本地高可用架构设计

公益讲座11：00分准时开始，请大家先浏览云技术微信公众号技术文章资料会在各群同步发布，已入群客户请勿重复入群！

20-17

数据库和云讲座群

甲骨文云技术公众号

ORACLE

本地高可用架构设计

黄嵩

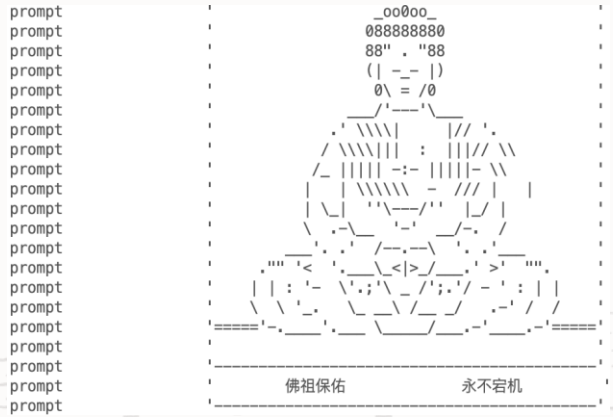
2022年6月24日

来自生活中的小故事

顾客：“上帝保佑我”
卖家：“顾客就是上帝”
顾客：“自己保佑自己？”
“画个圈圈诅咒你们……”

- 您的组织，期待您和强大IT系统的“保佑”才能合规运营，业务才能正常运转，才能完成经营目标，最终才能实现经营战略和企业使命。
- 您和您的IT系统，“灾难”来临时上帝来“保佑”？

  自己保佑自己。



为何将业务连续性作为重中之重

组织的经营战略需要

c.gb688.cn/bzgk/gb/showGb?type=online&hcno=B7DDC387ECA63A1C1CEAE15BE01E2A61

100%

附　录　C

（资料性附录）

某行业 RTO/RPO 与灾难恢复能力等级的关系示例

C.1　RTO/RPO 与灾难恢复能力等级的关系

表 C.1 说明信息系统灾难恢复各等级对应的 RTO/RPO 范围。

表 C.1　RTO/RPO 与灾难恢复能力等级的关系

灾难恢复能力等级	RTO	RPO
1	2 天以上	1 天至 7 天
2	24 小时以上	1 天至 7 天
3	12 小时以上	数小时至 1 天
4	数小时至 2 天	数小时至 1 天
5	数分钟至 2 天	0 至 30 分钟
6	数分钟	0

什么灾难与灾难恢复（DR）

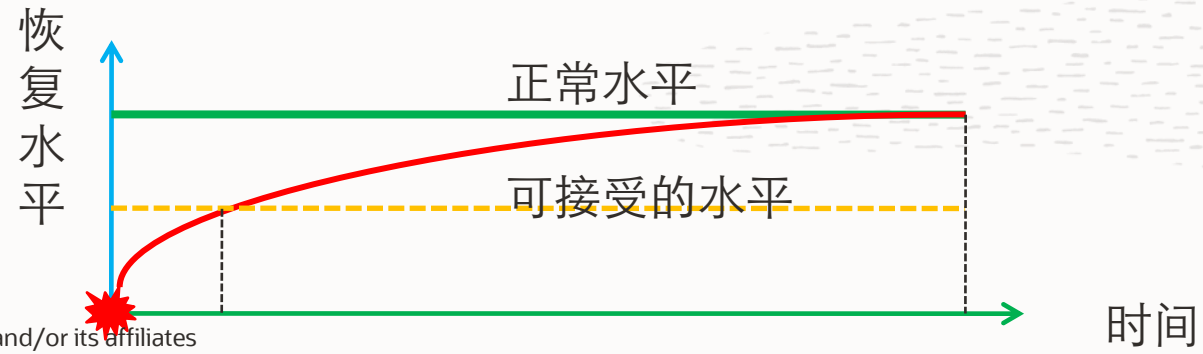
理论研究机构：
国际灾难恢复协会 DRI International
国际灾难恢复(中国)协会 DRI China

1. 灾难---由于人为或自然的原因，造成信息系统严重故障或瘫痪，使信息系统支持的业务功能停顿或服务水平不可接受、达到特定的时间的突发性事件。通常导致信息系统需要切换到灾难备份中心运行.
2. 灾难恢复 --- 为了将信息系统从灾难造成的故障或瘫痪状态恢复到可正常运行状态、并将其支持的业务功能从灾难造成的不正常状态恢复到可接受状态，而设计的活动和流程。

来源：<GB/T 20988-2007 信息安全技术 系统灾难恢复规范 >3 术语与定义

相关术语：

「业务影响分析（BIA）、业务连续性（BC）、业务连续性管理（BCM）、灾难恢复预案、灾难恢复规划、RTO、RPO等」参考：http://c.gb688.cn/bzgk/gb/showGb?type=online&hcno=B7DDC387ECA63A1C1CEAE15BE01E2A61



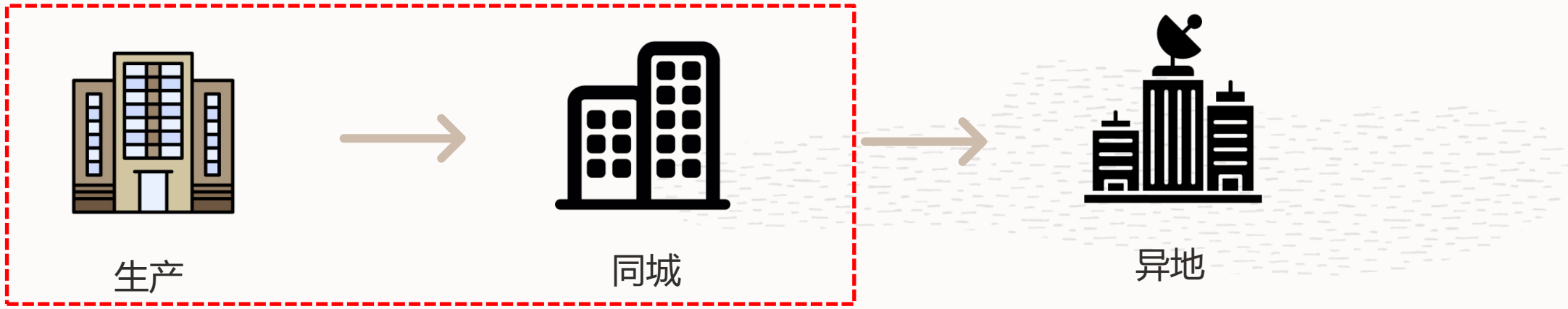
DR建设的国内标准

6级： 数据零丢失和远程集群支持	● 实现远程数据实时备份，实现零丢失 ● 应用软件可以实现实时无缝切换 ● 远程集群系统的实时监控和自动切换能力
5级： 实时数据传输及完整设备支持	●实现远程数据复制技术 ●备用网络也具备自动或集中切换能力
4级： 电子传输及完整设备支持	●配置所需要的全部数据和通讯线路及网络设备，并处于就绪状态 ●7*24 运行;更高的技术支持和运维管理
3级： 电子传输和部分设备支持	●配置部分数据,通信线路和网络设备 ●每天实现多次的数据电子传输 ●备用场地配置专职的运行管理人员
2级：备用场地支持	●预定时间调配数据,通信线路和网络设备 ●备用场地管理制度 ●设备及网络紧急供货协议
1级：基本支持	●每周至少做一次完全数据备份 ●制定介质存取、验证和转储的管理制度 ●完整测试和演练的灾难恢复计划

