
terprise-Ready Cluster Solu
Peter Cheng | 程勇
Open Solution Alliance Manage
2008 TargetSource

议程

- **企业面临的高可靠性挑战**
- 负载均衡解决方案
- JVM 集群解决方案
- 数据库集群解决方案

当前企业高可用性集群的挑战

最终使用者	“… 希望更专注于设计的本职工作少做或者不做 IT 工作”	• 高性能计算应用基于开放平台的提供与集成 • 简单易用的计算作业提交与监控
IT 系统管理人员	“… 部署管理一个高性能计算机群实在太困难了”	• 更加简单的部署与建立计算机集群 • 简单统一的集群管理环境并与现有的 IT 架构一致
应用开发者	“… 编程很困难 “… 没有足够好用的开发工具与环境”	• 零编程的集群配置模式 • 无需额外的二次开发

全面的集群解决方案

	服务器负载均衡	应用服务器集群	数据库集群
			
	Load Director	Terracotta	Continuent
系统			
用户群	政府、研究机构、大型商业企业、互联网企业		
典型客户	www.abc.com , www.real.com, www.eds.com , www.ctrip.com		

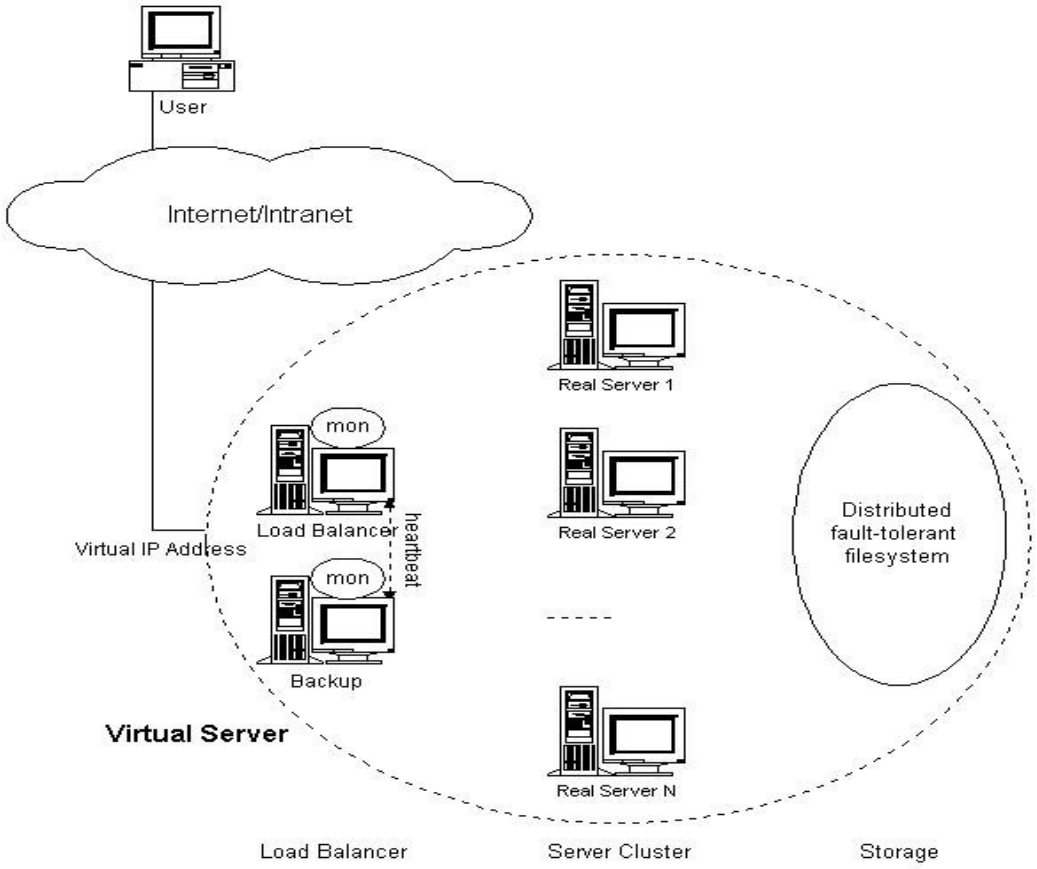
议程

- 企业面临的高可靠性挑战
- **负载均衡解决方案**
- JVM 集群解决方案
- 数据库集群解决方案
- Linux 操作系统高可靠性方案

核心技术：集群技术

虚拟服务器体系结构

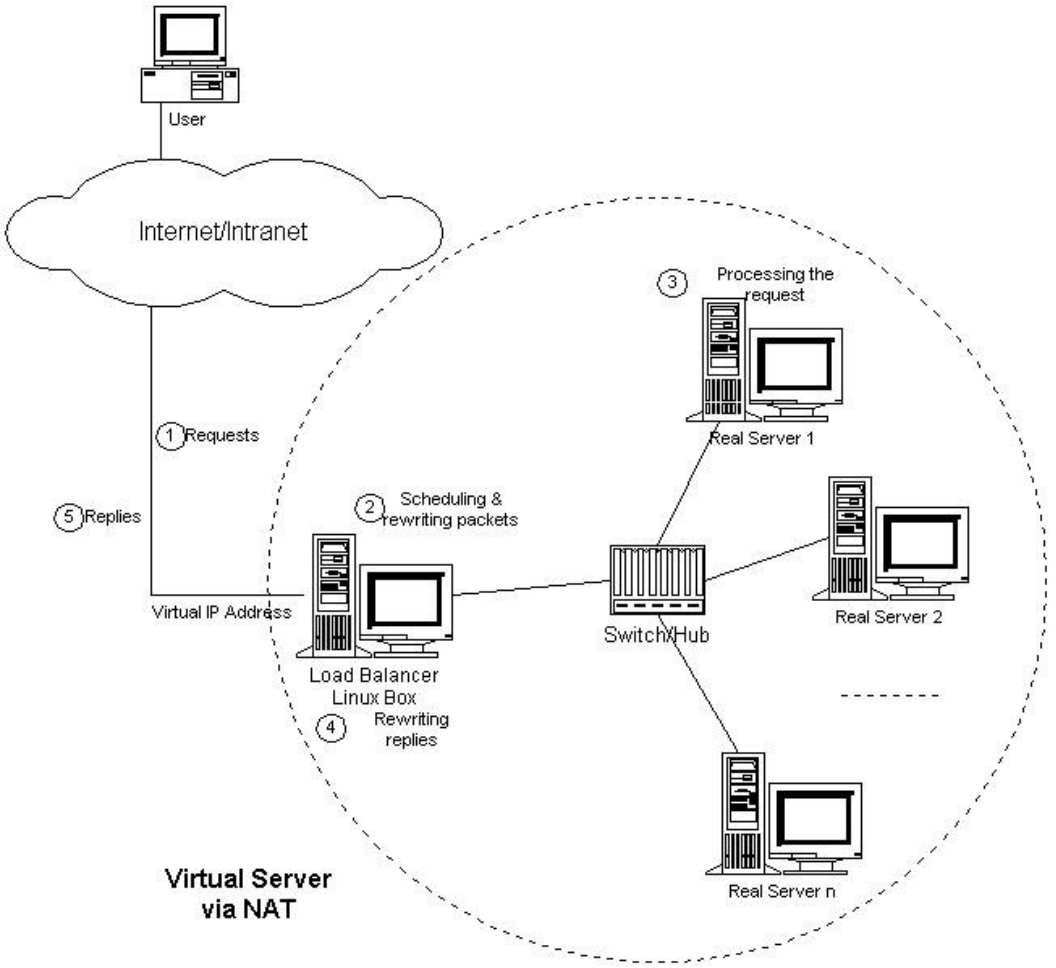
- 负载调度器
- 服务器池
- 后端存储



LVS 集群技术–技术简介 (1/3)

IP 负载平衡技术
—Virtual Server via NAT

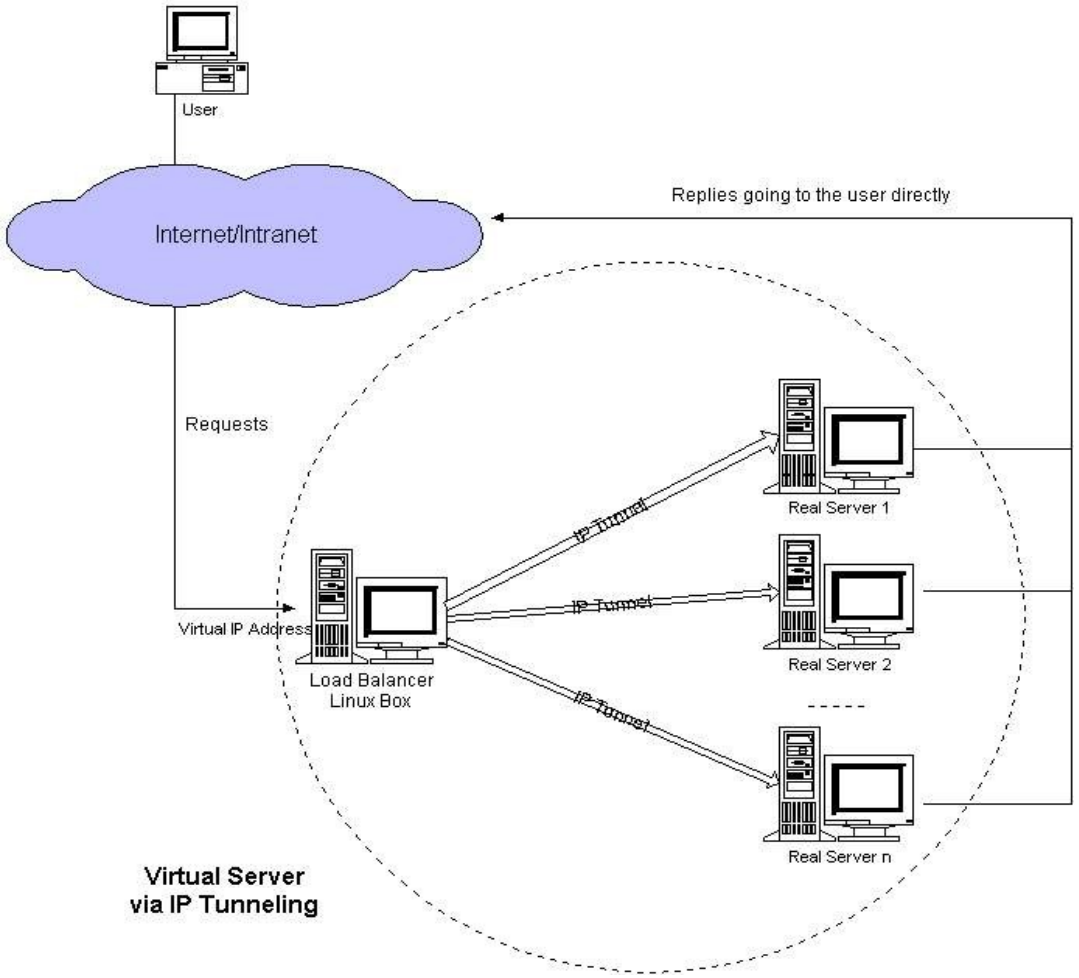
- 体系结构
- IP 隧道
- 工作流程
- 特　　点



LVS 集群技术–技术简介 (2/3)

