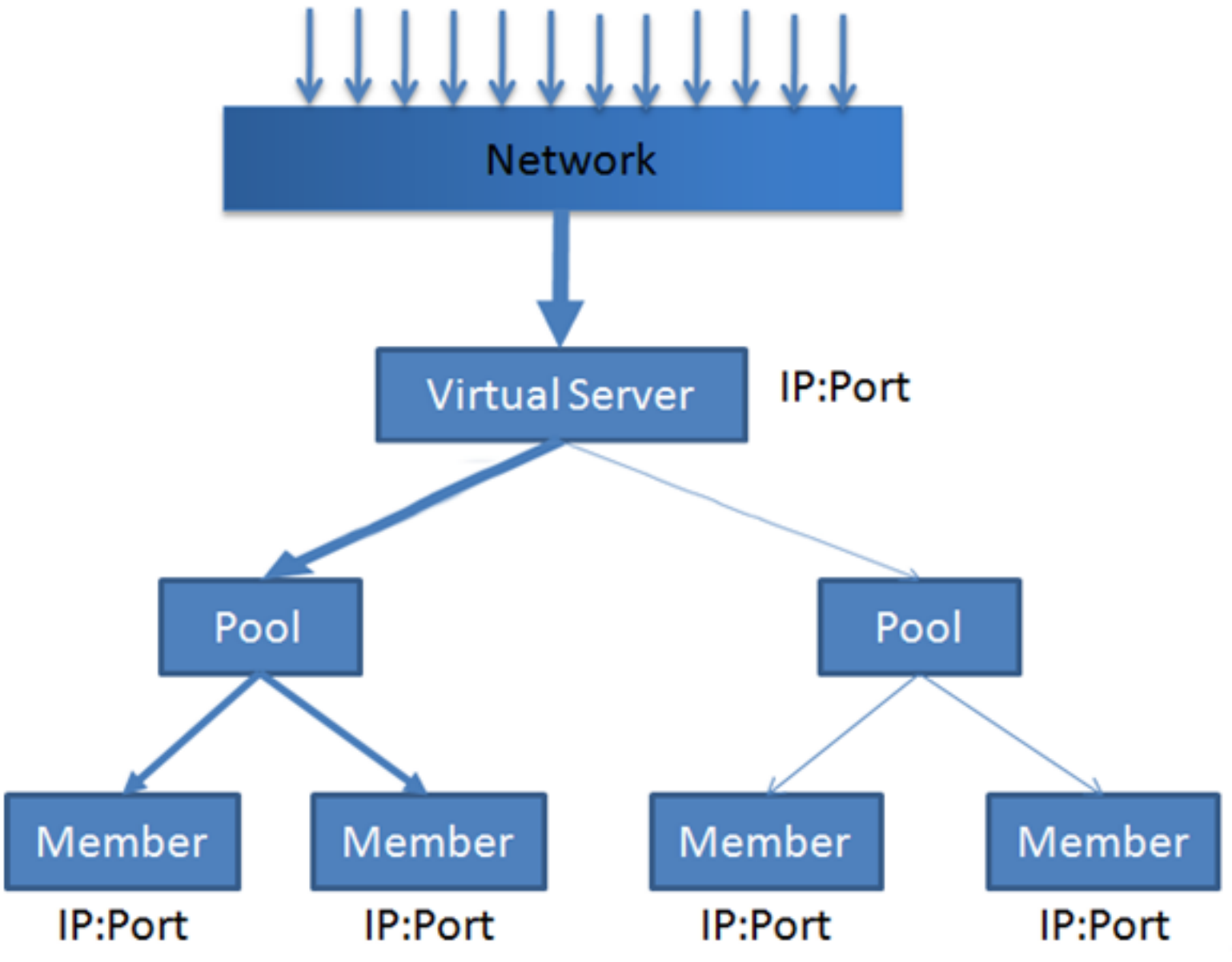
LVS 生产环境架构详解

在这篇文章中：

一、LVS 技术介绍

1.1 工作原理

LVS 是 Linux Virtual Server 的简写，意即 Linux 虚拟服务器，是一个开源的负载均衡流量调度器。LVS 集群采用 IP 负载均衡技术和基于内容请求分发技术，将用户请求按照一定策略分发到后端的 Server 上，从而将一组服务器构成一个高性能的、高可用的虚拟服务器。在特定的场景下，整个服务器集群的结构对客户是透明的，而且无需修改客户端和服务器端的程序。



