eBPF + Observability Day

CSCI 2952-F – Distributed Systems at Scale: Microservices Management

Palani Kodeswaran   Sayandeep Sen

IBM Research

Microservices: What’s awesome ?

Microservices: (not-so good parts)

1. Too many layers [Isolation, security and privacy]
2. Workloads move all over [Scalability, Elasticity]
3. No-organization has full end-to-end visibility [Business secrets]
4. Distance between application developer and deployment increases [Separation of concern]

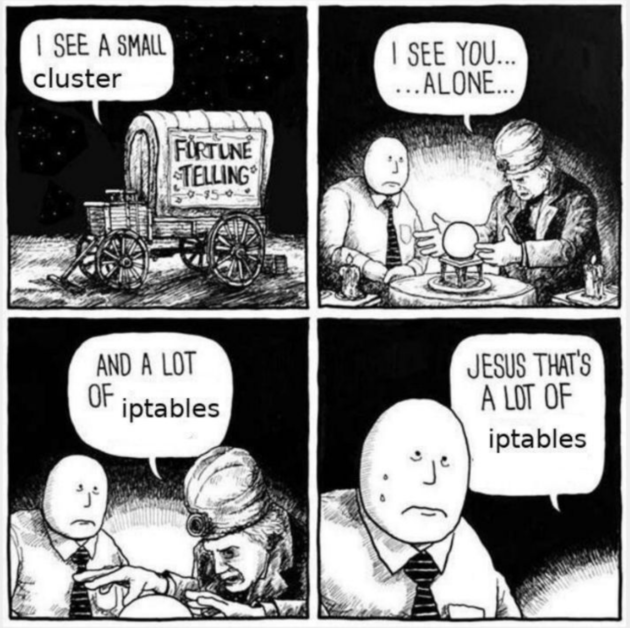
Microservices: (not-so good parts)

1. Too many layers [Isolation, security and privacy]
2. Workloads move all over [Scalability, Elasticity]
3. No-organization has full end-to-end visibility [Business secrets]
4. Distance between application developer and deployment increases [Separation of concern]

1. When problems occur, debugging becomes exceptionally hard!
2. What tools do people use?
3. Who is responsible for End-to-end System Health of all these services?
4. Do Operators know enough about what is running?

Debuggability

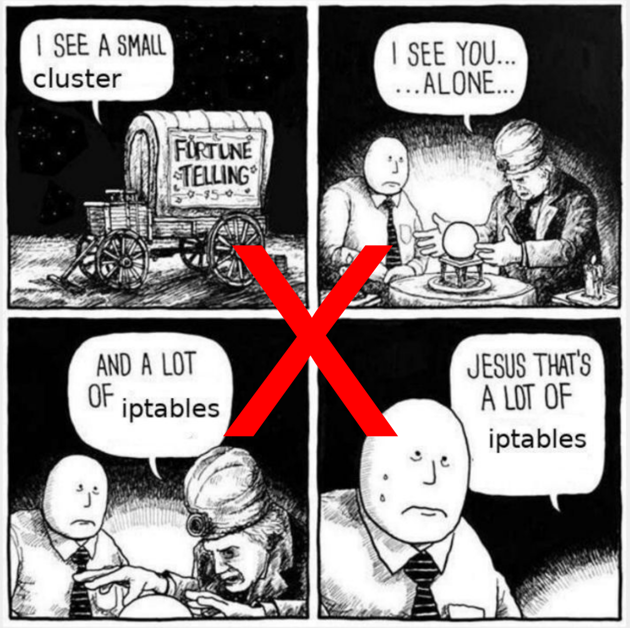
```
:OUTPUT ACCEPT [illegible]
:POSTROUTING ACCEPT [46778:335 10020]
COMMIT
*nat
:PREROUTING ACCEPT [0:0]
:INPUT ACCEPT [0:0]
:OUTPUT ACCEPT [31:1905]
:POSTROUTING ACCEPT [21:1305]
:DOCKER - [0:0]
:KUBE-MARK-DROP - [0:0]
:KUBE-MARK-MASQ - [0:0]
:KUBE-NODEPORTS - [0:0]
:KUBE-POSTROUTING - [0:0]
:KUBE-SEP-ARIYJBMSCT6NPKLC - [0:0]
:KUBE-SEP-EVB54GPOXM4P4KYH - [0:0]
:KUBE-SEP-JNEFDVSS622RF3KK - [0:0]
:KUBE-SEP-UHV3DTYFO2UR3QEF - [0:0]
:KUBE-SEP-PVCRDUMNZPYK3THF - [0:0]
:KUBE-SEP-RY4UHCSDDTRJ5BRD - [0:0]
:KUBE-SEP-YQP473NSN3FTS3LX - [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-JNEFDVSS622RF3KK -s 10.217.0.224/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-JNEFDVSS622RF3KK -p tcp -m tcp -j DNAT --to-destination 10.217.0.224:53
-A KUBE-SEP-UHV3DTYFO2UR3QEF -s 192.168.1.125/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-UHV3DTYFO2UR3QEF -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-YQP473NSN3FTS3LX -s 10.217.0.71/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-YQP473NSN3FTS3LX -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-JNEFDVSS622RF3KK
-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-UHV3DTYFO2UR3QEF
-A KUBE-SVC-TCOU7JCQXEZGVUNU -m statistic --mode random --probability 0.50000000000 -j KUBE-SEP-PVCRDUMNZPYK3THF
-A KUBE-SVC-TCOU7JCQXEZGVUNU -j KUBE-SEP-YQP473NSN3FTS3LX
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 10.109.209.138/32 -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 10.110.226.210/32 -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 10.96.213.37/32 -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 10.109.16.198/32 -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 10.103.194.218/32 -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 10.96.214.214/32 -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 10.108.221.155/32 -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
```



Source: reddit.com/r/networkingmemes/comments/8u7jyz/container_networking/

Debuggability

```
:OUTPUT ACCEPT [[illegible]]
:POSTROUTING ACCEPT [46778:135 10020]
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-JNETOVS5622RF3KK - [0:0]
:KUBE-SEP-UHV3DTYFO2URJQEF - [0:0]
:KUBE-SEP-PVCRDUMNZPYK3THF - [0:0]
:KUBE-SEP-RY4UHCSDDTRJSBRD - [0:0]
:KUBE-SEP-YQP473NSN3FTS3LX - [0:0]
:KUBE-SERVICES - [0:0]
:KUBE-SVC-ERIFXISQEP7F7OF4 - [0:0]
:KUBE-SVC-JD5MR3NA44DYORP - [0:0]
:K[illegible]MMFTMTXRNEY - [0:0]
:K[illegible]SVC-TC[illegible]CQXE2GVUNU - [0:0]
-A [illegible]ROUTING [illegible]
-A [illegible]ROUTING [illegible]
-A [illegible]TPUT -m [illegible] "kubernetes [illegible]" -j [illegible]SERVICES
-A [illegible]TPUT ! -d [illegible] -m addrtype [illegible]-type [illegible] DOCKER
-A [illegible]STROU[illegible] -m comment --comment "kubernetes postrouting rules" -j KUBE-POSTROUTING
-A [illegible] -s 172.1[illegible] -o d[illegible]SQUERADE
-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-JNETOVS5622RF3KK -s 10.217.0.224/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-JNETOVS5622RF3KK -p tcp -m tcp -j DNAT --to-destination 10.217.0.224:53
-A KUBE-SEP-UHV3DTYFO2URJQEF -s 192.168.1.125/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-UHV3DTYFO2URJQEF -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-RY4UHCSDDTRJSBRD -s 10.217.0.71/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-RY4UHCSDDTRJSBRD -p tcp -m tcp -j DNAT --to-destination 10.217.0.71:53
-A KUBE-SEP-YQP473NSN3FTS3LX -s 10.217.0.71/32 -j KUBE-MARK-MASQ
-A KUBE-SEP-YQP473NSN3FTS3LX -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-JD5MR3NA44DYORP
-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-JNETOVS5622RF3KK
-A KUBE-SVC-ERIFXISQEP7F7OF4 -j KUBE-SEP-RY4UHCSDDTRJSBRD
-A KUBE-SVC-JD5MR3NA44DYORP -m statistic --mode random --probability 0.50000000000 -j KUBE-SEP-ARIYJBMSCT6NPKLC
-A KUBE-SVC-JD5MR3NA44DYORP -j KUBE-SEP-EVB54GPOXM4P4KYH
-A KUBE-SVC-NPX46M4PTMTKRN6Y -j KUBE-SEP-UHV3DTYFO2URJQEF
-A KUBE-SVC-TCOU7JCQXEZGVUNU -m statistic --mode random --probability 0.50000000000 -j KUBE-SEP-PVCRDUMNZPYK3THF
-A KUBE-SVC-TCOU7JCQXEZGVUNU -j KUBE-SEP-YQP473NSN3FTS3LX
COMMIT
*filter
:INPUT ACCEPT [2938:623620]
:FORWARD DROP [0:0]
:OUTPUT ACCEPT [2893:6714
:DOCKER - [0:0]
:DOCKER-ISOLATION-STAGE-
:DOCKER-ISOLATION-STAGE-
:DOCKER-USER - [0:0]
:KUBE-EXTERNAL-SERVICES -
:KUBE-FIREWALL - [0:0]
:KUBE-FORWARD - [0:0]
:KUBE-SERVICES - [0:0]
-A INPUT -m conntrack --ct
-A INPUT -m conntrack --ct
-A INPUT -j KUBE-FIREWALL
-A FORWARD -m comment --
-A FORWARD -m conntrack --
-A FORWARD -j DOCKER-USE
-A FORWARD -j DOCKER-ISOL
-A FORWARD -o docker0 -m c
-A FORWARD -o docker0 -j D
-A FORWARD -i docker0 ! -o d
-A FORWARD -i docker0 -o d
-A FORWARD -o docker_gwb
-A FORWARD -o docker_gwb
-A FORWARD -i docker_gwbr
-A FORWARD -i docker_gwbr
-A OUTPUT -m conntrack --ct
-A OUTPUT -j KUBE-FIREWAL
-A DOCKER-ISOLATION-STAG
-A DOCKER-ISOLATION-STAG
-A DOCKER-ISOLATION-STAG
-A DOCKER-ISOLATION-STAG
-A DOCKER-ISOLATION-STAG
-A DOCKER-ISOLATION-STAG
-A DOCKER-USER -j RETURN
-A KUBE-FIREWALL -m comm
-A KUBE-FORWARD -m comm
-A KUBE-FORWARD -m comm
-A KUBE-FORWARD -s 10.217
-A KUBE-FORWARD -d 10.217
-A KUBE-SERVICES -d 10.99.3
-A KUBE-SERVICES -d 10.96.6
-A KUBE-SERVICES -d 10.104.
-A KUBE-SERVICES -d 10.98.8
-A KUBE-SERVICES -d 10.97.1
-A KUBE-SERVICES -d 10.106.
-A KUBE-SERVICES -d 10.104.
-A KUBE-SERVICES -d 10.104.
-A KUBE-SERVICES -d 10.106.
-A KUBE-SERVICES -d 10.109.
-A KUBE-SERVICES -d 10.106.
-A KUBE-SERVICES -d 10.111.
-A KUBE-SERVICES -d 10.110.
-A KUBE-SERVICES -d 10.98.9
-A KUBE-SERVICES -d 10.99.7
-A KUBE-SERVICES -d 10.108.
-A KUBE-SERVICES -d 10.97.3
-A KUBE-SERVICES -d 10.98.2
-A KUBE-SERVICES -d 10.107.
-A KUBE-SERVICES -d 10.98.6
-A KUBE-SERVICES -d 10.101.
-A KUBE-SERVICES -d 10.105.
-A KUBE-SERVICES -d 10.108.
-A KUBE-SERVICES -d 10.109.
-A KUBE-SERVICES -d 10.104.
-A KUBE-SERVICES -d 10.106.
-A KUBE-SERVICES -d 10.109.
-A KUBE-SERVICES -d 10.109.
-A KUBE-SERVICES -d 10.98.2
-A KUBE-SERVICES -d 10.105.
-A KUBE-SERVICES -d 10.99.3
-A KUBE-SERVICES -d 10.111.
-A KUBE-SERVICES -d 10.99.4
-A KUBE-SERVICES -d 10.104.
-A KUBE-SERVICES -d 10.98.7
-A KUBE-SERVICES -d 10.110.
-A KUBE-SERVICES -d 10.100.
-A KUBE-SERVICES -d 10.96.5
-A KUBE-SERVICES -d 10.101.
-A KUBE-SERVICES -d 10.100.
-A KUBE-SERVICES -d 10.111.
-A KUBE-SERVICES -d 10.100.
-A KUBE-SERVICES -d 10.111.69.90/32 -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 10.96.214.151/32 -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 10.101.122.17/32 -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 10.108.221.155/32 -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
```



Source: reddit.com/r/networkingmemes/comments/8u7jyz/container_networking/

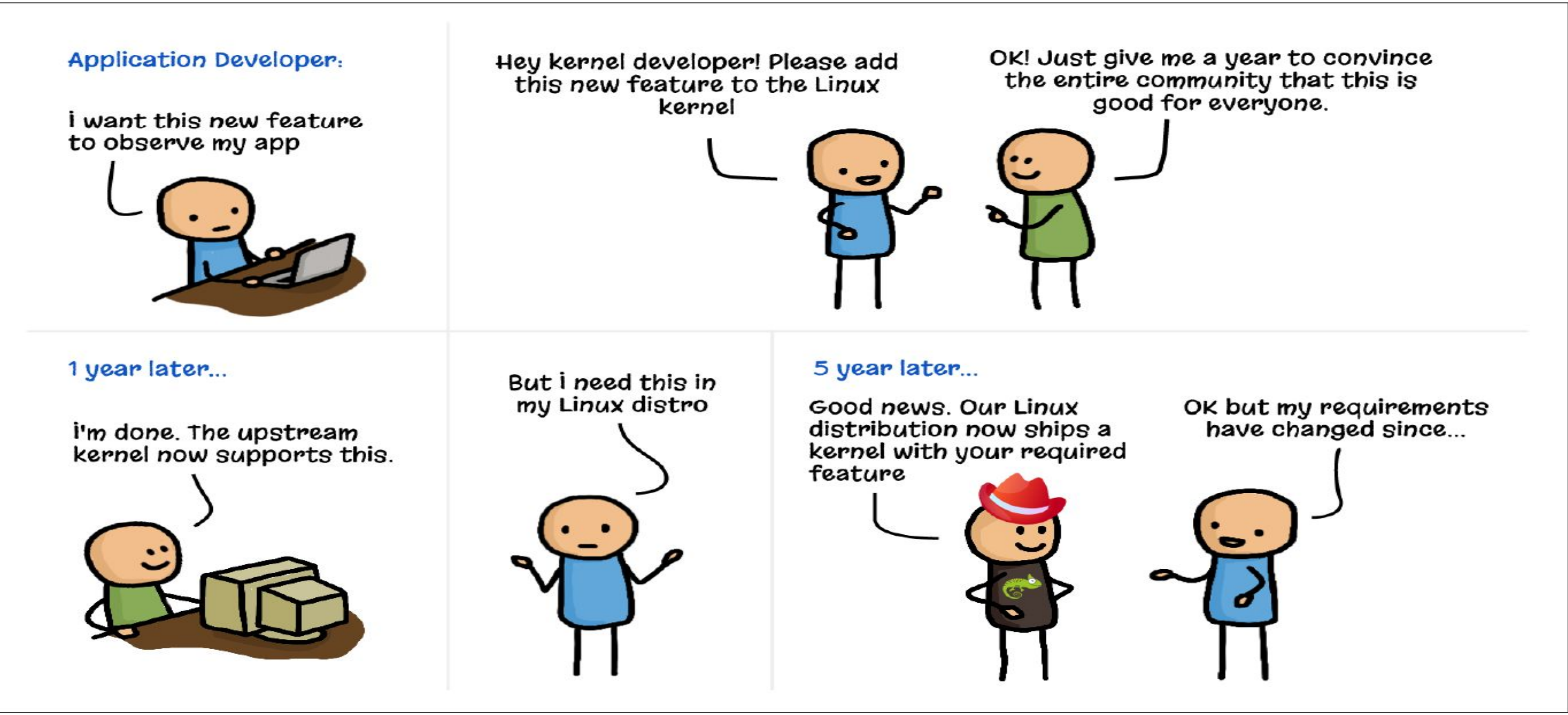
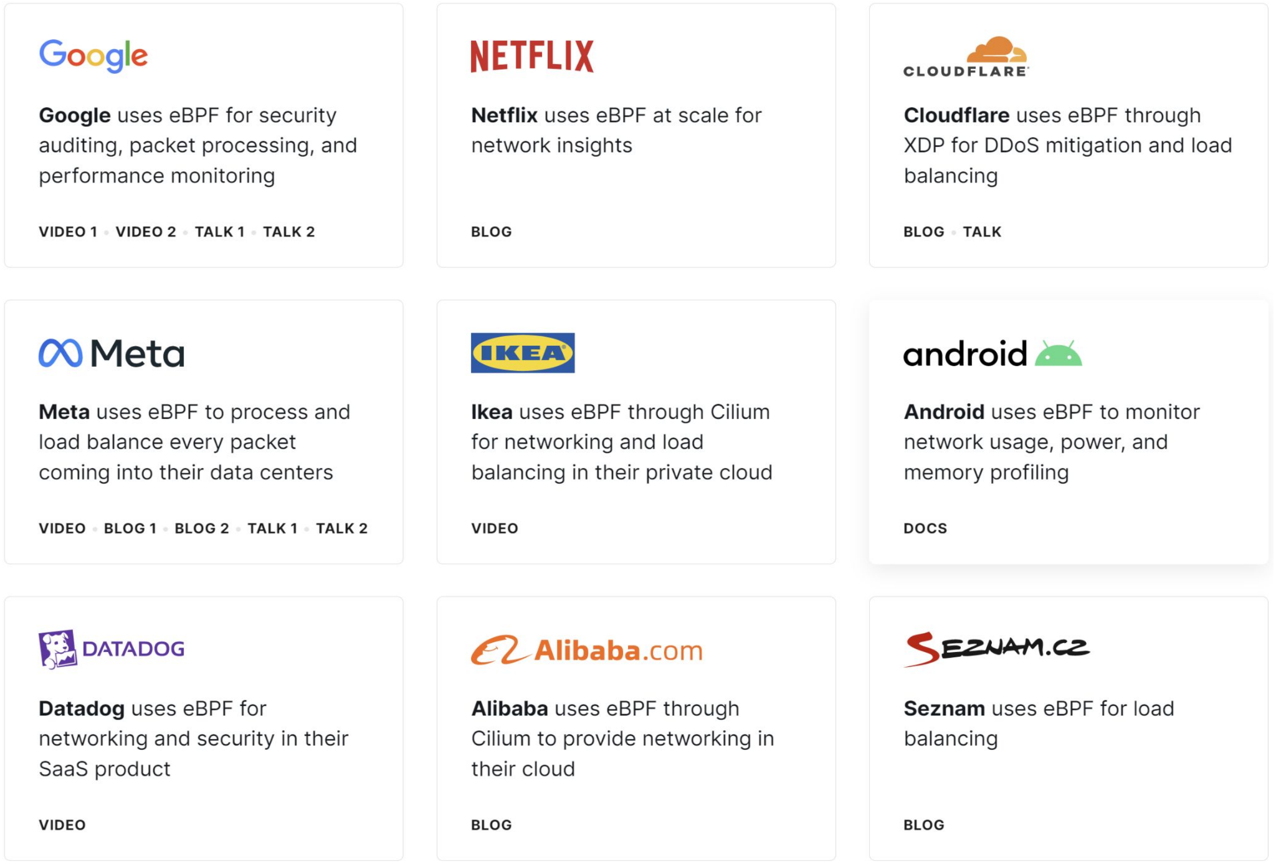

Application Developer:
i want this new feature to observe my app
Hey kernel developer! Please add this new feature to the Linux kernel
OK! Just give me a year to convince the entire community that this is good for everyone.
1 year later...
i'm done. The upstream kernel now supports this.
But i need this in my Linux distro
5 year later...
Good news. Our Linux distribution now ships a kernel with your required feature
OK but my requirements have changed since...

Figure 2-1. Adding features to the kernel (cartoon by Vadim Shchekoldin, Isovalent)

Enter eBPF



Google uses eBPF for security auditing, packet processing, and performance monitoring

VIDEO 1 · VIDEO 2 · TALK 1 · TALK 2

Netflix uses eBPF at scale for network insights

BLOG

Cloudflare uses eBPF through XDP for DDoS mitigation and load balancing

BLOG · TALK

Meta uses eBPF to process and load balance every packet coming into their data centers

VIDEO · BLOG 1 · BLOG 2 · TALK 1 · TALK 2

Ikea uses eBPF through Cilium for networking and load balancing in their private cloud

VIDEO

Android uses eBPF to monitor network usage, power, and memory profiling

DOCS

Datadog uses eBPF for networking and security in their SaaS product

VIDEO

Alibaba uses eBPF through Cilium to provide networking in their cloud

BLOG

Seznam uses eBPF for load balancing

BLOG

eBPF & Kubernetes - the future

"The Linux kernel continues its march towards becoming BPF runtime-powered microkernel."

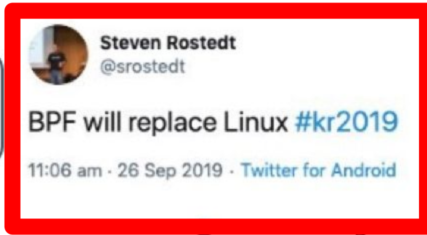
Tiny core kernel with user definable kernel functionality in BPF (instead of kernel modules)

Less security bugs & kernel crashes due to **smaller attack surface** and **safety-verified code**

Drastic reduction of 'static' feature creep for **better resource efficiency**

Kubernetes then ships **custom BPF**-tailored extensions to optimize needs for user workloads

Today's kube-proxy replacement through BPF is just a tiny dot in that universe ...


Steven Rostedt
@srostedt

BPF will replace Linux #kr2019

11:06 am · 26 Sep 2019 · Twitter for Android

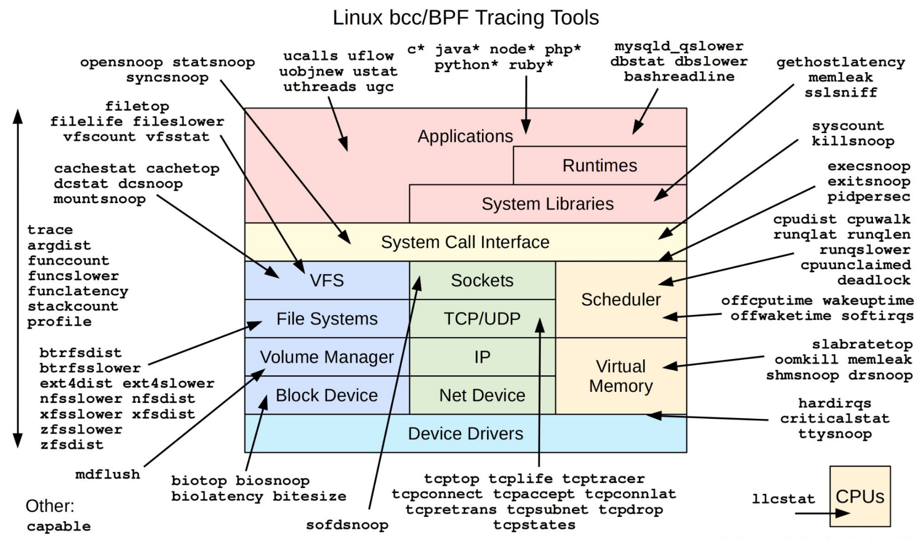
2013
2014
2015
2016
2017
2018
2019
2020
?

Toke Høiland-Jørgensen, https://lwn.net/ml/bufferbloat/87bls8bnsm.fsf@toke.dk/

eBPF: Full System Visibility (and much much more !!)

- Trace and debug of third party software!!
- Visibility of whole system deployment
- Bolt On (Application Transparent)

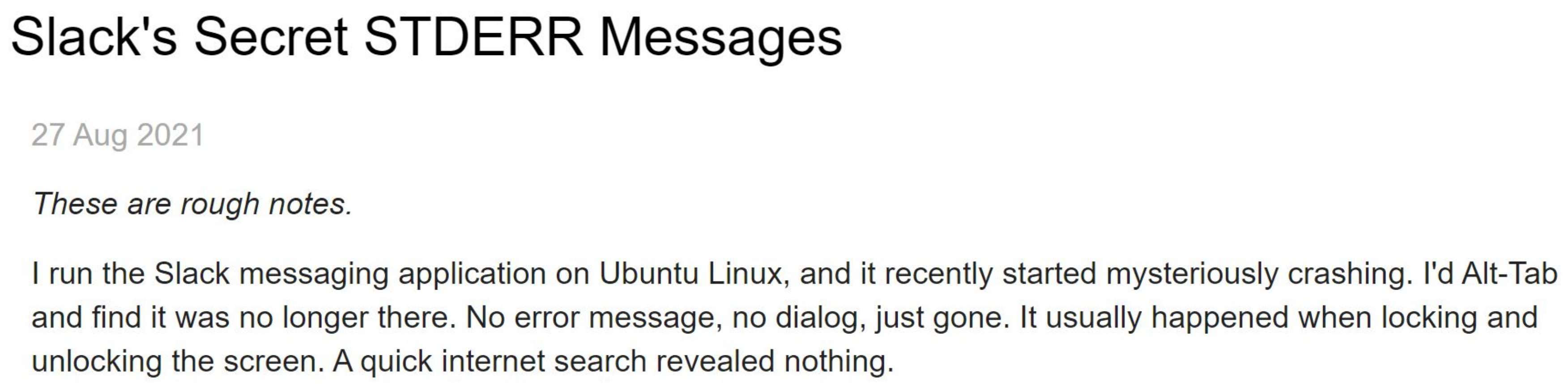
Besides applications in Networking, Security and Storage …


Linux bcc/BPF Tracing Tools
opensnoop statsnoop syncsnoop
ucalls uflow uobjnew ustat uthreads ugc
c* java* node* php* python* ruby*
mysqld_qslower dbstat dbslower bashreadline
gethostlatency memleak sslsniff
filetop filelife fileslower vfscount vfsstat
cachestat cachetop dcstat dcsnoop mountsnoop
trace argdist funccount funcslower funclatency stackcount profile
btrfsdist btrfsslower ext4dist ext4slower nfsslower nfsdist xfsslower xfsdist zfsslower zfsdist
mdflush
Other: capable
biotop biosnoop biolatency bitesize
sofdsnoop
tcptop tcplife tcptracer tcpconnect tcpaccept tcpconnlat tcpretrans tcpsubnet tcpdrop tcpstates
syscount killsnoop
execsnoop exitsnoop pidpersec
cpudist cpuwalk runqlat runqlen runqslower cpuunclaimed deadlock
offcputime wakeuptime offwaketime softirqs
slabratetop oomkill memleak shmsnoop drsnoop
hardirqs criticalstat ttysnoop
llcstat
CPUs
Applications
Runtimes
System Libraries
System Call Interface
VFS
Sockets
Scheduler
File Systems
TCP/UDP
Volume Manager
IP
Virtual Memory
Block Device
Net Device
Device Drivers

https://github.com/iovisor/bcc#tools 2019

Case Study : Slack Keeps Crashing

Debugging Slack



Slack's Secret STDERR Messages

27 Aug 2021

These are rough notes.

