统信软件技术有限公司
UnionTech Software Technology Co., Ltd.

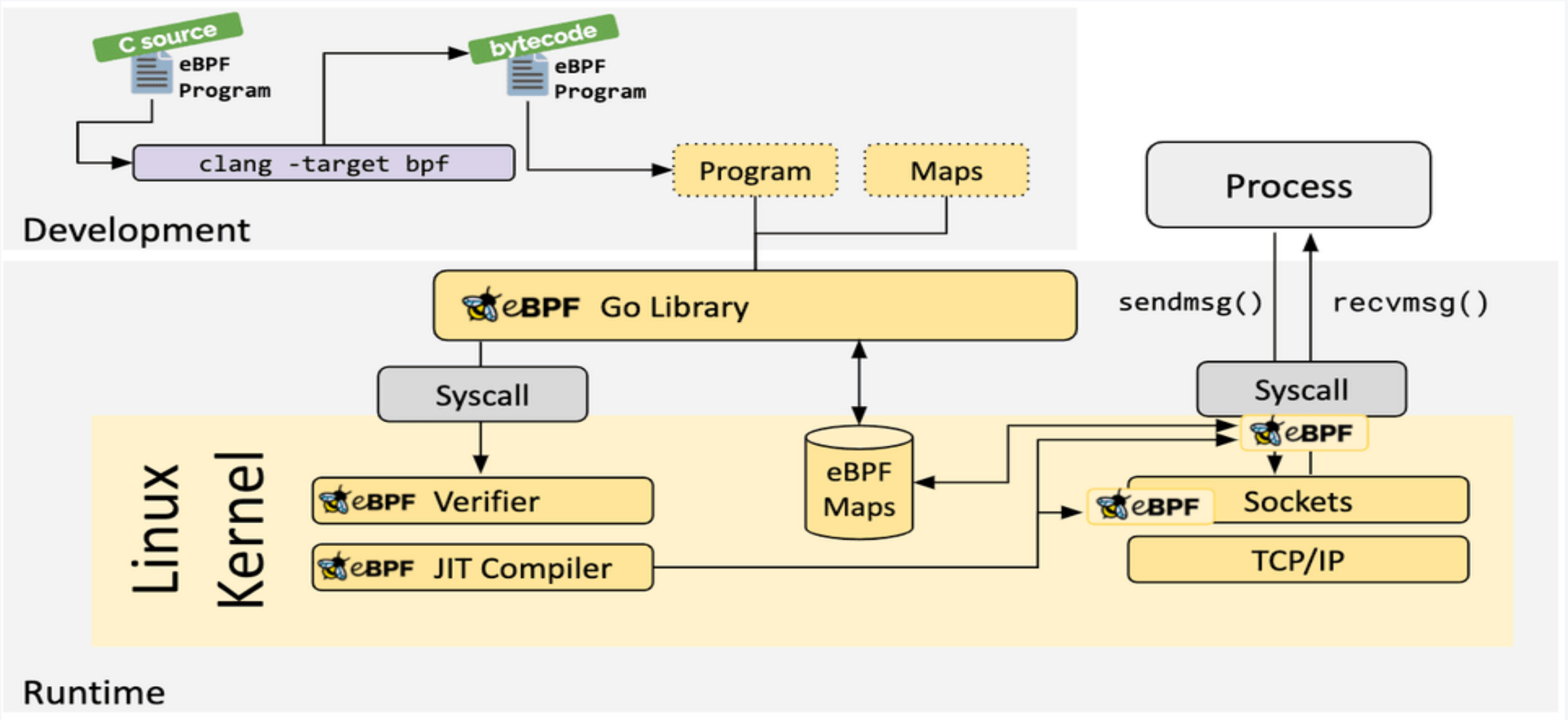
用eBPF扩充KVM的退出处理

郭克

why eBPF

基于其特权级别，内核一直以来都是观察程序运行，改变系统行为的理想场所。但是内核开发和调试困难，学习周期极长，进化极慢，对安全性和稳定性要求极高， 很难成为理想的创新操场。

How


C source
eBPF
Program
bytecode
eBPF
Program
clang -target bpf
Program
Maps
Development
Process
sendmsg()
recvmsg()
eBPF Go Library
Syscall
Syscall
eBPF
eBPF
Maps
Linux
Kernel
eBPF Verifier
eBPF
Sockets
eBPF JIT Compiler
TCP/IP
Runtime