什么是（本地）高可用

高可用性H.A.（High Availability）指的是通过尽量缩短因日常维护操作（计划）和突发的系统崩溃（非计划）所导致的**停机时间 & 数据不一致**，以提高**系统和应用的可用性**。衡量系统是否满足高可用，就是当一个或者多台部件在计划内/外停止服务的时候，**系统整体和服务依然正常可用**。他是业务连续性的至关重要保障。

DR系统在物理架构上通常会按需体现为【两地两中心】或者【两地（生产、同城）三（异地）中心】的部署格局。今天我们探讨的（本地）高可用架构设计，范围限定于生产数据中心内及同城灾备中心。



业务连续性
-本地高可用设计方法论

总体规划
- IT与业务现状、（关键）业务影响分析
- 明确范围与目标，形成总体规划方案

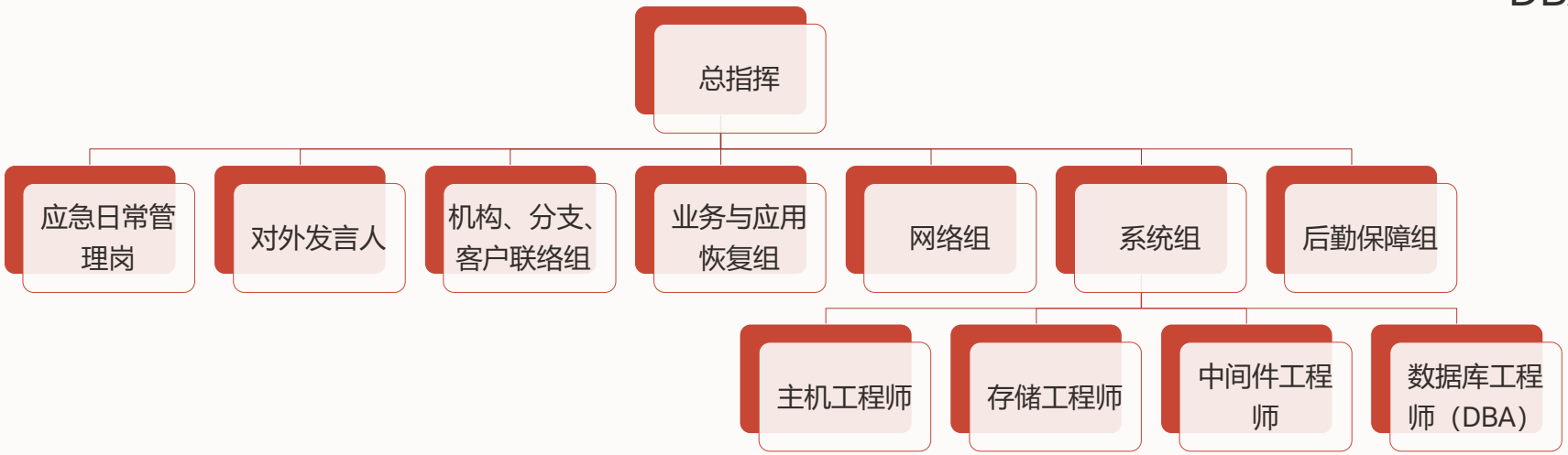
技术选型
- 确定灾备模式\数据保护策略、确定基础架构、数据复制方案等
- 必要的POC验证

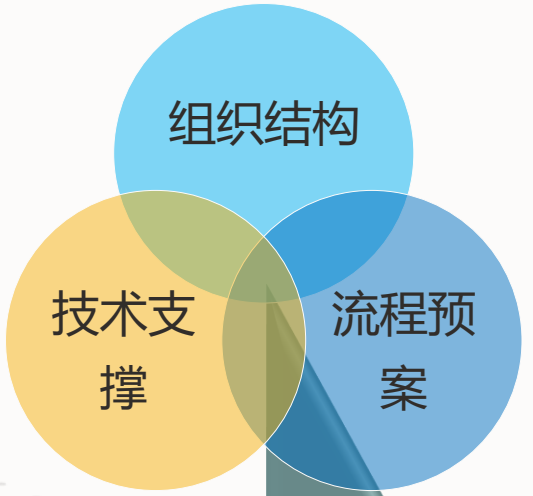
实施与应急预案
- 实施灾备技术方案，同步设计应急预案
- 明确事件范围及影响、设计事件流程（管理+技术）、演练检测

运行维护
- 日常运维监控
- 定期演练检测
- 必要的自动化工具

谁来主导业务连续性

架构师
系统管理员
DBA……？

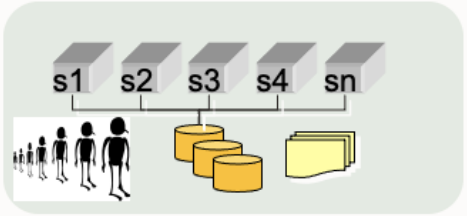




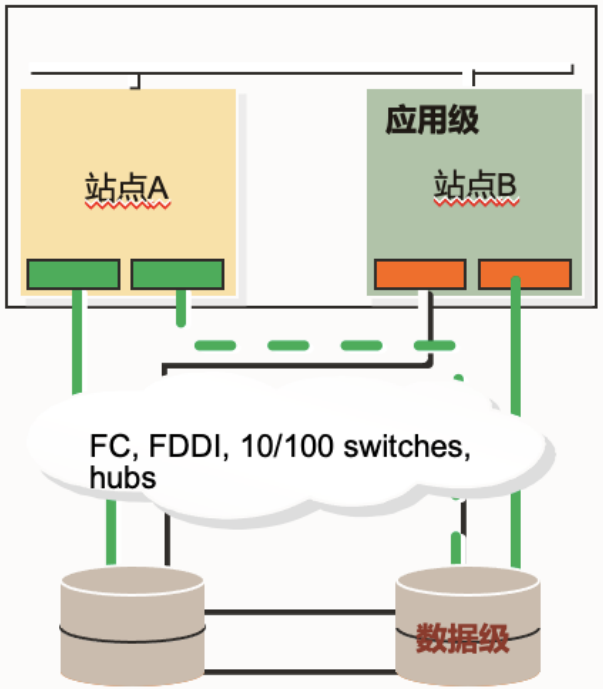
业务连续性管理（BCM）								
危机管理		应急管理		业务连续性规划（BCP）				
关联组织危机管理	危机通信及危机公关	紧急时间应急响应处置	灾难时间应急响应处置	风险分析和业务影响分析	恢复策略和方案（组织策略、业务策略、技术策略）	信息系统恢复预案	业务恢复预案	重建和回退计划

- 技术是实现载体
- 预案是行动指南
- 组织是灵魂

信息化决策者需要关注什么



业务级



组织战略	企业经营决策层	• 政策、法律法规 • 业务持续管理 • 风险,危机管理 • ……
组织架构和人员配置	客户 （业务连续性管理委员会）	• 组织架构 • 职责和责任 • 人员的意识和技能
流程管理 （业务与IT部门）	客户（业务、IT、管理部门）	• 业务应急和恢复流程 • IT应急和恢复流程 • 运维管理
应用架构 数据架构	客户（开发、运维）； **Oracle（MAA设计、演练支持、切换保障）**	• 应用系统架构 • 数据库系统设计 • 数据复制 • 备份和恢复
IT基础设施 及相关技术	客户（主机、存储、网络、安全）	• 网络架构（切换） • 时钟同步 • 安全区域划分 • 系统架构设计
数据中心基础设施	客户（机房设施）	• 机房环境设计（含电力、空调、消防等） • 环境监控 • 物理安全