IP 负载平衡技术
—Virtual Server via IP Tunneling

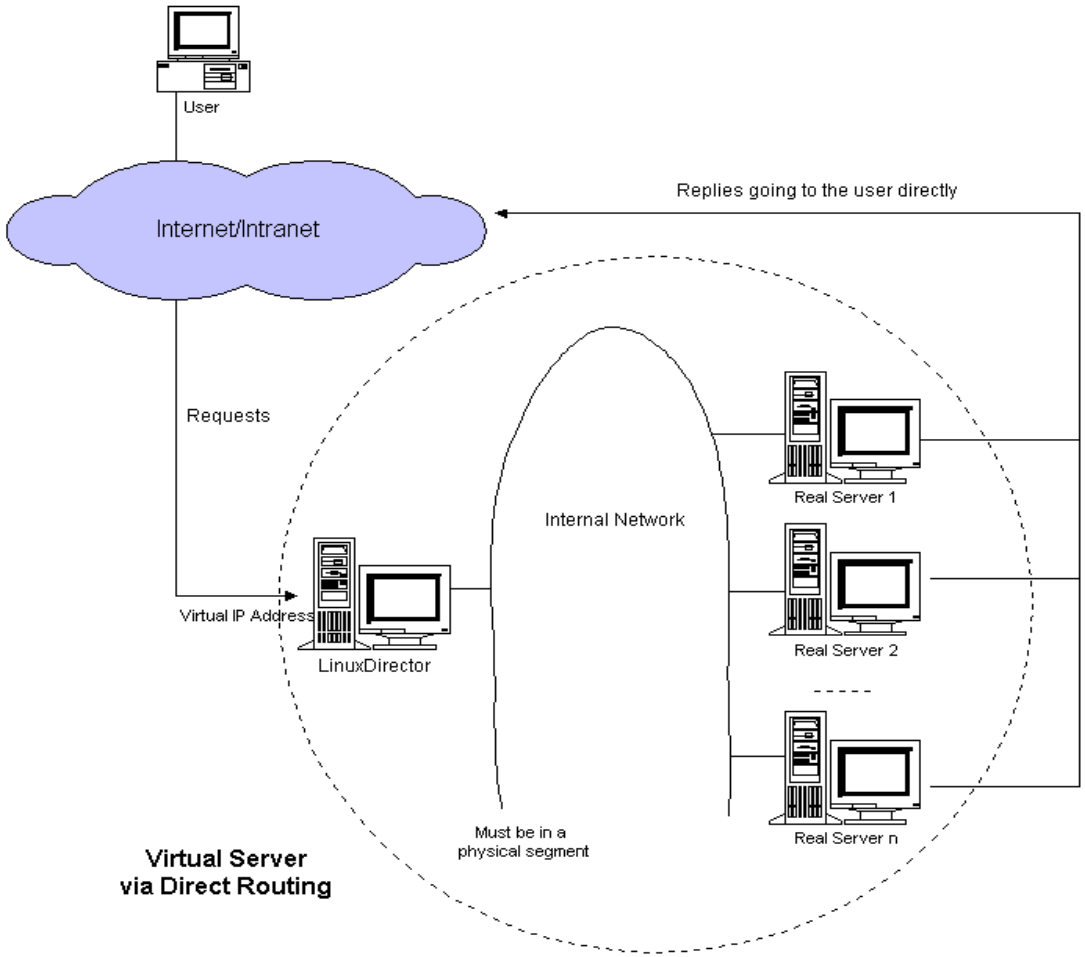
- 体系结构
- IP 隧道
- 工作流程
- 特　　点



LVS 集群技术–技术简介 (3/3)

IP 负载平衡技术
—Virtual Server via Direct Routing

- 体系结构
- 原理
- 工作流程
- 特　　点



LVS 集群技术–应用实例

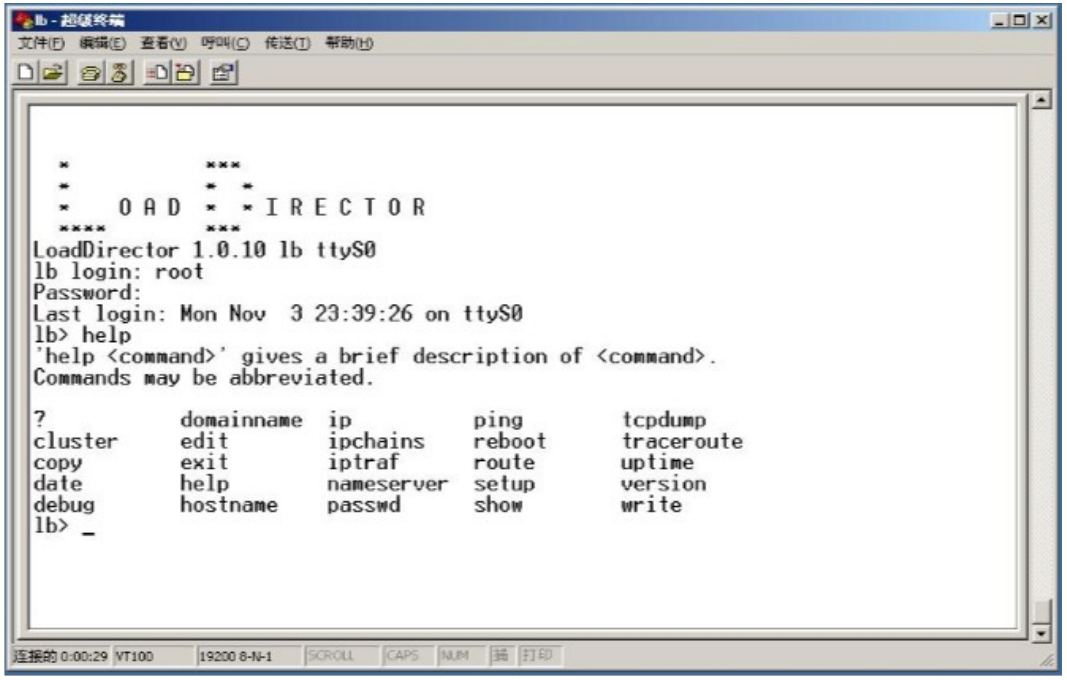
- ABC.com (American Broadcasting Corporation)
- www.real.com
- empas.com
- 英国国家 JANET Cache 网
- Dell
- EDS.com
- sourceforge.net & linux.com
- Sina.com
- QQ.com

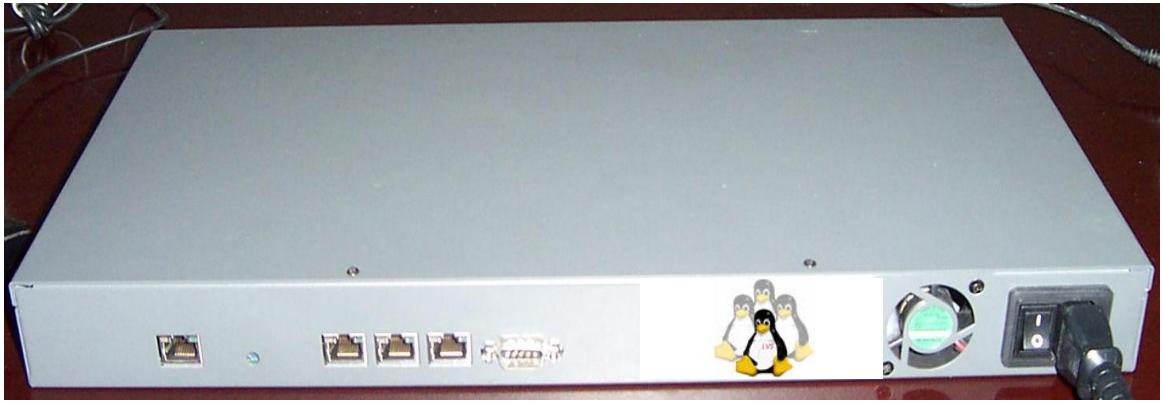
集群技术 – 用户评价

“We tried virtually all of the commercial load balancers, LVS beats them all for reliability, cost, manageability, you-name-it”
Jerry Glomph Black, Director, Internet & Technical Operations, Real Networks, Seattle Washington, USA
http://marc.theaimsgroup.com/?1=linux-virtual-server&m=95385809030794&w=2