I run the Slack messaging application on Ubuntu Linux, and it recently started mysteriously crashing. I'd Alt-Tab and find it was no longer there. No error message, no dialog, just gone. It usually happened when locking and unlocking the screen. A quick internet search revealed nothing.

From here: [Slack's Secret STDERR Messages (brendangregg.com)]

1. Enabling core dumps

I'm guessing it's core dumping and Ubuntu's apport is eating them. Redirecting them to the file system so I can then do core dump analysis using gdb(1), as root:

```
# cat /proc/sys/kernel/core_pattern
|/usr/share/apport/apport %p %s %c %d %P
# mkdir /var/cores
# echo "/var/cores/core.%e.%p.%h.%t" > /proc/sys/kernel/core_pattern
[...another crash...]
# ls /var/cores
#
```

This didn't work: No core file showed up. I may need to increase the core file size ulimits for Slack, but that might mean mucking around with its startup scripts; I'll try some other tracing first.

From here: Slack's Secret STDERR Messages (brendangregg.com)

2. exitsnoop

Using an eBPF/bcc tool to look for exit reasons:

```
# exitsnoop -t
TIME-AEST    PCOMM            PID    PPID   TID    AGE(s)  EXIT_CODE
13:51:19.432 kworker/dying    3663305 2      3663305 1241.59 0
13:51:30.948 kworker/dying    3663626 2      3663626 835.76  0
13:51:33.296 systemd-udevd    3664149 2054939 3664149 3.55    0
13:53:09.256 kworker/dying    3662179 2      3662179 2681.15 0
13:53:25.636 kworker/dying    3663520 2      3663520 1122.60 0
13:53:30.705 grep             3664239 6009   3664239 0.08    0
13:53:30.705 ps               3664238 6009   3664238 0.08    0
13:53:40.297 slack            3663135 1786   3663135 1459.54 signal 6 (ABRT)
13:53:40.298 slack            3663208 3663140 3663208 1457.86 0
13:53:40.302 slack            3663140 1786   3663140 1459.18 0
13:53:40.302 slack            3663139 1786   3663139 1459.18 0
13:53:40.303 slack            3663171 1786   3663171 1458.22 0
13:53:40.317 slack            3663197 1786   3663197 1458.03 0
13:53:44.827 gdm-session-wor  3664269 1778   3664269 0.02    0
[...]
```

This traced a Slack SIGABRT which happened around the same time as a crash. A strong lead.

From here: Slack's Secret STDERR Messages (brendangregg.com)

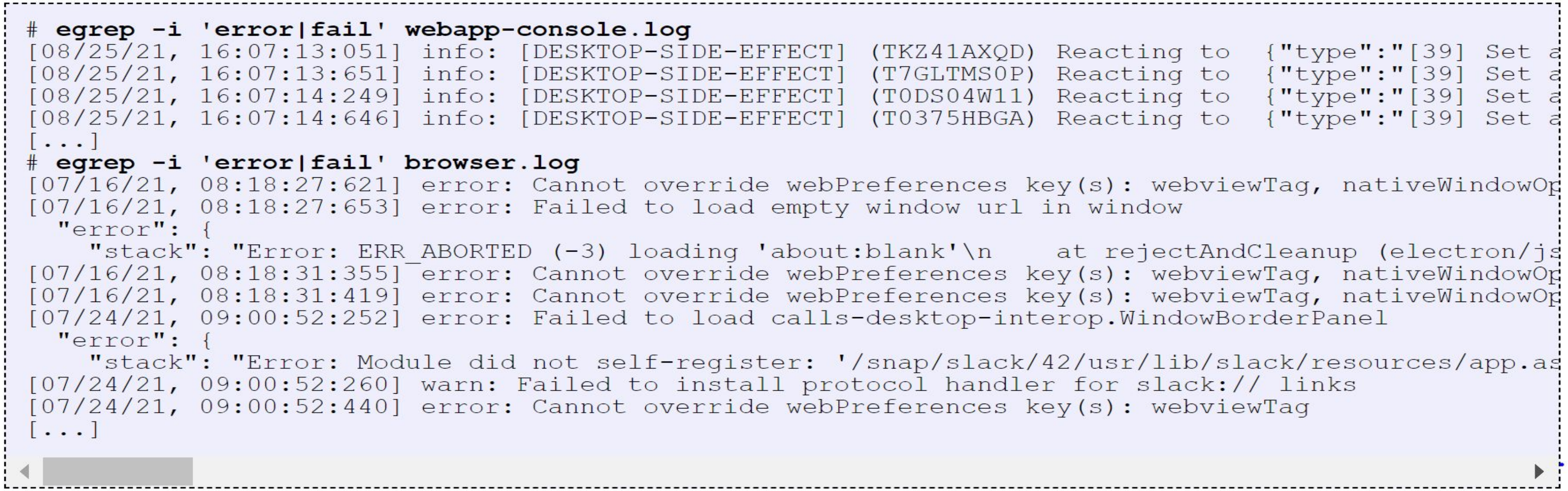
5. signals.bt

The signals.bt bpftrace tool from my BPF book traces the signal:signal_generate tracepoint, which should catch every type of generated signal, including tgkill(2). Trying it out:

```
# bpftrace /home/bgregg/Git/bpf-perf-tools-book/originals/Ch13_Applications/signals.bt
Attaching 3 probes...
Counting signals. Hit Ctrl-C to end.
^C
@[SIGNAL, PID, COMM] = COUNT

@[SIGPIPE, 1883, Xorg]: 1
@[SIGCHLD, 1797, dbus-daemon]: 1
@[SIGINT, 3665167, bpftrace]: 1
@[SIGTERM, 3665198, gdm-session-wor]: 1
@[SIGCHLD, 3665197, gdm-session-wor]: 1
@[SIGABRT, 3664940, slack]: 1
@[SIGTERM, 3665197, gdm-session-wor]: 1
@[SIGKILL, 3665207, dbus-daemon]: 1
@[SIGWINCH, 859450, bash]: 2
@[SIGCHLD, 1778, gdm-session-wor]: 2
@[, 3665201, gdbus]: 2
@[, 3665199, gmain]: 2
@[SIGWINCH, 3665167, bpftrace]: 2
@[SIGWINCH, 3663319, vi]: 2
@[SIGCHLD, 1786, systemd]: 6
@[SIGALRM, 1883, Xorg]: 106
```

Ok, there's the SIGABRT for slack. (There's a new sigsnoop(8) tool for bcc that uses this tracepoint as well.)

From here: Slack's Secret STDERR Messages (brendangregg.com)



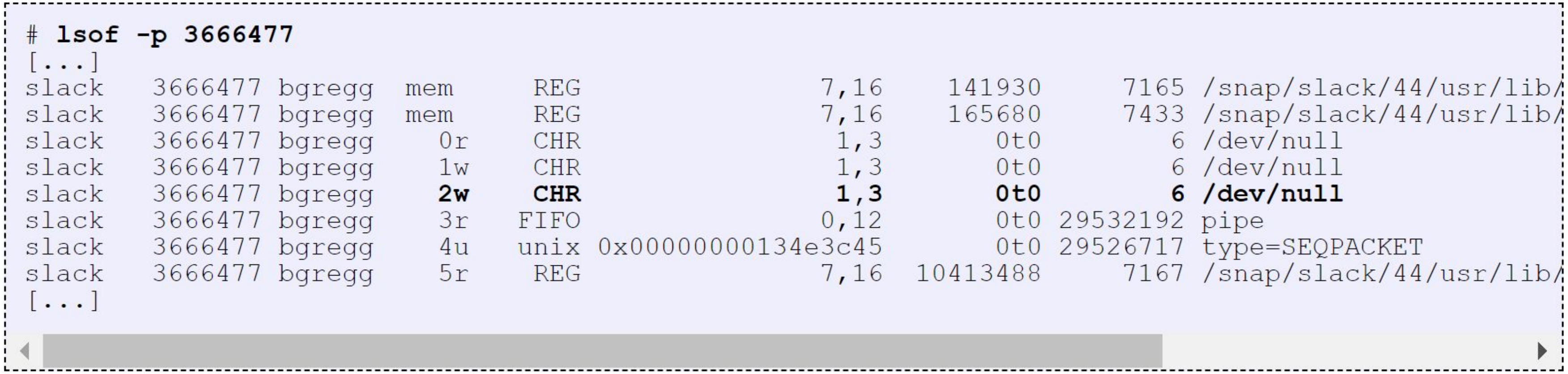
```
# egrep -i 'error|fail' webapp-console.log
[08/25/21, 16:07:13:051] info: [DESKTOP-SIDE-EFFECT] (TKZ41AXQD) Reacting to  {"type":"[39] Set a
[08/25/21, 16:07:13:651] info: [DESKTOP-SIDE-EFFECT] (T7GLTMS0P) Reacting to  {"type":"[39] Set a
[08/25/21, 16:07:14:249] info: [DESKTOP-SIDE-EFFECT] (T0DS04W11) Reacting to  {"type":"[39] Set a
[08/25/21, 16:07:14:646] info: [DESKTOP-SIDE-EFFECT] (T0375HBGA) Reacting to  {"type":"[39] Set a
[...]
# egrep -i 'error|fail' browser.log
[07/16/21, 08:18:27:621] error: Cannot override webPreferences key(s): webviewTag, nativeWindowOp
[07/16/21, 08:18:27:653] error: Failed to load empty window url in window
  "error": {
    "stack": "Error: ERR_ABORTED (-3) loading 'about:blank'\n    at rejectAndCleanup (electron/js
[07/16/21, 08:18:31:355] error: Cannot override webPreferences key(s): webviewTag, nativeWindowOp
[07/16/21, 08:18:31:419] error: Cannot override webPreferences key(s): webviewTag, nativeWindowOp
[07/24/21, 09:00:52:252] error: Failed to load calls-desktop-interop.WindowBorderPanel
  "error": {
    "stack": "Error: Module did not self-register: '/snap/slack/42/usr/lib/slack/resources/app.as
[07/24/21, 09:00:52:260] warn: Failed to install protocol handler for slack:// links
[07/24/21, 09:00:52:440] error: Cannot override webPreferences key(s): webviewTag
[...]
```

I browsed the logs for a while but didn't see a smoking gun. Surely it spits out some error message when crashing, like to STDERR...

From here: [Slack's Secret STDERR Messages (brendangregg.com)]

8. STDERR Tracing

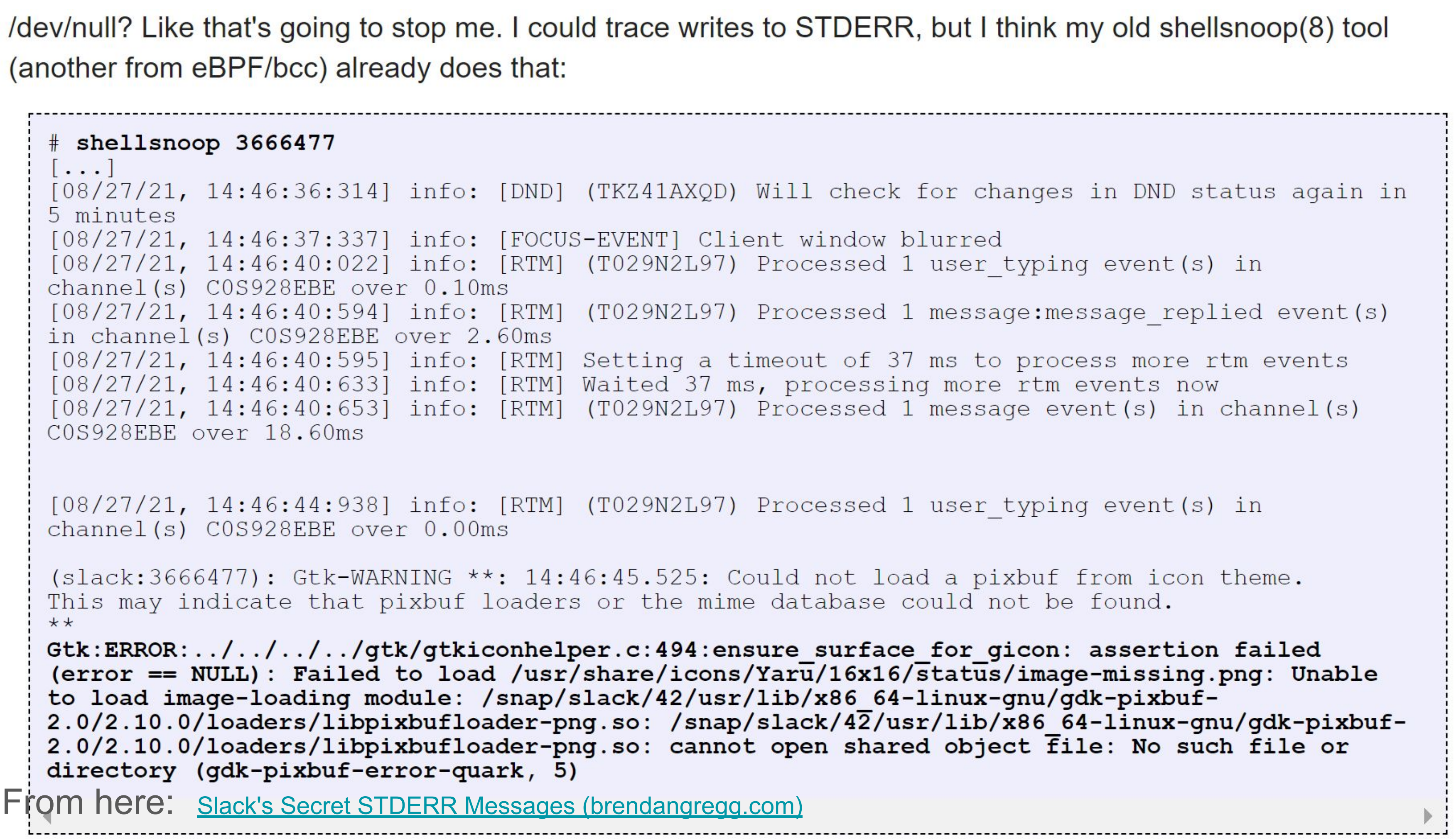
Where is STDERR written?



```
# lsof -p 3666477
[...]
slack   3666477 bgregg  mem     REG                7,16    141930     7165 /snap/slack/44/usr/lib/
slack   3666477 bgregg  mem     REG                7,16    165680     7433 /snap/slack/44/usr/lib/
slack   3666477 bgregg    0r    CHR                 1,3       0t0        6 /dev/null
slack   3666477 bgregg    1w    CHR                 1,3       0t0        6 /dev/null
slack   3666477 bgregg    2w    CHR                 1,3       0t0        6 /dev/null
slack   3666477 bgregg    3r   FIFO                0,12       0t0 29532192 pipe
slack   3666477 bgregg    4u   unix 0x00000000134e3c45       0t0 29526717 type=SEQPACKET
slack   3666477 bgregg    5r    REG                7,16  10413488     7167 /snap/slack/44/usr/lib/
[...]
```

/dev/null? Like that's going to stop me. I could trace writes to STDERR, but I think my old shellsnoop(8) tool (another from eBPF/bcc) already does that:

From here: Slack's Secret STDERR Messages (brendangregg.com)



/dev/null? Like that's going to stop me. I could trace writes to STDERR, but I think my old shellsnoop(8) tool (another from eBPF/bcc) already does that:

```
# shellsnoop 3666477
[...]
[08/27/21, 14:46:36:314] info: [DND] (TKZ41AXQD) Will check for changes in DND status again in
5 minutes
[08/27/21, 14:46:37:337] info: [FOCUS-EVENT] Client window blurred
[08/27/21, 14:46:40:022] info: [RTM] (T029N2L97) Processed 1 user_typing event(s) in
channel(s) C0S928EBE over 0.10ms
[08/27/21, 14:46:40:594] info: [RTM] (T029N2L97) Processed 1 message:message_replied event(s)
in channel(s) C0S928EBE over 2.60ms
[08/27/21, 14:46:40:595] info: [RTM] Setting a timeout of 37 ms to process more rtm events
[08/27/21, 14:46:40:633] info: [RTM] Waited 37 ms, processing more rtm events now
[08/27/21, 14:46:40:653] info: [RTM] (T029N2L97) Processed 1 message event(s) in channel(s)
C0S928EBE over 18.60ms


[08/27/21, 14:46:44:938] info: [RTM] (T029N2L97) Processed 1 user_typing event(s) in
channel(s) C0S928EBE over 0.00ms

(slack:3666477): Gtk-WARNING **: 14:46:45.525: Could not load a pixbuf from icon theme.
This may indicate that pixbuf loaders or the mime database could not be found.
**
Gtk:ERROR:../../../../gtk/gtkiconhelper.c:494:ensure_surface_for_gicon: assertion failed
(error == NULL): Failed to load /usr/share/icons/Yaru/16x16/status/image-missing.png: Unable
to load image-loading module: /snap/slack/42/usr/lib/x86_64-linux-gnu/gdk-pixbuf-
2.0/2.10.0/loaders/libpixbufloader-png.so: /snap/slack/42/usr/lib/x86_64-linux-gnu/gdk-pixbuf-
2.0/2.10.0/loaders/libpixbufloader-png.so: cannot open shared object file: No such file or
directory (gdk-pixbuf-error-quark, 5)
```

From here: [Slack's Secret STDERR Messages (brendangregg.com)]

It's the .so file that is missing, not the .png:

```
# ls -lh /usr/share/icons/Yaru/16x16/status/image-missing.png
-rw-r--r-- 1 root root 535 Nov  6  2020 /usr/share/icons/Yaru/16x16/status/image-missing.png
# ls -lh /snap/slack/42/usr/lib/x86_64-linux-gnu/gdk-pixbuf-2.0/2.10.0/loaders/libpixbufloader-pr
ls: cannot access '/snap/slack/42/usr/lib/x86_64-linux-gnu/gdk-pixbuf-2.0/2.10.0/loaders/libpixbu
```

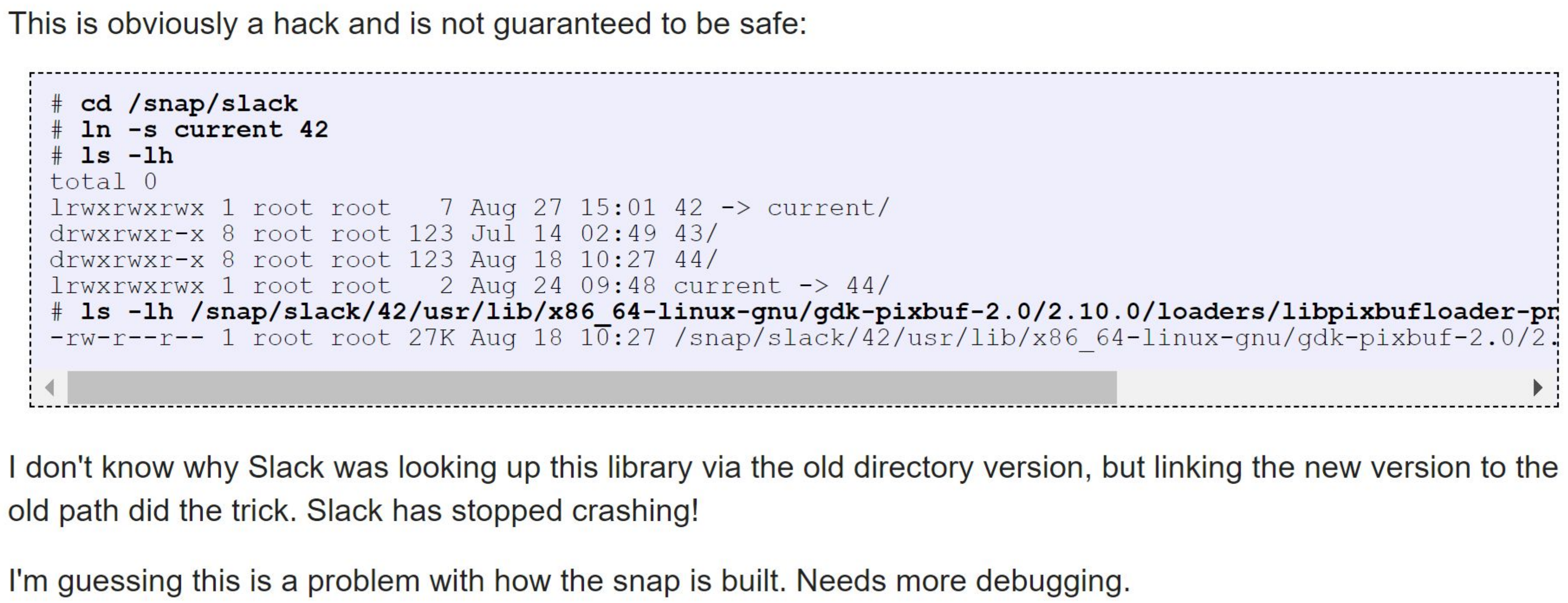
But there is a .so file with a similar path:

```
# ls -lh /snap/slack/
total 0
drwxrwxr-x 8 root root 123 Jul 14 02:49 43/
drwxrwxr-x 8 root root 123 Aug 18 10:27 44/
lrwxrwxrwx 1 root root   2 Aug 24 09:48 current -> 44/
# ls -lh /snap/slack/44/usr/lib/x86_64-linux-gnu/gdk-pixbuf-2.0/2.10.0/loaders/libpixbufloader-pr
-rw-r--r-- 1 root root 27K Aug 18 10:27 /snap/slack/44/usr/lib/x86_64-linux-gnu/gdk-pixbuf-2.0/2.
```

Hmm, I wonder...

From here: [Slack's Secret STDERR Messages (brendangregg.com)]

9. Workaround



This is obviously a hack and is not guaranteed to be safe:

```
# cd /snap/slack
# ln -s current 42
# ls -lh
total 0
lrwxrwxrwx 1 root root   7 Aug 27 15:01 42 -> current/
drwxrwxr-x 8 root root 123 Jul 14 02:49 43/
drwxrwxr-x 8 root root 123 Aug 18 10:27 44/
lrwxrwxrwx 1 root root   2 Aug 24 09:48 current -> 44/
# ls -lh /snap/slack/42/usr/lib/x86_64-linux-gnu/gdk-pixbuf-2.0/2.10.0/loaders/libpixbufloader-pr
-rw-r--r-- 1 root root 27K Aug 18 10:27 /snap/slack/42/usr/lib/x86_64-linux-gnu/gdk-pixbuf-2.0/2.
```

I don't know why Slack was looking up this library via the old directory version, but linking the new version to the old path did the trick. Slack has stopped crashing!

I'm guessing this is a problem with how the snap is built. Needs more debugging.

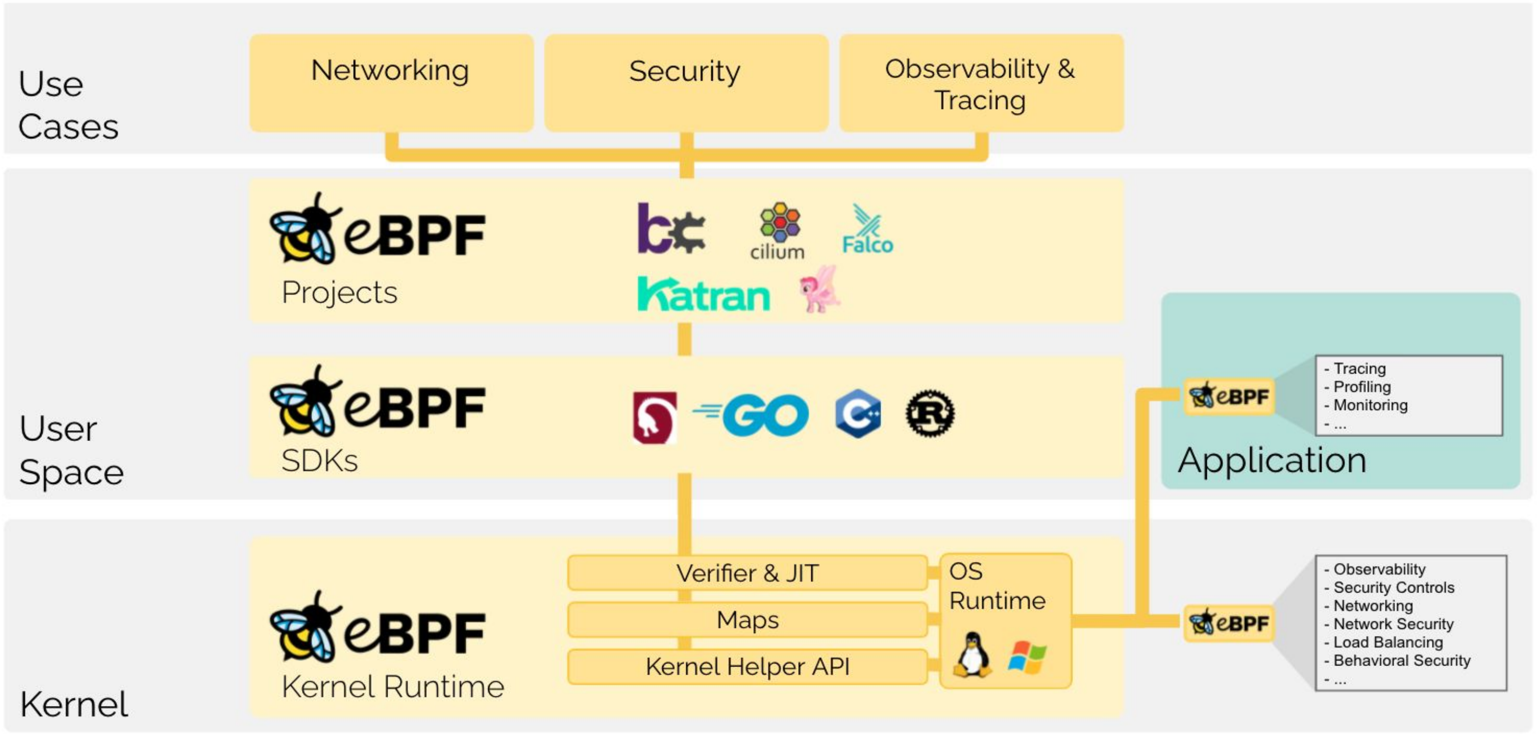
From here: Slack's Secret STDERR Messages (brendangregg.com)

