京东基于DPDK技术的高性能四层负载均衡器SKYLB

作者：京东商城基础平台部

已授权运维帮订阅号发布

开源版本近期功能完善后会告知大家，敬请关注运维帮

摘要

随着京东业务的高速增长，作为应用入口的负载均衡，大流量大并发带来的挑战越来越严峻。本文主要介绍了京东商城设计和实践的一套高可靠，高性能的负载均衡器，我们命名为SKYLB。是一个使用intel DPDK报文转发库，实现运行在通用X86服务器上自研的分布式负载均衡服务。配合网络路由器的OSPF或者BGP协议，组成承担京东数据中心核心四层负载均衡的集群。最大限度的发挥普通X86服务器硬件资源的性能，实现一套适合于京东商城业务的低成本，分布式，高性能，可扩展的智能负载均衡系统。

介绍

京东商城目前是国内最大的电商企业。京东的机房内部的流量爆炸式快速的增长。早在2016年初京东商城已经将所有的业务系统全部迁移到容器平台JDOS，线上百万+容器实例稳定运行。大流量的负载均衡的分配显得至关重要，也是京东商城新一代软件定义数据中心的关键基础服务设施之一。

负载均衡器一般介于网络上的路由器与后端服务器之间，负责将每个数据包通过一定的服务匹配，将其转发到后端服务器节点。充分考虑到京东商城数据中心全容器及全三层BGP组网的模型。以及基于DPDK的几乎达到网卡限速的性能，我们在设计负载均衡时，仅考虑实现了FULLNAT模式，即出向

和入向的流量均通过负载均衡器，基本数据流程图如下图1所示：

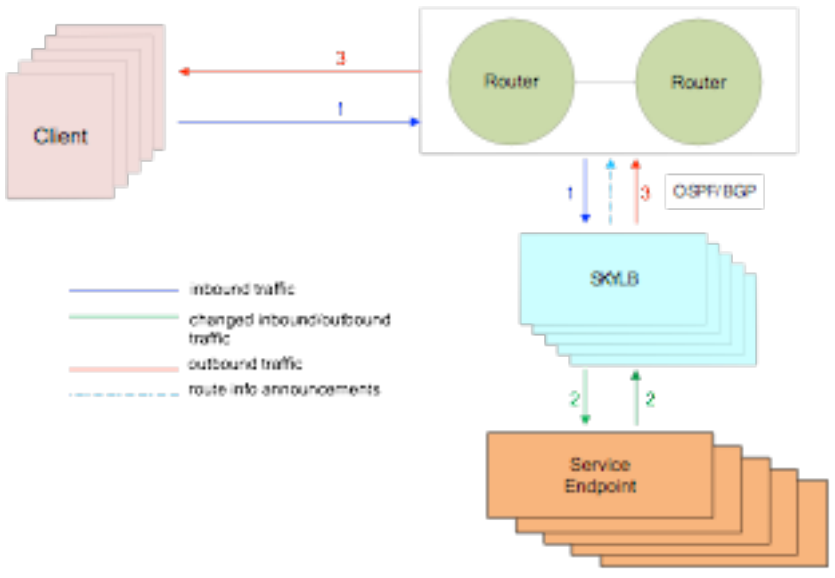

图1 负载均衡流程图

一般根据业务及流量的规模的不同阶段来选择使用不同的负载均衡，通常我们在负载均衡的选择上大致有以下两个方向：1)硬件负载均衡，如F5。CitrixNetscaler等；2)软件负载均衡，如基于LVS，Haproxy，Nginx等开源软件来实现的负载均衡。对于上述两种负载均衡的选择，各有优缺点，如下：

1） 硬件负载均衡

可扩展性受限，无法跟上业务流量增长的需求。以及如618、双十一大促等瞬间流量高峰。

虽然可以成对部署避免单点故障，但是一般只能提供1+1冗余。

缺乏互联网快速迭代的灵活性，升级成本昂贵。

一般只能根据网络的情况来设定负载均衡，无关乎实际系统和硬件的情况。

成本较高。

2） 基于开源软件的负载均衡

可以根据实际系统和应用的状态来合理的负载，迭代、扩容和部署相对方便。

单台负载均衡性能相对较差，需要用集群来支撑负载均衡的整体性能。

性价比较低。

我们的目标:

1. 设计实现一套高可靠、高性能、易维护及性价比高的L4负载均衡系统,。
2. 基于通用X86_64服务器框架，以及支持DPDK网卡硬件，易开发和移植。
3. 方便部署、删除和维护，集成到京东软件定义数据中心（JDOS2.0）系统，作为京东下一代软件定义数据中心的基础组件。
4. 负载均衡下的服务器基于系统应用负载流量均摊，负载均衡器提供N+1冗余，借助OSPF/BGP控制负载均衡器的流量负载。
5. 基于系统应用级别的探活，自动故障检测及流量快速恢复。

