



Container Networking Solutions

Agenda

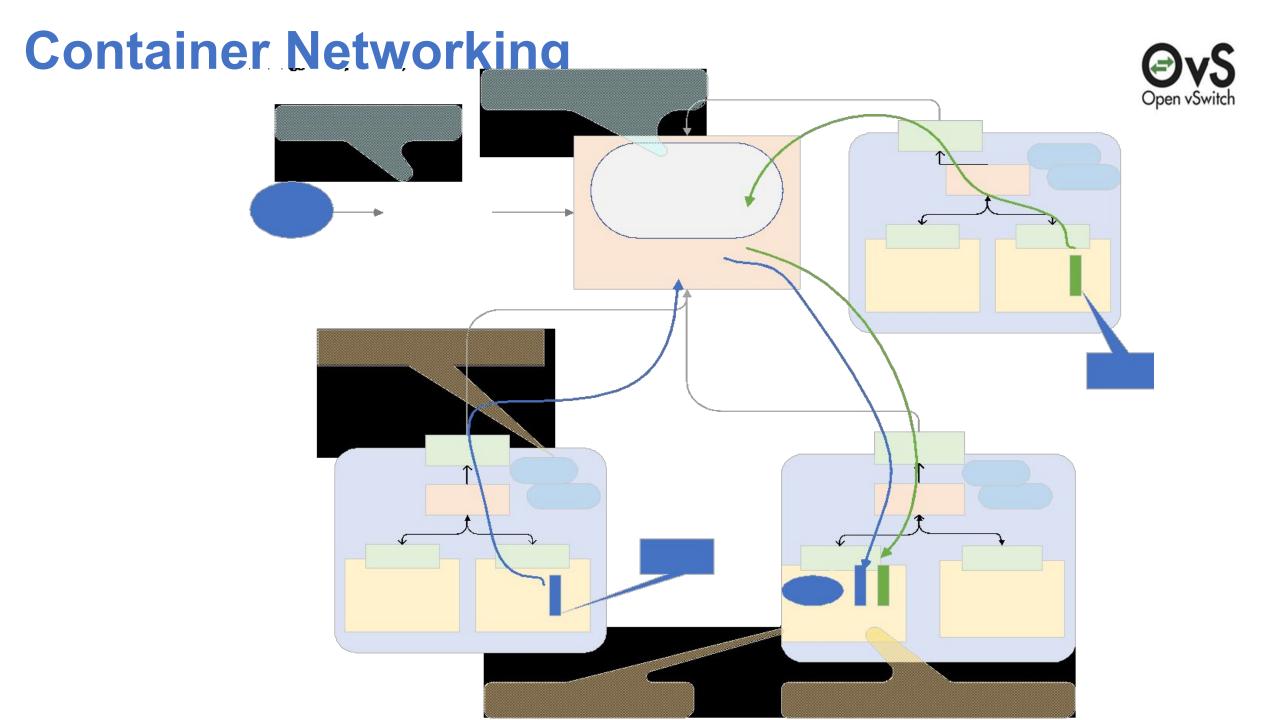


Introduction

- Container Network Functions & Interfaces
- Limitations
- Container Interface Classifier
- Community Solutions
- Our Proposed Solution
- Consolidation

Functionality Offload

- P4 Sample -Kubeproxy
- P4 Connection tracking
- P4 L4-L7 Classifier



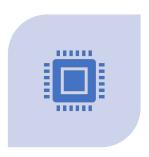












Device performance
Throughput,
Programmability,
Resource Scaling
has increased
significantly



effective utilization
of the system
resources (cores,
memory, network) to
improve
performance and
packing of containers



Ideal Container Network Solution



5. Existing interfaces for containers

Shared (Pod Interfaces)

- Stacked netdevs on a PCIe PF netdev
- Assigned to container namespaces
- Examples : MACVLAN, IPVLAN, bridge

Dedicated (Pod Interfaces)

- SR-IOV VFs
- Too heavy
 - Separate PCIE config space
 - HW based packet replication for broadcast, multicast – higher PCI BW utilization



Why Hardware Accelerate?

- End to end Maximize throughput
 - Avoid the SW long pole which limits how much a Server Pod can handle.
- End to end Native Scale out.
 - By Reduce Latency and Jitter
 - Dedicated resources takes away the need for OS to schedule on a shared resource....OS overhead for managing resources is gone.
 - Cpu scheduling, memory management etc
- Security & Isolation
 - Queue level isolation.

