



基于DPDK实现的LB支撑阿里 巴巴双11业务

阿里云—梁军（刺背）

什么是负载均衡

什么是负载均衡

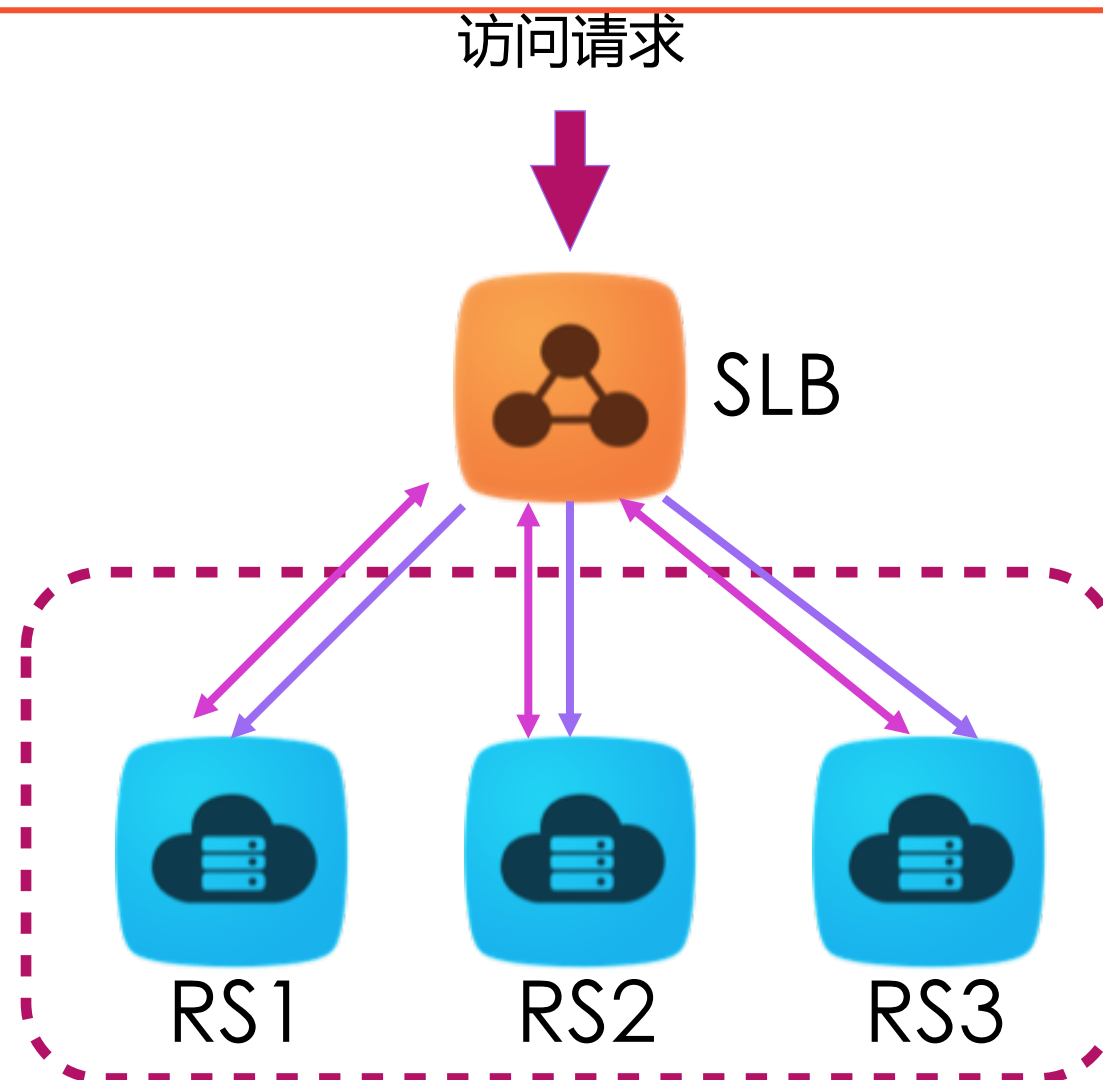
通过某种负载分担技术，将外部发送来的请求均匀分配到对称结构中的某一台服务器上，而接收到请求的服务器独立地回应客户的请求。

负载均衡解决了什么问题

分担负载，容灾切换。

负载均衡的实现

硬件负载均衡、**软件负载均衡**，DNS

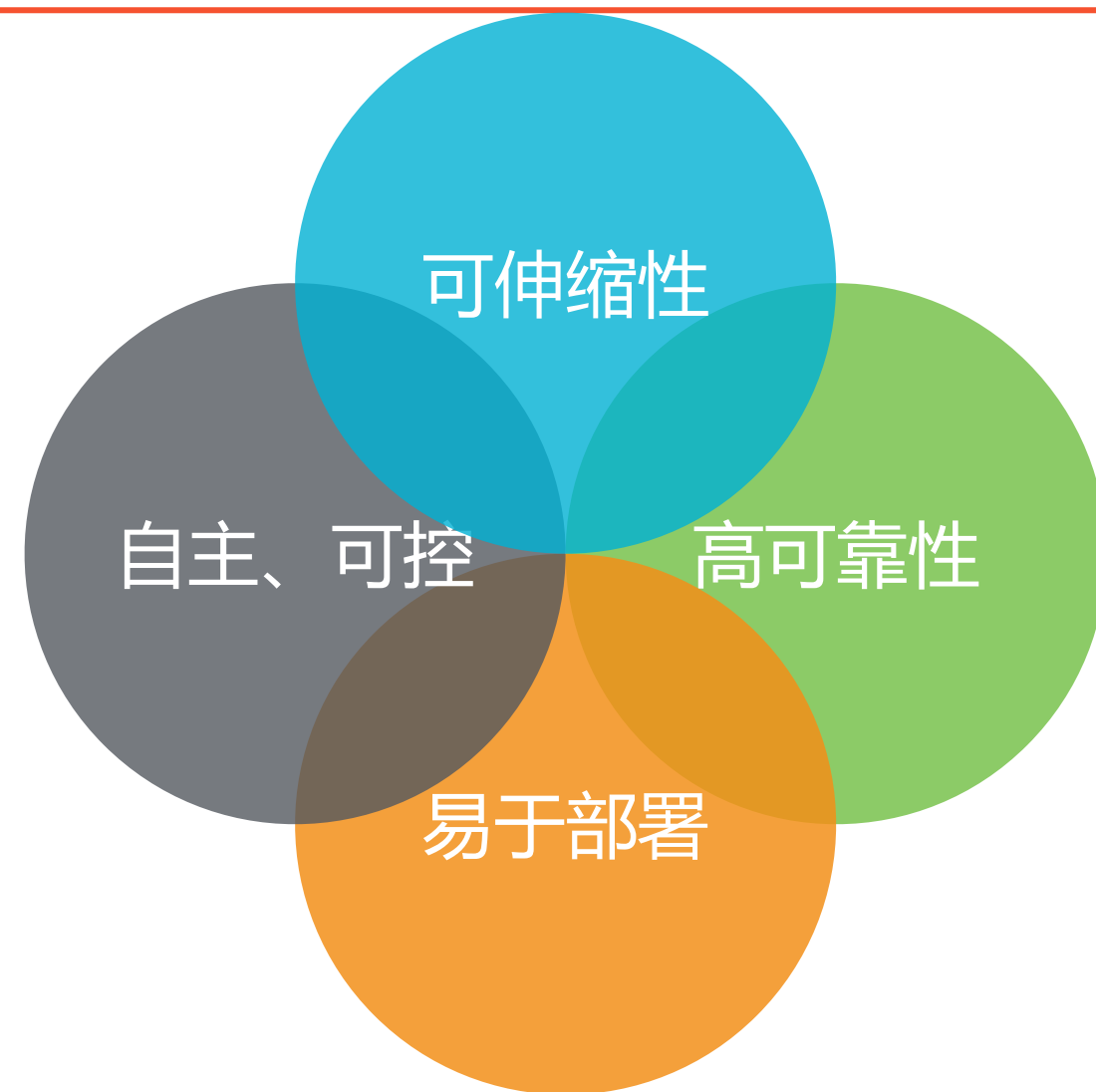


SLB (server load balancer)