本文主要介绍了SKYLB一种基于DPDK平台实现的快速可靠的软件网络负载均衡系统。不仅可以快速的横向扩展，还可以最大限度的提升负载均衡单个NIC的处理转发速度，来实现L4的负载均衡。借助DPDK的优势，如便利的多核编程框架、巨页内存管理、无锁队列、无中断poll-mode 网卡驱动、CPU亲和性等等来实现快速的网卡收发及处理报文，后续考虑TCP/IP 用户态协议实现和优化，进而实现L7负载均衡。

系统概览

工作场景

SKYLB部署在京东容器集群JDOS的前端，对于一个应用集群，发布一个或多个VIP到SKYLB服务上，当客户端需要访问应用节资源URL，首先通过域名访问JD智能分布式DNS服务（SKYDNS详见 https://github.com/ipdcode/skydns）, SkyDns会智能返回当前最近且状态正常且负载正常的VIP服务的IP,客户端就会向VIP去请求连接。

SKYLB节点上运行了一个路由发布服务agent，我们使用该agent与开启

OSPF/BGP的路由器做路由交互，当SKYLB的上层路由器接收到请求VIP的数据包时，路由器通过（OSPF/BGP）的等价多路径转发协议选择一个可以使用的发布该VIP的SKYLB节点，将报文转发给一个SKYLB节点。通过这种策略来实现VIP的发布和横向容量负载能力扩展。

当报文通过上述步骤到达SKYLB负载均衡后，通过常用的负载均衡策略(round robin，一致性hash ，最小连接数)，将数据包送到相应的后端容器服务。

系统架构

JingdongDatacenter Operating System(JDOS) 是基于JDOS提供物理机/虚拟机/容器的统一管理系统、配备灵活的网络互连、可靠的分布式共享存储，实现软件定义数据中心的计算资源统一管理和集群管理。通过使用JDOS，可以直接迅速得到任意需要的计算、存储、网络、安全等方面的资源和能力。SKYLB作为整个JDOS系统的一个重要组成部分，负责提供集群的L4负载均衡能力，通过restful API等接口与JDOS系统交互。用户可以通过统一调度管理平台便捷的创建、删除、迁移负载均衡系统。同时多个应用服务进行流量分发的负载均衡服务，从而扩展应用系统对外的服务能力，并且通过消除单点故障提高应用系统的可用性。

系统的基本架构如下图2所示，每一个集群配备一组可容灾的负载均衡控制中心,主要通过restful api接口负责与JDOS调度中心交互，接收vip的配置请求。同时我们在每一个SKYLB的节点上运行一个代理进程，该代理进程通过gRPC与控制中心连接。接收控制中心下达的创建及删除vip,后端server endpoint服务等一系列指令，通过load balancer 提供的命令行执行相应的指令。接收load balancer 关于流量及报文的监控信息，对于流量及监控进行告警，并且通知控制中心和调度中心进行扩容等操作。

