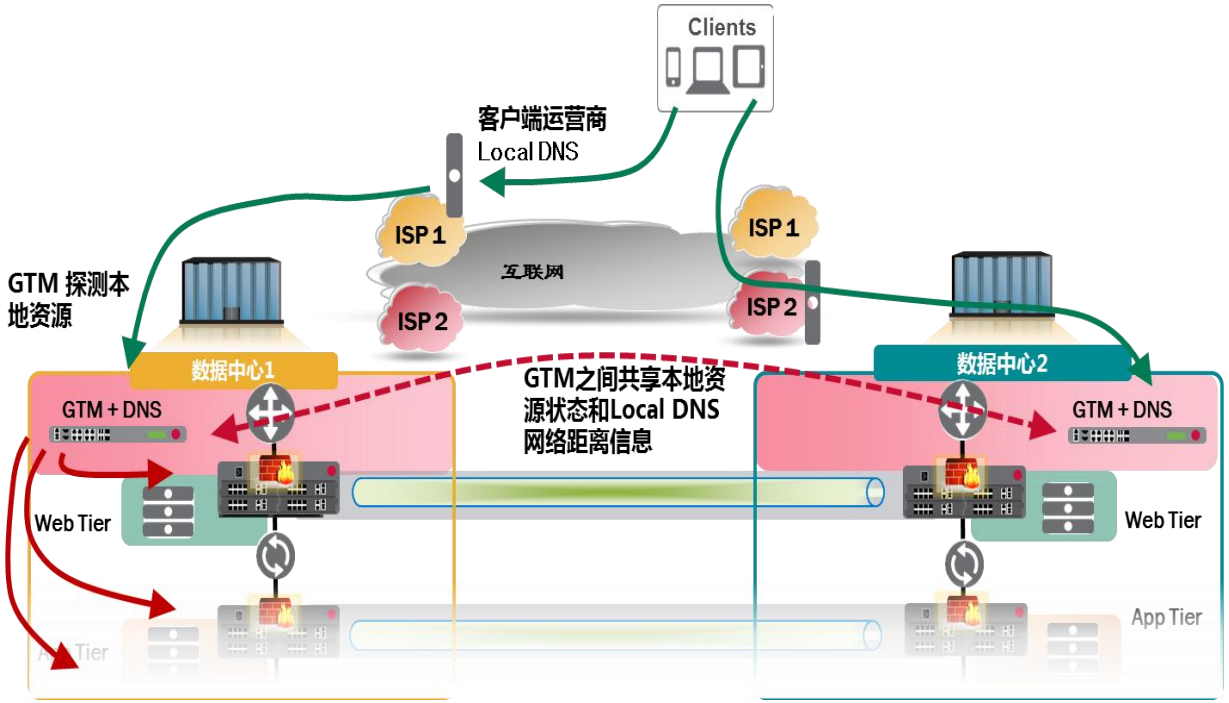
应用层双活要点

- 双活需要从接入、应用层、数据连接等层面考虑实现，才能实现“零”切换。
- 应用支持：建议构建统一管理的接口层或采用服务总线技术
- 实现应用自动重连机制，确保自动切换，减少人工切换。---支持数据库切换后应用的正常运行
- 双中心部署相同的应用集群方式，或跨中心的集群



接入层技术

- 采用全局负载均衡（如F5的GTM）、DNS、或前置CDN等技术实现跨中心灵活接入。
- 1、业务多中心并行模式：通过一组GSLB来对外提供服务，GSLB监控服务的状态，并通知组内其他设备，对于每一个DNS请求返回最佳结果，好的策略选择和配置方式可以最大幅度提高客户体验。
- 2、业务多中心互备模式：对于内网业务通过一组SLB来提供服务，实现DNS解析，负载分发和故障切换。



