

Anolis OS 优化 Virtio 协议 增强网络性能实践分享

衡琪

高性能网络 SIG

阿里云

/ 目录 /

Contents

- 01 Virtio 虚拟网卡简介
- 02 Virtio 虚拟网卡现状
- 03 Virtio 协议增强及应用
- 04 未来规划

Virtio 虚拟网卡简介-云上的网卡 Virtio-net

Virtio-net 是国内各种云平台、DPU 上虚拟机网卡
Virtio-net 属于 Virtio 标准下的一种网卡设备

Virtio 标准是一种 I/O 半虚拟化解决方案
标准化虚拟机(前端)与主机(后端)的交互方式

托管在 OASIS 组织，由 OASIS 的 Virtio 技术委员会
负责管理和维护，是一个国际性的开放标准



Virtio 虚拟网卡现状-挑战

性能

1

缺失中断聚合能力，
无法有效发挥后端
收包能力

2

Tunnel 报文队列集中
在一个 CPU 处理，无
法发挥多队列优势

功能

3

XDP 互斥于硬件校验和能力
XDP 互斥于 Jumbo frame

4

缺少运维统计信息，
硬件丢包信息统计，
不利于问题定位

Virtio 虚拟网卡现状-机遇

自研软硬件体系

长期活跃于社区

Nvidia
Oracle
Marvell
双周会

Virtio
Specification
投票权

Virtio 协议增强及应用- DIM 动态中断调节



中断参数不可调节
等待多久、累积多少个报文

解
决

动态中断调节, DIM
根据当前流量统计情况
动态调节中断参数

低负载流量

高负载流量

低载时延
无法优化

中断太多, 频繁
打断 CPU 收包进
程, 造成大量丢
包

效果

Nginx

10+%
吞吐

CPU

15 ~ 30%
sockperf
UDP

underload

时延降低
0.4 usec

Virtio 协议增强及应用- DIM 动态中断调节

01 队列级别的统计流量信息调整中断参数

3	14	246
个月	个版本	封邮件

制定 virtio 新标准、15+ 内核补丁

02

请求太多导致Kick 频繁，DMA 次数太多，影响 DPU 对多 VM、大队列场景的良好支持

请求合并(batch request)。后端对于控制命令处理效率提升 QP 倍！

03

流量情况复杂，决策可能失误

- 优化参数列表，对低、中载流量决策更加公平
- 更稳健的采样时间
- 添加新的局部最优点，有目的地优化低中载流量

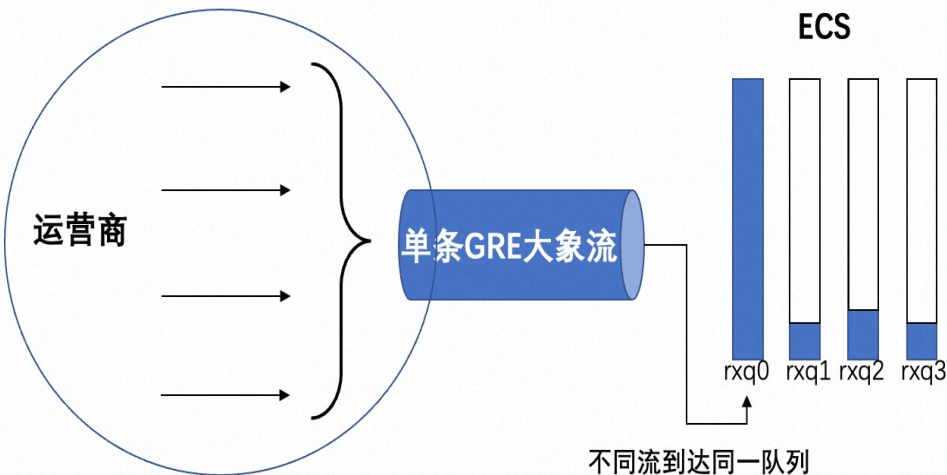
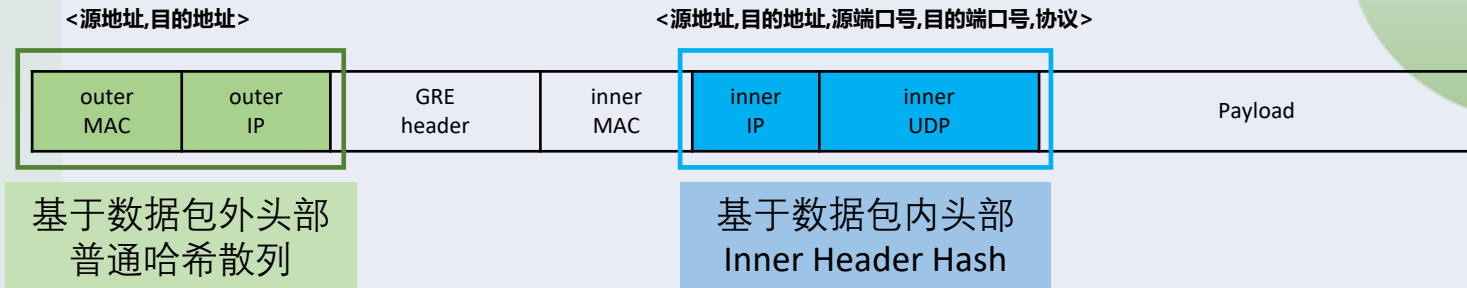
04