Summary

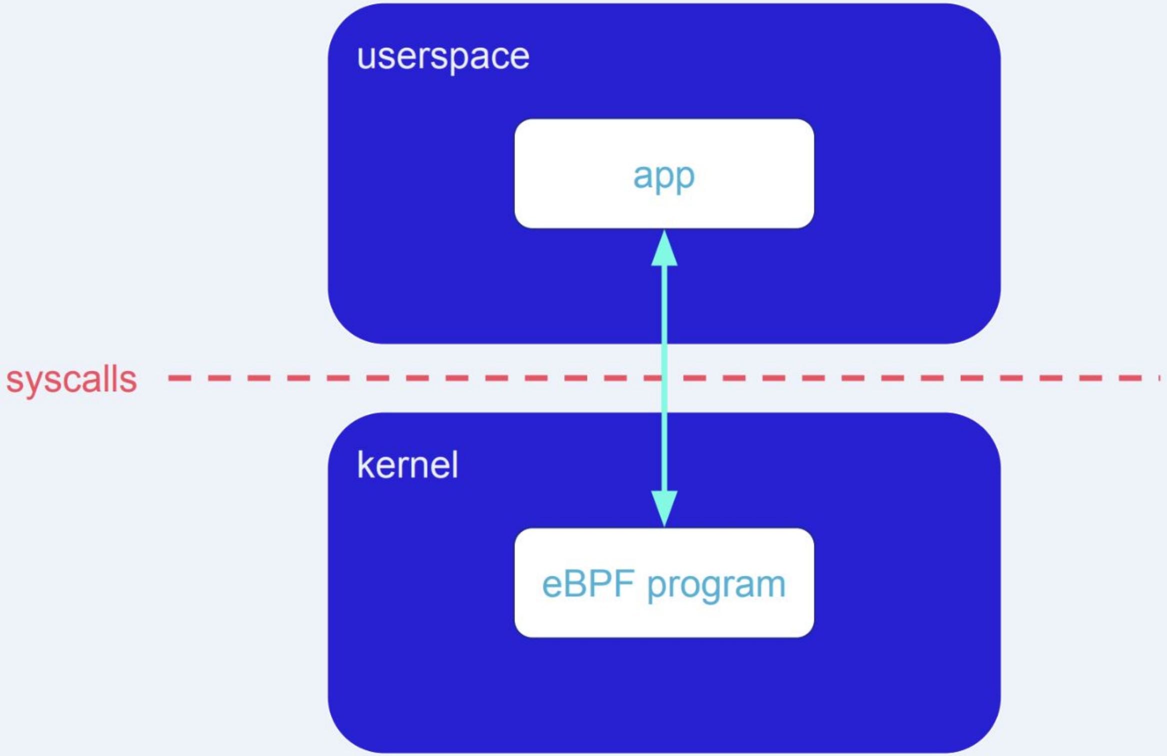
eBPF has super powers that gives it unprecedented flexibility and visibility to diagnose problems in even third party code!!!

eBPF Internals

What is eBPF



From here: [What is eBPF? An Introduction and Deep Dive into the eBPF Technology]()

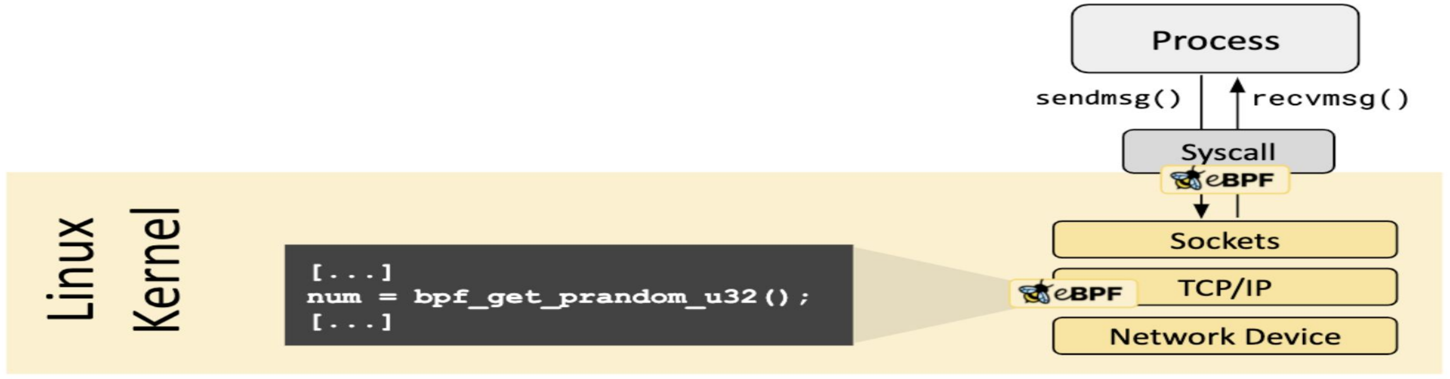

userspace
app
syscalls
kernel
eBPF program

man bpf

```
eBPF programs can be written in a restricted C that is compiled (using
the clang compiler) into eBPF bytecode.  Various features are omitted
from this restricted C, such as loops, global variables, variadic
functions, floating-point numbers, and passing structures as function
arguments.
```

Helper Calls

eBPF programs cannot call into arbitrary kernel functions. Allowing this would bind eBPF programs to particular kernel versions and would complicate compatibility of programs. Instead, eBPF programs can make function calls into helper functions, a well-known and stable API offered by the kernel.



```
[...]
num = bpf_get_prandom_u32();
[...]
```

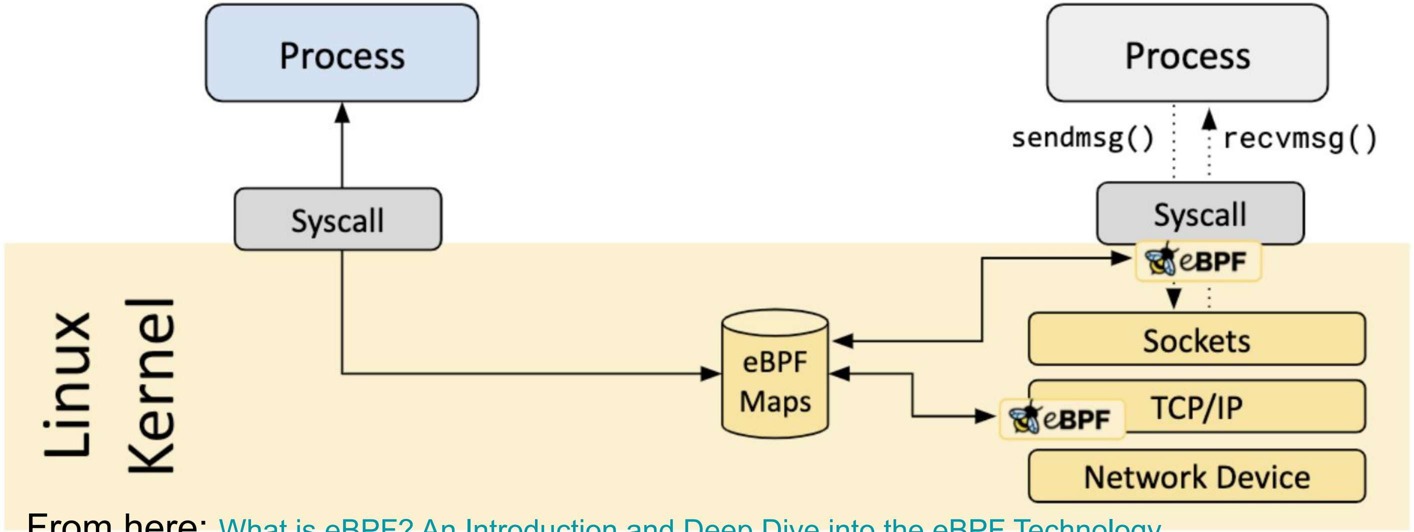
The set of available helper calls is constantly evolving. Examples of available helper calls:

- Generate random numbers
- Get current time & date
- eBPF map access
- Get process/cgroup context
- Manipulate network packets and forwarding logic

From here: [What is eBPF? An Introduction and Deep Dive into the eBPF Technology]

Maps

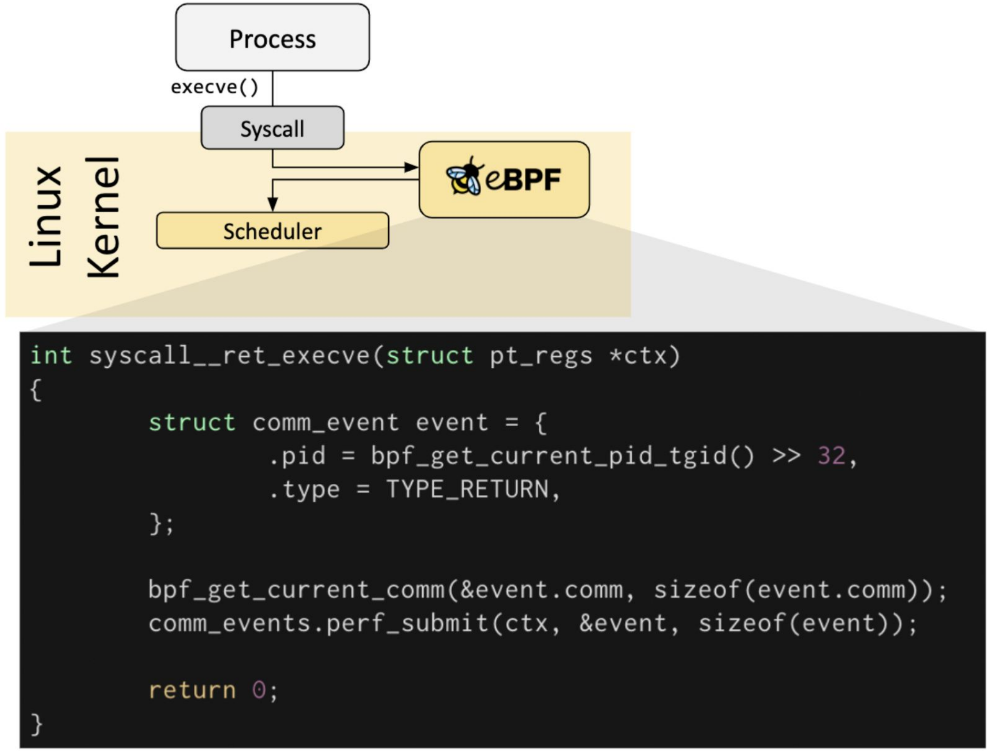
A vital aspect of eBPF programs is the ability to share collected information and to store state. For this purpose, eBPF programs can leverage the concept of eBPF maps to store and retrieve data in a wide set of data structures. eBPF maps can be accessed from eBPF programs as well as from applications in user space via a system call.



From here: [What is eBPF? An Introduction and Deep Dive into the eBPF Technology]

Hook Overview

eBPF programs are event-driven and are run when the kernel or an application passes a certain hook point. Pre-defined hooks include system calls, function entry/exit, kernel tracepoints, network events, and several others.

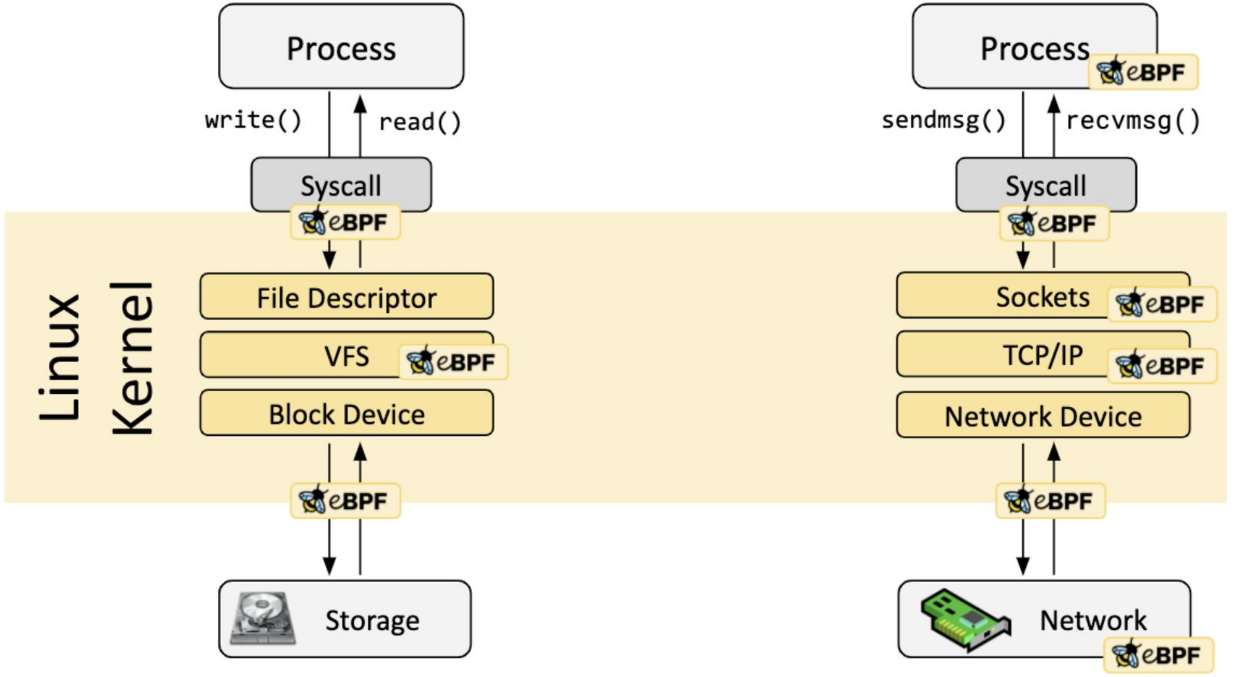


```
int syscall__ret_execve(struct pt_regs *ctx)
{
        struct comm_event event = {
                .pid = bpf_get_current_pid_tgid() >> 32,
                .type = TYPE_RETURN,
        };

        bpf_get_current_comm(&event.comm, sizeof(event.comm));
        comm_events.perf_submit(ctx, &event, sizeof(event));

        return 0;
}
```

From here: What is eBPF? An Introduction and Deep Dive into the eBPF Technology

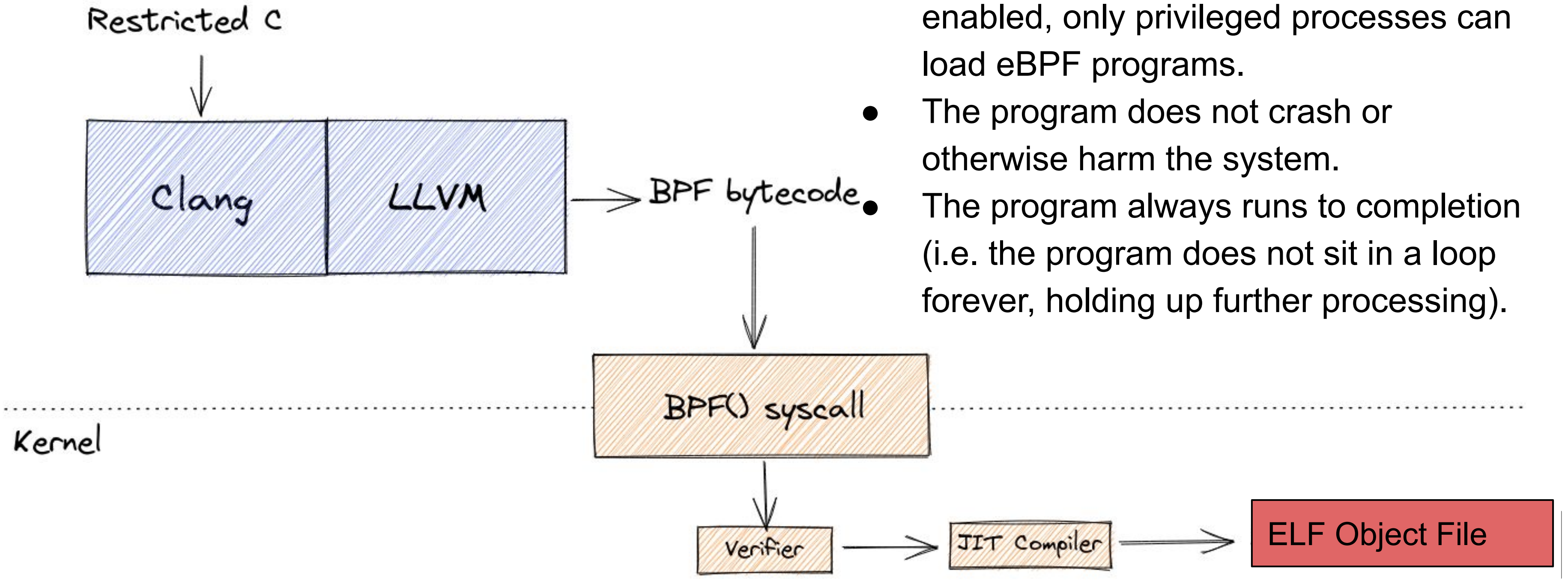
User Land probes & kernel probes

If a predefined hook does not exist for a particular need, it is possible to create a kernel probe (kprobe) or user probe (uprobe) to attach eBPF programs almost anywhere in kernel or user applications.



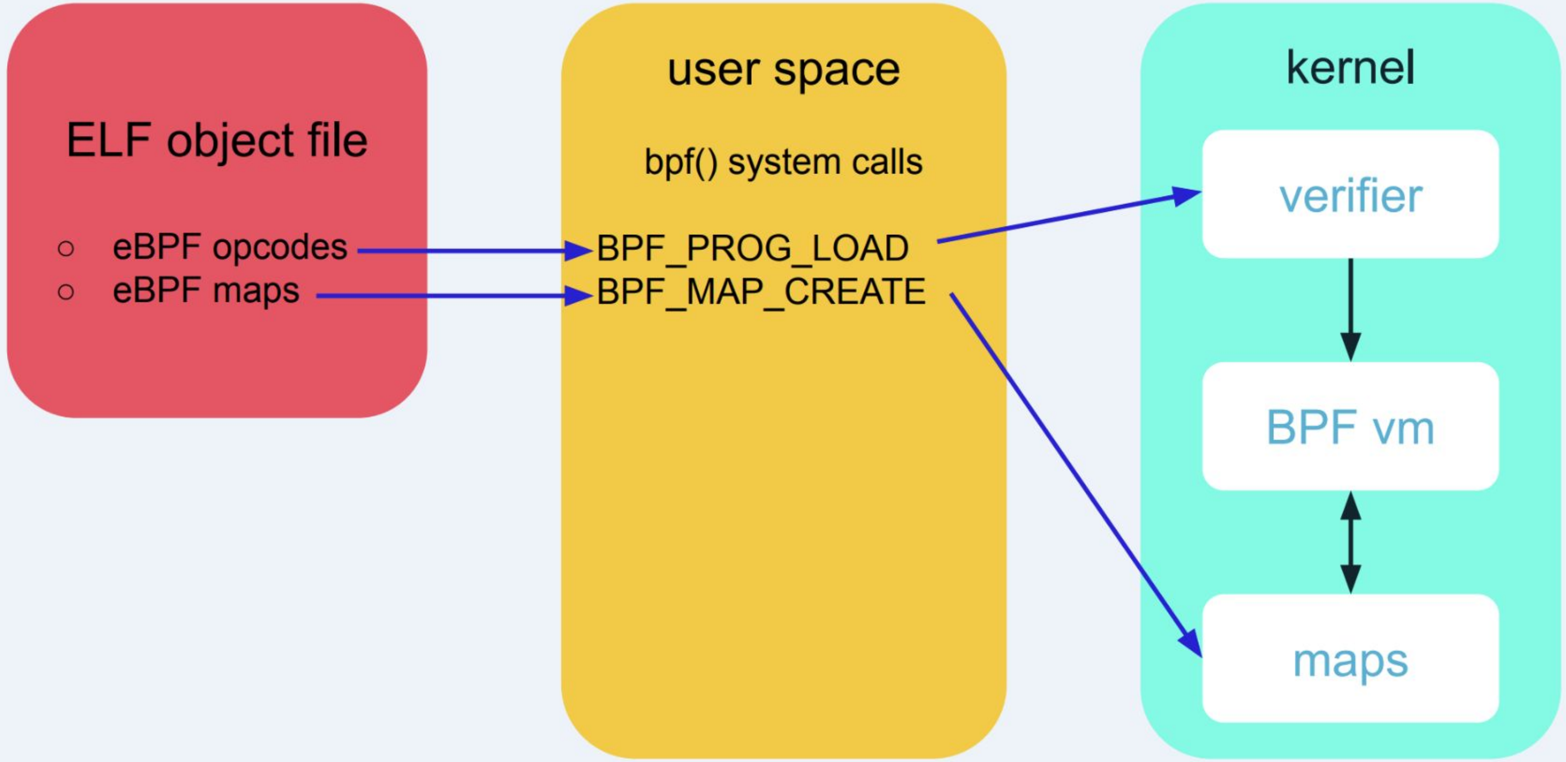
From here: [What is eBPF? An Introduction and Deep Dive into the eBPF Technology]

Tool chain



- The process loading the eBPF program holds the required capabilities (privileges). Unless unprivileged eBPF is enabled, only privileged processes can load eBPF programs.
- The program does not crash or otherwise harm the system.
- The program always runs to completion (i.e. the program does not sit in a loop forever, holding up further processing).

From here: [life-bpf-program.png]


ELF object file
○ eBPF opcodes
○ eBPF maps
user space
bpf() system calls
BPF_PROG_LOAD
BPF_MAP_CREATE
kernel
verifier
BPF vm
maps

ELF object file

- eBPF opcodes
- eBPF maps

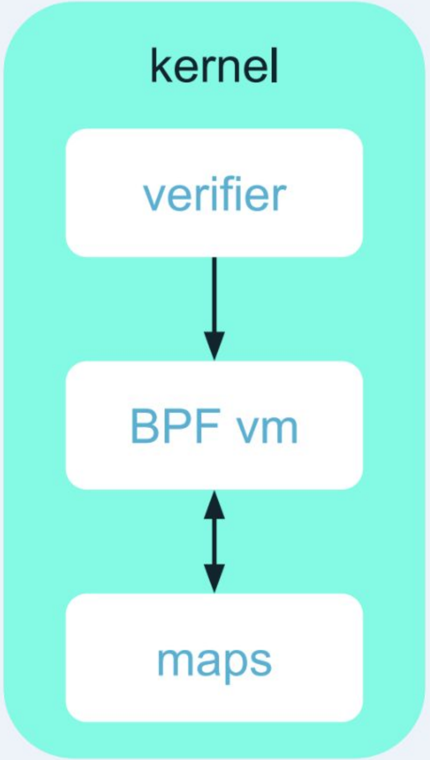
user space

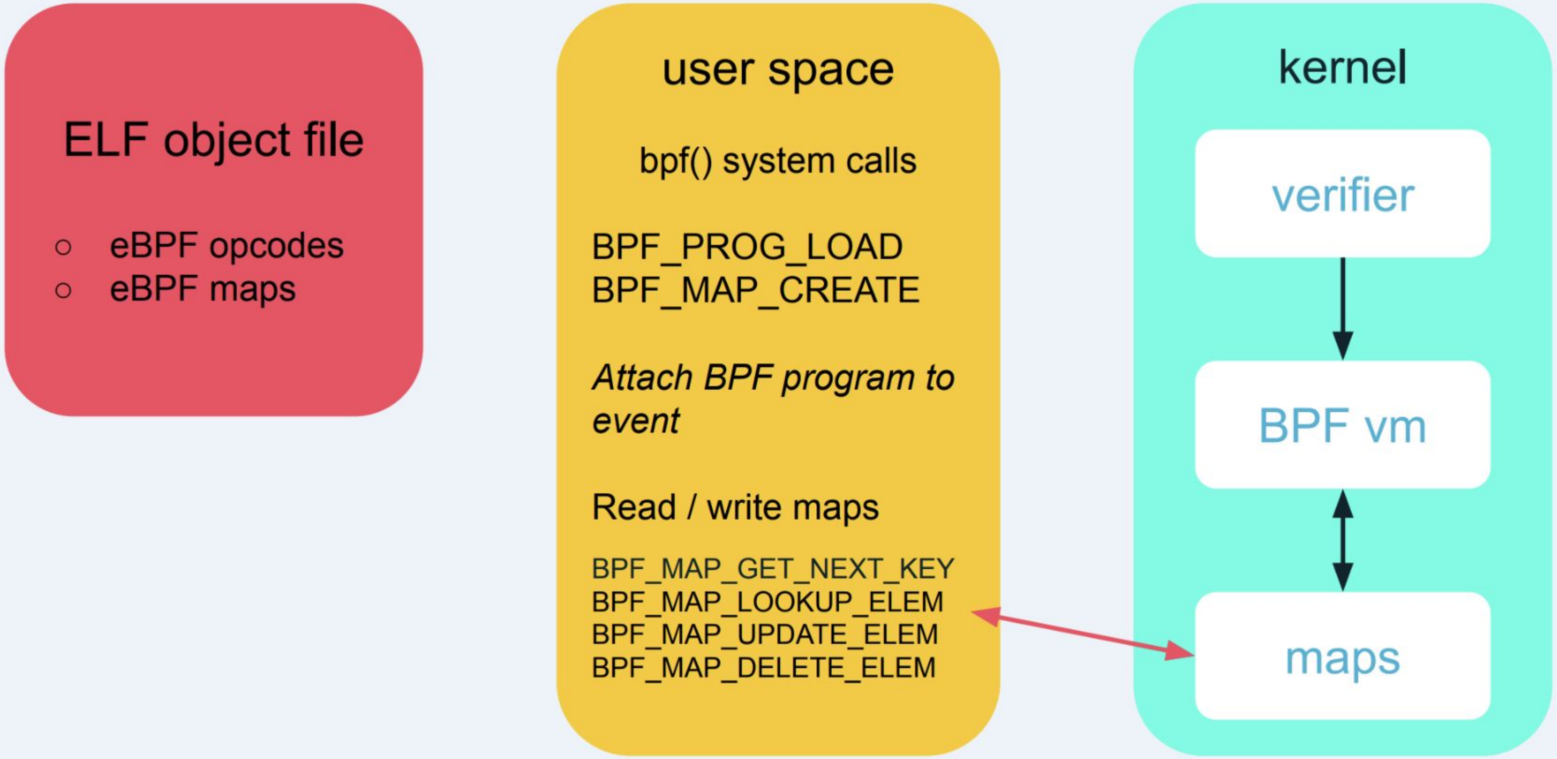
bpf() system calls

BPF_PROG_LOAD
BPF_MAP_CREATE

Attach BPF program to event

kernel





ELF object file

- eBPF opcodes
- eBPF maps

user space

bpf() system calls

BPF_PROG_LOAD
BPF_MAP_CREATE

Attach BPF program to event

Read / write maps

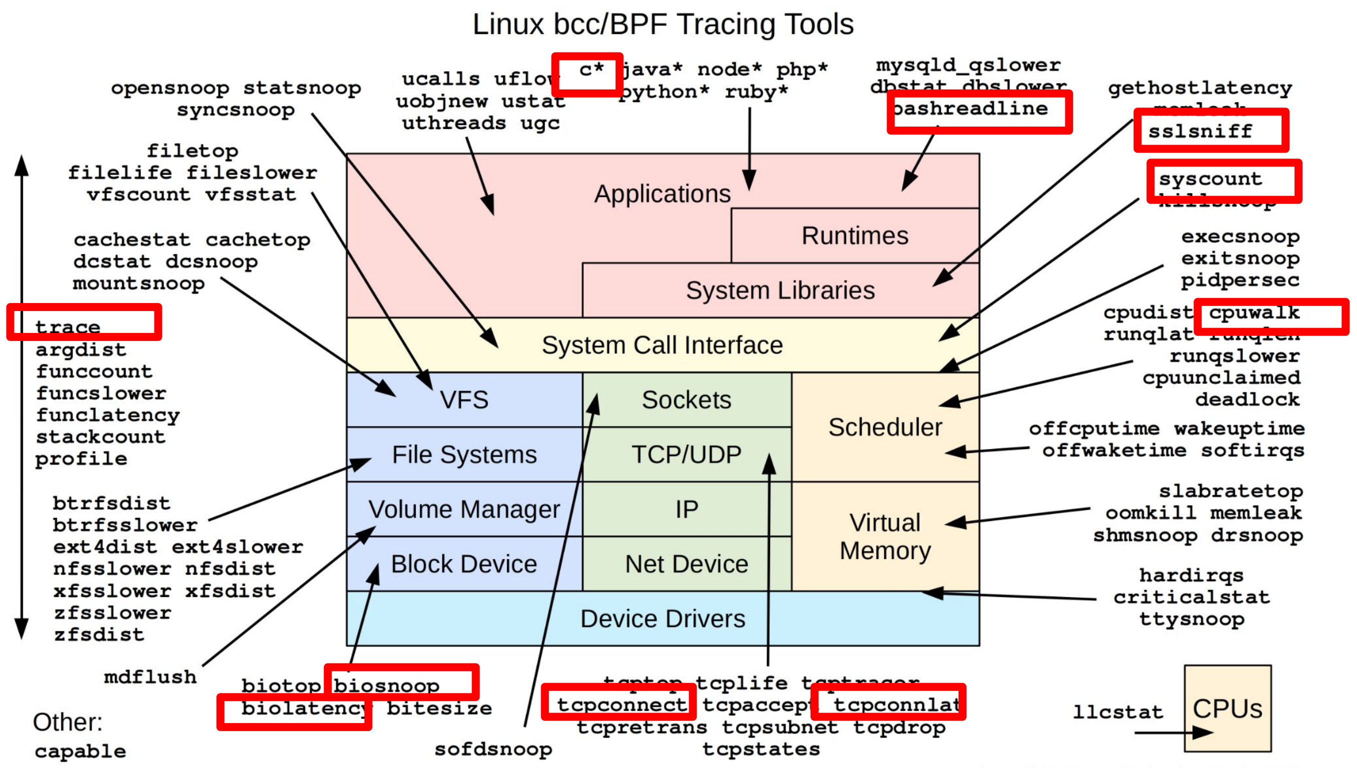
BPF_MAP_GET_NEXT_KEY
BPF_MAP_LOOKUP_ELEM
BPF_MAP_UPDATE_ELEM
BPF_MAP_DELETE_ELEM