➢**应用双活：**当单数据中心出现故障时，可以将请求引导向另一个可用的数据中心，实现双活高可用。

➢**智能流量控制：GSLB**根据后端服务器负载和链路状况实现不同站点间流量调配，链路优选，保证用户访问最佳性能服务器，确保访问质量，提升用户感知。

- 数据层
- 存储层
- 接入/应用层
- 虚拟化/云平台
- 技术关键点

云架构下的双活

云化后的变化：

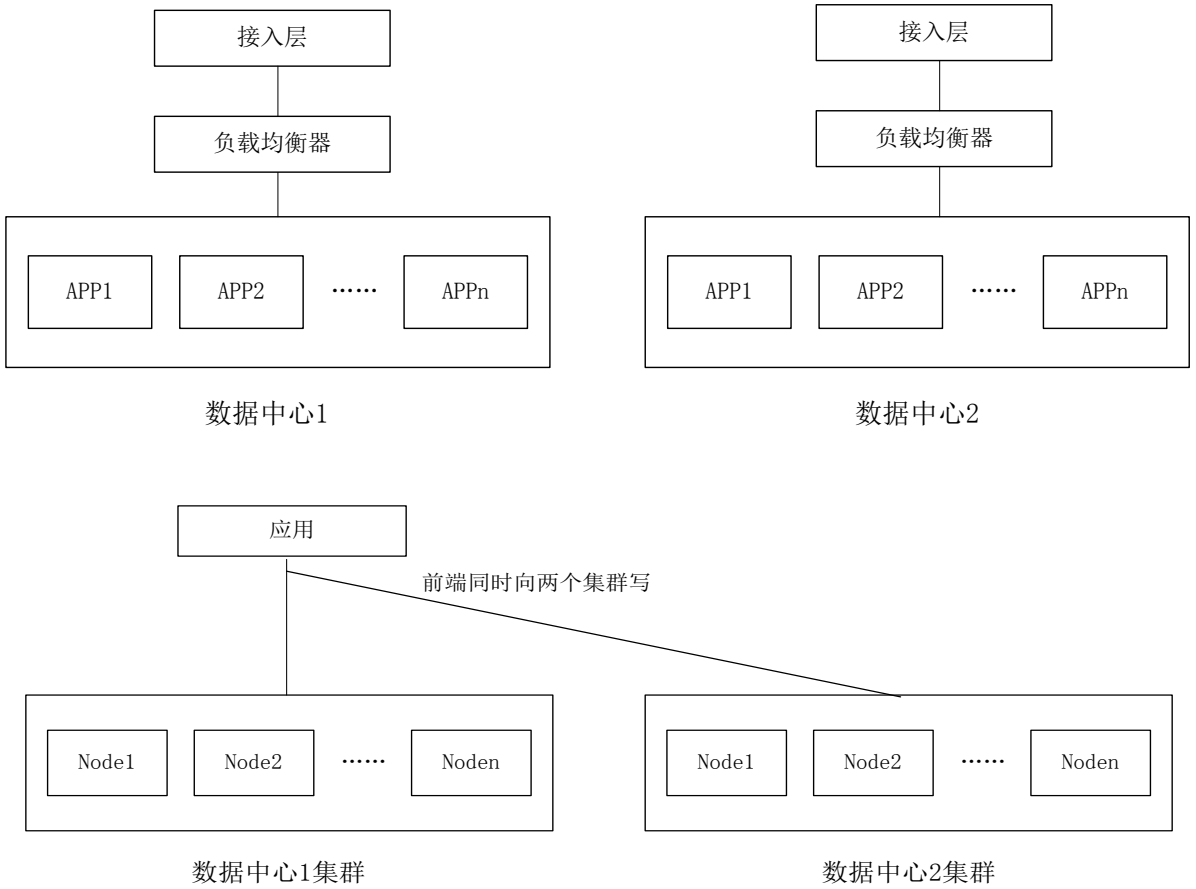
- 虚拟化技术
- 应用实现集群化和x86化

带来问题：难以沿用原有的双活设计方式，需要新考虑考虑集群化的业务连续性双活方案。

云化双活分类	方案描述	场景举例
传统基于负载均衡的双活架构	每个中心部署独立的云化应用集群，通过接入层负载均衡实现双活	Web集群等
基于分布式应用协调机制	构建一套跨中心应用集群，通过分布式应用协调如zookeeper实现跨中心的高可靠性集群，统一配置、统一管理和任务分配。	EBUS跨中心双活应用集群、分布式缓存等
hadoop、mpp等的双活机制	1、应用写两份方式实现双活 2、跨中心集群方式	大数据
虚拟化平台的跨中心双活（迁移）	1、跨中心虚拟机集群，可平滑迁移 2、每个中心一套集群，通过接入层构建负载均衡实现双活	云资源池

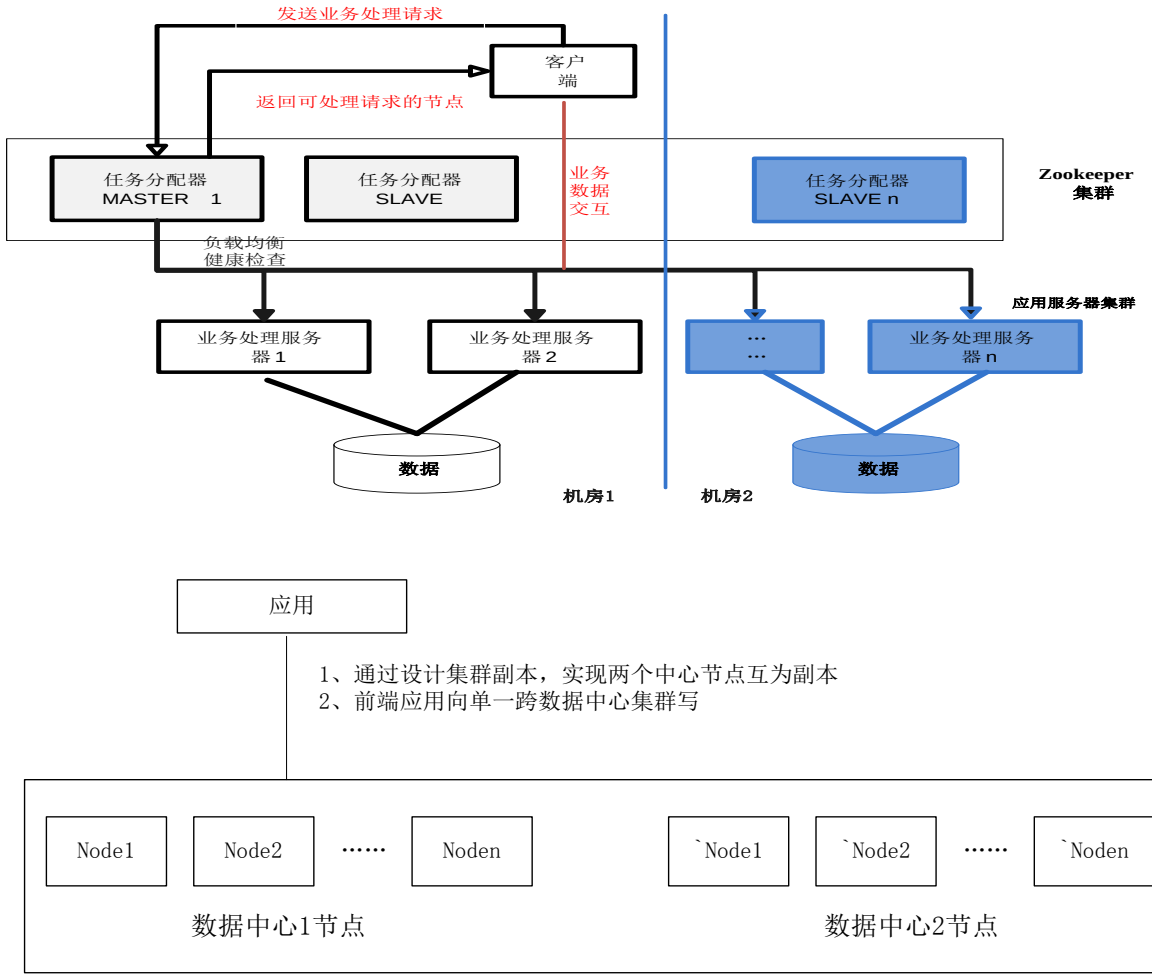
模式1、相互独立的双集群

- 在每个中心部署独立的云化应用集群
- 1、如Web类应用可通过接入层和负载均衡实现双活访问，
- 2、如hadoop或MPP集群应用可通过上层应用实现双集群数据同步，从而实现双活。