Assignable Container interfaces using X-IOV



(S-IOV)

Hardware Assisted Virtualization

Highly scalable and high- performance sharing of I/O devices across isolated domains

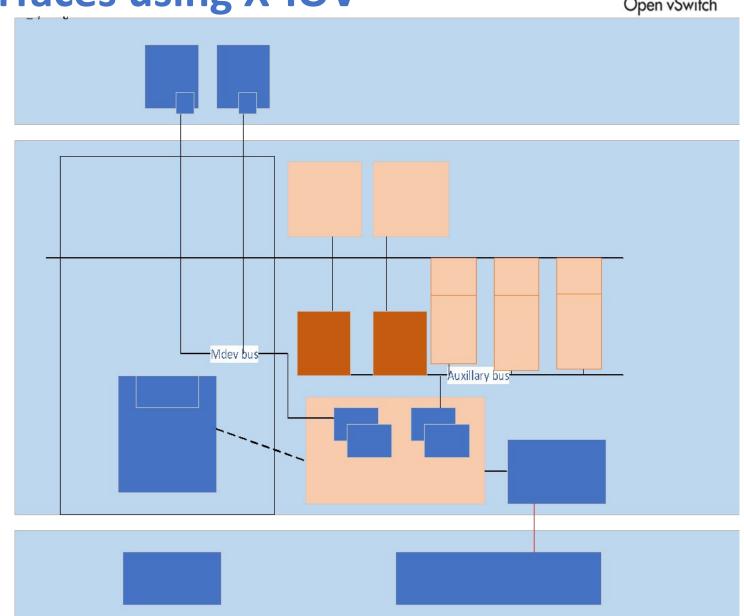
Assignable device Interfaces =>User Container Interfaces

Platform Scalability using PASID

Support Virtual Device Composition

RID and PASID identifies the address space associated with the request

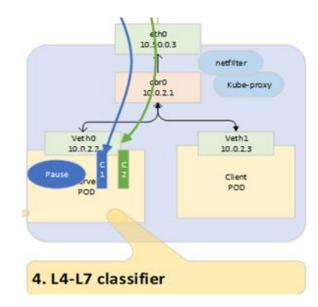
ADI Memory Mapped regions





4. Container Interface Classifier: Solution

- L4-L7 Classifier and forwarder in the HW
- This extends the HW Offloaded vSwitch Classification End Point to the Container Interface.



- Option1: HW Offload the classifier & forwarder
 - AF_XDP raw_socket bound to a HW vPort/QP through Side band filters.
 - Provide inline filters to be added in HW as part of TX packet.
 - ATR style in ADQ
- Option2: HW Offload the Classifier
 - Provide a meta data classification hint to kernel/user with a packet.
 - Flow mark or a 32bit hash value based on L4-L7 fields.



3. Node: Load Balancer, DNAT, CT

Existing Solutions

- Kube-proxy
 - Kernel Netfilters Not performant
 - Iptables O(n) chains proportional to size of cluster, in-place rule modifications not possible.
 - IPVS O(1) hash ipset but do not work well with other services requiring iptables for filtering
 - Kernel with eBF/XDP Accelerated
- Connection Tracking
 - Robust to syn floods but limited by max size
 - No flexibility, fixed hash algorithm and field selection for hash
- Kernel Overall not very flexible, latencies due to irq processing, context switching, slow API configuration interfaces



Community's Approach - eBPF/XDP

Benefits - Performant than the kernel

Designed as an alternative to DPDK.

Flexibility, code injected into the kernel

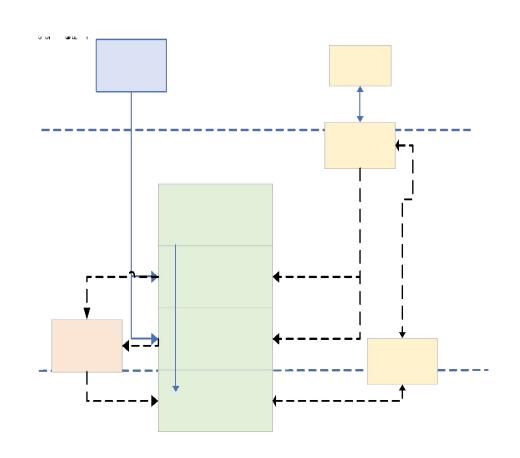
Ability to reload programs on-the-fly

Network Functions - Network Policy, Encryption, Load Balancer, Firewall, Monitoring etc