@lizrice

aqua

Hands ON

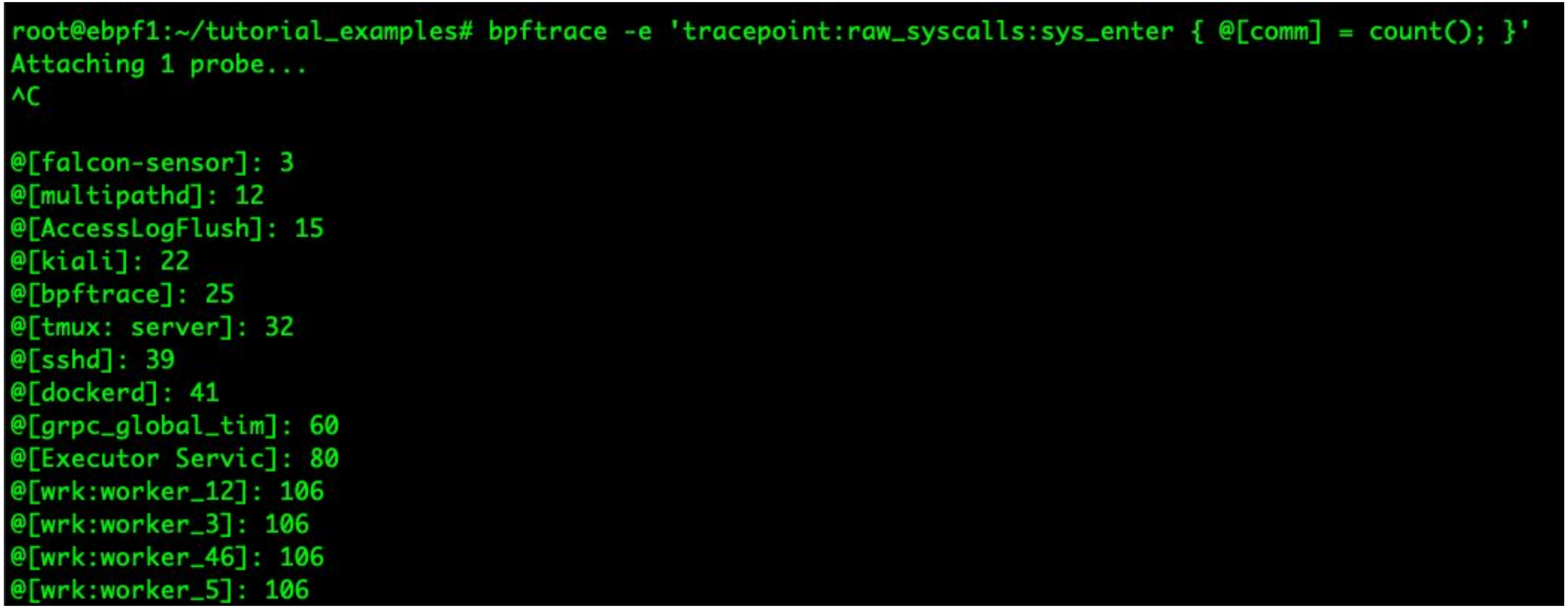
eBPF is Everywhere!!



https://github.com/iovisor/bcc#tools 2019

Tracing: Syscall Count

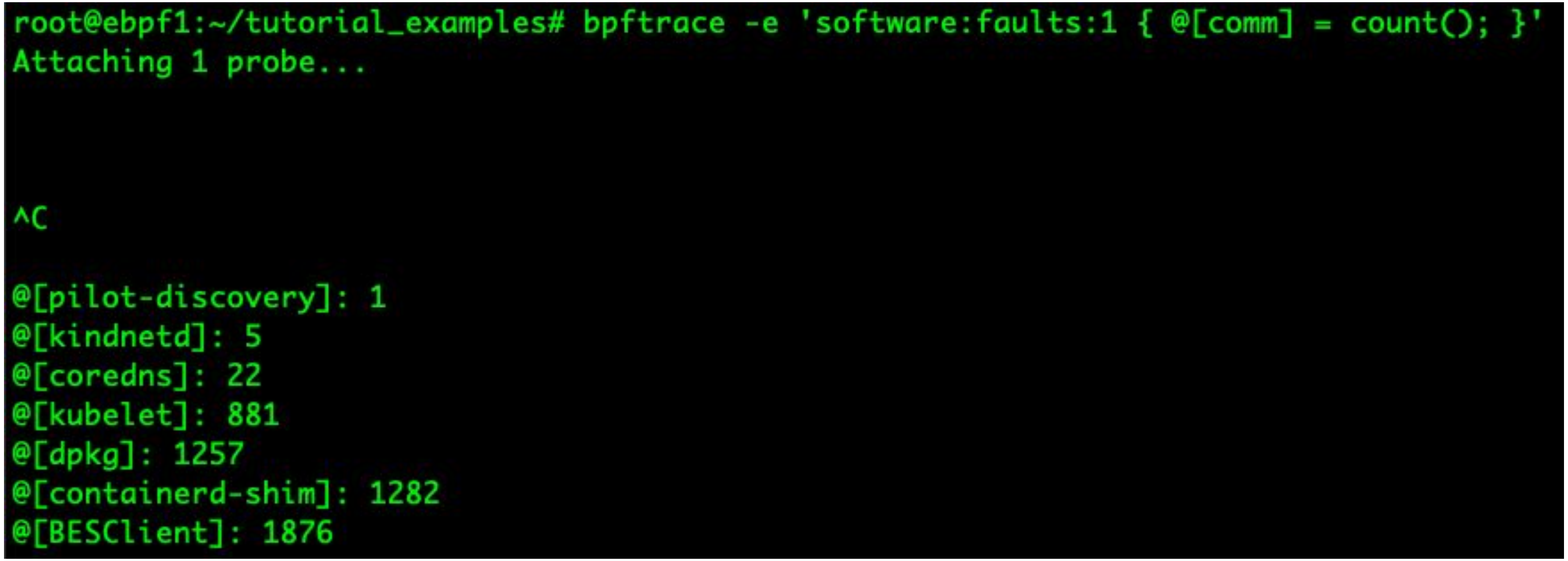
```
# Syscall count by program
  bpftrace -e 'tracepoint:raw_syscalls:sys_enter { @[comm] = count(); }'
```



```
root@ebpf1:~/tutorial_examples# bpftrace -e 'tracepoint:raw_syscalls:sys_enter { @[comm] = count(); }'
Attaching 1 probe...
^C

@[falcon-sensor]: 3
@[multipathd]: 12
@[AccessLogFlush]: 15
@[kiali]: 22
@[bpftrace]: 25
@[tmux: server]: 32
@[sshd]: 39
@[dockerd]: 41
@[grpc_global_tim]: 60
@[Executor Servic]: 80
@[wrk:worker_12]: 106
@[wrk:worker_3]: 106
@[wrk:worker_46]: 106
@[wrk:worker_5]: 106
```

Tracing: Software Faults

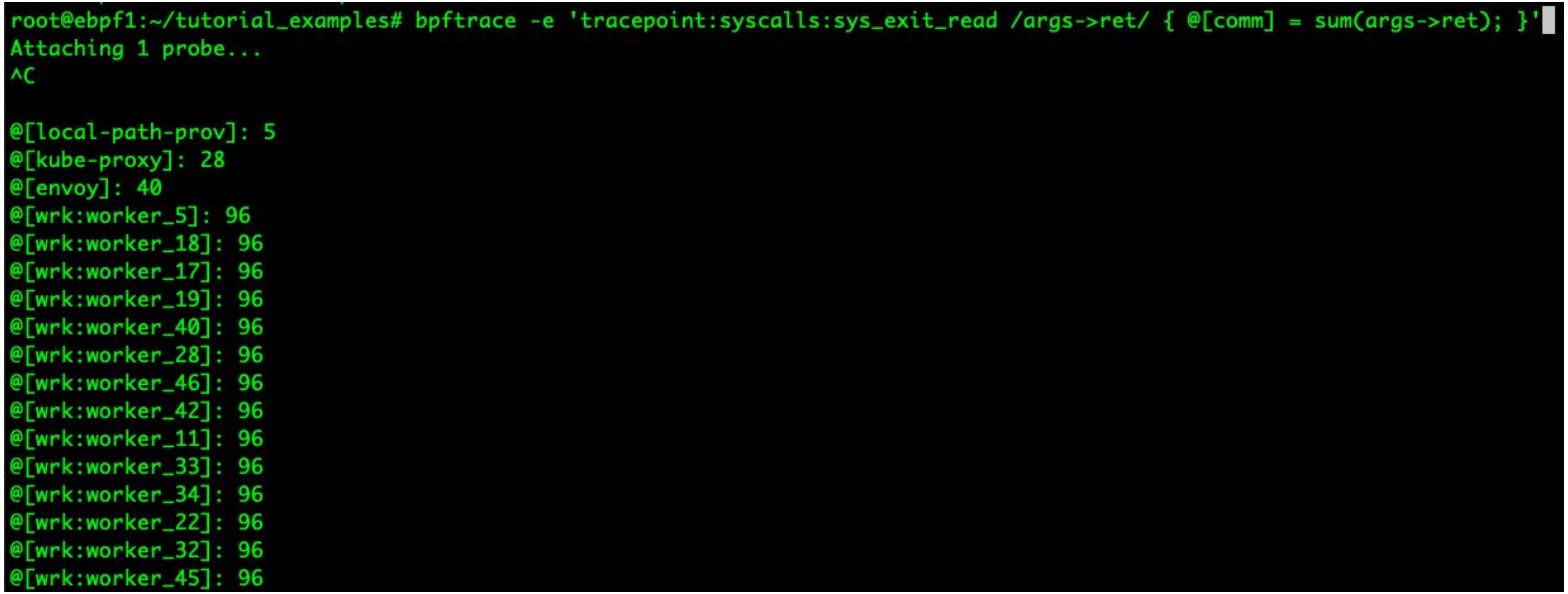
```
# Count page faults by process

  bpftrace -e 'software:faults:1 { @[comm] = count(); }'
```



```
root@ebpf1:~/tutorial_examples# bpftrace -e 'software:faults:1 { @[comm] = count(); }'
Attaching 1 probe...



^C

@[pilot-discovery]: 1
@[kindnetd]: 5
@[coredns]: 22
@[kubelet]: 881
@[dpkg]: 1257
@[containerd-shim]: 1282
@[BESClient]: 1876
```

Tracing: Read Bytes By Process

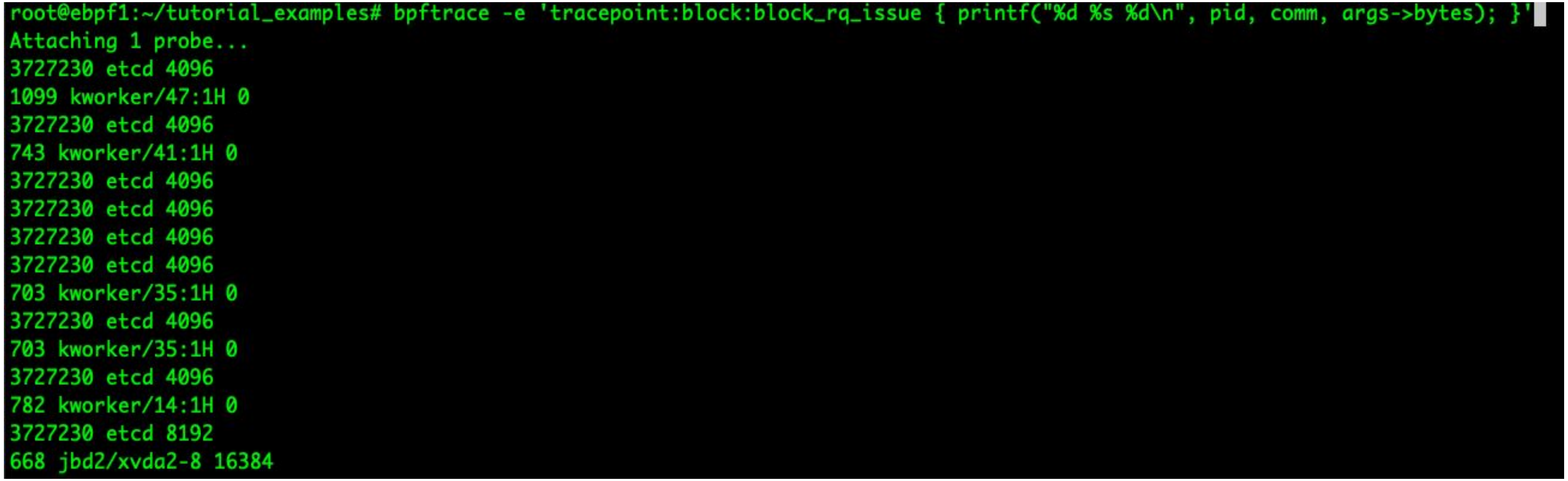
```
# Read bytes by process:

bpftrace -e 'tracepoint:syscalls:sys_exit_read /args->ret/ { @[comm] = sum(args->ret); }'
```



```
root@ebpf1:~/tutorial_examples# bpftrace -e 'tracepoint:syscalls:sys_exit_read /args->ret/ { @[comm] = sum(args->ret); }'
Attaching 1 probe...
^C

@[local-path-prov]: 5
@[kube-proxy]: 28
@[envoy]: 40
@[wrk:worker_5]: 96
@[wrk:worker_18]: 96
@[wrk:worker_17]: 96
@[wrk:worker_19]: 96
@[wrk:worker_40]: 96
@[wrk:worker_28]: 96
@[wrk:worker_46]: 96
@[wrk:worker_42]: 96
@[wrk:worker_11]: 96
@[wrk:worker_33]: 96
@[wrk:worker_34]: 96
@[wrk:worker_22]: 96
@[wrk:worker_32]: 96
@[wrk:worker_45]: 96
```

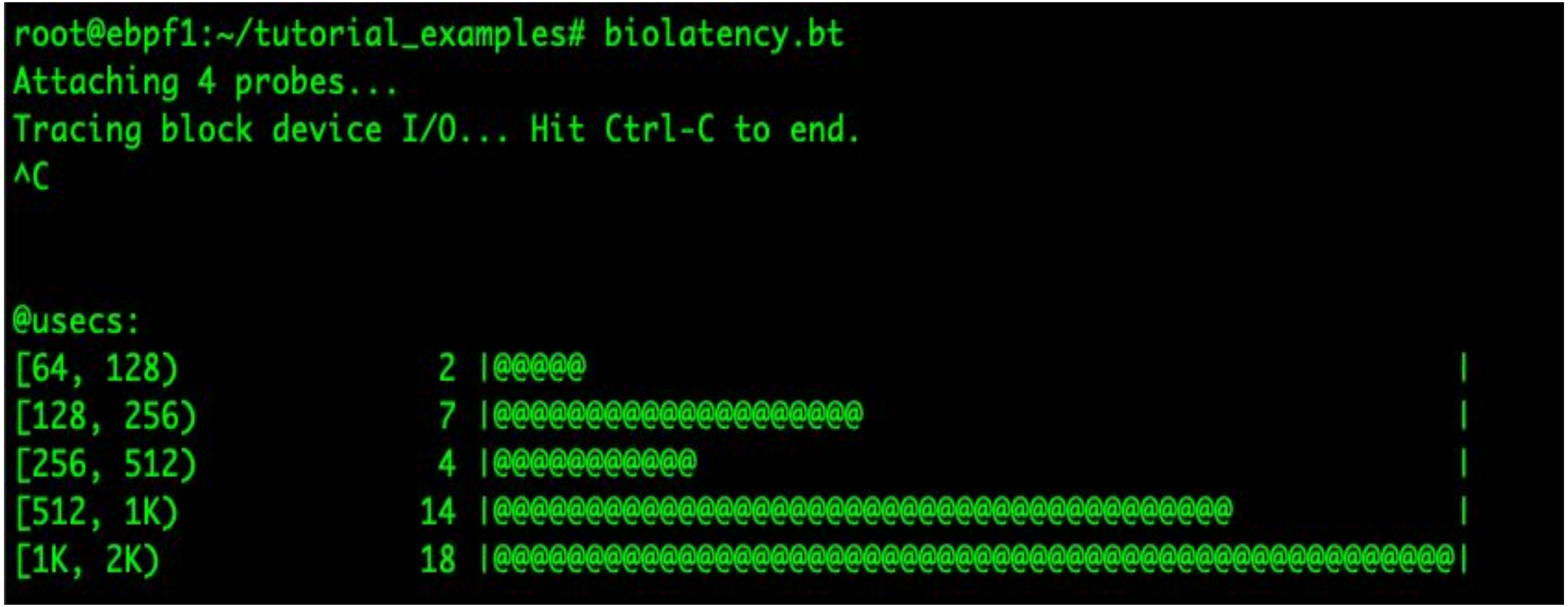
Tracing: Disk Size Read By Process

```
# Trace disk size by process

bpftrace -e 'tracepoint:block:block_rq_issue { printf("%d %s %d\n", pid, comm, args->bytes); }'
```



```
root@ebpf1:~/tutorial_examples# bpftrace -e 'tracepoint:block:block_rq_issue { printf("%d %s %d\n", pid, comm, args->bytes); }'
Attaching 1 probe...
3727230 etcd 4096
1099 kworker/47:1H 0
3727230 etcd 4096
743 kworker/41:1H 0
3727230 etcd 4096
3727230 etcd 4096
3727230 etcd 4096
3727230 etcd 4096
703 kworker/35:1H 0
3727230 etcd 4096
703 kworker/35:1H 0
3727230 etcd 4096
782 kworker/14:1H 0
3727230 etcd 8192
668 jbd2/xvda2-8 16384
```

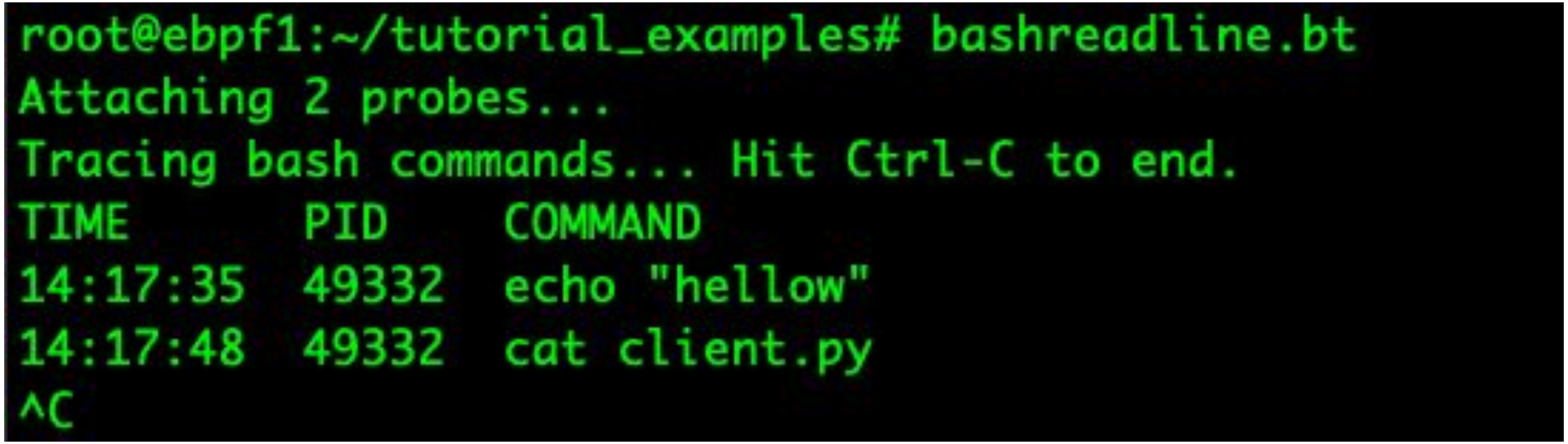
Tracing : Block IO Latency



```
root@ebpf1:~/tutorial_examples# biolatency.bt
Attaching 4 probes...
Tracing block device I/O... Hit Ctrl-C to end.
^C


@usecs:
[64, 128)              2 |@@@@@                                               |
[128, 256)             7 |@@@@@@@@@@@@@@@@@@@@                                |
[256, 512)             4 |@@@@@@@@@@@                                         |
[512, 1K)             14 |@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@            |
[1K, 2K)              18 |@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@|
```

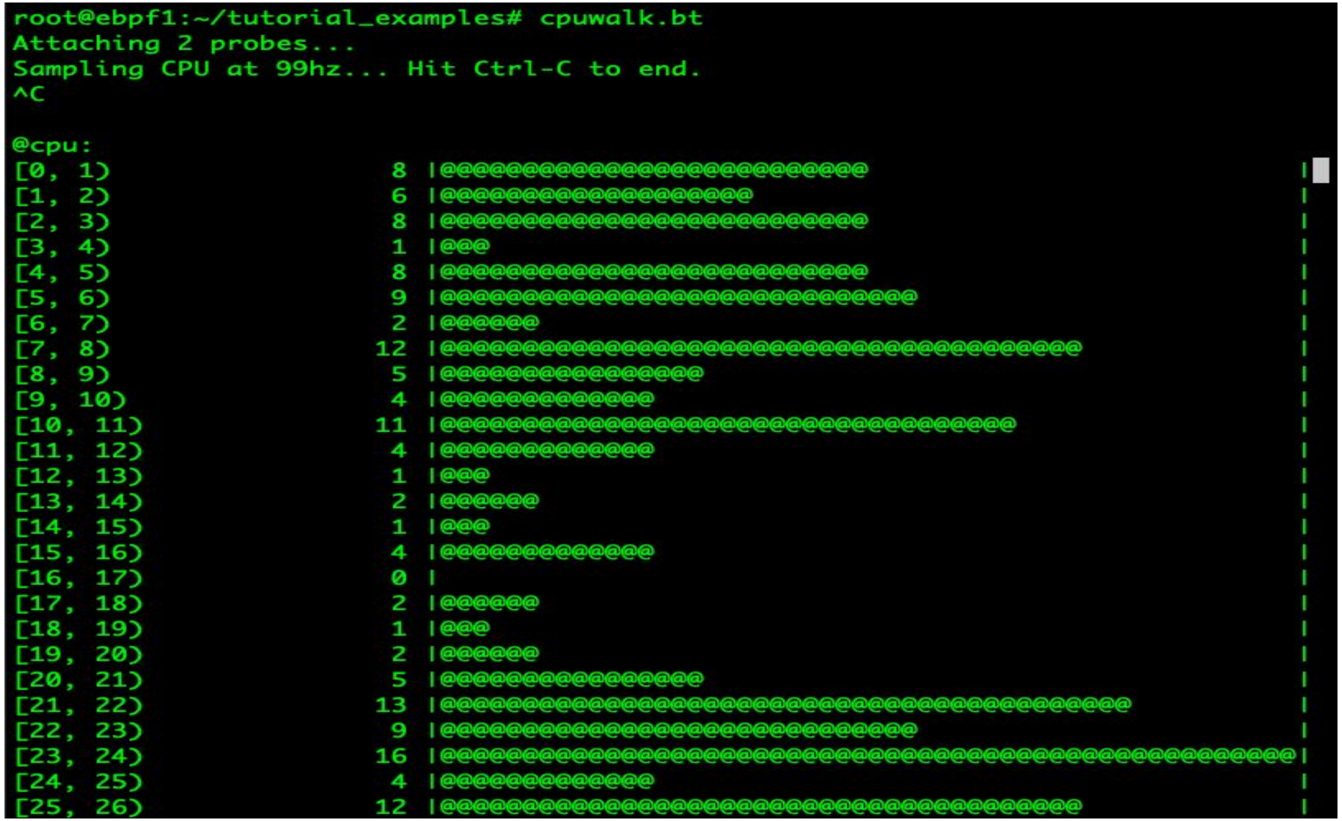
URetProbe : Bash ReadLine

#Print entered bash commands system wide
bashreadline.bt



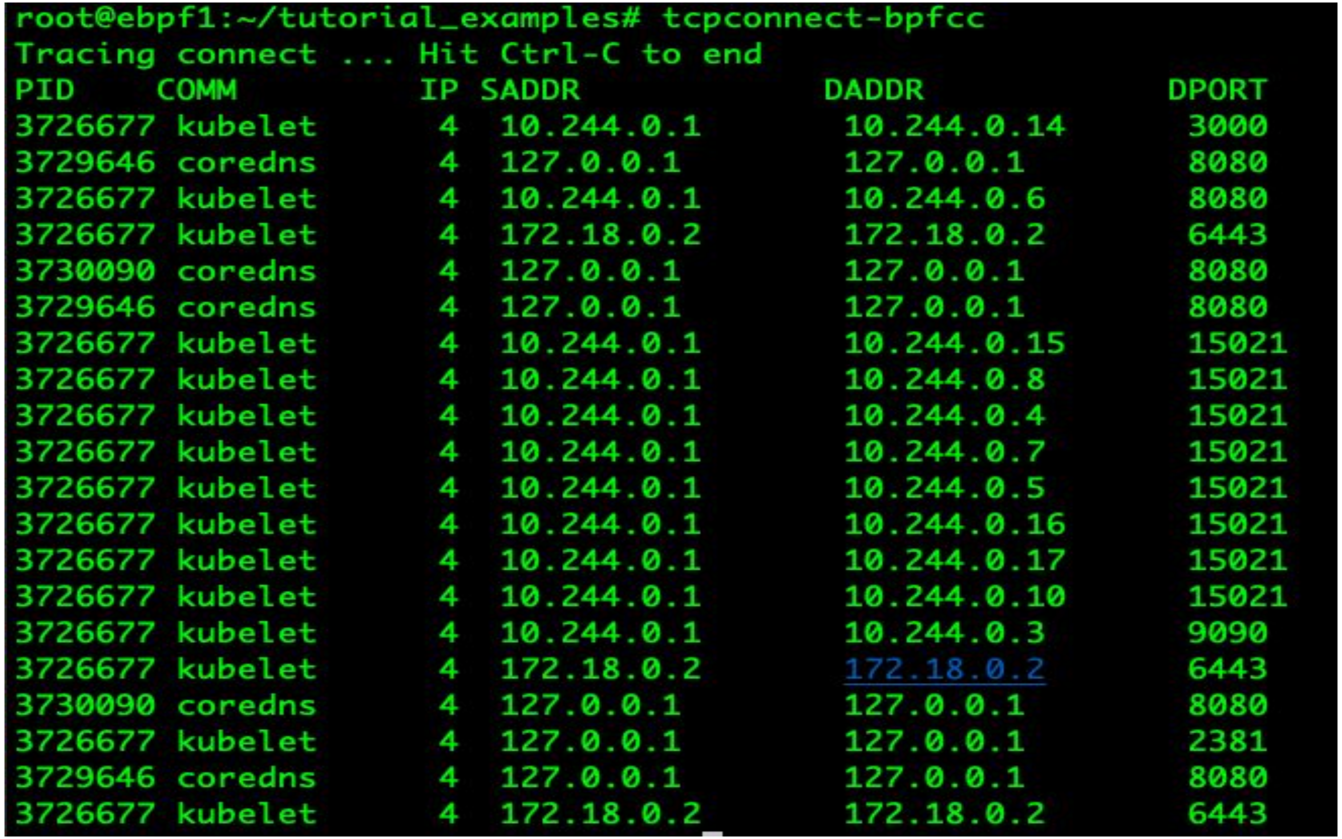
```
root@ebpf1:~/tutorial_examples# bashreadline.bt
Attaching 2 probes...
Tracing bash commands... Hit Ctrl-C to end.
TIME      PID    COMMAND
14:17:35  49332  echo "hellow"
14:17:48  49332  cat client.py
^C
```

