国家电网公司
STATE GRID
CORPORATION OF CHINA

# 双活技术交流

张顺仕

二〇一六年八月

国家电网公司
STATE GRID
CORPORATION OF CHINA

# 目录


1 同城双活产生背景

2 同城双活架构介绍

3 同城双活产品对比

4 双活应用研究介绍

5 异地容灾技术分析

国家电网公司
STATE GRID
CORPORATION OF CHINA

# 双活数据中心产生背景分析

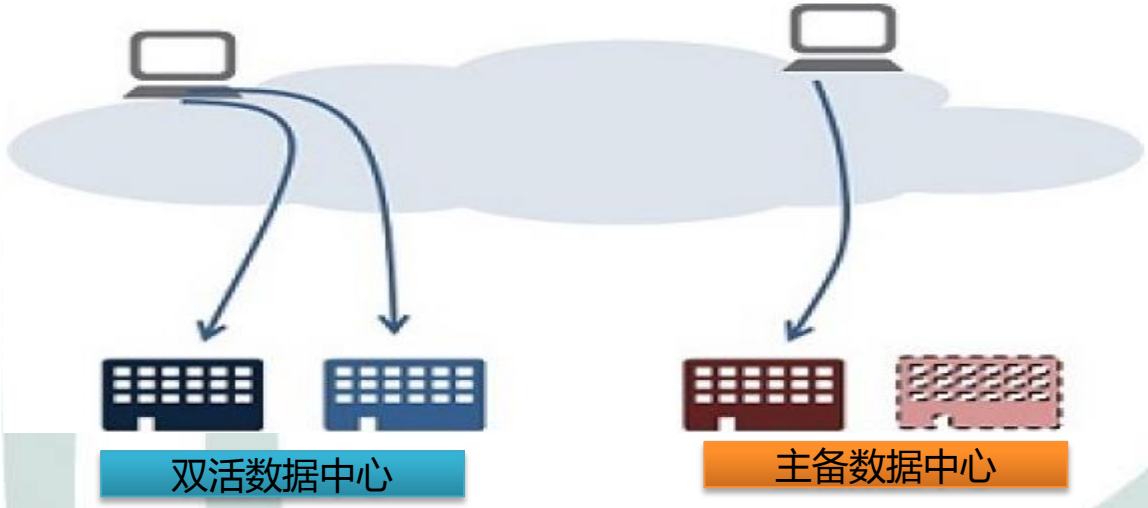
企业出于灾备（Disaster Recovery）的目的，一般都会建设两个（或多个）数据中心。一个是主数据中心用于承担用户的业务，一个（或多个）备份数据中心用于备份主数据中心的数据、配置、业务等。主备数据中心之间一般有冷备、热备、双活三种方式。

近年来，随着公司信息化建设的深入推进，业务应用系统已经成为公司生产经营的重要支撑。为了保障业务数据安全和业务系统连续性，网省公司已对营销、生产等重要系统建立了的异地灾备系统，但异地灾备存在RTO、RPO、切换复杂、运维代价大等问题。

为了提升系统的应急保障能力，进行适当的系统改造，建设业务同城双活系统，从网络架构层、应用层、数据库层以及存储层面分别来构建，从而达到一个完整的双活系统，保障运营系统的数据零丢失、业务不中断。

国家电网公司
STATE GRID
CORPORATION OF CHINA

# 双活数据中心的概念与特点




**双活数据中心特点：**

- 提高资源利用率；
- 数据零丢失；
- 实现的业务连续性。

双活的两个数据中心都处于运行当中，同时提供实时业务处理能力，且互为备份，所以称为“双活”；而传统的灾备（主备）数据中心，日常仅一个数据中心投入运行，另外一个数据中心处于非工作状态（不对外提供业务处理服务），只有当灾难发生时，主生产数据中心瘫痪，灾备中心才会启动并接管业务处理。