- 老版本基于内核实现，已经开源
- 新版本基于DPDK实现

优势

- 自主、可控
- 可伸缩性
- 高可靠性
- 易于部署



基于DPDK开发的阿里负载均衡

DPDK

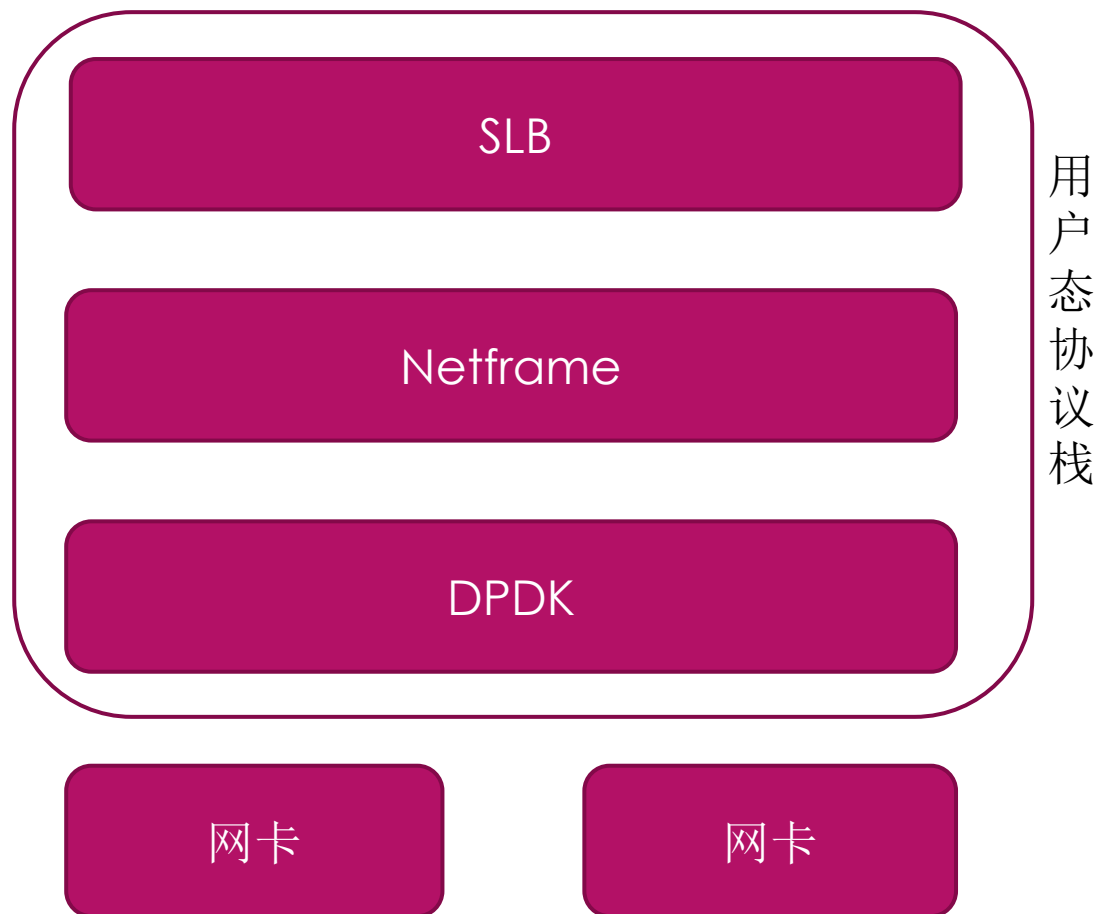
- 提供高性能的平台
- 用户态处理数据包能力

Netframe平台

- 基于DPDK自研的高性能2/3层用户态网络协议栈
- 提供了高效的开发平台

SLB (负载均衡)

