
# 爱奇艺DPDK网络优化实践

爱奇艺技术产品中心　余珂

爱奇艺 iQIY
悦 享 品 质

# Agenda

- 业务需求及问题
- DPDK 方案
- 4层LB（负载均衡）优化
- 7层LB（负载均衡）优化
- 虚拟化网络优化
- 未来的挑战

全球互联网技术大会 GLOBAL INTERNET TECHNOLOGY CONFERENCE
iQIYI 爱奇艺

# 面临问题

- 业务需求及问题
  - 网络高并发要求，如：千万级别的并发请求
  - 低延时用户体验
  - 低成本的服务器成本
  - 突发流量

全球互联网技术大会 GLOBAL INTERNET TECHNOLOGY CONFERENCE
爱奇艺 iQIYI

# DPDK方案

**1. What is DPDK**

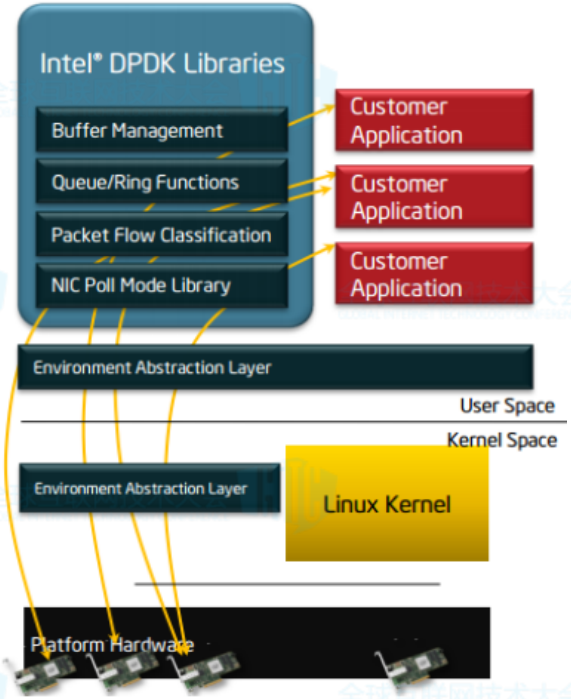
DPDK：数据面软件开发套件

开发数据包处理软件，增强数据面处理能力

**2. Usage**

基于X86平台开发：

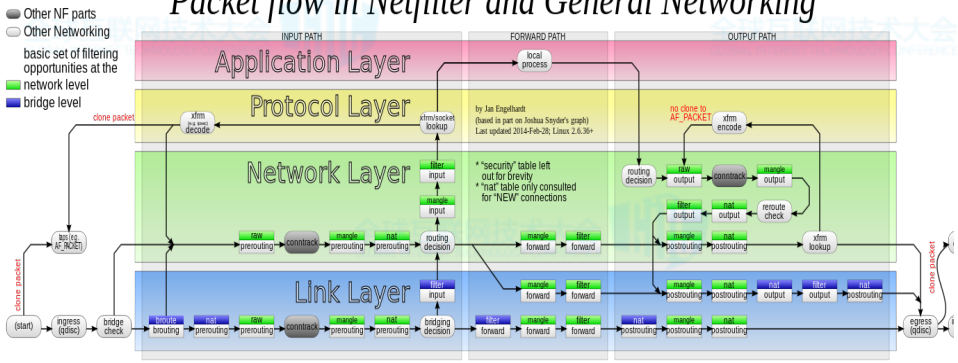
IPsec GW、FireWall、LoadBalancer、IPS、TCP/IP协议栈、虚拟交换机/路由器、tcpdump-like