代理进程同时还负责后端服务 server endpoint基于服务可用性的健康检查，及时根据后端服务的状态通过命令行进行添加和删除的操作。

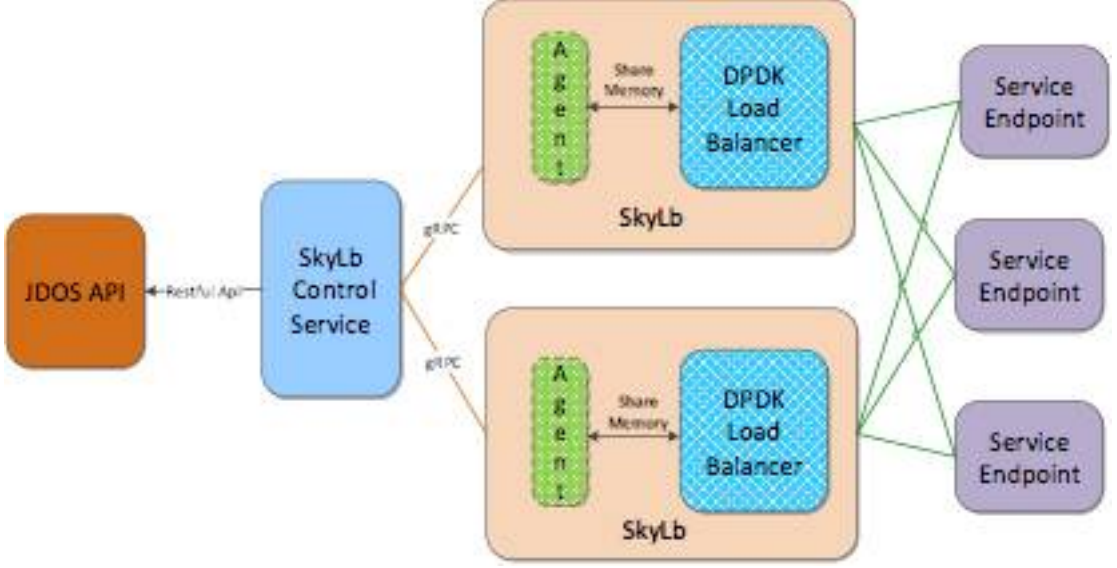

图2 系统架构图

优势

1）扩展性

支持动态添加和删除后端服务的容器，实现无缝的伸缩；在伸，缩过程中，对相关调用和访问者无影响。

2）高可用性

提供多活负载均衡，有多个VIP，它们对应一个域名，自研DNS服务SKYDNS会根据请求的客户端IP智能解析可用的VIP，返回给用户，从而实现更高的可用性；即使一个VIP不可用，也不会影响业务系统对外提供服务。同时借助OSPF/BGP等协议实现负载均衡的横向扩充

3）服务能力自动可调

SKYLB根据VIP实际接收流量的负载需要调整负载均衡的服务能力，比如流量、连接数的控制等指标。

功能特点

1）协议支持

负载均衡支持包含TCP、UDP协议的四层负载均衡，配备健全的链路跟踪机制，以及多种调度策略，用户可以根据服务的需要创建合适自己的负载均衡。

2）高可用性

支持容器的健康检查，除传统的IP+Port，并加入对URL检查，保证应用可用性： 健康检查频率可自定义；一旦探测到异常，则不会将流量再分配到这些异常实例，保证应用可用性。

3）集群部署，多层次容错机制： 负载均衡采用集群部署，支持热升级，机器故障和集群维护对用户完全透明，结合DNS使用还可支持全局负载均衡。

4）灵活性

支持多种流量调度算法，使得流量分配更均匀： 负载均衡支持加权轮询和最小连接数这两种调度算法，可根据自身需求选择相应的算法来分配用户访问流量，并支持设置后端容器权重，使得流量调度更均匀，提升负载均衡能力。

支持会话保持，满足用户个性化需求： 负载均衡通过IP地址实现会话保持，可将一定时间内来自同一用户的访问请求，转发到同一个后端容器上进行处理，从而实现用户访问的连续性。

5）易用性

提供多种管理途径，轻松操纵负载均衡： 用户可通过控制台轻松实现负载均衡器的配置、释放等功能。后续会开放标准的API或SDK提供给用户，从而自己开发对负载均衡的控制管理。

系统设计及技术实现

负载均衡模式选择

常用的负载均衡模式有DR，NAT，TUNNEL，FULLNAT。每种模式都有自己的优势和使用场景，业内对每种模式的分析比较文档很多，不再介绍。由

于JDOS容器网络需要VLAN隔离，而FULLNAT刚好支持LB和RS跨VLAN通信，结合我们自身容器集群的需求，我们在实现SKYLB时主要考虑支持FULLNAT模式。图3是SKYLB的FULLNAT负载均衡模式中数据包的流向图。SKYLB位于客户端和后端服务之间，对于客户端的请求报文，将目的地址替换成后端服务的地址，源地址替换成SKYLB的本地地址，对于后端服务的响应报文，将目的地址替换成客户端地址，源地址替换成SKYLB的VIP地址。

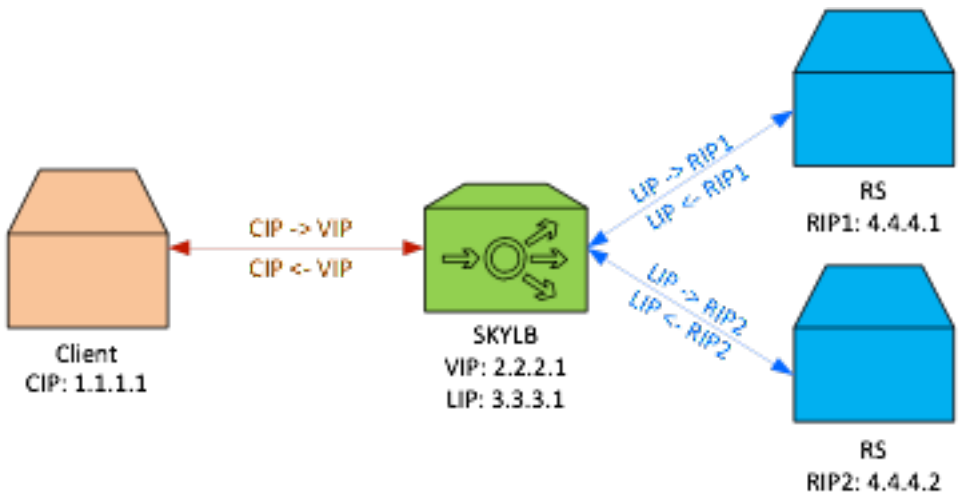

图3 FULLNAT模式下SKYLB的数据包流向图

借助DPDK实现高速转发

Data Plane DevelopmentKit（DPDK）：是运行在Linux 用户态，实现X86通用平台网络报文快速处理的库和驱动的集合，如下图4所示，其主要特点：