Plugins - Cilium, Callico, Katran (facebook), etc

Limitations -

- Not HW offloadable.
- Limited by performance and scalability.
- Kernel/user space context switching
- General purpose CPUs and Memory architecture is not ideal for Deep table Lookup
- Scale in CPU usage/CPU cache with load







Cilium does connect-time load balancing by hooking into the BPF & XDP/TC hook on the receive. When a program tries to connect to a Kubernetes service, Cilium intercepts the connection attempt, load balances with DNAT's to directly connect to the backend pod's IP instead

Throughput in queries/s – test run with fortio and ngnix. 2 clients, 2 services

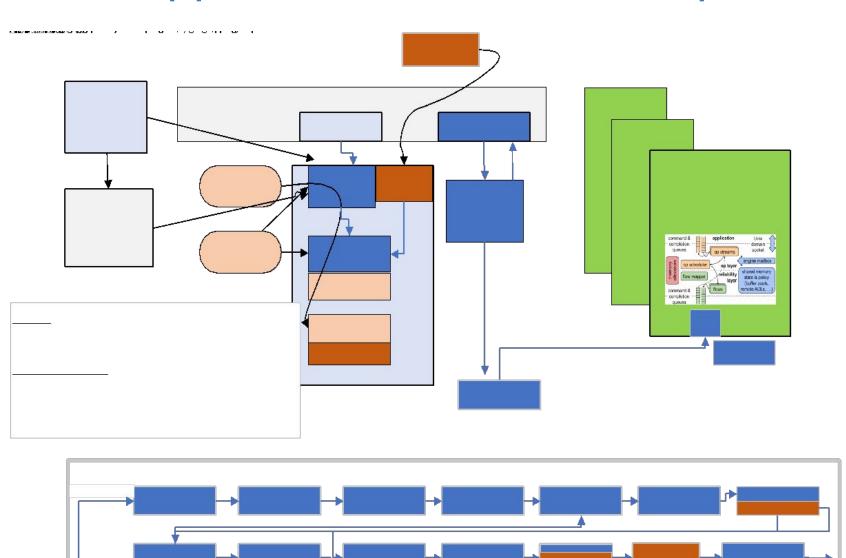
No XDP, 2 CLV interfaces			
Fortio + Nginx	Client 1	Client 2	Total
100B	47578.6	50411.5	97990.1
1500B	46174.4	49990.4	96164.8
XDP, 1 CLV interface			
Fortio + Nginx	Client 1	Client 2	Total
100B	64038.7	53970	118008.7
1500B	49623.5	48369.7	97993.2

No-XDP and XDP performed similar with 1000 connections. CPU consumption in both reaches to >8 cores with more sessions

XDP path today does not take advance of hardware XDP_REDIRECT queue designed to send packet to another interface, so no significant performance gains

Our Approach – The H/W Datapath





Programmable MAT tables. Contract between control plane and data plane for runtime control

Device capability defined by architecture. Eg wildcard match -TCAM, Exact match-DRAM. Compiler responsible for mapping.

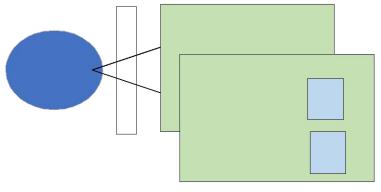
Parallel lookups, conditional actions & atomicity

Features are defined in the software. Faster introduction, verification, test and deployment.

Programmer defined 1) parser (parse graph); headers (inner/outer L2,L3,L4, app) and orders 2) how the output packet will look on the wire

Counters, meters, stateful registers, hash functions, ALU, TTLs and counters per entry, PRE