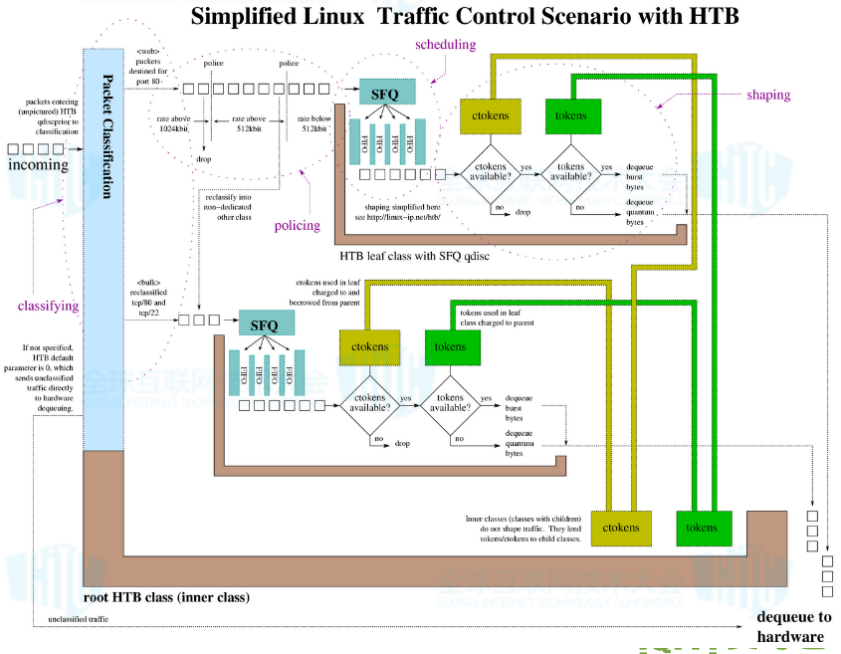



全球互联网技术大会 GLOBAL INTERNET TECHNOLOGY CONFERENCE

# 4层LB优化：为何LVS不够快

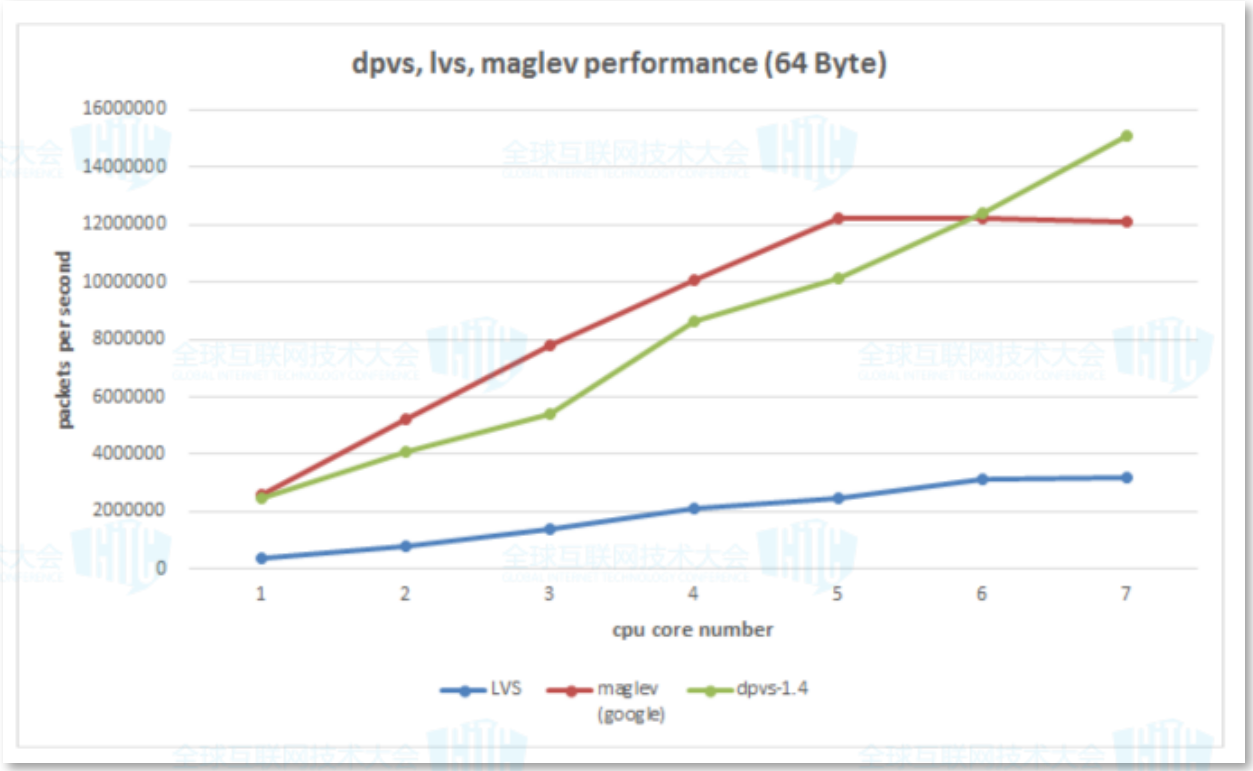
- Kernel是问题所在
- 资源共享及竞争
- IRQ风暴



# 4层LB优化：如何提高性能

- Kernel Bypass
- Share Nothing
- Zero Copy
- 轮询 vs. 中断
- 内存池
- NUMA aware
- Huge Page
- … …