国家电网公司
STATE GRID
CORPORATION OF CHINA

# 目录


1 同城双活产生背景

2 同城双活架构介绍

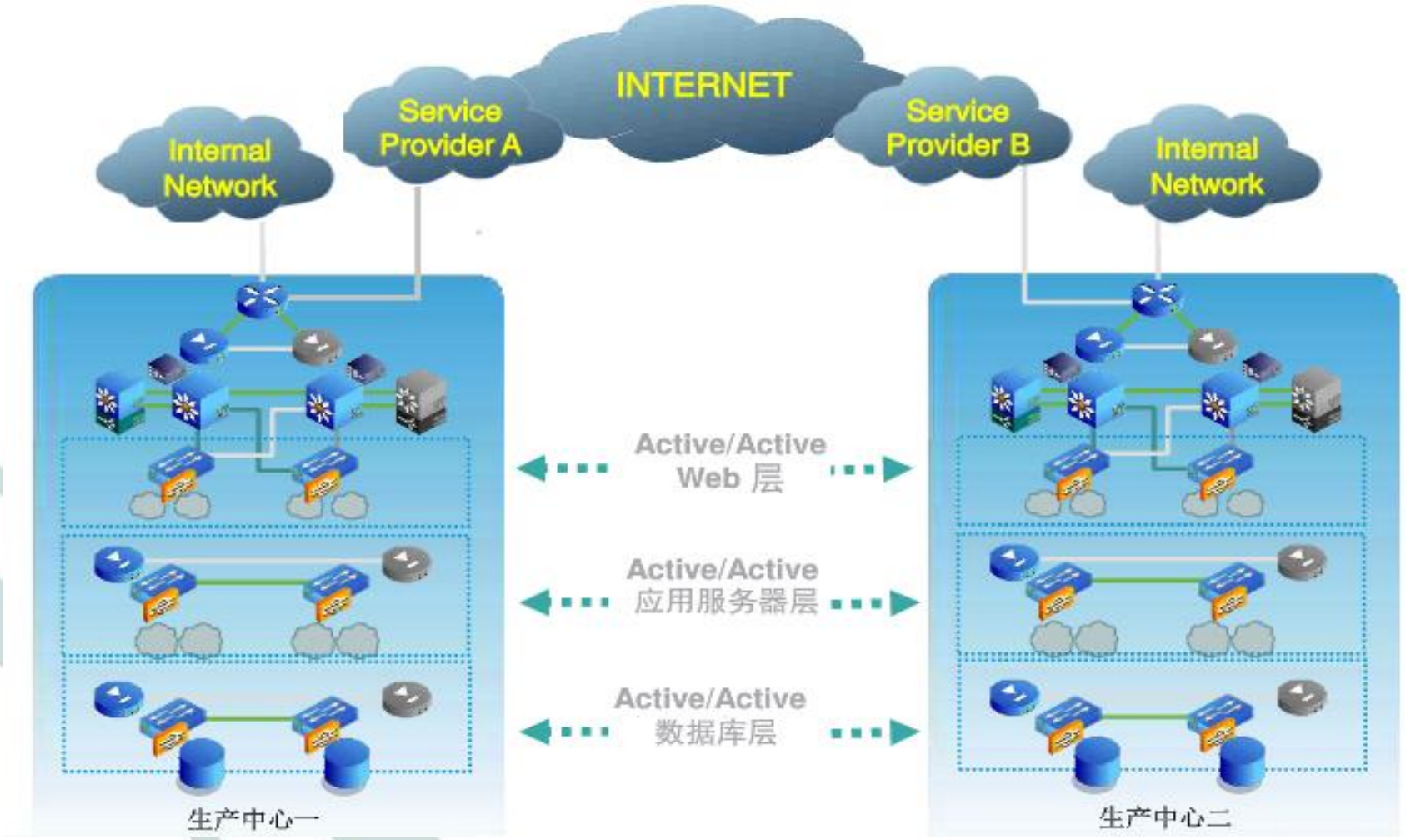
3 同城双活产品对比

4 双活应用研究介绍

5 异地容灾技术分析

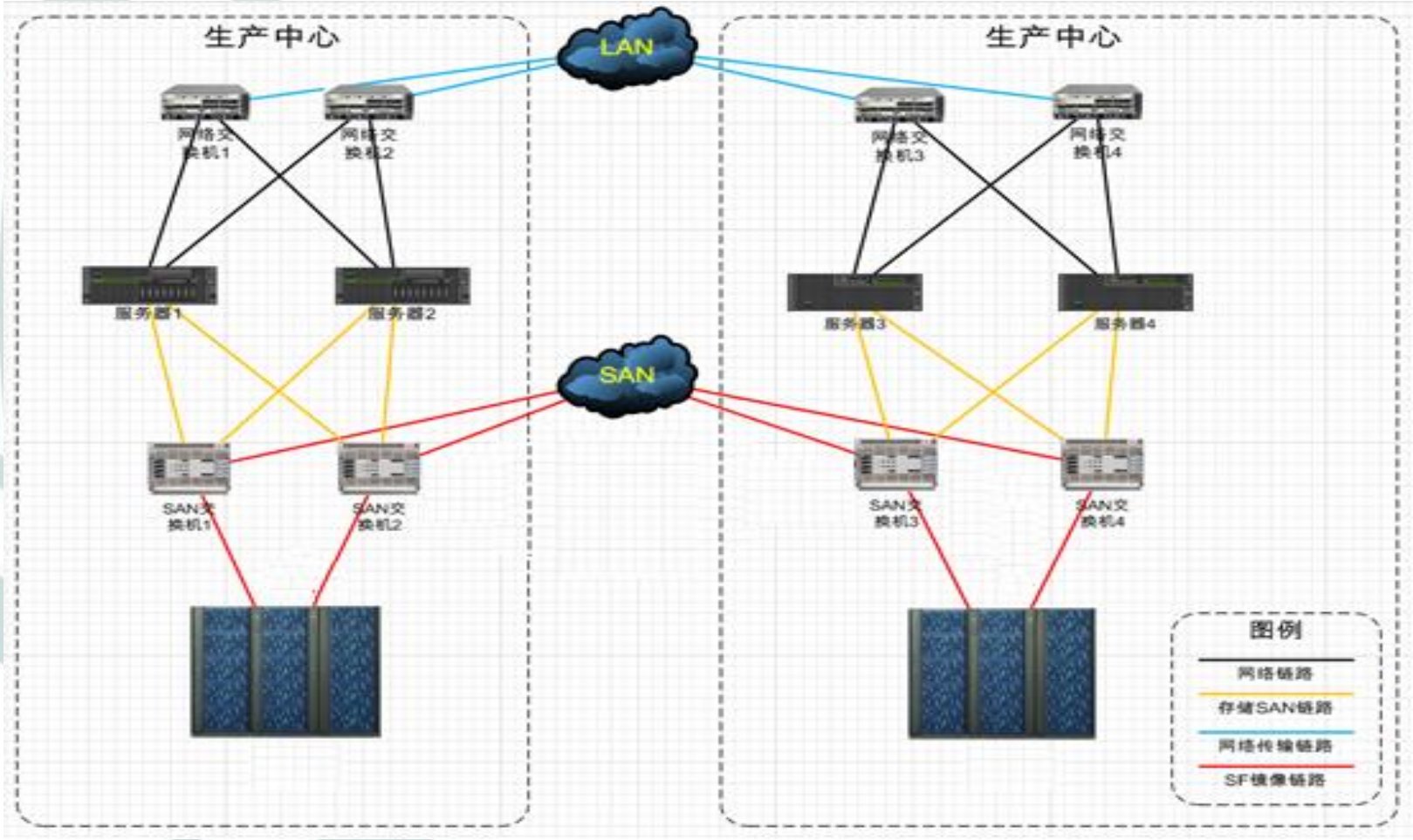
国家电网公司
STATE GRID
CORPORATION OF CHINA

# 同城双活逻辑结构



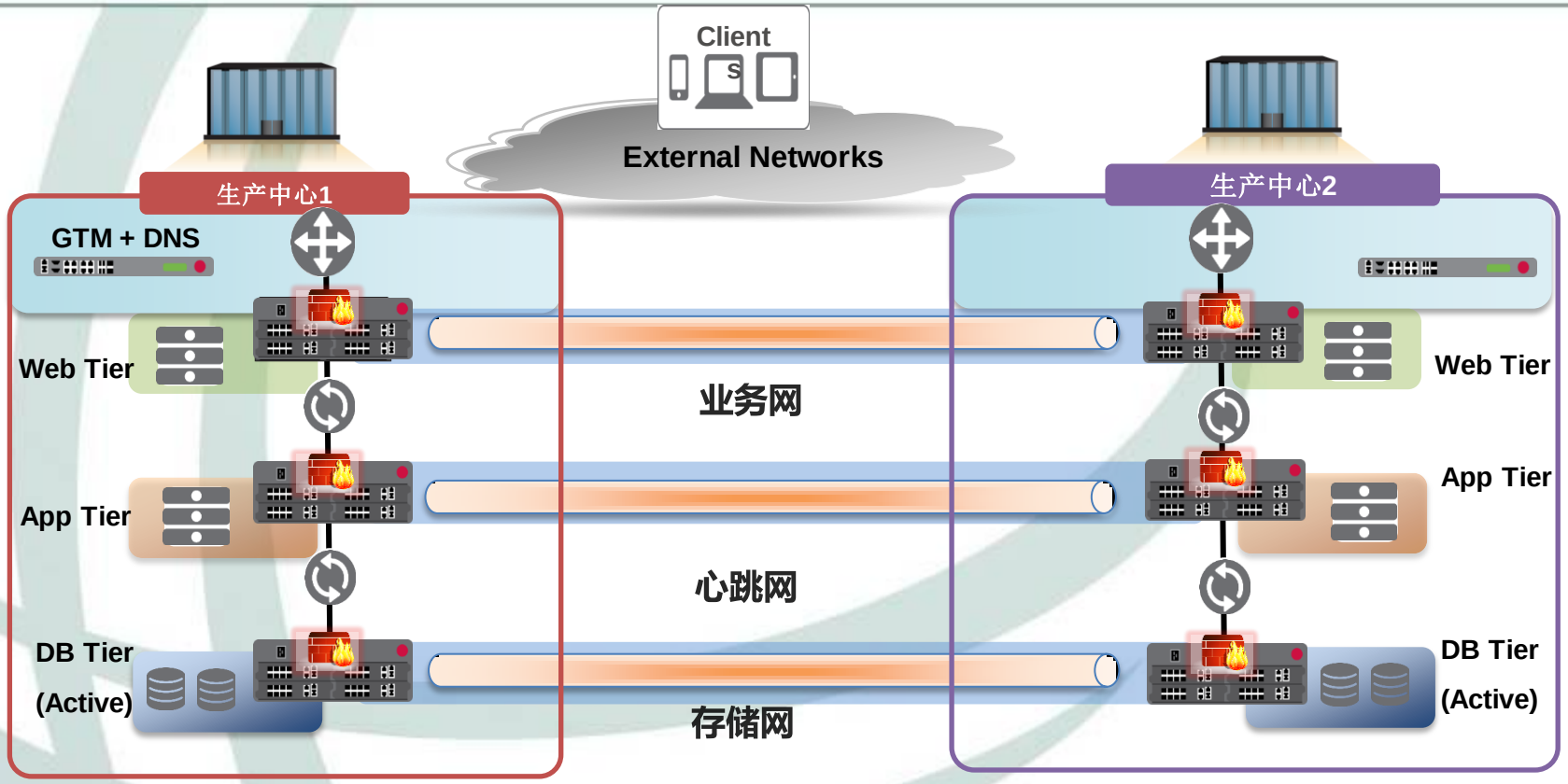
国家电网公司
STATE GRID
CORPORATION OF CHINA

# 同城双活架构图



国家电网公司
STATE GRID
CORPORATION OF CHINA

# 双活数据中心网络架构



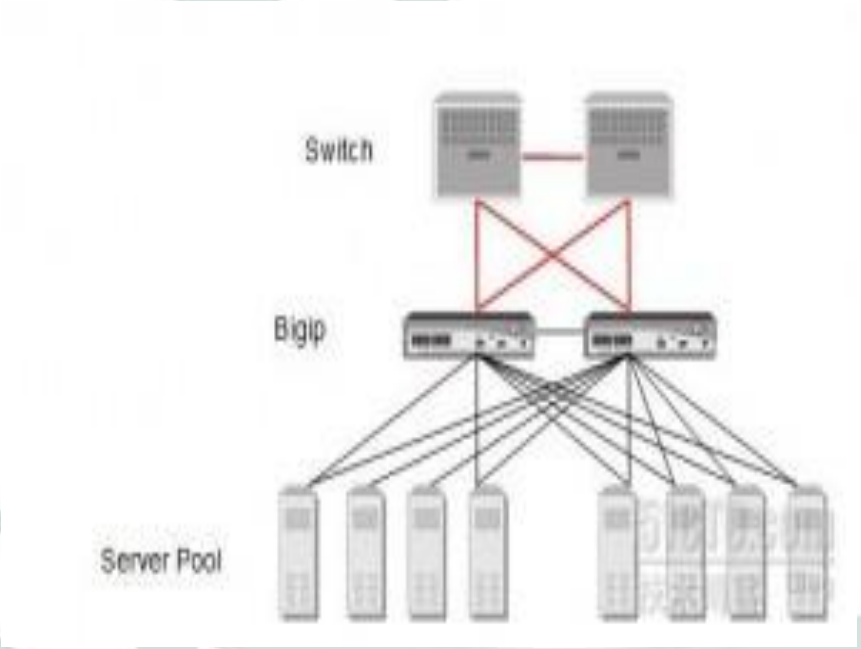

**业务网**：生产中心1与生产中心2的网络设备采用2条10G光纤链路互联，主要提供两中心间数据库服务器与应用服务器DCN网的互联互通。

**心跳网**：生产中心1与生产中心2的网络设备采用2条10G光纤链路互联，主要提供两中心间系统集群心跳和数据库RAC心跳的互联互通。

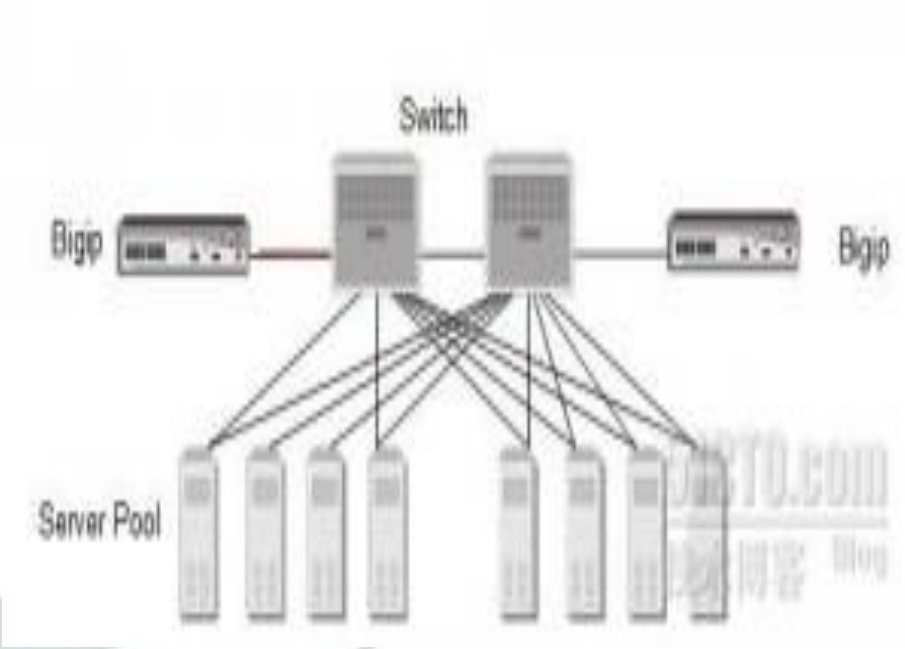
**存储网**：生产中心1与生产中心2的光交设备采用4条4G/8G光纤链路两两互联，主要提供两中心间服务器与异地存储的互通。

