ToprowDB 高可用性、灾备、数据复制技术及应用

华胜信泰

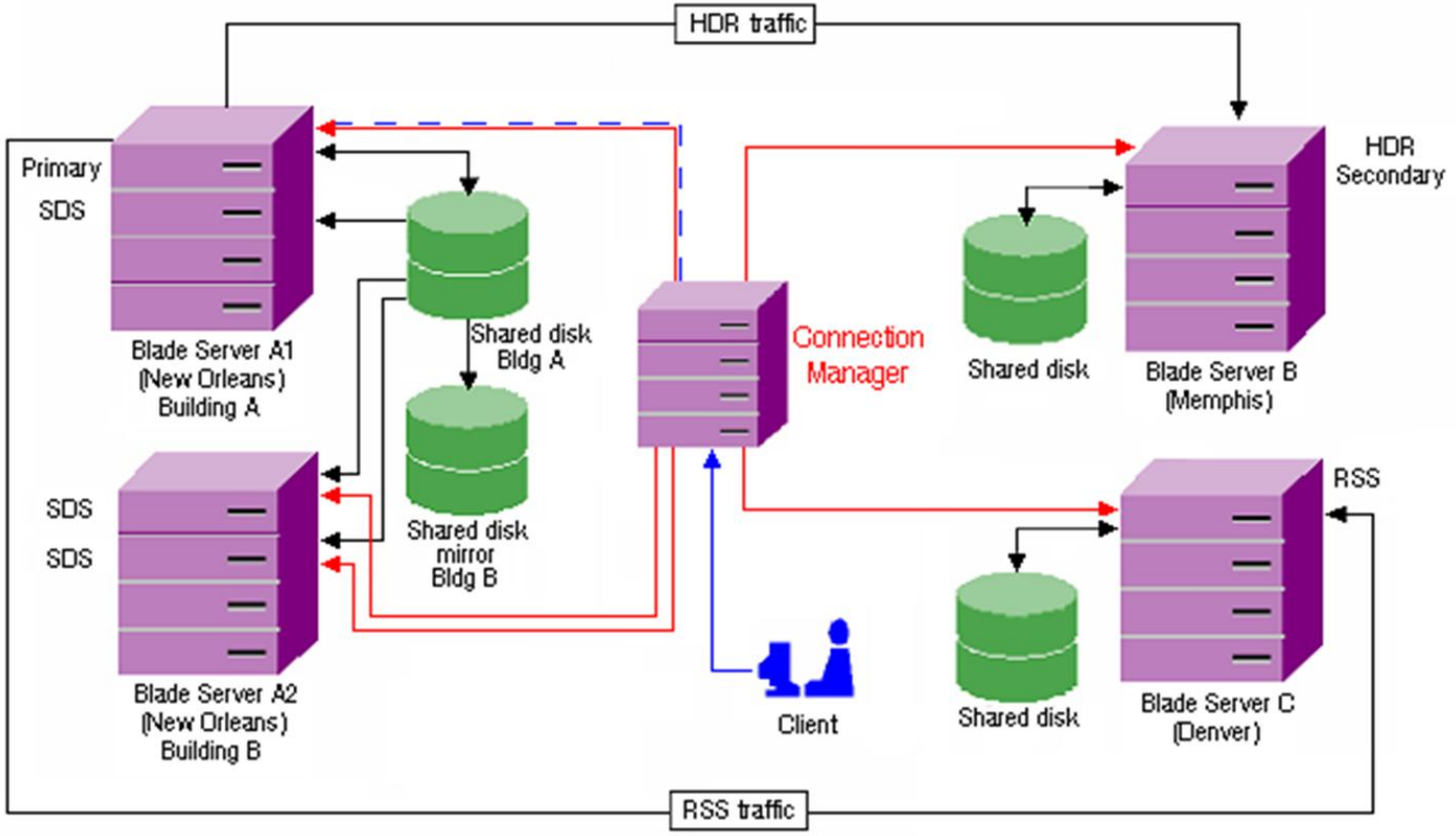
李俊旗

2016年4月

ToprowDB的由来

主要交流内容

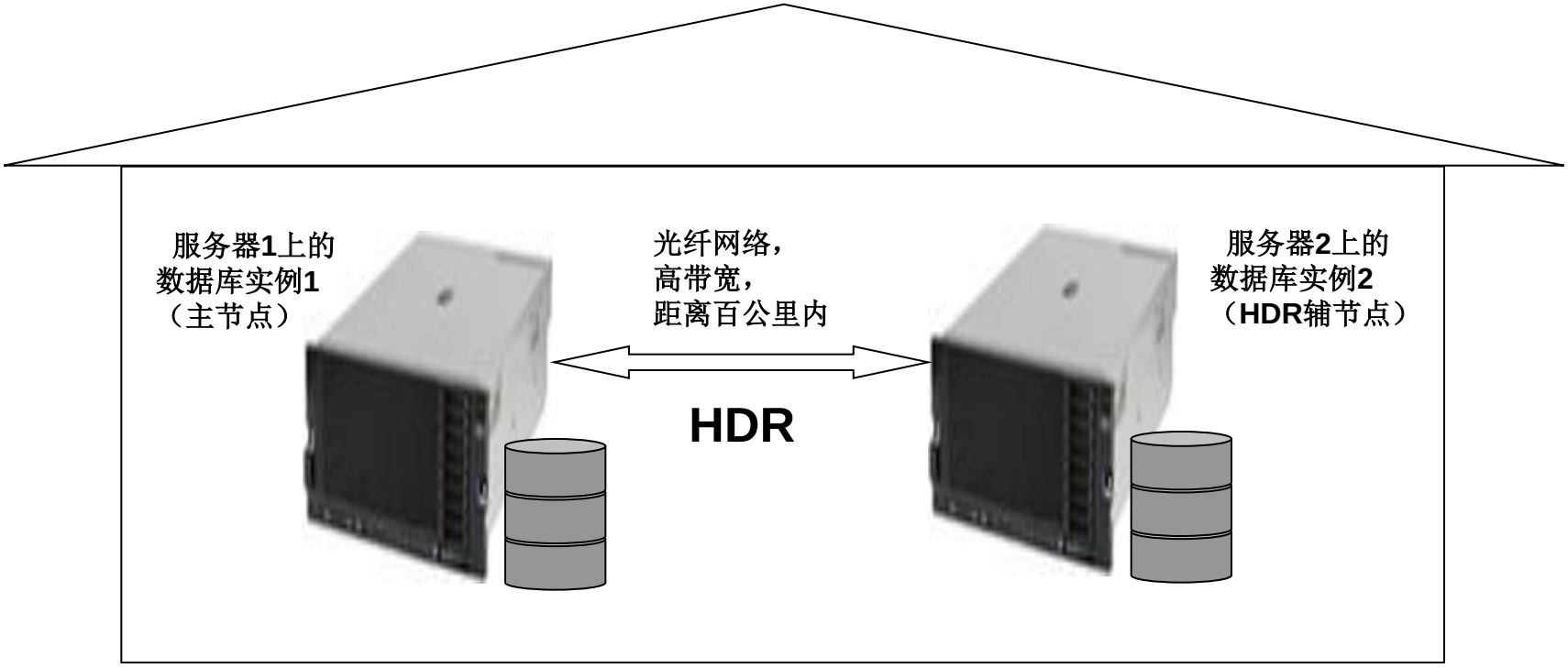
- HDR、RSS、SDS的使用场景、基本原理、差异比较、创建和故障恢复
- 正确的选择数据库高可用性解决方案、灾备方案和数据复制方案



ToprowDB高可用性、灾备和复制解决方案

- 1992年 Informix 6 提供了 HDR 技术
- 1994年 Informix 7 提供了 ER 技术
- 2006年 Informix 11 提供了 SDS、RSS、CLR 技术
- 2008年 Informix 11.5
  - HDR、SDS、RSS 备机都支持读写能力，提供了更强大的负载均衡
  - CM(Connection Manager)功能部件
    - ✓ SLA(Service Level Agreement) 功能，更好地实现负载均衡能力
    - ✓ FOC(Fail Over Connection) 功能，实现透明故障接管能力
- 所有这些对客户端应用来说都是是透明的
- HDR (High availability Data Replication, 高可用性数据复制)
- ER (Enterprise Replication 企业复制)
- RSS (Remote Standalone Secondary, 远程独立辅节点)
- SDS (Shared Disk Secondary,共享存储设备的集群)
- CLR (Continuous Log Restore，连续日志恢复)

HDR图示



服务器1和服务器2分别运行数据库实例

HDR高可用性数据复制

- **HDR 是Informix 数据库管理系统中历史最悠久的高可用性解决方案**，自IDS 6以来就包含在产品中，1992年
- **HDR 早在 Informix 7.3 时就被广泛使用**在电信业、银行业，已有二十几年的使用历史
- **HDR双机互备已经成为7*24系统的基本平台，**当任一主机发生故障时，可在8秒内切换至备机，企业的业务不受影响
- **大中华区的主要客户**
  - 电信业
  - 银行业
  - 保险业
  - 政府部门

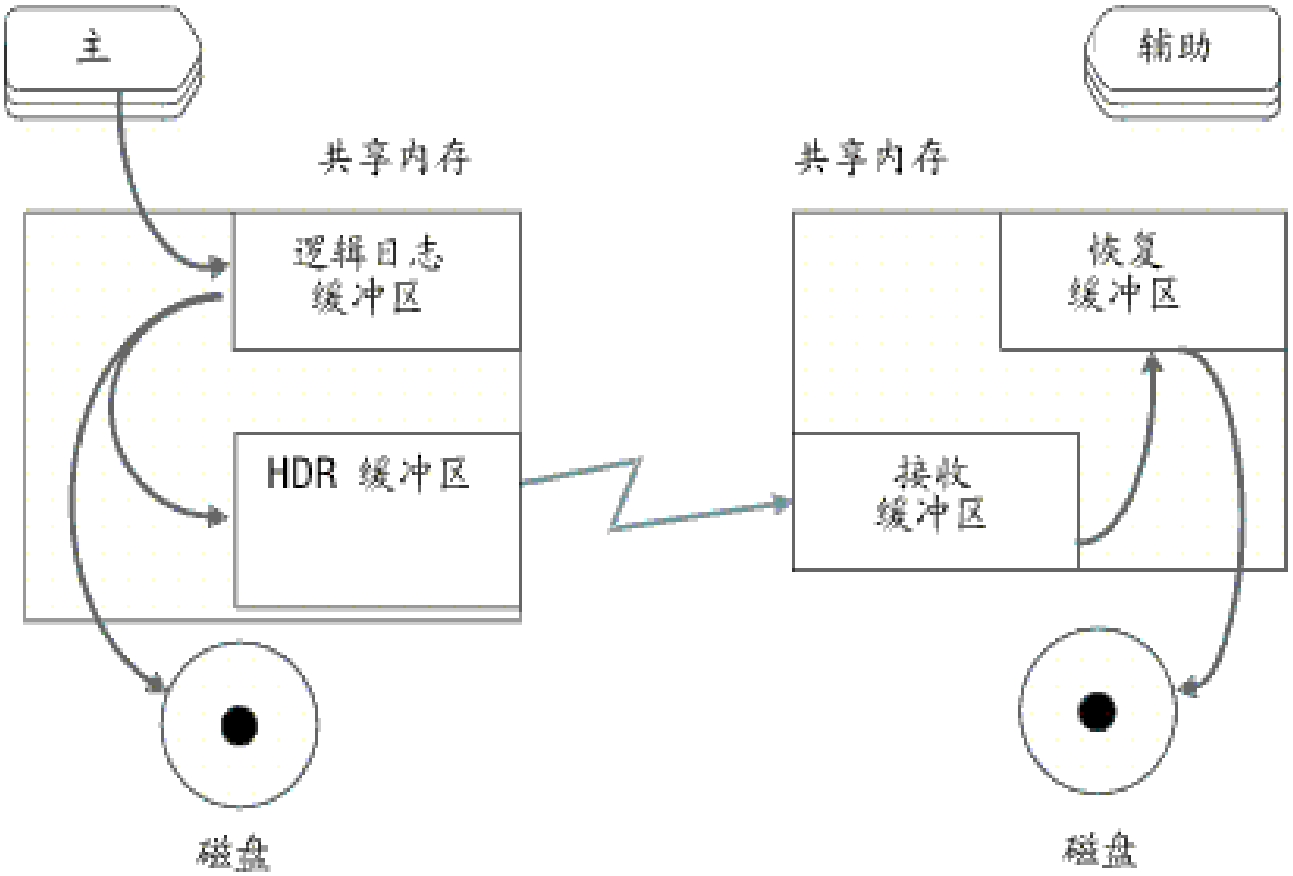
HDR特点

- 主节点和辅节点互为热备份
- 一个集群中最多只能有一个HDR辅节点
- 主节点和辅节点间需要传输逻辑日志，所以主节点和辅节点之间的网络的带宽需要足够大
- HDR辅节点可读写
- 容易配置，无需重大配置改变
- 可以承受硬件和软件故障
- 提供自动接管能力
  - 设置DRAUTO后，发生故障时可自动切换