风险分析（Risk Analyze）

风险分析是对当前IT环境和的面临潜在风险确分析，拟定应对风险的总体方向

主生产中心机房内事件

设备单点故障、电源系统故障、网络系统故障、机房其它设施如空调系统故障、存储子系统故障等

病毒破坏，人为事件破坏等

例行维护、升级、业务上线等

→ 机房达标，设备冗余，高可用架构、本地备份

主生产中心外部事件

建筑雷击、火灾、光缆中断等

→ 同城容灾、同城异地备份

城市级应急事件

数据中心区域的交通、电力、通讯及其它关键的IT设施遭到严重破坏
例如：自然灾害、恐怖袭击、电力及通讯光缆故障无法预估恢复时间

→ 异地容灾、跨城异地备份

《IT风险事件列表》需保持更新

业务影响分析（BIA）

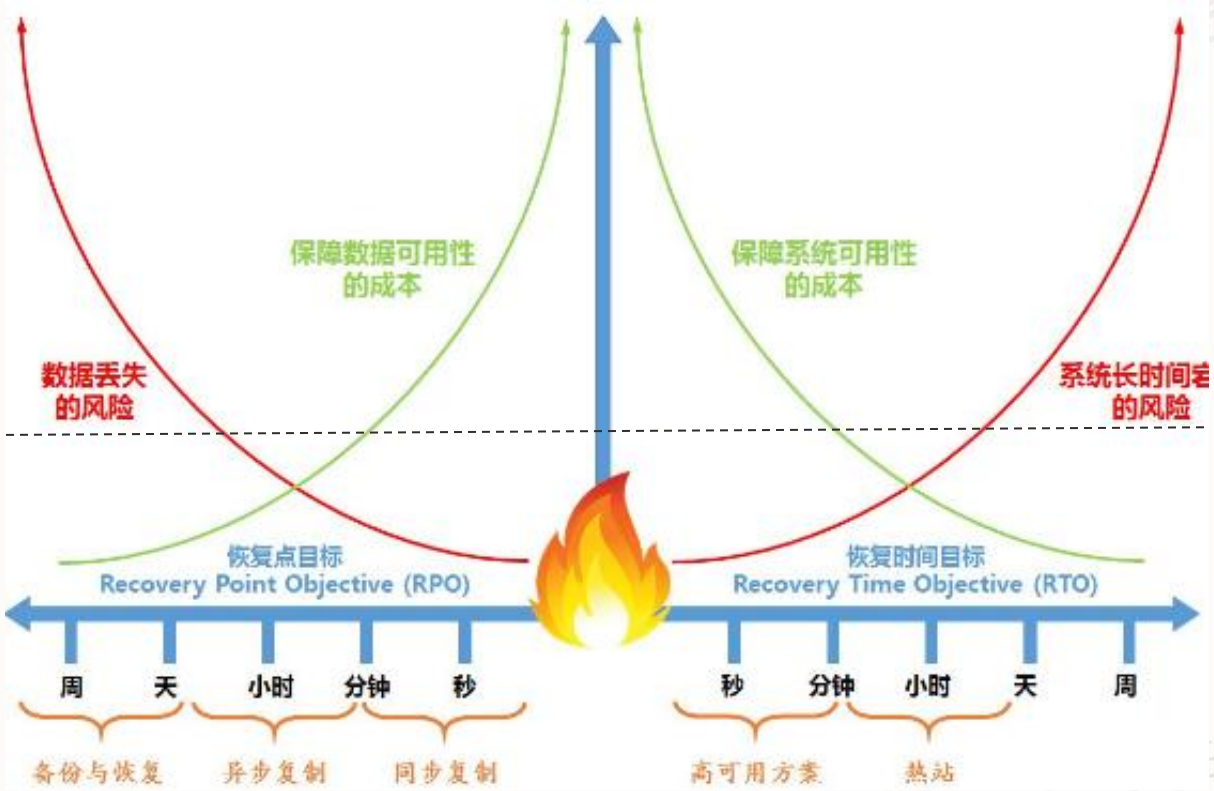
业务影响分析（Business Impact Analysis，BIA）分析业务功能及相关系统资源、评估特定灾难对各种业务功能的影响过程。

- 识别关键业务应用系统及其子模块
- 识别应用系统之间的相互关联及上下游关系（CMDB）
- 评估系统中断（时长）对关键业务的损失及影响（定性和定量）；
- 明确关键业务对连续需求（RTO和RPO）；
- 识别关键的服务时间段和可容忍的性能下降程度

业务影响分析补充：

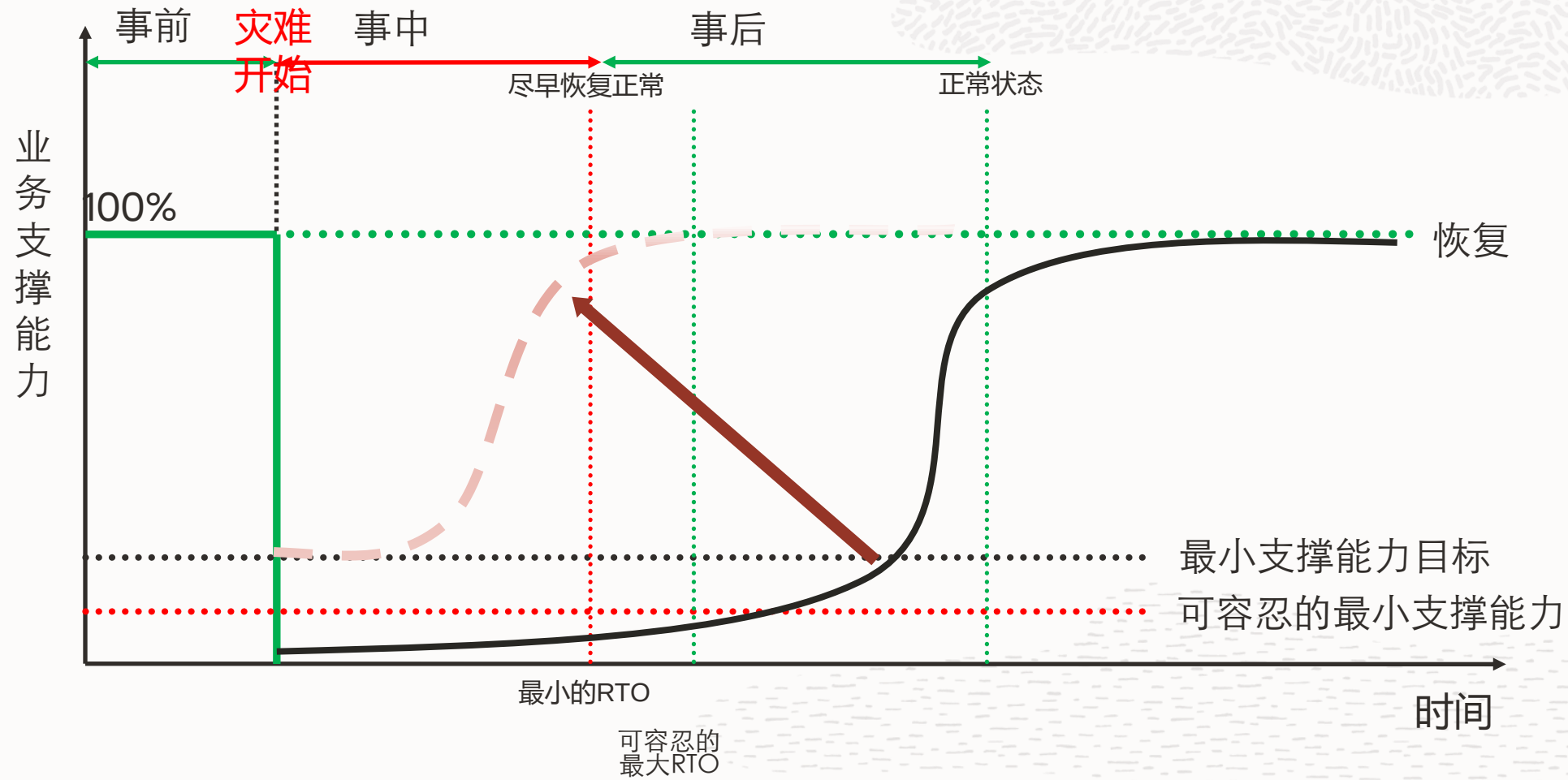
- 经济影响：直接损失、违约补偿金、账单损失、投资损失等
- 生产力损失：额外投入的员工人数*工作时间来弥补停机及数据丢失，如每个营业网点需要额外增加10人，10人天用于手工处理帐务，补录核对票据信息等。
- 声誉影响：竞争对手、客户尤其是重点价值客户、供应链、渠道、合作伙伴、证券市场等产生的负面影响。
- 财务指标影响：现金流、收入确认、应收账款、应付账款、短期盈利能力、短期抗风险能力等。