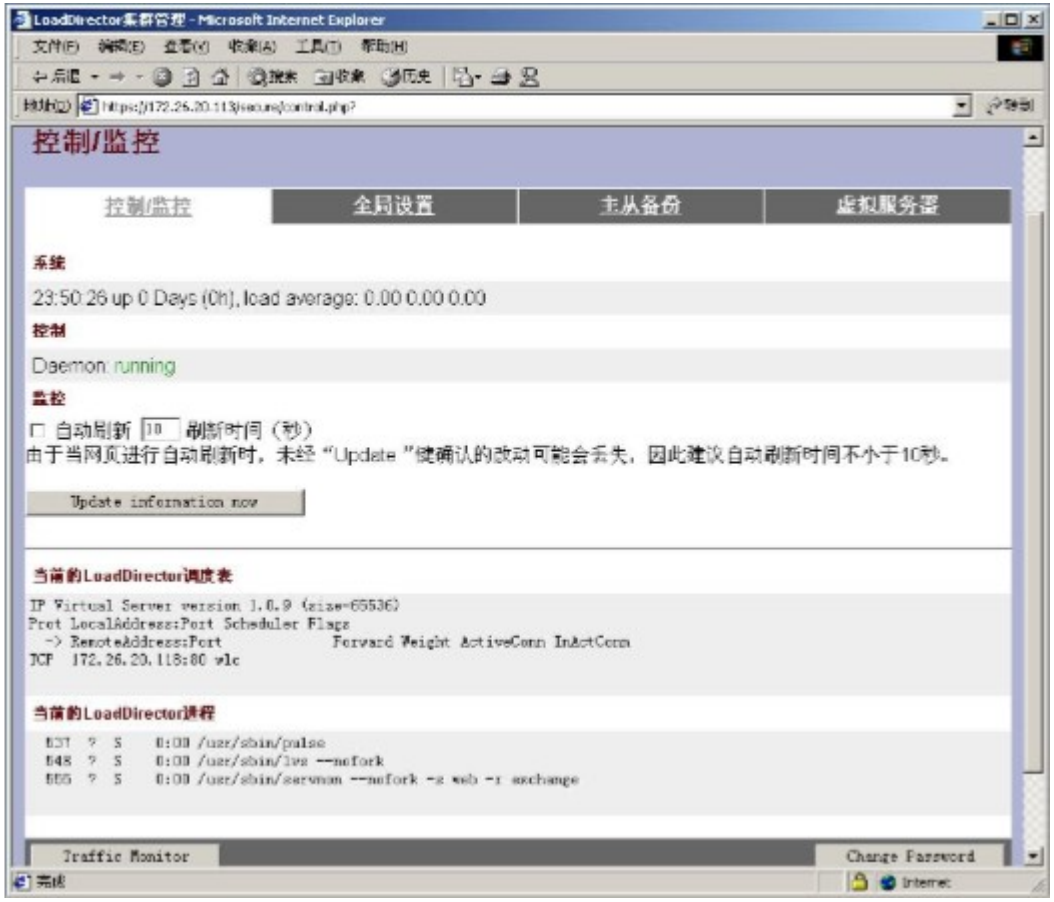
“I can say without a doubt that lvs toasts F5/BigIP solutions, at least in our real world implementations. I wouldn’t trade a good lvs box for a Cisco LocalDirector either”
Drew Streib, Information Architect, VA Linux Systems, USA
http://marc.theaimsgroup.com/?1=linux-virtual-server&m=95385694529750&w=2

Load Director CLI 配置界面





Load Director Web 管理监控界面



Load Director 集群技术–应用实例

- Load Director 用户
  - 携程网 ctrip.com
  - 浩方在线
  - Ku6.com
  - 世纪互联
  - 中国移动湖南分公司
  - 中国联通广东分公司
  - 广州白云机场
  - 湖南省信息中心
  - 晓通网络 www.xiaotong.com.cn
  - 万维易化 www.ezcross.com
  - 凯思昊鹏 www.hopen.com.cn
  - 鸿联九五 www.hl95.com
  - 河北邯郸信息网 www.hd.gov.cn
  - 轻点万维

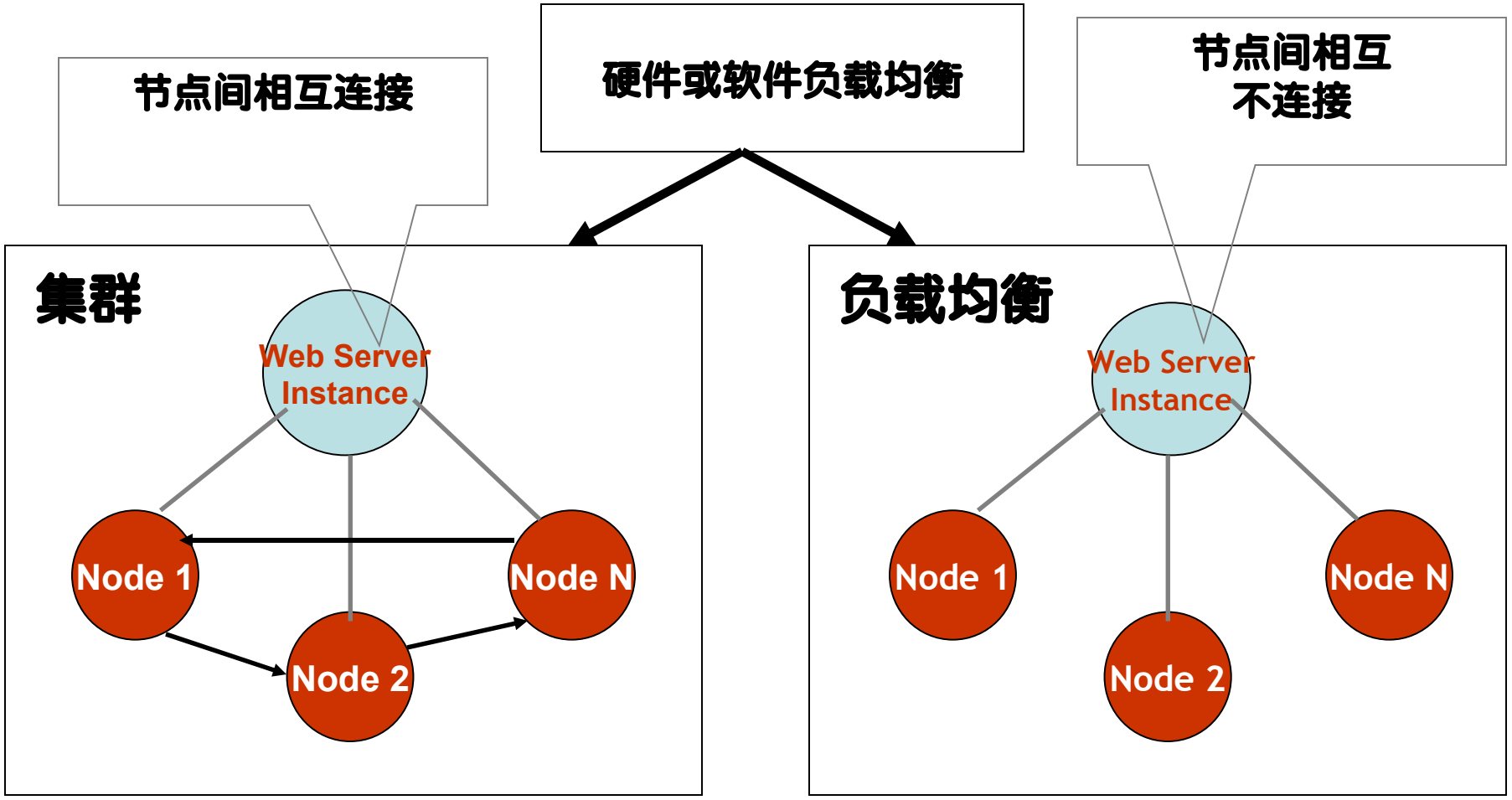
Load Director 竞争性分析

	Load Director 2000	F5 BIG/IP 2U
操作系统	• 支持任何 TCP/IP 操作系统 包括 :Linux, Solaris, AIX, HP UX, Mac/OS, Windows	• 支持 WIN NT 各种 Unix 平台和 Mac/OS
内存	• 512M，可扩展到 2G	• 512M 可扩展到 1G
网络端口 负载均衡技术	• Intel 10/100/1000 以太网口 3~5 个 • 网络地址转换 (NAT) • 直接路由 (Direct Routing) • IP 隧道 (IP Tunneling)	• Intel 10/100/1000 以太网口 3~5 个 • 网络地址转换 (NAT)
调度算法	• 轮叫和加权轮叫 • 最小连接和加权最小连接 • 基于局部性最小连接调度 • 带复制的基于局部性最小连接调度	• 轮叫 • 最小连接
最大系统 吞吐率	• 700Mbps(NAT) • 9Gbps(DR, TUN)	• 700Mbps(NAT)
最大可同 时连接数	• 500 万	• 500 万
价格	• LD2000 ￥119,500 • LD2000S ￥189,500	• F5 BIG-IP-3400 ￥370000 • F5 BIG-IP-1500 ￥230000

议程

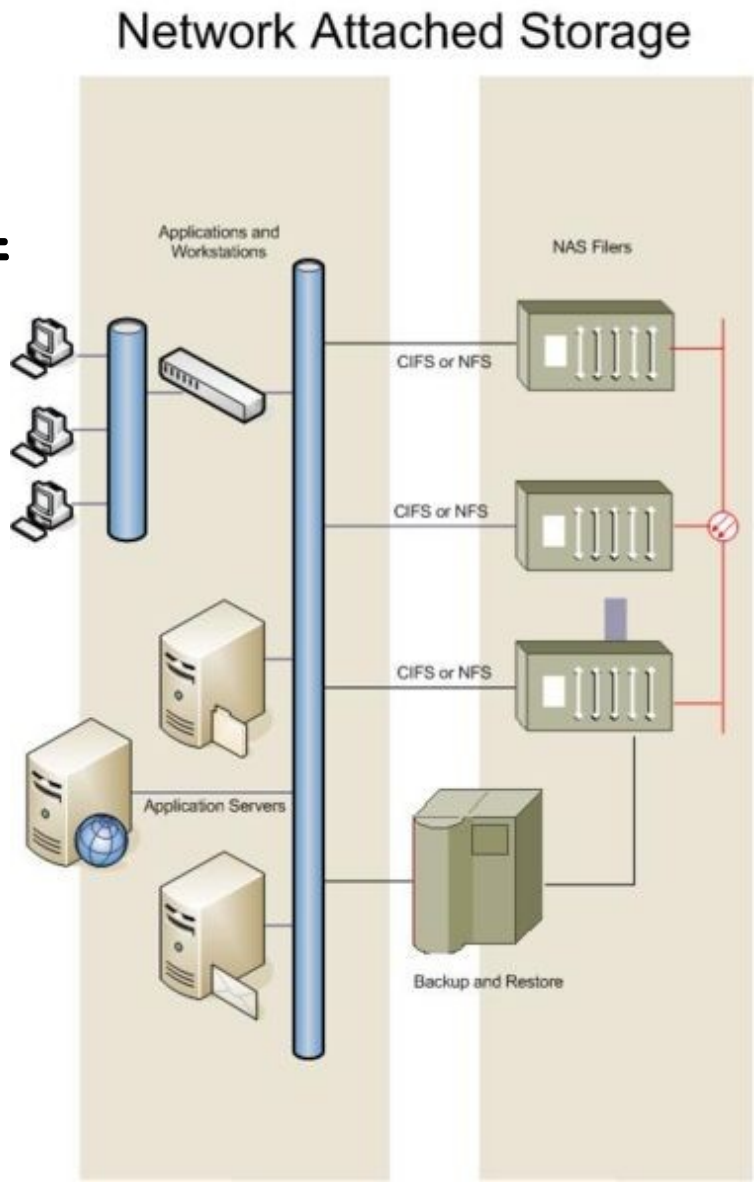
- 企业面临的高可靠性挑战
- 负载均衡解决方案
- **应用服务器集群解决方案**
- 数据库集群解决方案