国家电网公司
STATE GRID
CORPORATION OF CHINA

# 同城双活架构下的负载均衡模式




直连模式结构

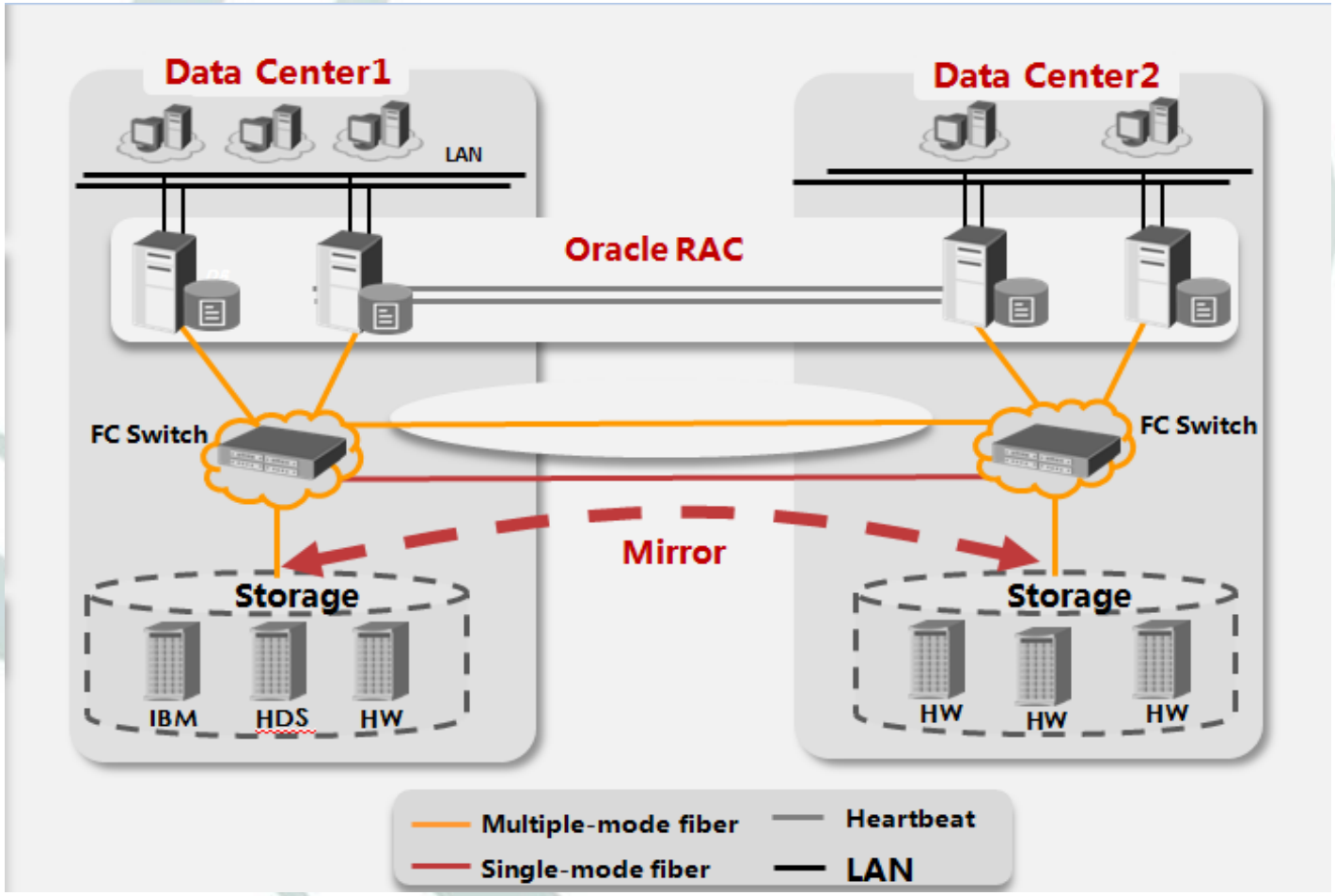



旁路模式结构

这两种的部署结构不一致，做双活的话是有差异，如果是直连结构，就需要两个机房的四个负载均衡做级联。如果是旁路部署，那么需要两个机房的四个交换机做级联。

国家电网公司
STATE GRID
CORPORATION OF CHINA

# 同城双活数据库结构图




**方案特点:**

- 业务双活访问，资源充分利用；
- 业务不中断，数据0丢失；
- 易扩展，可平滑升级为两地三中心；
- 存储统一管理，管理与维护成本低。

国家电网公司
STATE GRID
CORPORATION OF CHINA

# 数据库层双活实现——远程RAC

1. 两个机房的所有节点组成一个Oracle RAC集群

2. 两个机房之间通过DWDM设备提供基础链路承载

3. 同一机房内部的Oracle RAC两节点之间通过冗余的本地交换机，结合主机端的双网卡绑定技术，实现CRS心跳互联

4. 同一机房内部的Oracle RAC两节点之间通过冗余的本地交换机，结合Oracle自身的Load Balance功能，实现Cache Fusion互联

5. 同一机房内部主机与存储之间，通过冗余的SAN交换机，结合存储多路径管理软件功能，实现互联

6. 两个机房之间的Oracle RAC节点，依赖底层的存储镜像技术和以太网互联技术，通过Oracle自身的节点配置实现一个集群运行

7. 两个机房之间的以太网交换机，利用Trunking技术和VLAN聚合技术实现互联，并打通机房RAC节点之间的CRS心跳和Cache Fusion通道

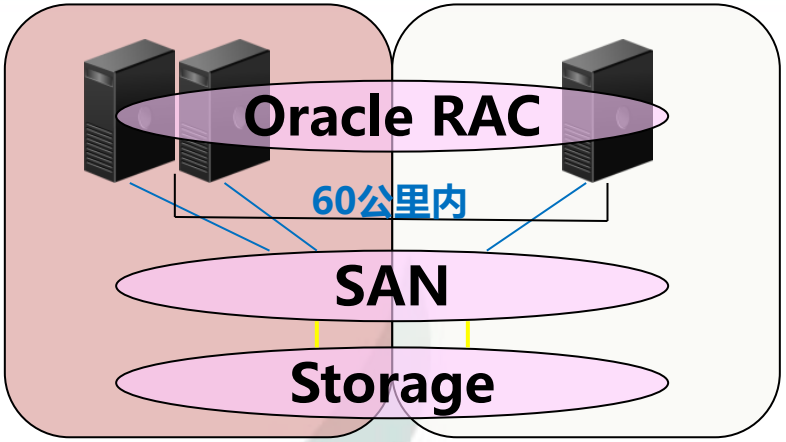
8. 两个机房存储，利用主机上的存储管理软件镜像功能，实现数据的同步，并保持对上层Oracle数据库的透明

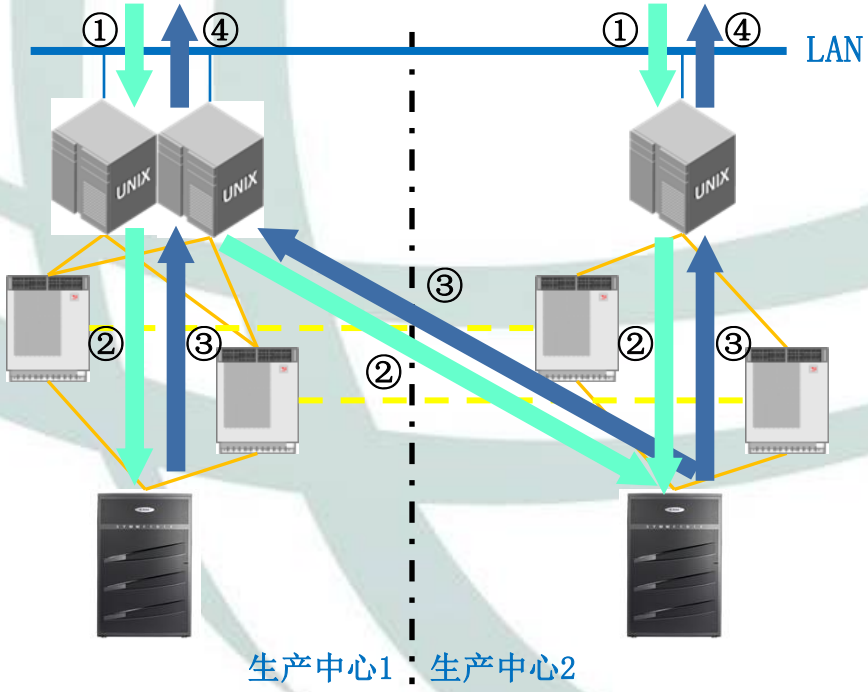
9. 两个机房之间的SAN交换机，利用广域I/O加速技术和ISL技术实现两个机房交换机之间的各自互联