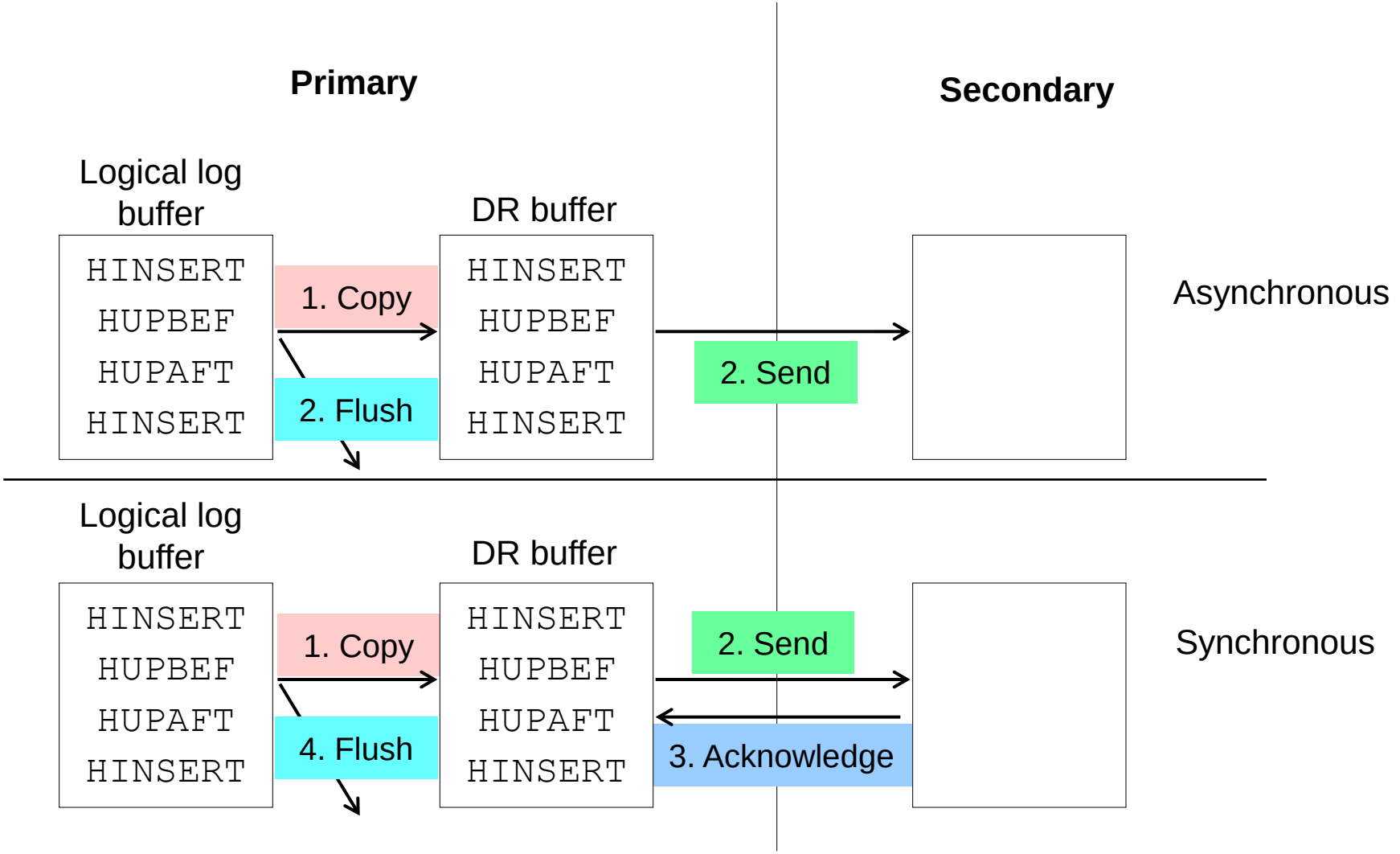
HDR配置要求

- 相同的数据库/硬件/操作系统
- 可设置为同步或异步通信方式
- 复制整个实例结构
- 仅带日志的数据库能被复制
- Primary (Read/Write)/ Secondary (Read/Write)
- Ontape或OnBar 工具恢复（支持管道方式）
- 故障切换: 自动 (DRAUTO, or CM) 或 手动
- 加密通信

HDR工作原理



同步和异步模式



轻松配置HDR

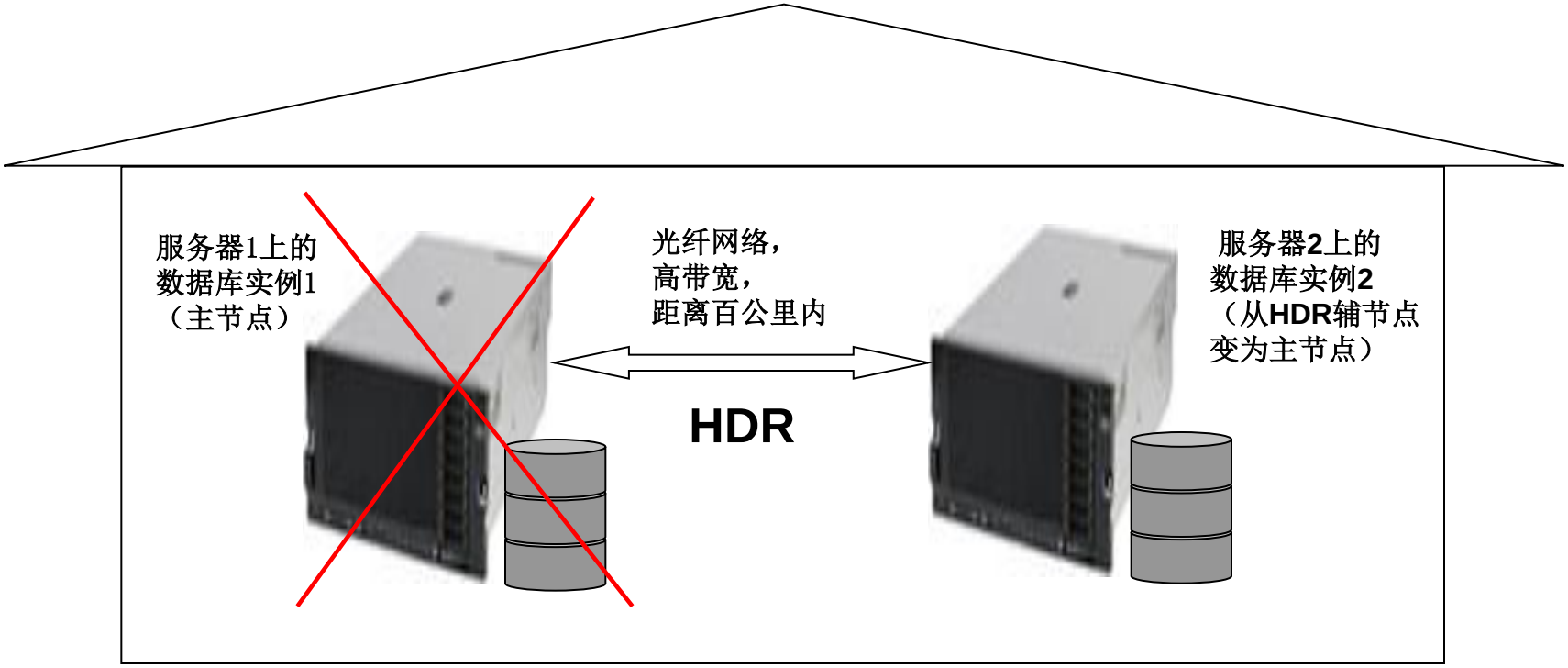
步骤	主节点	辅节点
1	onmode -d primary sec_name	onmode -ky
2	在主节点进行备份 ontape -s -L 0	
3		将备份文件拷贝到辅节点，在辅节点进行恢复 ontape -p
4		onmode -d secondary pri_name

HDR功能：故障切换与HDR自动重建

ONCONFIG参数DRAUTO

- 0：off
  - 主服务器发生故障时，不自动进行故障切换
- 1：RETAIN_TYPE
  - 主服务器发生故障时，自动进行故障切换
  - 主服务器恢复以后，主服务器仍为主服务器
- 2：REVERSE_TYPE
  - 主服务器发生故障时，自动进行故障切换
  - 主服务器恢复以后，主服务器退为辅服务器
- 3：Connection Manager
  - 由连接管理器决定如何进行故障切换

HDR故障场景2：主节点发生故障



服务器1发生故障后，数据库实例2被提升为主节点，
原先位于数据库实例1上的负载被自动转移到数据库实例2

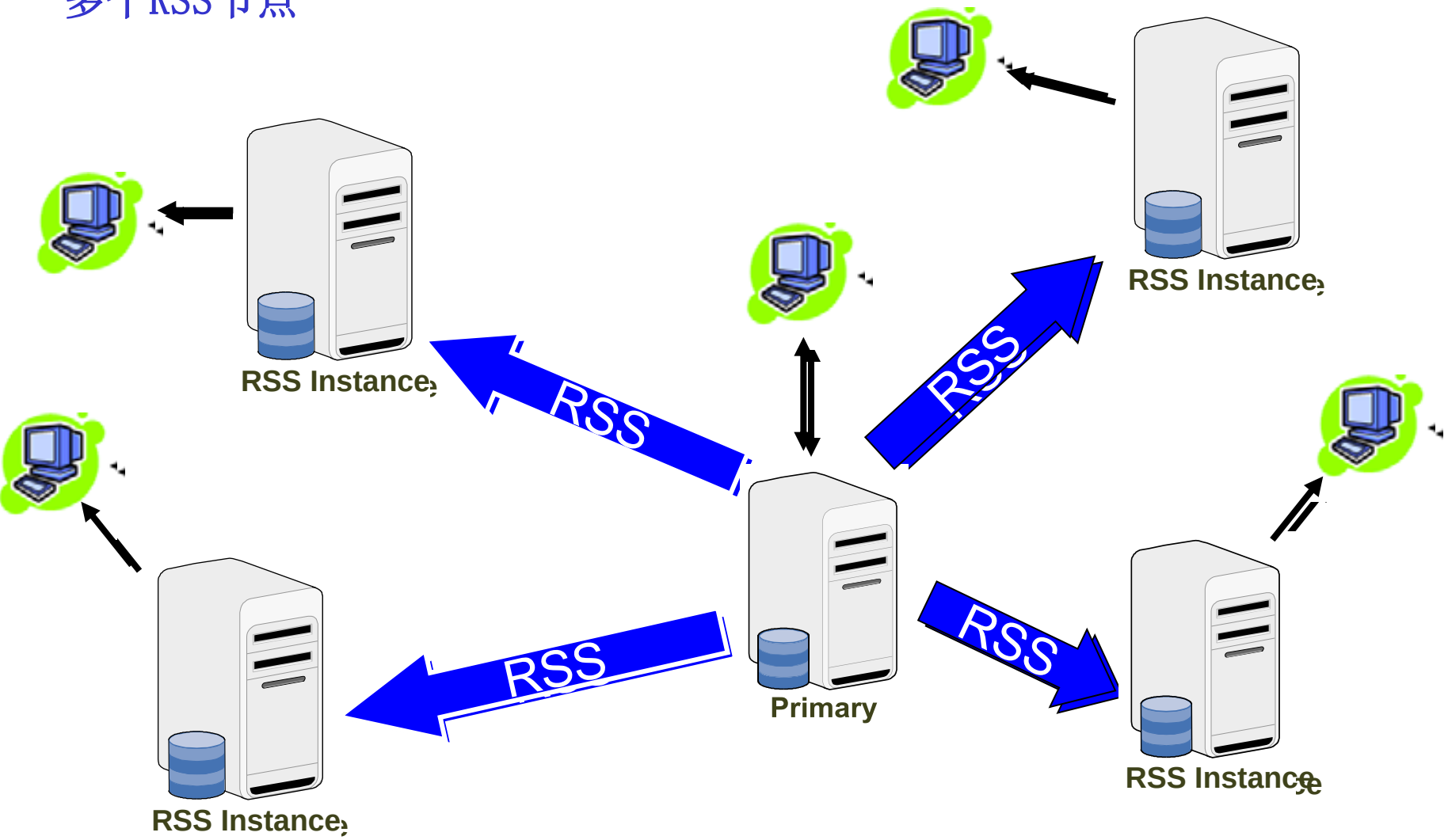
DRAUTO=2时的故障切换与HDR自动重建

DRAUTO=2：REVERSE_TYPE

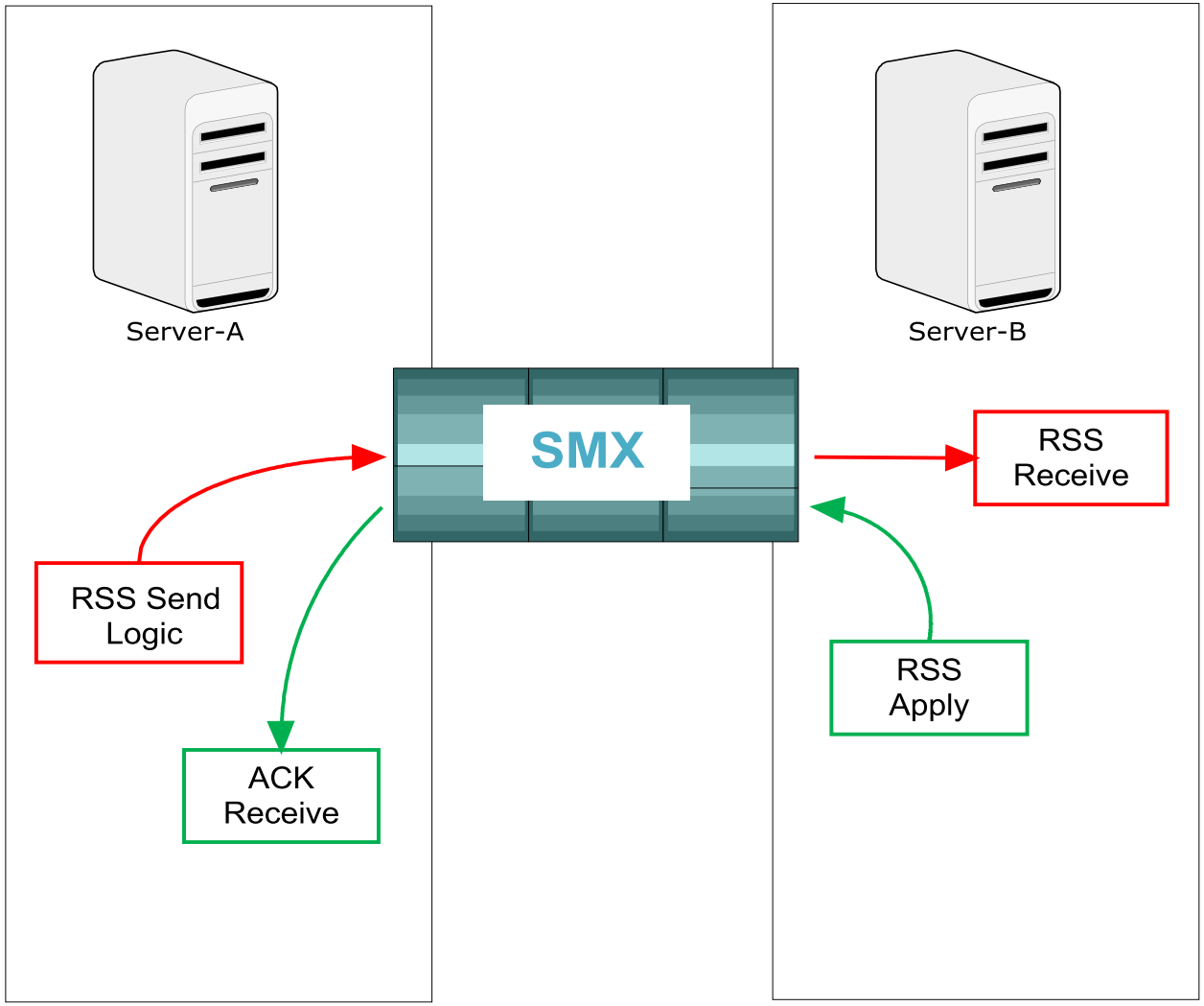
- 主服务器发生故障时，自动进行故障切换
- 主服务器恢复以后，主服务器退为辅服务器