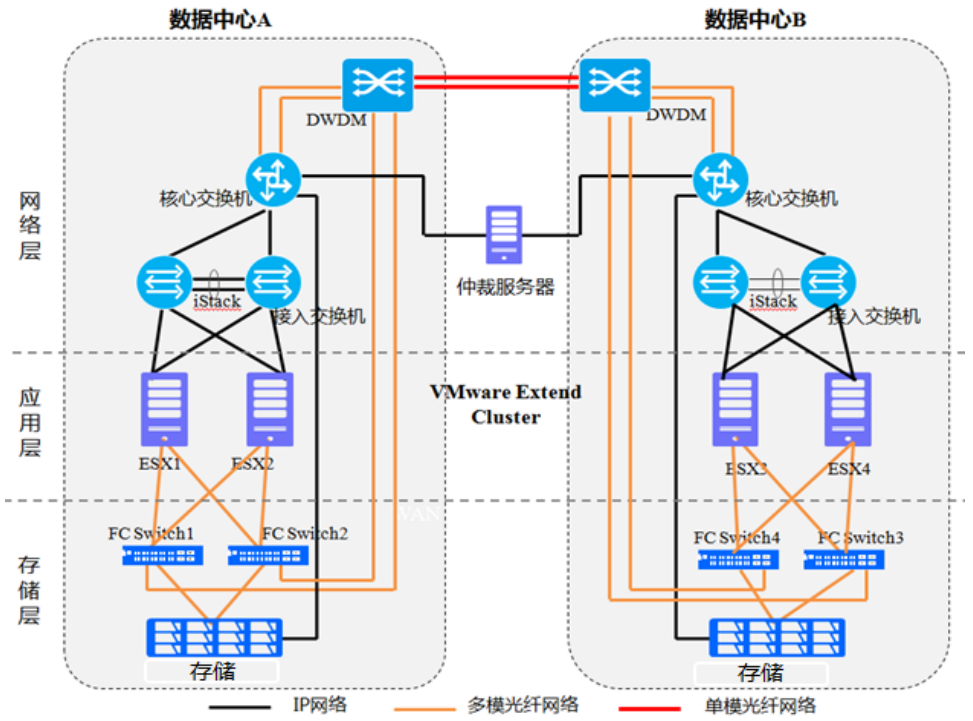
模式2、跨中心单集群模式

- 第一种是基于分布式应用协调机制：构建一套跨中心应用集群，通过分布式应用协调如Zookeeper实现跨中心的高可靠性集群，实现统一配置、统一管理和任务分配。
- 第二种是基于数据副本保护机制：如详单云和大数据的hadoop集群、大数据的MPP集群等，通过进行合理规划设计，确保任一中心节点都是完整的数据副本，由集群自动维护两个中心的数据副本同步机制来实现双活。



虚拟化云平台双活

- 基于存储阵列双活和VMware 跨站点集群功能实现虚拟化平台数据中心容灾解决方案，在阵列双活技术支撑下，通过VMware Cluster 的HA高可用功能实现故障业务切换保护，从而达到保证业务连续性的要求。
  - 网络站点间二层互联，采用波分传输，存储实现双活为上层提供共享存储；
  - 将两个数据中心服务器配置为一个集群，通过HA和DRS实现高可用和资源动态智能分配；
  - 服务器之间建议通过万兆以太网提供心跳服务与vMotion迁移流量，集群内的所有服务器需符合集群的兼容性规则。
  - 应用层：由四台服务器构建VMware ESXi Cluster。