# 双活数据同步过程

## 双活双中心技术方案

通过第三方卷管理软件实现数据库的远程RAC技术，可保持数据的实时同步，保证数据的一致性与完整性，实现数据库双活，从而实现业务双中心提供同时服务。




## 第三方卷管理的读写模式

数据读取流程：

①应用服务器向数据库服务器发出读请求

②数据库服务器读取最近存储中心存储的数据

③存储向数据库服务器确认数据读取完成

④数据库服务器向应用服务器确认数据读取完成

数据写入流程：

①应用服务器向数据库服务器发出写请求

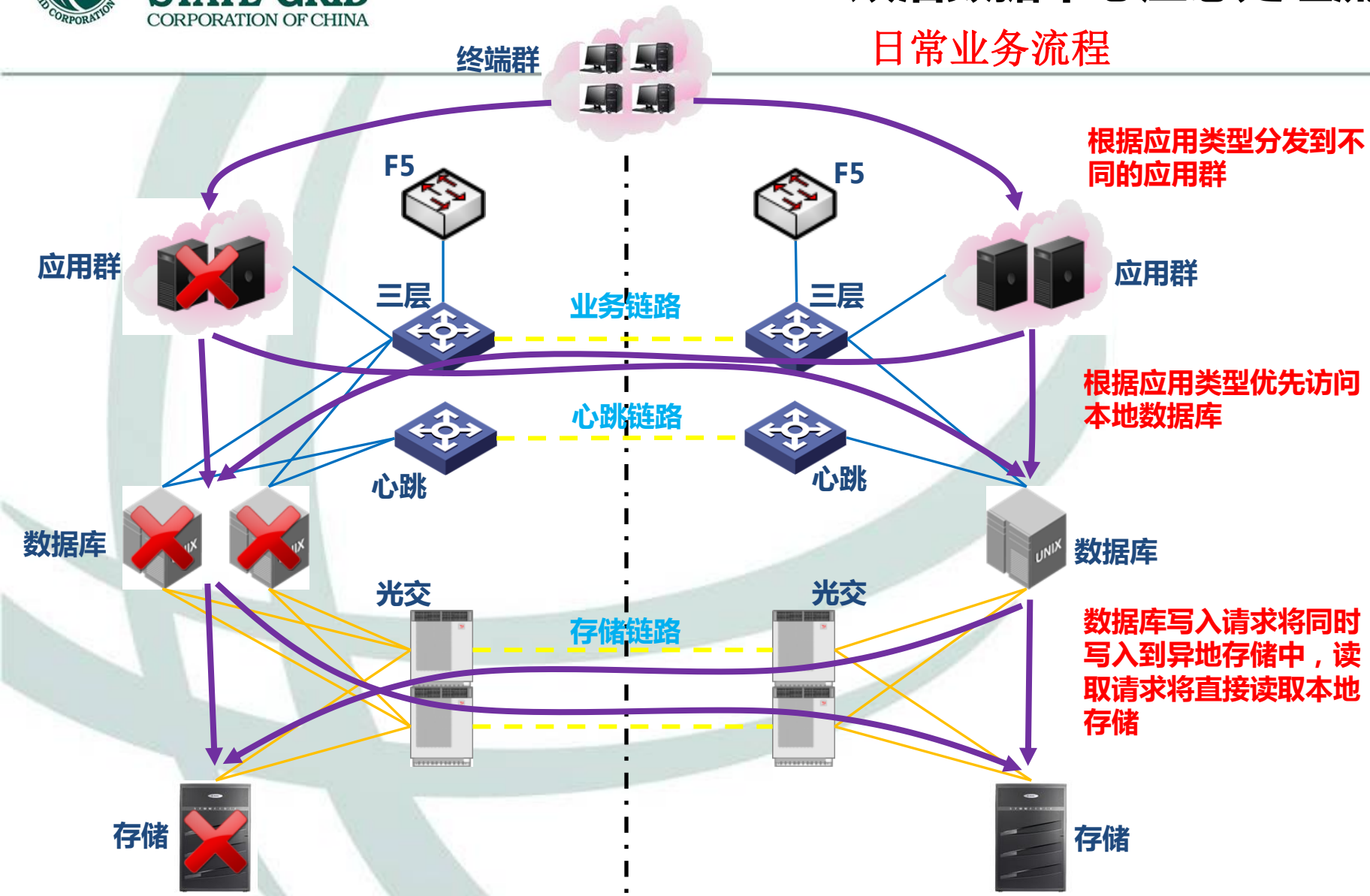
②数据库服务器向生产中心存储先后向数据库服务器确认数据写入完成

③数据库服务器向应用服务器确认数据写入完成

国家电网公司
STATE GRID
CORPORATION OF CHINA

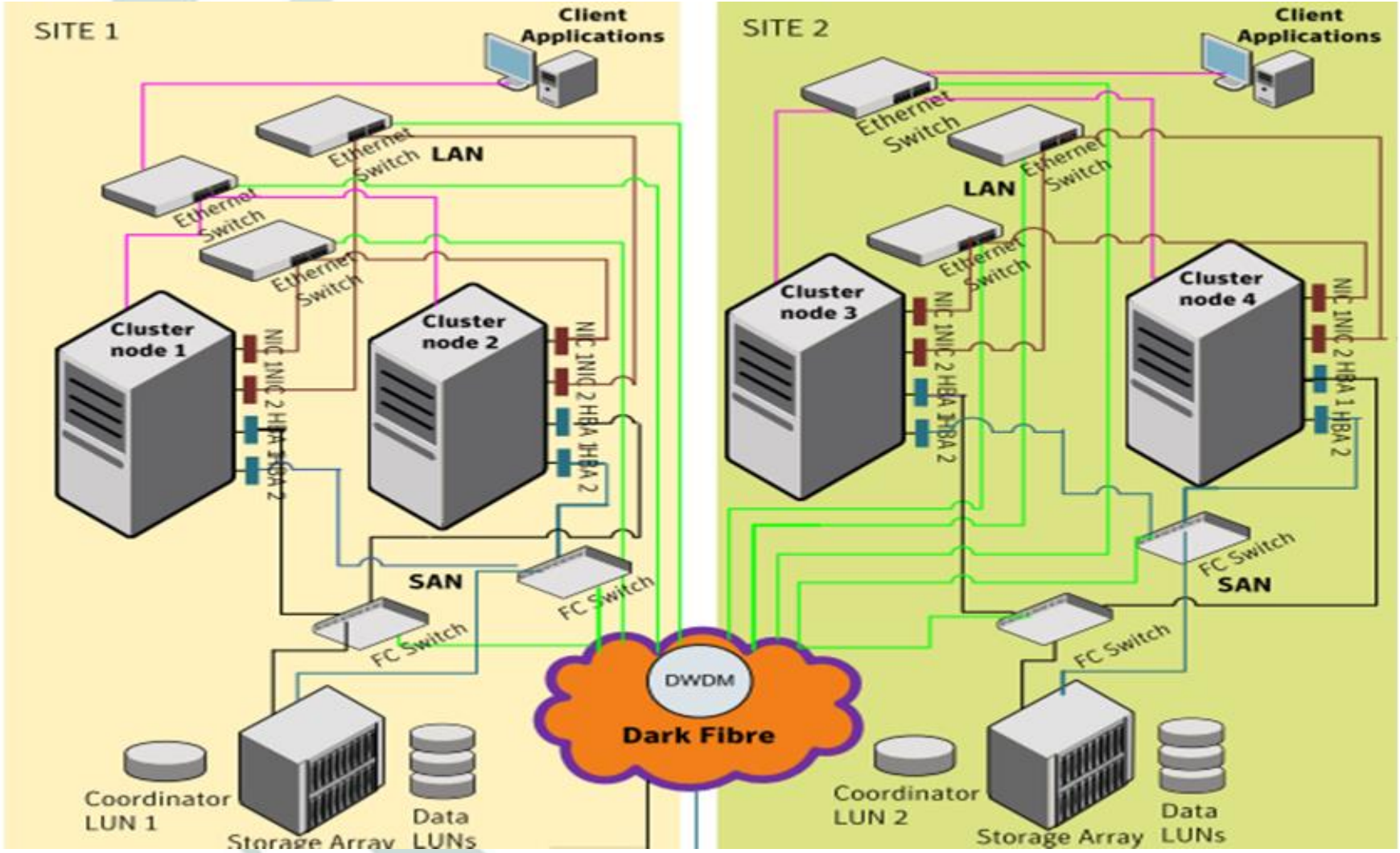
# 双活数据中心应急处理流程

## 日常业务流程



国家电网公司
STATE GRID
CORPORATION OF CHINA

# 典型的双活案例



国家电网公司
STATE GRID
CORPORATION OF CHINA

# 目录


1 同城双活产生背景

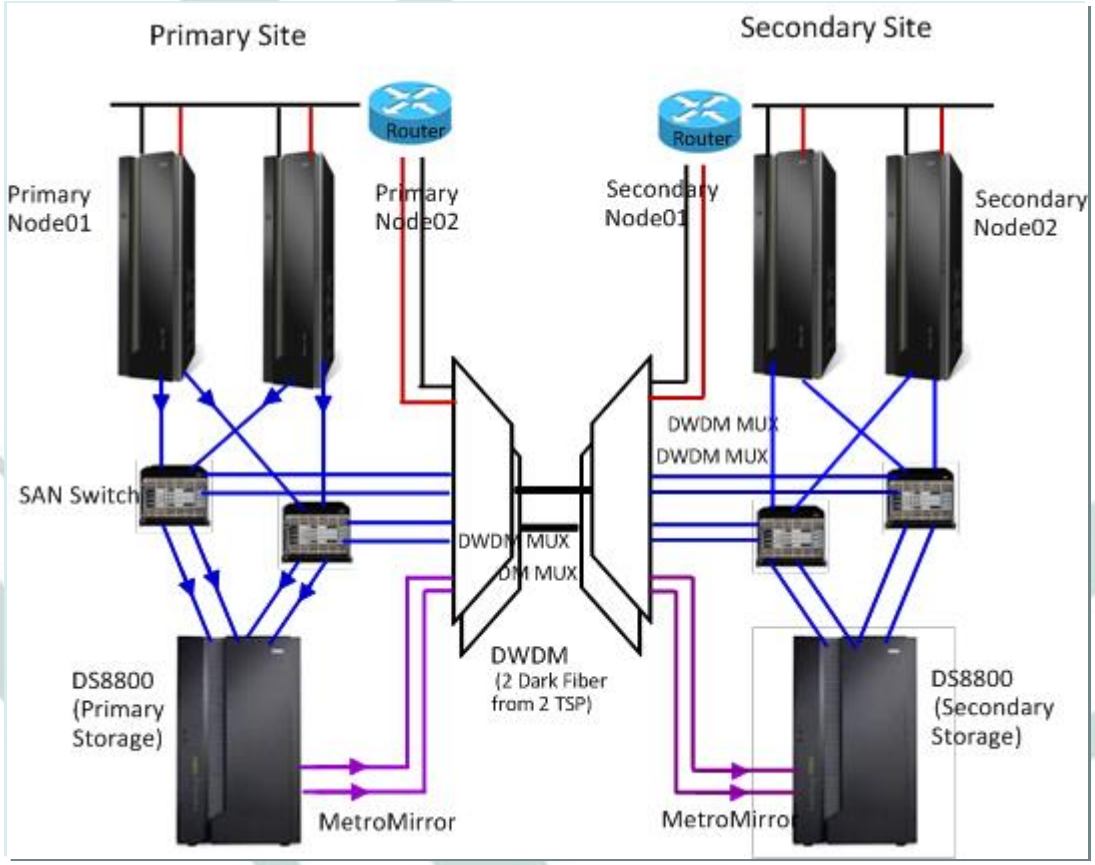
2 同城双活架构介绍

3 同城双活产品对比

4 双活应用研究介绍

5 异地容灾技术分析

# 远程RAC数据库技术-非ORACLE ASM



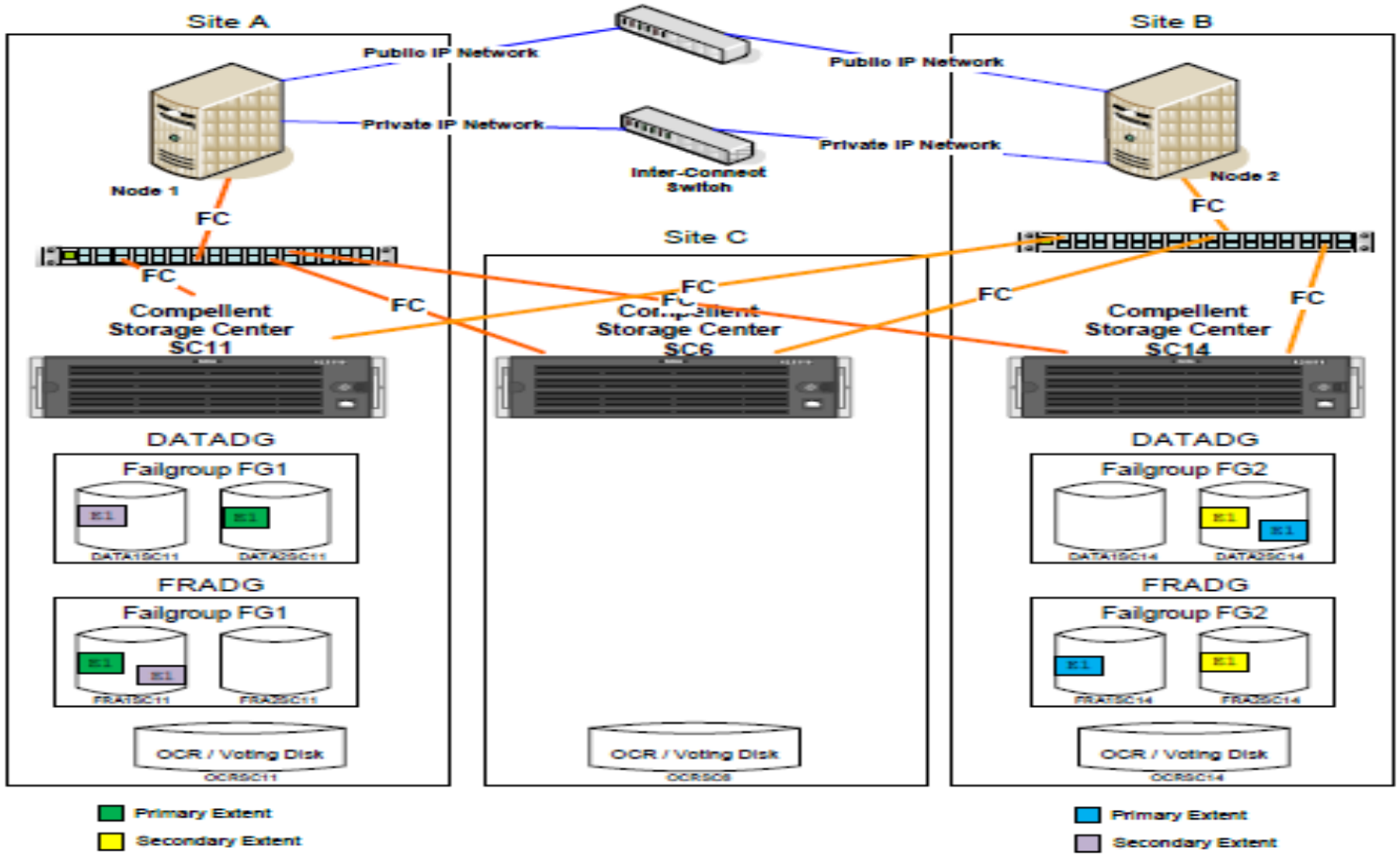

- **基于Oracle RAC/EXTENDED RAC技术**
- **多种实现方案**
  - 基于赛门铁克卷复制的解决方案
  - 基于EMC VPLEX METRO的解决方案
  - 基于IBM PowerHA System Mirror的解决方案

**主要特点：**

- 实现跨机房双活集群
- 消除存储的单点故障

# 远程RAC数据库技术-ORACLE ASM

## 基于Oracle ASM的解决方案



国家电网公司
STATE GRID
CORPORATION OF CHINA

# 远程RAC数据库技术实现对比

| 厂商/产品 | 实现原理 | 稳定性 | 性价比 | 运维复杂度 | 成功案例 | 备注 |
|---|---|---|---|---|---|---|
| Symantec / Storage Foundation | 通过软件实现底层存储级逻辑卷复制技术实现存储同步，从而实现数据库双活 | 良好 | 根据数据库节点数收费 | 运维复杂，需要厂商介入 | 目前电信，银行用的比较多，但大集中业务的双活案例很少 | |
| EMC / VPLEX METRO | 通过硬件实现底层存储级逻辑卷复制技术实现存储同步，从而实现数据库双活 | 良好 | 硬件设备独立收费 | 独立硬件，运维相对简单 | 大集中业务下的双活案例少 | |
| IBM / PowerHA System | 通过硬件实现底层存储级逻辑卷复制技术实现存储同步，从而实现数据库双活 | 良好 | 硬件设备独立收费 | 独立硬件，运维相对简单 | 大集中业务下的双活案例少 | |
| ORACLE/ ASM Normal Redundancy | 通过软件实现底层存储级逻辑卷复制技术实现存储同步，从而实现数据库双活 | 良好 | 免费 | 运维简单，跟普通的数据库管理无区别 | 大集中业务下的双活案例少 | |