抗风险能力与投资平衡



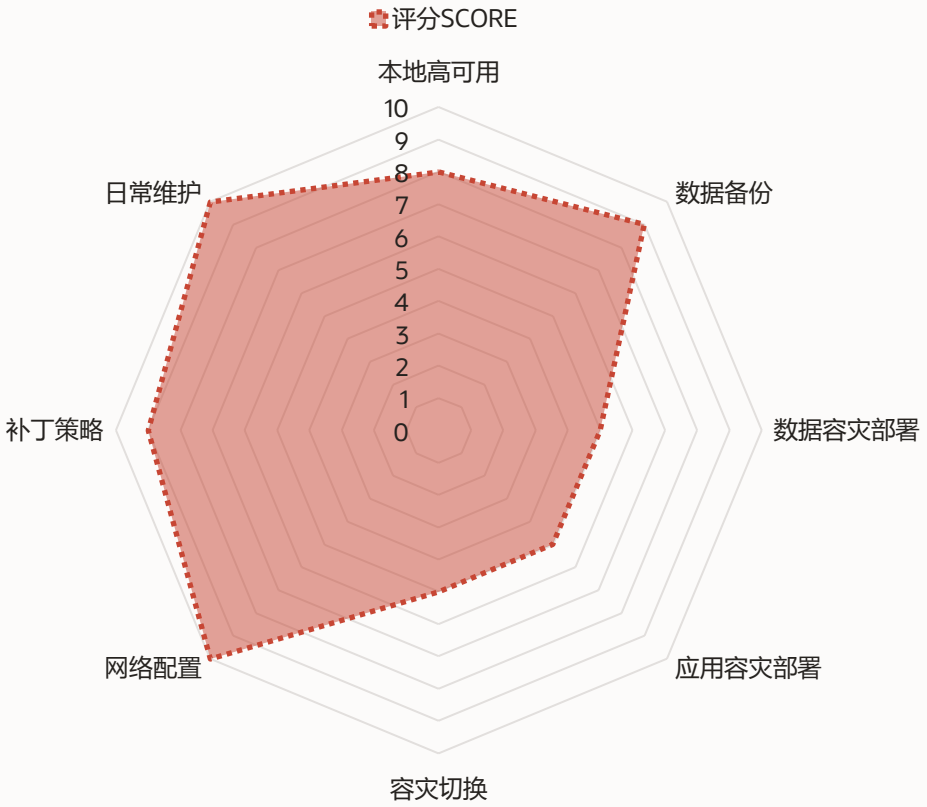
最佳平衡点衡量

业务连续性管理（BCM）的预期目标



至此，完成第一步，业务系统连续性管理总体规划

建议完成设计后与Oracle一起进行MAA架构健康风险评估



某客户规划阶段评估后的汇总

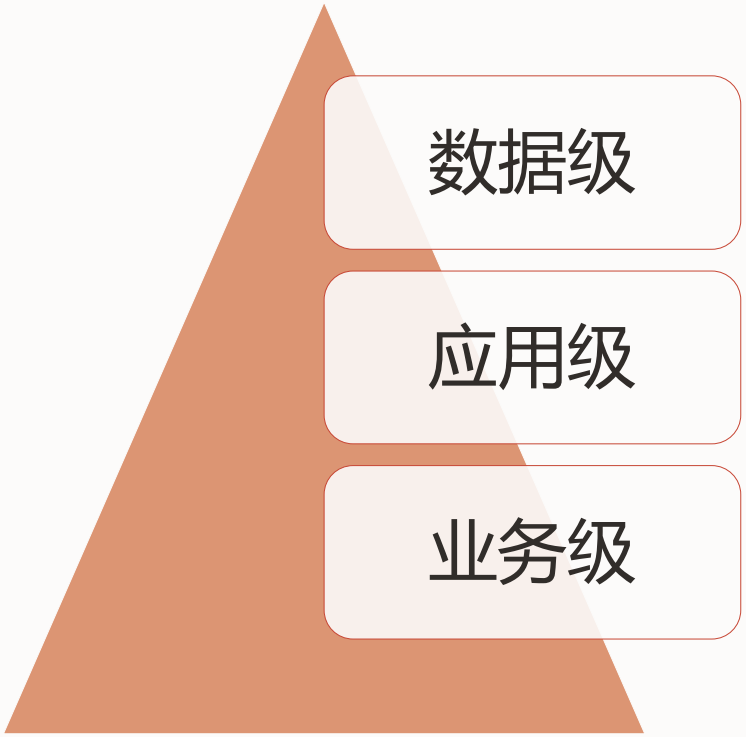
1.现有规划设计，整体表现良好,可以满足企业RTO以及RPO目标

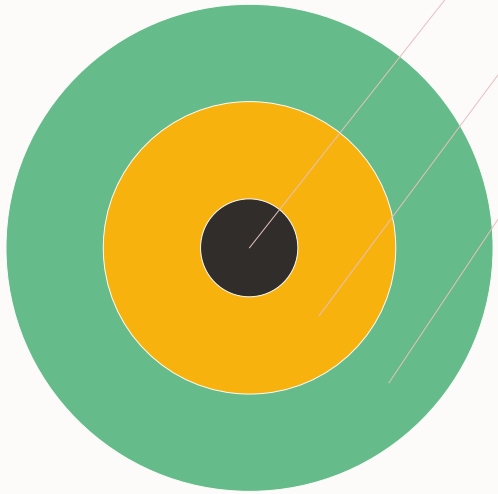
✓ 期望RTO<30分钟 RPO=0

实际的RTO & RPO （目前xx系统未正式上线，需要正式上线后实际演练获悉)

2.在本地高可用性（同数据中心内未部署本地ADG）、数据备份（目前备份方案有丢失1小时数据的风险）、数据容灾（异地容灾尚无规划）、应用容灾部署（本地应用容灾暂未规划）、补丁策略等方面有一定的提升空间