Tracing: Sample which CPUs are executing processes



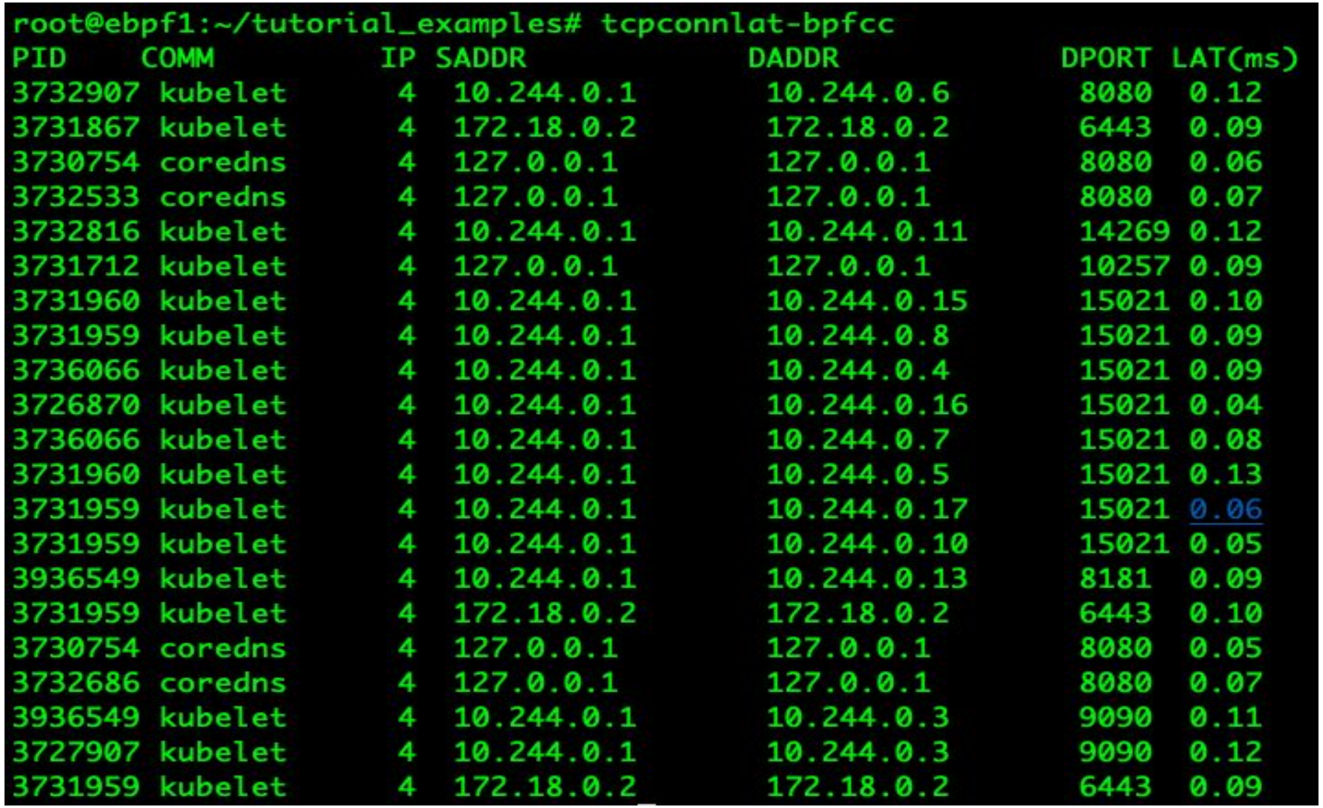
```
root@ebpf1:~/tutorial_examples# cpuwalk.bt
Attaching 2 probes...
Sampling CPU at 99hz... Hit Ctrl-C to end.
^C

@cpu:
[0, 1)                 8 |@@@@@@@@@@@@@@@@@@@@@@@@@@                          |
[1, 2)                 6 |@@@@@@@@@@@@@@@@@@@                                 |
[2, 3)                 8 |@@@@@@@@@@@@@@@@@@@@@@@@@@                          |
[3, 4)                 1 |@@@                                                 |
[4, 5)                 8 |@@@@@@@@@@@@@@@@@@@@@@@@@@                          |
[5, 6)                 9 |@@@@@@@@@@@@@@@@@@@@@@@@@@@@@                       |
[6, 7)                 2 |@@@@@@                                              |
[7, 8)                12 |@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@             |
[8, 9)                 5 |@@@@@@@@@@@@@@@@                                    |
[9, 10)                4 |@@@@@@@@@@@@@                                       |
[10, 11)              11 |@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@                 |
[11, 12)               4 |@@@@@@@@@@@@@                                       |
[12, 13)               1 |@@@                                                 |
[13, 14)               2 |@@@@@@                                              |
[14, 15)               1 |@@@                                                 |
[15, 16)               4 |@@@@@@@@@@@@@                                       |
[16, 17)               0 |                                                    |
[17, 18)               2 |@@@@@@                                              |
[18, 19)               1 |@@@                                                 |
[19, 20)               2 |@@@@@@                                              |
[20, 21)               5 |@@@@@@@@@@@@@@@@                                    |
[21, 22)              13 |@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@          |
[22, 23)               9 |@@@@@@@@@@@@@@@@@@@@@@@@@@@@@                       |
[23, 24)              16 |@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@|
[24, 25)               4 |@@@@@@@@@@@@@                                       |
[25, 26)              12 |@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@             |
```

Tracing: Active TCP Connections



```
root@ebpf1:~/tutorial_examples# tcpconnect-bpfcc
Tracing connect ... Hit Ctrl-C to end
PID     COMM         IP SADDR            DADDR            DPORT
3726677 kubelet       4  10.244.0.1       10.244.0.14      3000
3729646 coredns       4  127.0.0.1        127.0.0.1        8080
3726677 kubelet       4  10.244.0.1       10.244.0.6       8080
3726677 kubelet       4  172.18.0.2       172.18.0.2       6443
3730090 coredns       4  127.0.0.1        127.0.0.1        8080
3729646 coredns       4  127.0.0.1        127.0.0.1        8080
3726677 kubelet       4  10.244.0.1       10.244.0.15      15021
3726677 kubelet       4  10.244.0.1       10.244.0.8       15021
3726677 kubelet       4  10.244.0.1       10.244.0.4       15021
3726677 kubelet       4  10.244.0.1       10.244.0.7       15021
3726677 kubelet       4  10.244.0.1       10.244.0.5       15021
3726677 kubelet       4  10.244.0.1       10.244.0.16      15021
3726677 kubelet       4  10.244.0.1       10.244.0.17      15021
3726677 kubelet       4  10.244.0.1       10.244.0.10      15021
3726677 kubelet       4  10.244.0.1       10.244.0.3       9090
3726677 kubelet       4  172.18.0.2       172.18.0.2       6443
3730090 coredns       4  127.0.0.1        127.0.0.1        8080
3726677 kubelet       4  127.0.0.1        127.0.0.1        2381
3729646 coredns       4  127.0.0.1        127.0.0.1        8080
3726677 kubelet       4  172.18.0.2       172.18.0.2       6443
```

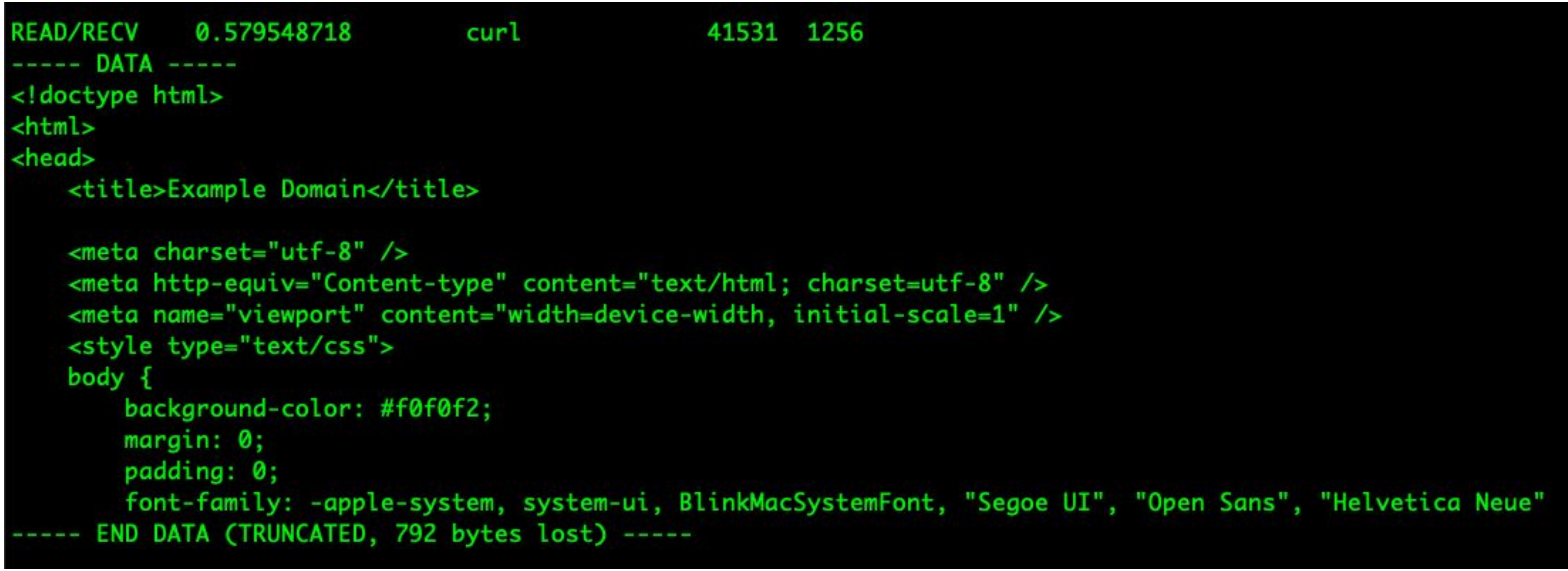
TCP Connection Latency



```
root@ebpf1:~/tutorial_examples# tcpconnlat-bpfcc
PID     COMM         IP SADDR            DADDR            DPORT LAT(ms)
3732907 kubelet       4  10.244.0.1       10.244.0.6       8080  0.12
3731867 kubelet       4  172.18.0.2       172.18.0.2       6443  0.09
3730754 coredns       4  127.0.0.1        127.0.0.1        8080  0.06
3732533 coredns       4  127.0.0.1        127.0.0.1        8080  0.07
3732816 kubelet       4  10.244.0.1       10.244.0.11      14269 0.12
3731712 kubelet       4  127.0.0.1        127.0.0.1        10257 0.09
3731960 kubelet       4  10.244.0.1       10.244.0.15      15021 0.10
3731959 kubelet       4  10.244.0.1       10.244.0.8       15021 0.09
3736066 kubelet       4  10.244.0.1       10.244.0.4       15021 0.09
3726870 kubelet       4  10.244.0.1       10.244.0.16      15021 0.04
3736066 kubelet       4  10.244.0.1       10.244.0.7       15021 0.08
3731960 kubelet       4  10.244.0.1       10.244.0.5       15021 0.13
3731959 kubelet       4  10.244.0.1       10.244.0.17      15021 0.06
3731959 kubelet       4  10.244.0.1       10.244.0.10      15021 0.05
3936549 kubelet       4  10.244.0.1       10.244.0.13      8181  0.09
3731959 kubelet       4  172.18.0.2       172.18.0.2       6443  0.10
3730754 coredns       4  127.0.0.1        127.0.0.1        8080  0.05
3732686 coredns       4  127.0.0.1        127.0.0.1        8080  0.07
3936549 kubelet       4  10.244.0.1       10.244.0.3       9090  0.11
3727907 kubelet       4  10.244.0.1       10.244.0.3       9090  0.12
3731959 kubelet       4  172.18.0.2       172.18.0.2       6443  0.09
```

Tracing SSL Data

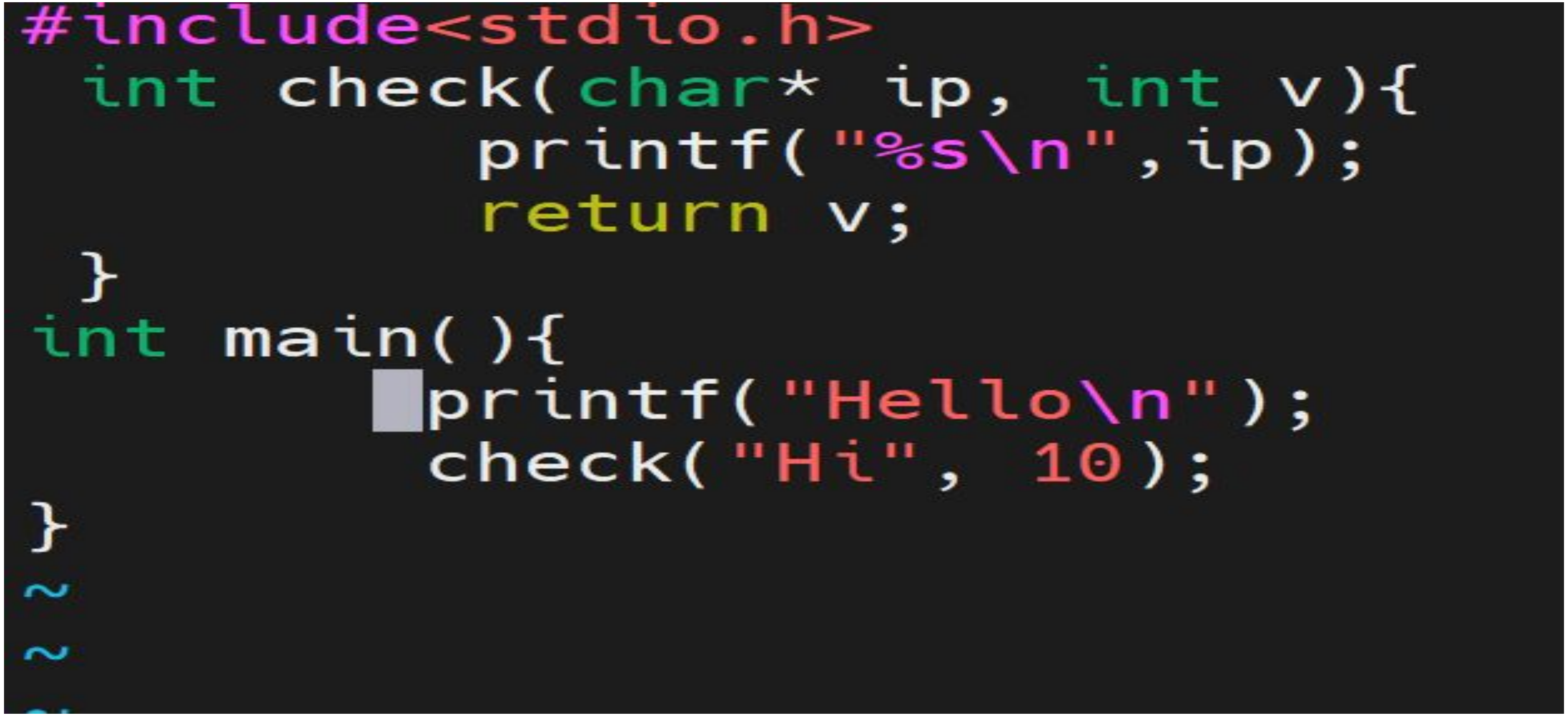
sslsniff-bpfcc



```
READ/RECV    0.579548718        curl             41531  1256
----- DATA -----
<!doctype html>
<html>
<head>
    <title>Example Domain</title>

    <meta charset="utf-8" />
    <meta http-equiv="Content-type" content="text/html; charset=utf-8" />
    <meta name="viewport" content="width=device-width, initial-scale=1" />
    <style type="text/css">
    body {
        background-color: #f0f0f2;
        margin: 0;
        padding: 0;
        font-family: -apple-system, system-ui, BlinkMacSystemFont, "Segoe UI", "Open Sans", "Helvetica Neue"
----- END DATA (TRUNCATED, 792 bytes lost) -----
```

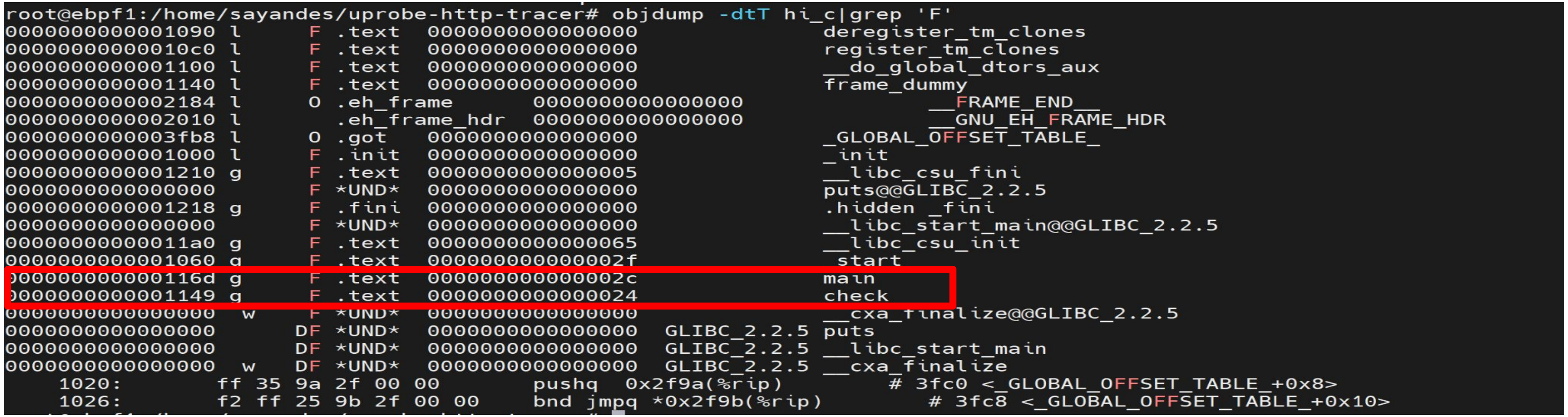
Tracing UserLand Code

Tracing With Uprobes: C code

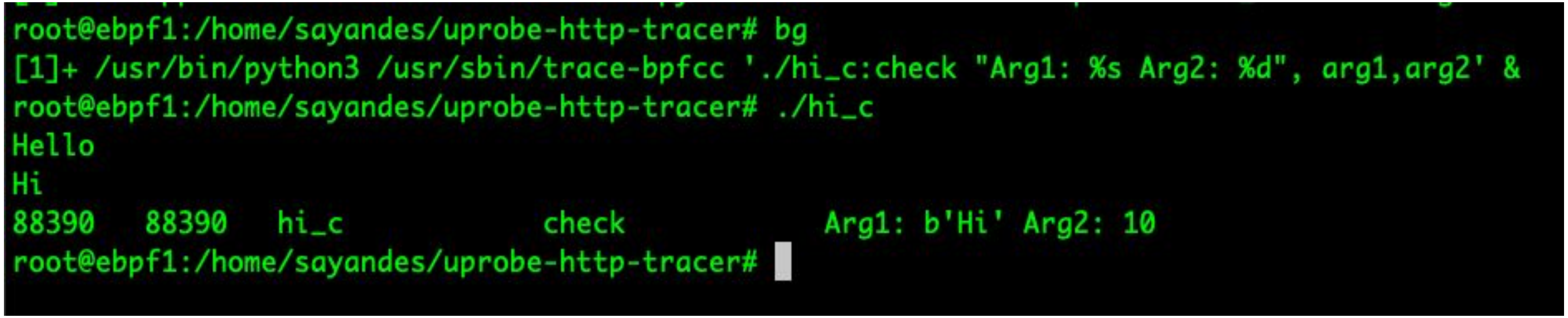


```
#include<stdio.h>
 int check(char* ip, int v){
          printf("%s\n",ip);
          return v;
 }
int main(){
         printf("Hello\n");
         check("Hi", 10);
}
~
~
```

Tracing With Uprobes:

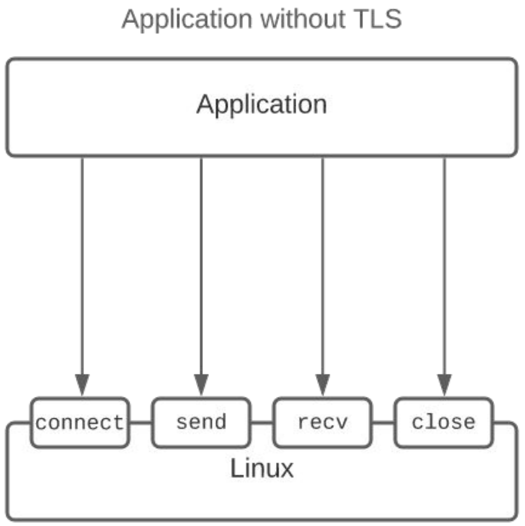


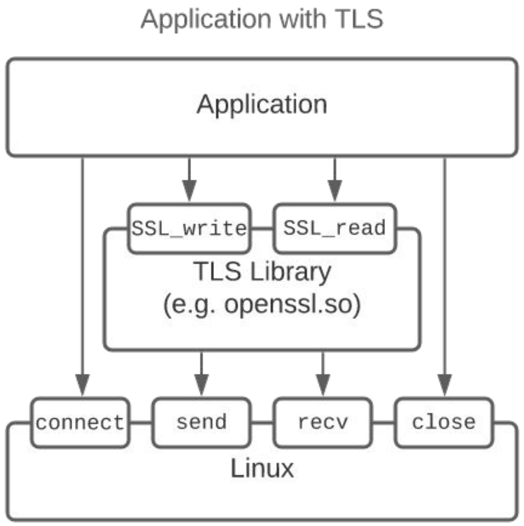
```
root@ebpf1:/home/sayandes/uprobe-http-tracer# objdump -dtT hi_c|grep 'F'
0000000000001090 l     F .text  0000000000000000              deregister_tm_clones
00000000000010c0 l     F .text  0000000000000000              register_tm_clones
0000000000001100 l     F .text  0000000000000000              __do_global_dtors_aux
0000000000001140 l     F .text  0000000000000000              frame_dummy
0000000000002184 l     O .eh_frame      0000000000000000              __FRAME_END__
0000000000002010 l       .eh_frame_hdr  0000000000000000              __GNU_EH_FRAME_HDR
0000000000003fb8 l     O .got   0000000000000000              _GLOBAL_OFFSET_TABLE_
0000000000001000 l     F .init  0000000000000000              _init
0000000000001210 g     F .text  0000000000000005              __libc_csu_fini
0000000000000000       F *UND*  0000000000000000              puts@@GLIBC_2.2.5
0000000000001218 g     F .fini  0000000000000000              .hidden _fini
0000000000000000       F *UND*  0000000000000000              __libc_start_main@@GLIBC_2.2.5
00000000000011a0 g     F .text  0000000000000065              __libc_csu_init
0000000000001060 g     F .text  000000000000002f              _start
000000000000116d g     F .text  000000000000002c              main
0000000000001149 g     F .text  0000000000000024              check
0000000000000000  w    F *UND*  0000000000000000              __cxa_finalize@@GLIBC_2.2.5
0000000000000000      DF *UND*  0000000000000000  GLIBC_2.2.5 puts
0000000000000000      DF *UND*  0000000000000000  GLIBC_2.2.5 __libc_start_main
0000000000000000  w   DF *UND*  0000000000000000  GLIBC_2.2.5 __cxa_finalize
    1020:       ff 35 9a 2f 00 00       pushq  0x2f9a(%rip)        # 3fc0 <_GLOBAL_OFFSET_TABLE_+0x8>
    1026:       f2 ff 25 9b 2f 00 00    bnd jmpq *0x2f9b(%rip)        # 3fc8 <_GLOBAL_OFFSET_TABLE_+0x10>
```