- 统一全阿里巴巴的负载均衡技术架构，电商，交易，阿里云等
- 支持业务的快速增长
- 支持双11等突发流量场景

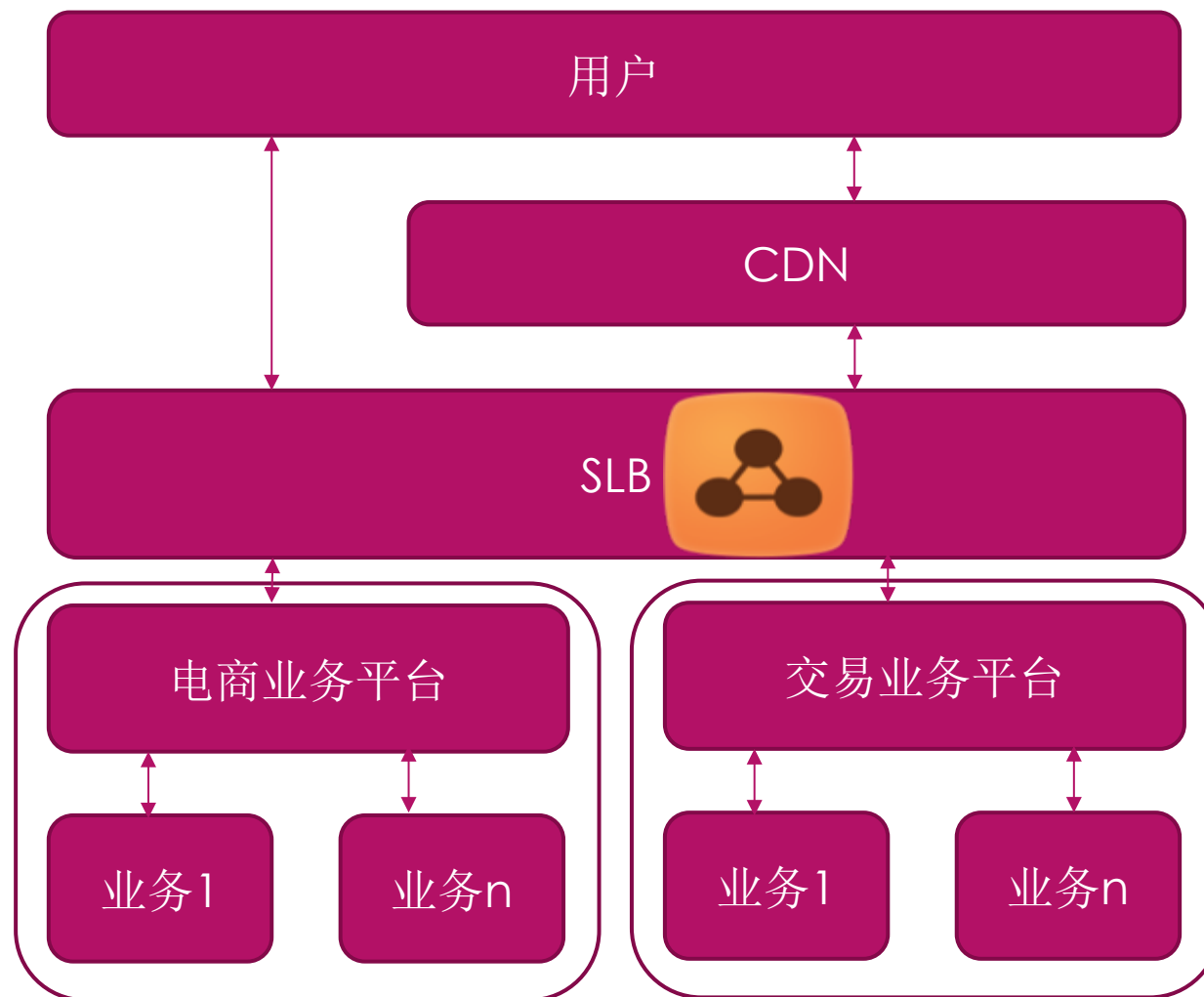


双11中负载均衡处于流量入口位置

流量入口

- CDN回源的入口
- 部分业务的直接入口

稳定性



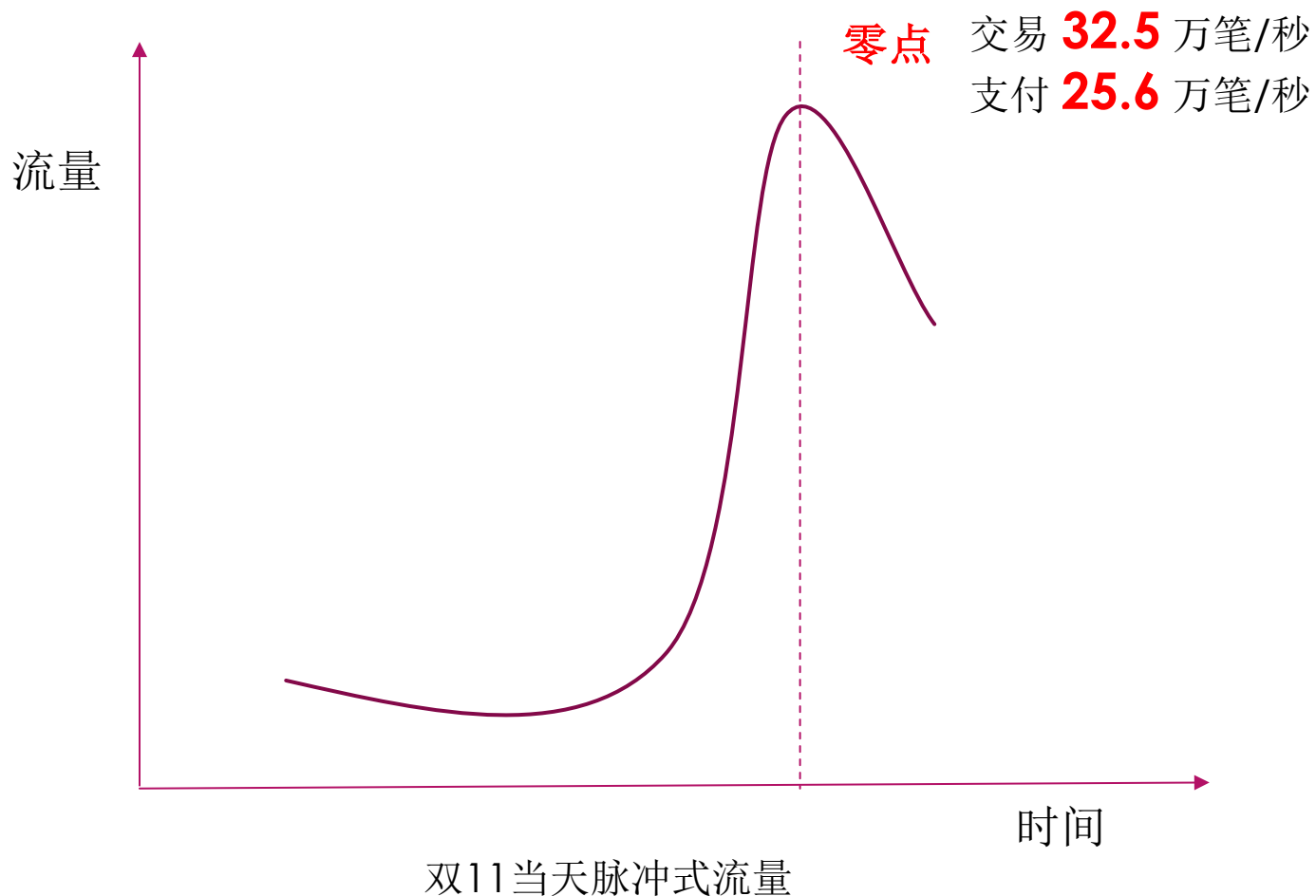
基于DPDK 的负载均衡支撑双11活动

支持双11，春节红包等大促活动

- 已支撑2016，2017的双11大促活动
- 也将支持2018年的双11大促

面临的挑战

- 如何应对**脉冲式**的流量模型
- 如何实现**业务无感**的版本升级与容灾
- 如何提供**高可用**的服务

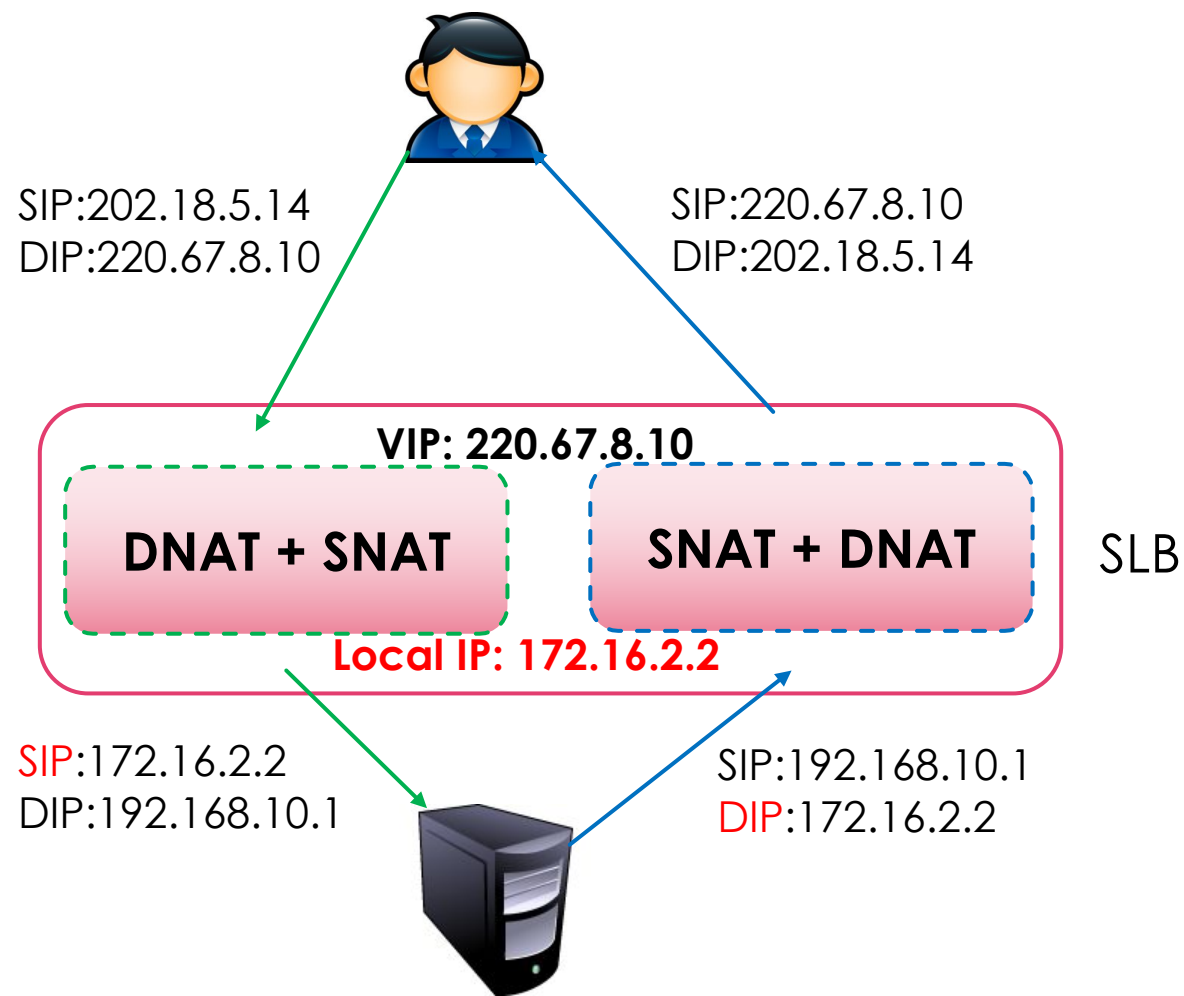