目前eBPF涉及核心领域

网络 安全 观测

bcc： 内核追踪工具

Cilium： 网络，安全与内核观测工具

bpftrace：与bcc类似的内核追踪工具

Falco：安全

Katran：数据包处理

Pixie：K8S观察工具。

Calico：容器网络与容器安全工具。

Tetragon：系统观测工具。

Pyroscope：性能优化工具。

eCapture：密文抓包工具

Parca：持续性能工具。

Hubble：K8S网络，安全与服务管理工具。

ingress-node-firewall：K8S防火墙。

netobserv：网络流量观测工具。

blixt：K8S L4 load balancer

bpfd：bpf管理工具

Apache SkyWalking：性能监测工具。

L3AF：bpf程序管理工具

Alaz：K8S monitor

如何用eBPF改变网络以外的内核功能

存储

- 实现访问nvme盘的额外功能，比如索引，聚合以及安全绕过一些linux内核存储软件栈。
- https://github.com/xrp-project/XRP

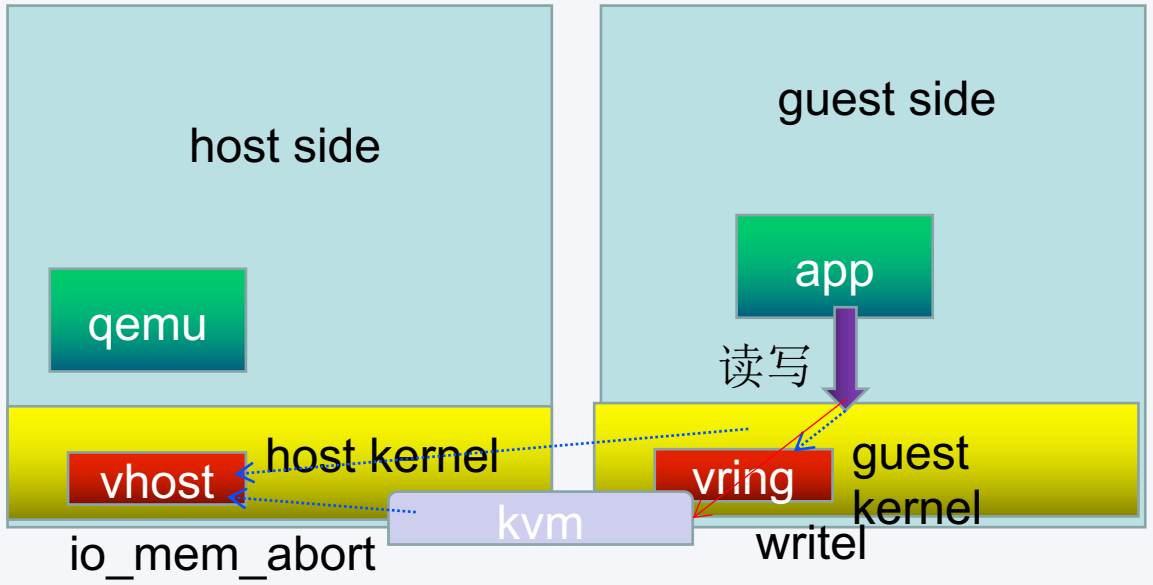
GPU

- 1.监测GPU性能
- 2. 实现一些GPU的功能。
- 3.GPU性能优化
- https://github.com/vchuravy/bpf_uvm

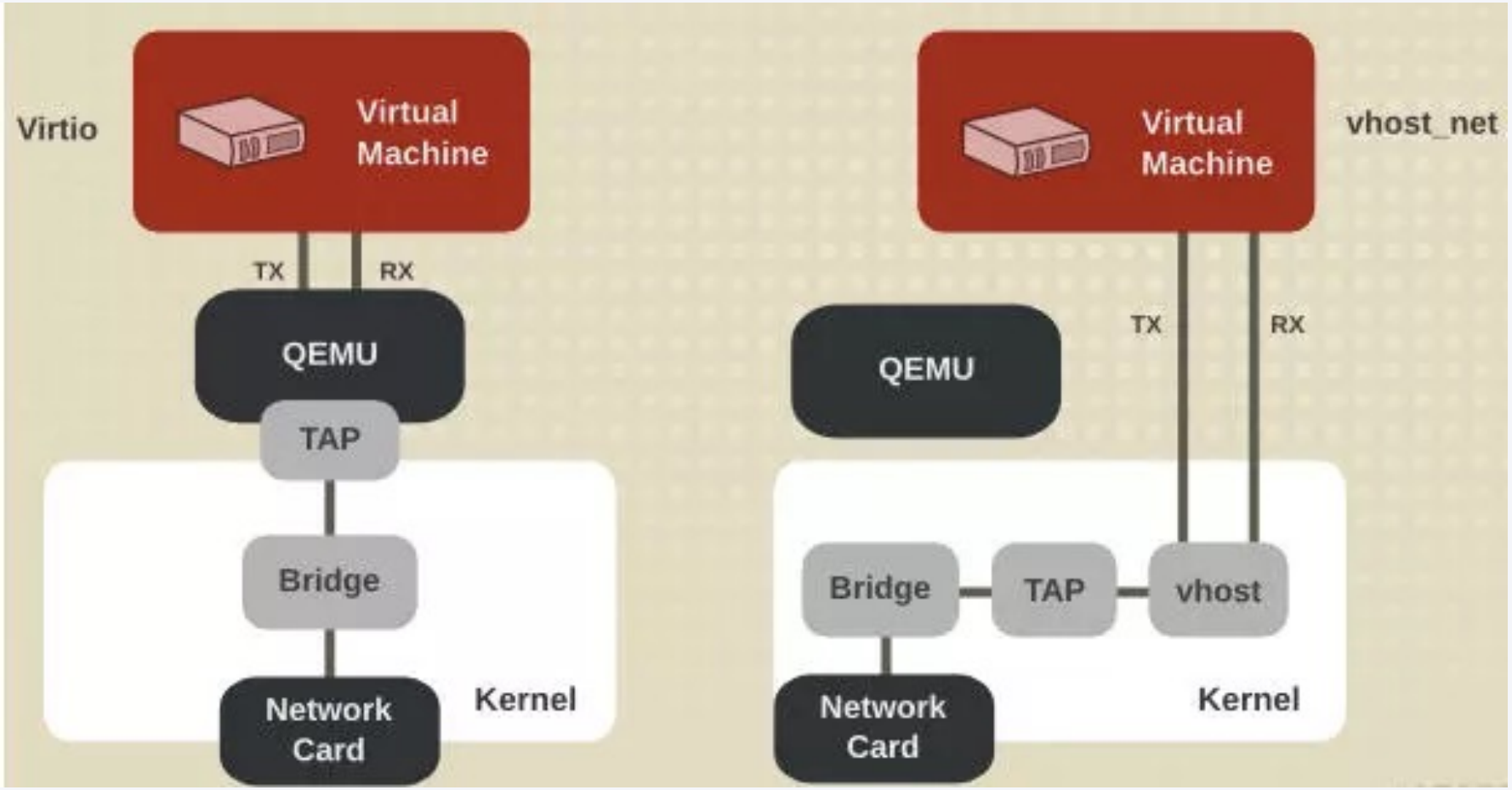
调度

- 利用ebpf进行调度策略更改。
- https://github.com/google/ghost-userspace

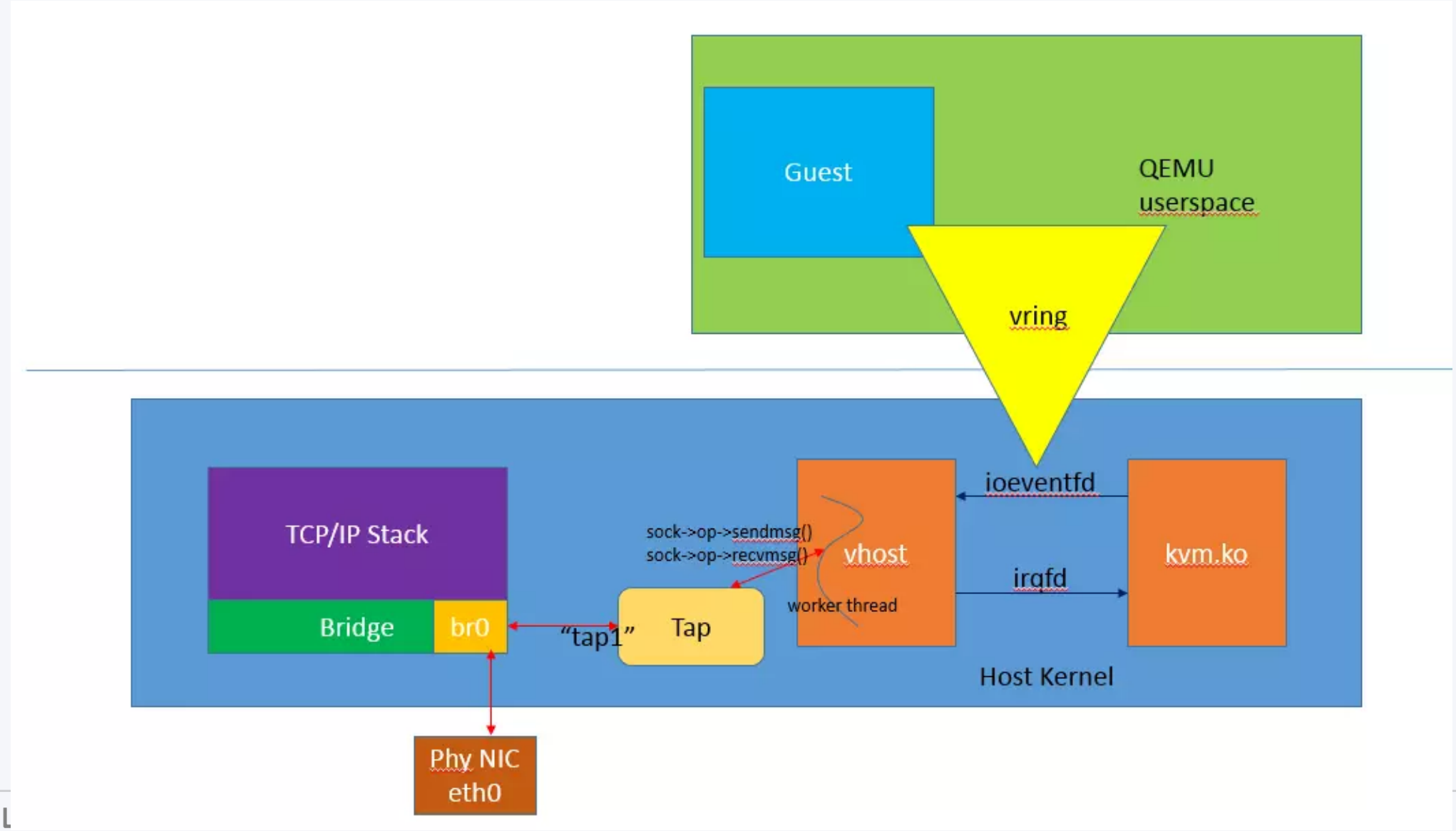
Why not vmm? try extend vhost?



看看vhost


Virtio
Virtual Machine
TX
RX
QEMU
TAP
Bridge
Network Card
Kernel
vhost_net
Virtual Machine
TX
RX
QEMU
Bridge
TAP
vhost
Network Card
Kernel

vhost交互数据流


Guest
QEMU
userspace
vring
TCP/IP Stack
Bridge
br0
"tap1"
Tap
sock->op->sendmsg()
sock->op->recvmsg()
vhost
worker thread
ioeventfd
irqfd
kvm.ko
Host Kernel
Phy NIC
eth0