负载均衡与集群的节点

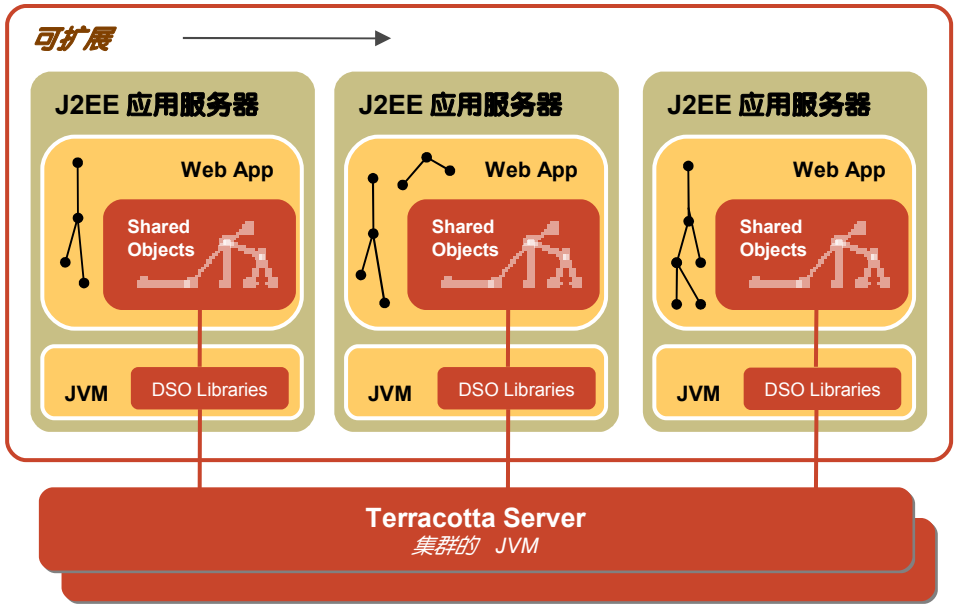

节点间相互连接
硬件或软件负载均衡
节点间相互
不连接
集群
Web Server
Instance
Node 1
Node N
Node 2
负载均衡
Web Server
Instance
Node 1
Node N
Node 2

什么是集群的 JVM?

- 多个 JVM 如何解决内存共享的问题? .
- JVM 内存共享比较满足一下两个基本条件
  - 1. 对于应用来说应该像 RAM 一样
  - 2. 必须作为一个基础设施服务



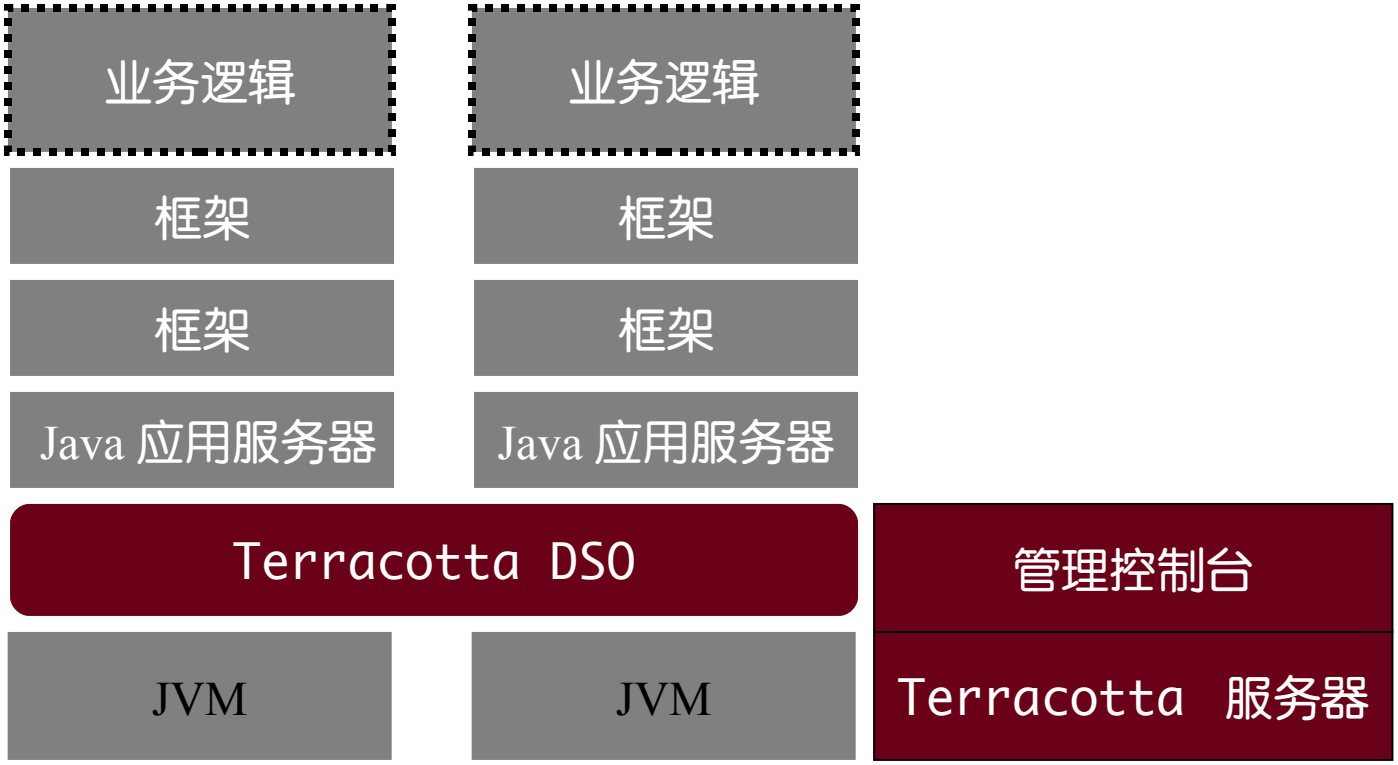
Terracotta JVM 集群架构： 集群你的 JavaEE 应用资源



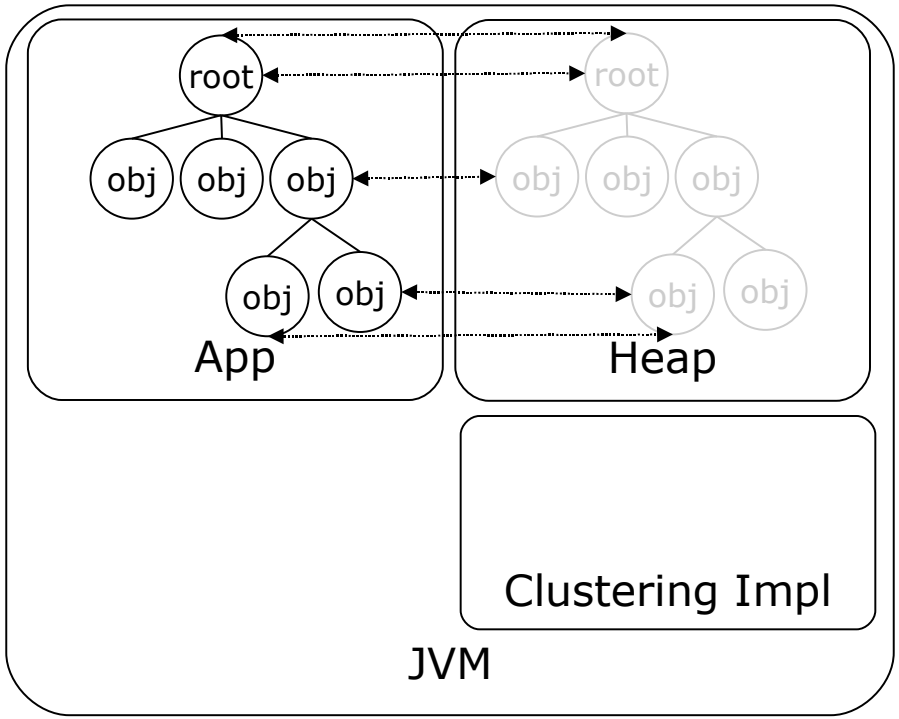
特性

- 内存堆级别的复制
- 中心化存储
- 虚拟内存
- 可监控和管理

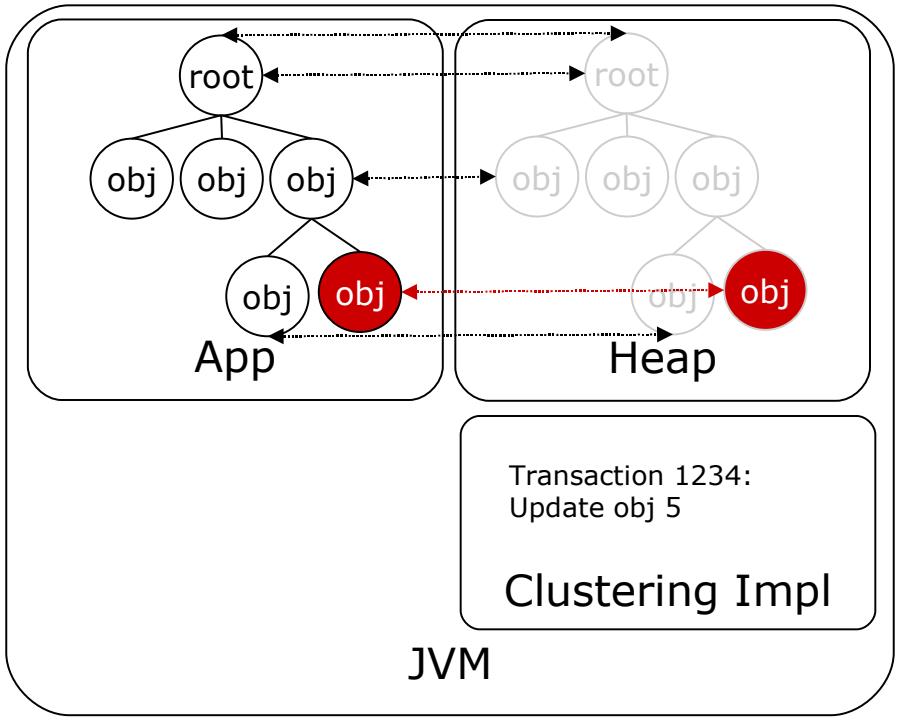
Terracotta 如何来实现?