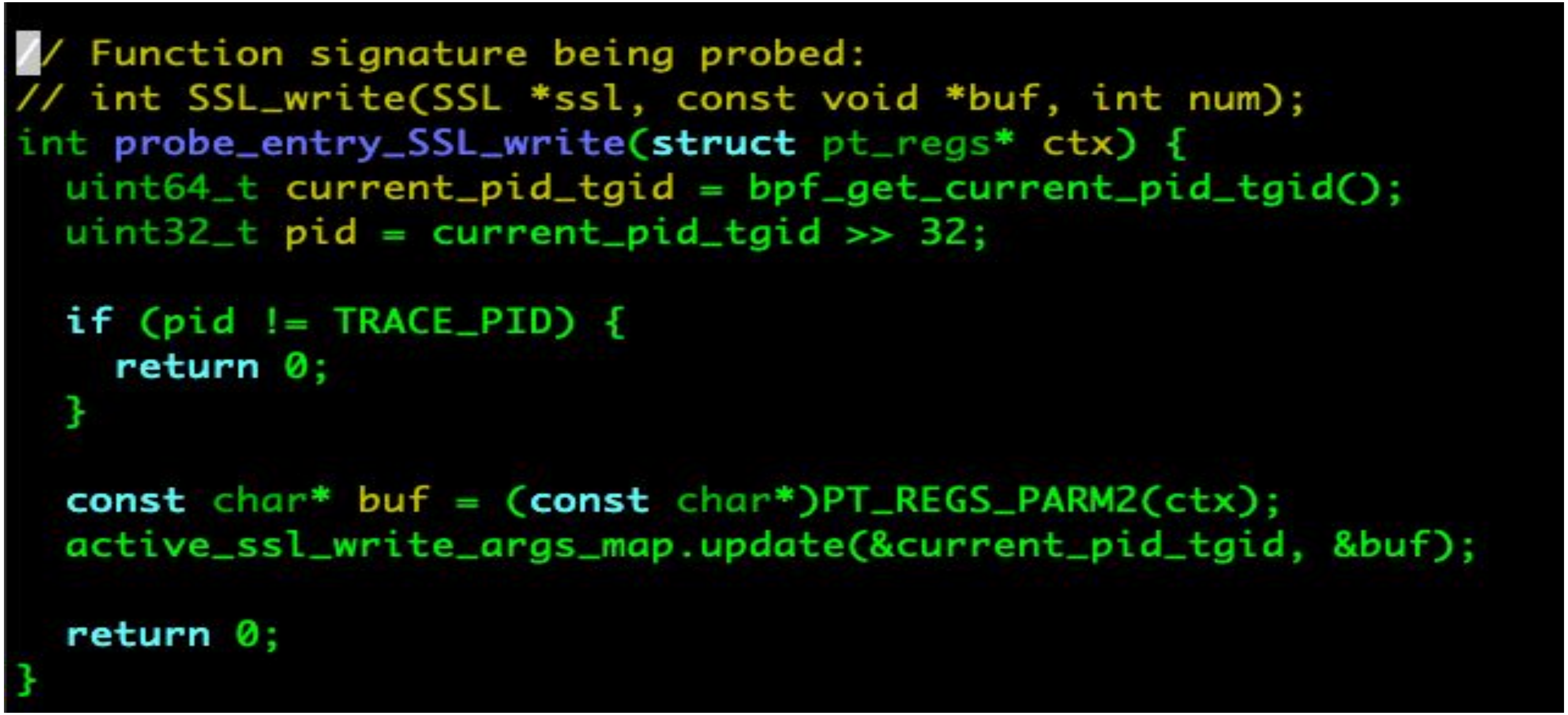
```
root@ebpf1:/home/sayandes/uprobe-http-tracer# bg
[1]+ /usr/bin/python3 /usr/sbin/trace-bpfcc './hi_c:check "Arg1: %s Arg2: %d", arg1,arg2' &
root@ebpf1:/home/sayandes/uprobe-http-tracer# ./hi_c
Hello
Hi
88390   88390   hi_c            check           Arg1: b'Hi' Arg2: 10
root@ebpf1:/home/sayandes/uprobe-http-tracer# 
```

Tracing TLS Connections



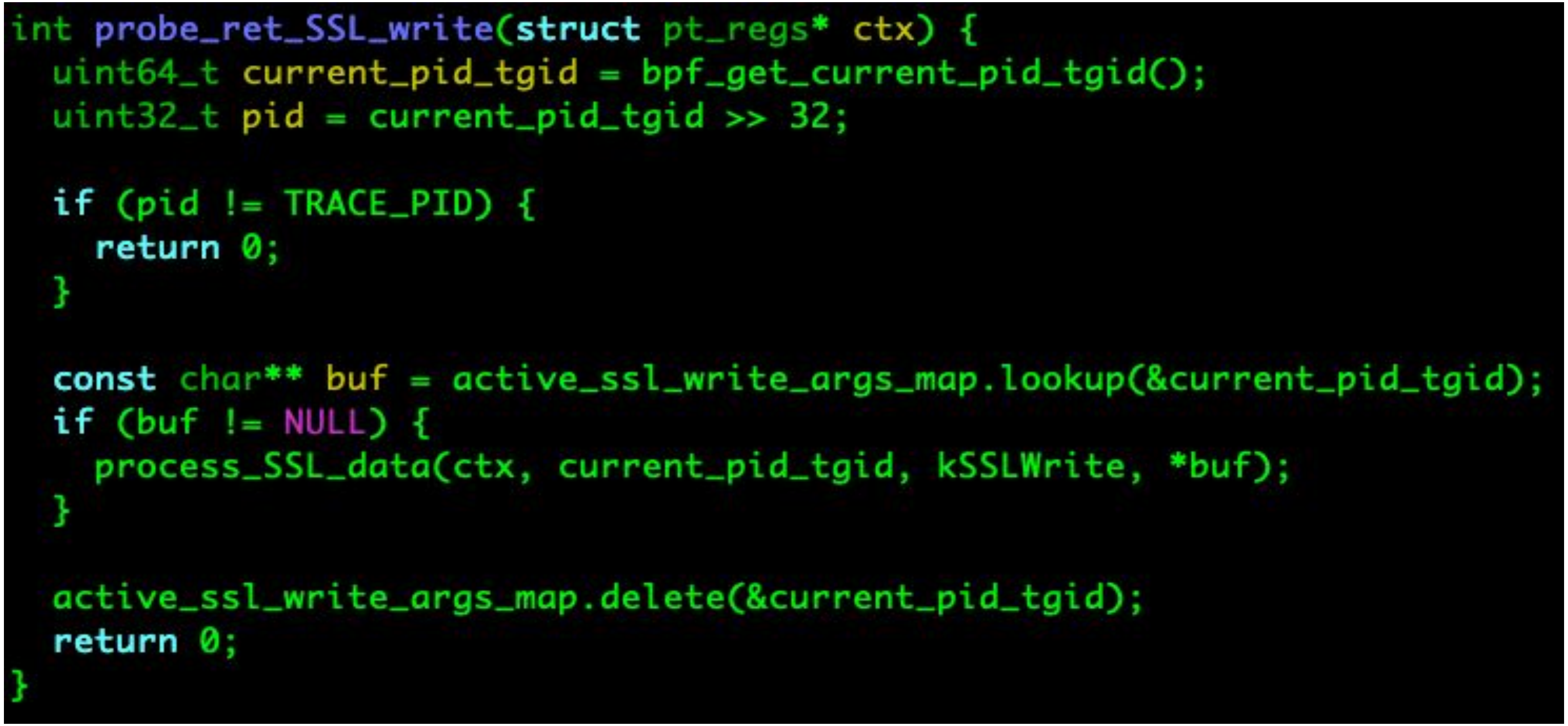


On SSL Function Entry



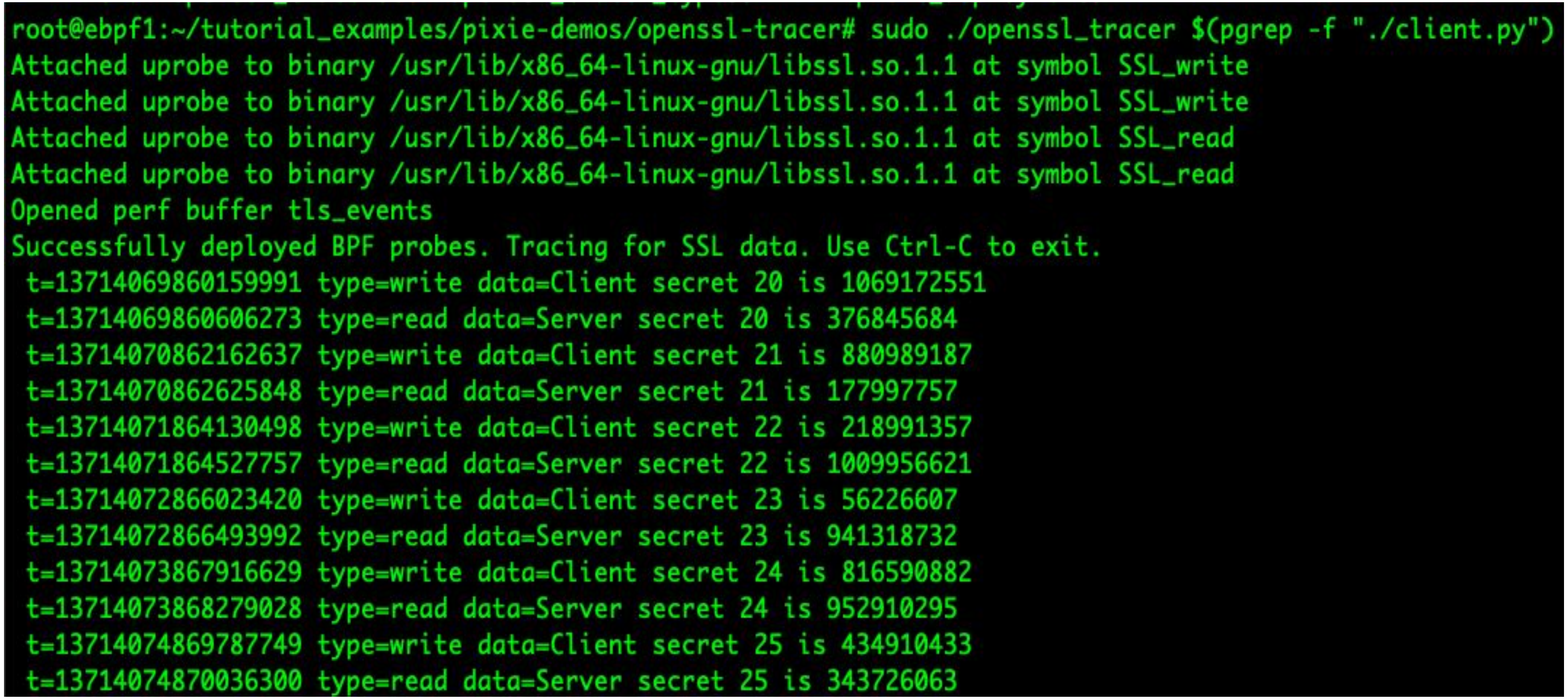
```
// Function signature being probed:
// int SSL_write(SSL *ssl, const void *buf, int num);
int probe_entry_SSL_write(struct pt_regs* ctx) {
  uint64_t current_pid_tgid = bpf_get_current_pid_tgid();
  uint32_t pid = current_pid_tgid >> 32;

  if (pid != TRACE_PID) {
    return 0;
  }

  const char* buf = (const char*)PT_REGS_PARM2(ctx);
  active_ssl_write_args_map.update(&current_pid_tgid, &buf);

  return 0;
}
```

On SSL Function Return



```
int probe_ret_SSL_write(struct pt_regs* ctx) {
  uint64_t current_pid_tgid = bpf_get_current_pid_tgid();
  uint32_t pid = current_pid_tgid >> 32;

  if (pid != TRACE_PID) {
    return 0;
  }

  const char** buf = active_ssl_write_args_map.lookup(&current_pid_tgid);
  if (buf != NULL) {
    process_SSL_data(ctx, current_pid_tgid, kSSLWrite, *buf);
  }

  active_ssl_write_args_map.delete(&current_pid_tgid);
  return 0;
}
```

Tracing OpenSSL with Uprobes



```
root@ebpf1:~/tutorial_examples/pixie-demos/openssl-tracer# sudo ./openssl_tracer $(pgrep -f "./client.py")
Attached uprobe to binary /usr/lib/x86_64-linux-gnu/libssl.so.1.1 at symbol SSL_write
Attached uprobe to binary /usr/lib/x86_64-linux-gnu/libssl.so.1.1 at symbol SSL_write
Attached uprobe to binary /usr/lib/x86_64-linux-gnu/libssl.so.1.1 at symbol SSL_read
Attached uprobe to binary /usr/lib/x86_64-linux-gnu/libssl.so.1.1 at symbol SSL_read
Opened perf buffer tls_events
Successfully deployed BPF probes. Tracing for SSL data. Use Ctrl-C to exit.
 t=13714069860159991 type=write data=Client secret 20 is 1069172551
 t=13714069860606273 type=read data=Server secret 20 is 376845684
 t=13714070862162637 type=write data=Client secret 21 is 880989187
 t=13714070862625848 type=read data=Server secret 21 is 177997757
 t=13714071864130498 type=write data=Client secret 22 is 218991357
 t=13714071864527757 type=read data=Server secret 22 is 1009956621
 t=13714072866023420 type=write data=Client secret 23 is 56226607
 t=13714072866493992 type=read data=Server secret 23 is 941318732
 t=13714073867916629 type=write data=Client secret 24 is 816590882
 t=13714073868279028 type=read data=Server secret 24 is 952910295
 t=13714074869787749 type=write data=Client secret 25 is 434910433
 t=13714074870036300 type=read data=Server secret 25 is 343726063
```

Use Cases

1. K8s Observability
2. Network Acceleration

Use Cases

1. **K8s Observability**
2. Network Acceleration

Pixie : Monitoring Kubernetes Clusters

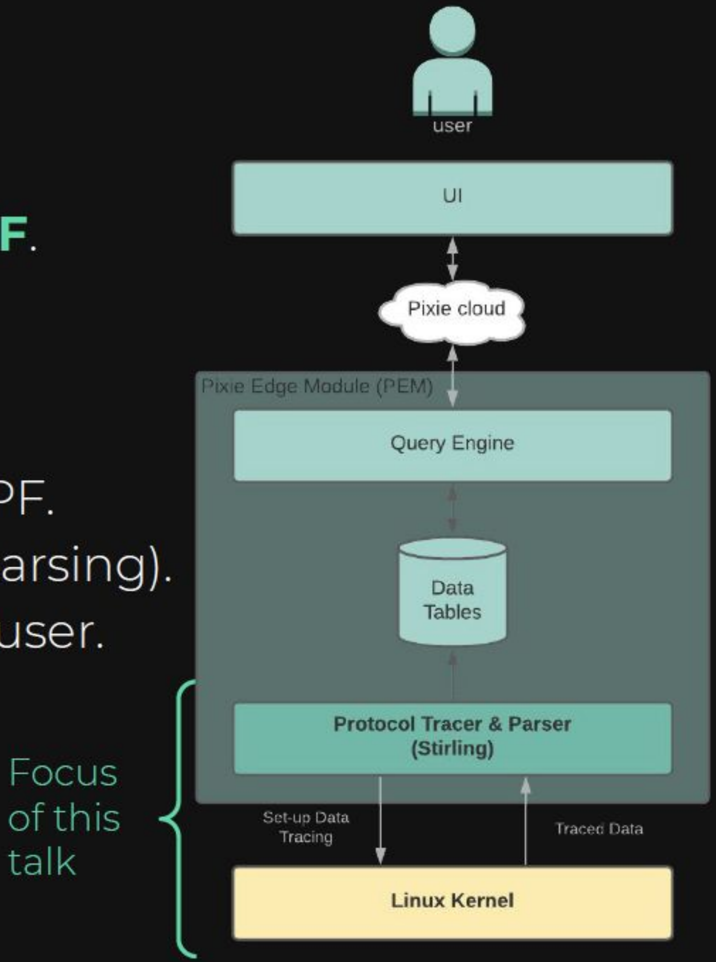
Slides From: https://lpc.events/event/11/contributions/945/

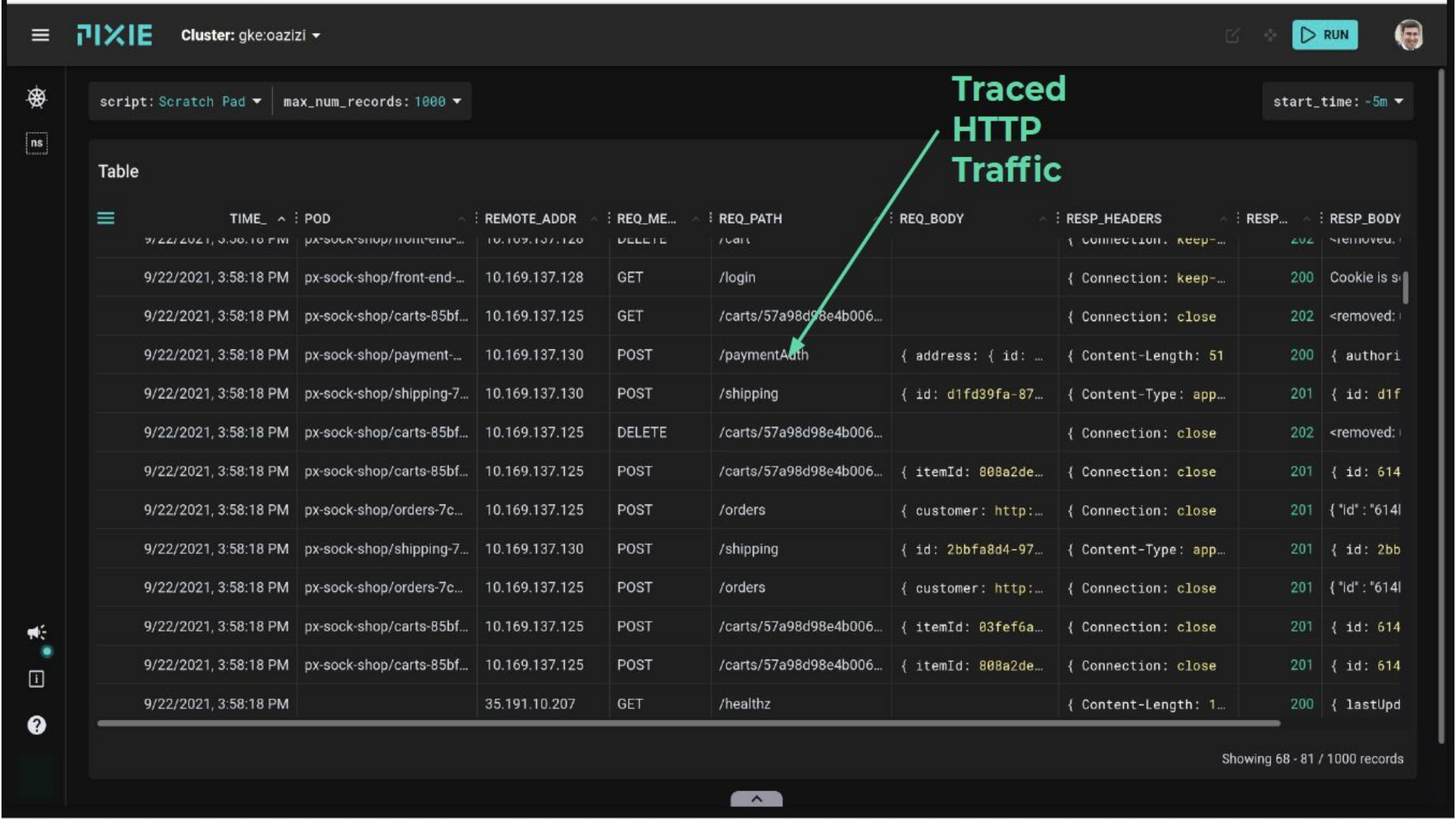
Overview

No instrumentation + low overhead ⇨ **eBPF**.

General approach:

- Capture data in kernel-space with eBPF.
- Process data in user-space (protocol parsing).
- Store data into tables for querying by user.



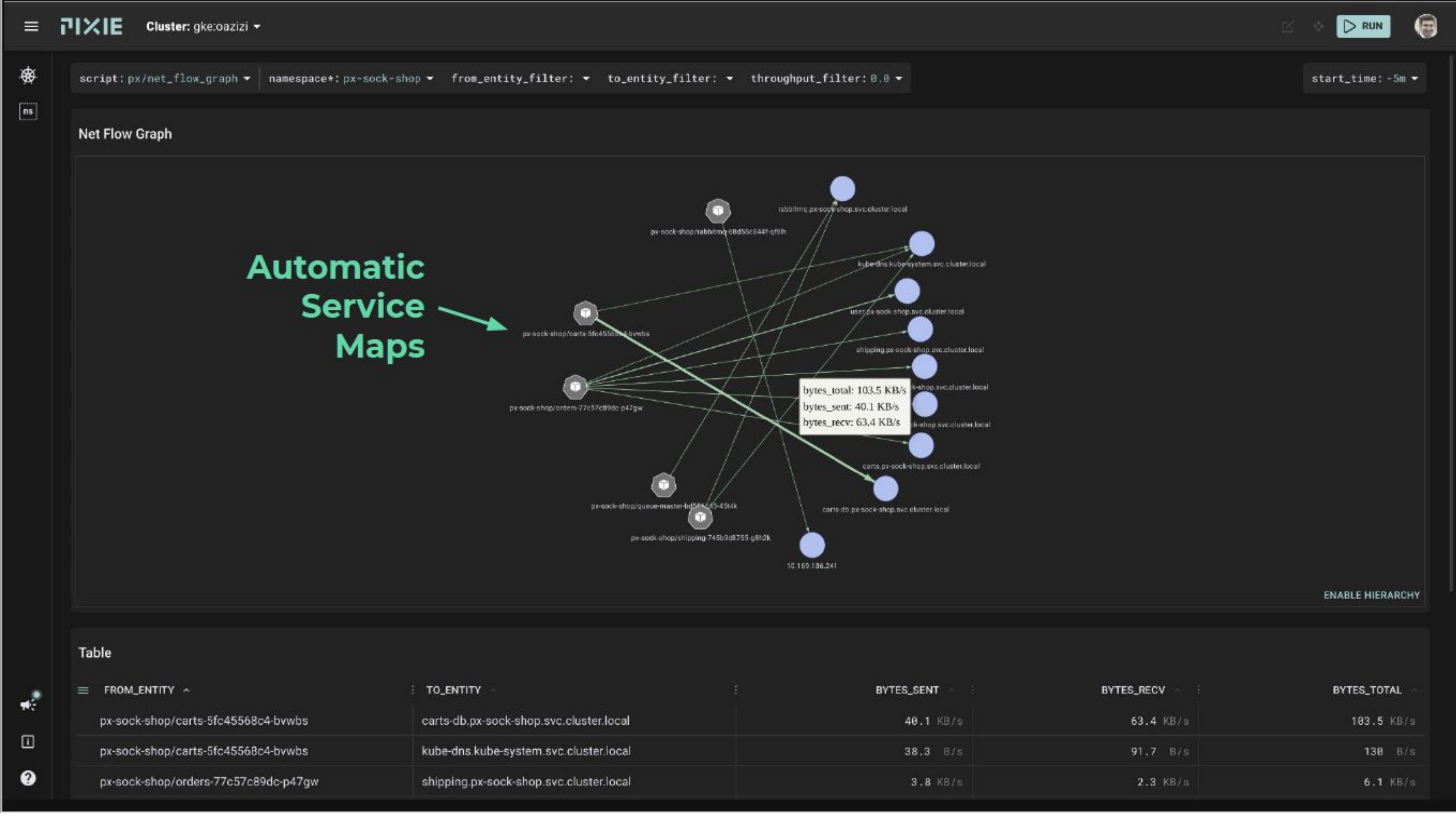


script: Scratch Pad max_num_records: 1000 start_time: -5m

Table

TIME_	POD	REMOTE_ADDR	REQ_ME...	REQ_PATH	REQ_BODY	RESP_HEADERS	RESP...	RESP_BODY
9/22/2021, 3:58:18 PM	px-sock-shop/front-end-...	10.169.137.128	DELETE	/cart		{ Connection: keep-...	202	<removed:
9/22/2021, 3:58:18 PM	px-sock-shop/front-end-...	10.169.137.128	GET	/login		{ Connection: keep-...	200	Cookie is s
9/22/2021, 3:58:18 PM	px-sock-shop/carts-85bf...	10.169.137.125	GET	/carts/57a98d98e4b006...		{ Connection: close	202	<removed:
9/22/2021, 3:58:18 PM	px-sock-shop/payment-...	10.169.137.130	POST	/paymentAuth	{ address: { id: ...	{ Content-Length: 51	200	{ authori
9/22/2021, 3:58:18 PM	px-sock-shop/shipping-7...	10.169.137.130	POST	/shipping	{ id: d1fd39fa-87...	{ Content-Type: app...	201	{ id: d1f
9/22/2021, 3:58:18 PM	px-sock-shop/carts-85bf...	10.169.137.125	DELETE	/carts/57a98d98e4b006...		{ Connection: close	202	<removed:
9/22/2021, 3:58:18 PM	px-sock-shop/carts-85bf...	10.169.137.125	POST	/carts/57a98d98e4b006...	{ itemId: 808a2de...	{ Connection: close	201	{ id: 614
9/22/2021, 3:58:18 PM	px-sock-shop/orders-7c...	10.169.137.125	POST	/orders	{ customer: http:...	{ Connection: close	201	{"id":"614l
9/22/2021, 3:58:18 PM	px-sock-shop/shipping-7...	10.169.137.130	POST	/shipping	{ id: 2bbfa8d4-97...	{ Content-Type: app...	201	{ id: 2bb
9/22/2021, 3:58:18 PM	px-sock-shop/orders-7c...	10.169.137.125	POST	/orders	{ customer: http:...	{ Connection: close	201	{"id":"614l
9/22/2021, 3:58:18 PM	px-sock-shop/carts-85bf...	10.169.137.125	POST	/carts/57a98d98e4b006...	{ itemId: 03fef6a...	{ Connection: close	201	{ id: 614
9/22/2021, 3:58:18 PM	px-sock-shop/carts-85bf...	10.169.137.125	POST	/carts/57a98d98e4b006...	{ itemId: 808a2de...	{ Connection: close	201	{ id: 614
9/22/2021, 3:58:18 PM		35.191.10.207	GET	/healthz		{ Content-Length: 1...	200	{ lastUpd

Showing 68 - 81 / 1000 records


PIXIE
Cluster: gke:oazizi
RUN
Automatic Service Maps
rabbitmq.px-sock-shop.svc.cluster.local
px-sock-shop/rabbitmq-68d55c844f-qf8lh
kube-dns.kube-system.svc.cluster.local
user.px-sock-shop.svc.cluster.local
shipping.px-sock-shop.svc.cluster.local
px-sock-shop/carts-5fc45568c4-bvwbs
px-sock-shop/orders-77c57c89dc-p47gw
carts.px-sock-shop.svc.cluster.local
carts-db.px-sock-shop.svc.cluster.local
px-sock-shop/queue-master-bd5
px-sock-shop/shipping-745b9d8755-g8h2k
10.169.136.241
ENABLE HIERARCHY

script: px/net_flow_graph   namespace*: px-sock-shop   from_entity_filter: to_entity_filter: throughput_filter: 0.0   start_time: -5m

Net Flow Graph

bytes_total: 103.5 KB/s
bytes_sent: 40.1 KB/s
bytes_recv: 63.4 KB/s

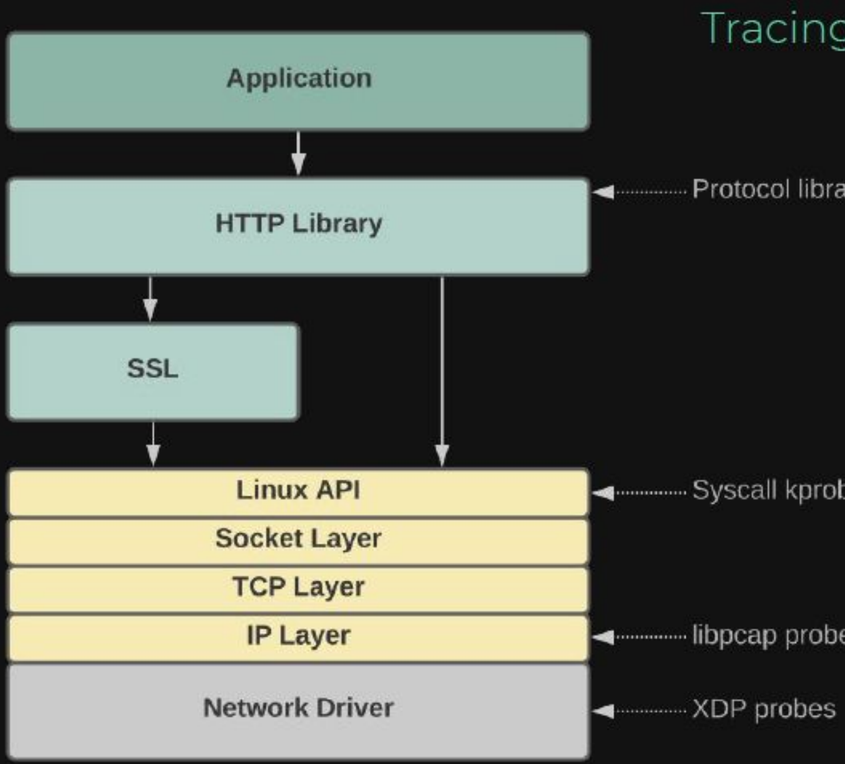
Table

FROM_ENTITY	TO_ENTITY	BYTES_SENT	BYTES_RECV	BYTES_TOTAL
px-sock-shop/carts-5fc45568c4-bvwbs	carts-db.px-sock-shop.svc.cluster.local	40.1 KB/s	63.4 KB/s	103.5 KB/s
px-sock-shop/carts-5fc45568c4-bvwbs	kube-dns.kube-system.svc.cluster.local	38.3 B/s	91.7 B/s	130 B/s
px-sock-shop/orders-77c57c89dc-p47gw	shipping.px-sock-shop.svc.cluster.local	3.8 KB/s	2.3 KB/s	6.1 KB/s

Where to Trace the Data?

