Liberating K8s from kube-proxy and iptables (and netfilter)

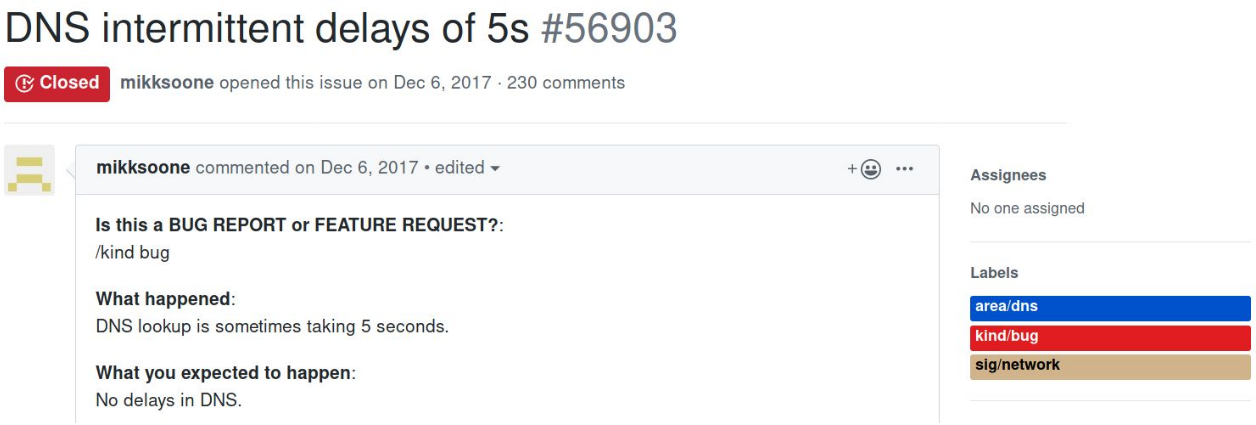
Martynas Pumputis, Cilium
(Daniel Borkmann, Thomas Graf, André Martins)

cilium

Performance

```
# perf top -a -e cycles:k

PerfTop:   16326 irqs/sec  (all, 4 CPUs)
-------------------------------------------------------------------------------
     8.79%  [kernel]            [k] native_sched_clock
     4.99%  [ip_tables]         [k] ipt_do_table
     3.09%  [e1000e]            [k] e1000_irq_enable
     2.51%  [nf_conntrack]      [k] __nf_conntrack_find_get
     2.03%  [kernel]            [k] fib_table_lookup
     1.98%  [kernel]            [k] sched_clock_cpu
     1.75%  [nf_conntrack]      [k] tcp_packet
     1.65%  [nf_conntrack]      [k] nf_conntrack_tuple_taken
     [...]
```

Reliability



DNS intermittent delays of 5s #56903

Closed mikksoone opened this issue on Dec 6, 2017 · 230 comments

mikksoone commented on Dec 6, 2017 • edited

Is this a BUG REPORT or FEATURE REQUEST?:
/kind bug

What happened:
DNS lookup is sometimes taking 5 seconds.

What you expected to happen:
No delays in DNS.

Assignees

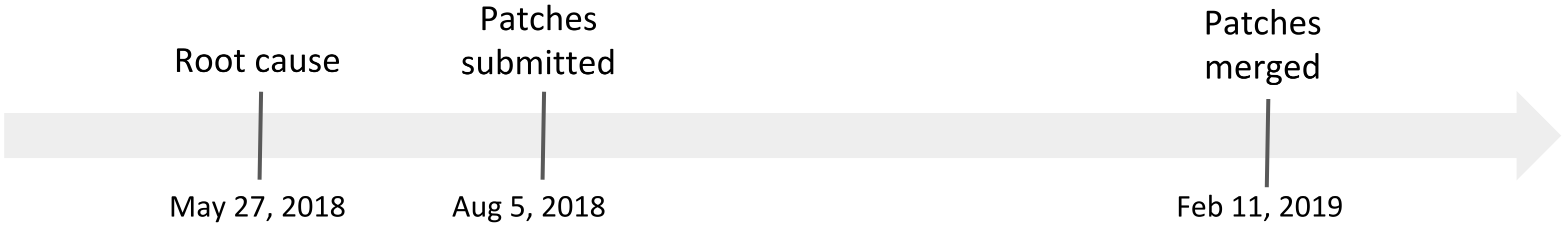
No one assigned

Labels

area/dns

kind/bug

sig/network



Root cause

May 27, 2018

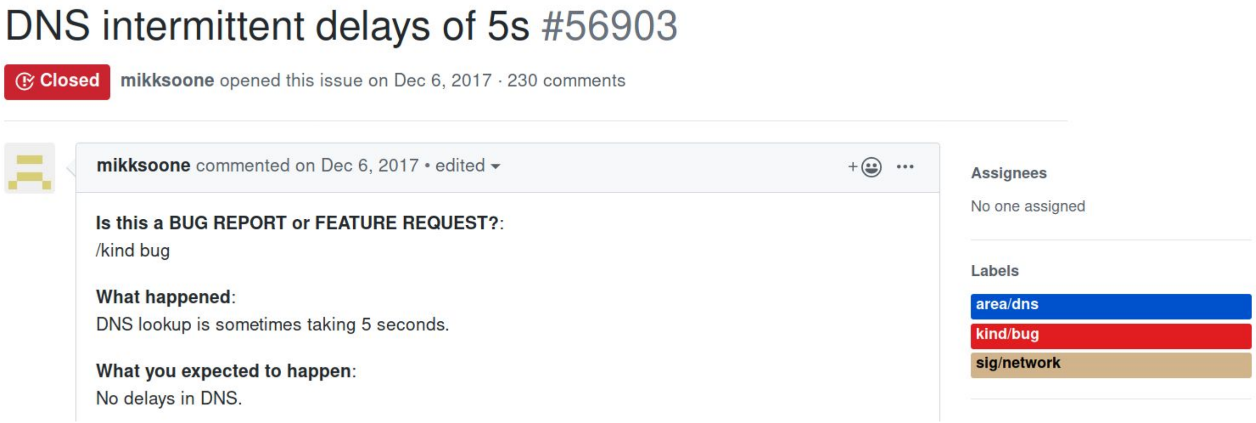
Patches submitted

Aug 5, 2018

Patches merged

Feb 11, 2019

Reliability



DNS intermittent delays of 5s #56903

Closed mikksoone opened this issue on Dec 6, 2017 · 230 comments

mikksoone commented on Dec 6, 2017 • edited

Is this a BUG REPORT or FEATURE REQUEST?:
/kind bug

What happened:
DNS lookup is sometimes taking 5 seconds.

What you expected to happen:
No delays in DNS.

Assignees

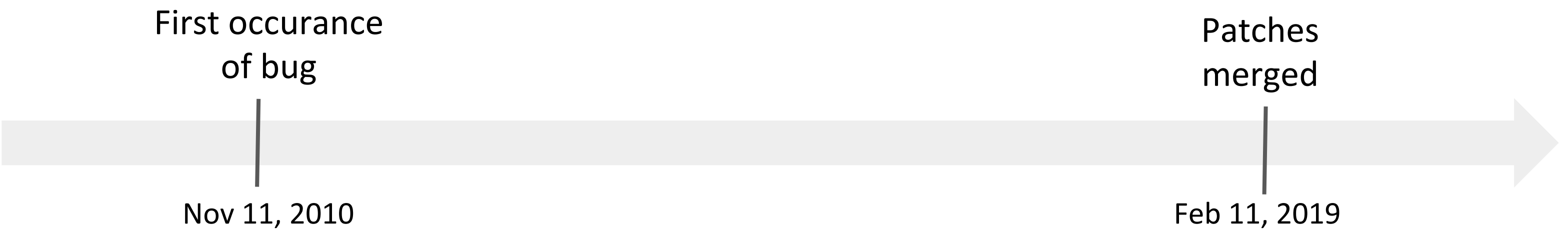
No one assigned

Labels

area/dns

kind/bug

sig/network



First occurance of bug

Nov 11, 2010

Patches merged

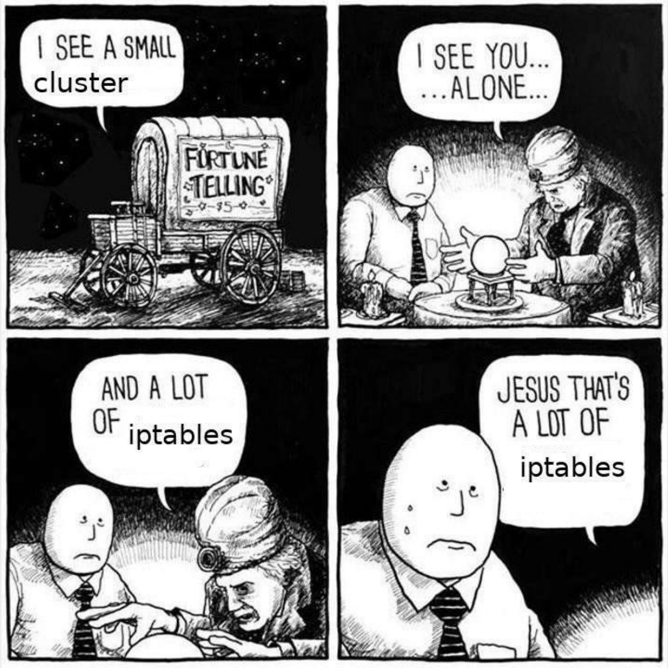
Feb 11, 2019

Debuggability

```
# iptables-save -c

*filter
:INPUT ACCEPT [0:0]
:FORWARD ACCEPT [0:0]
:OUTPUT ACCEPT [0:0]
[1:10] -A FORWARD -i eth0 -s 172.17.0.0/16 -j DROP
```