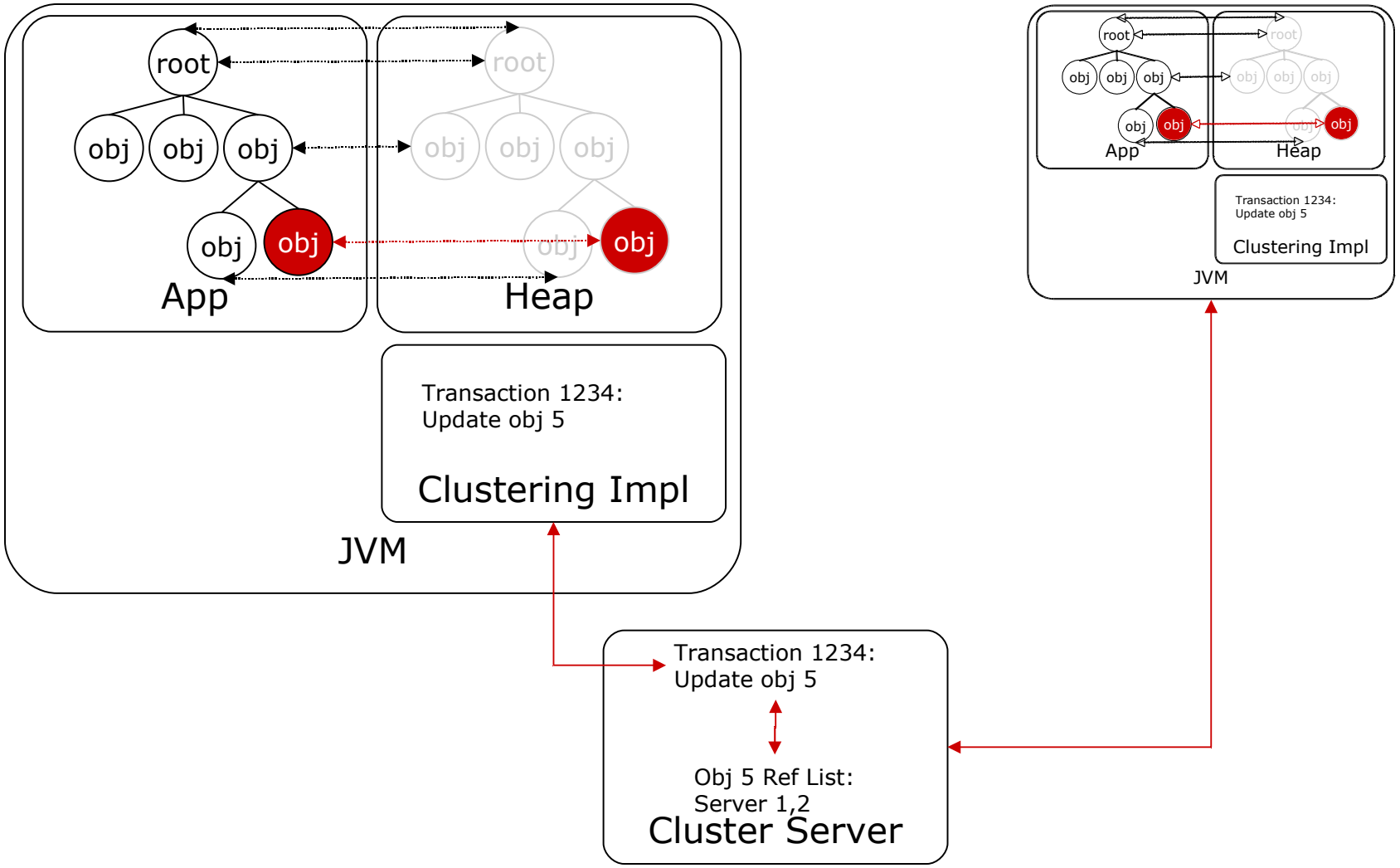
Terracotta 工作流程 (1/3)



Terracotta 工作流程 (2/3)



Terracotta 工作流程 (3/3)

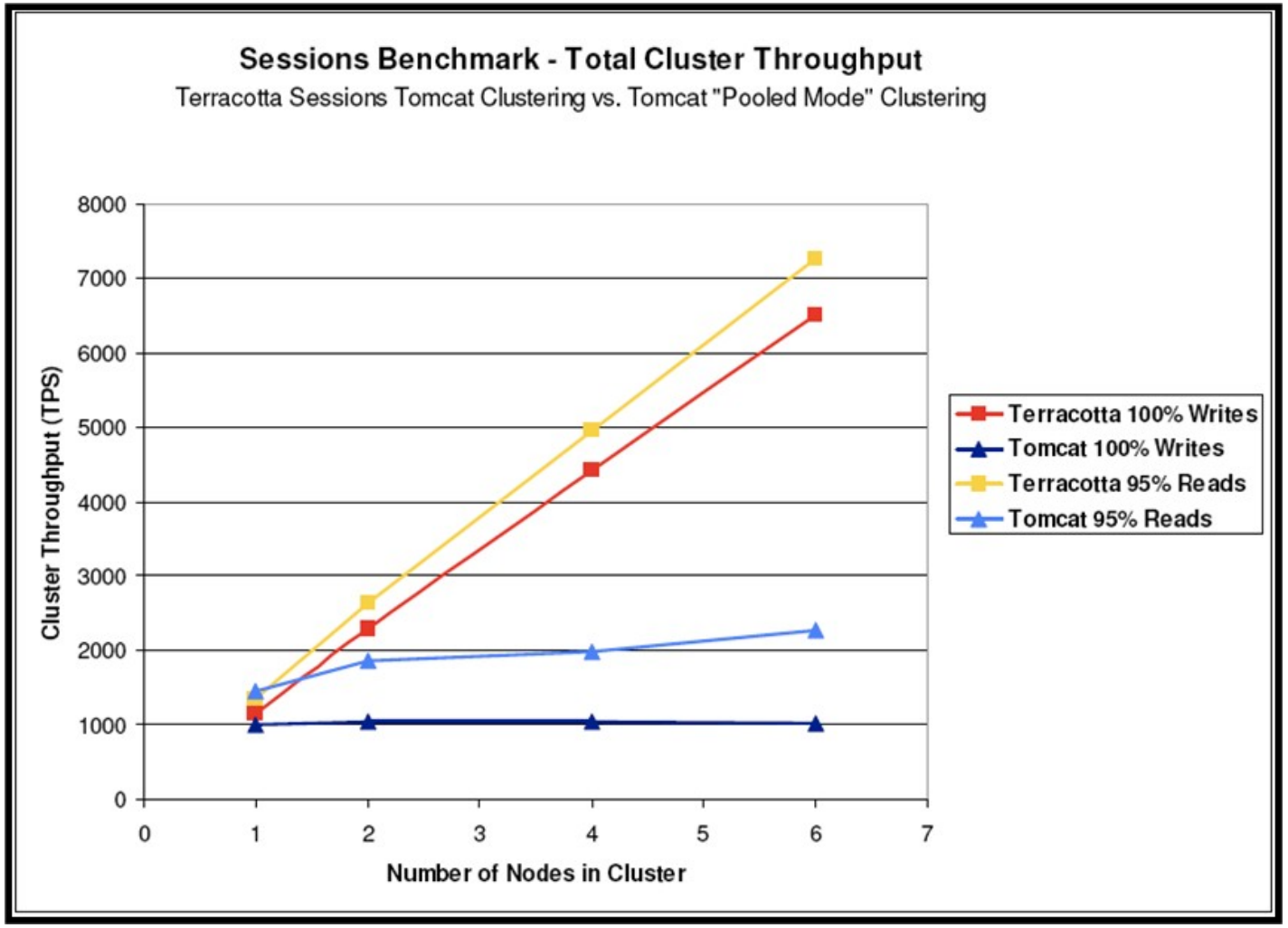


配置级管理，零编程 DEMO

Shared JTable (spreadsheet)

```
<terracotta-config>
  <dso>
    <server-host>localhost</server-host>
    <server-port>9510</server-port>
    <dso-client>

      <roots>
        <root>
          <field-name>demo.jtable.TableDemo.model</field-name>
        </root>
      </roots>

      <included-classes>
        <include><class-expression>demo..*</class-expression></include>
      </included-classes>

    </dso-client>
  </dso>

</terracotta-config>
```