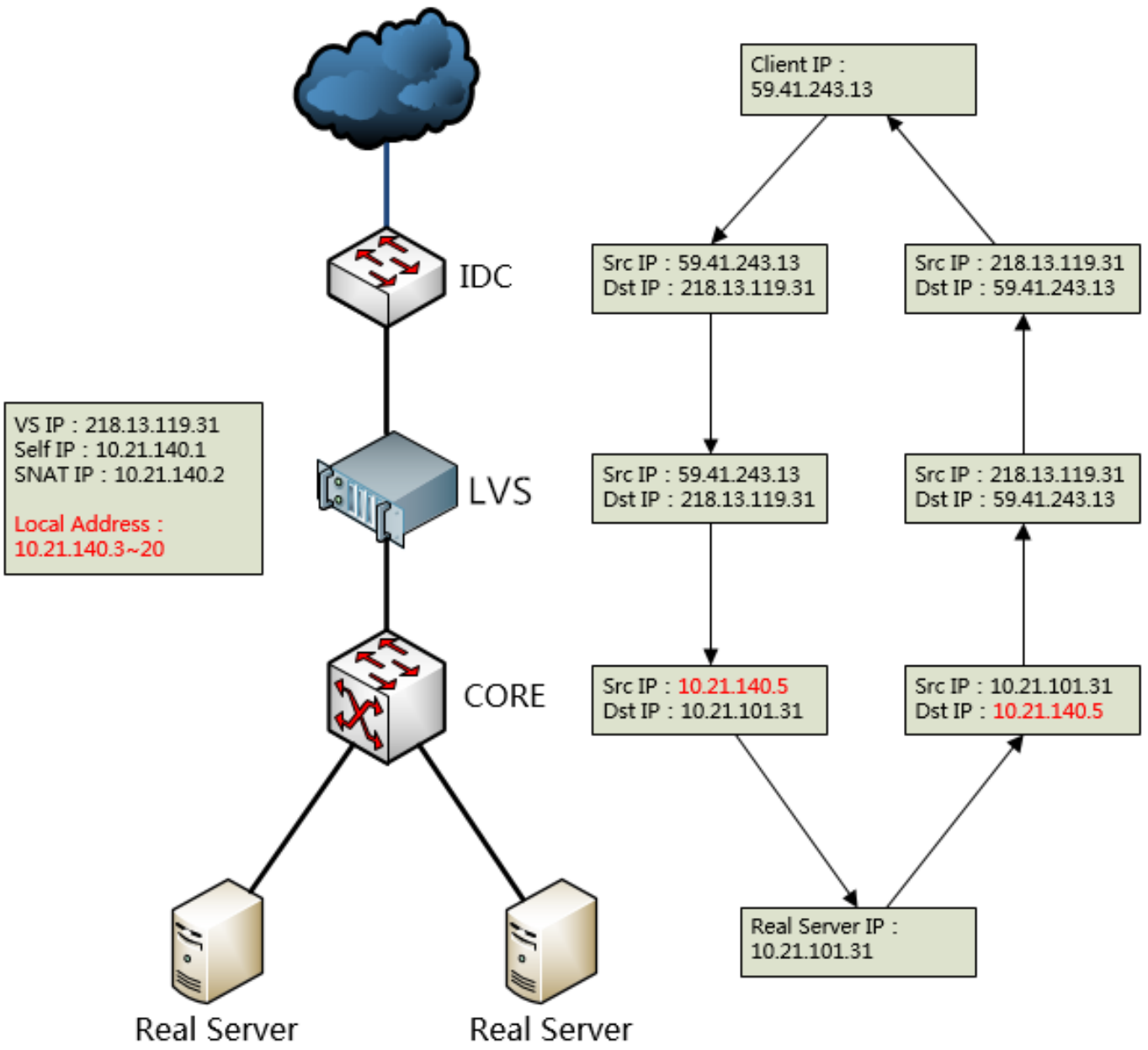
如图所示，前端调度器虚拟出 VS（Virtual Server）监听和接收请求，真正提供服务的是后端的 Member（亦称为 RealServer 或者 RS），数个 Member 组成一个 Pool，VS 的请求分发到 Pool 上，并在 Pool 当中的 Member 之间按一定策略分发轮询。LVS 与 F5 Big-IP LTM 不同之处在于，LVS 没有 Node 的