Ctrl vq 忙等待，造成 8 队列网卡 10+% 虚拟机 CPU 开销

命令批量化 + delayed work，完全消除额外的 CPU 开销

Virtio 协议增强及应用- Inner Header Hash

Tunnel 报文



背景

多条数据流被封装为单条 GRE 数据流，整个数据流拥有相同的外头部

问题

传统的外头部哈希，数据流全部被散列到相同队列处理，难以发挥多 CPU 处理优势

措施

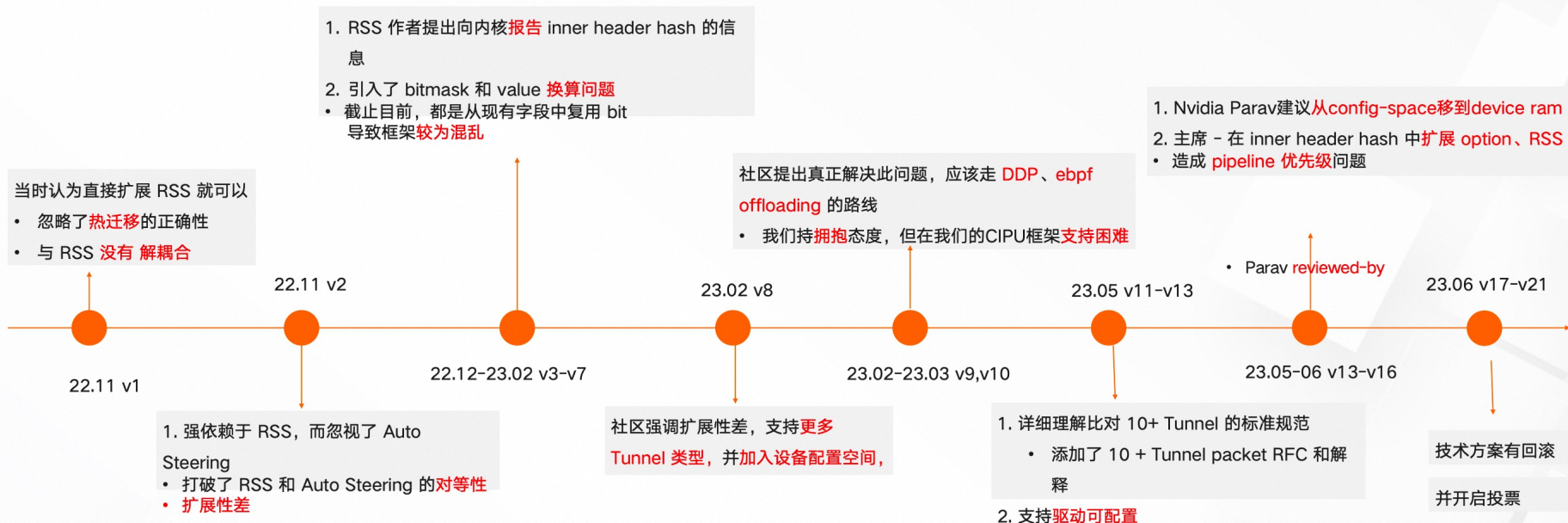
Inner Header Hash 彻底解决这种缺陷，针对传统隧道协议，充分发挥多队列、多 CPU 收包能力