Tomcat Http Session Cluster



竞争性分析 : Terracotta vs. Coherence

- **Terracotta: 分布式的 Java 运行环境**
  - 易于集成
  - 可视的中心化操作管理
  - 企业级的开放源码

- **Coherence: 开发者框架**
  - 入侵式的 API
  - 有限的可操作控制 – 非运行环境
  - Oracle 式的许可证

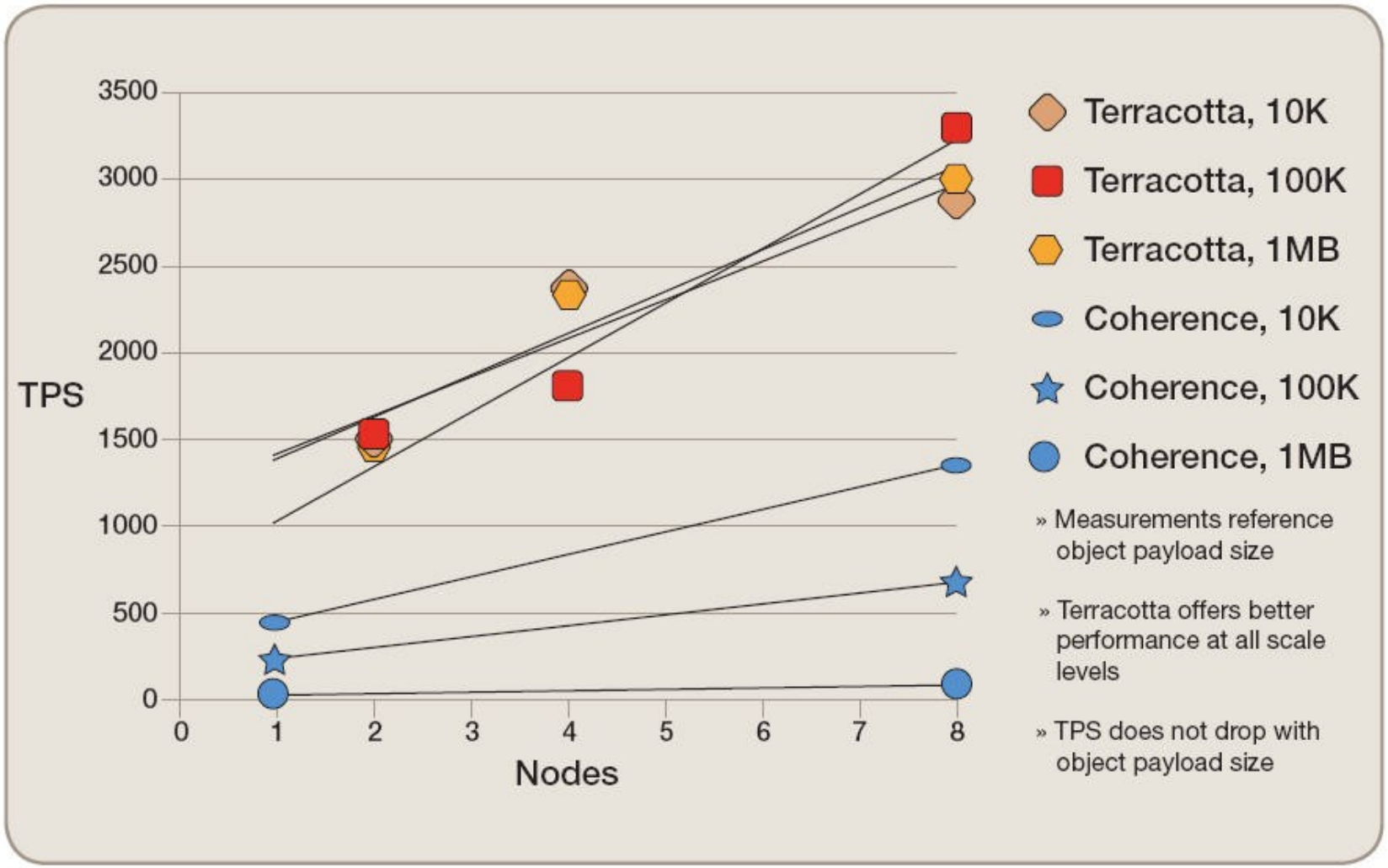
Terracotta 性能测试

- 测试环境
  - 文件系统： SAN/Linux ext3
  - JVM 内存堆大小 : 4GB
  - 硬件 : DL380
  - 存储类型： 内部磁盘 40G/HDB 335G Ext3
  - 内存： 12GB RAM
  - 网络： GigE
- 测试方法
  - 每秒交易数
  - CPU 利用率
  - 网络利用率
  - 内存特征

Terracotta 性能测试

- 评估标准
  - 高性能： 主机上服务能够处理交易的能力 (10,000TPS)
  - 易于使用： 技术应该使开发成本低廉并且保持简单。技术应该从应用中抽象分离出基础设施层所关注的部分 (集群，可扩展， HA)
  - 高可扩展： 服务应该能够满足业务增长的需求而扩展，并将运营和开发的成本降到最低。在相同的性能和 HA 下，离线主机系统能够处理 20 倍主机 (1MB) 系统容量的负载
  - 高可靠性： 服务应该满足跨企业的 99.999% 的 SLA
  - 运营监控与管理： 技术应该能够提供丰富的调试工具集，监控并管理这些集群服务

Terracotta vs Coherence 性能



测试结果关键点

- Terracotta 性能是 Oracle Coherence 的 6 倍
- Terracotta 实现应用需要 3 天而不是 6 周
- Terracotta 展现出卓越的运营监控和管理能力

Terracotta 竞争性分析

	Terracotta	Oracle Coherence
简单	• 选择适当的开发模式 • 选择最佳的框架和容器 • 自动判断工作负载	• 入侵式 API 必须显示编码 • 仅适用于数据缓存模式 • 缓存策略必须显示的配置
可扩展	• Delta 复制，精密的数据路由 = 高性能 • 有效的网络复制 • 在应用恶化服务器级别扩展	• 交易随可靠性和数据一致性规模化 • 复制是网络和 CPU 敏感的
高可用	• 没有单点故障 • 数据在磁盘上是持久化的	• 数据只在内存中复制 • 没有数据在哪存活概念
数据一致性	• 直接的共享状态模型 • 有保障的连贯更新 • ACID 兼容	• 高风险的数据失误 • 高扩展的交易一致性更新 • Get/Put 模式易损坏和出错
操作管理	• 服务器基础设施 • 特征鲜明，操作直接的控制台 • 开发与运营拥有公共的接口 • 持久化的数据意味着简单的提供 / 销毁	• 非运行时环境 • 有限的可视化与控制 • 没有数据在哪存活的概念 • 提供 / 销毁会带来数据损害
价格	• **$ 10000/JVM**	• **$ 30000/CPU**