Primary IDS_A	Second IDS_B
status: On-Line (Prim)	status: Updatable (Sec)
onmode -ky (关闭IDS_A)	
	status: On-Line (Prim) (IDS_B成为primary)
oninit -vy (启动IDS_A)	
status: Updatable (Sec) (IDS_A退为secondary)	status: On-Line (Prim)

多个RSS节点



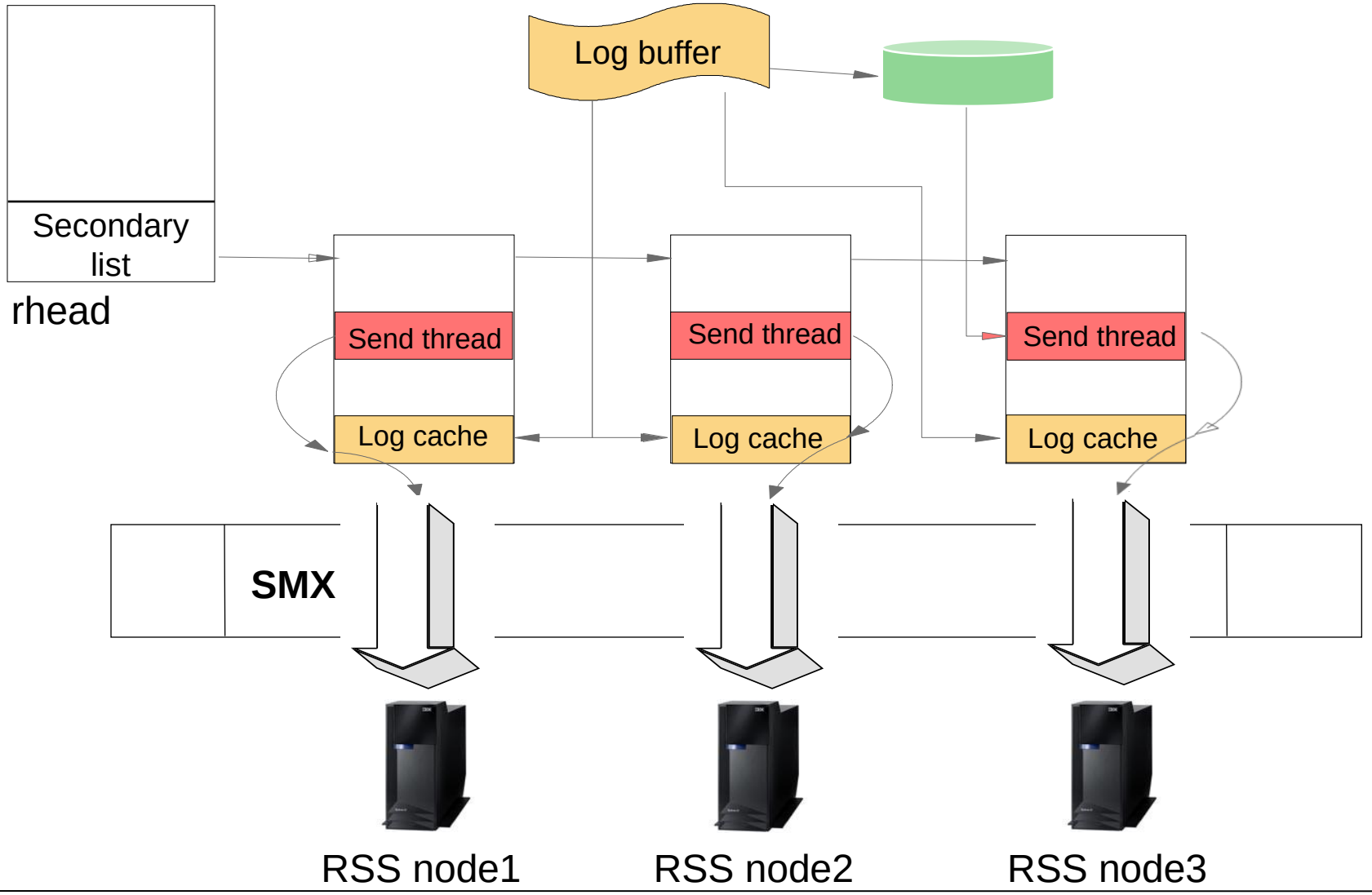
SMX (Server Multiplexer)



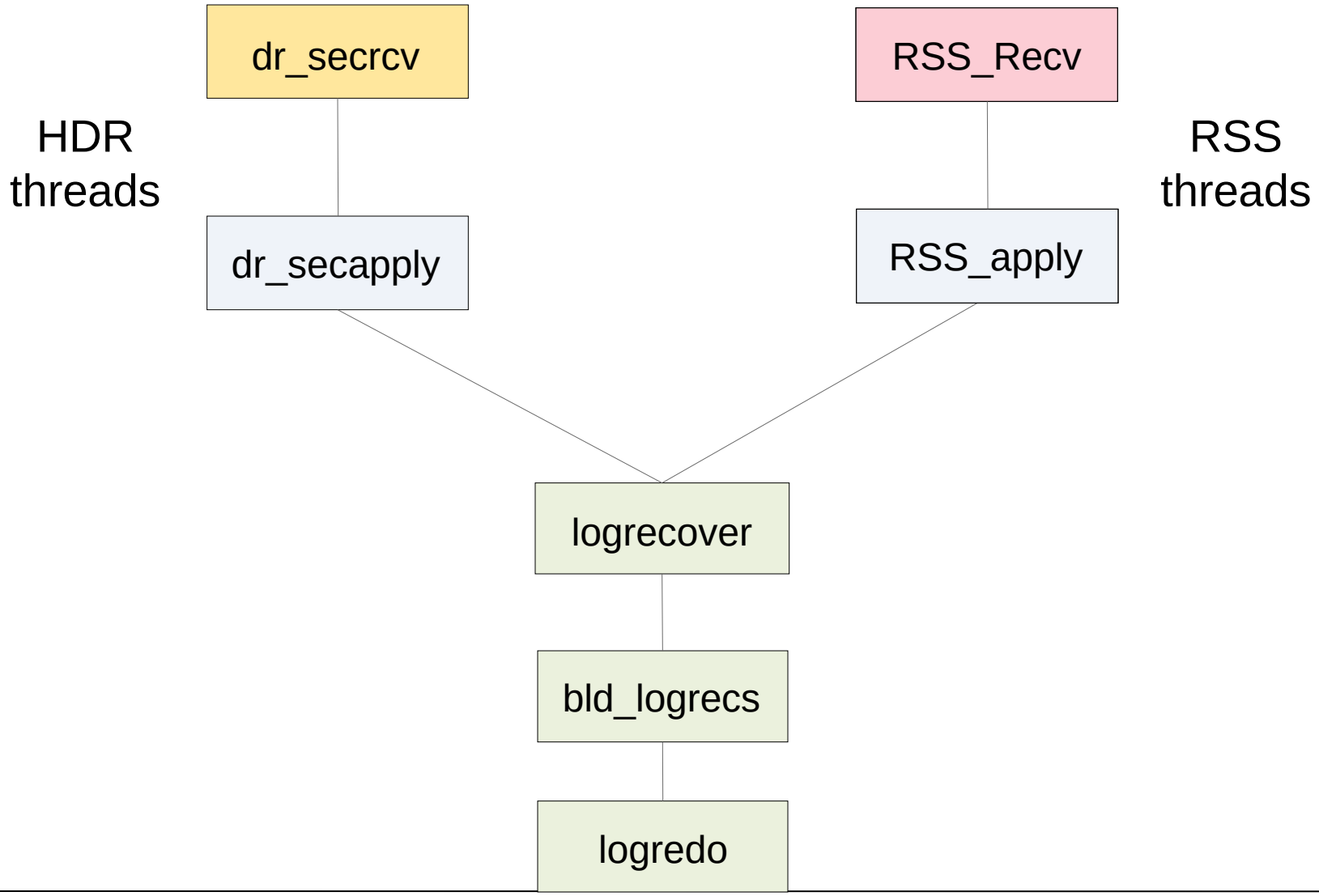
远程独立备份(RSS) 的特点

- Informix RSS在2006年就已推出(IDS 11)
- 类似于HDR
- 不同与 HDR
  - 可以建立多个 RSS 实例
  - 使用多道全双工通讯 (SMX) - 在低速网络上有更大吞吐
  - RSS对带宽的要求较低，主节点和辅节点之间可相距几百至几千公里
  - 只支持异步模式，不支持同步模式甚至对checkpoints
  - RSS辅节点可读写，承担部分业务，并和主节点相互备份；在某个节点发生故障时，该节点上的业务被转移到其它节点
  - 不能被升级为’主’(primary) - 但可以被升级为HDR ‘备’
    - ✓ 专注在灾备而非 HA，可以承受火灾、地震、海啸等
- **大中华区的主要客户：**电信业、银行业、保险

RSS Primary怎样工作



RSS secondary 怎样工作



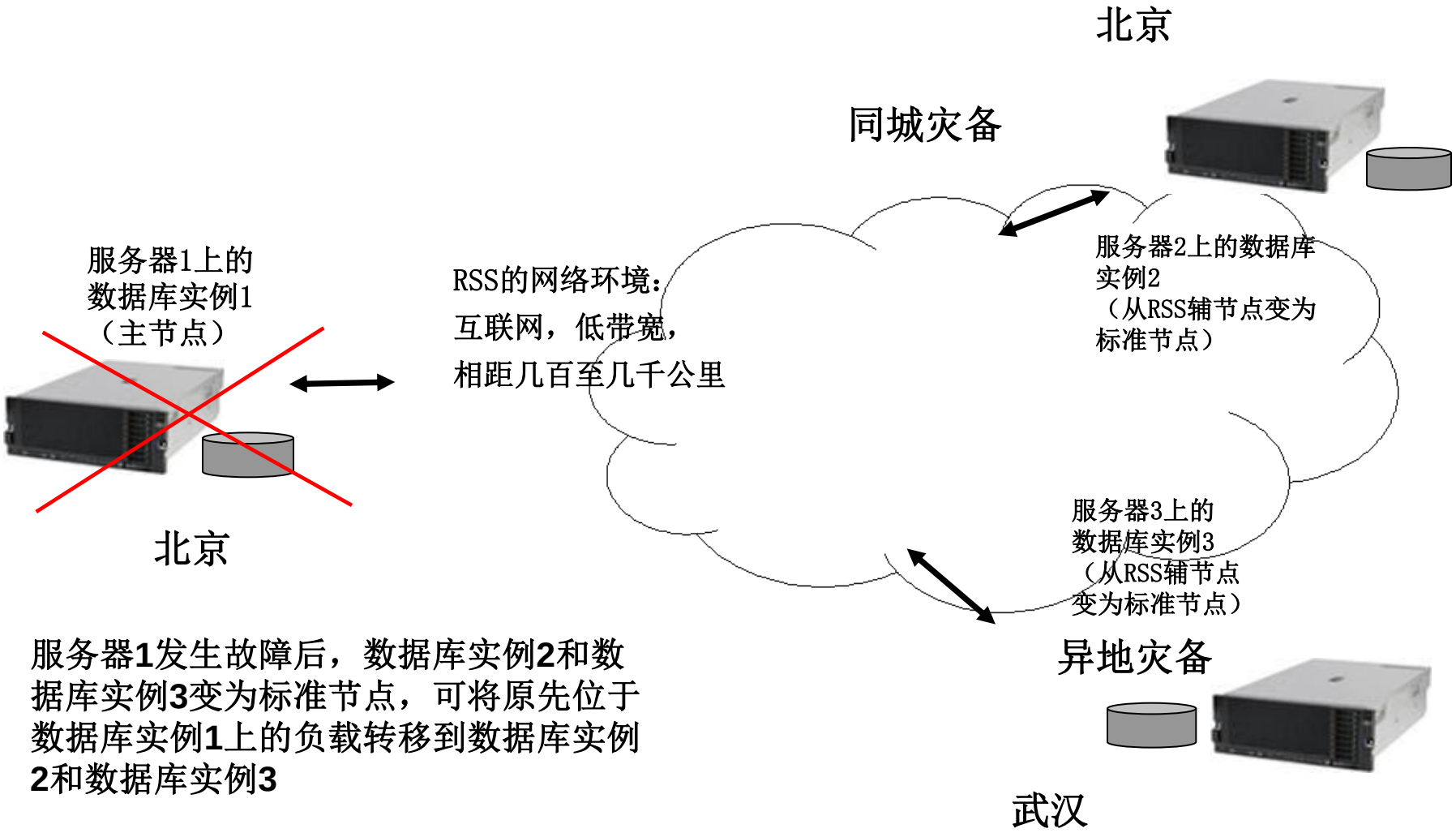
RSS：接管

- 规则
  - DRAUTO 对 RSS不起作用
  - RSS 不能直接与’主’相互接管
  - RSS 节点可转化成HDR ‘备’
  - HDR ‘备’可转化成 RSS
  - RSS节点可转化成标准的独立实例