- 多核编程框架及CPU亲和性

每个NUMA节点有单独的CPU和本地内存

CPU访问本地内存速度比访问远端内存快，避免CPU访问远端内存

注意网卡挂载的NUMA节点巨页内存管理

- 巨页(HugePage)

普通内存页面大小4KB，巨页内存页面大小2MB/1GB

减少页表项数目，降低TLB miss

使用大页面比使用4K的页面性能提高10%~15%

零拷贝，报文数据及转发内存零拷贝。

- 无锁队列

使用无锁队列，入队出队无需阻塞等待锁资源

- poll-mode网卡驱动

DPDK网卡驱动完全抛弃中断模式，基于轮询方式收包

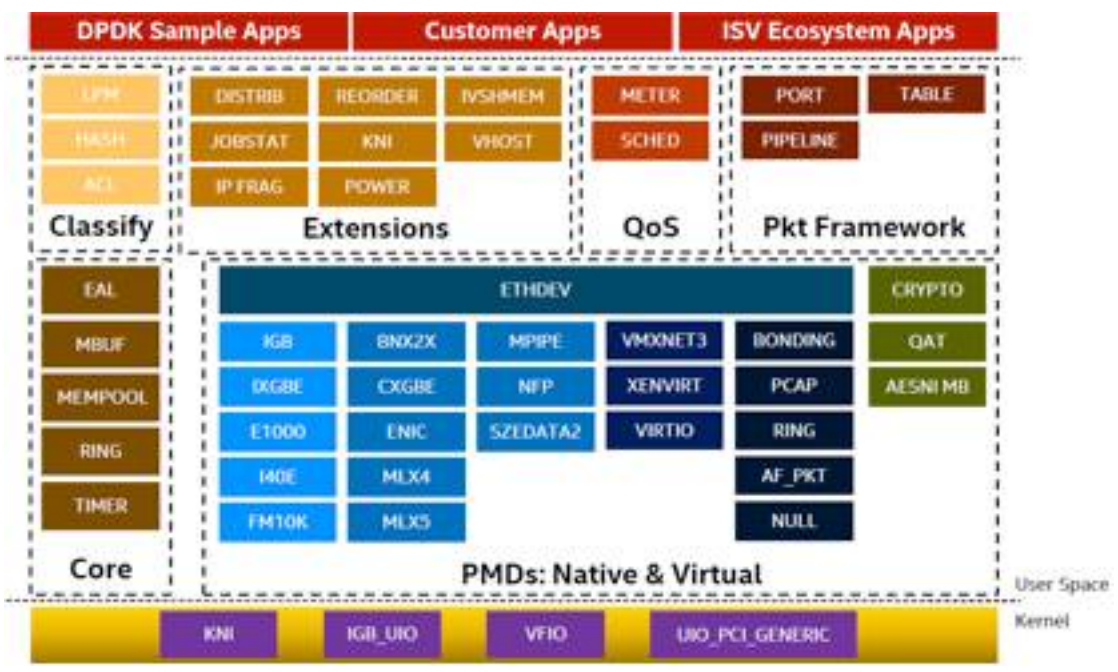

图4 DPDK相关模块

包处理架构实现

图5是SKYLB基于RTC数据包处理模型实现的架构。SKYLB选择一个核作为控制核，执行命令配置，与内核交换，ARP表维护等任务。其他的核作为工作核，每个工作核轮询网卡的一个RX队列，执行数据包处理任务。SKYLB利用网卡的RSS功能实现客户端请求报文的分流，利用网卡FDIR功能实现后端服务响应报文的分流。SKYLB对报文分流目的是要保证客户端的请求报文和其对应的服务端响应报文到达同一个工作核上。在这个目的达成的前提

下，SKYLB的业务实现会简单和高效很多。每个工作核维护一张session表，同于保存客户端和后端服务的连接信息。SKYLB不需要考虑对session表加锁，因为每个工作核的session表都是独立的。