确定灾备模式

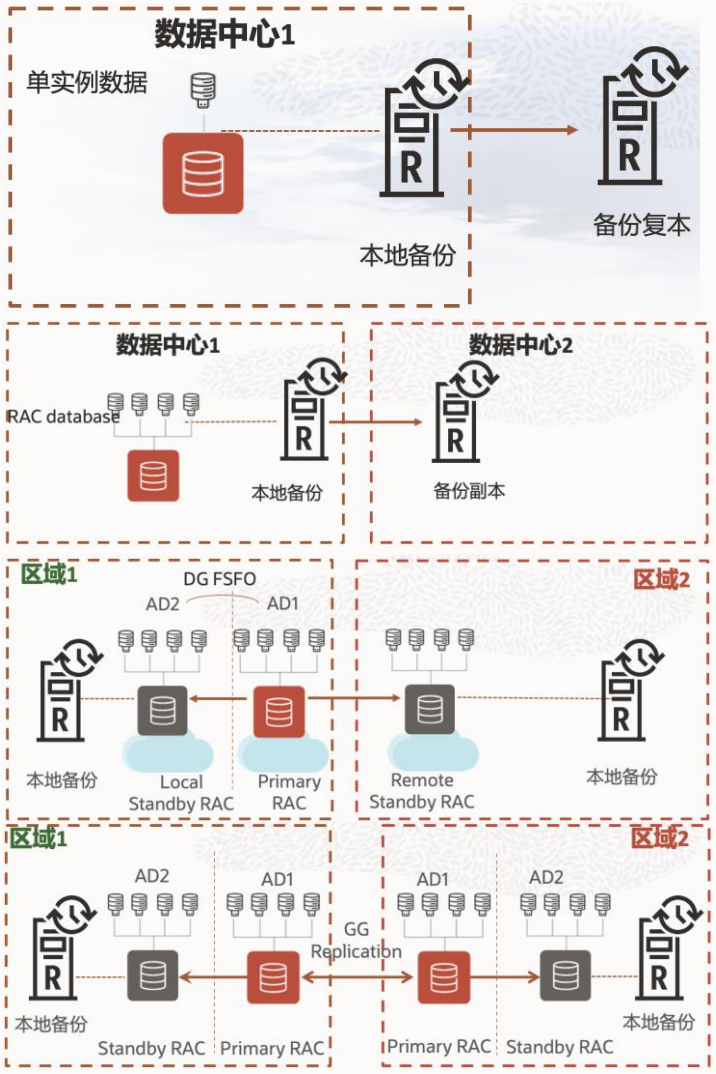




参考建议：

- 业务影响
- 业务连续性需求
- 可行的最小与最大投入
- 架构可靠性与先进性
- 可用性与可维护性
- 选最适合自己的

基础架构与数据保护



1. 数据中心可靠性、硬件冗余的设计是基本要求
2. 即使开发测试环境，系统损坏虽然不影响生产业务，但重新搭建环境准备数据也需要时间，除非开发测试部门可以中断任务。
3. **备份是最后一根救命稻草**，详情参见《数据备份与零数据丢失方案探讨》

功能	物理块损坏	逻辑块损坏
Dbverify, Analyze	物理块检查	块内和对象间一致性的逻辑检查
RMAN, ASM	物理块检查	块内逻辑检查
Active Data Guard	• Standby的连续物理块检查 • 强隔离，防止单点故障 • 自动修复物理损坏 • 自动数据库故障转移（写丢失保护）	• **检测写丢失损坏、自动关机和故障转移** • **Standby的块内逻辑检查**
Database	内存块和重做校验	内存块内检查，写丢失(shadow)保护
ASM	使用Extent Pairs自动损坏检测和修复	
Exadata	写入时HARD检查,自动磁盘擦洗和修复	写入时HARD检查

关于数据保护的最佳实践参考

1. ASM Technical Best Practices (Doc ID 265633.1)
2. **Best Practices for Corruption Detection, Prevention, and Automatic Repair - in a Data Guard Configuration (Doc ID 1302539.1)**
3. JDBC/thin Application Fails to Connect with ORA-01033 after Dataguard Switchover (Doc ID 2129131.1)
4. 使用DGMGRL(Dataguard Broker 命令行)执行12c Dataguard Swithover的最佳实践 (Doc ID 2440140.1)
5. How To Configure Client Failover For Data Guard Connections Using Database Services (Doc ID 1429223.1)

SERVICE=xxxdb LGWR SYNC AFFIRM VALID_FOR=(ONLINE_LOGFILES, PRIMARY_ROLE) DB_UNIQUE_NAME=xxxdb

SERVICE=RDRxxxDB LGWR ASYNC NET_TIMEOUT=60 COMPRESSION=ENABLE VALID_FOR=(ONLINE_LOGFILES, PRIMARY_ROLE) DB_UNIQUE_NAME=RDRxxxDB

选择合适的数据复制方案

基于存储的数据复制

通常为异步复制，同步复制须评估对性能的影响，主机层少量资源开销

通常要求主备同构，甚至同型号，成本高。

支持数据库及非数据库场景的数据复制

对带宽和延迟要求较高

校验数据通常需要停止复制

主端的损坏也会带入备端

复杂场景下的RTO和RPO是较大挑战

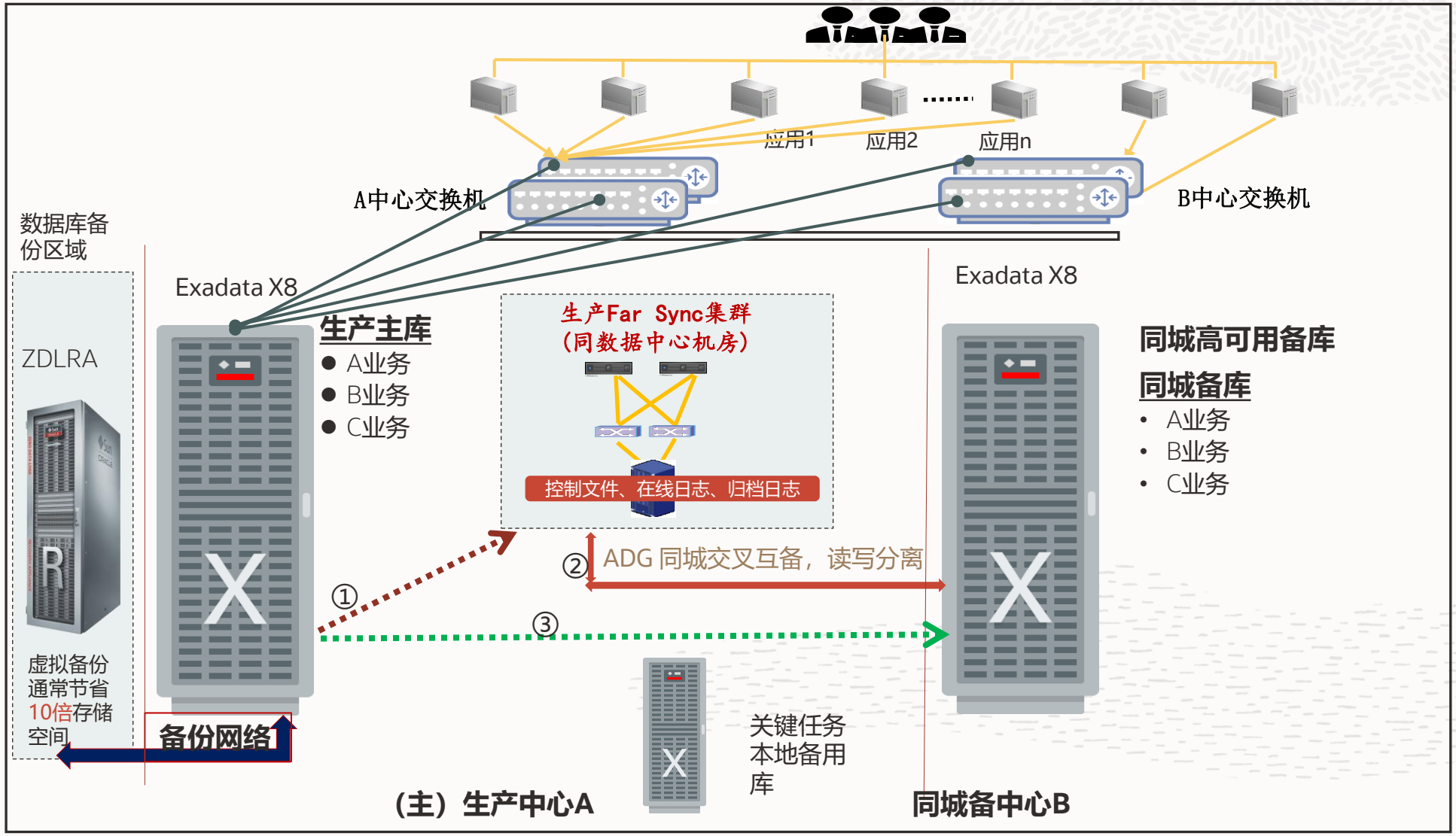
数据库软件的数据复制技术

基于数据库日志同步原理，实时或延迟应用日志
支持数据库内部数据复制
基础架构不要求同型号
数据实时校验，可用性高

参考原则

1. 数据复制以满足RPO、RTO目标为首要因素。
2. 其次须兼顾生产系统系统性能，确保用户体验。
3. 数据复制必须保障数据可用性。
4. 数据复制须在线数据验证。
5. 数据复制兼顾投资保护。