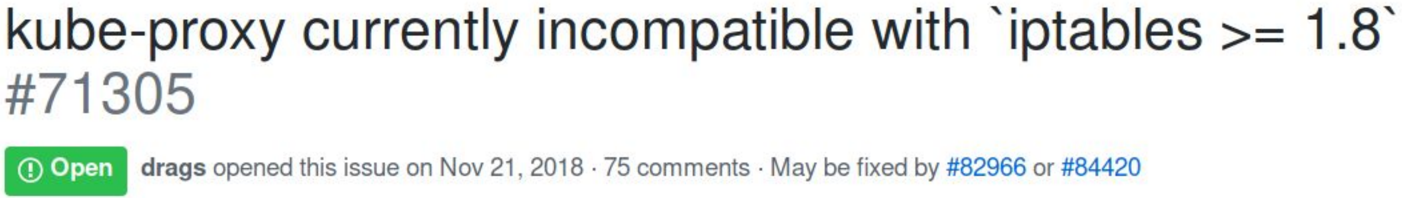
Debuggability

```
:OUTPUT ACCEPT [46709:33506771]
:POSTROUTING ACCEPT [46778:33510020]
COMMIT
*nat
:PREROUTING ACCEPT [0:0]
:INPUT ACCEPT [0:0]
:OUTPUT ACCEPT [31:1905]
:POSTROUTING ACCEPT [21:1305]
:DOCKER - [0:0]
:KUBE-MARK-DROP - [0:0]
:KUBE-MARK-MASQ - [0:0]
:KUBE-NODEPORTS - [0:0]
:KUBE-POSTROUTING - [0:0]
:KUBE-SEP-ARIYJBMSCT6NPKLC - [0:0]
:KUBE-SEP-EVB54GPOXM4P4KYH - [0:0]
:KUBE-SEP-JNEFDVS5622RF3KK - [0:0]
:KUBE-SEP-LHV3DTYFO2UR3QEF - [0:0]
:KUBE-SEP-PVCRDUMNZPYK3THF - [0:0]
:KUBE-SEP-RY4UHCSDDTRJ5BRD - [0:0]
:KUBE-SEP-YQP473NSN3FT53LX - [0:0]
:KUBE-SERVICES - [0:0]
:KUBE-SVC-ERIFXISQEP7F7OF4 - [0:0]
:KUBE-SVC-JD5MR3NA4I4DYORP - [0:0]
:KUBE-SVC-NPX46M4PTMTKRN6Y - [0:0]
:KUBE-SVC-TCOU7JCQXEZGVUNU - [0:0]
-A PREROUTING -m comment --comment "kubernetes service portals" -j KUBE-SERVICES
-A PREROUTING -m addrtype --dst-type LOCAL -j DOCKER
-A OUTPUT -m comment --comment "kubernetes service portals" -j KUBE-SERVICES
-A OUTPUT ! -d 127.0.0.0/8 -m addrtype --dst-type LOCAL -j DOCKER
-A POSTROUTING -m comment --comment "kubernetes postrouting rules" -j KUBE-POSTROUTING
-A POSTROUTING -s 172.17.0.0/16 ! -o docker0 -j MASQUERADE
-A POSTROUTING -s 172.18.0.0/16 ! -o docker_gwbridge -j MASQUERADE
-A DOCKER -i docker0 -j RETURN
-A DOCKER -i docker_gwbridge -j RETURN
-A KUBE-MARK-DROP -j MARK --set-xmark 0x8000/0x8000
-A KUBE-MARK-MASQ -j MARK --set-xmark 0x4000/0x4000
-A KUBE-POSTROUTING -m comment --comment "kubernetes service traffic requiring SNAT" -m mark --mark 0x4000/0x4000 -j MASQUERADE
-A KUBE-SEP-ARIYJBMSCT6NPKLC -s 10.217.0.224/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-ARIYJBMSCT6NPKLC -p tcp -m tcp -j DNAT --to-destination 10.217.0.224:9153
-A KUBE-SEP-EVB54GPOXM4P4KYH -s 10.217.0.71/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-EVB54GPOXM4P4KYH -p tcp -m tcp -j DNAT --to-destination 10.217.0.71:9153
-A KUBE-SEP-JNEFDVS5622RF3KK -s 10.217.0.224/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-JNEFDVS5622RF3KK -p tcp -m tcp -j DNAT --to-destination 10.217.0.224:53
-A KUBE-SEP-LHV3DTYFO2UR3QEF -s 192.168.1.125/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-LHV3DTYFO2UR3QEF -p tcp -m tcp -j DNAT --to-destination 192.168.1.125:6443
-A KUBE-SEP-PVCRDUMNZPYK3THF -s 10.217.0.224/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-PVCRDUMNZPYK3THF -p udp -m udp -j DNAT --to-destination 10.217.0.224:53
-A KUBE-SEP-RY4UHCSDDTRJ5BRD -s 10.217.0.71/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-RY4UHCSDDTRJ5BRD -p tcp -m tcp -j DNAT --to-destination 10.217.0.71:53
-A KUBE-SEP-YQP473NSN3FT53LX -s 10.217.0.71/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-YQP473NSN3FT53LX -p udp -m udp -j DNAT --to-destination 10.217.0.71:53
-A KUBE-SERVICES ! -s 10.217.0.0/16 -d 10.96.0.10/32 -p tcp -m comment --comment "kube-system/kube-dns:dns-tcp cluster IP" -m tcp --dport 53 -j KUBE-MARK-MASQ
-A KUBE-SERVICES -d 10.96.0.10/32 -p tcp -m comment --comment "kube-system/kube-dns:dns-tcp cluster IP" -m tcp --dport 53 -j KUBE-SVC-ERIFXISQEP7F7OF4
-A KUBE-SERVICES ! -s 10.217.0.0/16 -d 10.96.0.10/32 -p tcp -m comment --comment "kube-system/kube-dns:metrics cluster IP" -m tcp --dport 9153 -j KUBE-MARK-MASQ
-A KUBE-SERVICES -d 10.96.0.10/32 -p tcp -m comment --comment "kube-system/kube-dns:metrics cluster IP" -m tcp --dport 9153 -j KUBE-SVC-JD5MR3NA4I4DYORP
-A KUBE-SERVICES ! -s 10.217.0.0/16 -d 10.96.0.1/32 -p tcp -m comment --comment "default/kubernetes:https cluster IP" -m tcp --dport 443 -j KUBE-MARK-MASQ
-A KUBE-SERVICES -d 10.96.0.1/32 -p tcp -m comment --comment "default/kubernetes:https cluster IP" -m tcp --dport 443 -j KUBE-SVC-NPX46M4PTMTKRN6Y
-A KUBE-SERVICES ! -s 10.217.0.0/16 -d 10.96.0.10/32 -p udp -m comment --comment "kube-system/kube-dns:dns cluster IP" -m udp --dport 53 -j KUBE-MARK-MASQ
-A KUBE-SERVICES -d 10.96.0.10/32 -p udp -m comment --comment "kube-system/kube-dns:dns cluster IP" -m udp --dport 53 -j KUBE-SVC-TCOU7JCQXEZGVUNU
-A KUBE-SERVICES -m comment --comment "kubernetes service nodeports; NOTE: this must be the last rule in this chain" -m addrtype --dst-type LOCAL -j KUBE-NODEPORTS
-A KUBE-SVC-ERIFXISQEP7F7OF4 -m statistic --mode random --probability 0.50000000000 -j KUBE-SEP-JNEFDVS5622RF3KK
-A KUBE-SVC-ERIFXISQEP7F7OF4 -j KUBE-SEP-RY4UHCSDDTRJ5BRD
-A KUBE-SVC-JD5MR3NA4I4DYORP -m statistic --mode random --probability 0.50000000000 -j KUBE-SEP-ARIYJBMSCT6NPKLC
-A KUBE-SVC-JD5MR3NA4I4DYORP -j KUBE-SEP-EVB54GPOXM4P4KYH
-A KUBE-SVC-NPX46M4PTMTKRN6Y -j KUBE-SEP-LHV3DTYFO2UR3QEF
-A KUBE-SVC-TCOU7JCQXEZGVUNU -m statistic --mode random --probability 0.50000000000 -j KUBE-SEP-PVCRDUMNZPYK3THF
-A KUBE-SVC-TCOU7JCQXEZGVUNU -j KUBE-SEP-YQP473NSN3FT53LX
COMMIT
*filter
:INPUT ACCEPT [2938:623620]
:FORWARD DROP [0:0]
:OUTPUT ACCEPT [2893:671491]
:DOCKER - [0:0]
:DOCKER-ISOLATION-STAGE-1 - [0:0]
:DOCKER-ISOLATION-STAGE-2 - [0:0]
:DOCKER-USER - [0:0]
:KUBE-EXTERNAL-SERVICES - [0:0]
:KUBE-FIREWALL - [0:0]
:KUBE-FORWARD - [0:0]
:KUBE-SERVICES - [0:0]
-A INPUT -m conntrack --ctstate NEW -m comment --comment "kubernetes service portals" -j KUBE-SERVICES
-A INPUT -m conntrack --ctstate NEW -m comment --comment "kubernetes externally-visible service portals" -j KUBE-EXTERNAL-SERVICES
-A INPUT -j KUBE-FIREWALL
-A FORWARD -m comment --comment "kubernetes forwarding rules" -j KUBE-FORWARD
-A FORWARD -m conntrack --ctstate NEW -m comment --comment "kubernetes service portals" -j KUBE-SERVICES
-A FORWARD -j DOCKER-USER
-A FORWARD -j DOCKER-ISOLATION-STAGE-1
-A FORWARD -o docker0 -m conntrack --ctstate RELATED,ESTABLISHED -j ACCEPT
-A FORWARD -o docker0 -j DOCKER
-A FORWARD -i docker0 ! -o docker0 -j ACCEPT
-A FORWARD -i docker0 -o docker0 -j ACCEPT
-A FORWARD -o docker_gwbridge -m conntrack --ctstate RELATED,ESTABLISHED -j ACCEPT
-A FORWARD -o docker_gwbridge -j DOCKER
-A FORWARD -i docker_gwbridge ! -o docker_gwbridge -j ACCEPT
-A FORWARD -i docker_gwbridge -o docker_gwbridge -j DROP
-A OUTPUT -m conntrack --ctstate NEW -m comment --comment "kubernetes service portals" -j KUBE-SERVICES
-A OUTPUT -j KUBE-FIREWALL
-A DOCKER-ISOLATION-STAGE-1 -i docker0 ! -o docker0 -j DOCKER-ISOLATION-STAGE-2
-A DOCKER-ISOLATION-STAGE-1 -i docker_gwbridge ! -o docker_gwbridge -j DOCKER-ISOLATION-STAGE-2
-A DOCKER-ISOLATION-STAGE-1 -j RETURN
-A DOCKER-ISOLATION-STAGE-2 -o docker0 -j DROP
-A DOCKER-ISOLATION-STAGE-2 -o docker_gwbridge -j DROP
-A DOCKER-ISOLATION-STAGE-2 -j RETURN
-A DOCKER-USER -j RETURN
-A KUBE-FIREWALL -m comment --comment "kubernetes firewall for dropping marked packets" -m mark --mark 0x8000/0x8000 -j DROP
-A KUBE-FORWARD -m conntrack --ctstate INVALID -j DROP
-A KUBE-FORWARD -m comment --comment "kubernetes forwarding rules" -m mark --mark 0x4000/0x4000 -j ACCEPT
-A KUBE-FORWARD -s 10.217.0.0/16 -m comment --comment "kubernetes forwarding conntrack pod source rule" -m conntrack --ctstate RELATED,ESTABLISHED -j ACCEPT
-A KUBE-FORWARD -d 10.217.0.0/16 -m comment --comment "kubernetes forwarding conntrack pod destination rule" -m conntrack --ctstate RELATED,ESTABLISHED -j ACCEPT
-A KUBE-SERVICES -d 10.99.38.155/32 -p tcp -m comment --comment "default/nginx-59: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.96.61.252/32 -p tcp -m comment --comment "default/nginx-64: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.104.166.10/32 -p tcp -m comment --comment "default/nginx-67: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.98.85.41/32 -p tcp -m comment --comment "default/nginx-9: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.97.138.144/32 -p tcp -m comment --comment "default/nginx-17: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.106.49.80/32 -p tcp -m comment --comment "default/nginx-37: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.104.164.205/32 -p tcp -m comment --comment "default/nginx-5: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.104.25.150/32 -p tcp -m comment --comment "default/nginx-19: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.106.234.213/32 -p tcp -m comment --comment "default/nginx-88: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.109.209.136/32 -p tcp -m comment --comment "default/nginx-33: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.106.196.105/32 -p tcp -m comment --comment "default/nginx-49: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.111.101.6/32 -p tcp -m comment --comment "default/nginx-53: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.110.226.230/32 -p tcp -m comment --comment "default/nginx-79: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.98.99.136/32 -p tcp -m comment --comment "default/nginx-6: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.99.75.233/32 -p tcp -m comment --comment "default/nginx-7: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.108.41.202/32 -p tcp -m comment --comment "default/nginx-14: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.97.36.249/32 -p tcp -m comment --comment "default/nginx-99: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.98.213.37/32 -p tcp -m comment --comment "default/nginx-77: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.107.229.31/32 -p tcp -m comment --comment "default/nginx-92: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.98.64.251/32 -p tcp -m comment --comment "default/nginx-16: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.101.88.159/32 -p tcp -m comment --comment "default/nginx-31: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.105.71.74/32 -p tcp -m comment --comment "default/nginx-41: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.108.92.226/32 -p tcp -m comment --comment "default/nginx-63: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.109.252.234/32 -p tcp -m comment --comment "default/nginx-18: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.104.118.66/32 -p tcp -m comment --comment "default/nginx-30: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.106.224.55/32 -p tcp -m comment --comment "default/nginx-83: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.109.16.199/32 -p tcp -m comment --comment "default/nginx-100: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.109.231.213/32 -p tcp -m comment --comment "default/nginx-61: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.98.27.250/32 -p tcp -m comment --comment "default/nginx-95: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.105.42.108/32 -p tcp -m comment --comment "default/nginx-12: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.99.35.236/32 -p tcp -m comment --comment "default/nginx-20: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.111.42.123/32 -p tcp -m comment --comment "default/nginx-21: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.99.47.225/32 -p tcp -m comment --comment "default/nginx-22: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.104.184.242/32 -p tcp -m comment --comment "default/nginx-51: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.98.77.93/32 -p tcp -m comment --comment "default/nginx-68: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.110.169.113/32 -p tcp -m comment --comment "default/nginx-72: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.100.231.169/32 -p tcp -m comment --comment "default/nginx-90: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.96.58.51/32 -p tcp -m comment --comment "default/nginx-4: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.101.132.61/32 -p tcp -m comment --comment "default/nginx-25: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.100.64.242/32 -p tcp -m comment --comment "default/nginx-39: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.111.154.81/32 -p tcp -m comment --comment "default/nginx-50: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.100.179.151/32 -p tcp -m comment --comment "default/nginx-96: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.111.69.30/32 -p tcp -m comment --comment "default/nginx-35: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.96.35.212/32 -p tcp -m comment --comment "default/nginx-38: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.96.61.26/32 -p tcp -m comment --comment "default/nginx-84: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.96.229.244/32 -p tcp -m comment --comment "default/nginx-87: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.104.247.138/32 -p tcp -m comment --comment "default/nginx-66: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.96.214.153/32 -p tcp -m comment --comment "default/nginx-11: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.102.208.205/32 -p tcp -m comment --comment "default/nginx-55: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.106.35.32/32 -p tcp -m comment --comment "default/nginx-58: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.107.174.56/32 -p tcp -m comment --comment "default/nginx-65: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.98.142.83/32 -p tcp -m comment --comment "default/nginx-2: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.106.248.222/32 -p tcp -m comment --comment "default/nginx-15: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.98.202.86/32 -p tcp -m comment --comment "default/nginx-34: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.96.57.213/32 -p tcp -m comment --comment "default/nginx-71: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.98.33.199/32 -p tcp -m comment --comment "default/nginx-69: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.101.93.81/32 -p tcp -m comment --comment "default/nginx-75: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.99.199.226/32 -p tcp -m comment --comment "default/nginx-78: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.103.122.17/32 -p tcp -m comment --comment "default/nginx-10: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.103.194.216/32 -p tcp -m comment --comment "default/nginx-27: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.97.117.130/32 -p tcp -m comment --comment "default/nginx-32: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.98.254.254/32 -p tcp -m comment --comment "default/nginx-56: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.100.164.89/32 -p tcp -m comment --comment "default/nginx-29: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.106.187.33/32 -p tcp -m comment --comment "default/nginx-42: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.111.68.111/32 -p tcp -m comment --comment "default/nginx-44: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.97.54.135/32 -p tcp -m comment --comment "default/nginx-46: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.106.128.46/32 -p tcp -m comment --comment "default/nginx-13: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.108.223.155/32 -p tcp -m comment --comment "default/nginx-26: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.108.101.195/32 -p tcp -m comment --comment "default/nginx-62: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.102.124.200/32 -p tcp -m comment --comment "default/nginx-73: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
-A KUBE-SERVICES -d 10.101.141.155/32 -p tcp -m comment --comment "default/nginx-93: has no endpoints" -m tcp --dport 80 -j REJECT --reject-with icmp-port-unreachable
```