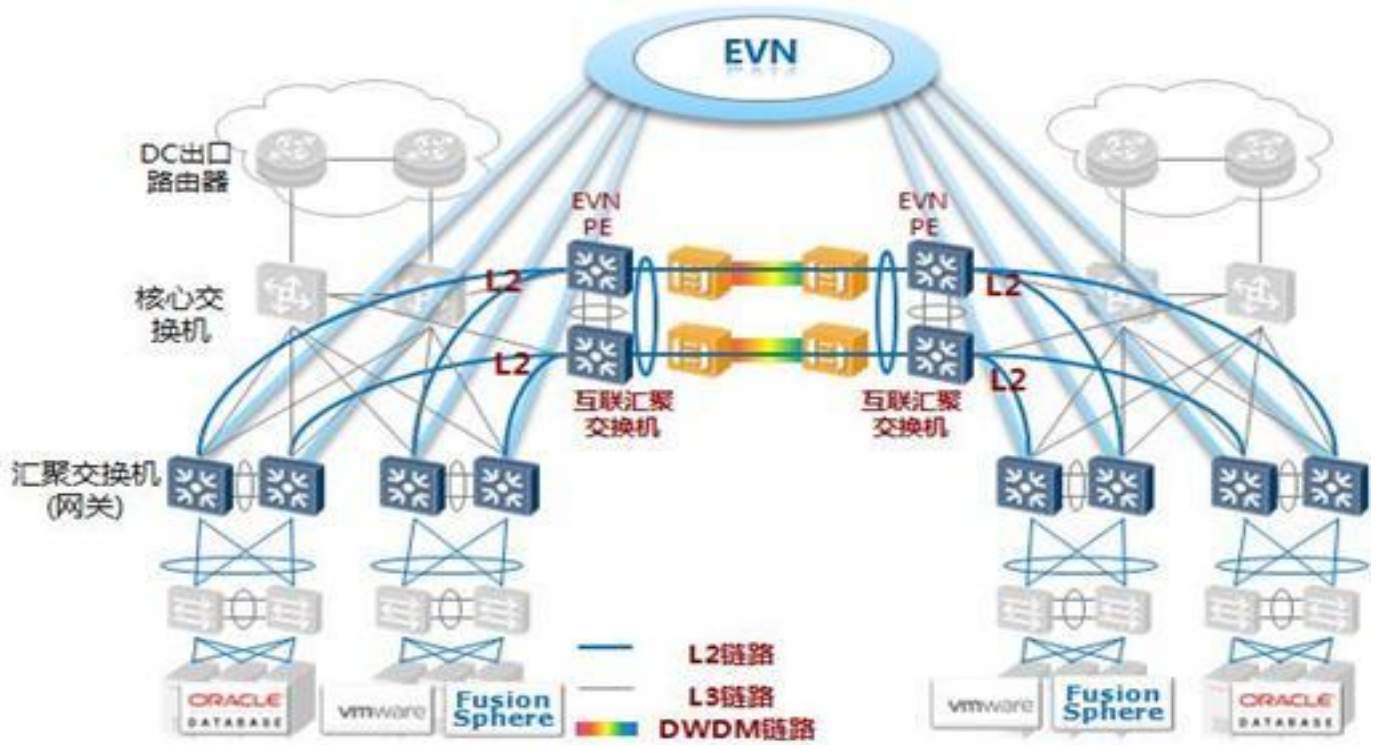
- 数据层
- 存储层
- 接入/应用层
- 虚拟化/云平台
- 技术关键点

1、跨中心大二层网络

- 方案1： EVN/OTV/EVI 技术

以EVN为例，每个中心部署互联汇聚交换机，网关交换机通过链路聚合接入该互联汇聚交换机，互联汇聚交换机通过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。数据中心间三层互通，二层域完全隔离ARP广播、未知单播限制在本数据中心。