轻松配置RSS

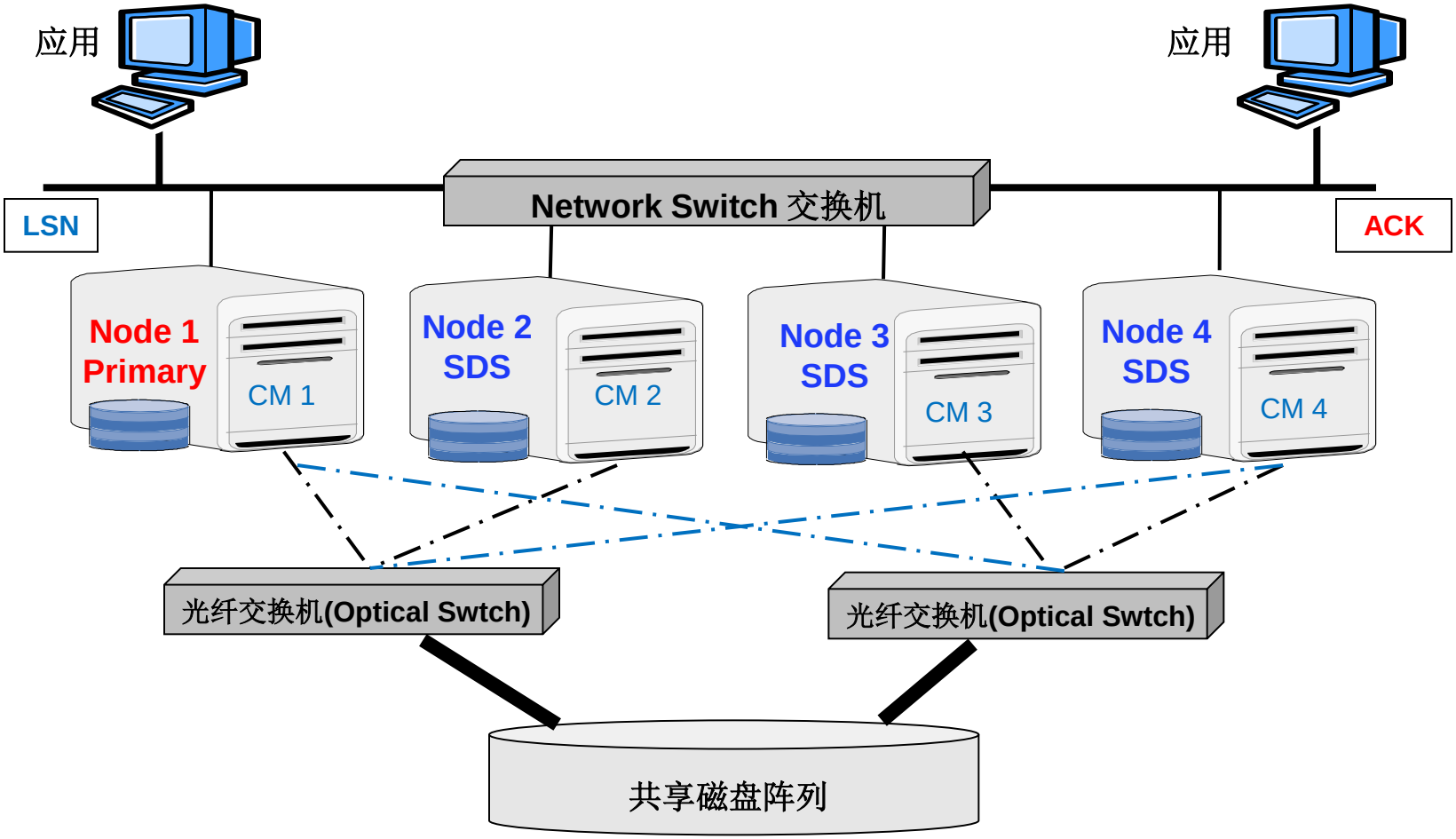
步骤	主节点	辅节点
1	onmode -d add RSS rss_name	onmode -ky
2	在主节点进行备份 ontape -s -L 0	
3		将备份文件拷贝到辅节点，在辅节点进行恢复 ontape -p
4		onmode -d RSS pri_name

RSS故障场景：主节点发生故障



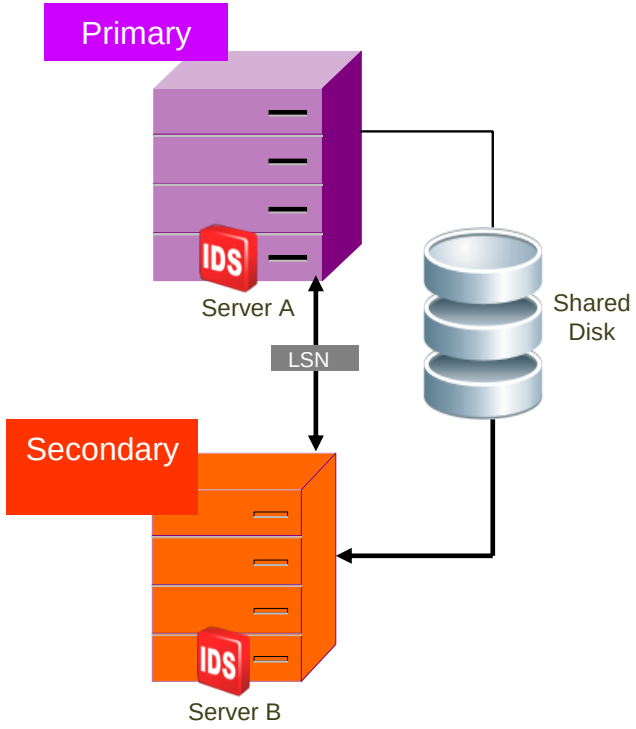
服务器1发生故障后，数据库实例2和数据库实例3变为标准节点，可将原先位于数据库实例1上的负载转移到数据库实例2和数据库实例3

SDS单一主节点(Primary)架构 - 多节点共享磁盘



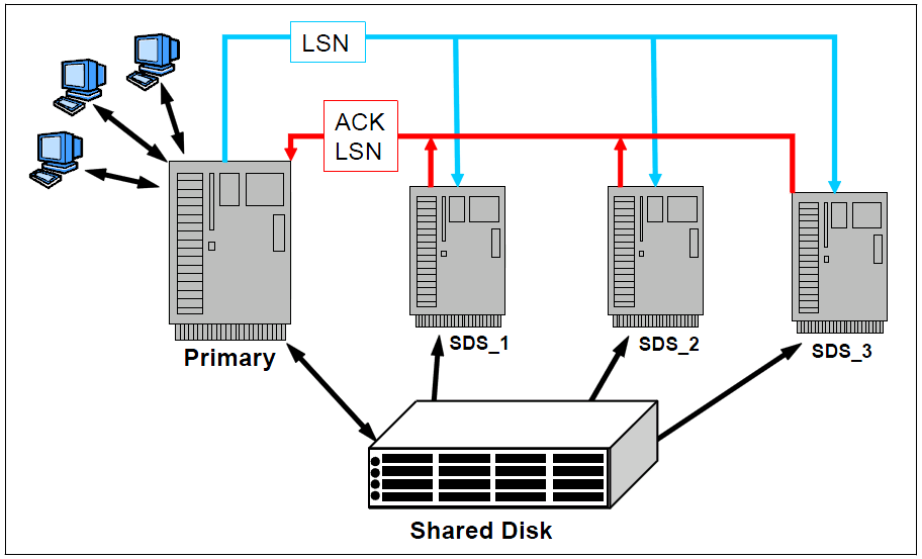
SDS共享磁盘辅节点的集群

- Informix SDS在2006年就已推出(IDS 11)
- 一个集群中可以有多个SDS辅节点
- 所有的节点共享一个存储设备
- SDS辅节点可读
- 高可用性，可以承受软硬件故障、当某一节点出现故障时，其他节点将自动、快速接管；在实际系统峰值负载情况下，经过各种可能的异常故障场景测试，SDS数据库均可以在预期时间内(<1分钟)完成切换，应用程序可在<2分钟内完成切换
- 高可扩展性，当集群中有4个SDS辅节点时，增加新的SDS辅节点后，集群的性能仍会得到较大的提升；企业可根据数据增长的需要，往集群中添加新的SDS辅节点
- SDS无需特殊硬件支持，安装配置简单、快速
- 应用透明性，无需调整您的应用程序
- 在中国，已有银行等核心系统使用SDS



SDS工作原理

- 主节点和 SDS 辅节点共享磁盘
- 主节点仅需向辅节点发送日志号(LSN)，主节点无需向辅节点发送日志



1. **Primary：** 写逻辑日志
2. **Primary：** 将Log Sequence Number (LSN) 发送到 secondary servers
3. **SDS_1,2,3:** 根据接收到的LSN从磁盘读取相应的log并在buffer中重做,但不写回到磁盘
4. **SDS_1,2,3:** 重做后发送ACK回primary server
5. **Primary：** 确认ACK

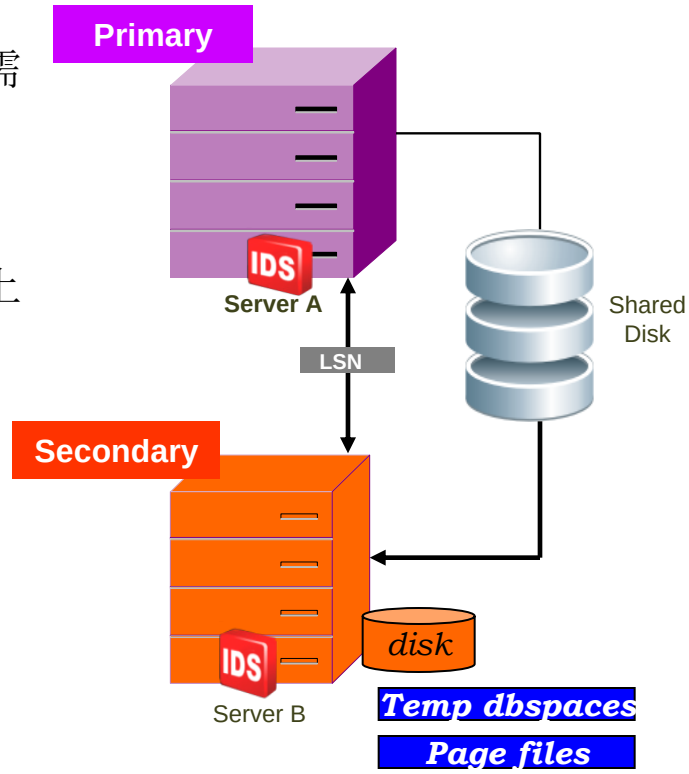
SDS工作原理(续)

Temporary Dbspaces

- Secondary 上一些特定的SQL执行时需要TempDBS ，如物化视图、hash join、排序等操作
- Secondary 不能共享使用Primary定义的 TempDBS 需要定义本地的 TempDBS

Page Files

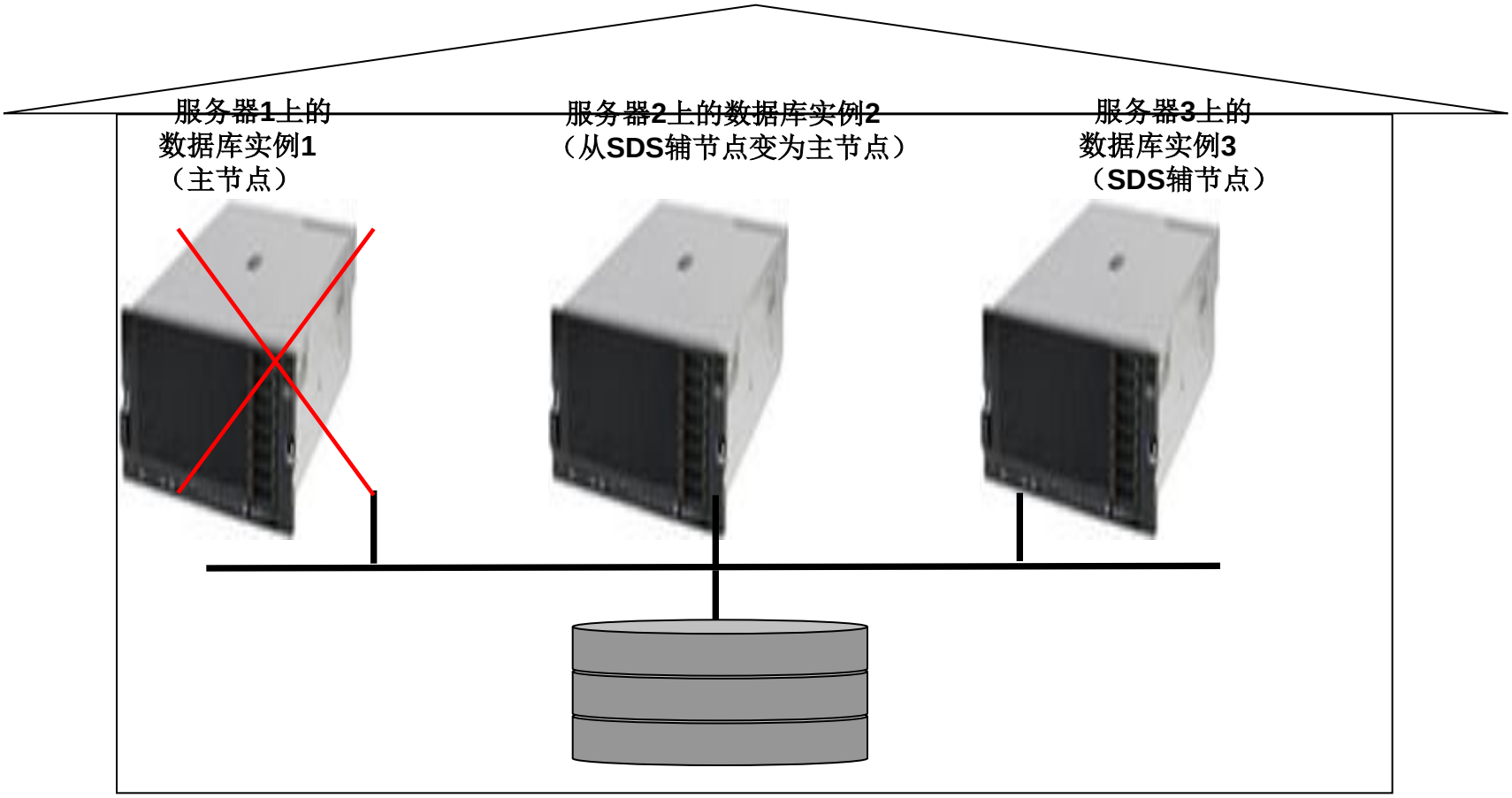
- Secondary 把从逻辑日志收到的数据变化在Buffer上重做
- 将Buffer的内容写入本地文件（Page Files）
- 当发生Checkpoint时，Page Files内容将被清空
- 需要定义2个Page Files