双11负载均衡使用FNAT转发模式

部署灵活

- 摆脱网络的限制，网络可达即可

跨机房容灾



FNAT转发模型

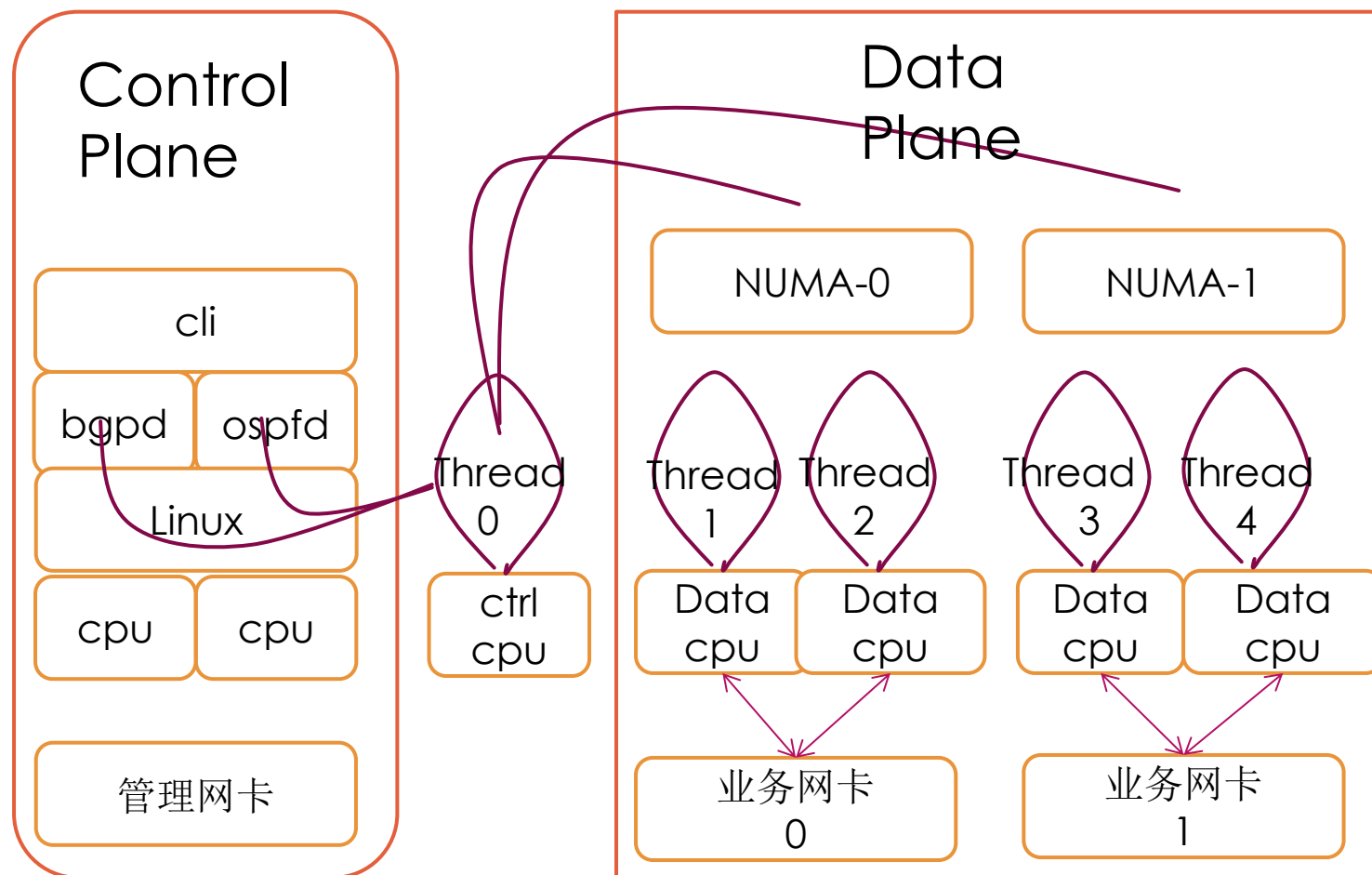
高性能的负载均衡——基于DPDK平台

利用DPDK平台的优势，
打造极轻量的高效数据转发面

- 无中断
- 无抢占
- 无系统调用
- 无锁
- CPU cache
- TLB
- NUMA

控制面与转发面分离设计
原则

- 轻量的转发面
- 复杂的控制面



SLB的系统架构图

高性能的负载均衡——软硬件结合的转发架构

RSS (Receive Side Scaling)

- 从client 的数据包打散到不同的CPU

FDIR(Flow Director)