https://www.reddit.com/r/networkingmemes/comments/8u7jyz/container_networking/

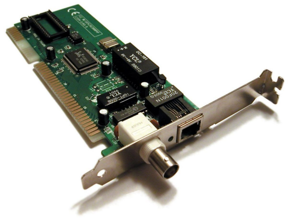
Compatibility

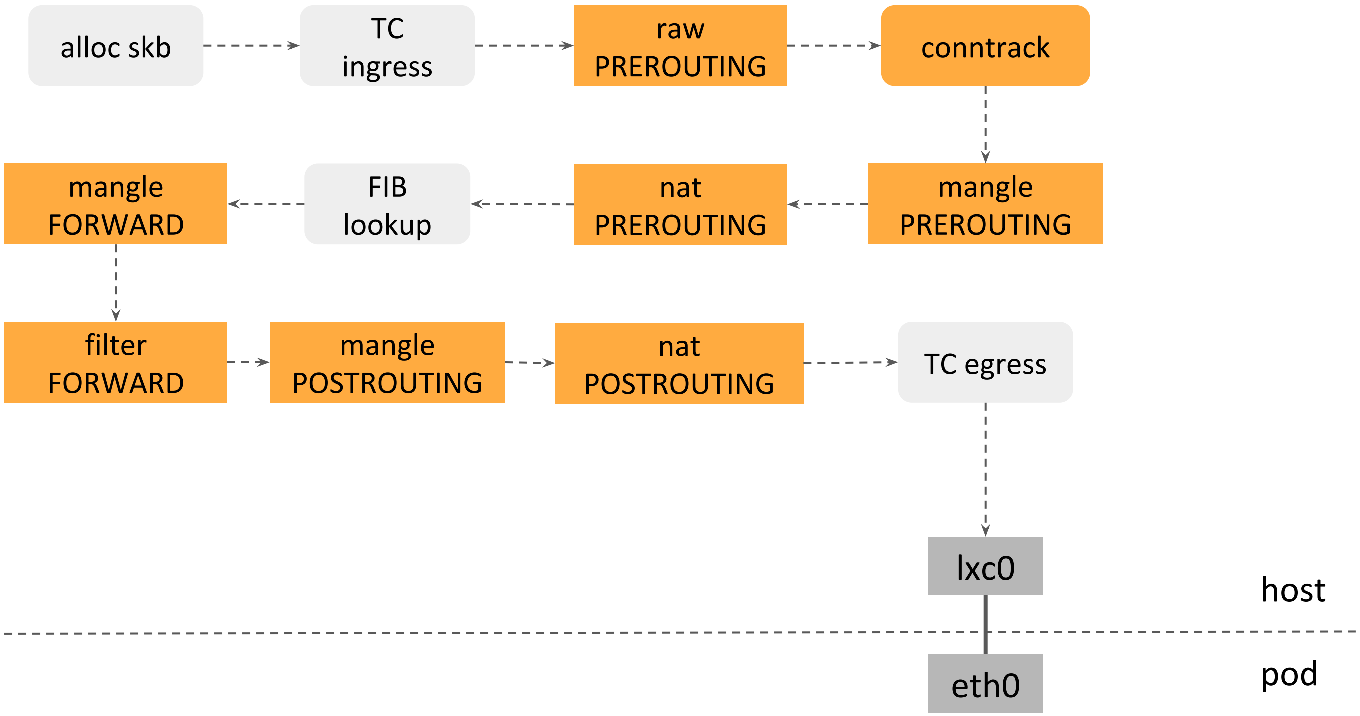


kube-proxy currently incompatible with `iptables >= 1.8` #71305

drags opened this issue on Nov 21, 2018 · 75 comments · May be fixed by #82966 or #84420

Packet flow





alloc skb

TC ingress

raw PREROUTING

conntrack

mangle PREROUTING

nat PREROUTING

FIB lookup

mangle FORWARD

filter FORWARD

mangle POSTROUTING

nat POSTROUTING

TC egress

lxc0

eth0

host

pod

CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=122201

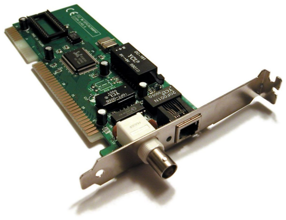
ClusterIP with iptables

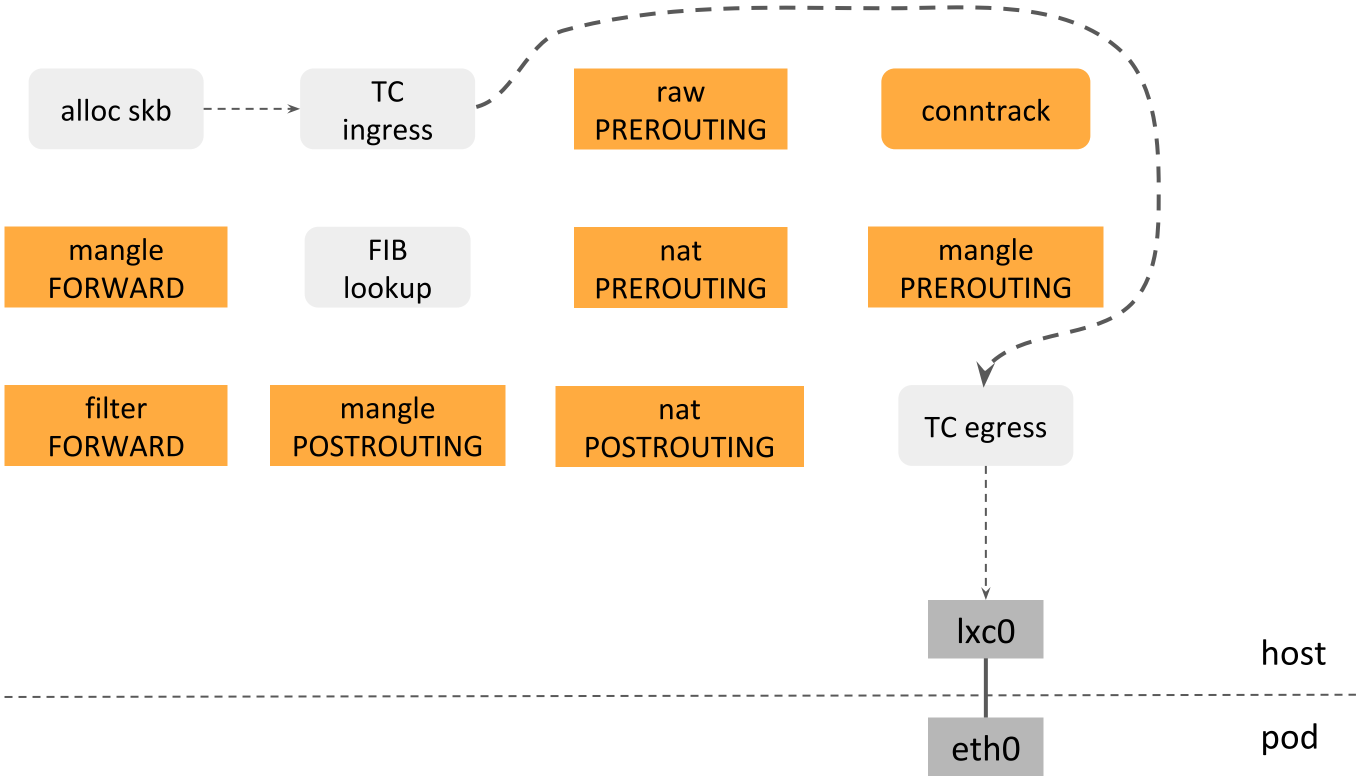
```
$ kubectl get svc nginx
NAME   TYPE        CLUSTER-IP   EXTERNAL-IP   PORT(S)
nginx  ClusterIP   3.3.3.3      <none>        80/TCP

$ kubectl get endpoints nginx
NAME   ENDPOINTS
nginx  1.1.1.1:80, 1.1.2.2:80
```

```
-t nat -A PREROUTING -m conntrack --ctstate NEW -j KUBE-SERVICES

-A KUBE-SERVICES ! -s 1.1.0.0/16 -d 3.3.3.3/32 -p tcp -m tcp --dport 80 -j KUBE-MARK-MASQ
-A KUBE-SERVICES -d 3.3.3.3/32 -p tcp -m tcp --dport 80 -j KUBE-SVC-NGINX

-A KUBE-SVC-NGINX -m statistic --mode random --probability 0.50 -j KUBE-SEP-NGINX1
-A KUBE-SVC-NGINX -j KUBE-SEP-NGINX2

-A KUBE-SEP-NGINX1 -s 1.1.1.1/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-NGINX1 -p tcp -m tcp -j DNAT --to-destination 1.1.1.1:80
-A KUBE-SEP-NGINX2 -s 1.1.2.2/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-NGINX2 -p tcp -m tcp -j DNAT --to-destination 1.1.2.2:80
```