国家电网公司
STATE GRID
CORPORATION OF CHINA

# 目录


1 同城双活产生背景

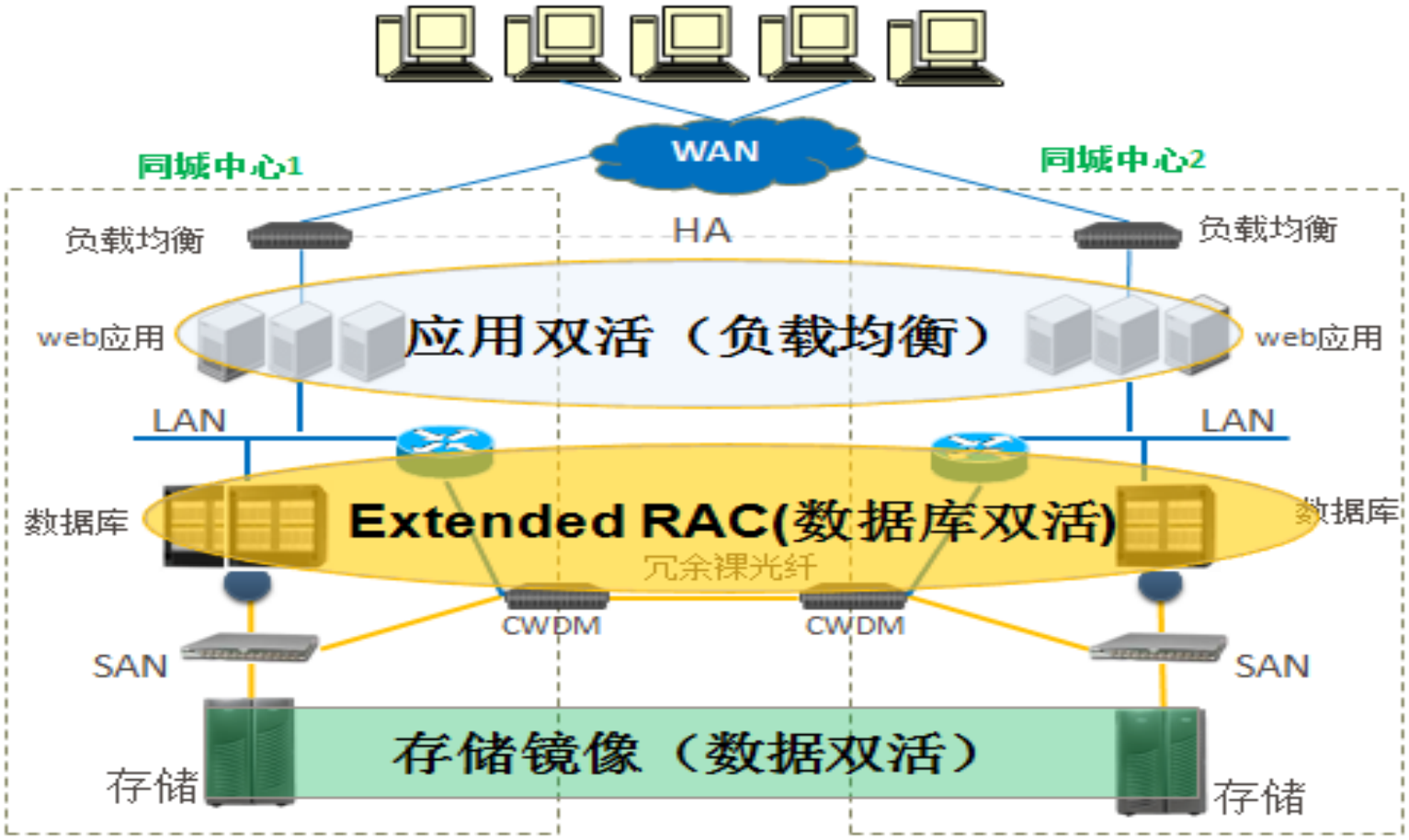
2 同城双活架构介绍

3 同城双活产品对比

4 双活应用研究介绍

5 异地容灾技术分析

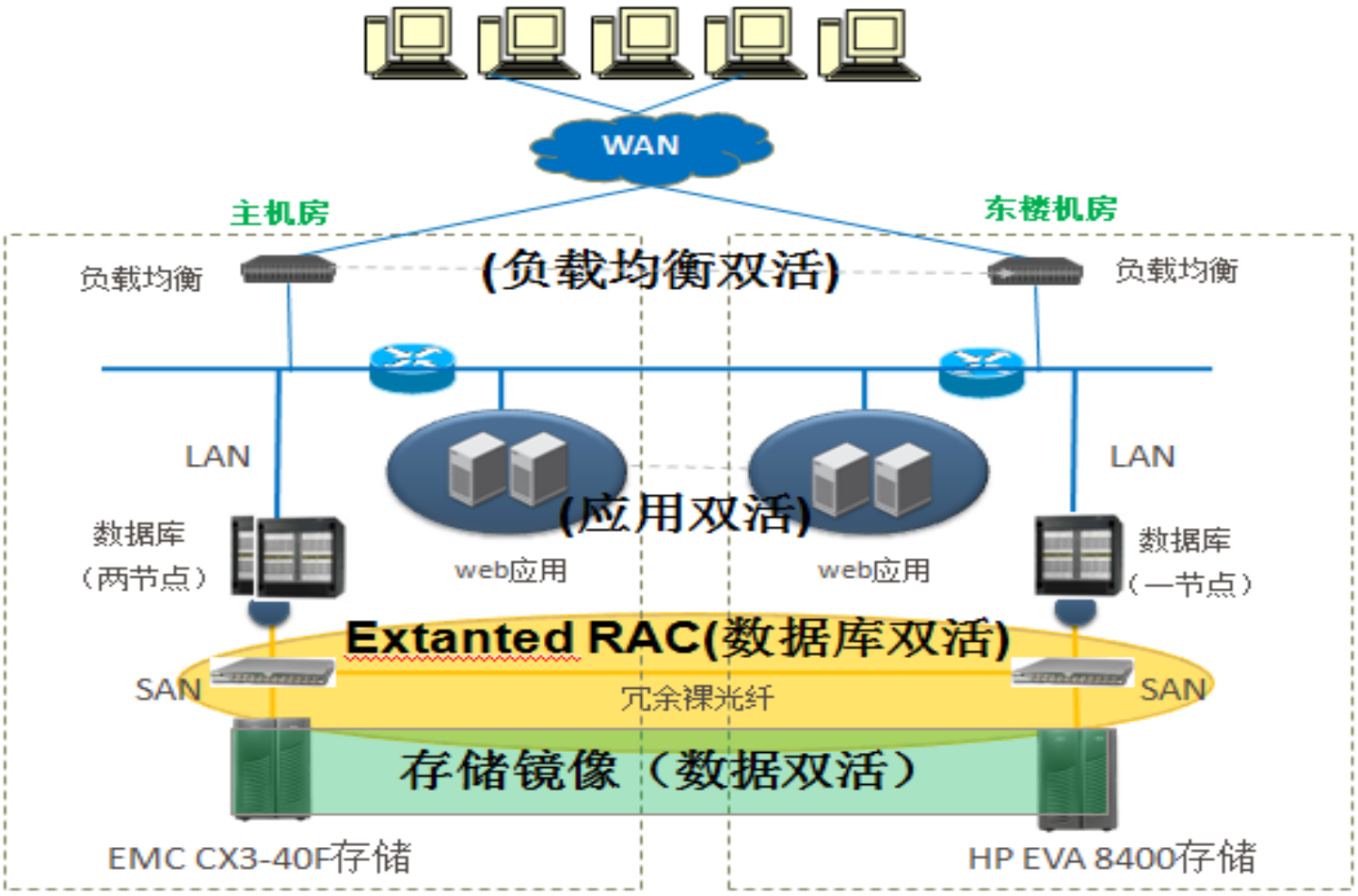

# 实验环境双活设计




备注：各中心机房设备均为双链路,避免单点故障。

国家电网公司
STATE GRID
CORPORATION OF CHINA

# 演练环境双活系统架构图

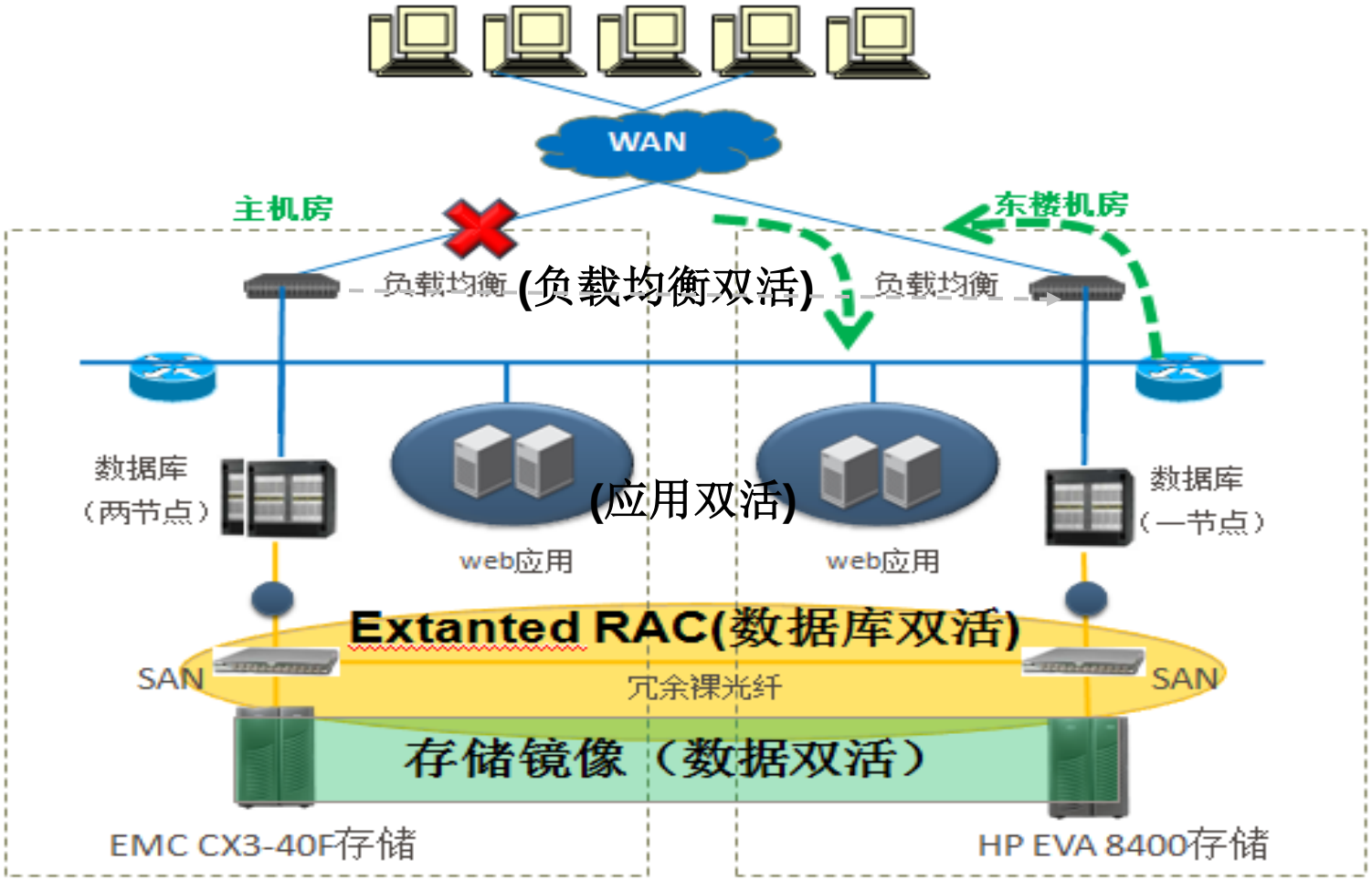



备注：目前测试环境负载均衡2台，网络交换机2台，SAN交换机2台，数据库3台。

实际的生产环境应该是负载均衡4台，网络交换机4台，SAN交换机4台，数据库4台。

国家电网公司
STATE GRID
CORPORATION OF CHINA

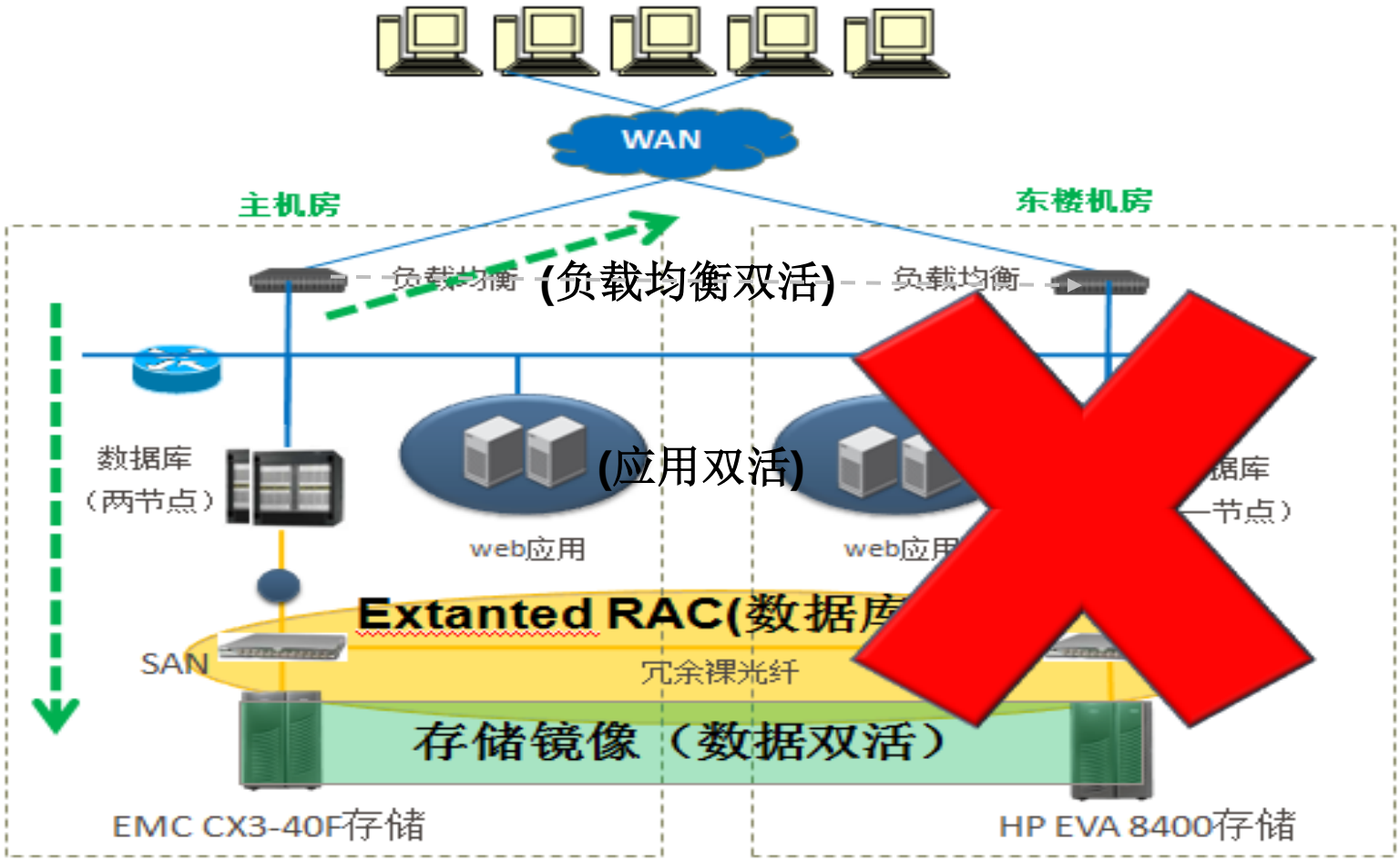
# 演练环境系统双活测试场景介绍




备注：模拟主生产中心负载均衡损坏，数据流向示意图。

国家电网公司
STATE GRID
CORPORATION OF CHINA