- 特定的目的IP地址到特定的CPU

FNAT

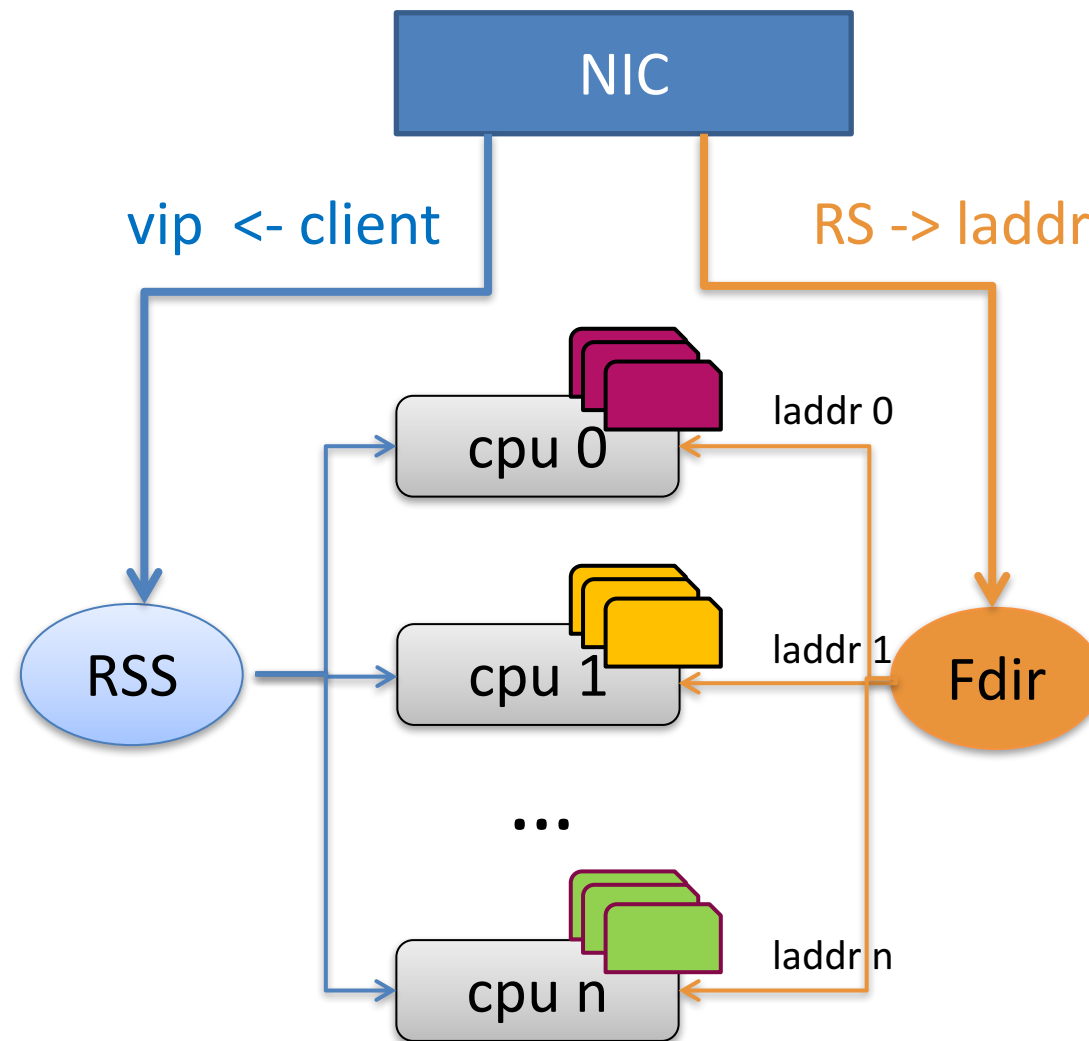
- 每个core都有自己的 local address

进出同CPU

- 利用网卡的RSS/FDIR功能实现同一个流进出同Core的转发架构

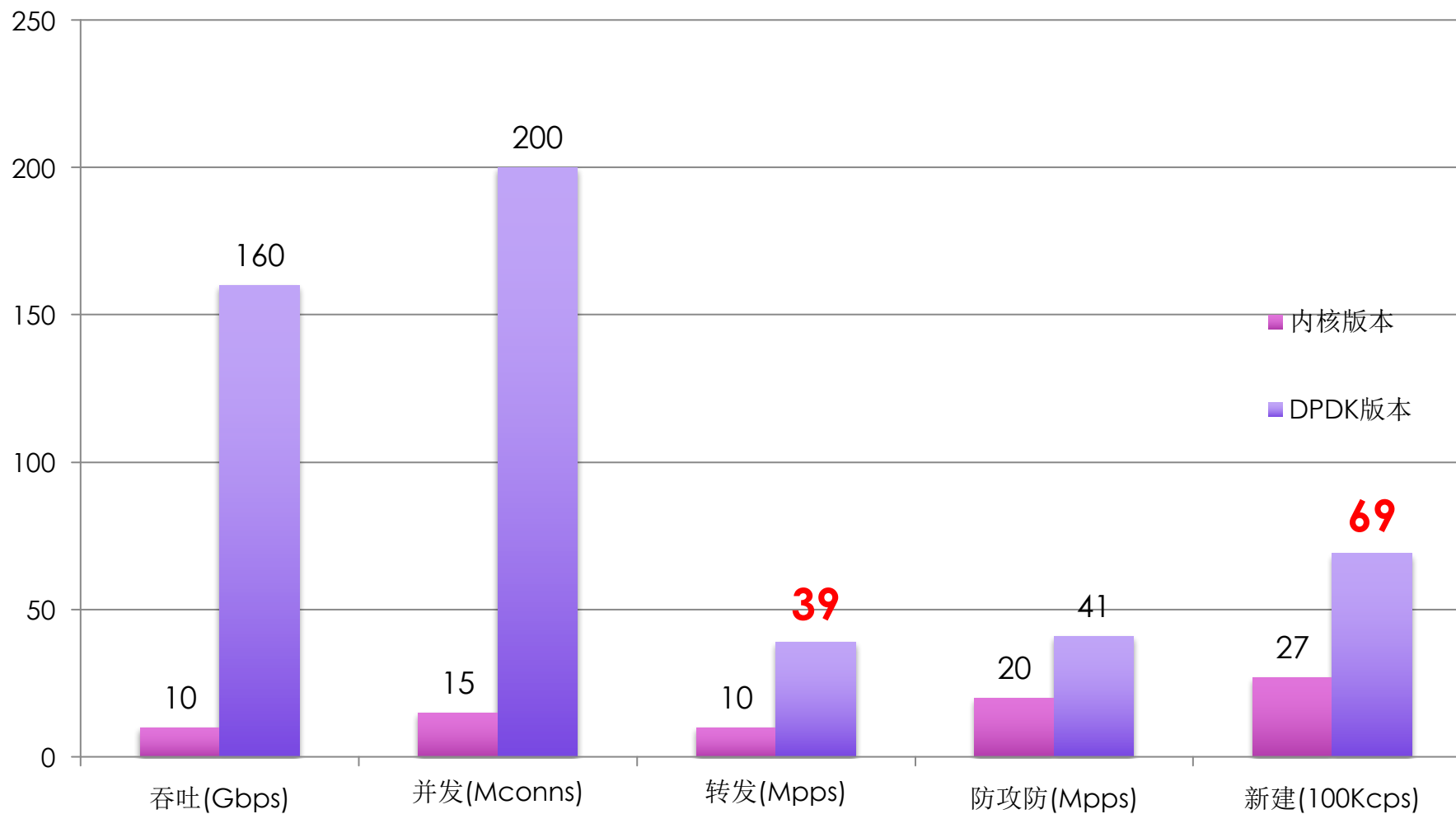
多CPU并发处理

- 数据结构Per-CPU化，CPU 完全并行化处理数据包，性能大幅提升



SLB并发的数据包处理架构

基于DPDK版本的负载均衡性能大幅提升



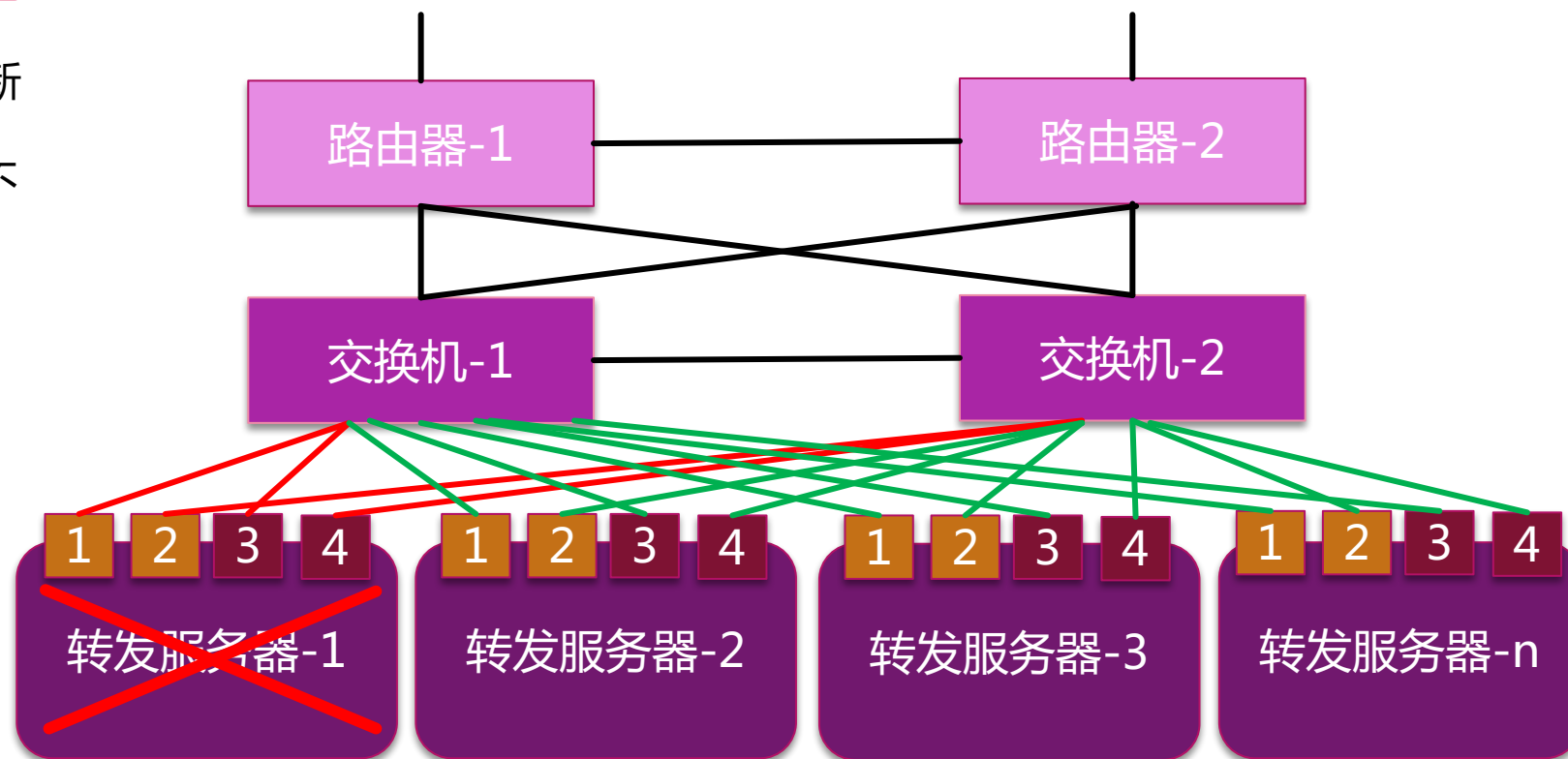
无感容灾与升级——连接断开问题

问题

- 单机故障时，ECMP 重新选路导致**绝**
大数连接中断
- 单网卡或者链路故障时，ECMP 重新选路导致连接中断
- 交换机故障时，切换交换机，选路不同导致连接中断
- 集群通过轮转升级版本时，导致 ECMP 重新选路导致连接中断