轻松配置SDS

步骤	主节点	SDS辅节点
1	设置onconfig参数SDS_TIMEOUT	SDS_ENABLE设为1 SDS_PAGING SDS_TEMPDBS
2	oninit -ivy	
3	设置 SDS主节点： onmode -d set SDS primary pri_name	
4		oninit -vy

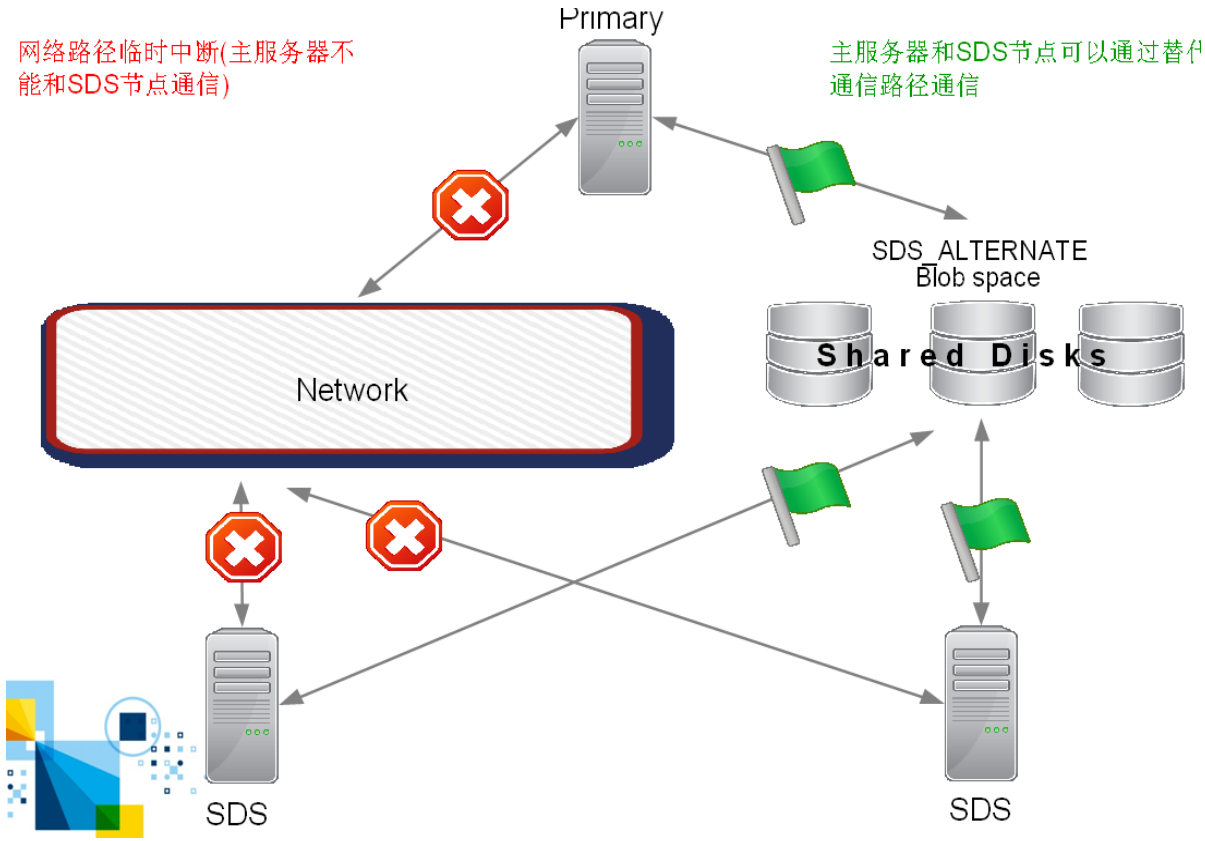
SDS故障场景：主节点发生故障



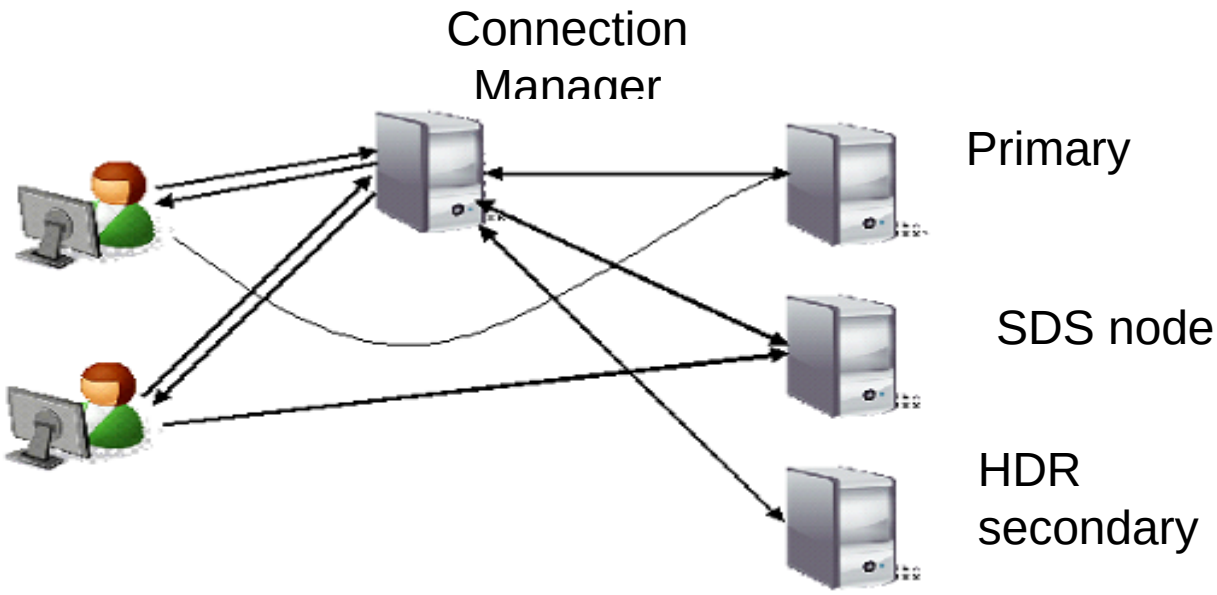
服务器1发生故障后，数据库实例2被提升为主节点，原先位于数据库实例1上的负载被自动转移到数据库实例2和数据库实例3

SDS网络故障的处理

- 若主服务器和**SDS**节点的网络连接不可用，则有另一种通过共享磁盘交换数据的方法：
  - **SDS_ALTERNATE** 设置为一个专用的BLOB space的名字



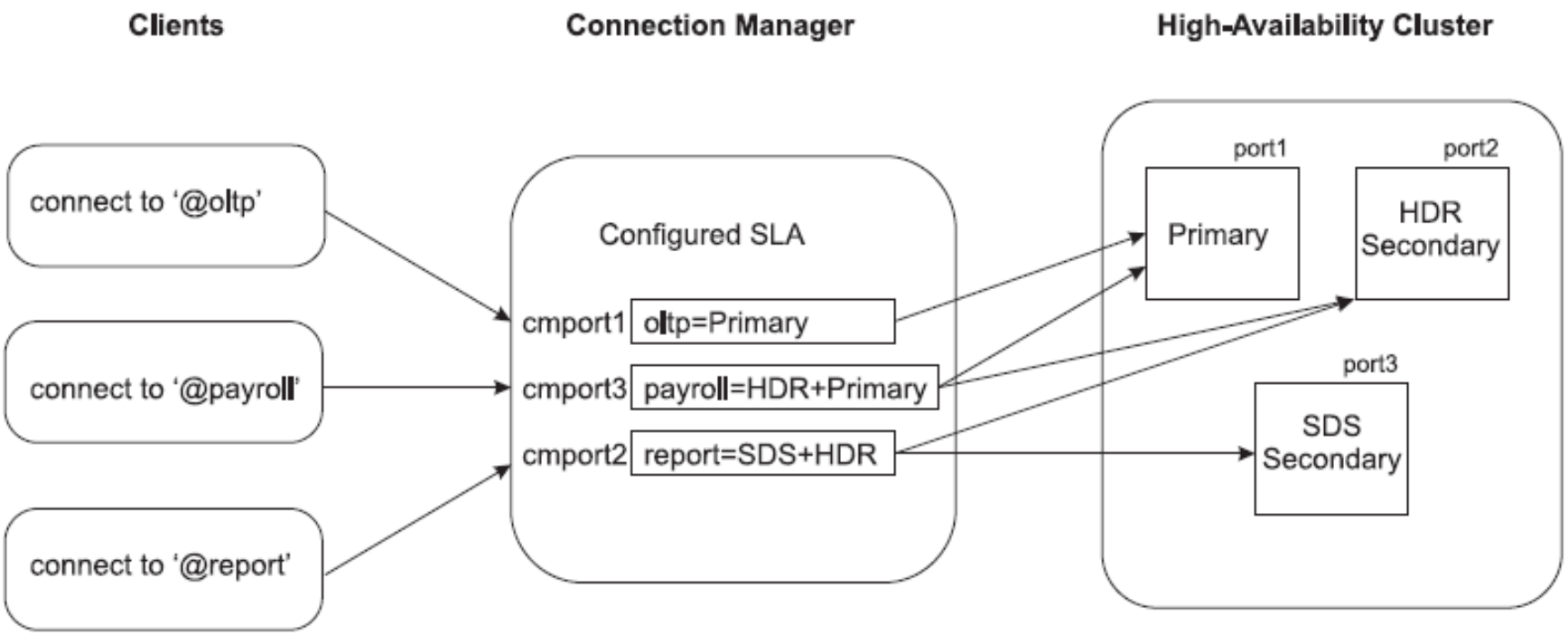
CM连接管理器



- CM（Connection Manager，连接管理器）
  - 将客户端请求分发给各个数据库服务器节点
  - 基于服务层协议--*Service Level Agreements*（SLA）
  - 管理负载均衡
  - CM监测集群里各数据库服务器节点的心跳
  - 提供故障转移的功能，对发生故障时的节点切换顺序进行管理（FOC）

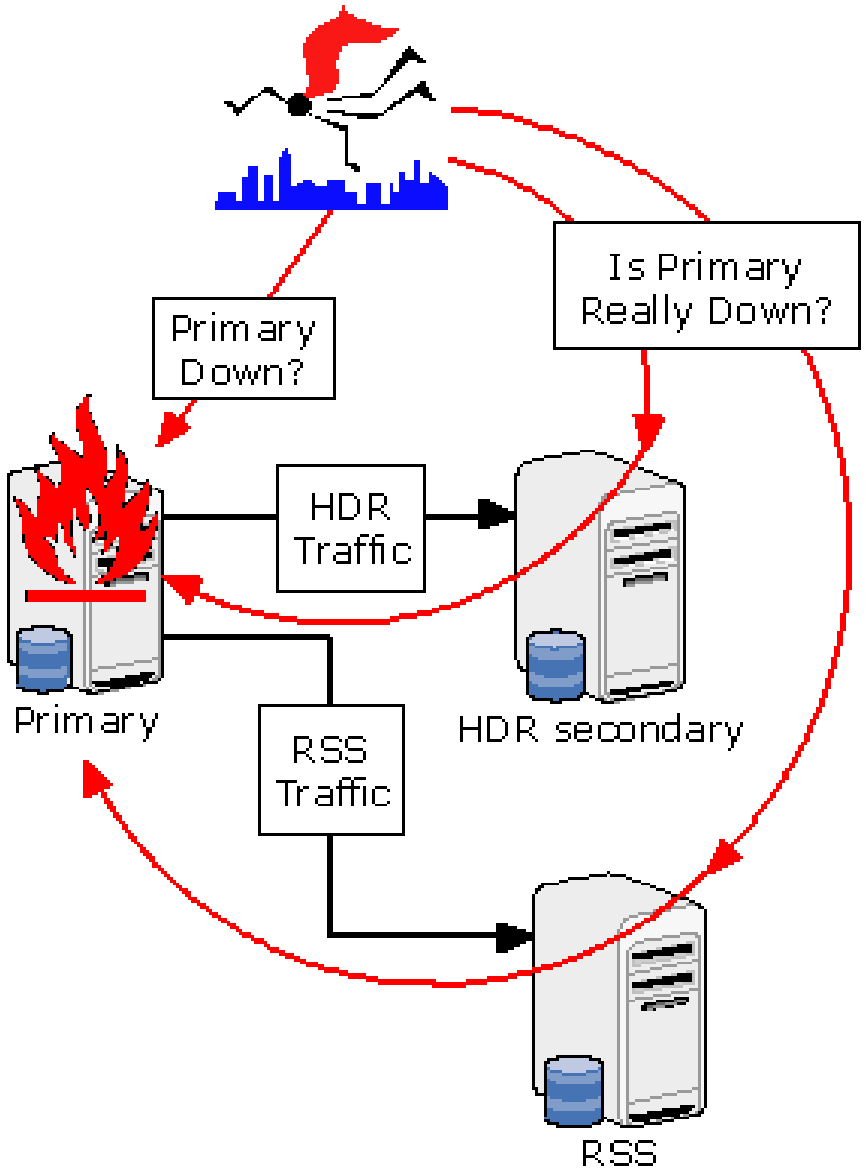
SLA 路由实例

- 使用如下 **SLAs进行连接:**
  - SLA oltp=primary
  - SLA payroll=HDR+primary
  - SLA report=SDS+HDR



CM 仲裁功能

- 实现在高可用性集群环境中节点故障自动切换逻辑
- 也被称作故障切换仲裁者
- 为了主节点的故障切换监控所有节点
- 当确认主节点故障时，运行故障切换（比如，转换备节点为主节点）
- 支持故障切换到一个RSS节点