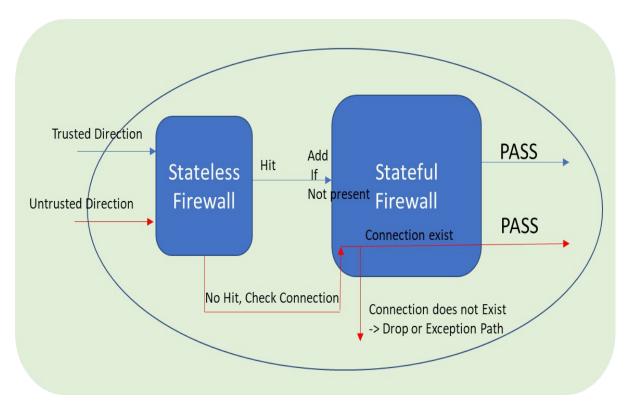
Node LB Data Plane P4



Mark Balancer +

```
extern ActionSelector{
       /// Construct a selection table for a given ActionProfile.
      ActionSelector(ActionProfile action_profile,
                               Hash< > hash,
                               SelectorMode t mode,
                               bit<32> max group size,
                               bit<32> num groups);
       ActionSelector(bit<32> size, Hash< > hash, SelectorMode t mode);
control simple lb(inout headers hdr,
       inout metadata meta,
       switch uint32 t lb table size,
       inout standard metadata t standard metadata) {
       Hash<switch_uint32_t>(HashAlgorithm_t.CRC32) selector_hash;
       ActionSelector(
              1024, selector_hash, SelectorMode_t.FAIR) pod_selector;
       // Pick an entry and apply DNAT
       action set_nhop(bit<48> pod_dmac, bit<32> pod_ipv4, bit<9> port) {
             hdr.ethernet.dstAddr = pod_dmac;
             hdr.ipv4.dstAddr = pod ipv4;
             hdr.tcp.dstPort = port;
              standard metadata.egress spec = port;
       table 1b {
              key =
                    hdr.ipv4.dstAddr : exact;
                    hdr.tcp.dstPort : exact;
                    hdr.ipv4.srcAddr : selector;
                    hdr.ipv4.dstAddr : selector;
                    hdr.ipv4.tcp.srcPort : selector;
                    hdr.ipv4.tcp.dstPort : selector;
                    hdr.ipv4.protocol : selector
              actions = {
                     NoAction;
```

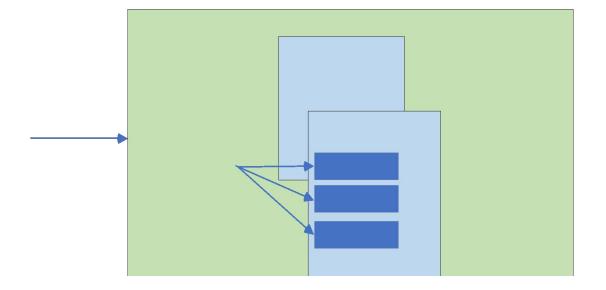
CT P4 Data Plane



```
C: > Users > asinghai > OneDrive - Intel Corporation > Documents > MEV > 

☐ CT mev.p4
      struct ct_aging_table_hit_params_t {
          FlowId t flow id:
          ExpireTimeSelection_t expire_time;
      action ct_aging_table_hit (
                                                      // See "Note 6"
          @per entry state in FlowId t flow id, // See "Note 0"
          @per_entry_state inout ExpireTimeSelection_t expire_time)
          if (modify expire soon on hit) {
              expire time = new expire time selection; // called TIME SEL in MEV
              modify entry to expire soon();
                                                // See "Note 2"
           } else if (update expire time) {
              expire time = new_expire_time_selection; // called TIME_SEL in MEV
              restart expire timer();
                                                // sweep count = 0
          } else {
              restart_expire_timer();
      action ct aging table miss() {
          FlowId t my flow id;
          bool add succeeded;
          if (add on miss) {
                                                // See "Note 4"
              my flow id = allocate flow id(); // See "Note 1"
              add succeeded =
                  add_entry("ct_aging_table_hit", // name of action
                            (ct_aging_table_hit_params_t)
                            {flow id = my flow id,
                     expire time = new expire time_selection});
              // add entry() initializes the new entry as if
      table ct aging table {
          key = {
              // P4 developer gets to select the key fields they want,
              // e.g. This is a 5 tuple plus a field like zone id , used to guarantee
              // that entries in different IP private address domains are
              // @tableonly and @defaultonly are standard annotations
              // defined in the P4 16 language specification.
```

L4-L7 P4 Classifier



```
control L4 L7 classifier(inout headers hdr,
            inout metadata meta,
            switch_uint32_t lb_table_size,
           inout standard_metadata_t standard_metadata) {
            // Forward to a Container Interface
            action set queue(bit<16> queue id) {
                 standard metadata.egress spec = queue id;
10
11
            // Set a flow ID meta data
            action set_flow_id(bit<32> mark) {
                 standard metadata.flow id = mark;
            // exact match using n-tuple
            table 1b {
                   key = {
                          hdr.ipv4.dstAddr : exact;
                          hdr.udp.dstPort : exact;
                          hdr.udp.srcPort : exact;
                          hdr.ipv4.srcAddr : exact;
                          hdr.udp.quic.cid : exact
                   actions = {
                          NoAction;
                          set_queue;
                          flow count.count();
                          set flow id;
                   const default action = Drop;
                   counters = flow counter;
                   size = 1b table size; // can be configured
                   implementation = pod selector;
```