影响

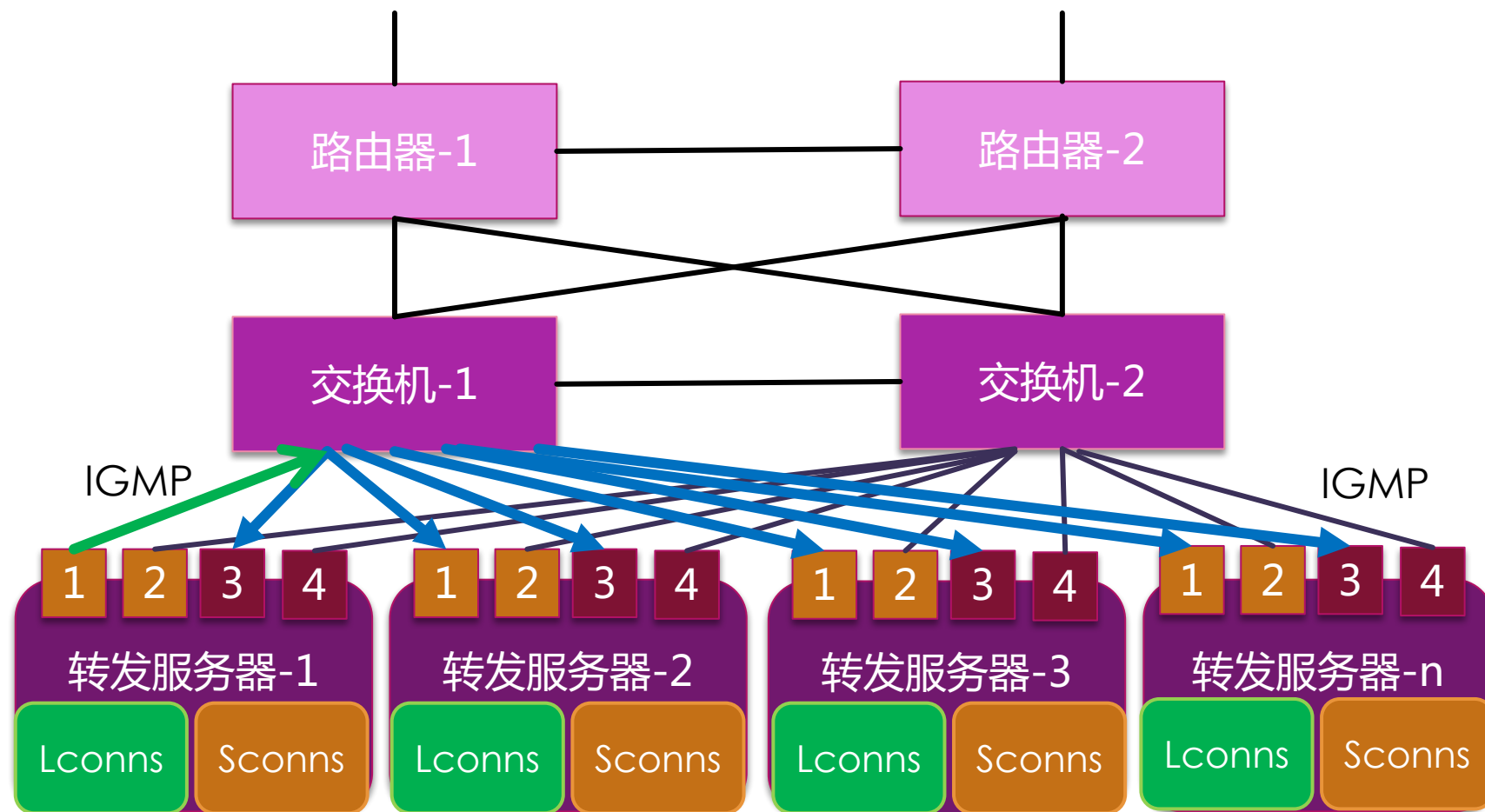
- 交易直接下跌
- 用户访问超时，异常



SLB容灾升级示意图

无感容灾与升级——组播会话同步

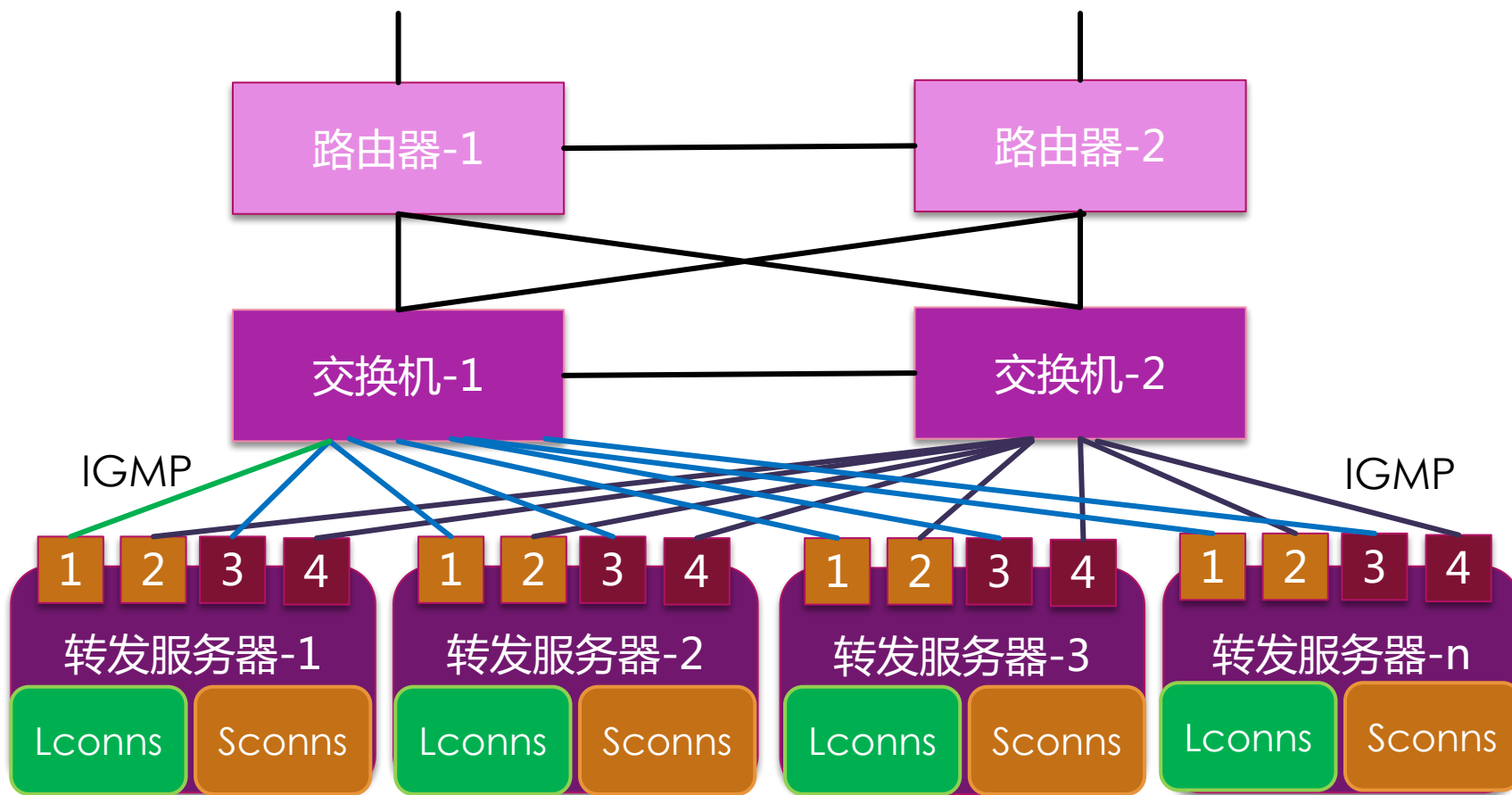
- 同一组转发服务器加入同一个组播组
- 利用组播解决机器的上线与下线
- 高性能：发送一份，组播复制多份
- 转发面简单、高效