Many options in the software stack:

We preferred tracing as close to the application layer as possible.



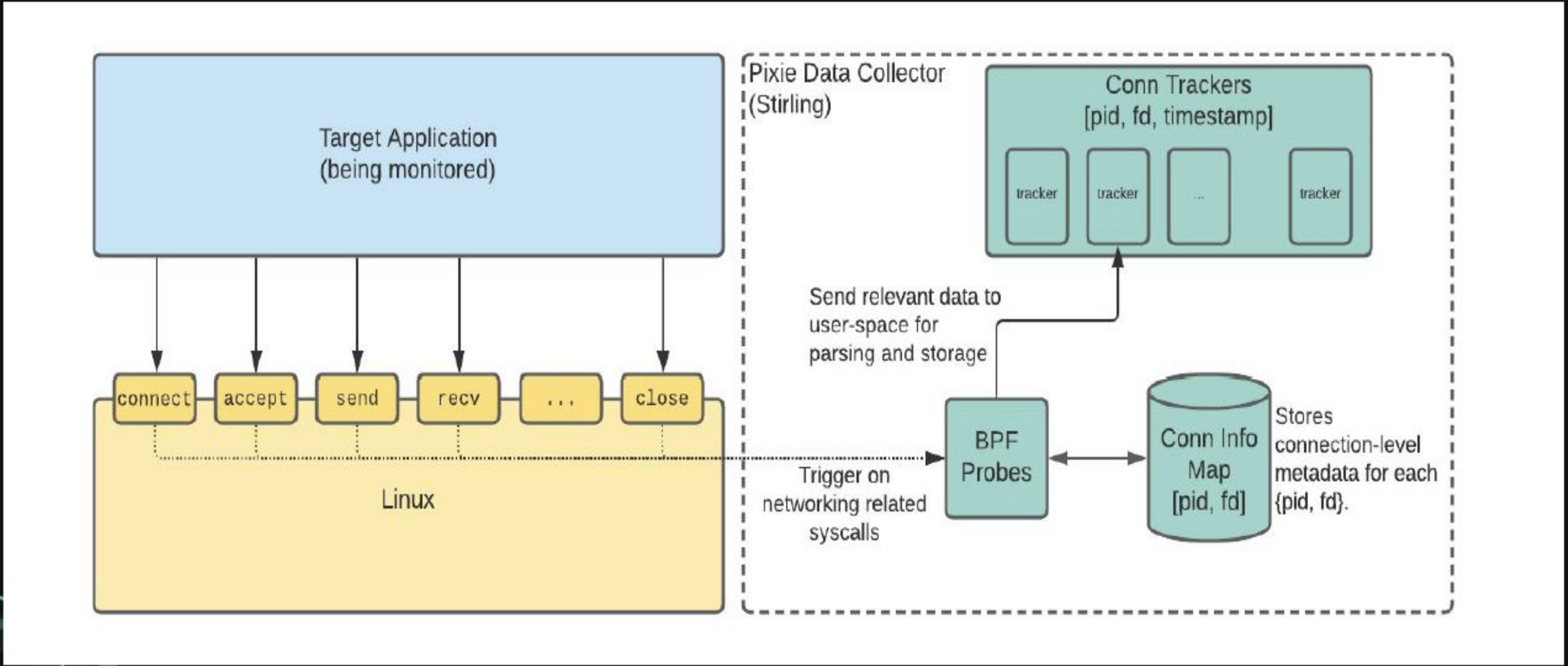
Approaches Compared

	protocol library uprobes	syscall kprobes	libpcap/XDP
Tracing overhead	Low	Low	Low
Scalability & Stability	Uprobes per library, Probe targets may change	High	High
Parsing effort	None	Protocol parsing	Packet processing & protocol parsing
SSL tracing	Cleartext available	Data encrypted	Data encrypted

We chose to use syscall kprobes on functions such as send() & recv().

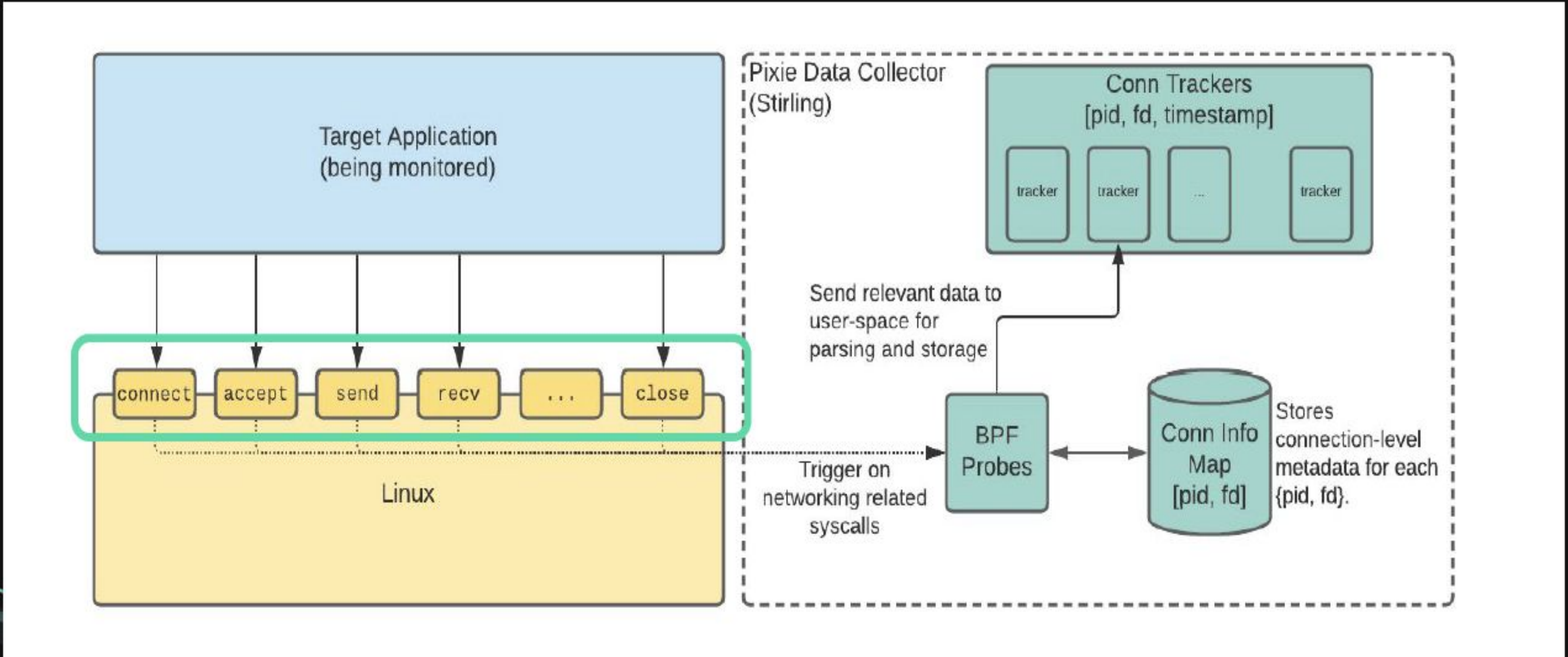
- Rationale: close to the application layer, but stable API.

Architecture



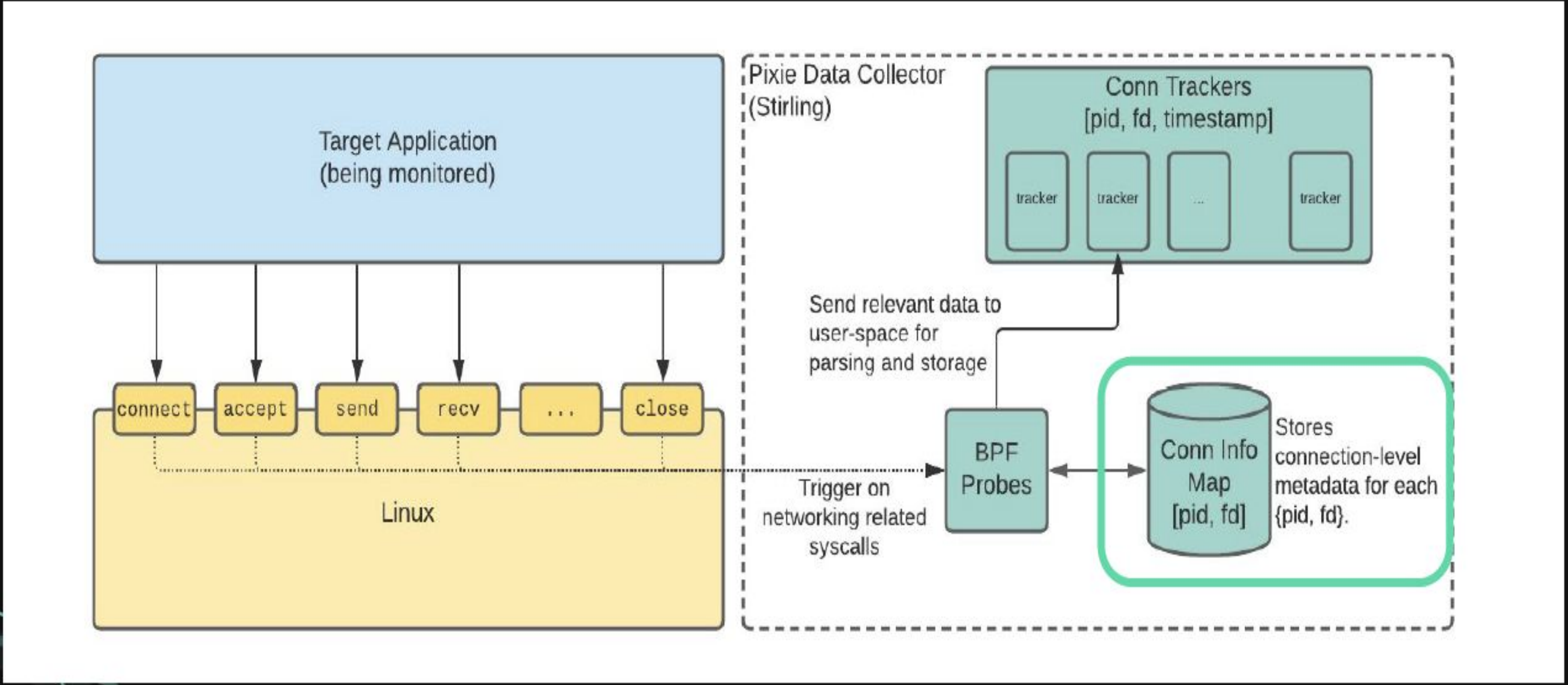
Architecture

1 - Setup probes on network related syscalls.



Architecture

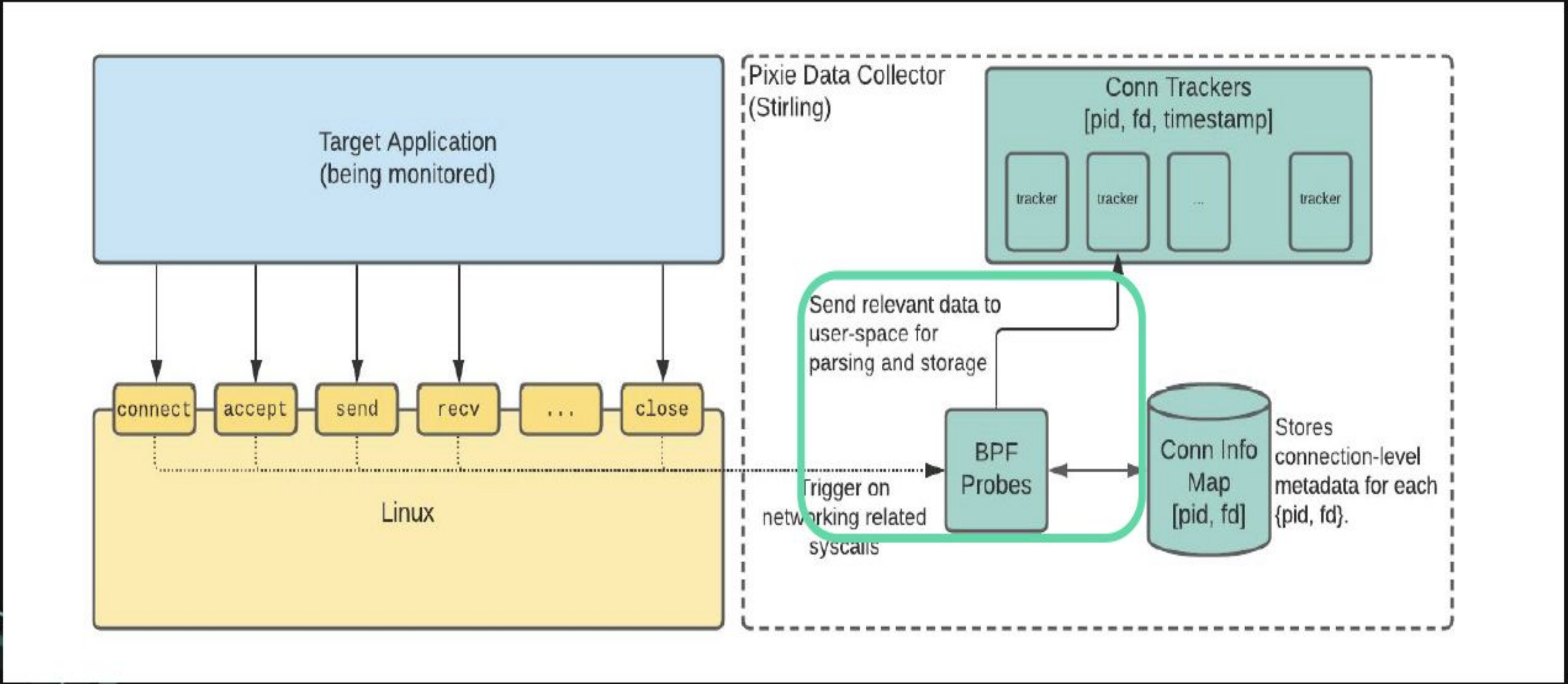
2 - Record connection metadata in BPF maps.



Target Application (being monitored)
connect
accept
send
recv
...
close
Linux
Pixie Data Collector (Stirling)
Conn Trackers [pid, fd, timestamp]
tracker
tracker
...
tracker
Send relevant data to user-space for parsing and storage
Trigger on networking related syscalls
BPF Probes
Conn Info Map [pid, fd]
Stores connection-level metadata for each {pid, fd}.

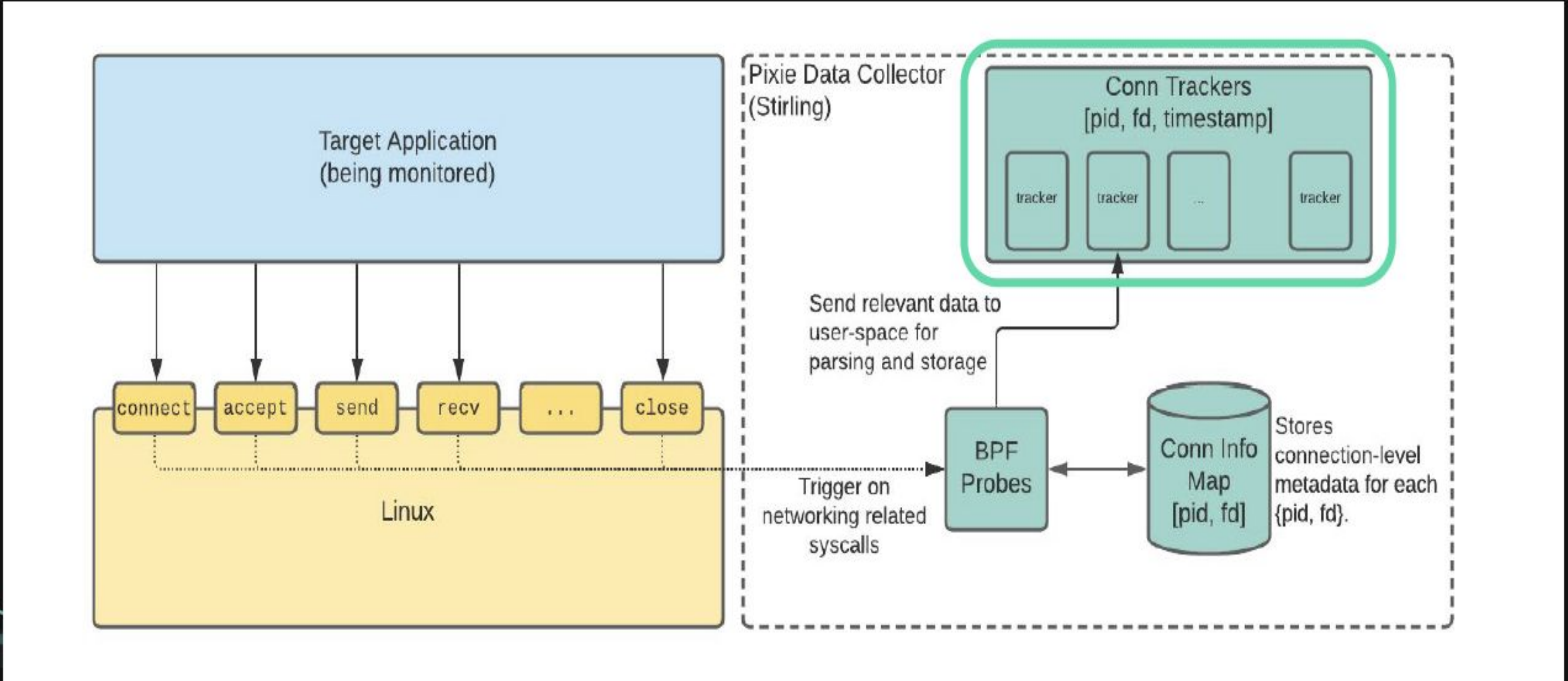
Architecture

3 - Infer protocol with basic rule-based classification as a simple filter. Transfer connection information and data through two perf buffers.



Architecture

**4 - Track connections in user-space with ConnTrackers.
Parse ConnTracker data into structured messages.**



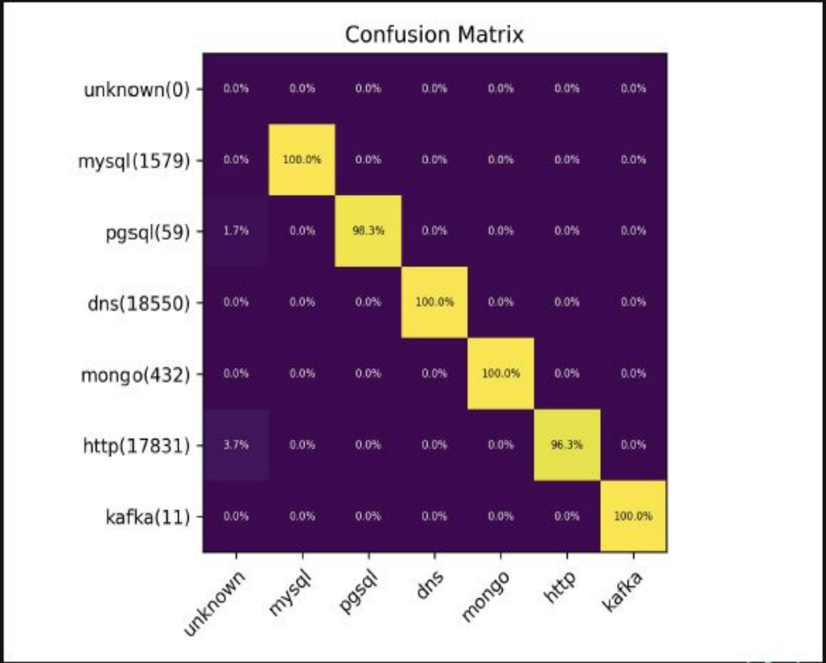
eBPF-Side Protocol Inference

To filter data transfers to user-space, we apply protocol inference in BPF.

- Just a filter: False positives are okay.
- Example for HTTP:

```
static __inline enum MessageType infer_http_message(const char*
buf, size_t count) {
 ...

 if (buf[0] == 'H' && buf[1] == 'T' && buf[2] == 'T' && buf[3] == 'P') {
        return kResponse;
 }
 ...
```

Likelihood that our inference eventually identifies the right protocol

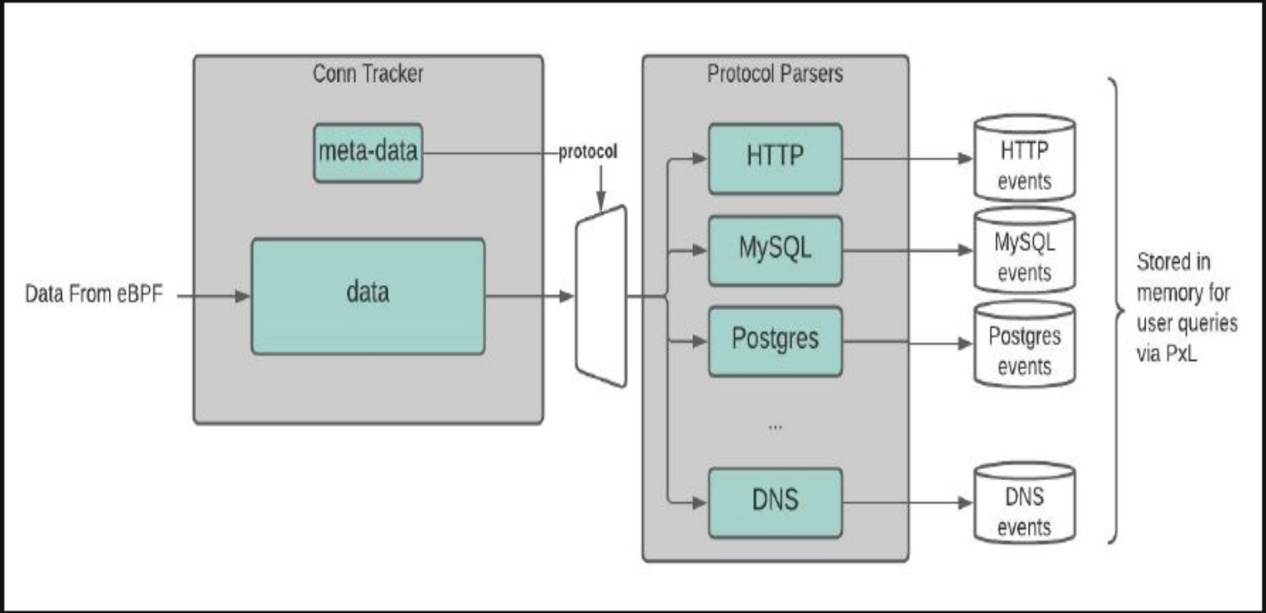


Confusion Matrix

	unknown	mysql	pgsql	dns	mongo	http	kafka
unknown(0)	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
mysql(1579)	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
pgsql(59)	1.7%	0.0%	98.3%	0.0%	0.0%	0.0%	0.0%
dns(18550)	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
mongo(432)	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
http(17831)	3.7%	0.0%	0.0%	0.0%	0.0%	96.3%	0.0%
kafka(11)	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%

Pluggable Protocol Parsers

Architecture consists of pluggable protocol parsers



Supported Protocols List
HTTP MySQL Postgres Redis Cassandra Kafka NATS DNS gRPC*