Virtio 协议增强及应用- Inner Header Hash

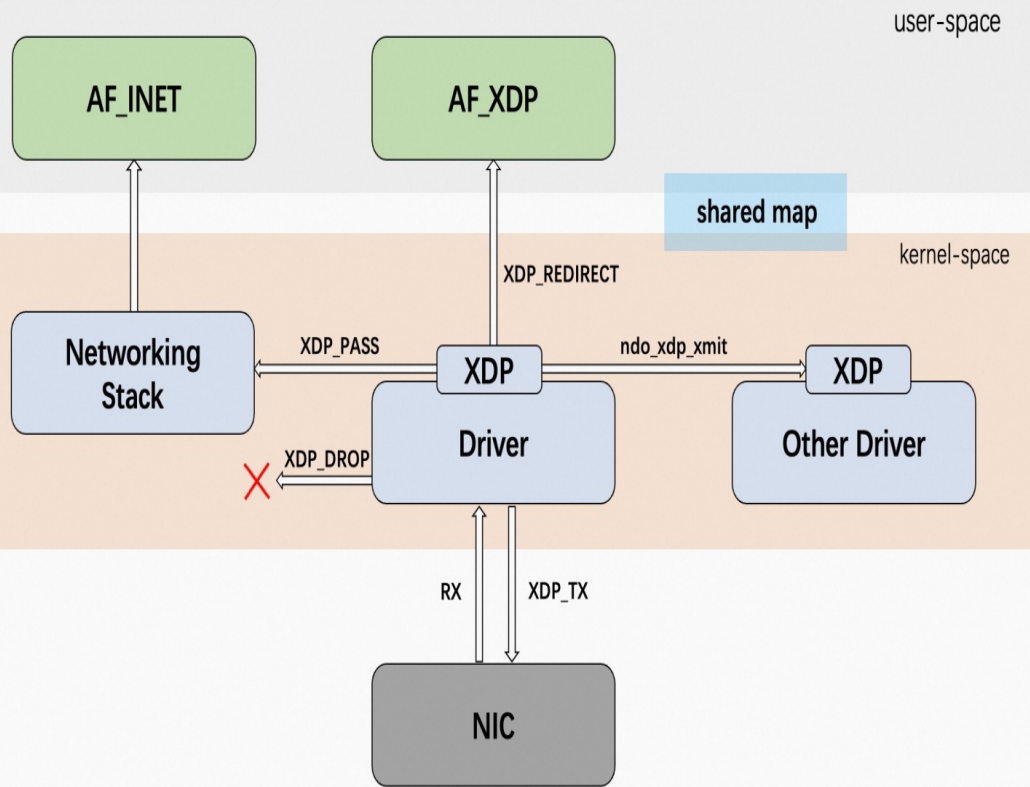
	动态设备定制 DDP	ebpf offloading	制定新 virtio 规范
Inner Header Hash 需求	✓	✓	✓
动态配置	✓	✓	✓
扩展性	好	好	差
占用硬件资源	多 永久占用设备资源, 即使虚拟机不使用	多 硬件支持 ebpf 程序 解析, JIT 编译器	少
对硬件架构入侵性	强	强	弱
最终方案			✓

Virtio 协议增强及应用- Inner Header Hash

8 个月 3 种方案 21 个版本 445 封邮件



Virtio 协议增强及应用- XDP (eXpress Data Path)