SLB组播会话同步示意图

无感容灾与升级——UDP组播的可靠性保障

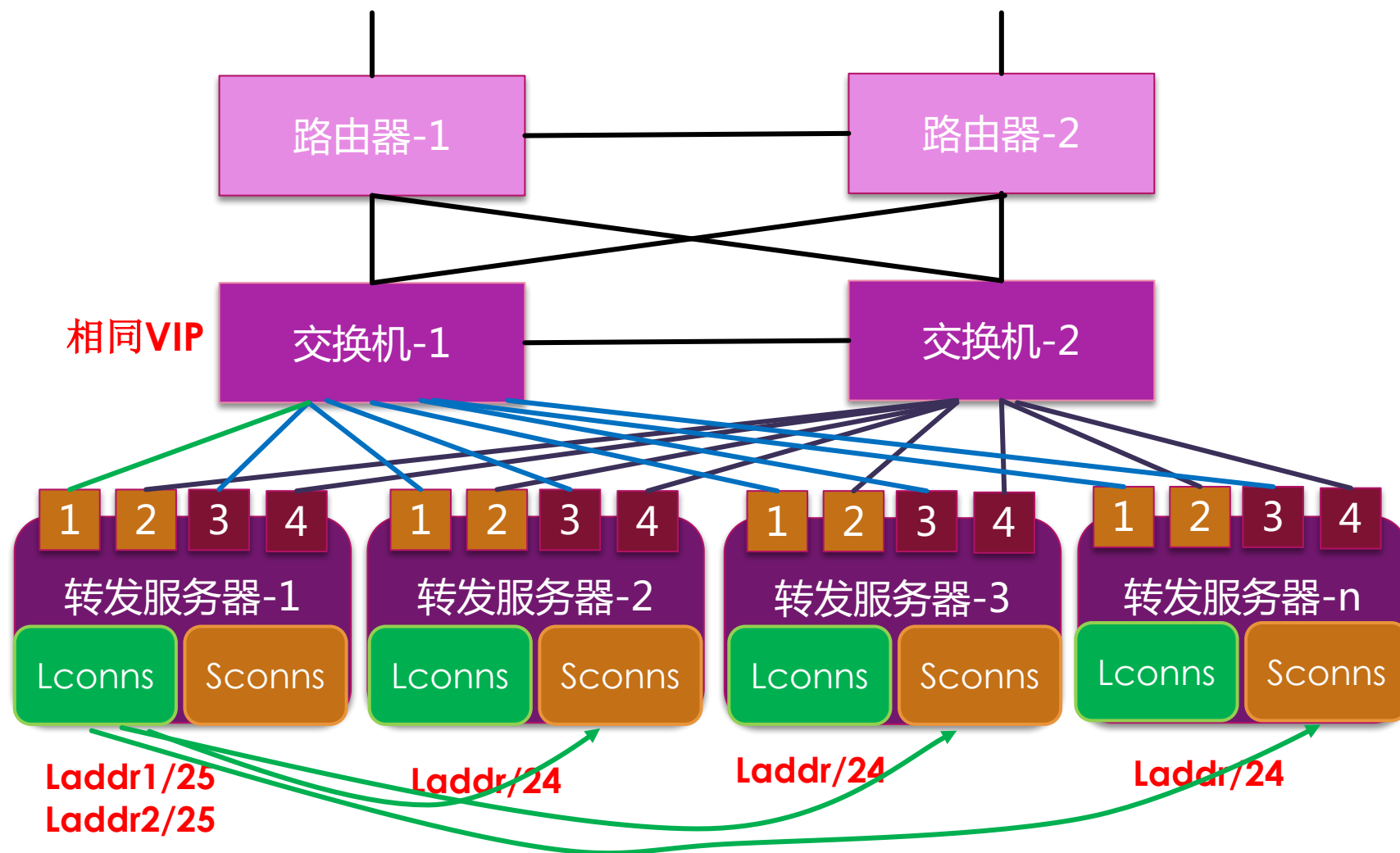
- 拓扑简单可靠，不易丢包。
- 会话同步的双保险
- 新建连接同步与定时全量同步
- RR发送同步包
- 同步包Round Robin的发送，解决链路异常，容灾场景问题
- 定时监控
- 定时发送组播包探测报文，保障组播功能正常