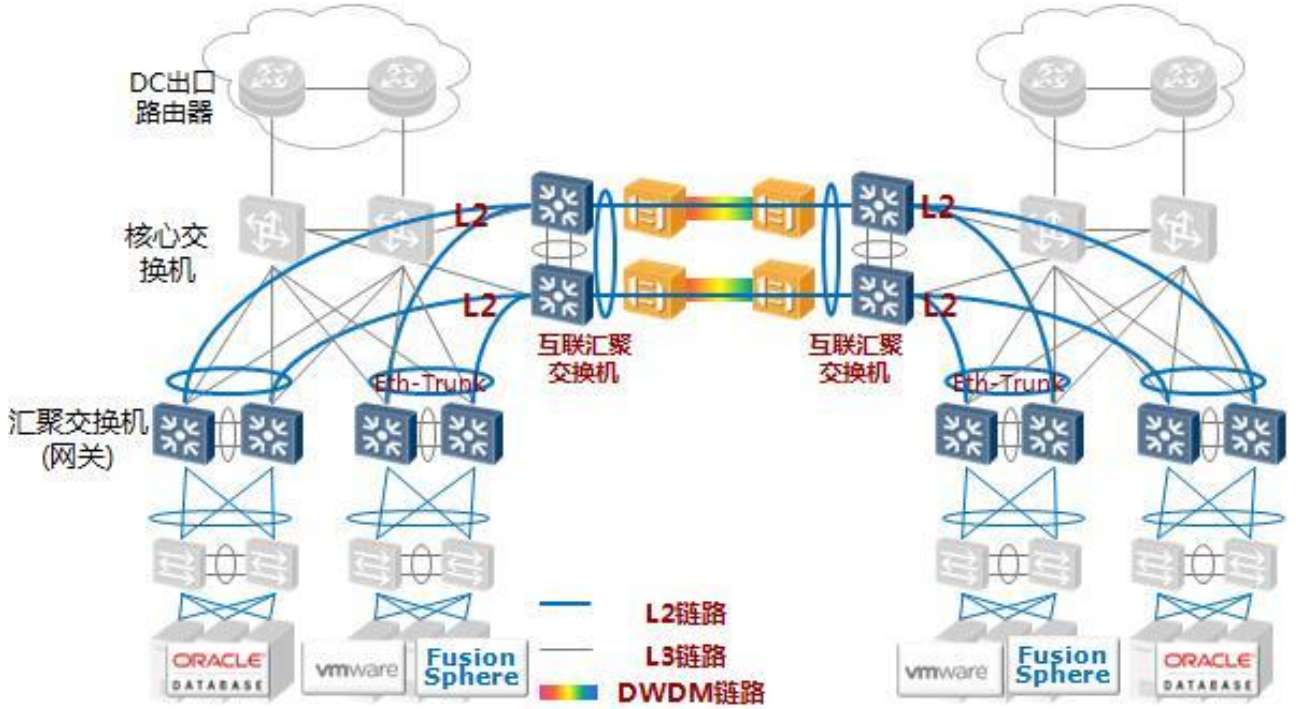
Mac IN IP



1、跨中心大二层网络

- 方案2：采用二层光纤直连技术打通。

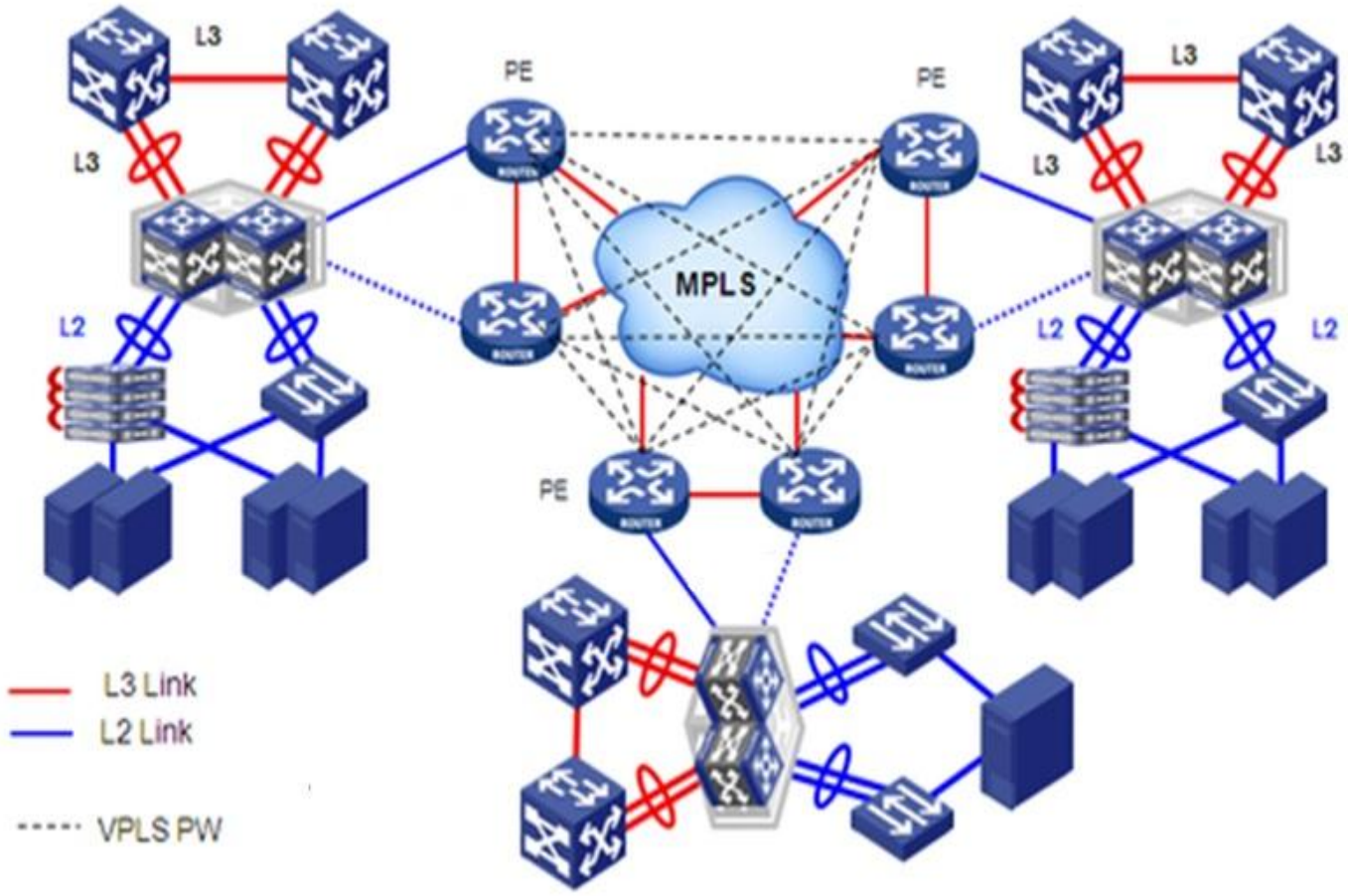
每个中心部署互联汇聚交换机，中心内的汇聚（网关）交换机通过链路聚合接入该互联汇聚交换机，互联汇聚交换机通过链路聚合接入波分设备，链路聚合保证整网无二层环路。同时在汇聚互联交换机配置二层风暴抑制



1、跨中心大二层网络

■ 方案3：采用基于MPLS网络的VPLS互联。

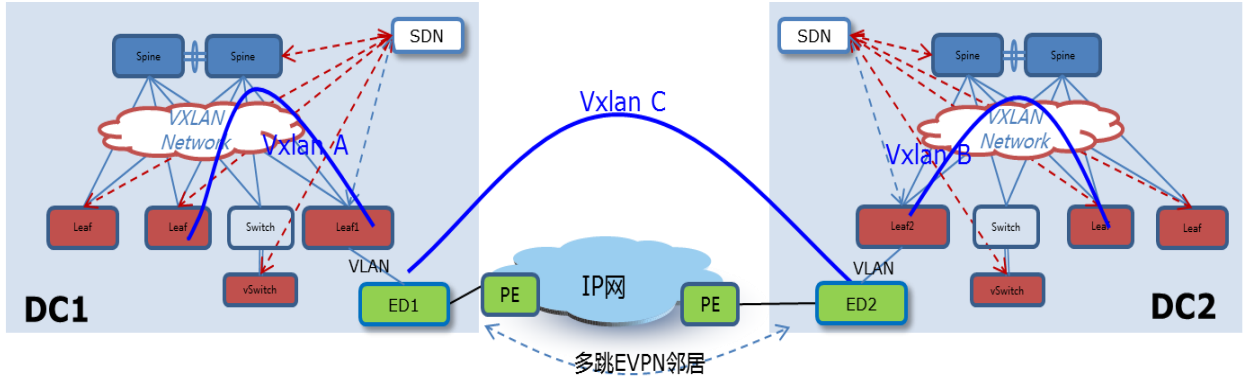
每个中心的核心交换机与专用的MPLS域专用网络直连，通过MPLS专属网络的本地PE设备与对端中心的机房PE设备之间建立VPN，将各个PE设备所互连的二层网络通过MPLS VPN方式建立二层互通。