nat
PREROUTING

Packet flow





alloc skb

TC ingress

raw PREROUTING

conntrack

mangle FORWARD

FIB lookup

nat PREROUTING

mangle PREROUTING

filter FORWARD

mangle POSTROUTING

nat POSTROUTING

TC egress

lxc0

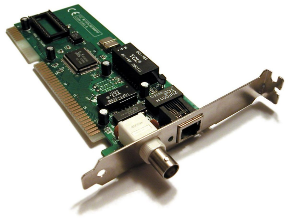
eth0

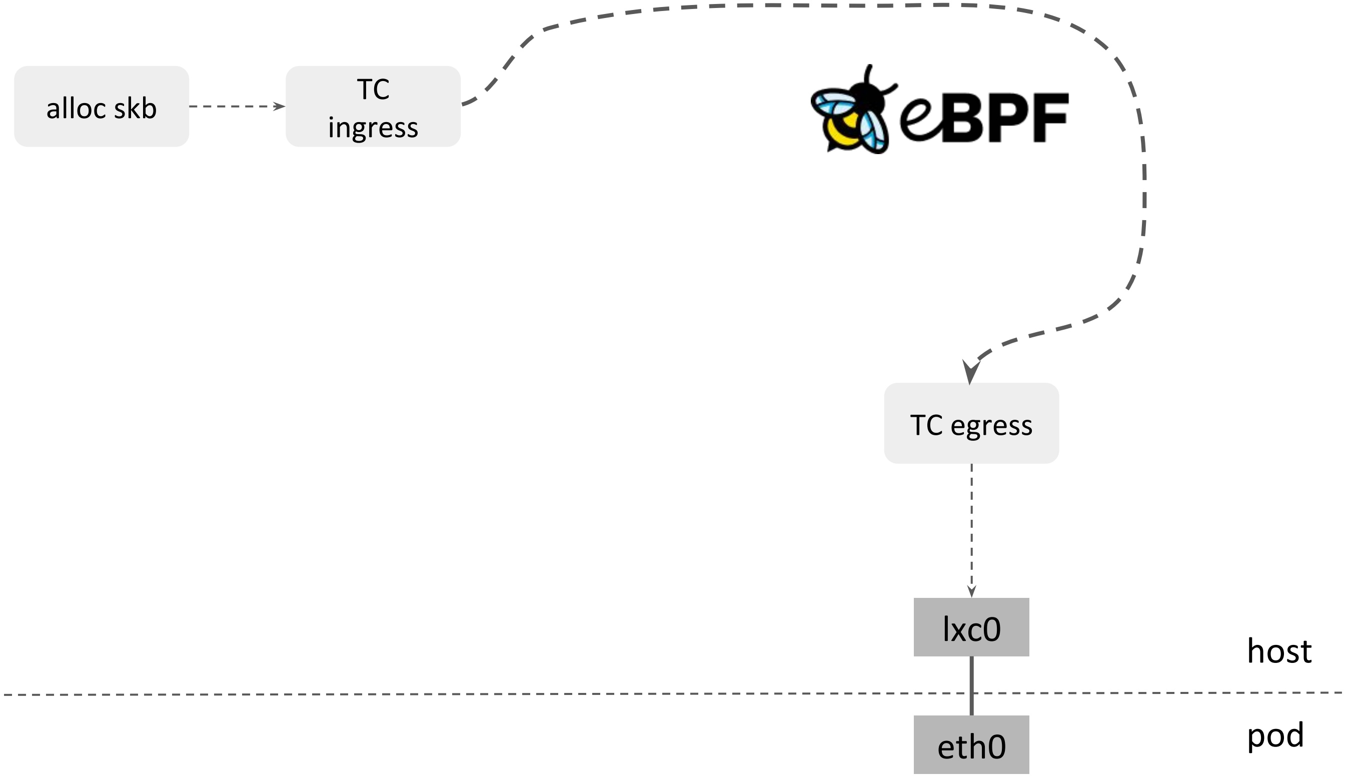
host

pod

CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=122201

Packet flow





alloc skb

TC ingress

eBPF

TC egress

lxc0

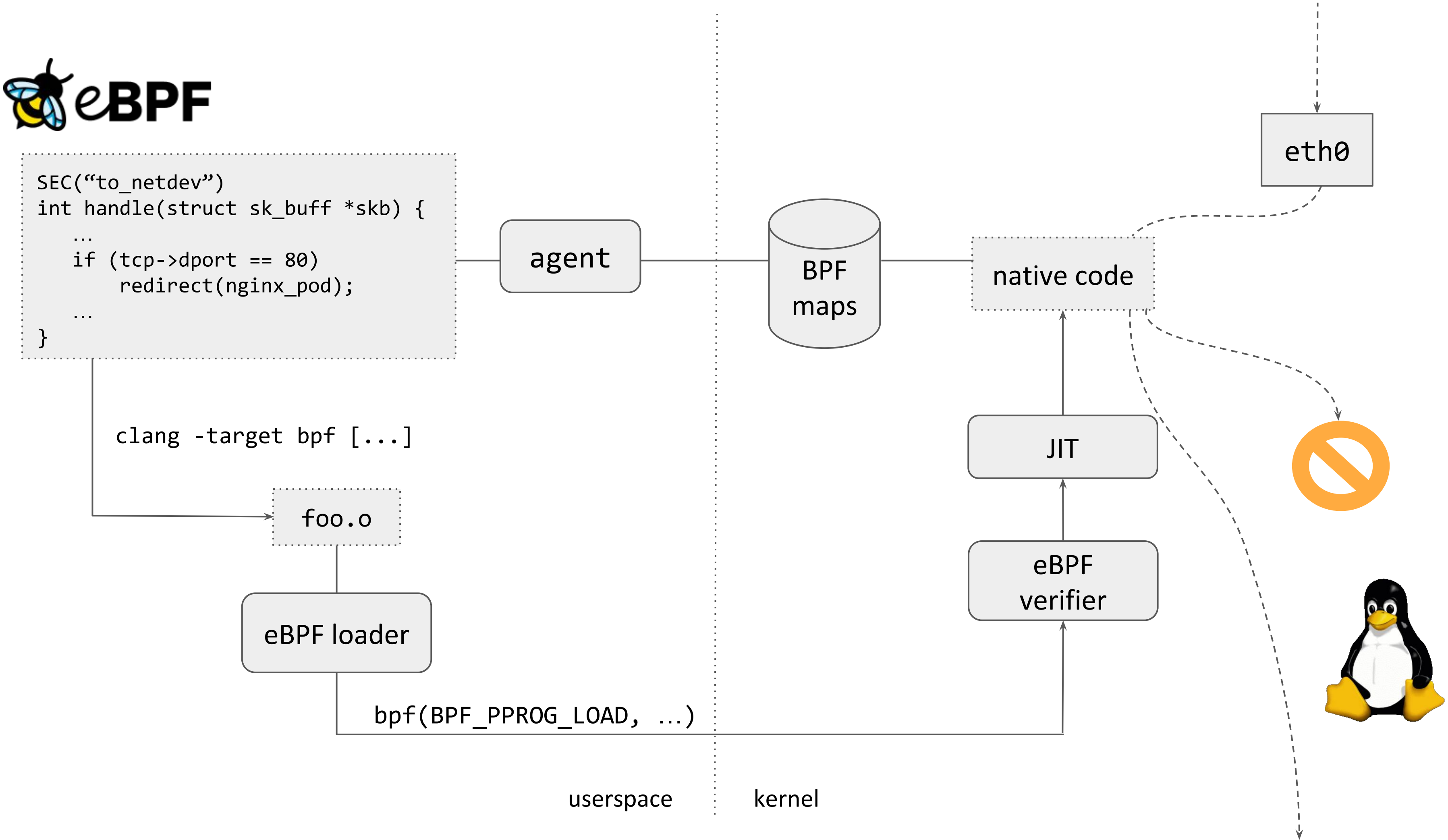
eth0

host

pod

CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=122201

eBPF



```
SEC(“to_netdev”)
int handle(struct sk_buff *skb) {
   …
   if (tcp->dport == 80)
       redirect(nginx_pod);
   …
}
```

```
clang -target bpf [...]
```

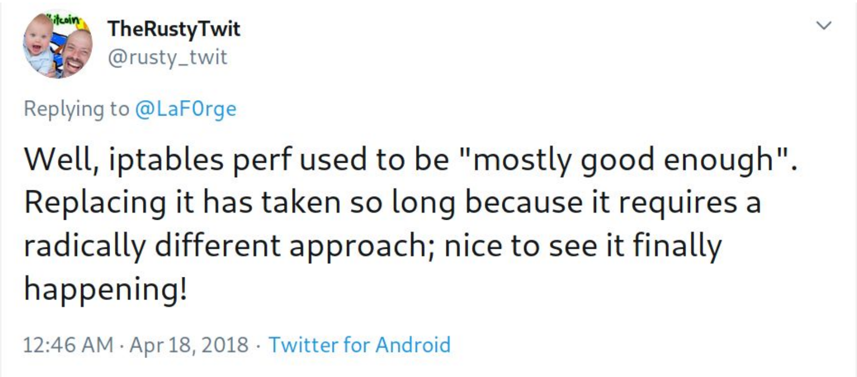
```
bpf(BPF_PPROG_LOAD, …)
```

eBPF

268 contributors (Jan 2016 to Nov 2019):

- 443 Daniel Borkmann (Cilium; maintainer)
- 242 Alexei Starovoitov (Facebook; maintainer)
- 210 Jakub Kicinski (Netronome)
- 195 Andrii Nakryiko (Facebook)
- 161 Yonghong Song (Facebook)
- 151 Stanislav Fomichev (Google)
- 145 Quentin Monnet (Netronome)
- 144 Martin KaFai Lau (Facebook)
- 139 John Fastabend (Cilium)
- 118 Jesper Dangaard Brouer (Red Hat)
- [...]

Users:



TheRustyTwit
@rusty_twit

Replying to @LaF0rge

Well, iptables perf used to be "mostly good enough". Replacing it has taken so long because it requires a radically different approach; nice to see it finally happening!