SLB组播会话同步示意图

无感容灾与升级——会话同步功能的路由策略

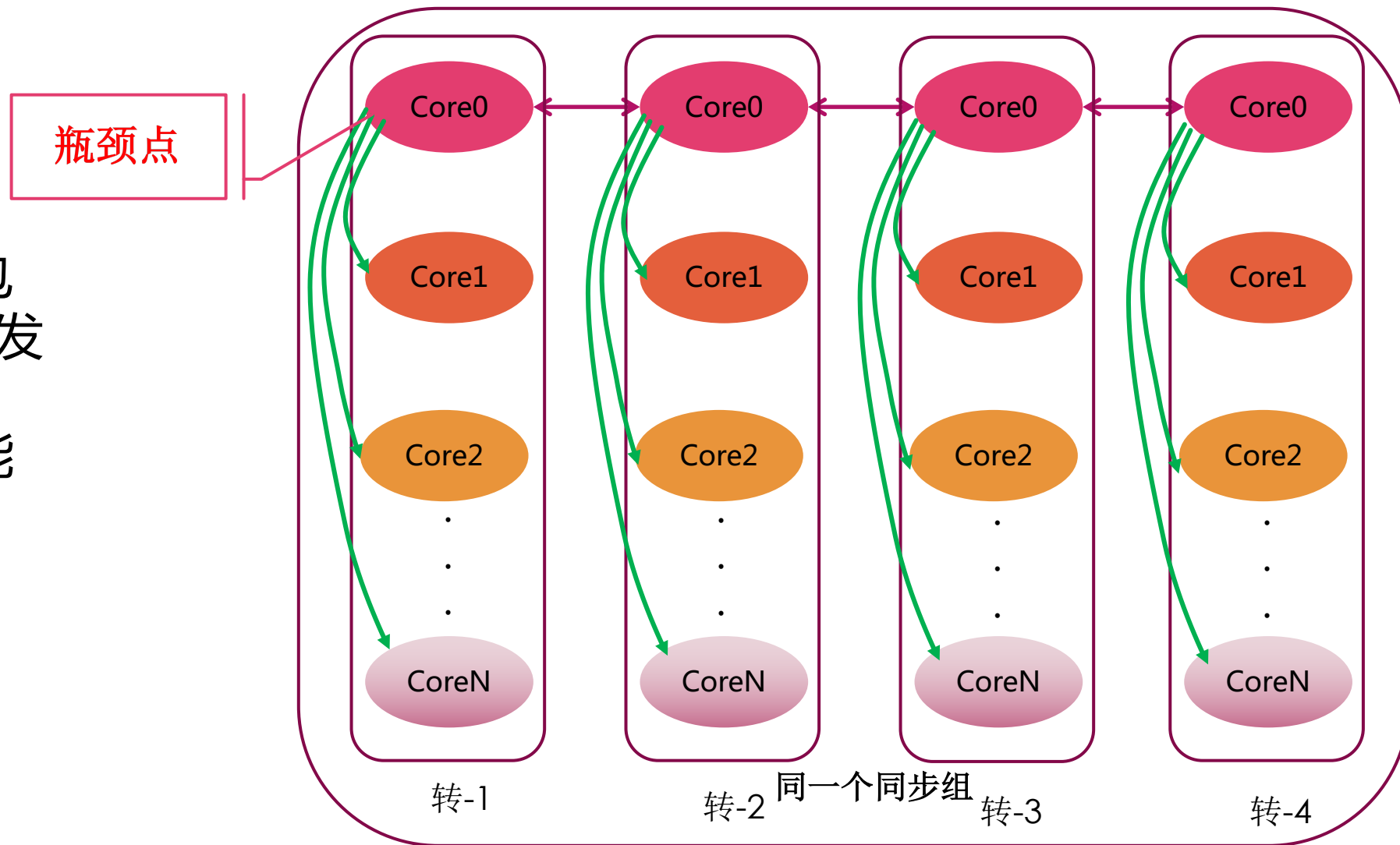
- VIP多台服务器ECMP切换时从client -> slb 的流量进行ECMP重选路到不同的服务器
- Laddr使用大小路由发布，切换时从RS->slb 的流量进行ECMP重选路到不同的服务器
- 切换时，流量可能进出同CORE不同台



SLB路由策略示意图

无感容灾与升级——会话同步遇到的性能问题

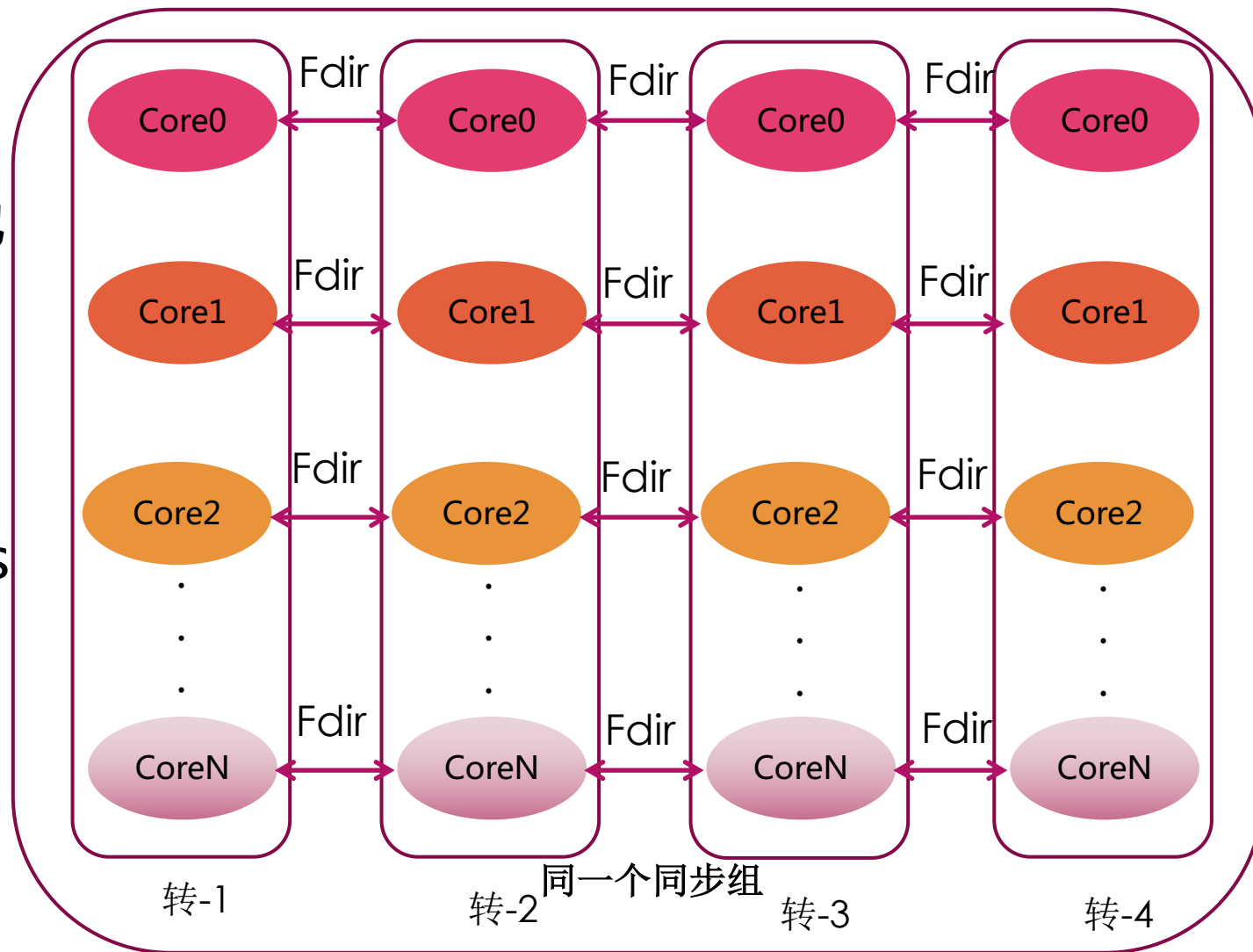
- 单机的同步数据包由CPU0接收与分发
- 单CPU处理成性能瓶颈



SLB单CPU同步示意图

无感容灾与升级——并发同步解决性能问题

- 软硬件结合的方案 (FDIR)
- 相同编号的CPU有同一个组播地址。
- 不同的CPU有不同的组播地址
- 利用FDIR功能，实现PER CPU session同步
- 消除单核瓶颈
- 提升同步性能



SLB并发同步示意图

“

谢谢

”