1、跨中心大二层网络

■ 方案4：基于Overlay网络的大二层互联。

以Vxlan实现方式为例，每个中心通过单独的ED设备与Underlay网络连接，在每个中心内部业务数据通过VXLAN进行业务交换，涉及到跨中心业务互访时，将通过与ED设备直连的Leaf设备剥离VXLAN标签转换为VLAN业务后，由ED设备再次进行VXLAN封装，从而通过大二层透传到对端中心的ED设备剥离VXLAN标签，由对端中心的Leaf设备重新封装VXLAN标签。



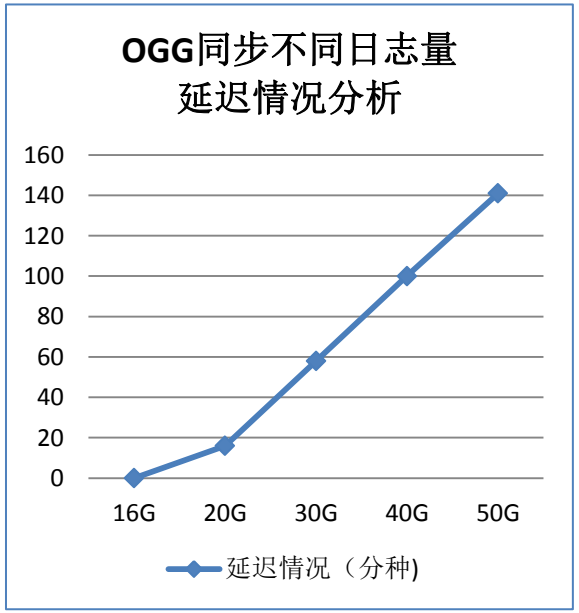
1、跨中心大二层网络

技术种类	组网方式	优势	劣势	适用模式
基于MPLS网络的二层互联	VPLS	1、标准化程度高，兼容性强，能够兼容大部分的MPLS网络 2、CE实现双归属，HA性能高，保证数据中心间互联的高可靠性 3、价格优势	技术比较复杂，部署及运维管理难度较大	适用于跨地域多中心互联
光纤直连	VPLS Over GRE	与基于MPLS网络的VPLS优势相同	1、技术比较复杂，部署及运维管理难度较大 2、需要部署QoS来保证带宽，时延难以保证	数据中心间只有IP互联网络时
大二层互连	OTV/EVN/EVN	1、网络改动较小 2、配置简单	1、各厂商私有协议，在涉及多品牌网络环境中难以实现对接	适用于多地域的中心互连
基于Overlay网络的二层互联	VXLAN	1、支持Overlay网络，可以跨裸光纤、MPLS或IP网络实现二层互联 2、配置简单 3、提高系统的HA性能	1、各厂商的私有协议，需要数据中心间采用同品牌设备 2、案例较少	适用于跨地域多中心互联

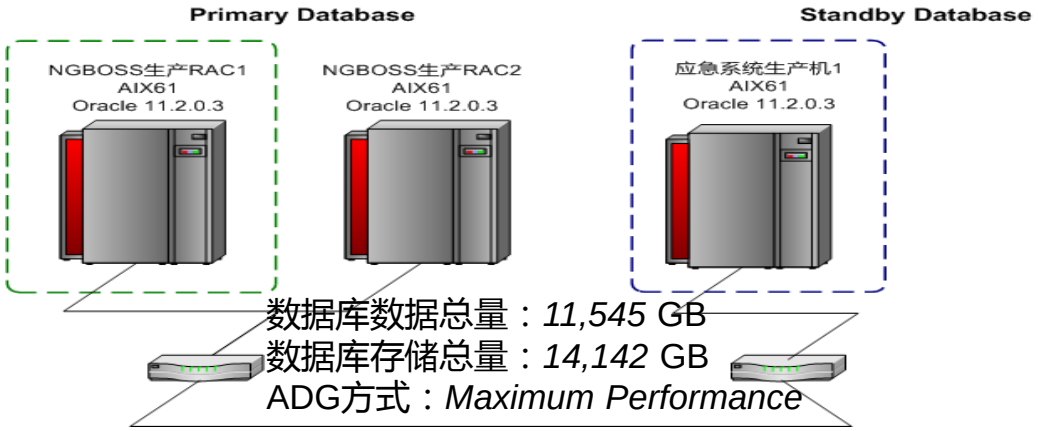
2、 关于GoldenGate

■Oracle GoldenGate 性能瓶颈在数据同步环节，即在复制进程Replicat入库速度，因为在容灾端恢复数据过程是执行逻辑SQL，比较消耗资源：

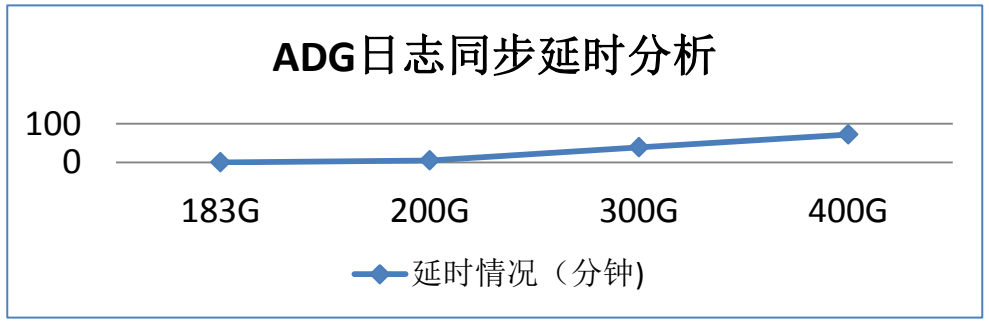
- **抽取进程（Extract）**：该进程主要瓶颈在于LCR(logical change record)转换为UDF环节，主要优化建议：
  - 拆分Extract进程，建议同一个schema下表尽量在一个进程组中
  - 优化进程参数如eofdelay、flushsecs等
  - I/O部分建议增加日志读取间隔3s，增加内存刷新时间3s
- **投递进程（Pump）**：带宽优化和IO优化：
  - 复制的表最好有主键或唯一索引，减少生产日志量
  - 数据传输过程启用数据压缩特性，减少带宽需求量
  - 适当增大TCP缓存
  - 增加队列读取间隔为3s，内存刷新时间为5s
- **复制/应用进程（Replicat）**：该环节出现性能问题较多，需要重点优化：
  - 合并小交易减少事物数量，减少写checkpoint file/table次数
  - 大交易拆分（maxtransops参数），提高写入速度
  - 基于表或Range等拆分replicat进程