# 演练环境系统双活测试场景介绍




备注：模拟单个机房掉电、火灾，数据流向示意图。

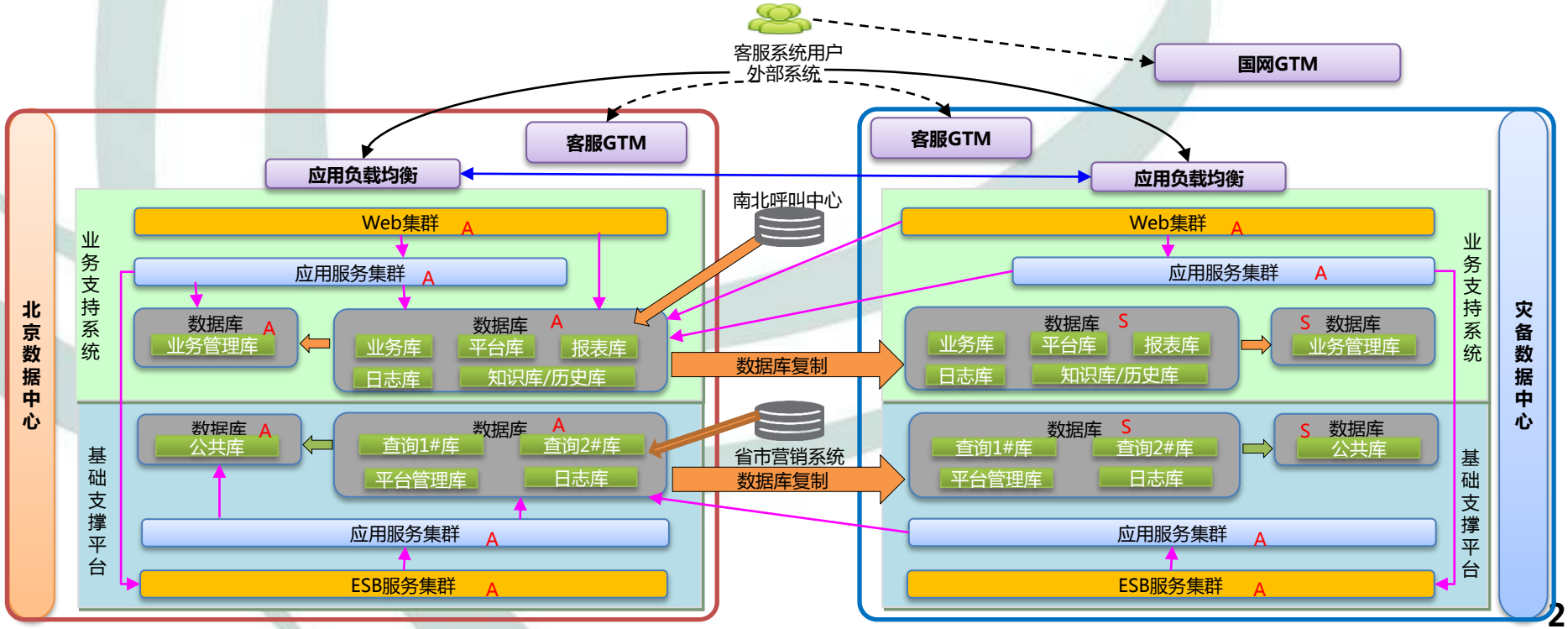
国家电网公司
STATE GRID
CORPORATION OF CHINA

# 实验环境双活各测试场景报告

| 测试场景 | 模拟损坏设备 | 业务影响 | 切换时间 | 备注 |
| --- | --- | --- | --- | --- |
| 模拟单机房同类设备全部损坏 | 负载均衡 | 无影响 | 无切换时间 | |
| | WEB服务器 | 无影响 | 无切换时间 | |
| | 数据库 | 无影响 | 无切换时间 | |
| | 存储 | 无影响 | 无切换时间 | |
| 模拟单机房全部损坏 | 负载均衡<br>WEB服务器<br>数据库<br>存储 | 无影响 | 无切换时间 | |

# XX公司准双活方案介绍

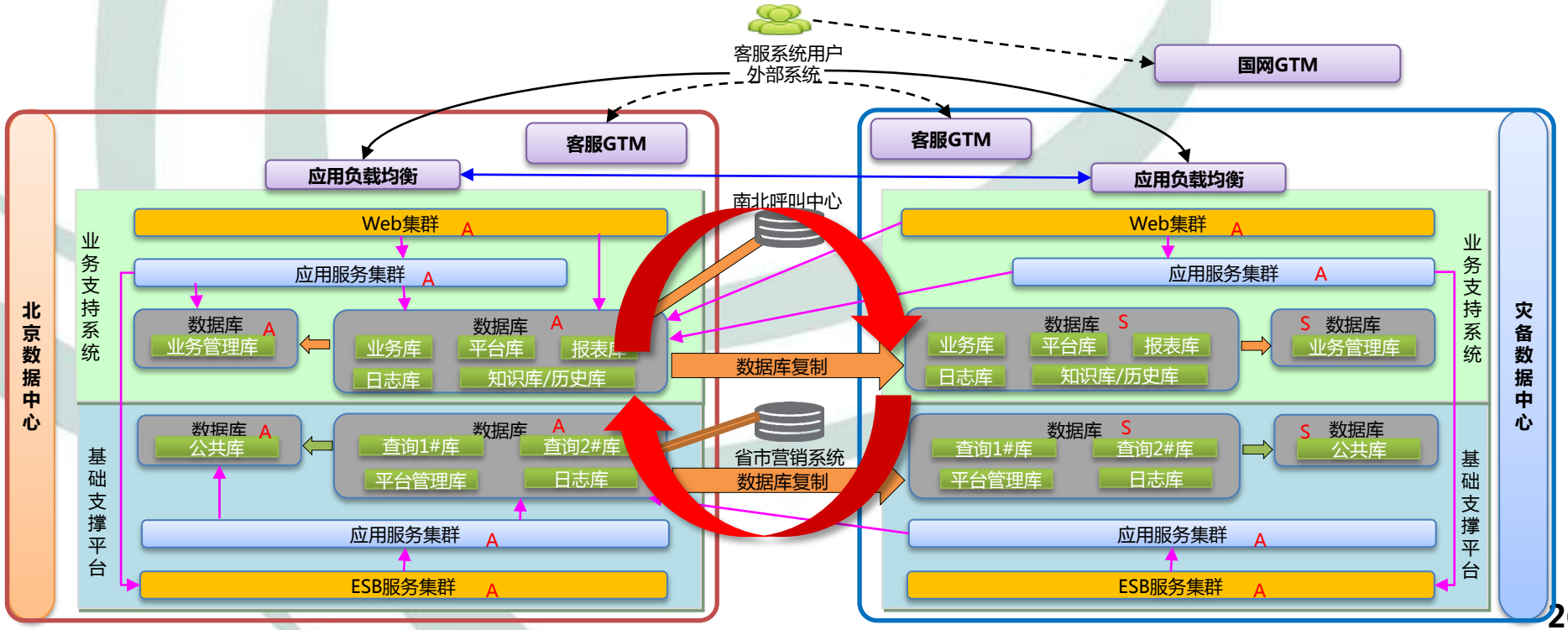
- ❖接入层：用户通过客服GTM分发业务到两个中心；
- ❖应用层：生产和灾备WEB和应用同时提供服务；
- ❖数据层：生产数据库提供两中心业务服务，灾备应用跨中心访问生产数据库；
  生产数据库通过ADG，将数据复制到灾备数据库；
- ❖资源层：在保证技术实现的基础上，灾备端使用资源虚拟化方式合理分配物理资源。