XDP作为 eBPF 程序的一种类型
Linux 内核中一种高性能数据包处理技术



应用价值

在防火墙、重定向、时延监控、零拷贝 (AF_XDP) 等场景具有重要使用价值。



缺点：支持不完善

1. 不支持与 Jumbo Frame 等大型帧共存
2. 不支持与硬件校验和卸载共存

Multi-buffer XDP

推送社区 11 个补丁支持与大型帧共存

Setup multi-buffer XDP

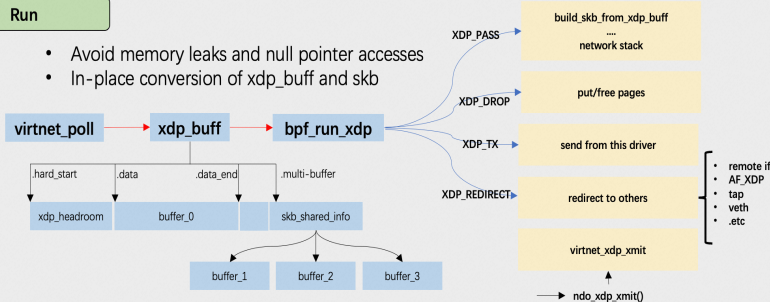
- Single-buffer XDP can work fine
- Run compatible with single-buffer XDP and can also handle single-buffer packets
- Allow larger MTU

Fill buffers for device

- Avoid the hole mechanism to get the desired result
- Focus on asynchronicity when producing and consuming buffers
- Correct meaning of truesize to help readers

Run

- Avoid memory leaks and null pointer accesses
- In-place conversion of xdp_buff and skb



支持与硬件校验和卸载共存

2023 年 5 月

提交 Survey 与 Virtio 社区商讨方案



2023 年 6 月

尝试重新探测校验和，但是会有一些 Corner Case 造成丢包



2023 年 11 月

计划制定新的 Virtio 规范从源头解决，经过 8 个版本，彻底修复了自 2011 年就存在的历史性问题

2023 Linux 社区云上网卡对于 XDP、AF-XDP 支持情况

	Basic	Redirect	ndo_xmit	multi-buf rx	multi-buf tx	AF_XDP <u>zerocopy</u>
ENA	✓	✓				
GVE	✓	✓	✓		✓	
virtio-net(社区)	✓	✓	✓	✓	✓	
virtio-net(Anolis)	✓	✓	✓	✓	✓	✓

- ENA, GVE 数据来源 netdevconf 0x17

- 龙蜥贡献

- 龙蜥优化增强, 适配

Virtio 协议增强及应用- Device stats

问题



运维时候的烦恼：突然开始丢包，由于丢包原因复杂，难以定位



更易掌握设备运行状态
获取设备统计信息便于调试

措施

2021.8 ~ 2023.10：支持 Virtio 网卡驱动获取设备的统计信息

接收队列	发送队列	控制队列
通知和中断数量	通知和中断数量	发送命令的数量
没有可用空间丢包	发送缓冲区过短	成功相应的数量
校验和错误被丢包	校验和卸载的报文数量	
超过限速丢包	超过限速丢包	
合包相关统计信息	TSO 相关统计信息	
未知原因丢包	未知原因丢包	



Virtio 协议增强 现在/未来规划

01

Receive Flow Filter

- 支持更强大的数据流过滤规则
- 支持更细粒度的数据流控制
- 作为 Virtio 数据流处理的基础设施，以支持 TC 等强大工具

02

Accelerate RFS

作为 RFS 的硬件加速版本
最大程度的减少 cache miss 带来的性能损失，提升数据性能

03

Dynamic Queue Creation

运行时创建队列，这对 XDP、ARFS 等场景具有重要意义，避免资源损耗

04

Inline descriptor

- 控制命令将内联间接描述符，提升命令处理速度
- Virtio-net-hdr 被内联，提升数据传输效率

05

Timestamp

支持硬件时间戳，便于调试

THANKS

S U B H E A D I N G



【龙蜥】高性能网络 SIG 开发者 & 用户群


【龙蜥】
高性能

【龙蜥】高性能网络 SIG...

218人



此二维码365天内有效 (2025年01月18日前)

 钉钉扫一扫群二维码，立即加入群聊

欢迎加入