3、 关于ADG



服务器	型号	CPU	内存
Primary Server	IBM P780 × 2	40 × 2	280G × 2
Standby Server	IBM P780 × 1	24	120G



- 日志产生量（采集于2015年4月初）
  - 日均产生归档量 ***1,300*** **GB,**其中 1 节点 ***600*** **GB,2**节点 ***700*** **GB**
  - **1**天日志的峰值为 ***1705*** **GB,** 1 节点峰值 ***811*** **GB,2**节点峰值 ***911*** **GB**
  - 单个小时日志峰值为 ***183*** **GB,1**节点峰值 ***90*** **GB,2**节点峰值 ***96*** **GB**
- 网络流量
  - 采用千兆网，传输日志平均占用带宽为 ***16.24*** **MB/s**，单个小时内峰值为 ***52*** **MB/s**
- 应用时延（Transport Lag + Apply Lag）
  - 异步方式传送日志，平均延时 ***0.65*** 秒，正常业务处理期间时延小于***10*** 秒
  - 生产库中产生大量I/O的维护操作，比如添加数据文件，会导致目标库应用时延相应增加，可通过调整维护作业时间窗口加以避免。

3、 Extend RAC关键参数

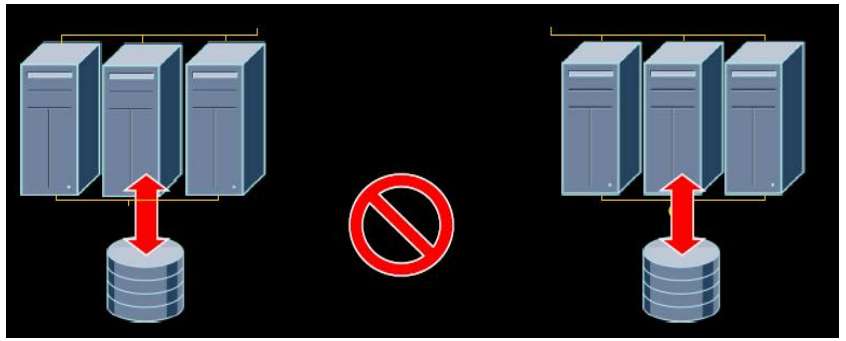
		参数名称	建议值	说明
GPFS Cluster 关键参数	GPFS集群性能参数	mmchconfig pagepool	204M	
		mmchconfig prefetchThreads	128	
		mmchconfig maxMBpS	8192	
		mmchconfig worker1Threads	475	
	心跳参数	mmchconfig minquorumnodes	2	
		leaseRecoveryWait	15	Default 35s
		TotalPingTimeout	20s	Default 120s
	网络故障系统恢复关键参数	mmchconfig failureDetectionTime	20	
		mmchconfig leaseRecoveryWait	15	
		mmchconfig totalPingTimeout	20	
		mmchconfig leaseDuration	15	
	I/O性能	mmchconfig readReplicaPolicy	local	本地优先读
Oracle RAC 和ASM参数	RAC仲裁	站点距离>10km， 网络传输用DWDM		
		Disk timeout （Disk Heartbeat IOT）	250s	默认120s
		Misscount（Network Heartbeat）	200s	默认 30s
	ASM参数	asm_preferred_read_failure_group	dg1.fg1,dg2.fg2..	ASM本地优先读
		_asm_hbeatiowait	120s	ASM磁盘心跳超时时间

注意：

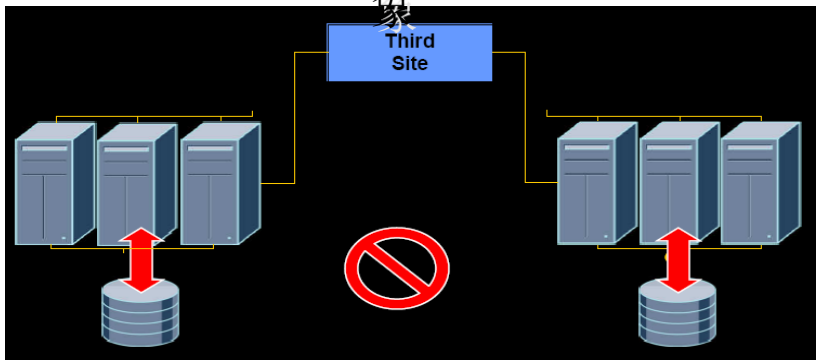
- 关于RAC仲裁和GPFS仲裁，保证RAC的磁盘仲裁要晚于GPFS的仲裁，使得在网络故障情况下GPFS提前RAC做出判定。
- ORACLE RAC的心跳参数：misscount是RAC网络心跳时间， disktimeout是表决盘的心跳时间

4、 防止“脑裂”现象

- 1、由于数据中心间距离远，网络稳定性相比同机房差，必须需要额外进行冗余设计，如网络连接、内部网络、san连接等。2个数据中心间网络不稳定情况下，无论存储虚拟化技术还是Oracle的RAC均可能出现“脑裂”现象，造成访问中断，数据不一致现象发生，需要仔细设计，如采用互联环状全冗余架构等、完善的仲裁机制等。
- 2、对跨中心间的网络带宽、存储访问带宽利用率不能超过30%。
- 3、双活由多层软硬件组成，如数据库RAC、远程文件系统、存储等，需要仔细规划他们之间的心跳参数，确保越低层的心跳超时时间越高。