概念，也没有 Profile，并且只提供 4 层的负载均衡。

1.2 调度模式

1.2.1 FullNat

LVS 除了支持传统的 DR、TUN、NAT 模式之外， LVS 内核，支持一种新的负载均衡模式：Fullnat。Fullnat 区分于传统 NAT 模式的地方在于数据包经过调度器的时候，源目 IP 和端口都被转换了，后端 Member 看到的数据包来源是调度器的 Local Address。



采用 Fullnat 模式的优点是可以更加灵活的适应不同的网络环境，对网络架构的改动需求最小，而且支持跨集群调度。缺点在于如果后端服务器需要获知真实的客户端 IP 地址，需要打内核补丁。当调度器与后端服务器建立 tcp 连接的时候，三次握手包的包头中会携带写入真实客户端 IP 地址的 address of tcp option 信息，只有后端服务器打了补丁才能获取到客户端 ip。

1.2.2 NAT

在网络环境许可的情况下，采用 NAT 模式可以带来 10%左右的性能提升，并且不需要配置 Local Address，由于 Local Address 需要与内网互联网段（LVS 与内网核心互联地址段）在同一子网内，使用 NAT 模式能够减少 IP 地址占用。

1.2.3 内网负载均衡

LVS 支持内网负载均衡功能。在实际应用场景中，可能会遇到在内网当中的服务器也需要负载均衡的情况，比如中间件层调用存储、前端调用数据库等等。内网负载均衡需要网络环境支持，在内网核心上将地址段路由下一跳指向 LVS 的 Self IP。需要注意的是，内网负载均衡必须是 Fullnat 模式，内网负载均衡不支持 NAT 模式。

1.3 SNAT

传统的 LVS 不具备 SNAT 功能，当使用 NAT 模式进行负载均衡的时候，后端服务器正常的公网访问需求仍然需要通过额外的设备解决。LVS 可以合并 SNAT，提供了 All-in-one 模式，使用中需要将内网核心的默认网关指向 LVS 的 SNAT IP。所有数据包在流入 LVS 的时候都会在 session 表中检查一遍，不存在会话的数据包根据策略路由从 SNAT IPool 池中选取 IP 作为源地址转发出去.

1.4 会话保持

LBG 的 tcp ESTABLISHED 状态默认 90s 超时，90 秒内没有数据传输就断开连接。在 4 层负载均衡场景下，LVS 默认启用会话保持功能，同一个 client ip 转发到同一个后端服务器，600 秒无数据传输后过期，不同的 client ip 或者会话过期则按调度算法重新分派。在 7 层负载均衡场景下，默认不启用会话保持。可以开启插入 cookie 会话保持功能，默认过期时间 3600 秒。 1.5 7 层负载均衡 可以使用 nginx 或 Tengine 提供 7 层负载均衡功能。

二、LVS 组网模式

2.1 并行旁挂

在已有成熟的网络架构中部署 LVS，采用并行旁挂的方式不需要网络有太多的变动，而且相比 DR 模式需要在 VLAN 配置 IP 地址不同，Fullnat 模式的只需要单独的一个 VLAN，就能够进行负载均衡。并行旁挂模式分为单臂和直挂两种。单臂的情况下，对链路的压力很大，当流量高峰期容易导致链路跑满，因此一般采用直挂的方式，将 LVS 分别直挂在 IDC 出口和内网核心。所有需要负载均衡的数据流量经由 IDC 出口转发给 LVS 服务器，其它流量不受影响。

2.2 串行路由

串行组网架构是应用最为广泛的负载均衡组网架构。只有在串行组网架构当中，才可以使用 LVS 的 NAT 模式。使用 NAT 模式可以做到对后端服务器完全透明，并且调度性能和网络吞吐也比并行旁挂模式要好。

数据的访问流向均先经过 IDC 交换机，通过 IDC 交换机路由到 LVS 上，根据 LVS 配置的负载均衡策略对流量进行负载均衡，再经由内网核心交换机和机柜接入交换机到达相应的服务器，返回的数据流亦然。