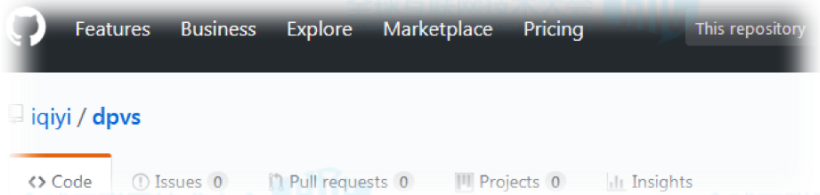


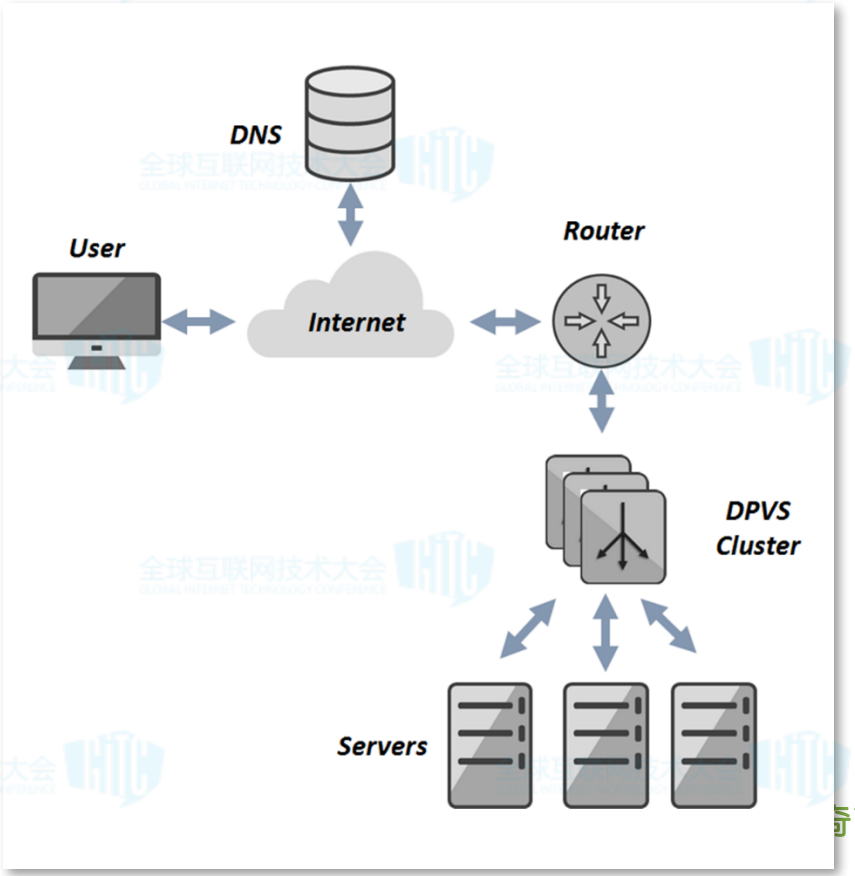

全球互联网技术大会 GLOBAL INTERNET TECHNOLOGY CONFERENCE
iQIYI 爱奇艺

# 4层LB优化：DPDK + LVS = DPVS

- 使用DPDK绕过Kernel
- 用户态IP协议栈
- 用户态实现LVS功能