12:46 AM · Apr 18, 2018 · Twitter for Android

```
$ kubectl -n kube-system delete ds kube-proxy
```

kube-proxy

1. ClusterIP

- In-cluster access via virtual IP

2. NodePort

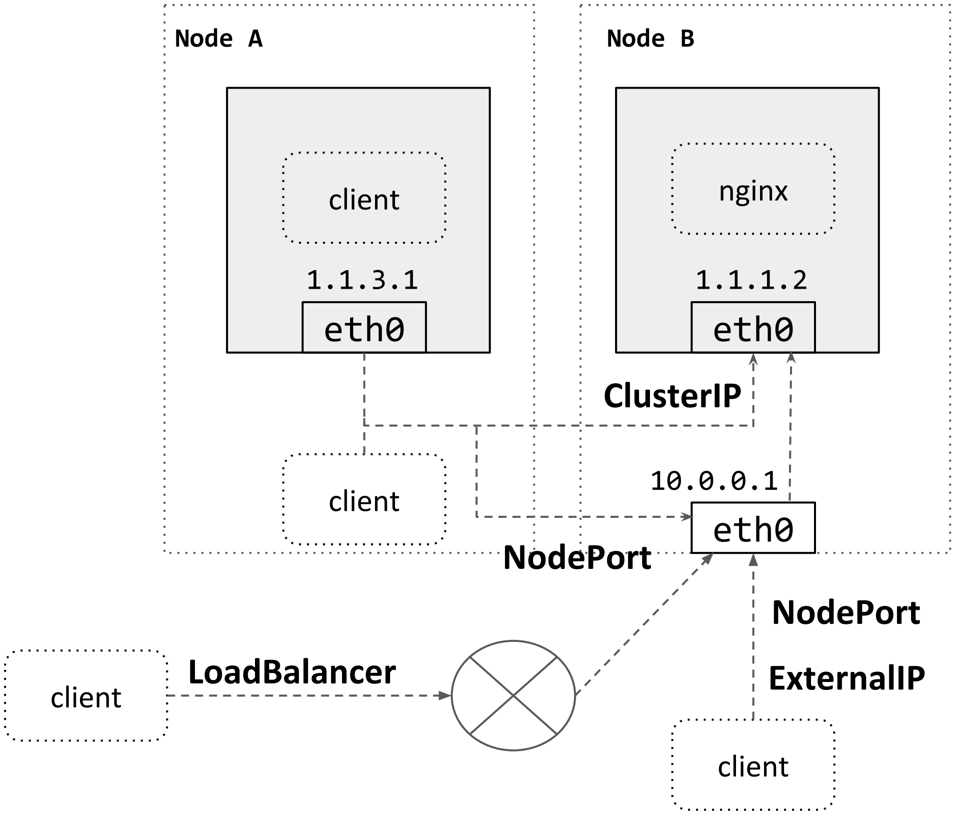
- Access from outside / inside via node IP + port

3. ExternalIP

- Access from outside via external IP

4. LoadBalancer

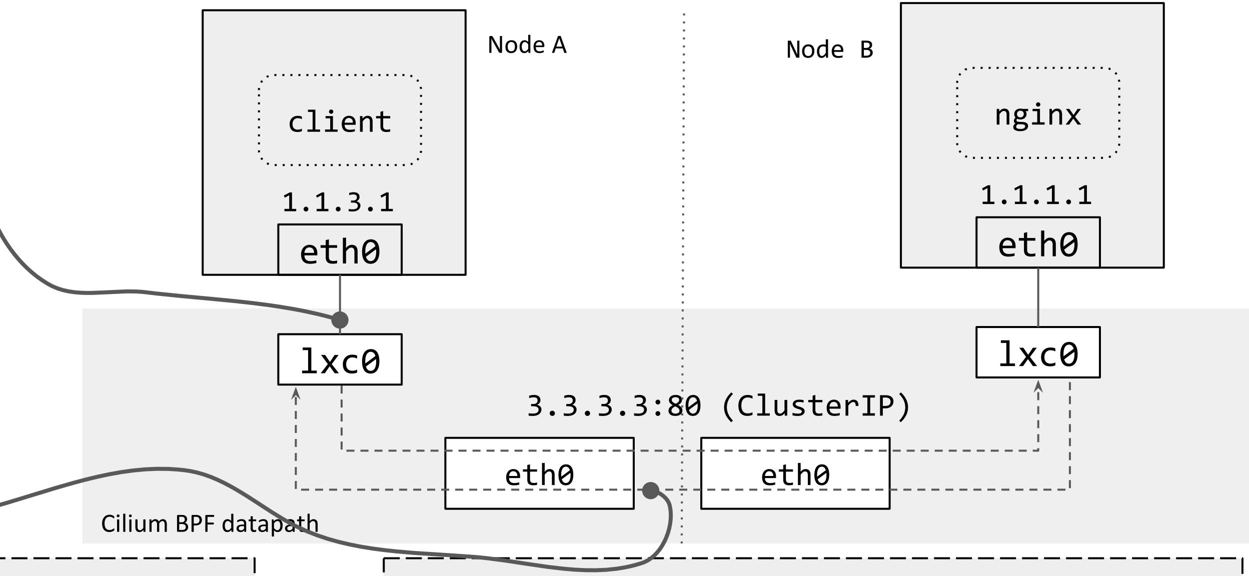
- Access from outside via external LB



ClusterIP (pod to pod) in Cilium

1. Lookup dst in SVC map
2. If found:
   a. Create SVC CT
   b. DNAT
3. Create Egress CT

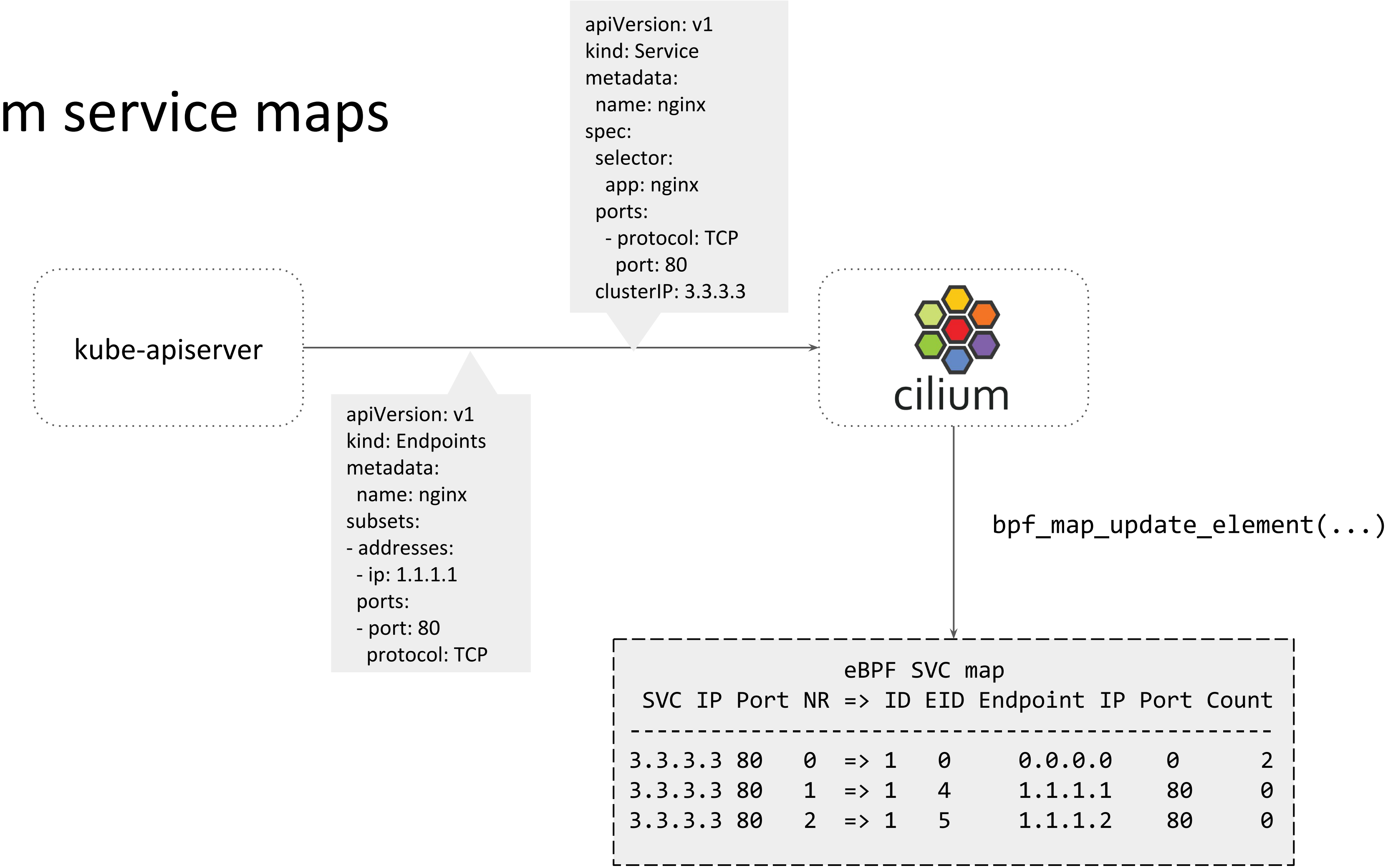
1. Lookup Egress CT
2. If found:
   a. Rev-NAT xlation
3. Redirect to lxc0



```
                 eBPF SVC map
 SVC IP Port NR => ID EID Endpoint IP Port Count
-----------------------------------------------
3.3.3.3 80   0  => 1   0     0.0.0.0    0      2
3.3.3.3 80   1  => 1   4     1.1.1.1    80     0
3.3.3.3 80   2  => 1   5     1.1.1.2    80     0
```

```
                eBPF conntrack LRU map
 srcIP   sPort  dstIP   dPort   Type    => EID|SVCID
-------------------------------------------------
1.1.3.1 4321  3.3.3.3  80     SVC     =>  4
1.1.3.1 4321  1.1.1.1  80     Egress  =>       1
1.1.1.1 80    1.1.3.1  4321   Ingress =>
```

Cilium service maps



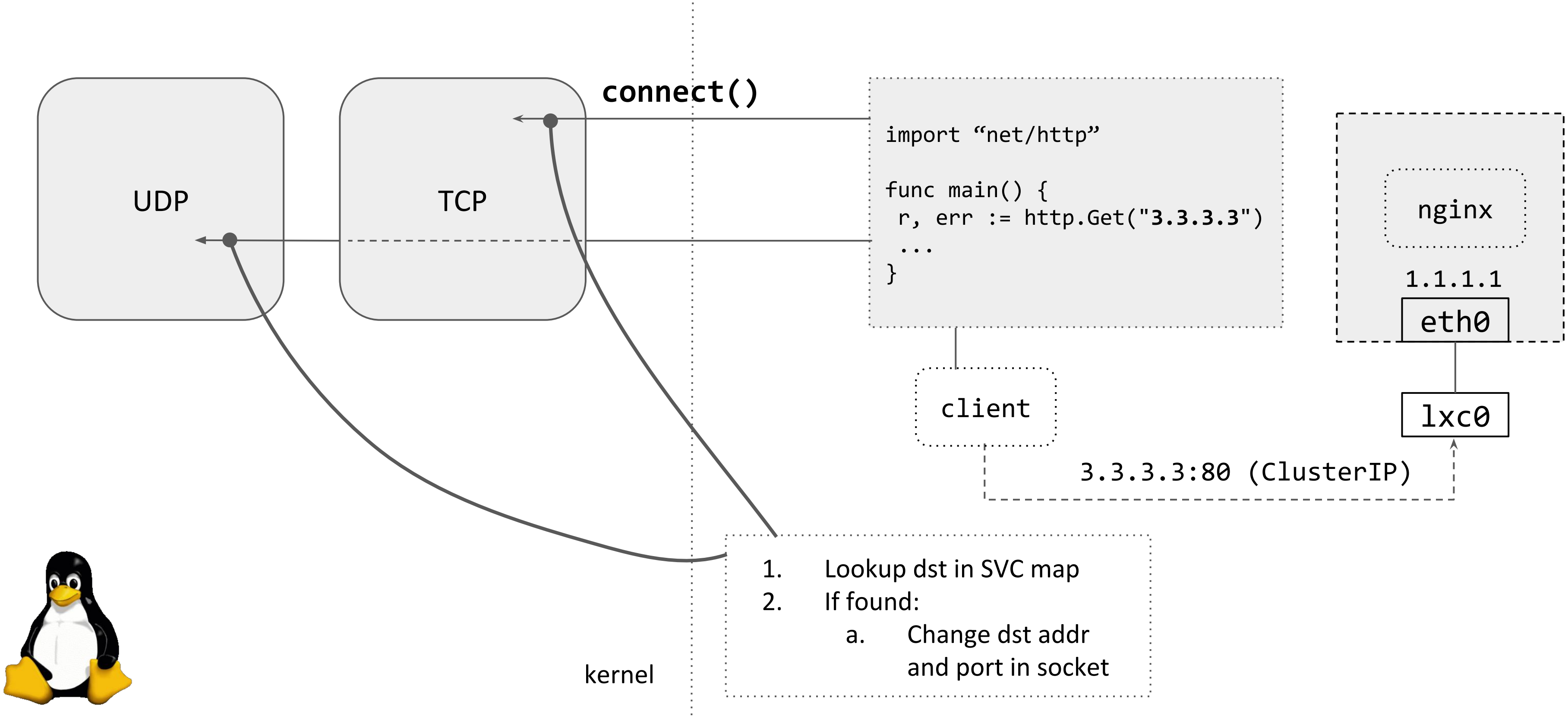
```
apiVersion: v1
kind: Service
metadata:
  name: nginx
spec:
  selector:
    app: nginx
  ports:
    - protocol: TCP
      port: 80
  clusterIP: 3.3.3.3
```

```
apiVersion: v1
kind: Endpoints
metadata:
  name: nginx
subsets:
- addresses:
  - ip: 1.1.1.1
  ports:
  - port: 80
    protocol: TCP
```