国家电网公司
STATE GRID
CORPORATION OF CHINA

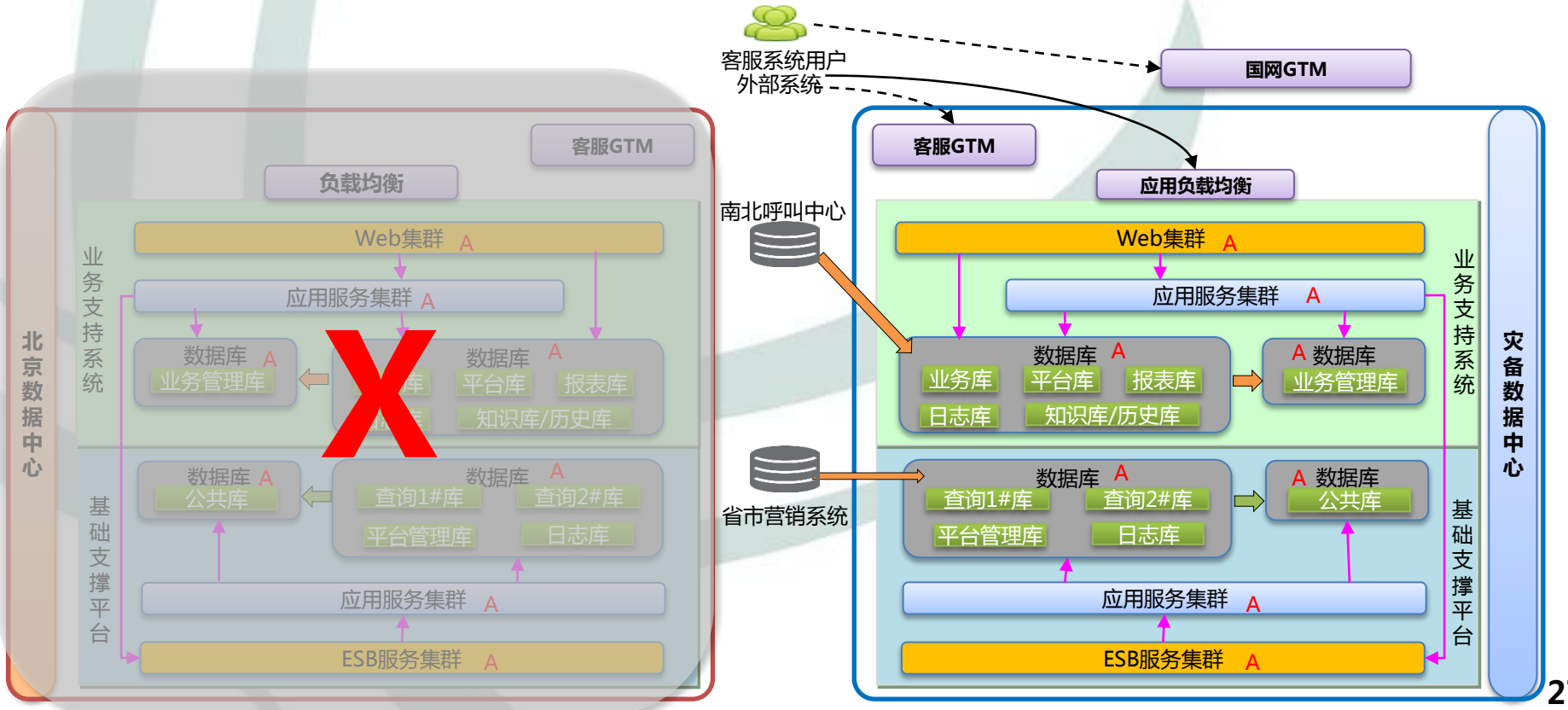
# XX公司准双活方案介绍

**日常主动运维分类：**客户端升级，应用升级，数据库升级，混合升级等
目前的灾备模式下，对于仅**应用升级和客户端升级**的场景可以采取**滚动升级**方式，无需停止业务对外服务。
一部分用户或一个中心进行升级，成功后对外提供服务。接下来在进行另一个中心升级。
可以最大限度满足主动运维场景下的，业务最大连续性保证。确保业务最大的可用性。



# XX公司客服准双活方案介绍

- 用户可以通过客服GTM自动切换和域名刷新至灾备中心；
- 灾备中心数据库进行切换，接管数据层服务；
- 南北呼叫中心、省市营销修改OGG目标端。



国家电网公司
STATE GRID
CORPORATION OF CHINA

# 目录


1 同城双活产生背景

2 同城双活架构介绍

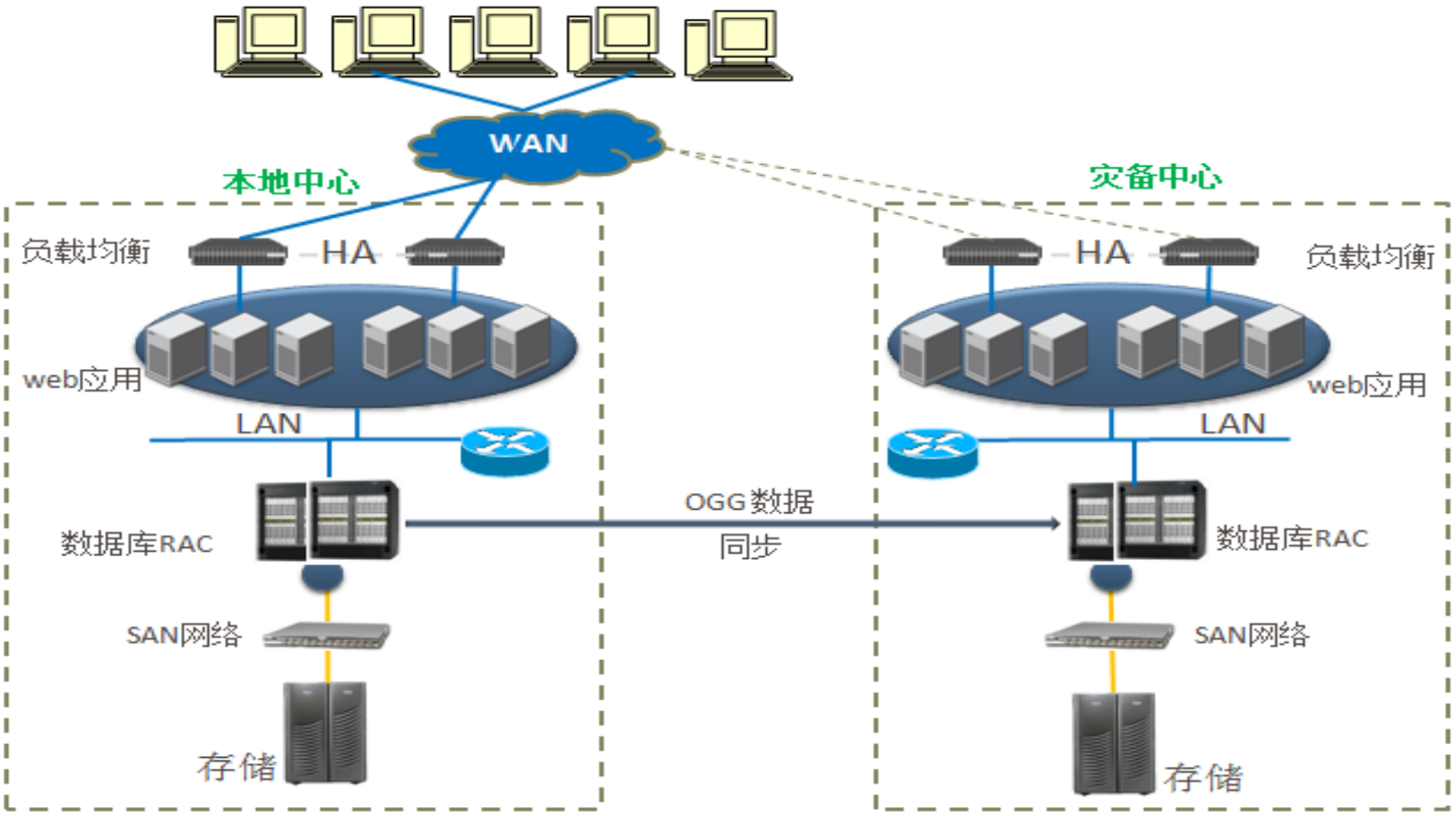
3 同城双活产品对比

4 双活应用研究介绍

5 异地容灾技术分析


国家电网公司
STATE GRID
CORPORATION OF CHINA

# 同城双活情况下的异地容灾




备注：异地容灾系统，应用层和数据层都需灾备，主要是数据库的数据要实时同步，国网现在用的是OGG，建议采用ORACLE DATA GUARD。

# 现有远程RAC技术缺陷

现有的远程RAC技术，虽然许多厂商都有解决方案，但大部分都是基于数据库底层存储的复制技术，通过实现底层存储的HA，从而达到双活的目的，这种方式有个很大的缺陷，当主存储出现数据库坏块或者存储写入异常，将会以同步方式将错误写到另一个存储。

ORACLE 10g ASM使用NORMAL 冗余的方式实现数据写入两个存储，但代价高，而且BUG较多，生产系统很少用，ORACLE官方不推荐。

综上，如果建设双活系统，建议本地或者远程配置实时的ORACLE ACTIVE DATA GUARD，避免这种物理损坏导致的不可用，从而有效保证数据库的高可用性。

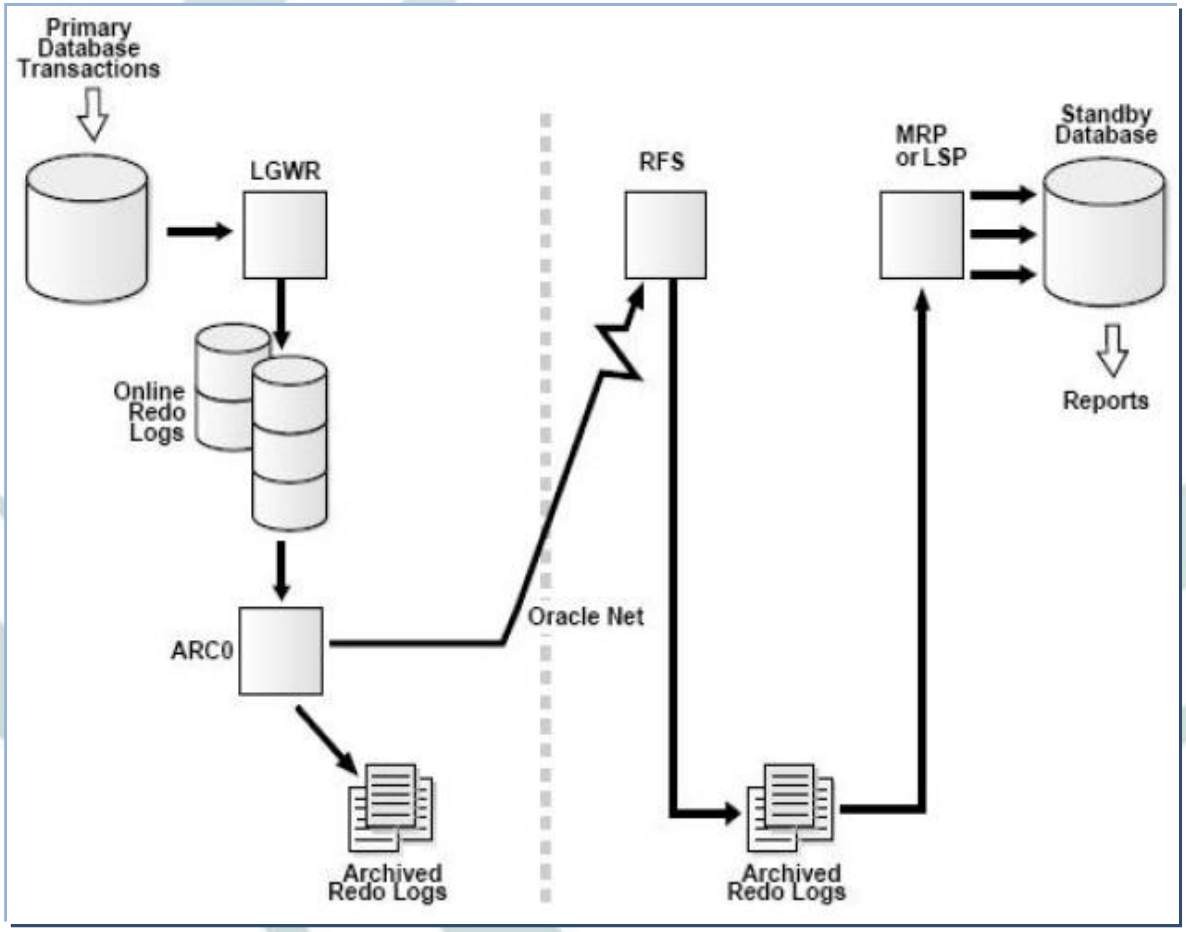
就目前成熟的案例来说，远程RAC+本地ORACLE ACTIVE DATA GUARD是属于非常高端的成熟案例，双活系统保障单生产中心故障，DATA GUARD保障ORACLE数据块级和存储写入异常导致的存储损坏。