Enterprise-Ready Cluster Solution

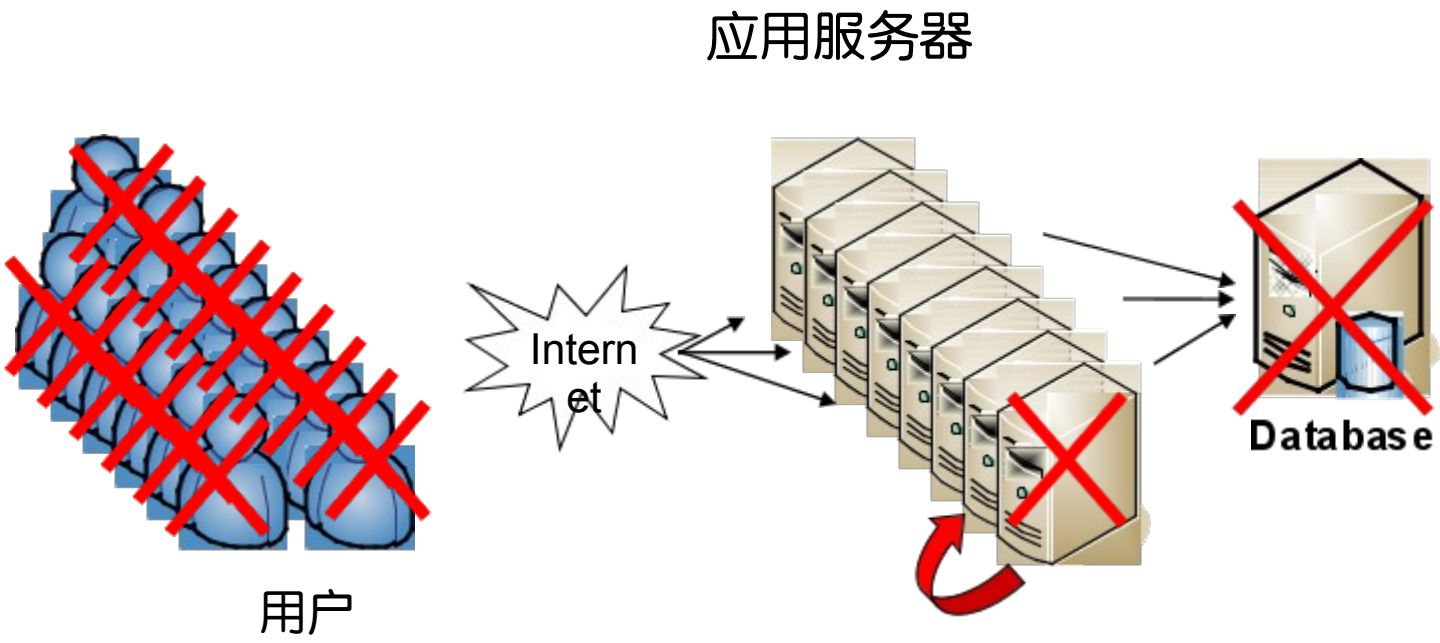
企业级用户

议程

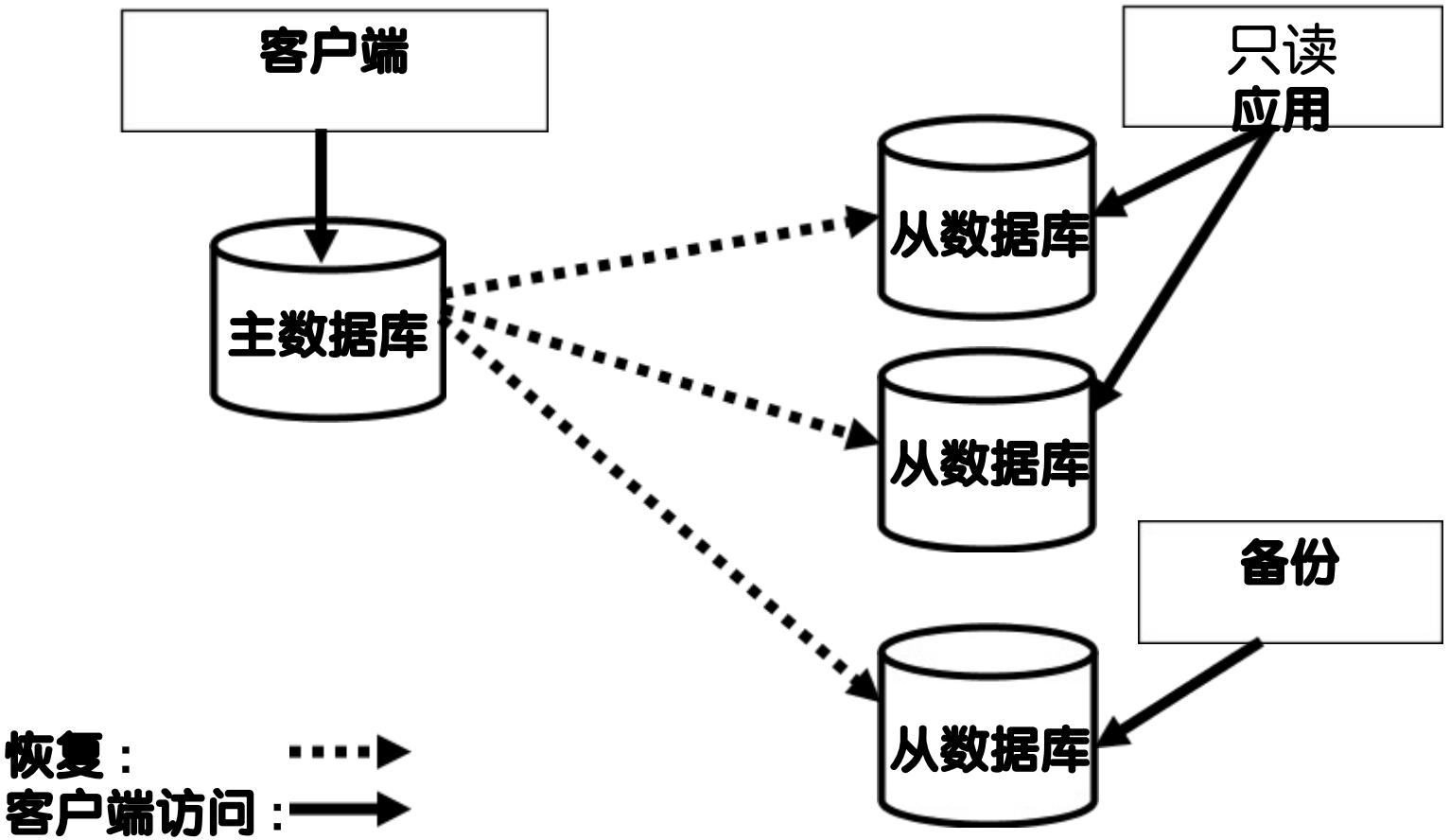
- 企业面临的高可靠性挑战
- 负载均衡解决方案
- 应用服务器集群解决方案
- **数据库集群解决方案**

DB 的高可用性问题

- 客户端连接到应用服务器
- 应用都基于数据库来构建
- 大量的工具解决应用层的可用性问题
- **数据库仍然有单点故障**



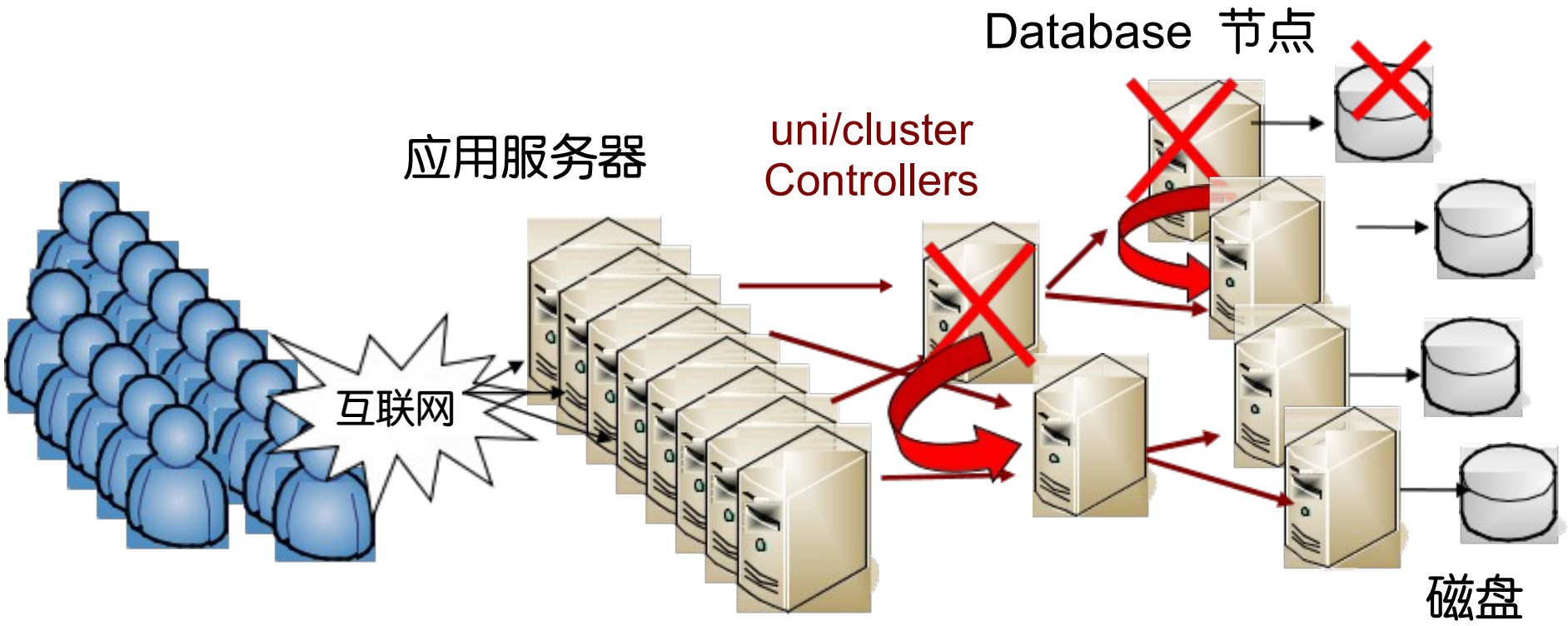
DBMS 集群常见解决方案 例如：MySQL



DBMS 集群解决方案 – 限制

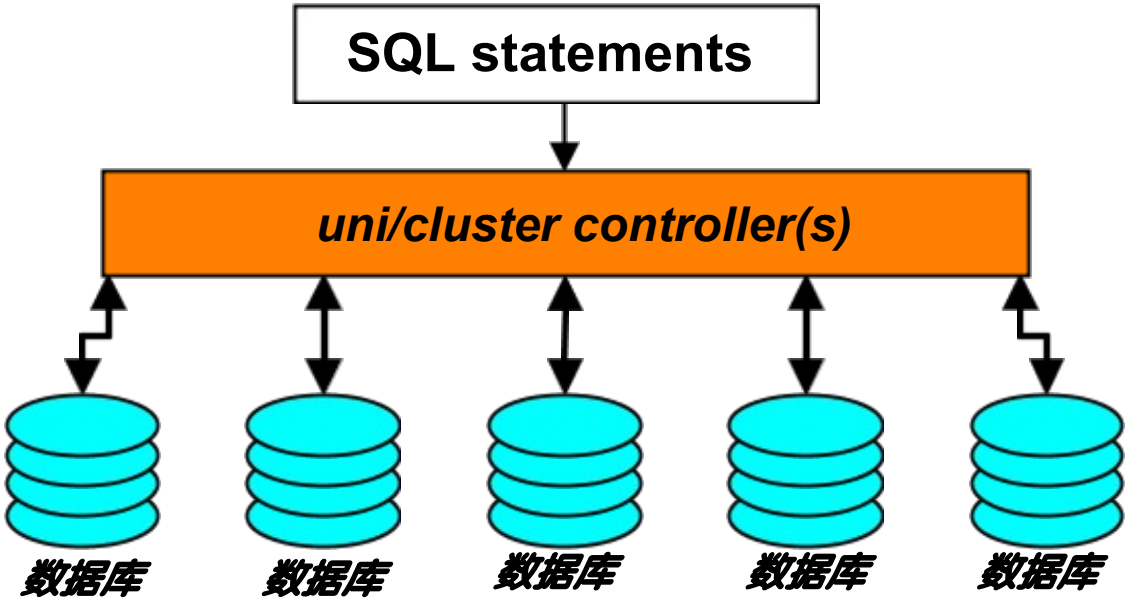
- 主数据仍然会是单点故障
- 故障恢复不是自动的，并不能透明的完成
- 异步更新 – 从数据库总是在后端
- 配置和管理非常复杂
- 没有一致的复制管理方式
  - MySQL vs. PostgreSQL