`bpf_map_update_element(...)`

```
                   eBPF SVC map
 SVC IP Port NR => ID EID Endpoint IP Port Count
------------------------------------------------
3.3.3.3 80   0  => 1  0    0.0.0.0    0      2
3.3.3.3 80   1  => 1  4    1.1.1.1    80     0
3.3.3.3 80   2  => 1  5    1.1.1.2    80     0
```

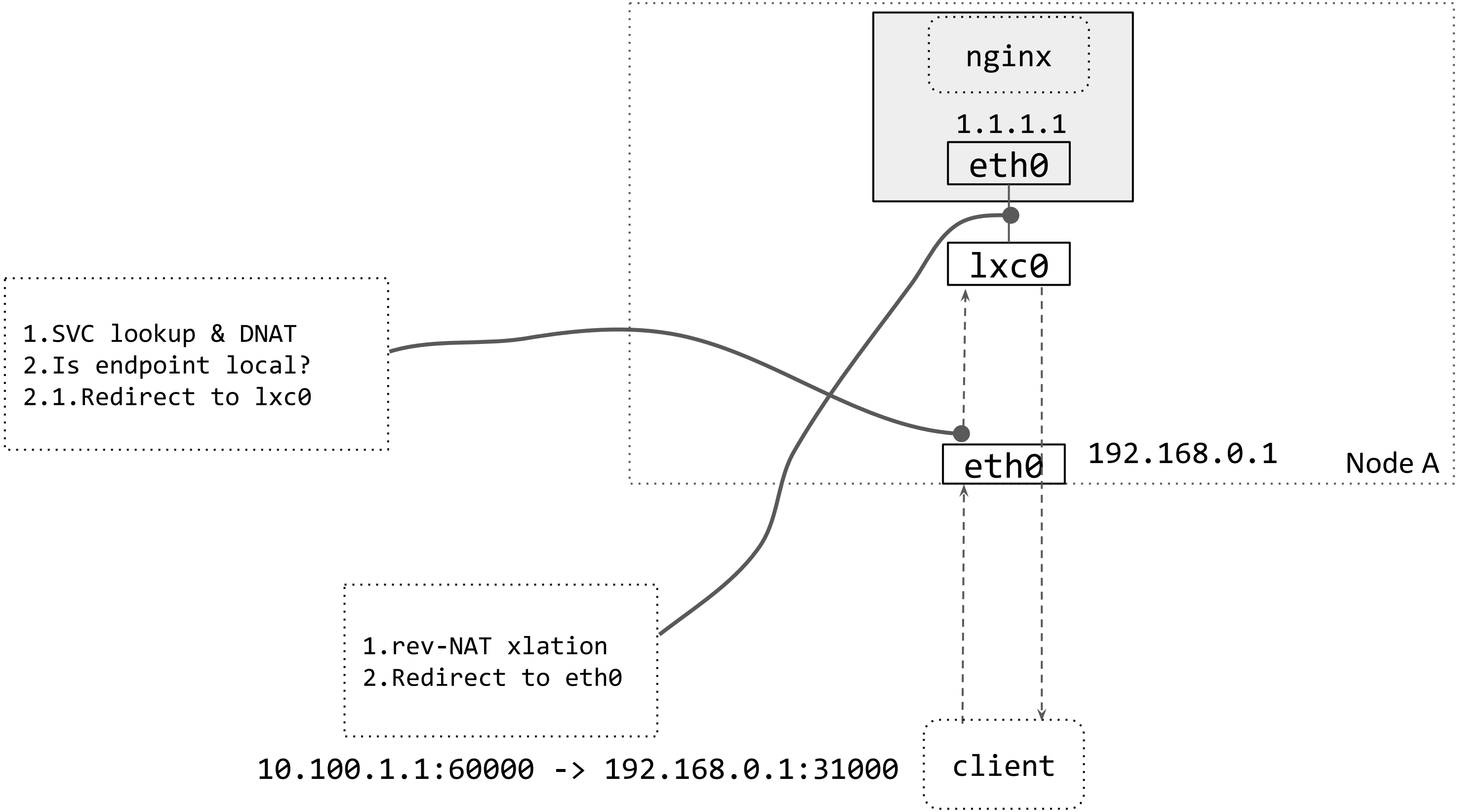
ClusterIP (host or pod to pod) in Cilium



```
import “net/http”

func main() {
 r, err := http.Get("3.3.3.3")
 ...
}
```