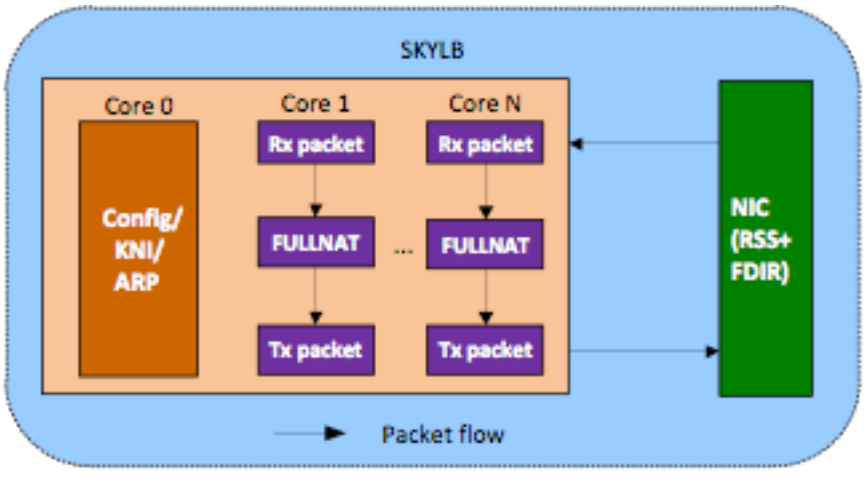

图5 SKYLB的RTC包处理模型框架图

我们设计SKYLB每个工作核独占至少一个LIP，并将LIP信息写入网卡配置。图6是网卡对IP报文分流的流程图。图中dst_ip即SKYLB为每个工作核分配的LIP。网卡对后端服务响应报文的目的地址LIP匹配成功，将报文送到绑定的RX队列，没有匹配的报文则可以认为是客户端的请求报文，按RSS流程分配RX队列。

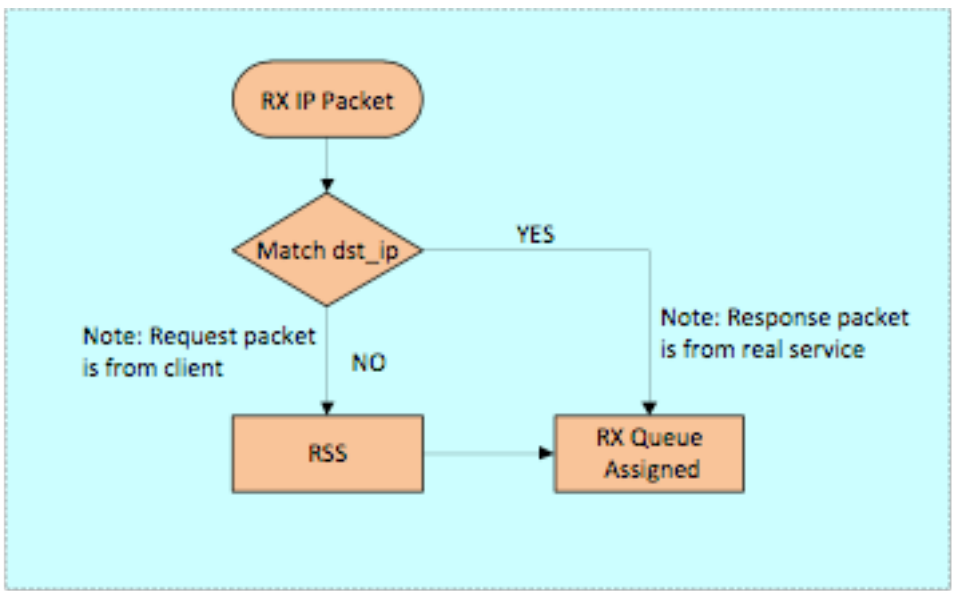

图6 网卡对IP报文分流过程图

SKYLB在启动过程中还会为每个物理口创建一个KNI接口，控制核负责轮询KNI接口。该接口主要用于外部程序quagga向路由器发布VIP信息，Agent检查后端服务健康状态。