两个中心间网络全断下会出现“脑裂”现象



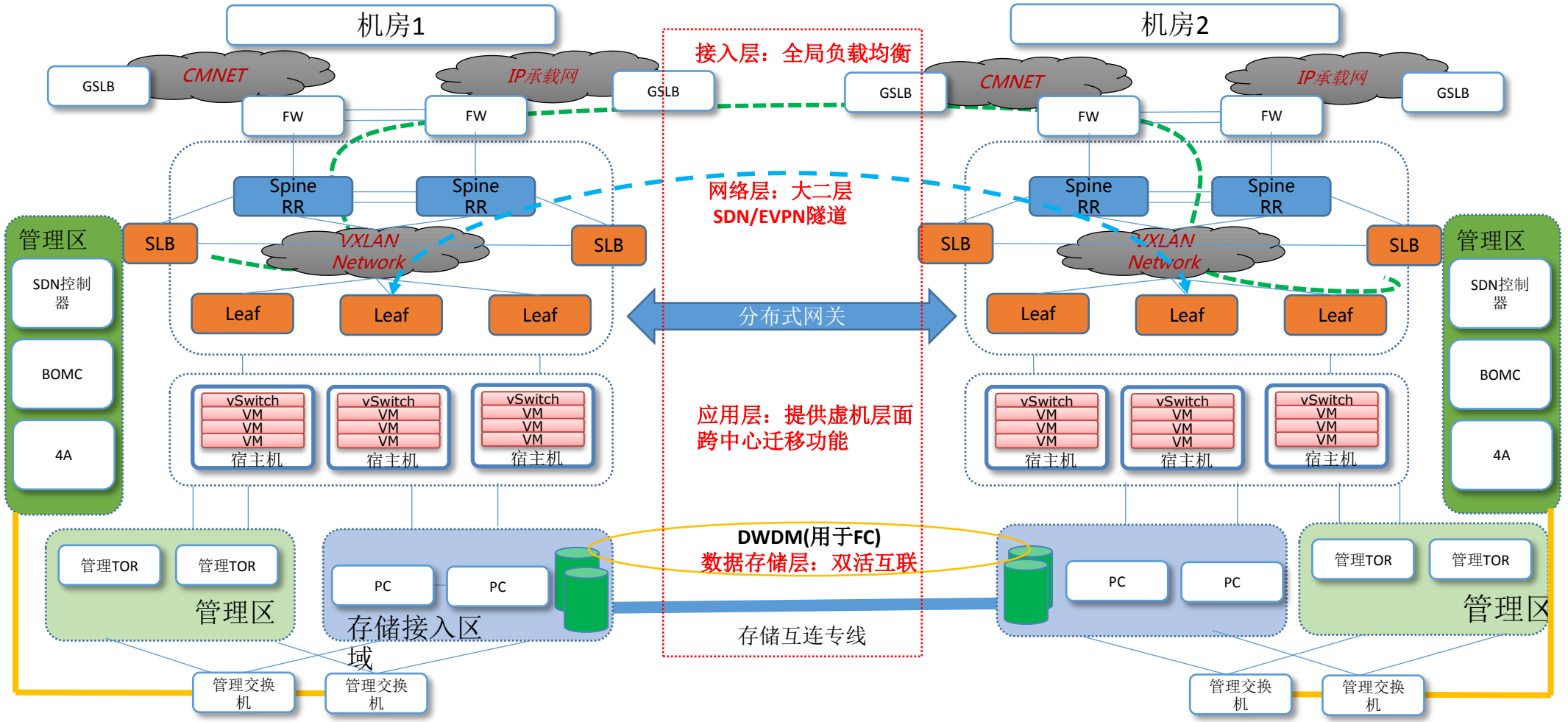
通过增加第三节点，防止两个中心间网络全断场景

5、 全面的计划内外测试场景

■ 双活涉及到跨中心网络层，数据层和存储层，故障场景相比较传统架构更多，更复杂，相互之间存在多种依赖关系，需要充分设计故障测试场景：

	类型	场景描述	执行操作	实际测试结果
计划内	计划内切换存储		手工切换	所有节点的上的I/O暂停3s后继续。
计划外	站点内链路故障	RAC1 单条光纤故障	down port	RAC1节点上的IO暂停近20秒后继续，其他节点的IO不受影响，存储不切换
		RAC1两条光纤故障	down port	RAC1节点上的IO暂停近20秒后继续，其他节点的IO不受影响，存储发生切换
		RAC1所有光纤故障	down port	RAC1节点被踢出Oracle RAC集群，其他节点的IO不受影响，存储不切换
	存储故障	主站点存储故障	Disable zone	所有节点的数据库IO均暂停30秒后继续
	网卡故障	RAC1 RAC私网故障	ifconfig * down	RAC1被踢出，其余节点I/O暂停30秒继续
		RAC1 NFS网卡故障	ifconfig * down	所有节点I/O不受影响
	节点故障	RAC1故障	halt -q	其他节点的数据库IO暂停30秒后继续
	站点故障	主站点故障	halt-q; disable zone	备站点I/O暂停30秒后继续，存储切换
	站点间网络故障	生产中心间所有网络故障（不包括NFS网卡）	ifconfig * down	主站点IO暂停30秒继续，备站点的节点被踢出Oracle集群
	FC链路故障	两个站点间FC链路故障	Disable zone	主站点I/O暂停30秒后继续，备站点的节点被踢出Oracle集群
	脑裂	FC链路及网络同时故障	Disable zone ifconfig * down	主站点I/O暂停30秒后继续，备站点的节点被踢出Oracle集群，备存储转为可读写

一个双活数据中心架构例子




DBAplus
www.dbaplus.cn
THANK YOU