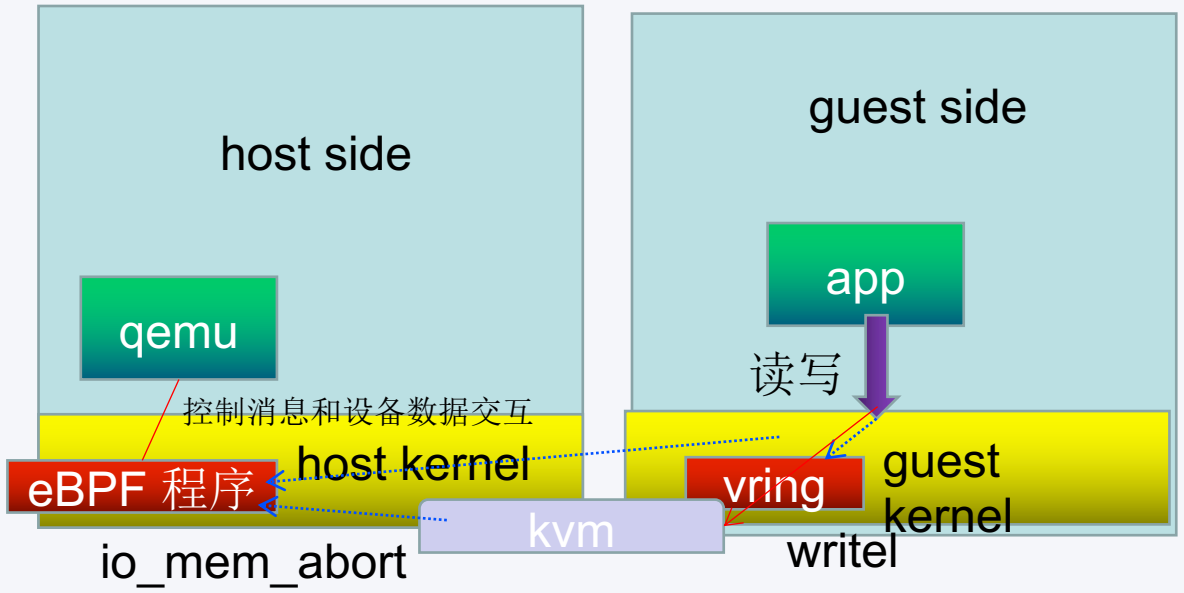
vhost 目前存在的问题

Vhost 目前广泛被用在加速virtio设备。
- 他把设备相关的大部分代码实现在内核里。
- 只支持virtio; 其它类型的设备要么返回给qemu处理，要么自己写代码模拟。
- qemu为了管路需要仍然需要和内核态的设备同步信息。
- 不能热更新。

我们能用eBPF快速扩展vhost吗?

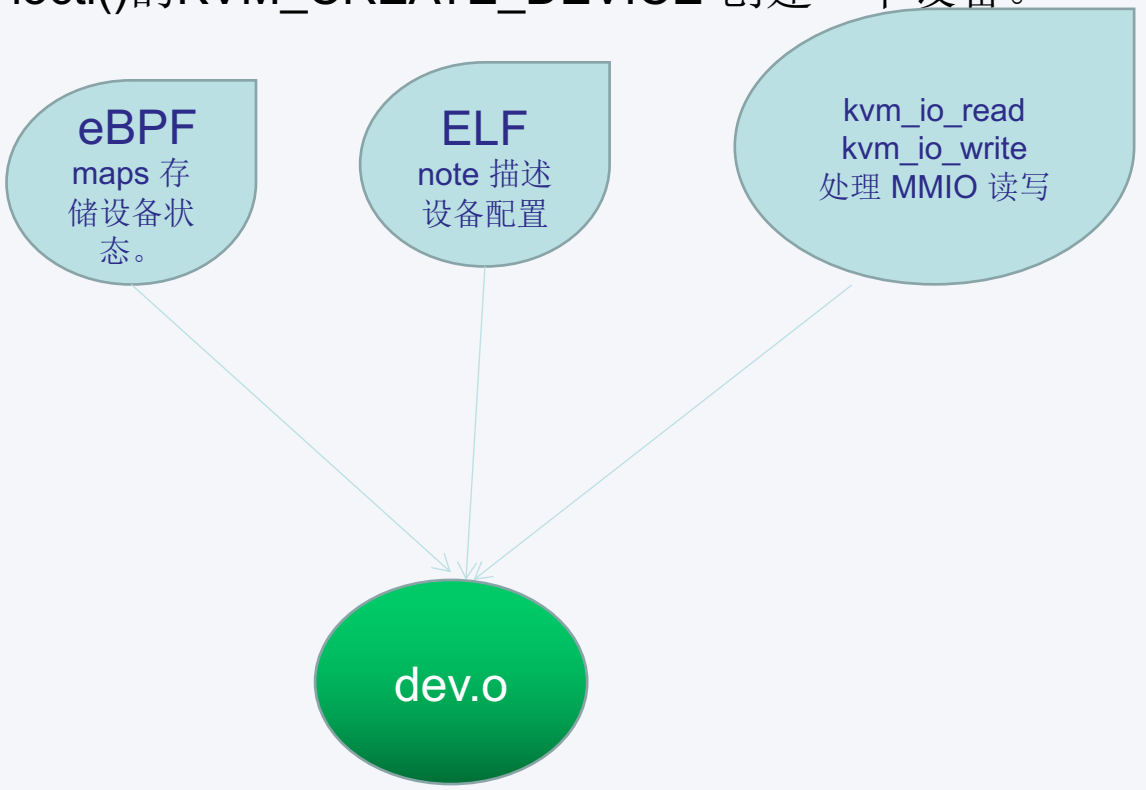
1. 通过形式化验证保证程序沙箱化不破坏内核功能。
2. 实时加载运行。
3. 灵活的ABIs
4. 好的权限控制。
5. 开发迅速。