SKYLB目前支持的负载均衡调度算法有一致性hash，round robin和最小连接数算法。

Session五元组，SKYLB采用五元组来实现会话的管理功能，如下图7 所示：

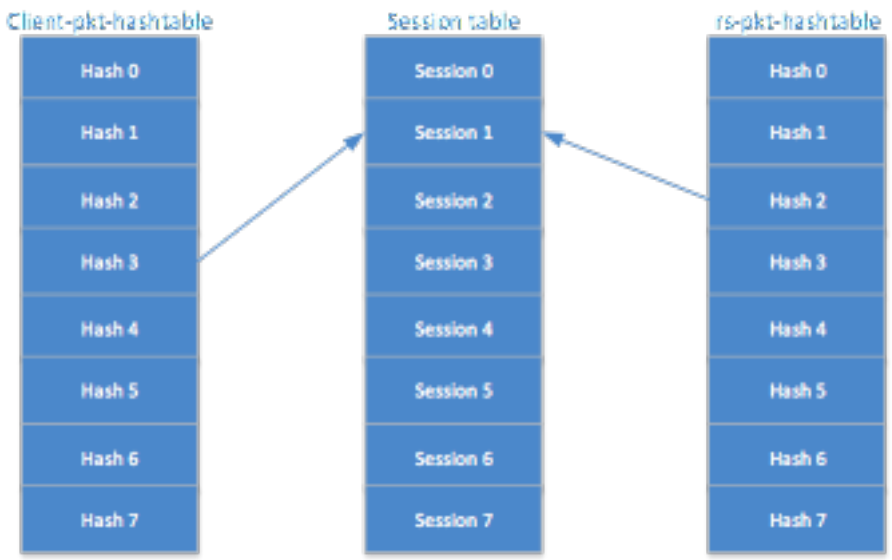

图7 SKYLB 五元组管理Session

负载均衡冗余设计

SKYLB使用BGP或者OSPF的模式组成集群，通过上述协议将数据包散列到集群中各个节点上，保证单台SKYLB故障或者恢复后能动态的将机器添加及删除。其冗余实现设计如下图6所示：

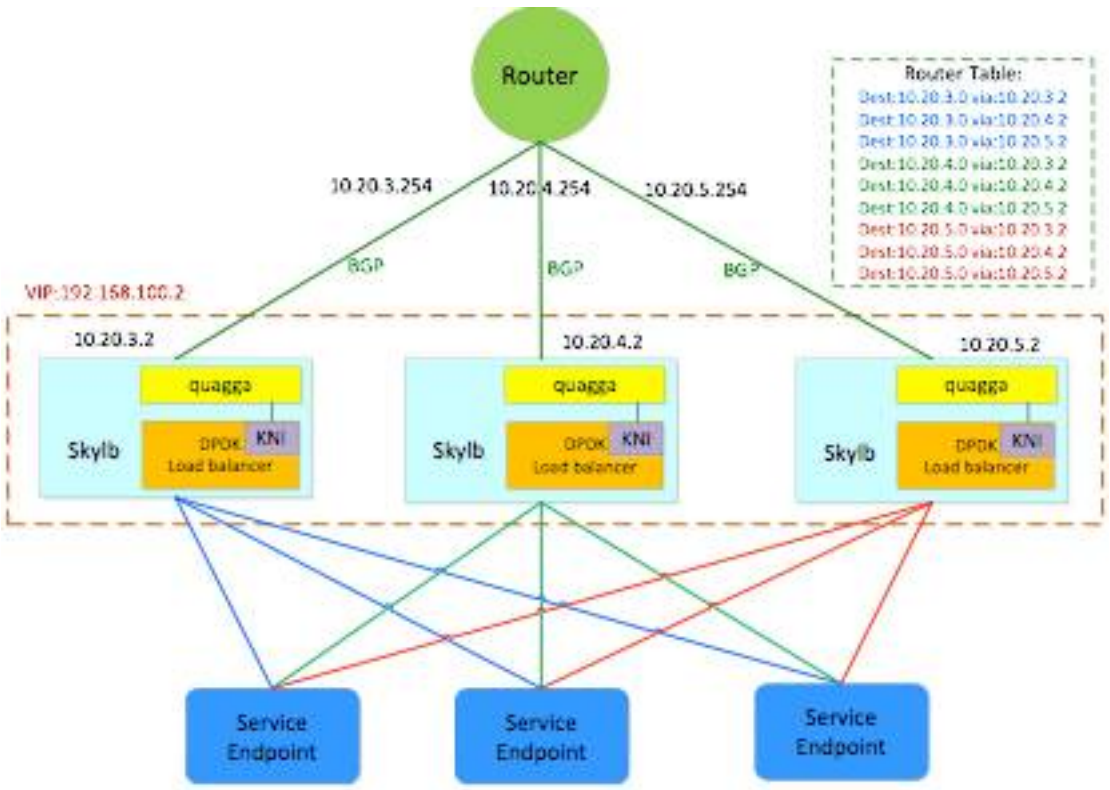

图8 SKYLB的冗余设计

性能优化实践

良好的流程设计是性能提升的关键，这方面的流程涉及和业务相关，缺乏共性，因此不做详细阐述。主要介绍SKYLB性能优化过程中使用的其他优化方法。

恰当地使用rte_prefetch0()，可以减少cache-miss次数，避免当前函数成为性能热点。性能调优工具perf可以帮助我们分析应该在哪处代码使用预取。例如我们使用perf发现报文处理函数中有一个处代码是性能热点，该代码用于读取新报文的类型字段并判断，分析认为很可能是cache-misses造成的。在进入报文处理函数前使用rte_prefetch0()，有效避免该函数成为热点。