串行组网架构整体网络结构比较单一整齐，业务数据流走向清晰可见，易于设计、部署实施，及后续的维护、管理，相关故障的排查。不过由于 LVS 位于出访流量的必经之路，除非有相应的路由策略，否则所有的对公网访问流量也需要经过 LVS，因此需要与 LVS 的 SNAT 配合使用。

三、LVS 高可用性

3.1 主播模式

在冗余方面，LVS 分别支持主备模式和集群模式。在主备模式下，LVS 可采用成熟的开源软件 Keepalived 实现冗余功能。在 LVS 主备方案实施当中，一台为主机正常提供服务，另外一台提供热备份。当主机离线时，备机会自动接管所有 VS，接替主机承担负载均衡的职责。 Keepalived 参照了 VRRP 协议实现故障切换。LVS 的 Self IP 与 F5 的 Self IP 在概念上并不相同。LVS 不需要像 F5 那样为每台设备的每个网段配置 Self IP，再配置一个不同于 Self IP 的 Float IP 对外提供服务。在 LVS 中，Self IP 直接对外提供服务，Fullnat 模式下还拥有不会随主备切换的 Local Addres。在正常情况下，主机对外宣告 Self IP，备机没有配置 IP，保持静默。如果主机发生故障，备机会

在一秒钟内检测到，产生故障切换事件，通过发送免费 ARP 宣告自己拥有 Self IP，引发流量切换。

Keepalived 可以指定某个网络接口运行 VRRP 实例，为了避免 VRRP 影响现有网络，可以采用单独的心跳线传输 Failover 流量。备机通过监听 VRRP 通告确认主机是否存活，如果主备机因为意外同时 ACTIVE，会导致严重的网络故障，并且需要人为干预才能恢复。为了避免单根链路故障而导致的意外故障切换，建议心跳线采用两根链路捆绑，可以大大降低故障几率。LVS 支持人为的进行主备机倒换，但是并不具备 F5 的会话镜像功能，因此在主备机倒换和故障切换之后，所有会话的连接性都会丢失。

3.2 集群模式

集群模式采用了 LVS ospf 方案，利用开源的软路由软件 quagga，对 IDC 接入交换机宣告 VIP 的主机路由信息，通过 OSPF 等价路由的特性可以提供最多八台 LVS All-active 的集群服务。在集群模式下，LVS 可以横向扩展，自由伸缩，但是会增加网络的复杂性。

单 OSPF 区域，LVS 只能使用 DR 模式，与内网核心使用 TRUNK 连接，在每一个服务 VLAN 宣告 IP，并且所有后端服务器必须配置 LVS 为网关，对网络的要求和后端服务器的变动需求非常大。如果要使用 NAT 模式，需要 LVS 与内网核心也建立 OSPF 邻居关系，将 SNAT IP 同时宣告给内网核心。受限于 ospf 调度算法，集群模式有可能无法提供无感知的伸缩特性。如果三层设备不支持 ospf 调度一致性 hash，那么当某台 LVS 离线的时候，所有长连接都会丢失。目前只有 Cisco 设备支持一致性 Hash 算法。

3.3 监控检查

Keepalived 可以对 Member（RS 后端服务器）做健康检查。当某个 RS 服务不可用，Keepalived 会自动将其从 Pool 中剔除。LVS 目前支持的健康检查方式有 TCP、HTTP、SSL。

四、LVS 网卡配置

LVS 系统拥有以下几种接口类型，在实施设计阶段要予以考虑。业务接口可根据网络架构和应用场景行衡量。

端口名称	端口用途	端口类型	端口数量
Failover	主备故障切换心跳线	电口	2
Interal	内网业务流量接	电口\光口	1~8
Ext	外网业务流量接口	电口\光口	1~8

业务流量接口采用多千兆电口捆绑的方式，需要注意对端三层设备的链路聚合负载均衡模式的选择，以防聚合中单根链路跑满的情况发生。心跳线的作用是备机探测主机是否存活的关键，一旦这条链路丢失，LVS 系统就会出现双 Active 的情况，所以使用两根心跳线捆绑的方式可以让出现心跳线物理损坏的几率大大降低。