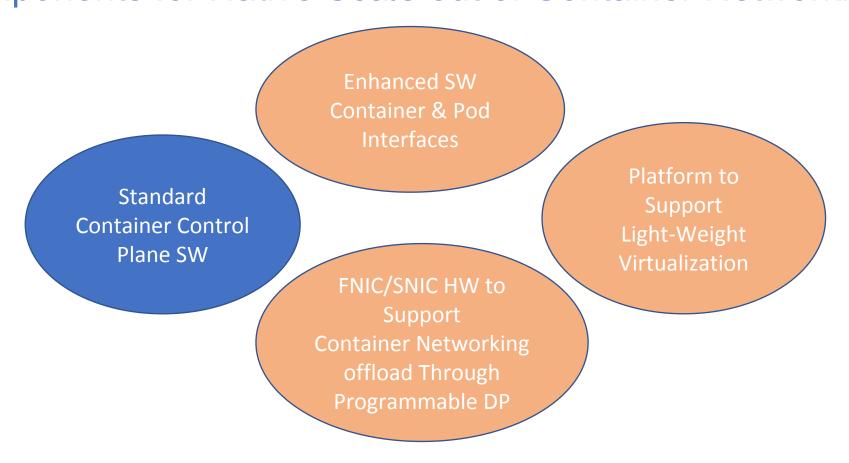


Opens

- Not every eBPF program can be HW offloaded as is. We are looking at all use cases.
- We would like to get community support in converting some well defined XDP implementations to p4 programs.
- P4 extensions or externs is an option for complete packet transformations like Crypto, checksum, packet replication etc.

Conclusion: Components for Native Scale out of Container Networking







Contacts

- P4 Code will be on github soon...
- Please email for more info...
- Contacts: Anjali Singhai <u>anjali.singhai@intel.com</u>
 Nupur Jain <u>nupur.jain@intel.com</u>
- Intel Container Networking Team:

,	n Daly,
Amritha Nambiar, Sri Pawel Szymanski, Kir Shaopeng He, Lia	dong Li, dhar Samudrala, ran Patil, ang Cunming , win Verplanke



eBPF Implementation Characterization - Cilium

Stack trace (NO XDP)

vmlinux	615.317s
nginx	55.704s
[Unknown]	24.101s
[Outside any known module]	24.101s
cls_bpf_classify	21.866s
$ abla$ tcf_classify \leftarrow tcf_classify_in	1.173s
$ abla$ tcf_classify \leftarrow tcf_classify \leftarrow	1.042s
[Unknown]	0.020s
$rac{1}{2}$ func@0x7a5bd4 \leftarrow func@0x79	Os
func@0x7a0570	Os
func@0x3527ee1	Os
ice.ko	23.259s
ip_tables.ko	20.728s
libc-2.13.so	11.371s
amplxe-perf	8.445s
cls_bpf.ko	8.339s
containerd-shim	8.209s

Stack trace (XDP)

10 6266	E0 704 E00
28.626s	52,704,500,000
28.626s	52,704,500,000
17.400s	29,923,000,000
3.194s	21,148,500,000
1.463s	609,500,000
1.273s	621,000,000
0.236s	322,000,000
0.035s	23,000,000
0.015s	11,500,000
0.005s	0
0.005s	34,500,000
Os	11,500,000
25.519s	26,553,500,000
17.896s	32,798,000,000
11.096s	22,482,500,000
7.943s	12,788,000,000
7.367s	8,533,000,000
1 1 1 1 1 1 1 7	7.400s 3.194s 3.463s 3.273s 3.236s 3.035s 3.005s 3.005s 3.005s 3.5.519s 3.7.896s 1.096s 7.943s

- XDP vs No XDP, CPU utilization is quite similar
- Benefits of XDP is being able to bypass the kernel stack in case of redirect.
- Redirect to external port requires dedicate HW Redirect TX queue.
- XDP Benefits can be derived from dedicated HW resources.