- 返程数据亲和性等难点

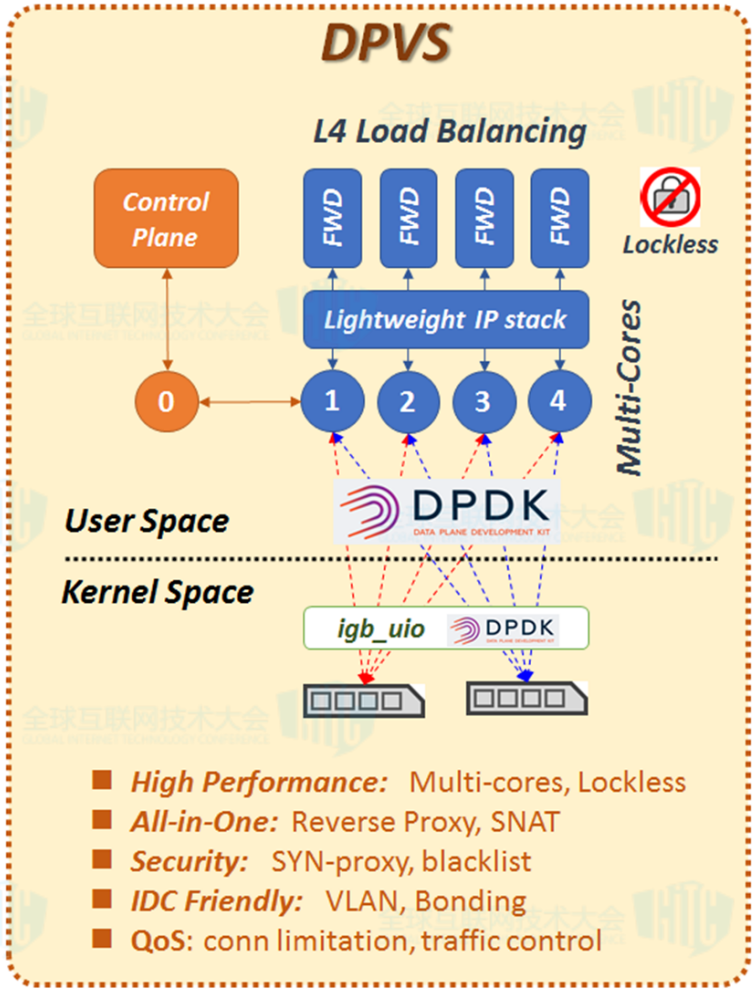
https://github.com/iqiyi/dpvs



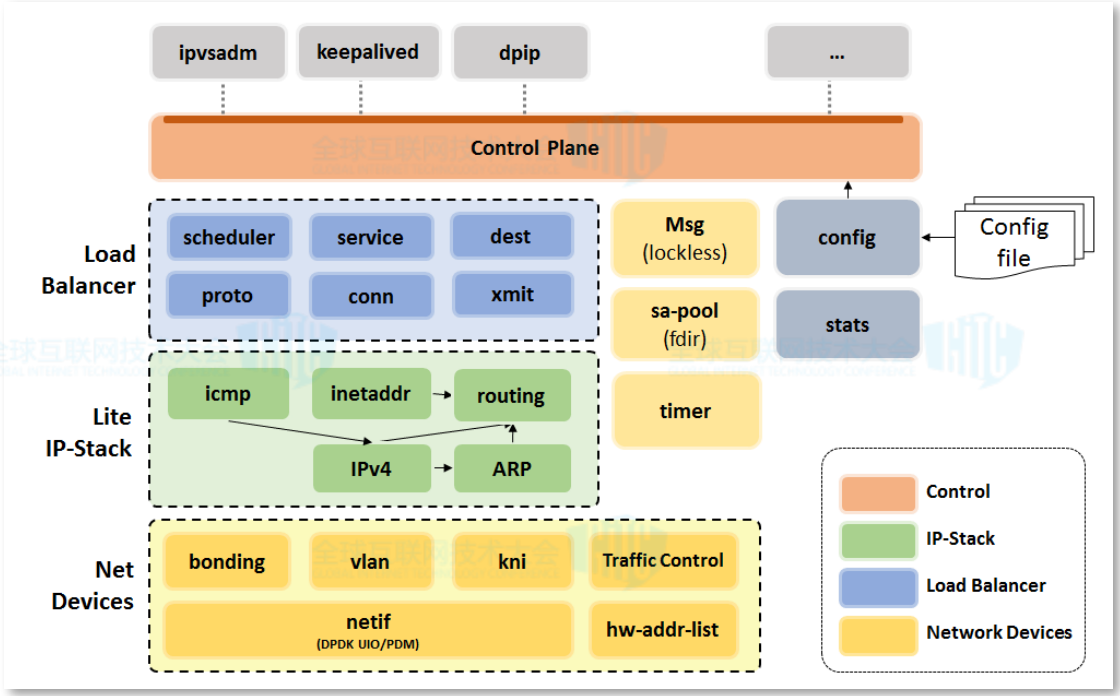
# 4层LB优化：DPVS架构

- 用户态实现
- Master/Worker
- Lockless
- 网卡队列/CPU绑定
- 跨CPU无锁通信
- 虚拟设备：bond/vlan/kni