本地高可用架构案例参考



Far Sync Configuration Best Practices (Doc ID 2678239.1)

本地高可用-Service

具体配置需根据实际情况灵活调整:

透明应用连续性（Transparent Application Continuity，要求19C)

$ srvctl add service -db mydb -service coredb -preferred coredb,coredb2 - failover_restore **AUTO** -failoverretry 1 -failoverdelay 3 -commit_outcome **TRUE** -failovertype **AUTO** -replay_init_time 600 -retention 86400 -notification TRUE -drain_timeout 300 -stopoption IMMEDIATE

基于DG环境的TAC示例:

$ srvctl add service -db mydb -service coredb -preferred coredb1 -available coredb2 -failover_restore AUTO -failoverretry 1 -failoverdelay 3 - commit_outcome TRUE -failovertype AUTO -replay_init_time 600 -retention 86400 -notification TRUE -role PRIMARY /PHYSICAL_STANDBY -drain_timeout 300 - stopoption IMMEDIATE

应用连接串配置

将推荐的连接串与内置的超时，重试和延迟一起使用，以便传入的连接在中断期间不会看到错误。

要求Oracle driver 版本 12.2及以上:

```
Alias (or URL) =
  (DESCRIPTION =
   (CONNECT_TIMEOUT=90)
   (RETRY_COUNT=50)(RETRY_DELAY=3)(TRANSPORT_CONNECT_TIMEOUT=3)
   (ADDRESS_LIST =
     (LOAD_BALANCE=on)
     (ADDRESS = (PROTOCOL = TCP)(HOST=primary-scan)(PORT=1521)))
   (ADDRESS_LIST =
     (LOAD_BALANCE=on)
     (ADDRESS = (PROTOCOL = TCP)(HOST=standby-scan)(PORT=1521)))
  (CONNECT_DATA=(SERVICE_NAME = <您的服务名>)))
```

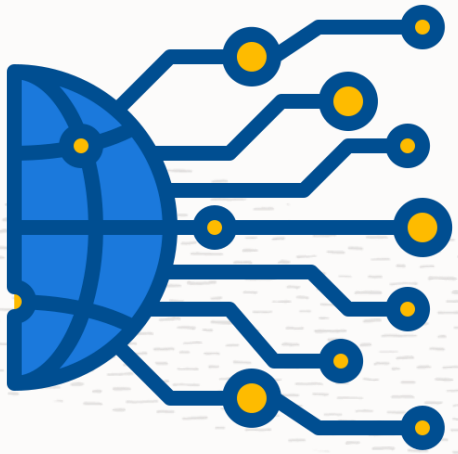
网络带宽评估

How To Calculate The Required Network Bandwidth Transfer Of Redo In Data Guard Environments (Doc ID 736755.1)

Redo Transport Compression in a Data Guard Environment (Doc ID 729551.1)

Assessing and Tuning Network Performance for Data Guard and RMAN (Doc ID 2064368.1)

- 用于计算网络带宽的公式（假设保守的 TCP/IP 网络开销为 30%）是：
- Required bandwidth = ((Redo rate in Megabytes per sec. / 0.70) * 8)= bandwidth in Mbps
- RAC环境下，需要考虑每个计算节点的传输带宽需求
- 同时评估Oracle之外的数据传输需求
- 月初、月中、月末、结息日、年结等在重要业务高峰期的峰值
- 日志压缩对CPU资源的开销须考虑
- 网络的冗余须慎重决策（多链路，多运营商）
- 距离、延迟与投入资金的平衡



设计网络切换

网络切换的本质是对外服务IP地址（服务）的跨中心转移，核心需求是网络切换过程对应用程序完全透明；无需在应用端进行任何配置改变。

基于IP地址的网络切换模式

- 主要适用场景：

以IP方式对外提供服务访问的子系统

- 要点：

1. 生产中心和备用中心对应子系统的对外服务网络采用相同的IP地址保证网络切换前后，对机构侧应用程序完全透明，无需在应用侧进行任何改变.

2. 在正常情况下，由生产中心IP地址对外提供业务服务，备用中心IP地址处于shutdown状态.

3. 当两中心之间需要进行网络切换时，

- 首先关闭生产中心的对外服务IP地址
- 然后激活备用中心的IP地址
- 当路由正常生效以后，转发到原生产中心的业务请求就自动流转到备用中心，从而实现中心间网络切换

基于DNS域名服务的网络切换模式

- 主要适用场景：

以DNS方式对外提供服务访问的子系统。

- 关键点：

1. 正常情况下，DNS服务器将业务子系统的域名解析为生产中心的IP地址，由生产中心对外提供业务服务

2. 当主备中心之间需要进行业务切换时：

首先，改变DNS服务器的A记录，使同一个域名重新解析为备用中心的IP地址

当DNS cache刷新之后，原来转发到生产中心的业务请求就会自动流转到备用中心，从而实现中心间网络切换

时钟同步服务

主、备系统的系统时间不一致，可能会导致业务切换后，支付、信用卡、计费、报送等对时间敏感类业务程序出现异常。

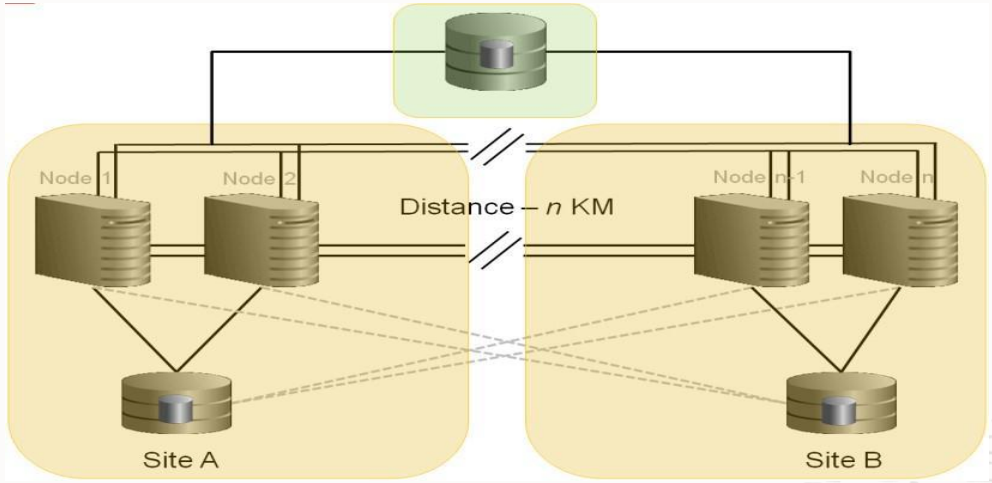
引入时钟同步，保证主、备中心的相关服务器的系统时间一致。

外部时钟源:

采用北斗 或 GPS全球定位系统作为时钟源

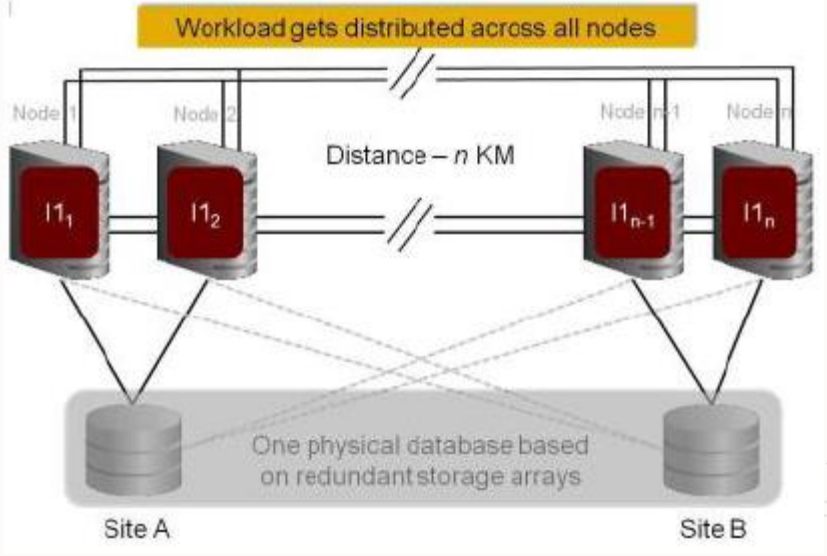
关于双活（Active-Active）架构的探讨

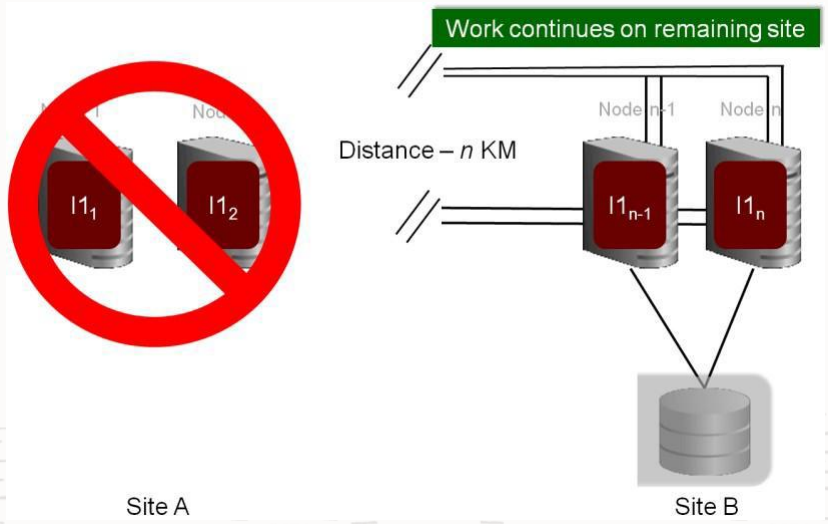
Extend RAC也叫远距离集群、扩展集群，其中的大多数或全部节点并不都在本地，通常相互之间有一定的距离。这种集群有许多名称，包括“园区集群”、“城域集群”、“地域集群”、“扩展集群”和“远距离集群”。



双活（Active-Active）架构知多少

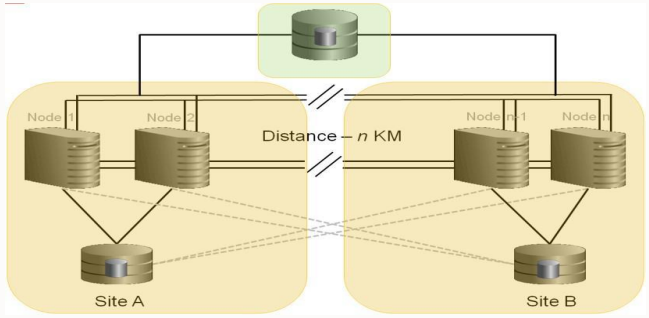
- 一个站点故障，另外一个站点自动接管
- 资源全面利用
  - 能够将全部运算分布到所有节点
  - 整体上是一套数据库，没有任何数据刷新延迟
  - 换言之，视为HA，不作为DR，也需要额外的DR保护，延迟是因为需要等待