eBPF Implementation Characterization – Cilium - cont



Cilium chains in iptables

```
# Generated by iptables-save v1.6.1 on Tue Dec 1 13:22:26 2020
:PREROUTING ACCEPT [339745057:40541787768]
:OUTPUT ACCEPT [1966765:7841806195]
:CILIUM_OUTPUT_raw - [0:0]
:CILIUM PRE raw - [0:0]
-A PREROUTING -m comment --comment "cilium-feeder: CILIUM PRE raw" -j CILIUM PRE raw
-A OUTPUT -m comment --comment "cilium-feeder: CILIUM OUTPUT raw" -j CILIUM OUTPUT raw
-A CILIUM OUTPUT raw -o lxc+ -m mark --mark 0xa00/0xfffffeff -m comment --comment "cilium: NOTRACK for proxy return traff
-A CILIUM OUTPUT raw -o cilium host -m mark --mark 0xa00/0xfffffeff -m comment --comment "cilium: NOTRACK for proxy retur
-A CILIUM_PRE_raw -m mark --mark 0x200/0xf00 -m comment --comment "cilium: NOTRACK for proxy traffic" -j NOTRACK
# Completed on Tue Dec 1 13:22:26 2020
# Generated by iptables-save v1.6.1 on Tue Dec 1 13:22:26 2020
:PREROUTING ACCEPT [339745170:40541800199]
:INPUT ACCEPT [1689410:7607891639]
:FORWARD ACCEPT [338055760:32933908560]
:OUTPUT ACCEPT [1966765:7841806195]
:POSTROUTING ACCEPT [340022530:40775715026]
:CILIUM POST mangle - [0:0]
:CILIUM PRE mangle - [0:0]
-A PREROUTING -m comment --comment "cilium-feeder: CILIUM PRE mangle" -j CILIUM PRE mangle
-A POSTROUTING -m comment --comment "cilium-feeder: CILIUM POST mangle" -j CILIUM POST mangle
-A CILIUM PRE mangle -m socket --transparent -m comment --comment "cilium: any->pod redirect proxied traffic to host prox
-A CILIUM PRE mangle -p tcp -m mark --mark 0xd7ae0200 -m comment --comment "cilium: TPROXY to host cilium-dns-egress prox
-A CILIUM PRE mangle -p udp -m mark --mark 0xd7ae0200 -m comment --comment "cilium: TPROXY to host cilium-dns-egress prox
# Completed on Tue Dec 1 13:22:26 2020
# Generated by iptables-save v1.6.1 on Tue Dec 1 13:22:26 2020
*filter
:INPUT ACCEPT [1689410:7607891639]
: FORWARD DROP [0:0]
:OUTPUT ACCEPT [1966765:7841806195]
:CILIUM FORWARD - [0:0]
:CILIUM INPUT - [0:0]
:CILIUM OUTPUT - [0:0]
-A INPUT -m comment --comment "cilium-feeder: CILIUM INPUT" -i CILIUM INPUT
-A FORWARD -m comment --comment "cilium-feeder: CILIUM FORWARD" -j CILIUM FORWARD
-A OUTPUT -m comment --comment "cilium-feeder: CILIUM OUTPUT" -j CILIUM OUTPUT
-A CILIUM FORWARD -o cilium host -m comment --comment "cilium: any->cluster on cilium_host forward accept" -j ACCEPT
-A CILIUM FORWARD -i cilium host -m comment --comment "cilium: cluster->any on cilium host forward accept (nodeport)" -j
-A CILIUM FORWARD -i lxc+ -m comment --comment "cilium: cluster->any on lxc+ forward accept" -j ACCEPT
-A CILIUM FORWARD -i cilium net -m comment --comment "cilium: cluster->any on cilium net forward accept (nodeport)" -j AC
-A CILIUM INPUT -m mark --mark 0x200/0xf00 -m comment --comment "cilium: ACCEPT for proxy traffic" -j ACCEPT
-A CILIUM OUTPUT -m mark --mark 0xa00/0xfffffeff -m comment --comment "cilium: ACCEPT for proxy return traffic" -i ACCEPT
-A CILIUM OUTPUT -m mark ! --mark 0xe00/0xf00 -m mark ! --mark 0xd00/0xf00 -m mark ! --mark 0xa00/0xe00 -m comment --comm
# Completed on Tue Dec 1 13:22:26 2020
# Generated by iptables-save v1.6.1 on Tue Dec 1 13:22:26 2020
:PREROUTING ACCEPT [0:0]
. THIDLIT ACCEPT [A.A]
```