- 安全相关：synproxy,黑名单,限流



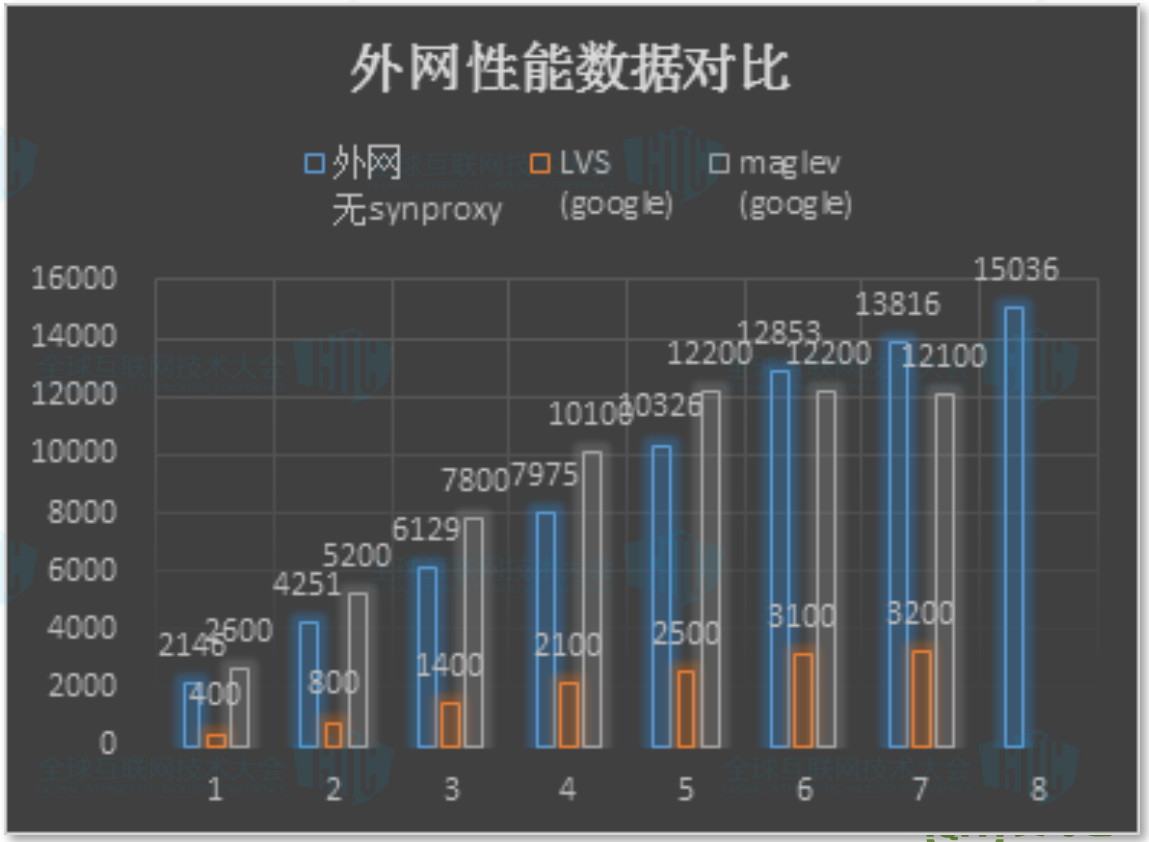
# 4层LB优化：DPVS功能特性

- TCP/UDP负载均衡，SNAT
- 高可用
  - 集群化、健康检查
- 高可扩展
  - 水平扩展LB/RS
- 轻量级IP栈
- 无锁化设计
- 虚拟接口
  - Bond/VLAN
- 安全相关
  - Synproxy/黑名单/限流