恰当地使用likely()和unlikely()，可以减少分支预测失败的次数。我们在SKYLB代码的一处分支语句中使用unlikely()优化，性能提升明显。分支预测优化点可以借助perf分析确定，也可以根据自己对代码执行流程的理解确定。

尽量减少锁的使用。SKYLB中配置信息不经常变化，我们没有单独为每个可能争用的资源加锁，而是只用一把DPDK提供的读写锁，每个读线程在加读锁后，处理一批报文，然后释放读锁。既简化流程，又减少了操作读写锁的开销(DPDK读写锁的开销并不是很大)。

性能数据

测试环境：

CPU：Intel(R) Xeon(R) CPU E5-2640 v3

NIC：intel 82599ES 10-GigabitSFI/SFP+ Network Connection

测试配置：

负载均衡模式：FULLNAT

调度算法：一致性hash

配置：CPU占用8核 内存占用4G

性能测试数据：

1）UDP发包，测试转发性能，我们使用了SKYDNS作为后端服务，客户端采用UDP请求DNS

线程数	Avg（ms）时延	TPS（笔/秒）	错误数	错误数
3000	1	4343070	0	0.00%

表1 SKYLB基于UDP的DNS性能测试数据

2）NGINX作为后端服务，压测HTTP性能。

配置：CPU占用8核 内存占用4G

线程数	Avg（ms）时延	TP99（ms）	TPS（笔/秒）	错误数	错误数
1100	0	2	2101170	0	0.00%

表2 SKYLB HTTP性能测试数据

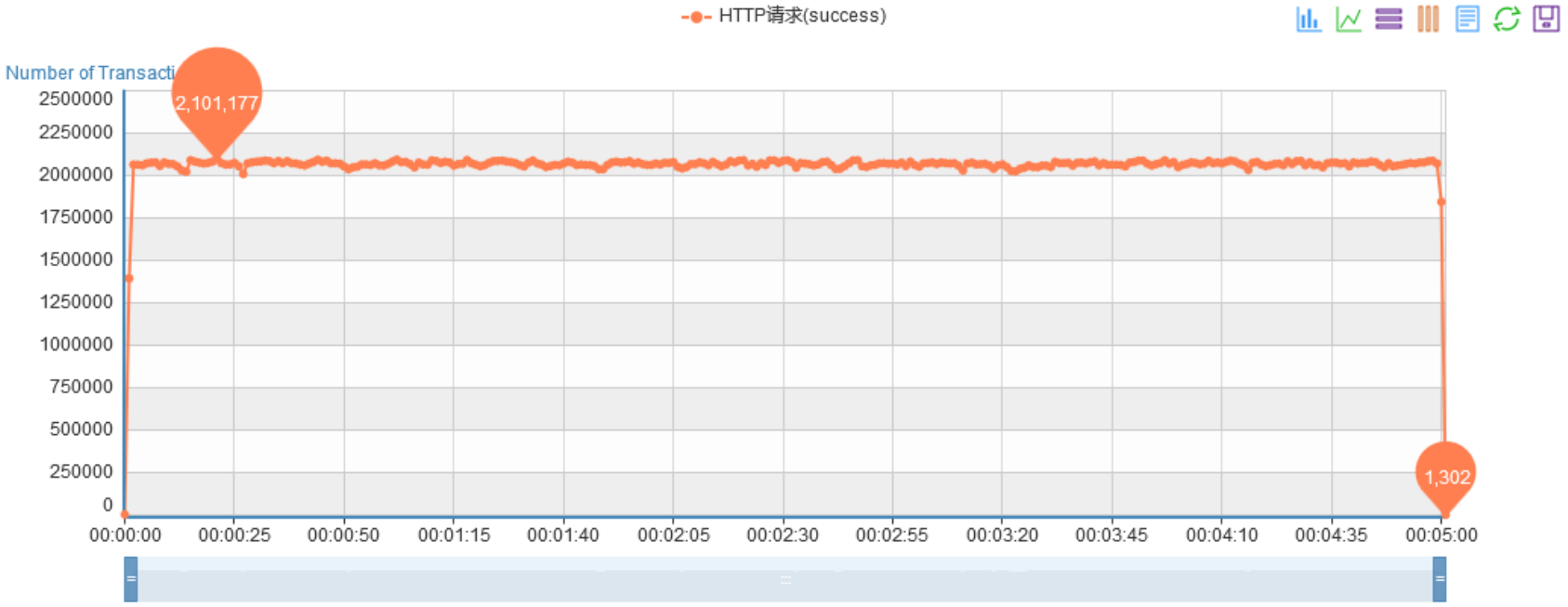

图9 SKYLB HTTP 性能测试指标图