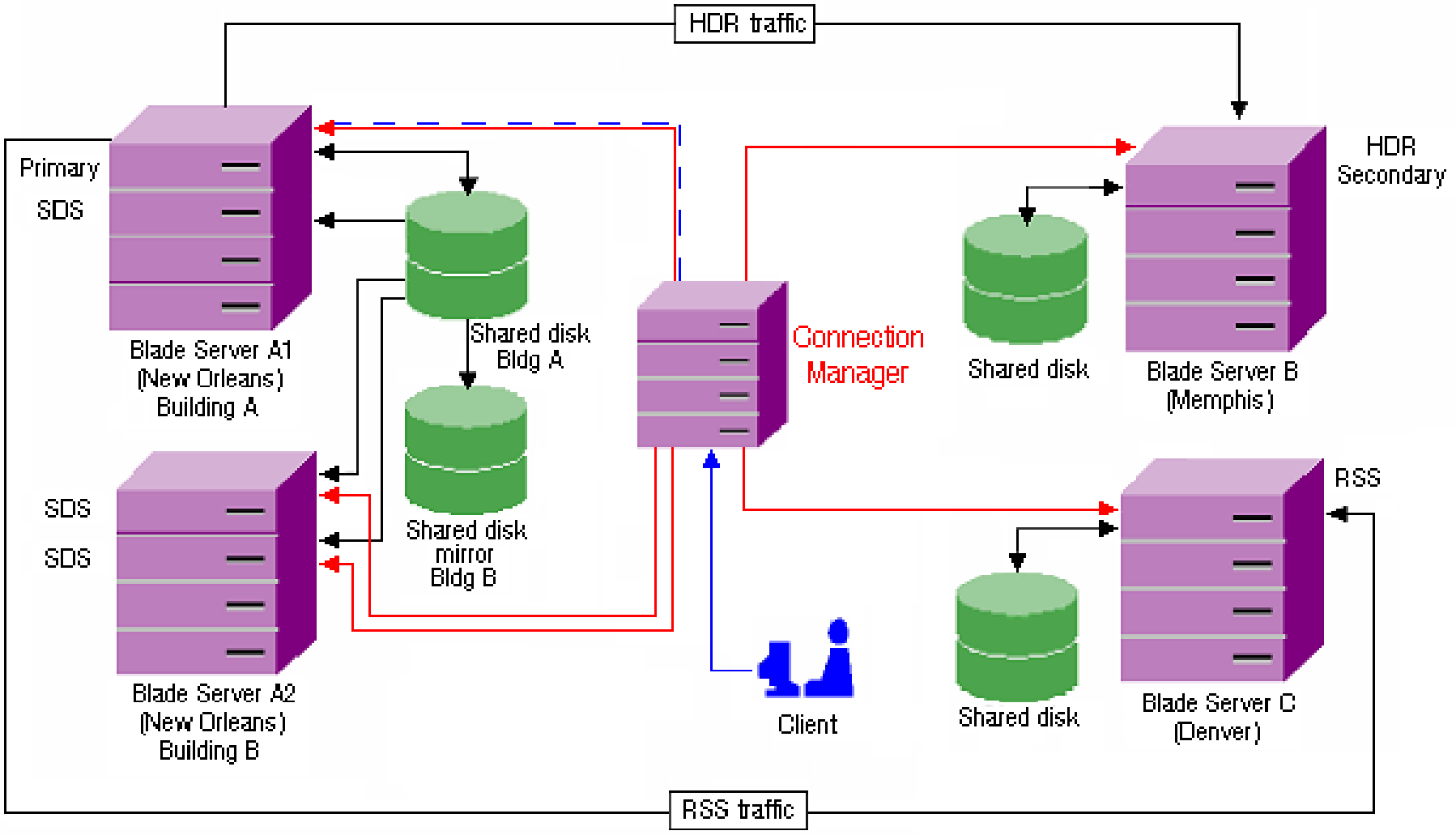
用CM 做数据库集群中的故障转换

- CM 配置文件中FOC 配置是一个或多个 Primary, SDS, HDR, RSS, 或一组由加号分开并在括号中的服务器类型
- FOC 定义故障切换的顺序

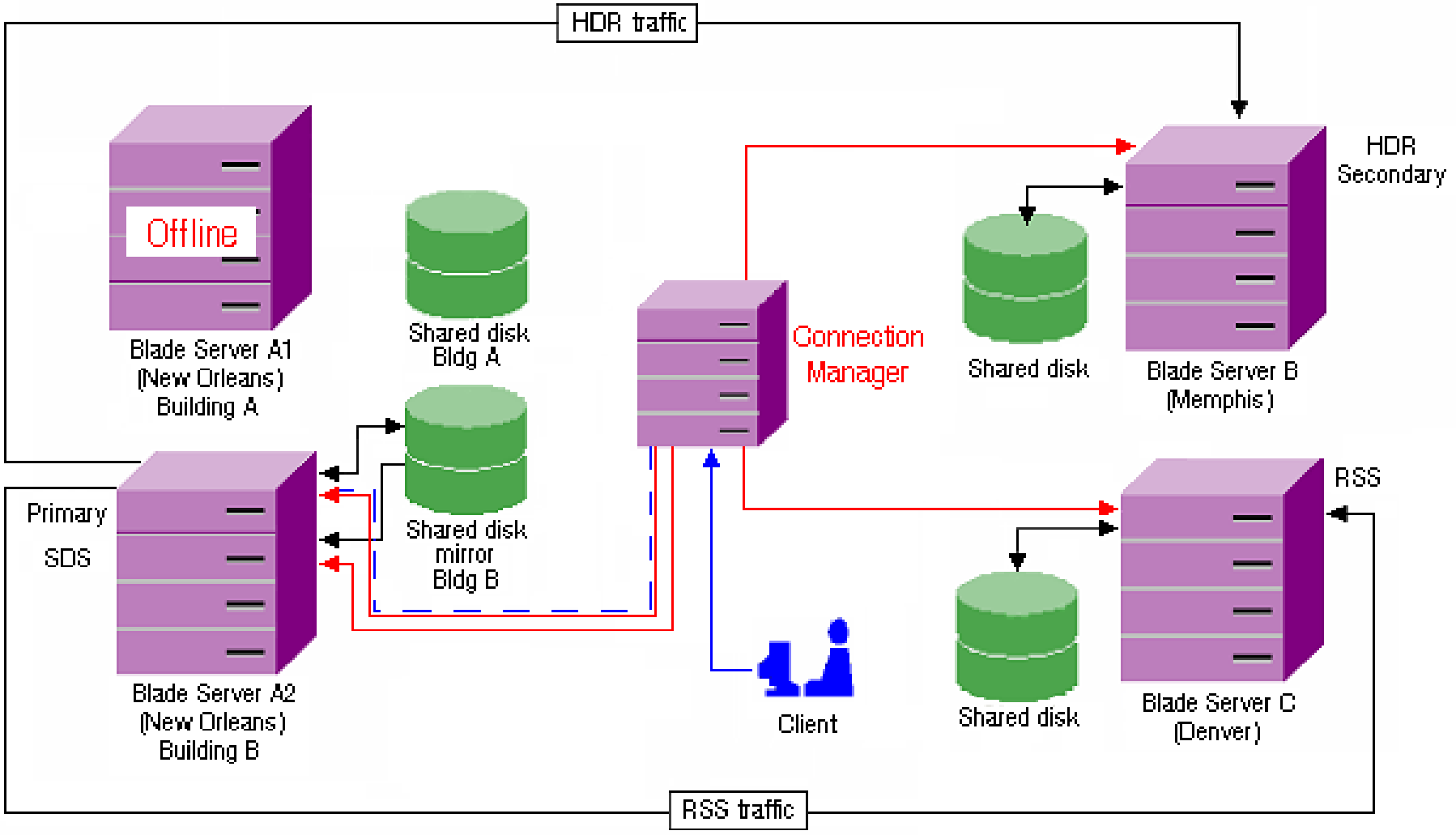
```
FOC RSS_node1+RSS_node2,60

    FOC             node1+(SDS+node2+HDR+node3)+node4+RSS,0
```

- 如果 Primary 失败, CM 首先试图转换 RSS_node1 为 Primary
- 如果不能转换，将转换 RSS_node2 为 Primary
- 在括号中的节点或服务器，命名的节点或服务器有较高的优先，其次服务器类型类，依次为SDS, HDR和RSS.