Uni/Cluster 数据库解决方案

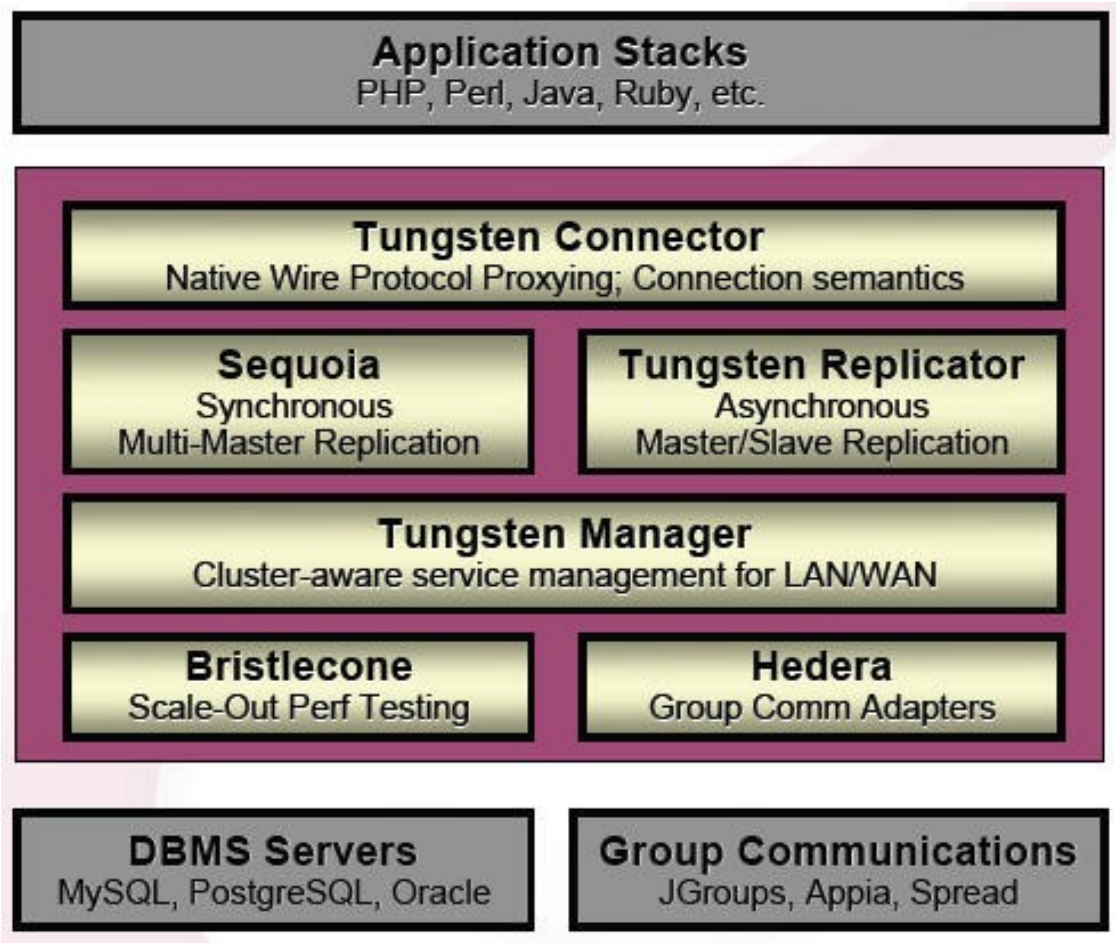


冗余的数据库集群架构

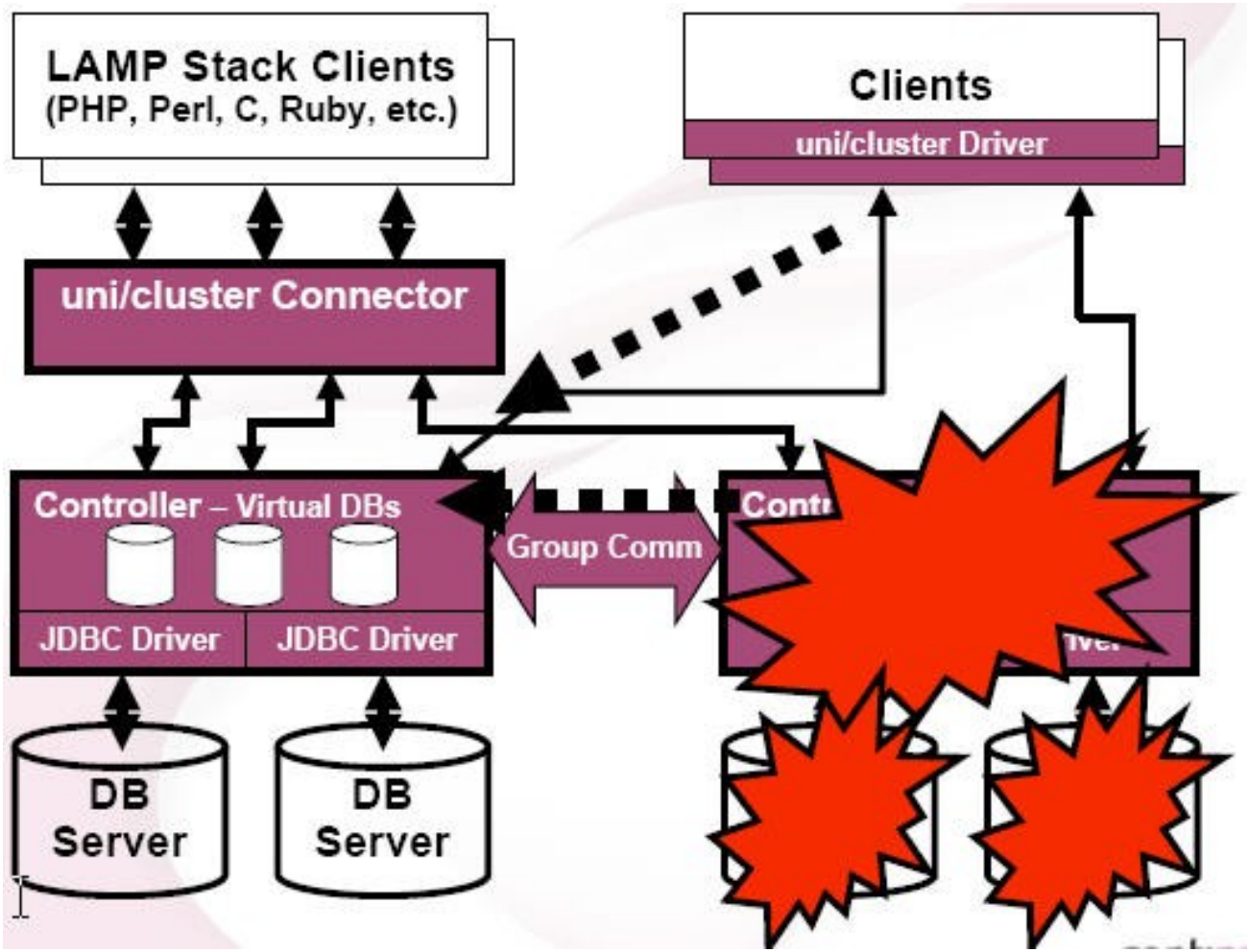
- **和磁盘 RAID 概念一样**
- **uni/cluster 数据库集群中间件**
  - *透明性*：为客户端提供数据库的单一视图 client
  - *可扩展性*：平衡后台数据库的负载
  - *高可靠性*：管理实务，重新恢复和自动故障切换 .



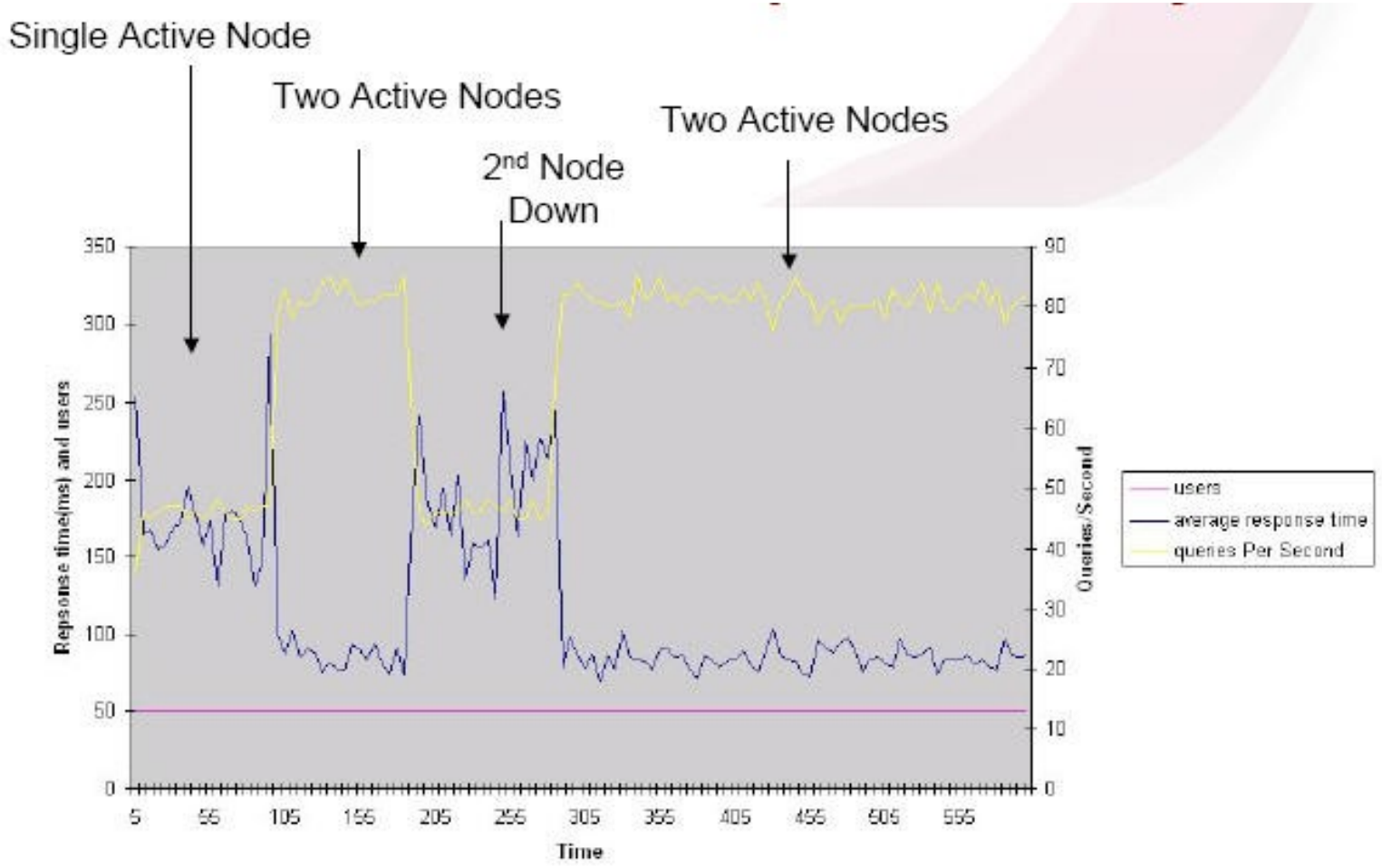
Tungsten


Application Stacks
PHP, Perl, Java, Ruby, etc.
Tungsten Connector
Native Wire Protocol Proxying; Connection semantics
Sequoia
Synchronous
Multi-Master Replication
Tungsten Replicator
Asynchronous
Master/Slave Replication
Tungsten Manager
Cluster-aware service management for LAN/WAN
Bristlecone
Scale-Out Perf Testing
Hedera
Group Comm Adapters
DBMS Servers
MySQL, PostgreSQL, Oracle
Group Communications
JGroups, Appia, Spread

Failover/Recover



uni/cluster - Non-stop Availability



Uni/Cluster GUI

Continuent 竞争性分析

	Continuent	Oracle RAC
数据库	• 支持 MYSQL, PostgreSQL • Oracle	• 仅适用于 Oracle
可扩展	• 多层次的扩展框架， 可解决不同层面的 集群问题	• 独立的集群产品， 仅针对 Oracle 数据库
高可用	• 没有单点故障 • 错误故障自动恢复	• 数据请求负载均衡 • 错误故障自动恢复
数据一致性	• 直接的共享状态模型 • 有保障的连贯更新 • ACID 兼容	• 高风险的数据失误 • 高扩展的交易一致性更新 • Get/Put 模式易损坏和出错
操作管理	• 提供统一的管理 监控客户端	• 依赖于 Oracle DB 的控制台
价格	• **¥ 112000/2CPU/5*8**	• **¥ 330000/License**