国家电网公司
STATE GRID
CORPORATION OF CHINA

# OGG数据同步缺陷

国网现有的应用级灾备数据库容灾使用的是ORACLE OGG同步技术，该技术保障了数据的不丢失，但OGG在生产出问题需要切换时，OGG容灾库是无法保证运营的，虽然数据一致，但统计值，索引都可能跟生产不一致，直接导致执行计划不一致，从而无法正常切换，所以ORACLE推荐用ACTIVE DATA GUARD，物理结构一致,当生产出问题，将备库激活打开，这样就不存在上述问题。此外OGG正常切换，无法直接进行回切，而ADG能够正常的切换与回切。

# REDO APPLY技术




**代表：**

Oracle公司的物理 DataGuard

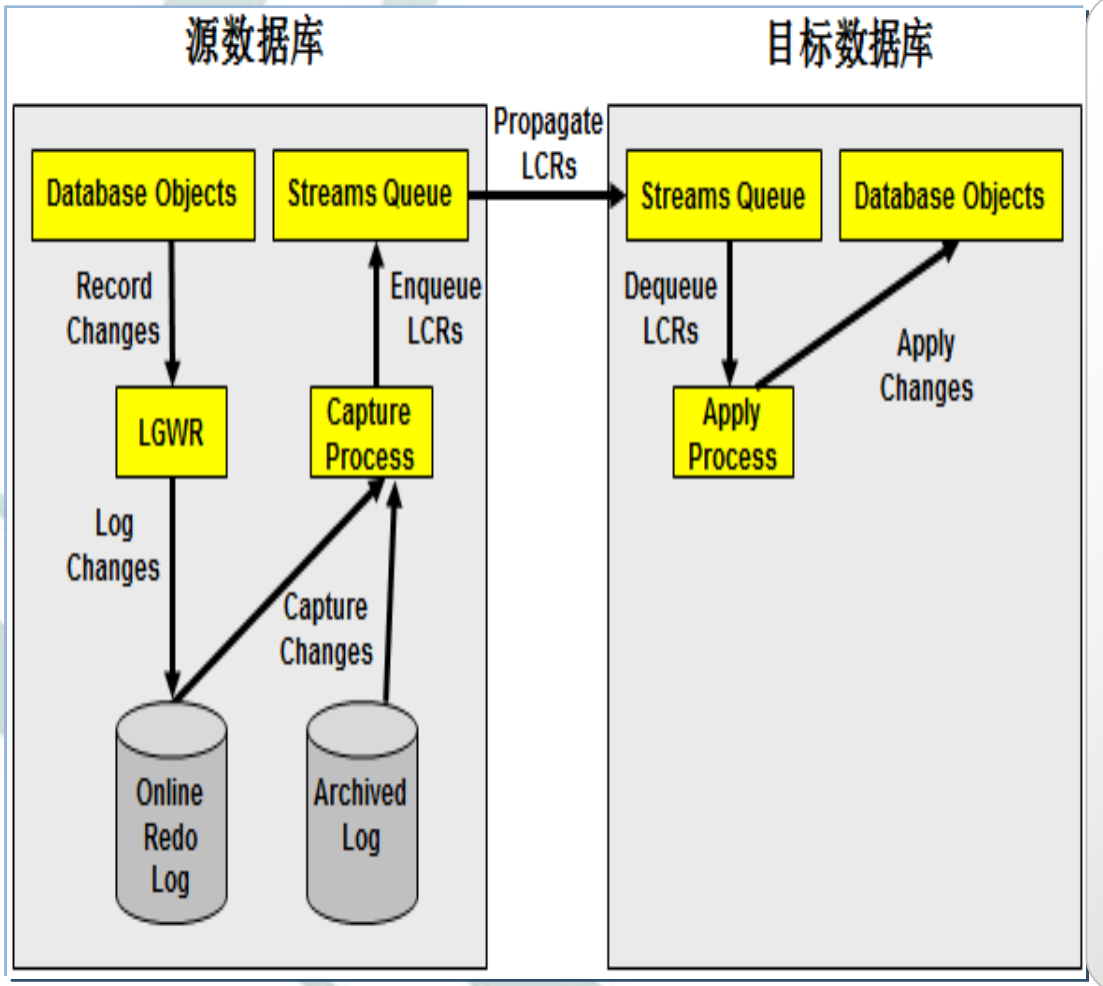
**优点：**

- 性能较好
- 不倚赖特定的存储
- 不需要很高的网络带宽
- 免费

**缺点：**

- 备库不能正常读写
- 服务器不能跨平台（11g后有限支持）

# SQL APPLY技术




**代表：**

- Oracle公司的逻辑 DataGuard、Streams、Golden Gate
- DSG公司的RealSync
- Quest公司的SharePlex

**优点：**

- 除了逻辑DataGuard外，可以跨平台
- 备库为独立的数据库，可以正常访问
- 配置灵活，可以选择性复制
- 不需要很高的网络带宽
- 可实现全功能双活复制

**缺点：**

- 性能一般，易出现较大数据延时，影响系统切换时间，实施效果取决于源系统的优化效果
- 对DDL支持不佳，由于数据字典变动而导致复制异常
- 对维护要求较高，经常需要进行数据故障处理
- 某些数据类型不支持
- 除了逻辑DataGuard和Streams外，需要单独购买

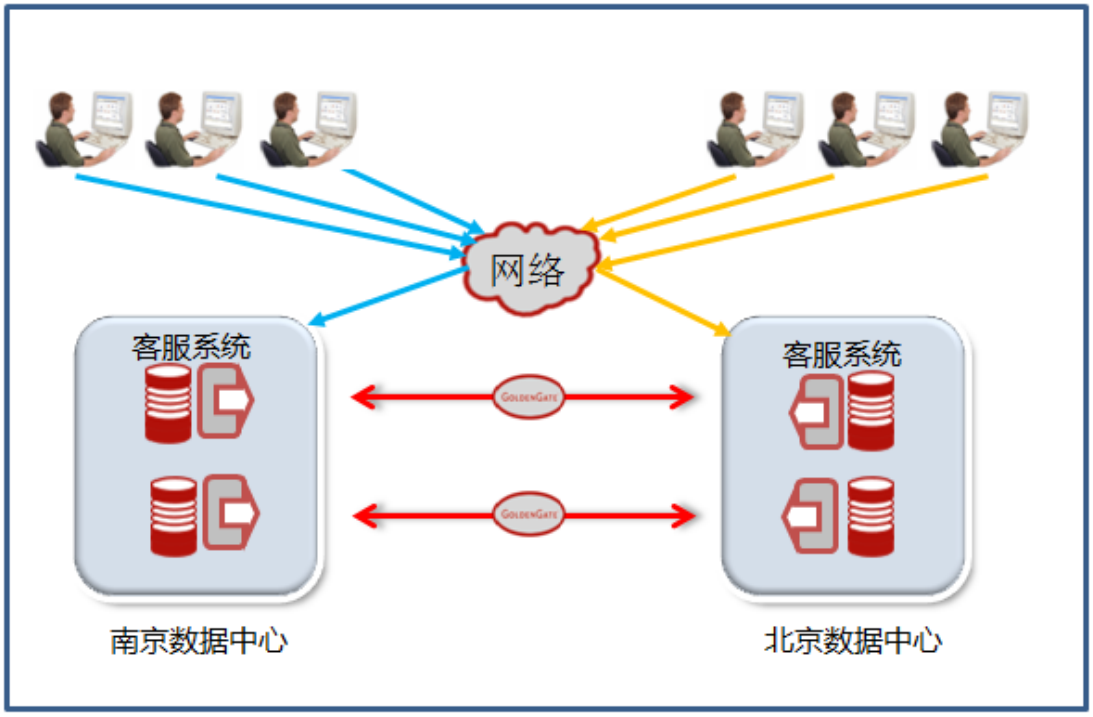
国家电网公司
STATE GRID
CORPORATION OF CHINA

# ADG与OGG方案比较

| 项目 | 物理Data Guard | DSG/OGG |
| --- | --- | --- |
| 基本原理 | 利用redo log中的物理信息进行复制，属于数据块级别的复制，备库是主库数据块的副本，主备库物理结构完全一致。 | 把redo log中的逻辑信息翻译成sql，并在备库执行，实现主备库数据的一致性。 |
| 整体性能 | 数据库的底层操作，数据延迟小 | 大事务复制效率较低，有延迟现象 |
| 数据可用性 | 备库是只读方式打开 | 备库是独立数据库，完全open |
| 事务一致性 | 始终维护事务的连续性 | 不支持 |
| 复制方式 | 有同步、异步复制方式 | 异步方式 |
| 数据压缩、传输和带宽占用 | 利用TCP/IP传输数据变化，支持日志压缩和传输压缩，带宽占用低 | 带宽占用较高 |
| 资源占用 | 主机资源占用在1%-5%之间 | 资源占用在10%左右 |
| 异构支持 | 支持异构存储，不支持异构操作系统和数据库 | 支持异构操作系统、数据库 |
| 维护工作量 | 日常维护简单 | 当应用版本不稳定时，日常维护量较大 |
| 数据类型支持 | 数据类型兼容性好 | 有限制，有些类型兼容性不是很好 |
| 主备角色切换 | 可以自由切换 | 不支持 |
| 数据转换、整合和分发 | 不支持 | 支持 |
| 数据崩溃保护 | 会在数据应用前对数据块进行验证 | 存在数据冲突问题 |
| 典型应用 | 应急（比较适合本地容灾） | 数据分发/集中、ETL（DW/BI） |

国家电网公司
STATE GRID
CORPORATION OF CHINA

# 双数据中心架构示意图




## 双数据中心

客服系统需要双业务数据中心同时对外提供业务服务。

## 容灾

在一个数据中心出现灾害时，另一个数据中心可以进行业务接管，同时需要满足必要的RTO/RPO指标。

## 业务接管方式

业务接管后，相关的服务请求将直接通过网络映射或则集群监控的方式，将相关的业务服务请求直接连到接管数据中心的服务层，避免应用服务远程访问数据库导致的服务时间延迟。

## 按业务管理区域数据分离原则

由于业务在流程上是有一定的关联关系的，同时为了减少相关业务软件逻辑调整以及不必要的数据冲突，建议将现在国家电网客服系统的应用服务按不同的省公司管理范畴进行分离，例如，南京数据中心负责南方各个省的客服系统数据服务，北京数据中心负责北方各个省。

国家电网公司
STATE GRID
CORPORATION OF CHINA

# 业务数据库调整原则

- ✓ 所有的业务数据表都必须包含PK或则非空的UI
- ✓ 禁止出现PK/UI字段的update操作
- ✓ 禁止TRIGGER的使用
- ✓ 业务数据表最好可SCHEMA分别处理，或则以按地域进行分区表的处理
- ✓ 对应流水号数据类型的处理采用SEQUENCE步长阶梯式调整，避免主键冲突
- ✓ 综合业务数据表，无法分离的，应该加上必要的逻辑判定字段，同时明确数据冲突的处理逻辑，类似，先到先得，后到先得，还是用户优先等等。

国家电网公司
STATE GRID
CORPORATION OF CHINA

# 谢　谢