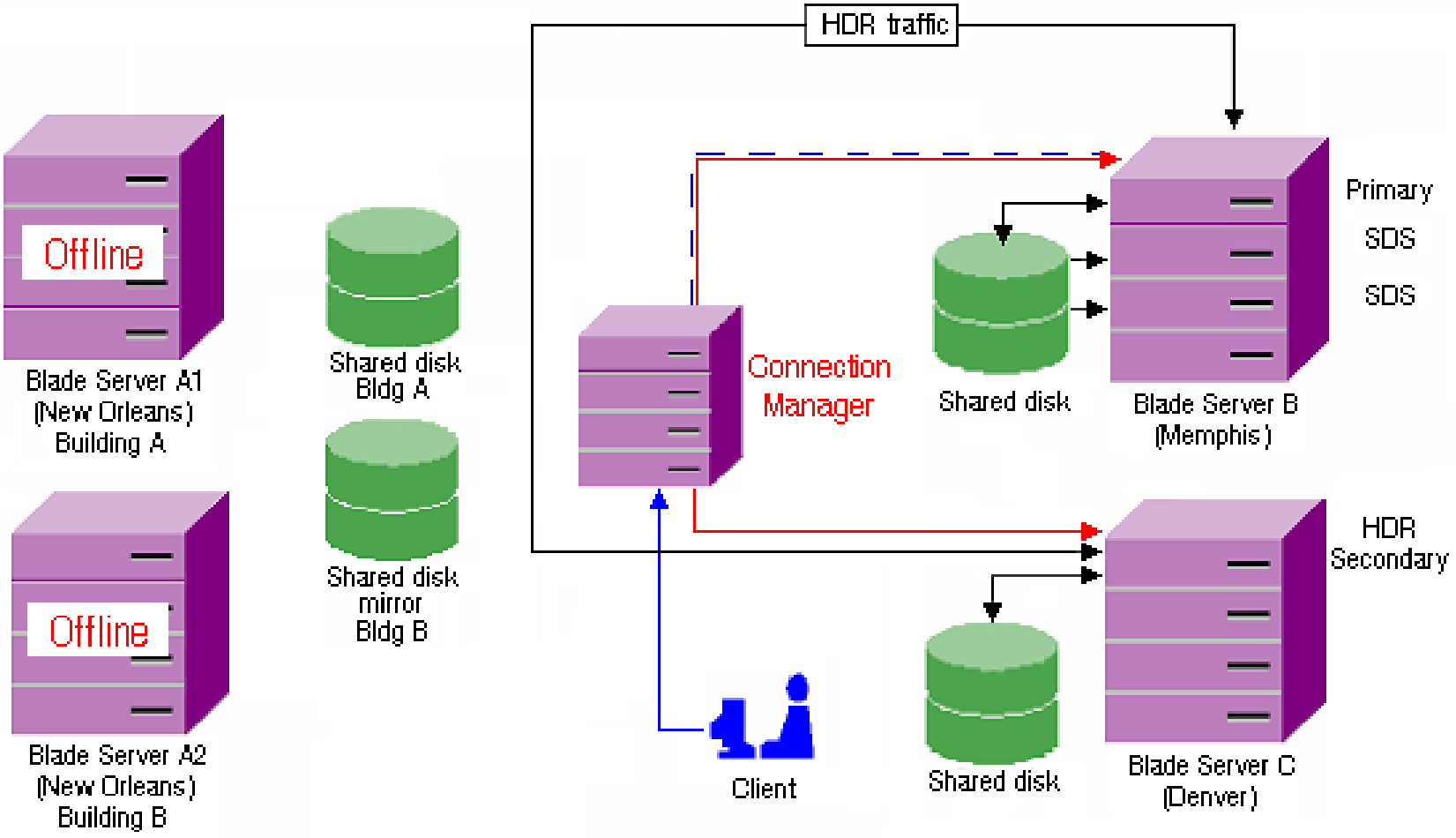
一个复杂集群方案



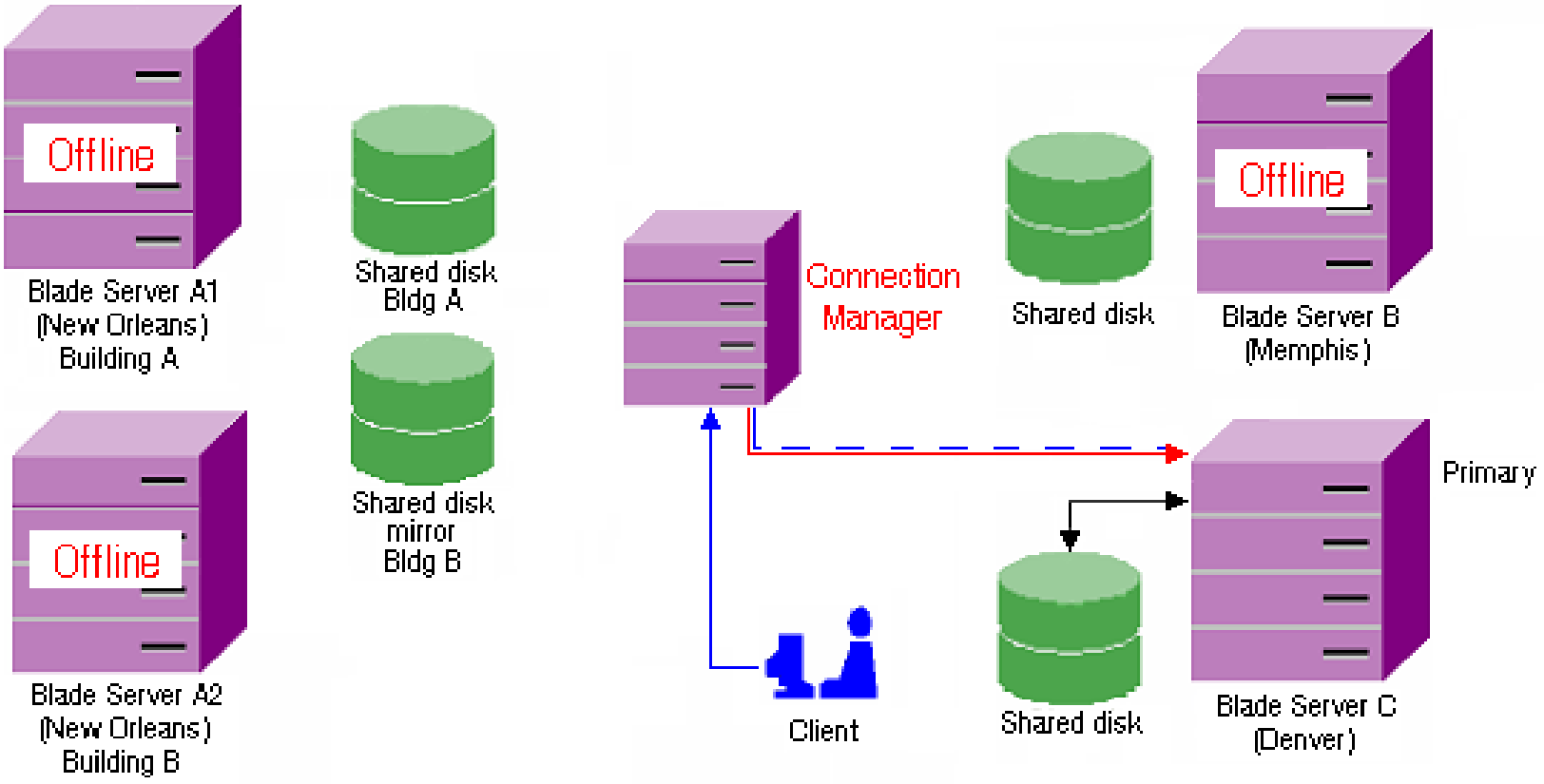
当建筑A发生水管破裂



当城市New Orleans遭受飓风



当城市Memphis发生地震



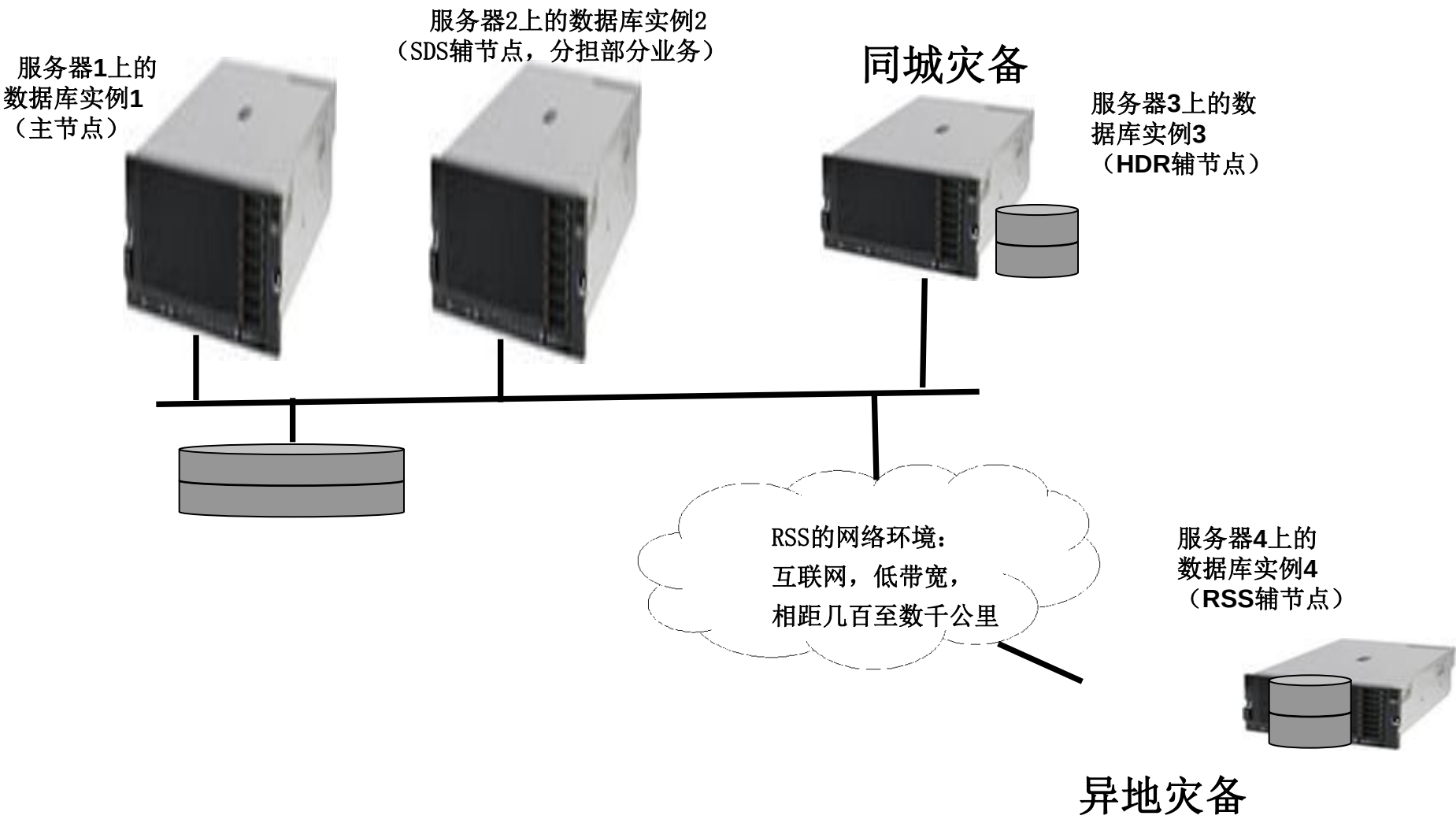
总 结

- 每一种类型的辅节点都有它自己的特点，我们需要根据实际情况选择合适类型的辅节点
  - 如果要建立本地双机互备，那么选择HDR辅节点
  - 如果要建立远程热备份节点，那么选择RSS或HDR辅节点
  - 如果要建立可扩展性强的集群，那么选择SDS辅节点
- 在一个集群里包含多种类型的辅节点，从而聚合各种类型辅节点的优点，更好的为业务服务
  - HDR + RSS ：既有本地备份，又有远程备份
  - SDS + RSS ：既有很强的可扩展性，又有很强的容灾能力
- ToprowDB提供了极好的高可用性，可以确保您的业务不间断的运行

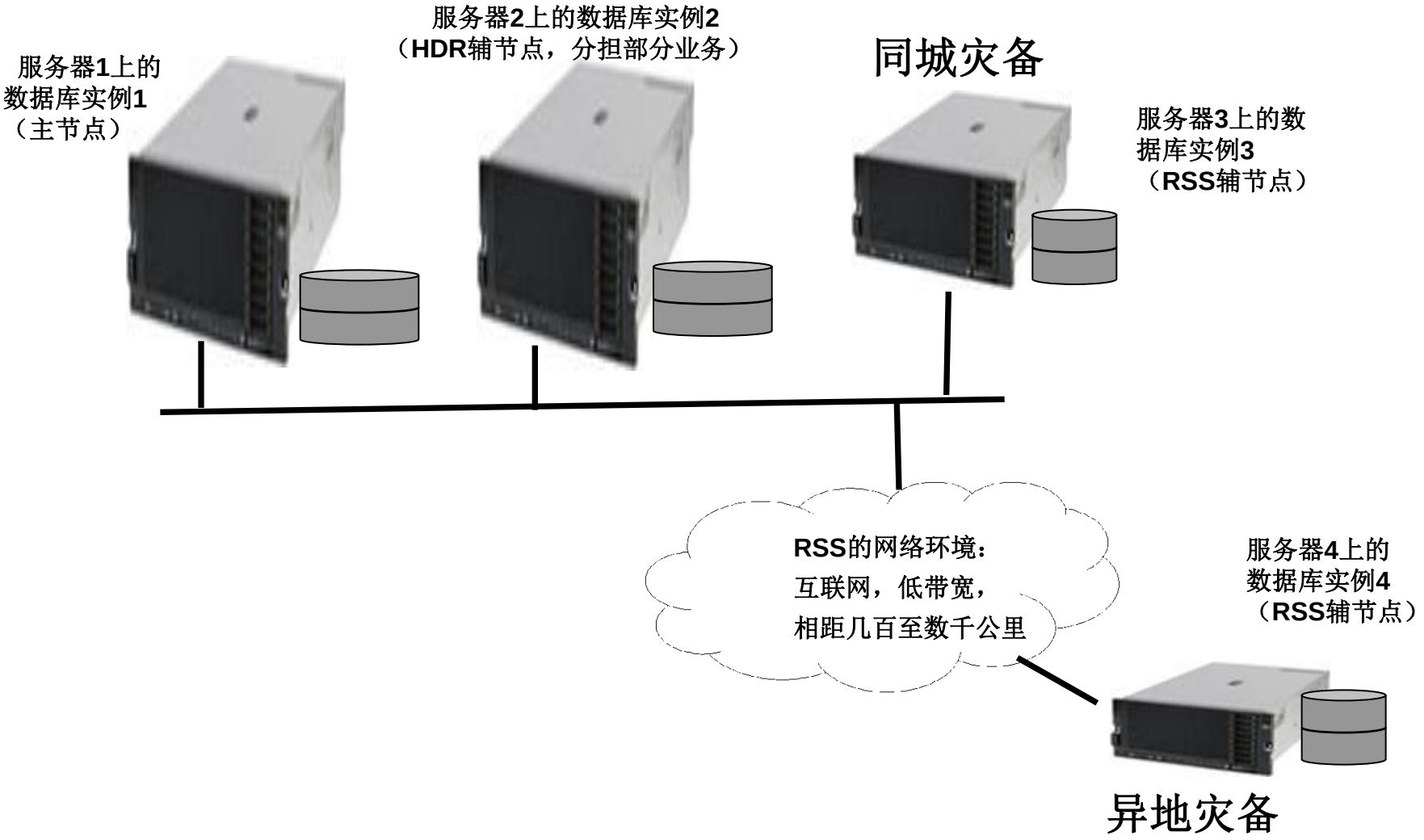
不同类型辅节点的对比

	HDR	RSS	SDS
网络连接	光纤连接	互联网或专线	光纤连接
节点间的距离	百公里内	数千公里	数米内，同一个机房
对带宽的要求	高	低	高
辅节点的最大数目	1个	多个	多个
存储设备	独立存储设备	独立存储设备	共享存储设备
承受服务器的软硬件故障	是	是	是
承受天灾	否	是	否