1. Lookup dst in SVC map
2. If found:
    a. Change dst addr and port in socket

NodePort with service endpoint on local node in Cilium

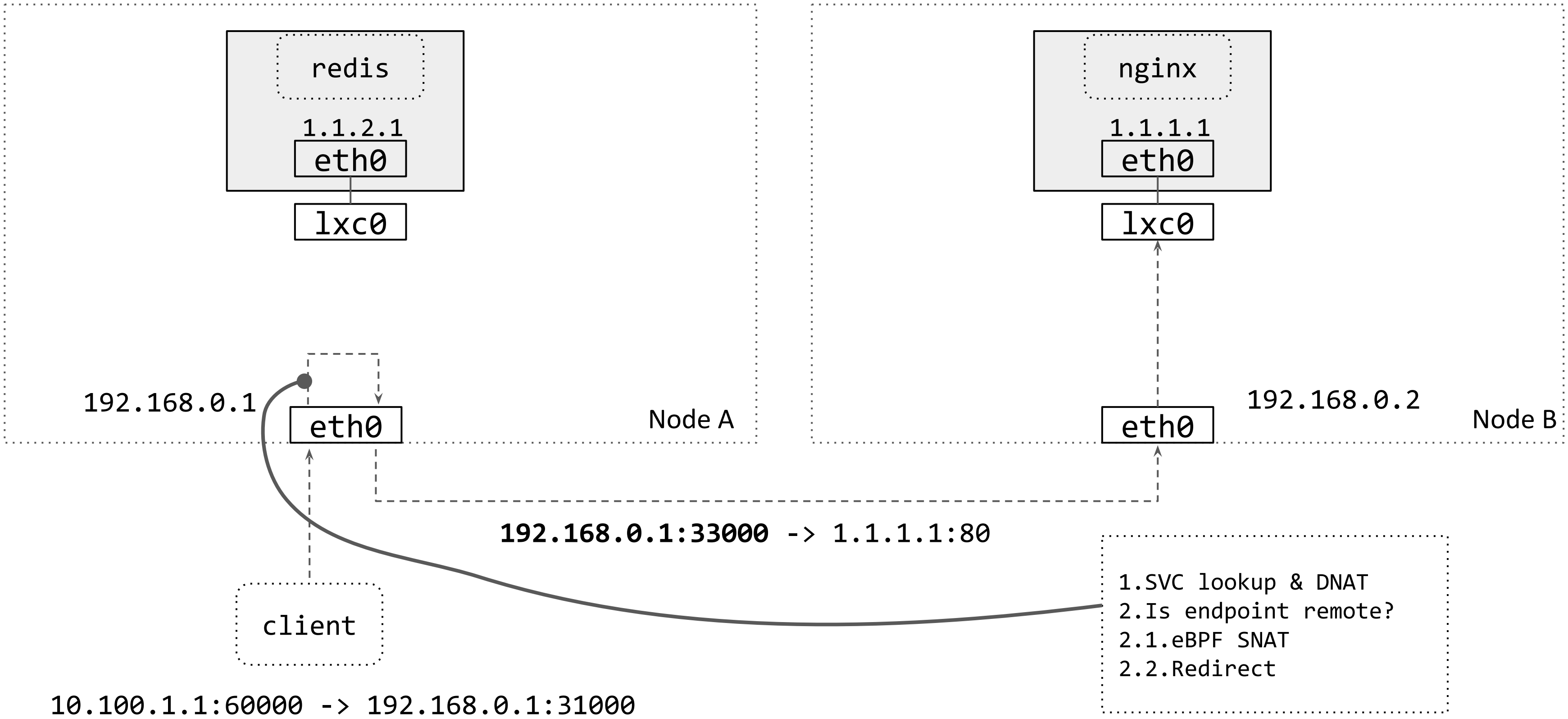


1.SVC lookup & DNAT
2.Is endpoint local?
2.1.Redirect to lxc0

1.rev-NAT xlation
2.Redirect to eth0

10.100.1.1:60000 -> 192.168.0.1:31000

NodePort with service endpoint on remote node in Cilium

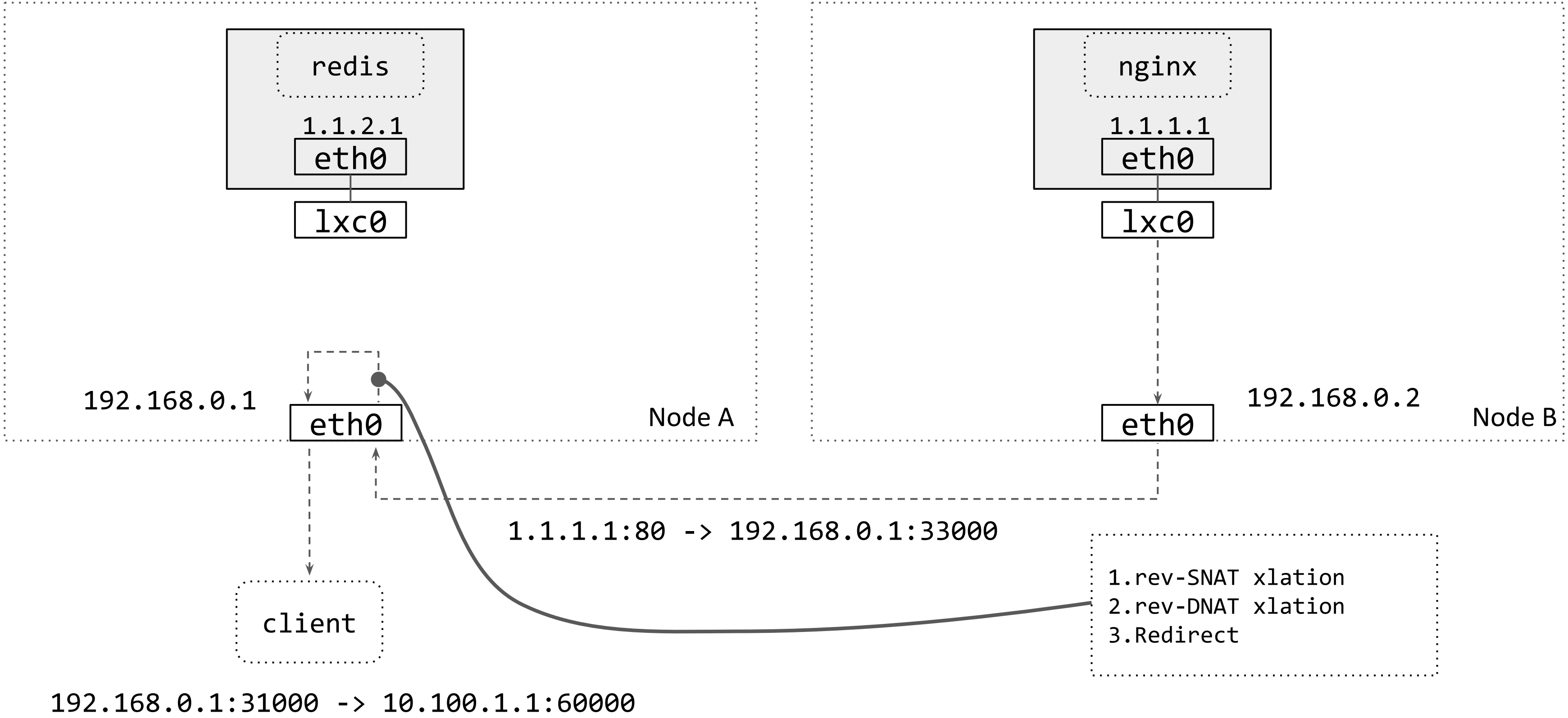


192.168.0.1:33000 -> 1.1.1.1:80

1.SVC lookup & DNAT
2.Is endpoint remote?
2.1.eBPF SNAT
2.2.Redirect

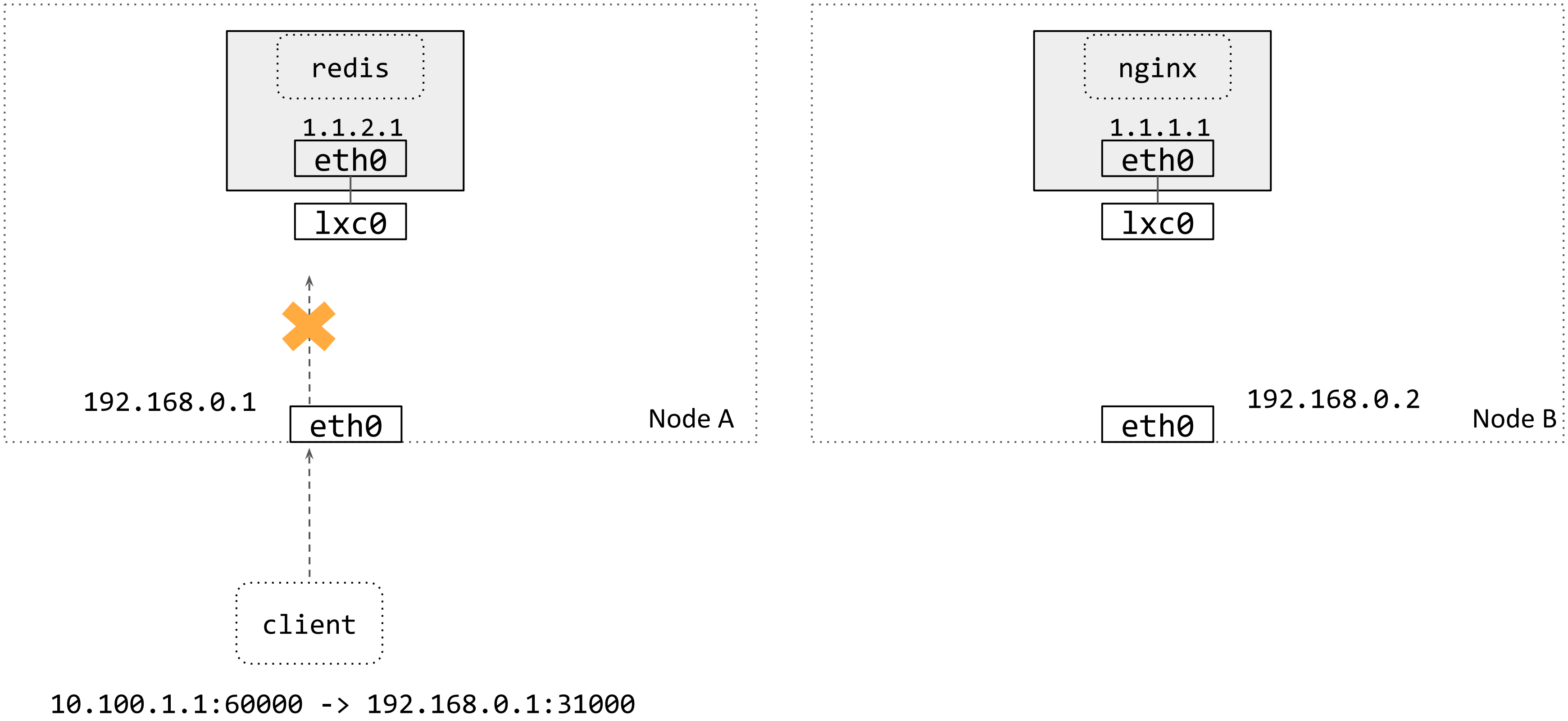
10.100.1.1:60000 -> 192.168.0.1:31000

NodePort with service endpoint on remote node in Cilium

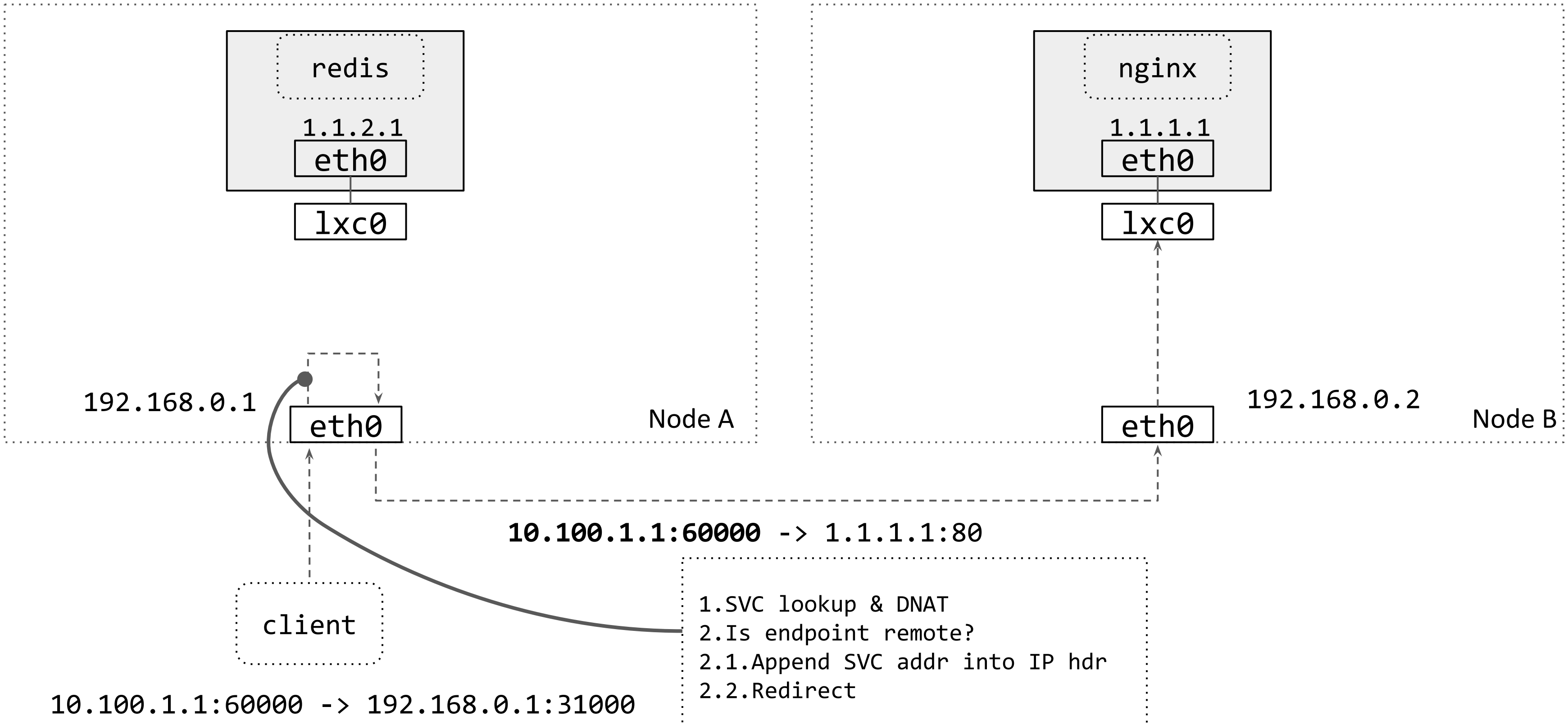


1.rev-SNAT xlation
2.rev-DNAT xlation
3.Redirect

NodePort externalTrafficPolicy=Local

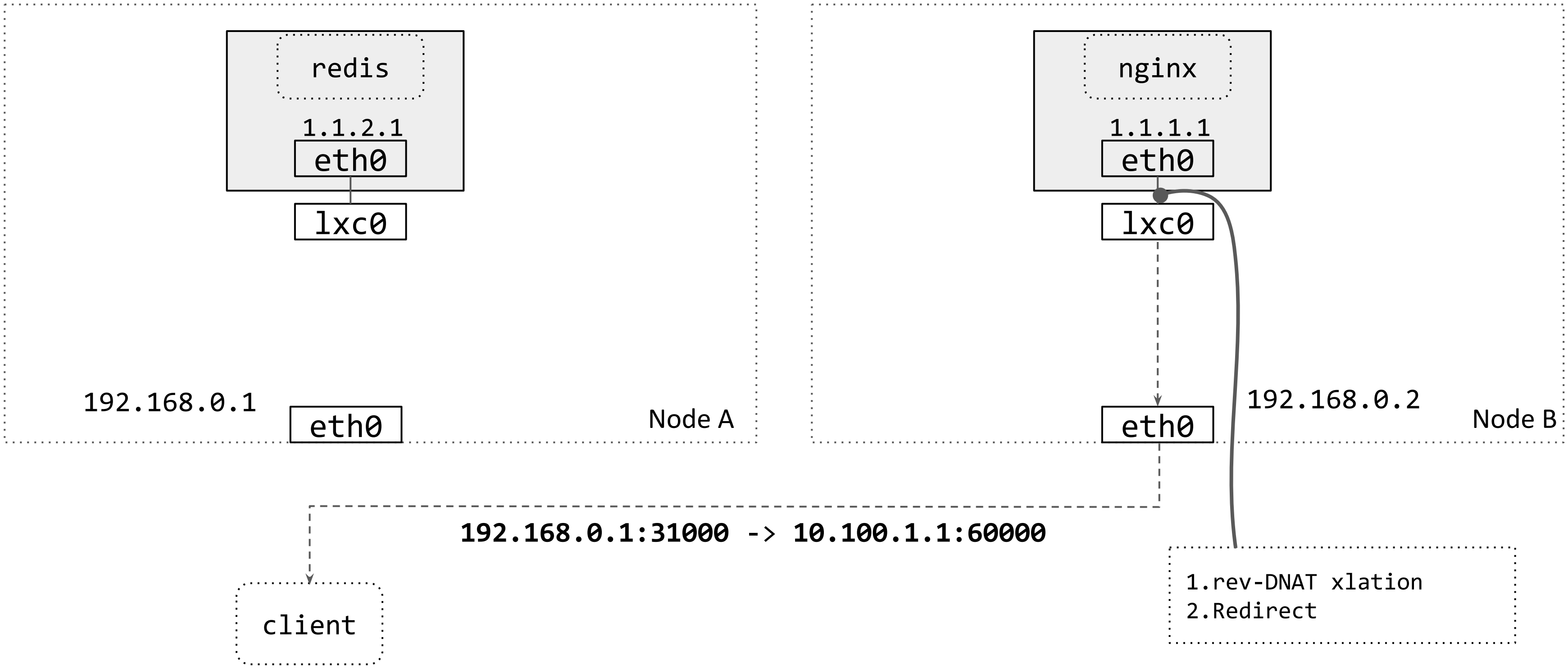


NodePort (DSR) in Cilium

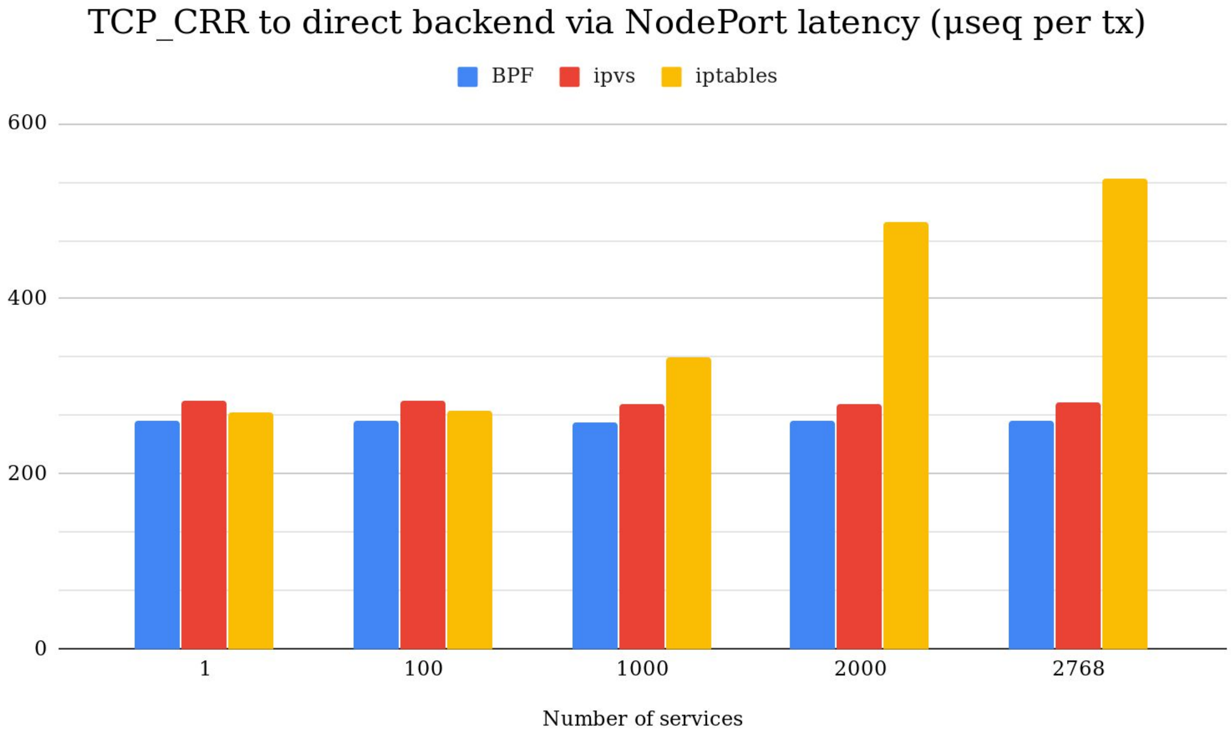


1.SVC lookup & DNAT
2.Is endpoint remote?
2.1.Append SVC addr into IP hdr
2.2.Redirect

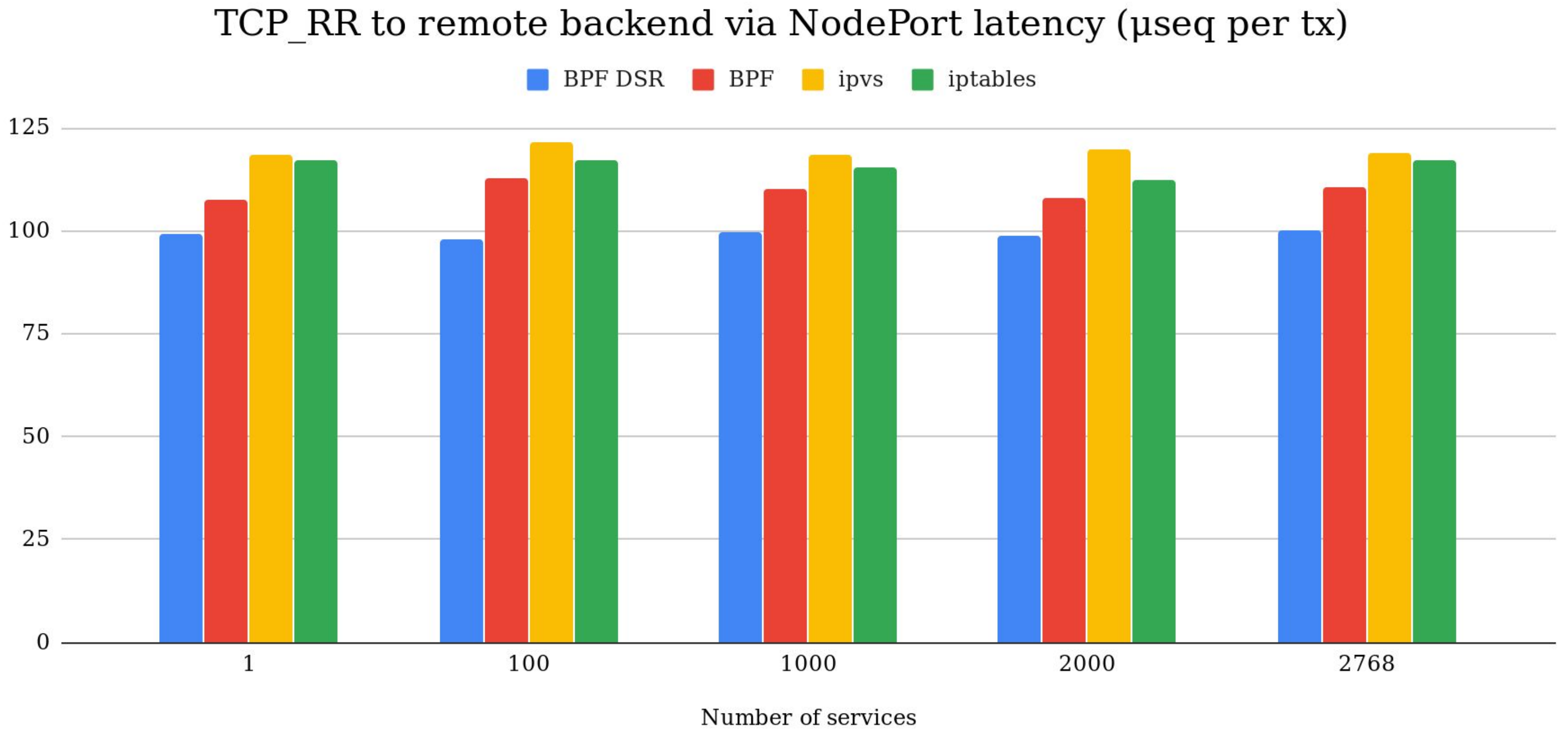