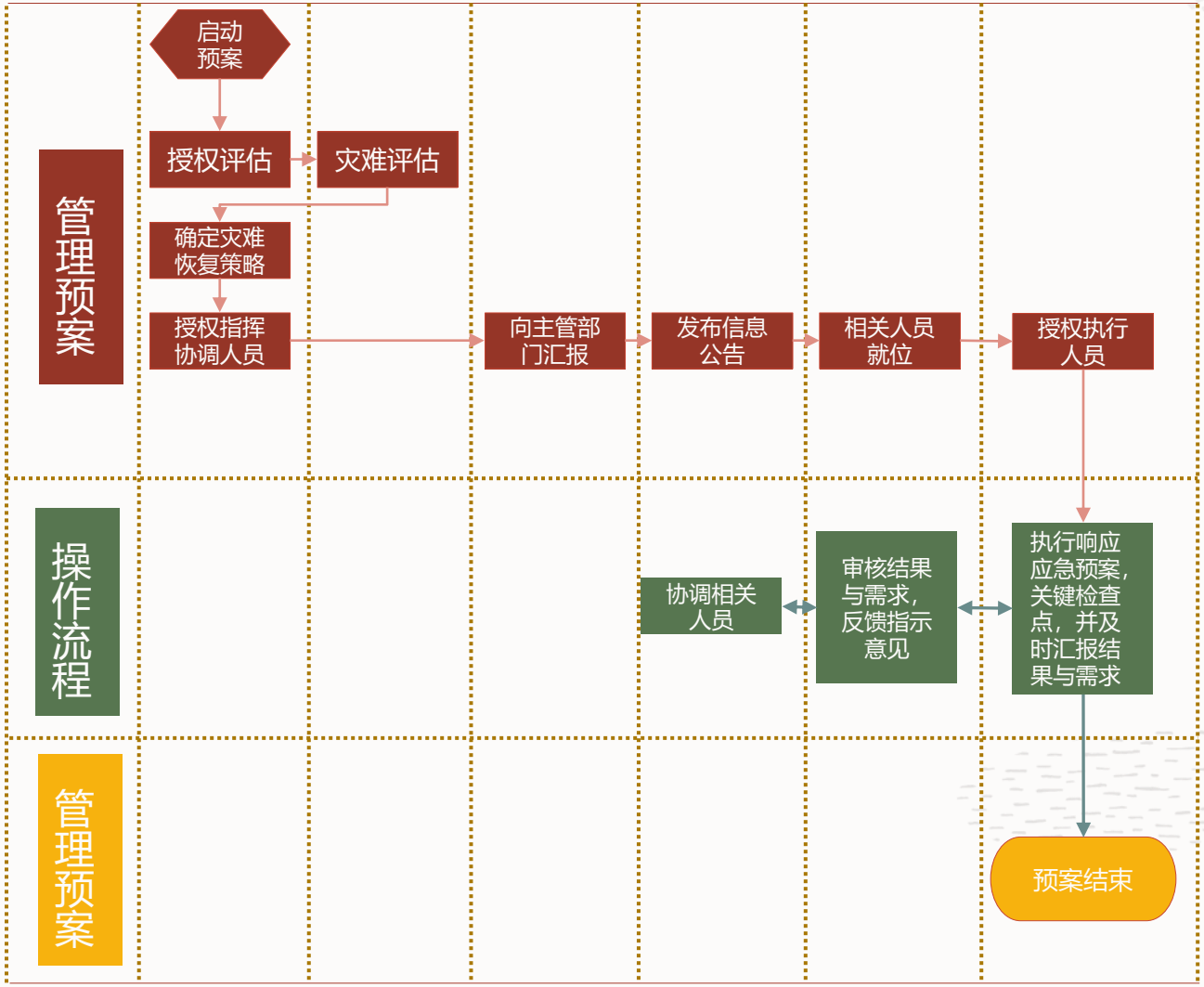
双活（Active-Active）架构需要考虑的因素

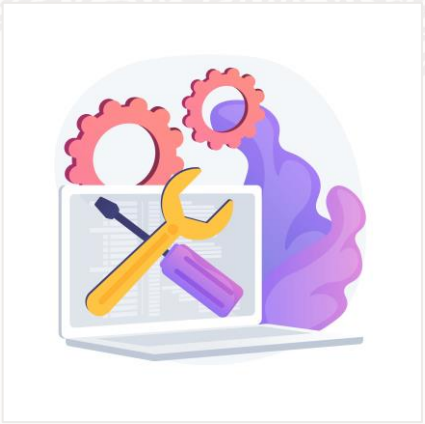
- 通常将一组节点置于站点 A
- 通常将另一组节点置于站点 B
- 需要在第三个站点放置仲裁表决磁盘
- 要求在节点和站点之间使用快速专用连接实现 Oracle RAC 实例间的通信
- **心跳网、存储数据传输网质量与效率直接影响整体性能与稳定性（流量越大，并发越大，影响越显著）**
- 距离限制（网络稳定性与延迟）



Tablespace	Reads	Av Rds/s	Av Rd(ms)	Av Blks/Rd	1-bk Rds/s	Av Writes(ms)	Buffer Waits	Av Buf Wt(ms)
APPS_TS_TX_DATA	4,031,975	1,129	41.45	12.48	265	737.85	1,309,837	15.65
CUD	576,317	161	125.34	1.02	14	640.04	1,408	6.76
APPS_TS_TX_IDX	234,381	66	17.19	1.00	64	582.56	83,617	67.23
XXT_DATA_D	239,147	67	117.92	3.13	5	940.00	0	0.00
XXT_INDEX_X	38,432	11	136.59	1.00	0	376.14	0	0.00
SYSTEM	27,947	8	39.03	1.38	7	433.05	13	55.38
TEMP1	13,139	4	7.66	1.43	0	204.41	0	0.00
UNDOTBS1	1,776	0	29.56	1.00	0	400.45	110	39.55
UNDOTBS2	7,414	2	21.65	1.00	2	2,110.00	151,084	1.90
SYSAUX	1,947	1	19.31	2.69	0	340.40	0	0.00
APPS_TS_INTERFACE	822	0	30.05	1.21	0	493.39	159	409.50
APPS_TS_SEED	1,462	0	22.72	1.13	0	340.24	117	71.28
APPS_TS_ARCHIVE	254	0	47.44	1.00	0	516.17	13	0.77
APPS_TS_MEDIA	177	0	272.03	1.25	0	357.93	0	0.00
XXT_DATA_D2	352	0	177.33	1.00	0	624.12	19	0.00
CTXD	180	0	58.72	1.07	0	465.17	0	0.00
APPS_TS_QUEUES	72	0	126.81	1.17	0	349.09	0	0.00
TEMP2	40	0	14.25	1.00	0	255.49	0	0.00
APPS_TS_SUMMARY	23	0	156.96	1.00	0	419.68	0	0.00
APPS_TS_TOOLS	7	0	585.71	1.00	0	95.00	0	0.00
APPS_UNDOTS1	5	0	662.00	1.00	0	1,055.00	0	0.00
TBS_GGS	6	0	548.33	1.00	0	240.00	0	0.00
APPS_TS_NOLOGGING	2	0	30.00	1.00	0	410.00	0	0.00
ODM	2	0	425.00	1.00	0	20.00	0	0.00
OLAP	2	0	30.00	1.00	0	20.00	0	0.00
OWAPUB	2	0	30.00	1.00	0	20.00	0	0.00
PORTAL	2	0	30.00	1.00	0	20.00	0	0.00

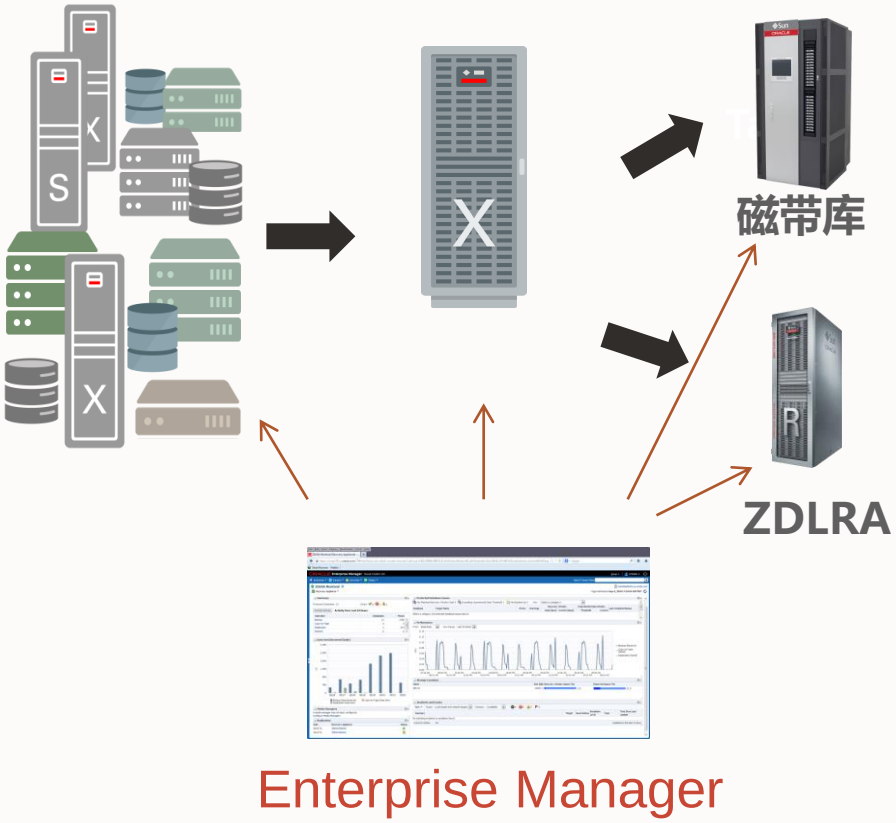
应急预案与演练





演练完成后，更新MAA健康风险模型评分，明确改进的方向！

运行维护建议



EM 12c, EM 13c: How to Manually add Database Target in Cloud Control (Doc ID 1396361.1)

建议使用Enterprise Manager作为基于Oracle的管理方案

- 加快事件解决速度、使用动态 Runbooks和智能事件压缩
- 识别数据库性能变化通过自动工作负载分析
- 为容量预测和性能趋势创建EM Warehouse
- 新大批量运维的界面简化修补并加强安全性

BTW，来自MAA评估总结的一点分享：

补丁基线管理：建议每半年或一年评估并更新一次补丁基线，可寻求Oracle帮助完成补丁评估及版本更新。

总结

1 业务连续性来自于严苛的合规要求及业务连续运营需求

2 风险分析、业务影响分析是高可用设计的重要依据。
演练是检测高可用成效的磨刀石。

3 企业MAA架构健康风险评估，展现业务连续性健康状况，找出业务连续提升方向

数据库和云讲座群

20-17

甲骨文云技术公众号

即刻扫描二维码
与甲骨文技术专家1V1深入交流