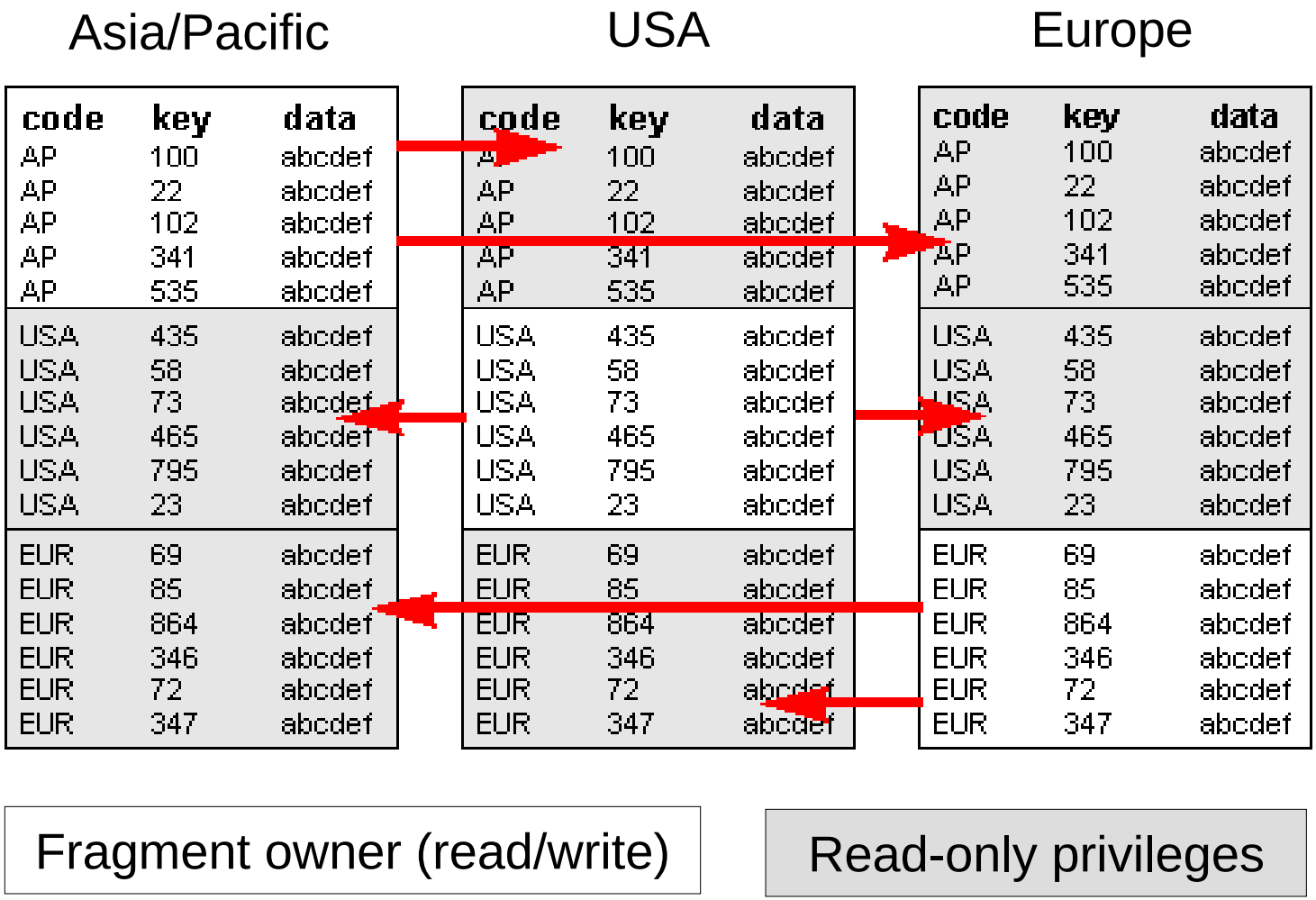
2地3中心部署方案一（Primary+SDS+HDR+RSS）



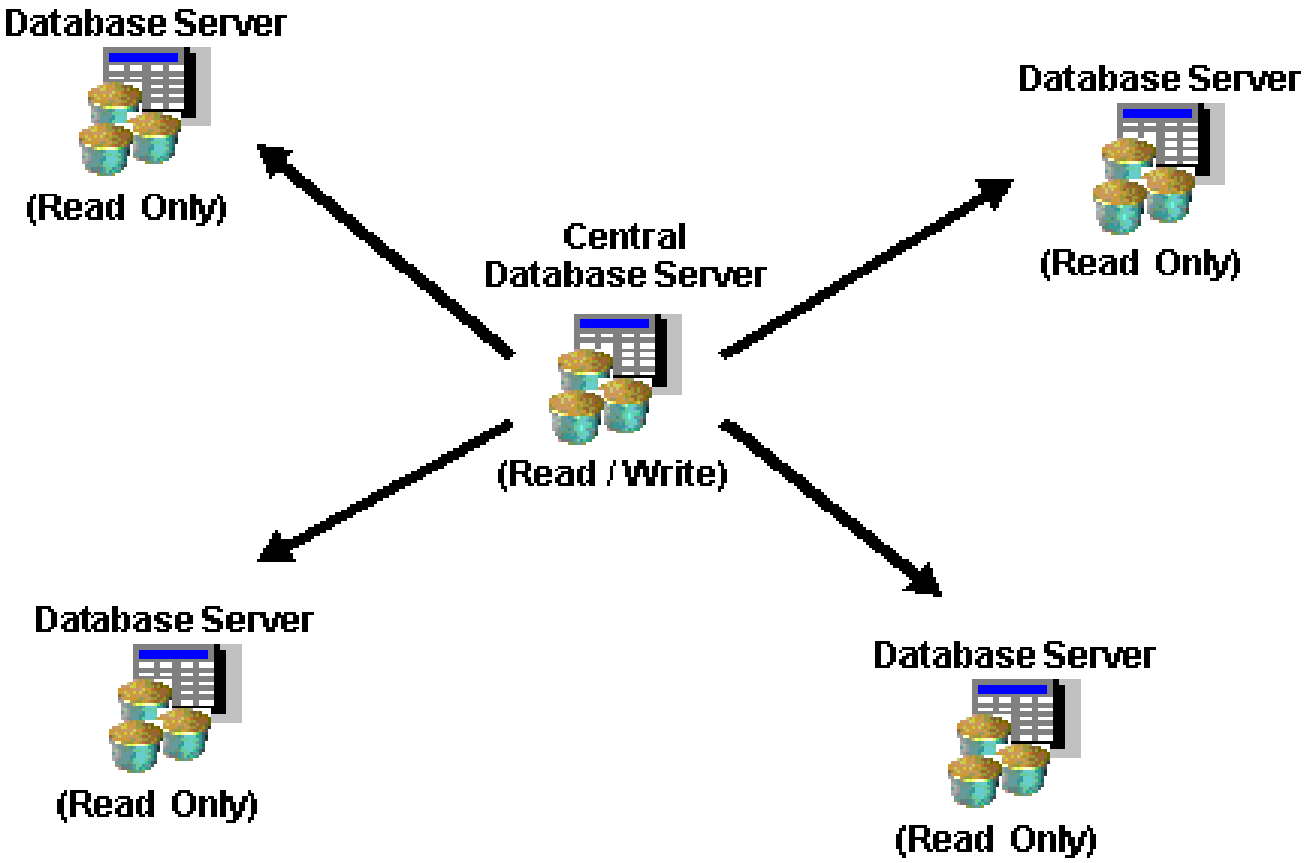
2地3中心部署方案二（Primary + HDR + 2个RSS）



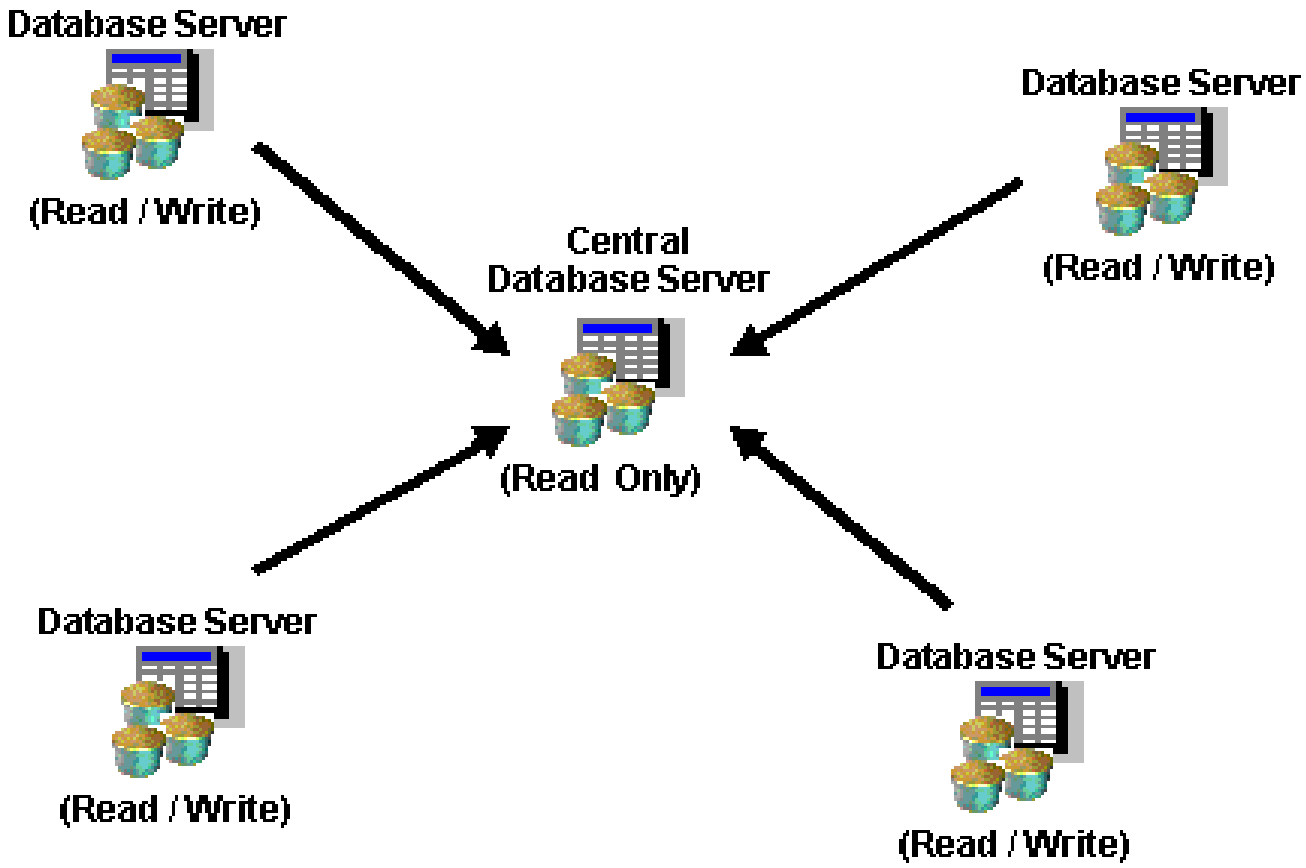
负载分区



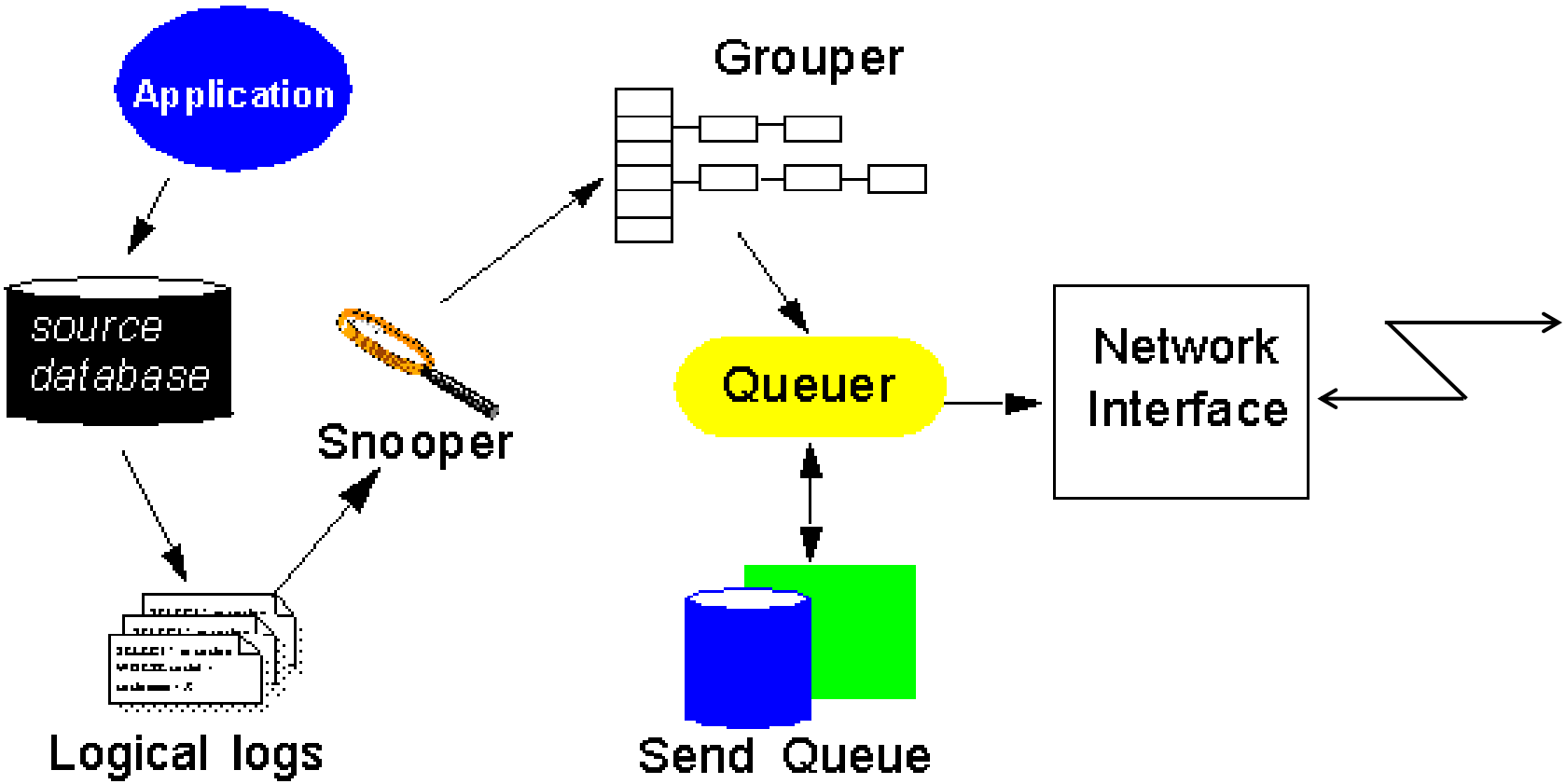
数据发布



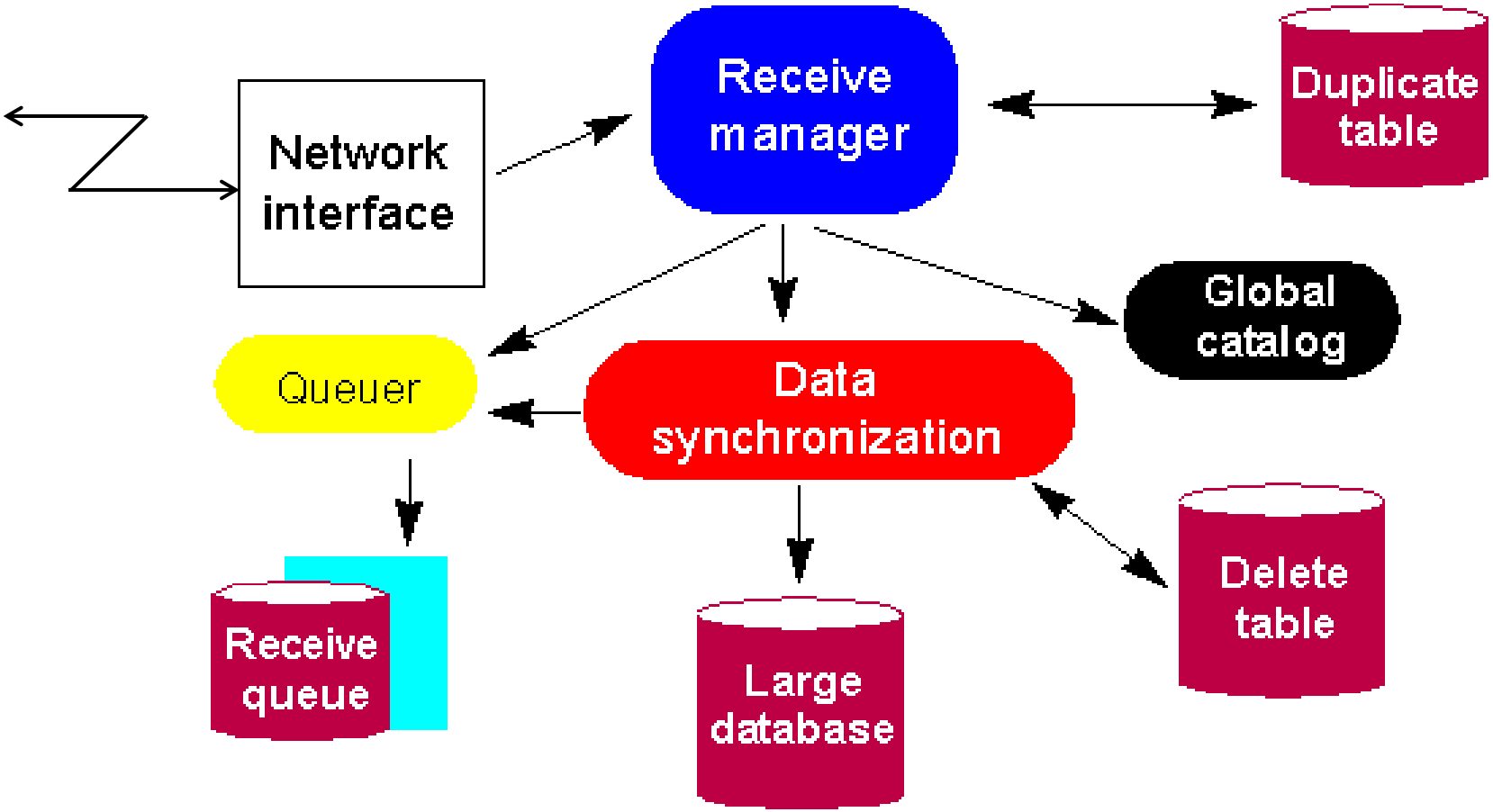
数据整合



源服务器数据流



目标服务器数据流



Vsettan
华胜信泰

感谢观赏

THANK YOU