NodePort (DSR) in Cilium



Performance (lower is better)



Performance (lower is better)



Summary

Performance

- Better performance and latency over kube-proxy (ipvs and iptables)

Reliability

- Less LOC in datapath
- No need to wait for a new kernel release to fix a bug

Debuggability

- Better tooling for introspection and troubleshooting

Compatibility

- No more exec iptables

Customization

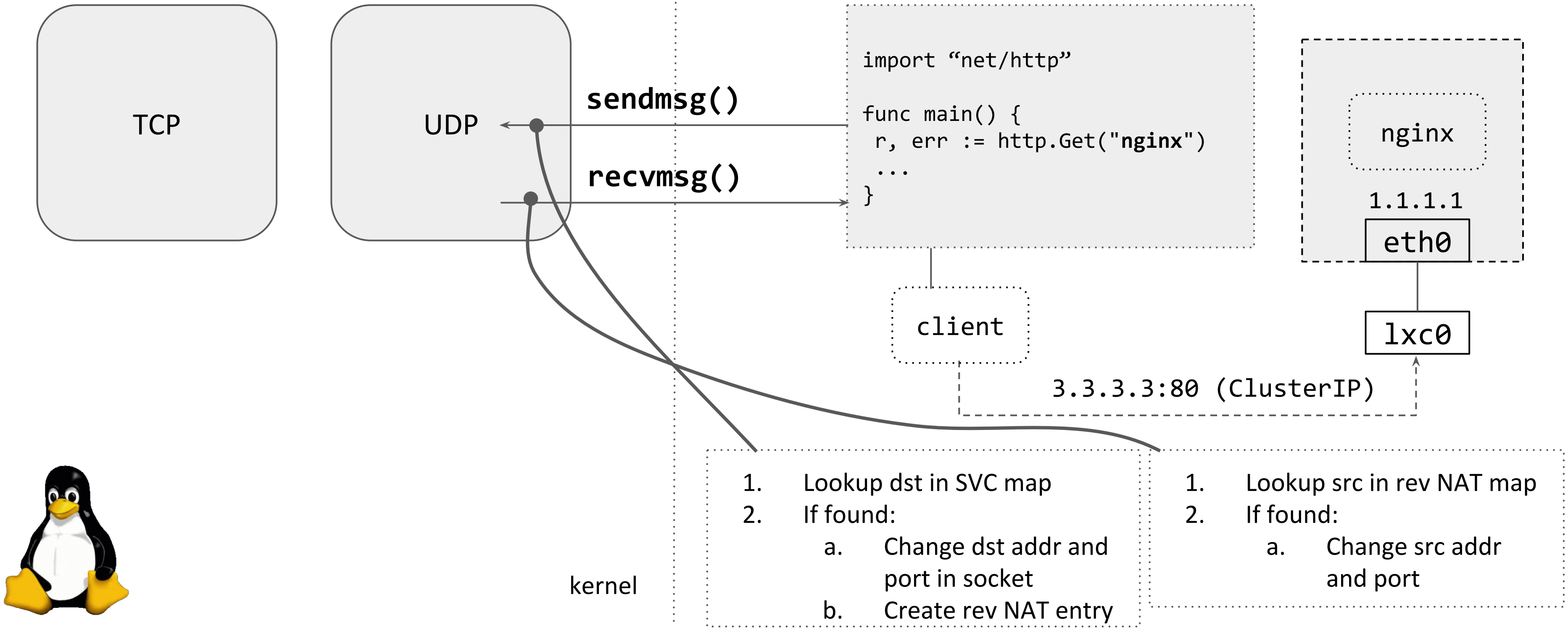
- Ability to change LB behaviour

Want to liberate yourself from kube-proxy?
Come to our booth (S93)!

https://cilium.link/kubeproxy-free
https://github.com/cilium/cilium

cilium

ClusterIP (host to pod)



TCP

UDP

sendmsg()

recvmsg()

```
import “net/http”

func main() {
 r, err := http.Get("nginx")
  ...
}
```

client

nginx

1.1.1.1

eth0

lxc0

3.3.3.3:80 (ClusterIP)

kernel

1. Lookup dst in SVC map
2. If found:
    a. Change dst addr and port in socket
    b. Create rev NAT entry

1. Lookup src in rev NAT map
2. If found:
    a. Change src addr and port