Cilium's own conntrack table

TCP IN 12.91.212.200:37386 -> 10.0.0.241:80 expires=17272887 RxPackets=18796 RxBytes=2057656 RxFlagsSeen=0x1b LastRxReport=17251794 TxPackets=18915 TxBytes=4544990 TxFlagsSeen=4 TCP OUT 13.91.212.200:33224 -> 10.0.0.55:80 expires=17275420 RxPackets=83205 RxBvtes=19555155 RxFlagsSeen=0x1a LastRxReport=17254320 TxPackets=80142 TxBvtes=9897097 TxFlagsSeen= TCP OUT 12.91.212.202:30007 -> 12.91.212.200:36342 service expires=17272760 RxPackets=0 RxBvtes=6 RxFlagsSeen=0x00 LastRxReport=0 TxPackets=0 TxBvtes=0 TxFlagsSeen=0x1f LastTxRu TCP OUT 12.91.212.202:38900 -> 12.91.212.200:30008 expires=17274846 RxPackets=2 RxBytes=120 RxFlagsSeen=0x14 LastRxReport=17253753 TxPackets=1 TxBytes=74 TxFlagsSeen=0x02 LastT; TCP OUT 12.91.212.200:43362 -> 10.0.0.246:80 expires=17275288 RxPackets=141178 RxBytes=33482100 RxFlagsSeen=0x1a LastRxReport=17254194 TxPackets=137700 TxBytes=17005280 TxFlagsSeen=0x1a LastRxReport=17254194 TxPackets=17254194 TxPackets=1 TCP IN 13.91.212.200:49604 -> 10.0.0.55:80 expires=17271556 RxPackets=90047 RxBytes=9859771 RxFlagsSeen=0x1b LastRxReport=17250463 TxPackets=91583 TxBytes=21846155 TxFlagsSeen=4 TCP IN 13.91.212.200:55058 -> 10.0.0.55:80 expires=17273171 RxPackets=52349 RxBytes=5757922 RxFlagsSeen=0x1b LastRxReport=17252078 TxPackets=53989 TxBytes=49437539 TxFlagsSeen=4 TCP OUT 12.91.212.200:43004 -> 10.0.0.241:80 expires=17275006 RxPackets=1016 RxBvtes=227571 RxFlagsSeen=0x1a LastRxReport=17253912 TxPackets=923 TxBvtes=113493 TxFlagsSeen=0x1e TCP IN 12.91.212.200:43512 -> 10.0.0.246:80 expires=17275288 RxPackets=140974 RxBytes=17409857 RxFlagsSeen=0x1e LastRxReport=17254195 TxPackets=145276 TxBytes=34327857 TxFlagsSe TCP OUT 13.91.212.202:48040 -> 13.91.212.200:30008 expires=17274814 RxPackets=2 RxBytes=120 RxFlagsSeen=0x14 LastRxReport=17253721 TxPackets=1 TxBytes=74 TxFlagsSeen=0x02 LastT: TCP IN 12.91.212.200:44096 -> 10.0.0.241:80 expires=17275421 RxPackets=102519 RxBytes=12660607 RxFlagsSeen=0x1e LastRxReport=17254328 TxPackets=104458 TxBytes=24884841 TxFlagsSe TCP IN 12.91.212.200:43468 -> 10.0.0.246:80 expires=17275288 RxPackets=130334 RxBytes=16095587 RxFlagsSeen=0x1e LastRxReport=17254195 TxPackets=133864 TxBytes=31706643 TxFlagsSe TCP OUT 12.91.212.202:30007 -> 12.91.212.200:42992 service expires=17275006 RxPackets=0 RxBvtes=6 RxFlagsSeen=0x00 LastRxReport=0 TxPackets=0 TxPackets=0 TxFlagsSeen=0x1e LastTxRu TCP OUT 13.91.212.202:30008 -> 13.91.212.200:54258 service expires=17272887 RxPackets=0 RxBytes=9 RxFlagsSeen=0x00 LastRxReport=0 TxPackets=0 TxBytes=0 TxFlagsSeen=0x1b LastTxRu TCP OUT 12.91.212.202:54410 -> 12.91.212.200:30007 expires=17274731 RxPackets=2 RxBytes=120 RxFlagsSeen=0x14 LastRxReport=17253638 TxPackets=1 TxBytes=74 TxFlagsSeen=0x02 LastT: TCP OUT 12.91.212.202:54426 -> 12.91.212.200:30007 expires=17274731 