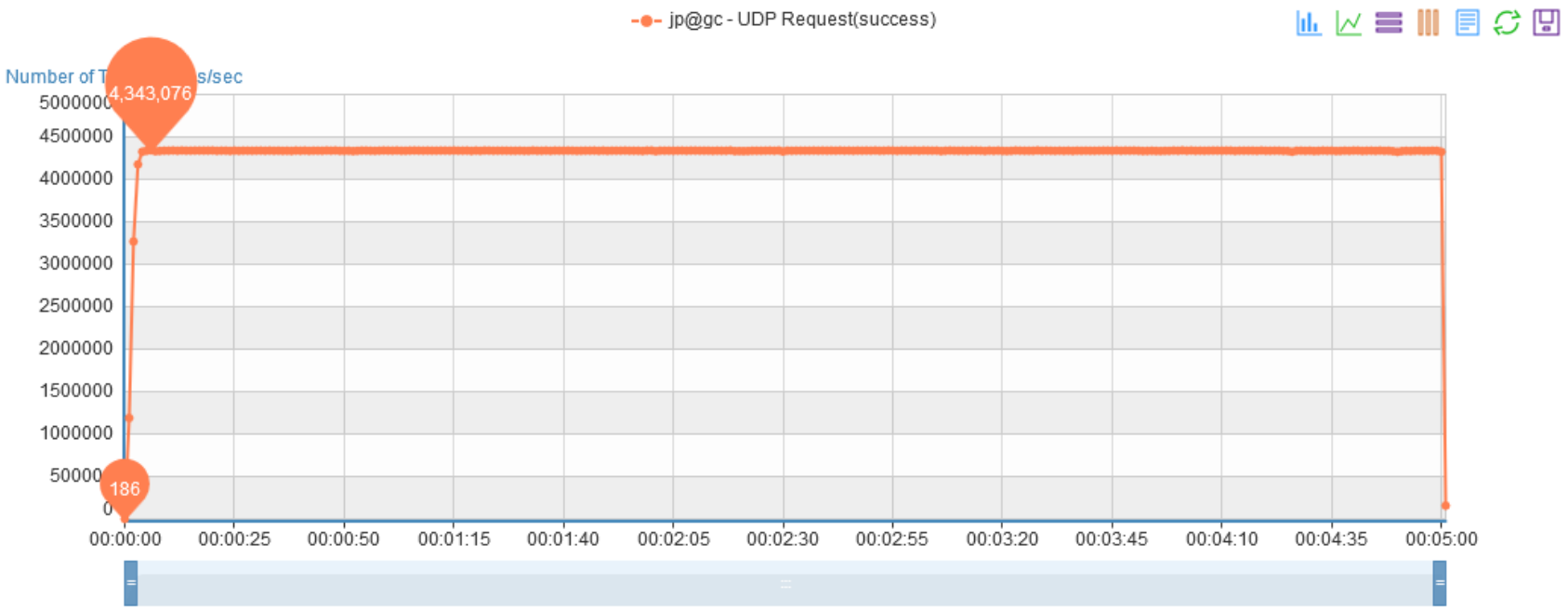

图9 SKYLBhttp 性能测试指标图

总结

本文主要介绍了SKYLB，一种的基于intel DPDK平台开发的负载均衡器。其接近网卡线速处理及转发能力，灵活的部署，多样的监控以及可靠的稳定性。基本上覆盖所有4层负载均衡的业务处理需求，配合集群管理以及调度，作为京东数据中心操作系统（JDOS）的一个重要的组成部分，在京东数据中心发挥至关重要的作用。

参考资料

Google Maglev：A Fast and Reliable SoftwareNetwork Load Balancer。

DPDK：http://dpdk.org/doc/guides/prog_guide/

运维帮现在提供全套ZABBIX监控解决方案，包括：

搭建ZABBIX系统、优化ZABBIX系统、升级ZABBIX系统。细节包括优化报警，优化监控项，代写监控模板、代写低级发现，协助设置微信、邮件、短信报警等服务。

我们提供远程电话支持，现场技术支持等全方位的服务，我们已协助30多家中大型企业进行了ZABBIX升级改造及技术支持，如果您的企业需要一套完整的Zabbix监控解决方案，添加微信（yunweibang55）或扫描本文最下方的微信号码就可以联系到我们。

欢迎加入「运维帮地方群」，现在有北京地方群、上海地方群、深圳地方群、成都地方群、广州地方群、杭州地方群。入群请先加群秘书（长按识别下方二维码），加群秘书时请告知所在城市及公司。

主流云厂商都已和运维帮达成战略合作，不管是1台还是100台，都可以享受到价格优惠，请联系群秘书。

群秘书微信，扫描下方二维码