try eBPF to extend vhost? 基本模型

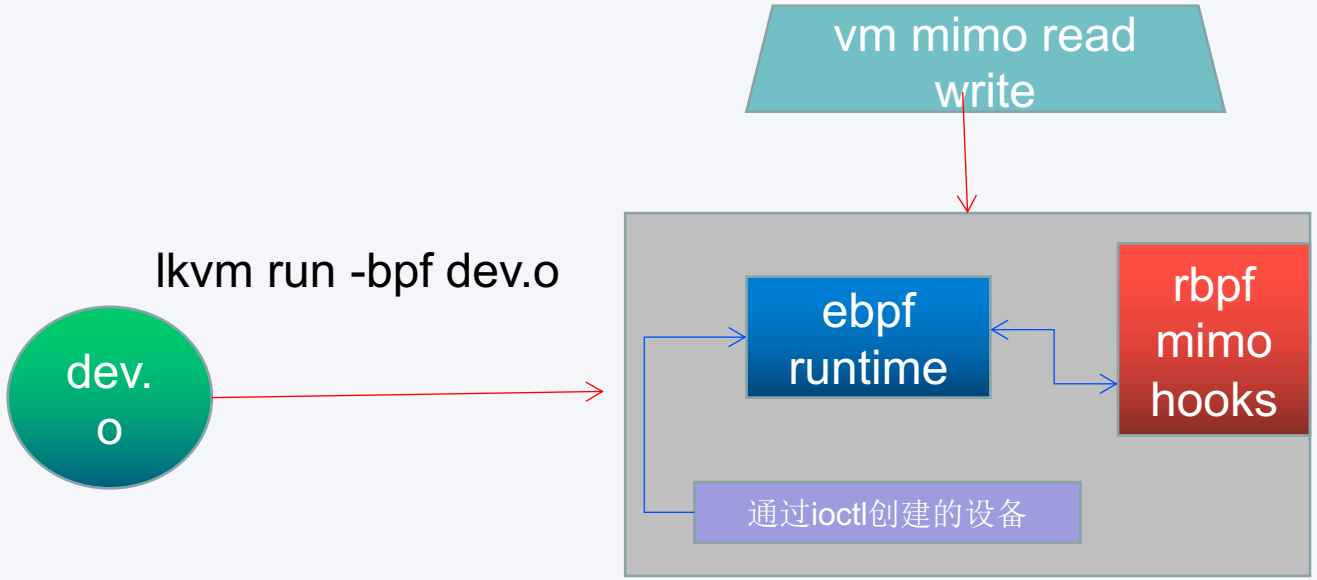


怎么实现呢?

通过kvm ioctl()的KVM_CREATE_DEVICE 创建一个设备。



How it works?


vm mimo read
write
lkvm run -bpf dev.o
dev.
o
ebpf
runtime
rbpf
mimo
hooks
通过ioctl创建的设备

Make it works

host kernel：
1. 实现一个bpf相关的kvm设备。 done
2. 设备映射的内存区域。done
3. 转发中断。 todo
4. 新的bpf程序类型实现。
5. 增加eBPF verifier codegen 规则。
6. Scheduler helpers。

qemu：
ELF note 解析和设备树产生。
Libbpf 抽取和加载程序。
初始化BPF相关的设备。
注入代码的逻辑。

guest kernel：
一个能演示的驱动done。

统信软件技术有限公司
UnionTech Software Technology Co., Ltd.

中国操作系统领创者
给世界更好的选择

统信软件官方微信公众号

统信软件官方社群

统信软件技术有限公司
UnionTech Software Technology Co., Ltd.

中国操作系统领创者
给世界更好的选择

统信软件官方微信公众号

统信软件官方社群

统信软件技术有限公司
UnionTech Software Technology Co., Ltd.

中国操作系统领创者
给世界更好的选择

统信软件官方微信公众号

统信软件官方社群