# 4层LB优化：DPVS性能数据

- 模式：fnat, tcp
- Client: wrk
- RS: nginx
- 内网：单网卡线速
- 外网：双网卡、双向



# 7层LB优化: DPDK+Nginx

- 需要完整用户态TCP/IP协议栈
- 开源协议栈调研：
  - ODP/OFP
  - mTCP：性能较好，协议栈功能过于简单，开发量大；
  - **F-stack**：性能优秀，功能完整的解决方案；
  - Seastar：C++，重构APP工作量大。
- 性能考虑？
- 功能考虑？
- 社区是否活跃？
- 便于开发、维护？

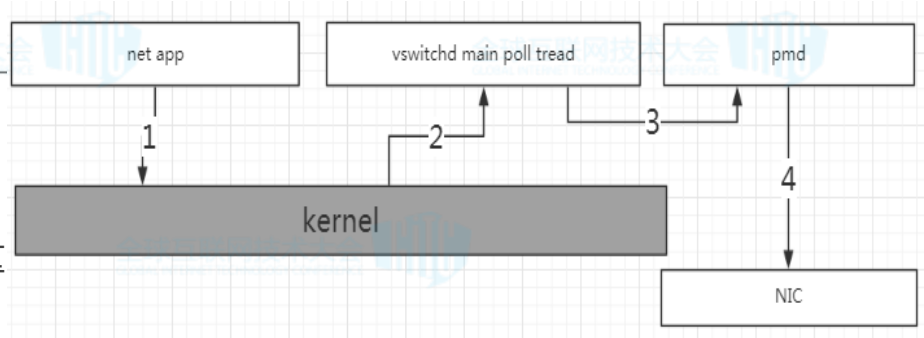
全球互联网技术大会 GLOBAL INTERNET TECHNOLOGY CONFERENCE
爱奇艺 iQIYI

# 虚拟化网络优化：OVS-DPDK状态

- 对比Legacy ovs性能，用ovs-dpdk，可以获得类似SR-IOV性能
- ovs-dpdk同 neutron，nova集成，解决overlay网络下时延问题
- ovs使用dpdk vhost-user 和vhostuserclient（auto reconnect）port
  - ovs >2.6.1， qemu >2.7
- Dpdk bonding
  - use bond4 lacp，need set phy switch
- Use 2 logical core for pmd threading
  - In same phy core or not ?

爱奇艺

# 虚拟化网络优化: 问题#1




- 物理机之间的流量和延时问题
  - 现象：
    - Glance的image同步或其他大流量
    - 某些共享存储使用的kernel协议栈
  - 原因：走物理机的管理网络报文路径太长
    - 不同于同在kernel下的ovs datapath
    - Userspace的datapatch，从协议栈，到依赖ovs-vswitchd的loop
  - 解决
    - 非sriov方式
      - Ovs local port tap/kni device -> use pure pmd thread + new++ isolated core
    - Sriov方式
      - 见后文