RxPackets=2 RxBytes=120 RxFlagsSeen=0x14 LastRxReport=17253638 TxPackets=1 TxBytes=74 TxFlagsSeen=0x02 LastT; TCP OUT 12.91.212.202:30007 -> 12.91.212.200:35150 service expires=17272401 RxPackets=0 RxBytes=8 RxFlagsSeen=0x00 LastRxReport=0 TxPackets=0 TxPtagsSeen=0x1f LastTxRi TCP IN 12.91.212.200:42930 -> 10.0.0.246:80 expires=17275006 RxPackets=1174 RxBvtes=144442 RxFlagsSeen=0x1e LastRxReport=17253913 TxPackets=1292 TxBvtes=289662 TxFlagsSeen=0x1a TCP IN 13.91.212.200:33946 -> 10.0.0.55:80 expires=17275768 RxPackets=24996 RxBytes=3086447 RxFlagsSeen=0x1a LastRxReport=17254672 TxPackets=25670 TxBytes=6078920 TxFlagsSeen=0: TCP OUT 12.91.212.202:30007 -> 12.91.212.200:38580 service expires=17273496 RxPackets=0 RxBytes=6 RxFlagsSeen=0x00 LastRxReport=0 TxPackets=0 TxBytes=0 TxFlagsSeen=0x1b LastTxRv TCP IN 13.91.212.200:34256 -> 10.0.0.55:80 expires=17275769 RxPackets=21454 RxBytes=2649240 RxFlagsSeen=0x1a LastRxReport=17254675 TxPackets=22597 TxBytes=5255183 TxFlagsSeen=0x1a LastRxReport=17254675 TxPackets=17254675 TxPackets= TCP OUT 13.91.212.200:33150 -> 10.0.0.55:80 expires=17275420 RxPackets=80644 RxBytes=19005294 RxFlagsSeen=0x1a LastRxReport=17254316 TxPackets=77972 TxBytes=9629102 TxFlagsSeen TCP OUT 12.91.212.202:30007 -> 12.91.212.200:44004 service expires=17275421 RxPackets=0 RxBytes=8 RxFlagsSeen=0x00 LastRxReport=0 TxPackets=0 TxBytes=0 TxFlagsSeen=0x1e LastTxRu TCP OUT 13.91.212.202:30008 -> 13.91.212.200:55654 service expires=17273480 RxPackets=0 RxBytes=9 RxFlagsSeen=0x00 LastRxReport=0 TxPackets=0 TxBytes=0 TxFlagsSeen=0x1b LastTxRu TCP OUT 12.91.212.202:55190 -> 12.91.212.200:30007 expires=17274765 RxPackets=2 RxBytes=120 RxFlagsSeen=0x14 LastRxReport=17253672 TxPackets=1 TxBytes=74 TxFlagsSeen=0x02 LastT: TCP OUT 12.91.212.202:39914 -> 12.91.212.200:30008 expires=17274929 RxPackets=2 RxBvtes=120 RxFlagsSeen=0x14 LastRxReport=17253836 TxPackets=1 TxBvtes=74 TxFlagsSeen=0x02 LastT; TCP OUT 12.91.212.202:38932 -> 12.91.212.200:30008 expires=17274846 RxPackets=2 RxBytes=120 RxFlagsSeen=0x14 LastRxReport=17253753 TxPackets=1 TxBytes=74 TxFlagsSeen=0x02 LastT; TCP OUT 12.91.212.202:38888 -> 12.91.212.200:30008 expires=17274846 RxPackets=2 RxBytes=120 RxFlagsSeen=0x14 LastRxReport=17253753 TxPackets=1 TxBytes=74 TxFlagsSeen=0x02 LastT; TCP OUT 12.91.212.202:30007 -> 12.91.212.200:44106 service expires=17275421 RxPackets=0 RxBytes=8 RxFlagsSeen=0x00 LastRxReport=0 TxPackets=0 TxBytes=0 TxFlagsSeen=0x1e LastTxRe TCP IN 12.91.212.200:44194 -> 10.0.0.246:80 expires=17275421 RxPackets=115638 RxBytes=14280861 RxFlagsSeen=0x1e LastRxReport=17254328 TxPackets=118303 TxBytes=28101171 TxFlagsSi TCP OUT 12.91.212.202:30007 -> 12.91.212.200:42722 service expires=17274983 RxPackets=0 RxBvtes=8 RxFlagsSeen=0x00 LastRxReport=0 TxPackets=0 TxBvtes=0 TxFlagsSeen=0x1e LastTxRu