全球互联网技术大会 GLOBAL INTERNET TECHNOLOGY CONFERENCE

iQIYI 爱奇艺

# 虚拟化网络优化：问题#2

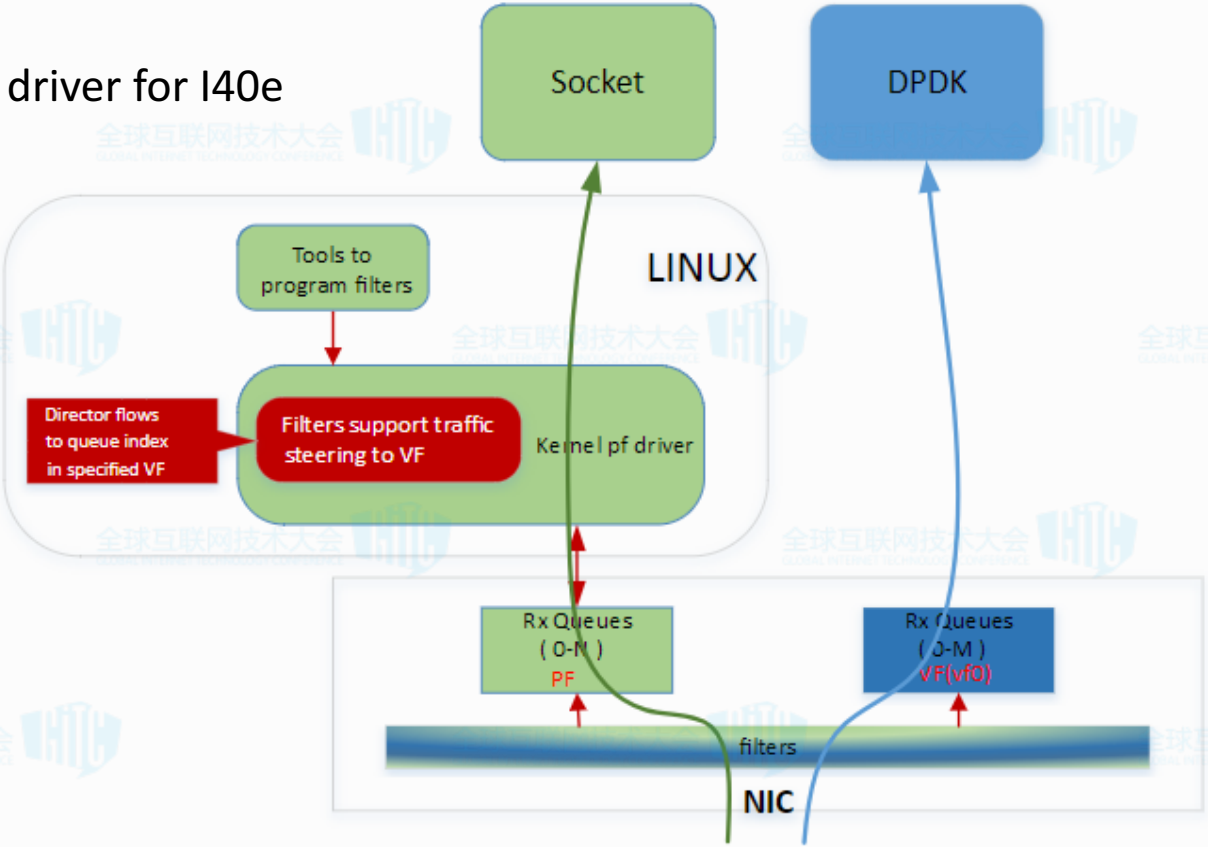
- 单网卡场景使用ovs dpdk后，如何隔离管理网络和数据网络
  - ovs-dpdk app 重启导致管理网络也中断，重要数据无法回传，idrac时间太长
  - Sriov
    - Dpdk dpdk bifurcated driver for i40e
      - L2 L3 flow scheduler
    - Customed ixgbe driver for 82599/x540T
      - Set VF promiscuous mode and unicast hash bit make PF work as a bridge
      - Vlan isolated the PF VF

全球互联网技术大会 GLOBAL INTERNET TECHNOLOGY CONFERENCE

iQIYI 爱奇艺

# 虚拟化网络优化：问题#2

bifurcated driver for I40e



# 虚拟化网络优化：问题#3

- ovs dpdk进程重启的时间优化问题
  - 重启时，hugepage内存初始化占据主要时间
  - 除去vm需要的内存，仅仅从小范围逐渐扩大查找dpdk需要连续hugepage内存mapping（4G左右）
  - 重启时间，从170s降到最快5s

全球互联网技术大会 GLOBAL INTERNET TECHNOLOGY CONFERENCE
iQIYI 爱奇艺

# 未来的挑战

- 4层LB（DPVS）：
  - 进一步优化，25G，40G，100G网卡
  - 流量控制，按VIP，源IP等模式进行限流
  - 采样功能，按VIP，流特征等进行数据采样
- 7层LB（DPDK-Nginx）：
  - 完善基本功能
  - DPDK-Nginx：完善部署、监控等，上线
- 虚拟化网络：
  - Security group 放到userspace
  - 避开Neutron DVR中对的kernel 依赖，尝试其他使用openflow的distributed方式

全球互联网技术大会 GLOBAL INTERNET TECHNOLOGY CONFERENCE

iQIYI 爱奇艺


悦　　享　　品　　质