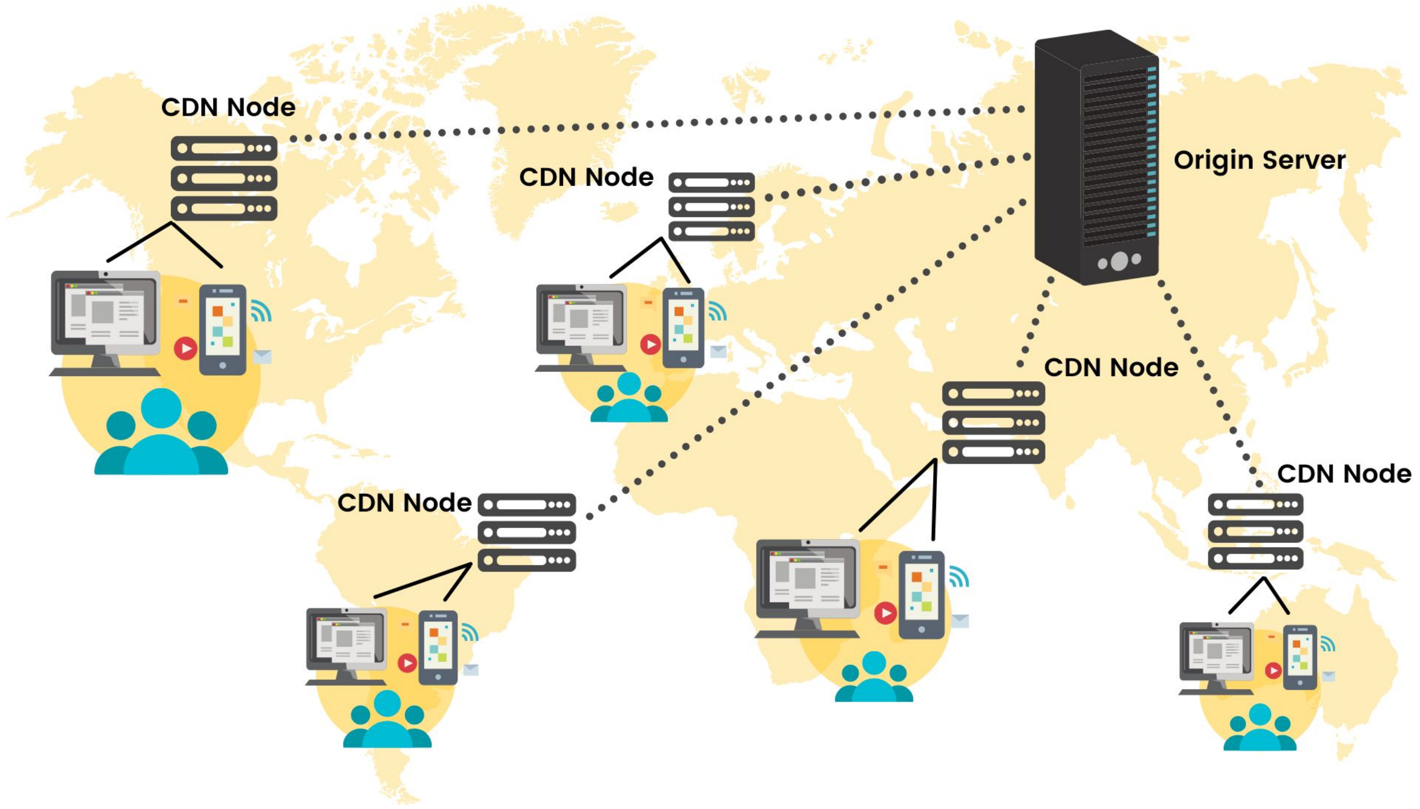
*gRPC is traced with dedicated uprobes

Bonus Material

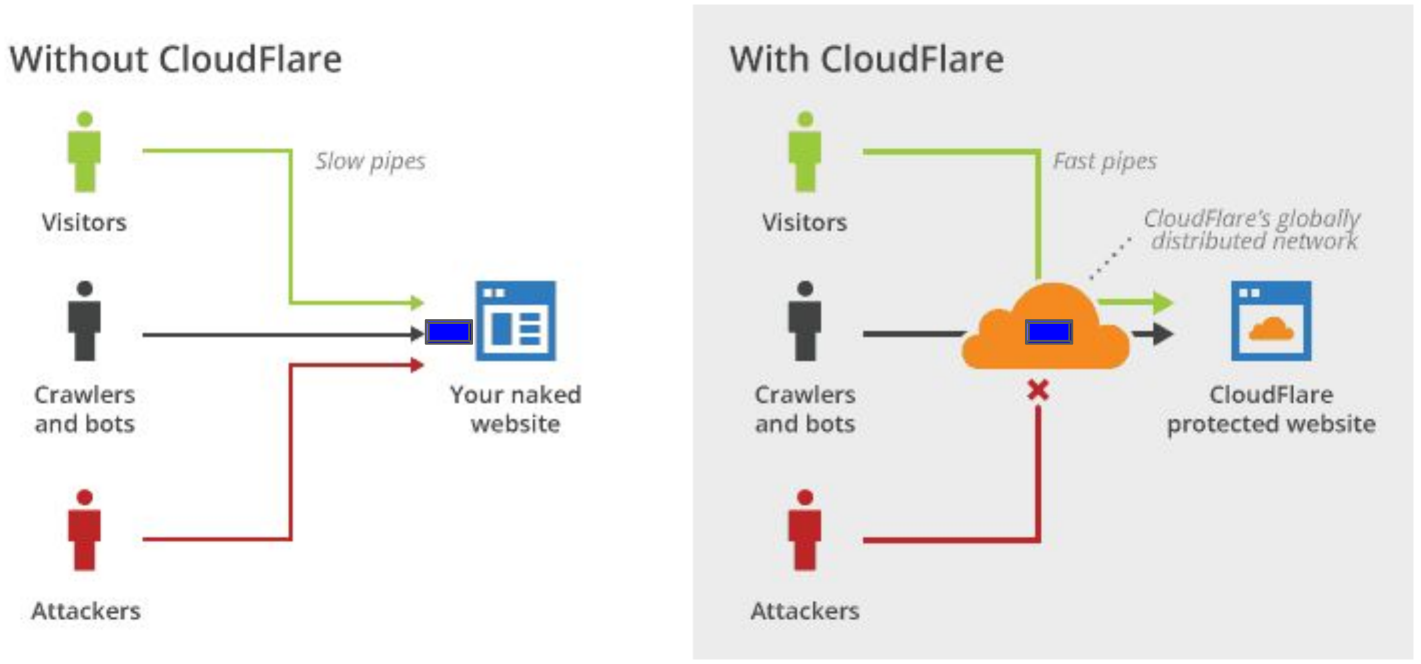
Use Cases

1. K8s Observability
2. **Network Acceleration**

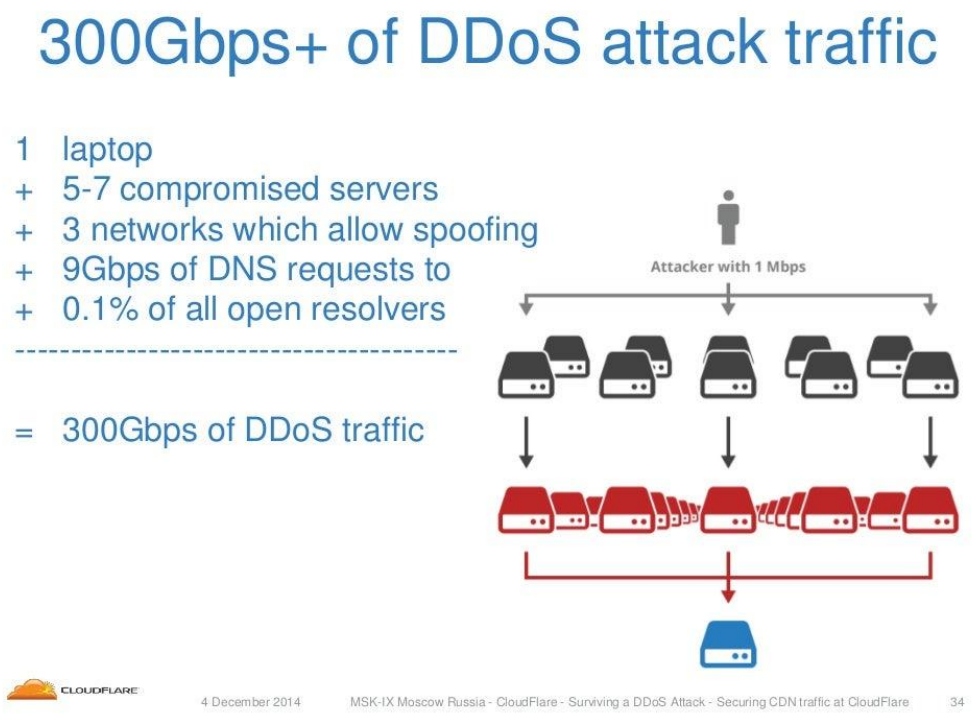
Content Distribution Networks



Cloudflare CDN



Cloudflare CDN

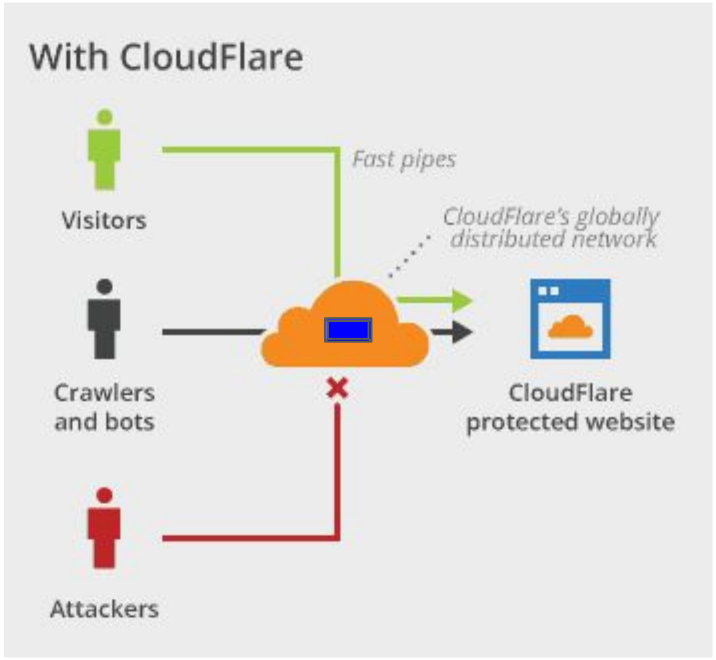


300Gbps+ of DDoS attack traffic

1 laptop
+ 5-7 compromised servers
+ 3 networks which allow spoofing
+ 9Gbps of DNS requests to
+ 0.1% of all open resolvers

= 300Gbps of DDoS traffic

Attacker with 1 Mbps


With CloudFlare
Visitors
Fast pipes
CloudFlare's globally distributed network
Crawlers and bots
CloudFlare protected website
Attackers

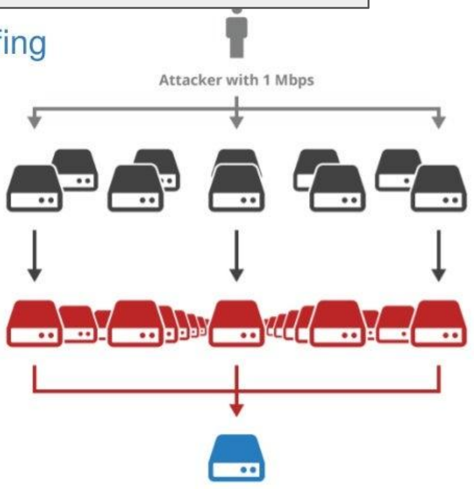
Cloudflare CDN

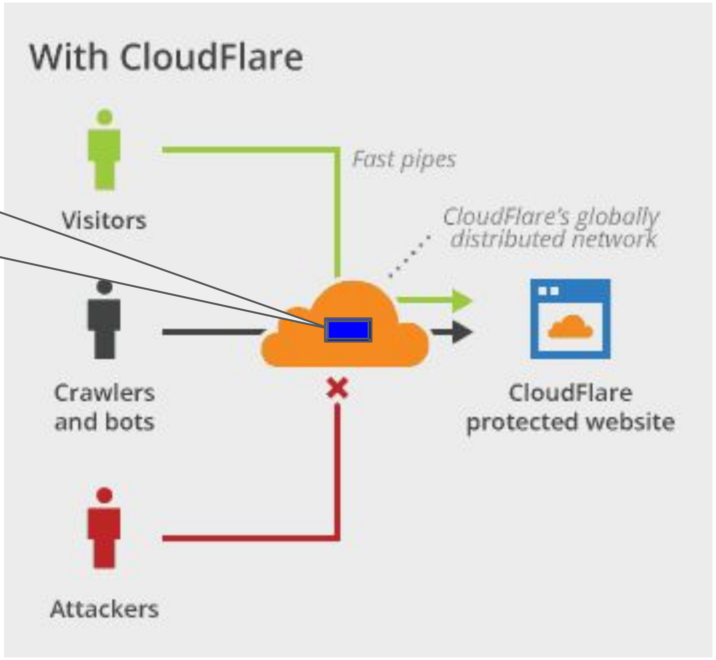
1. Edge of infra.
2. HW/SW ?
3. Sophisticated Detection
4. Drop pkts.

300Gbps+ ffic

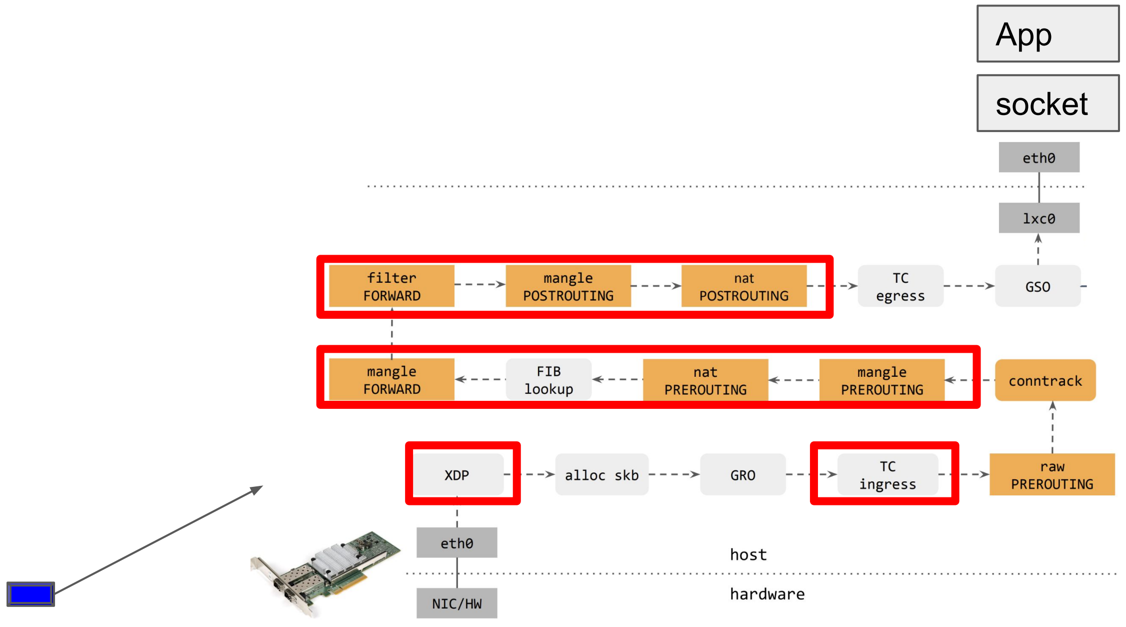
1 laptop
+ 5-7 compromised servers
+ 3 networks which allow spoofing
+ 9Gbps of DNS requests to
+ 0.1% of all open resolvers

= 300Gbps of DDoS traffic

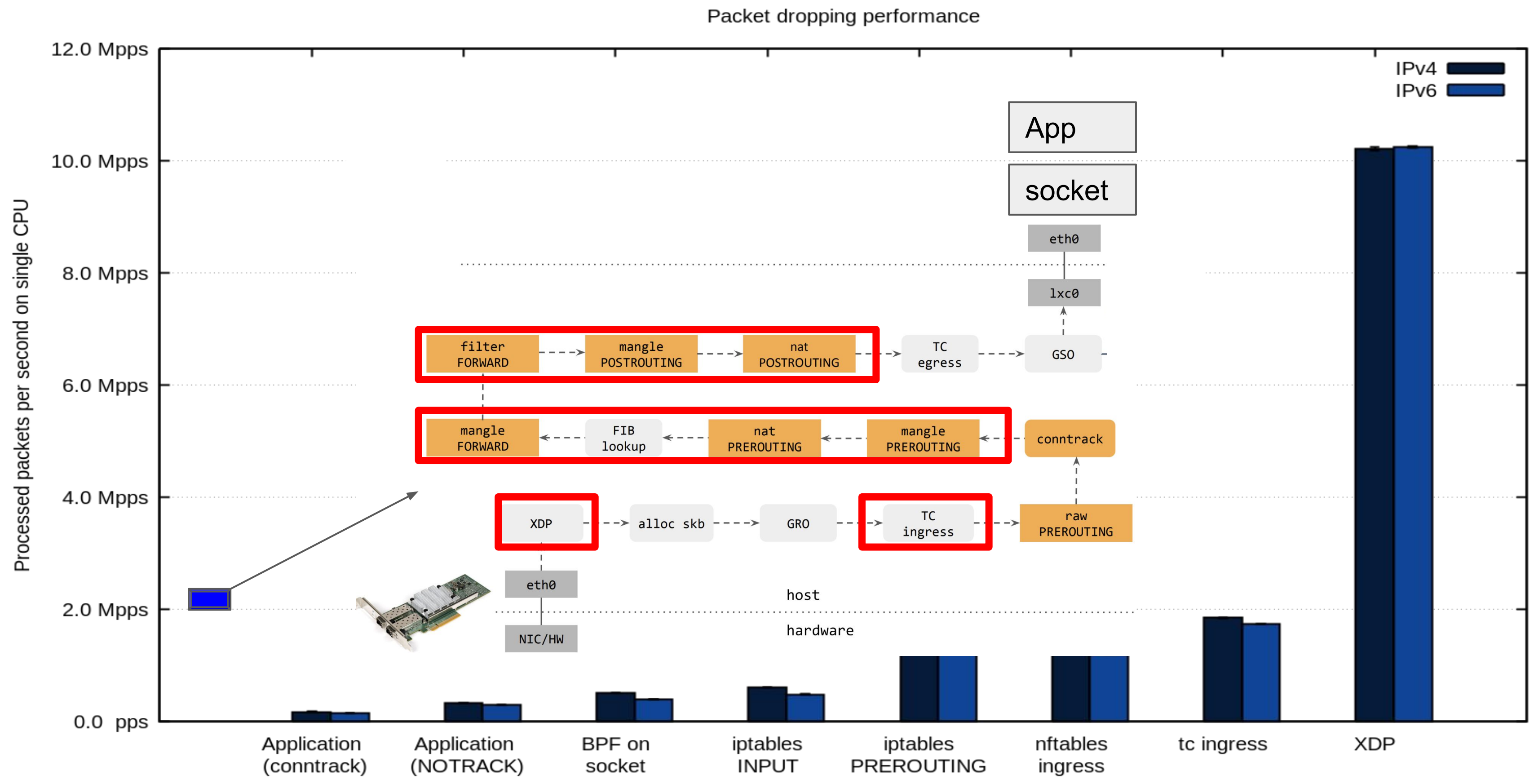



With CloudFlare
Visitors
Fast pipes
CloudFlare's globally distributed network
Crawlers and bots
CloudFlare protected website
Attackers

Packet Dropping for DDoS Mitigation

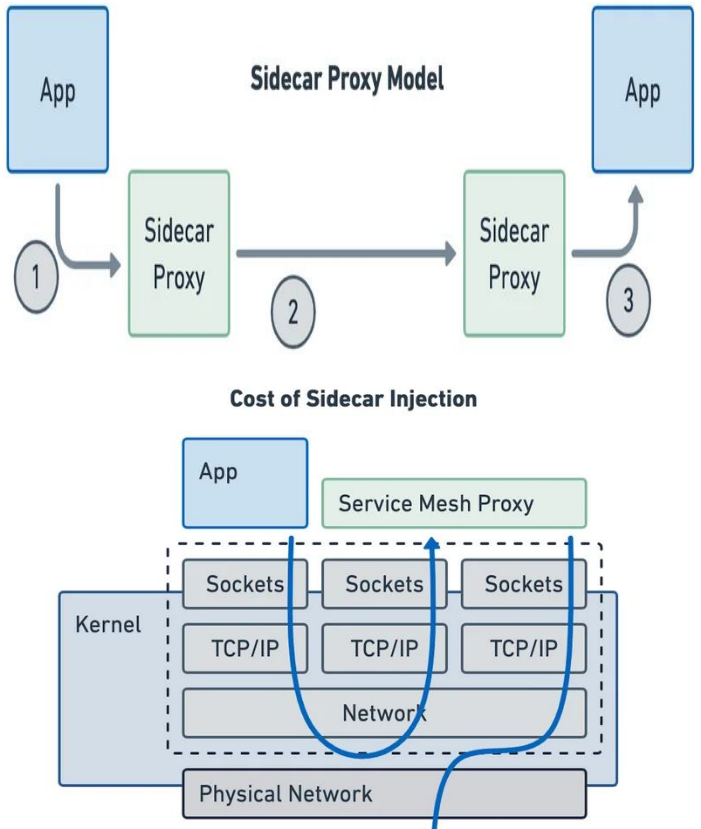


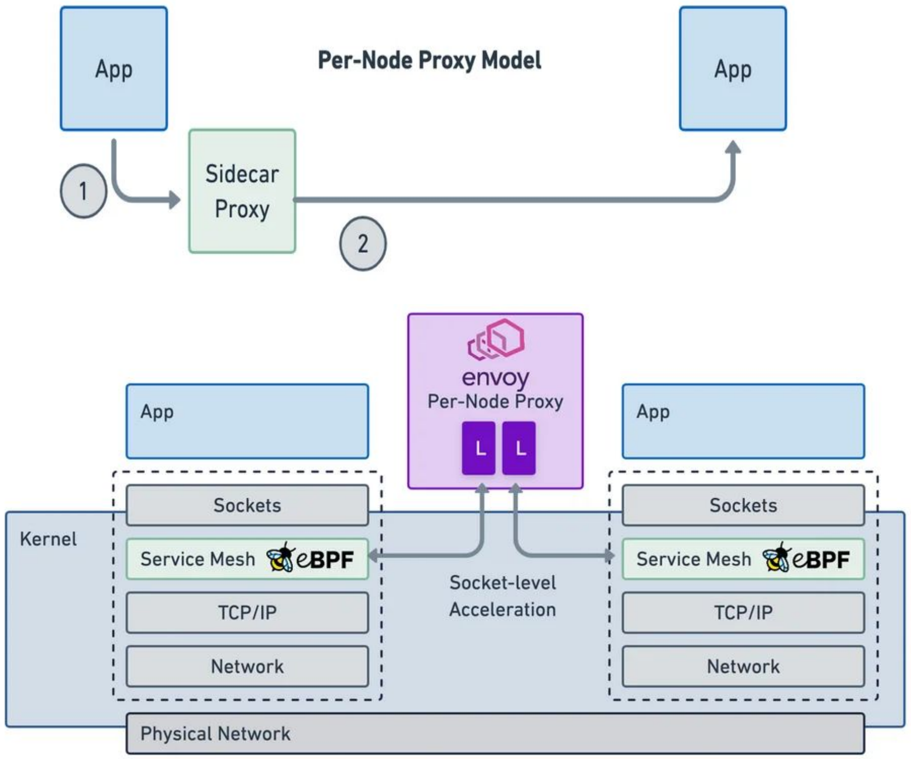
From Here: How to drop 10 million packets per second (cloudflare.com)



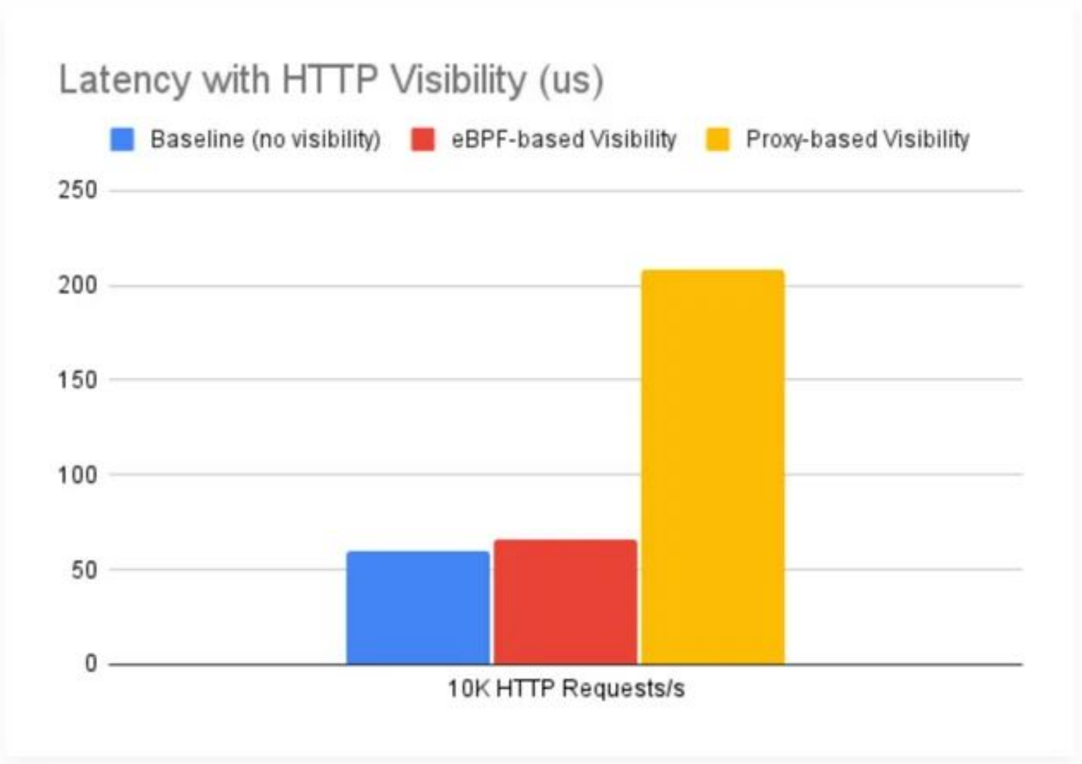
From Here: [How to drop 10 million packets per second (cloudflare.com)]

De-duplicate Network Stack Traversal with eBPF SockMap





Low Cost Service Mesh with eBPF



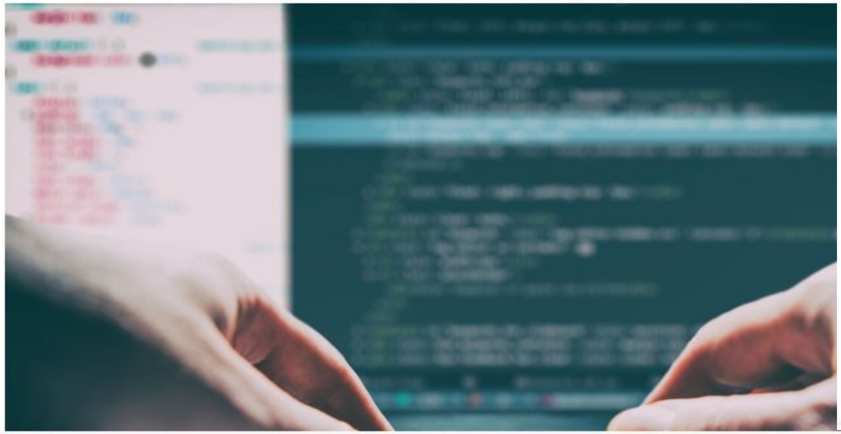
Conclusion

a) Where else is eBPF used : Storage, Security, DBMS,Scheduling etc.

b) What next -- eBPF startups mal

Groundcover lands $20M to he companies monitor app perfor

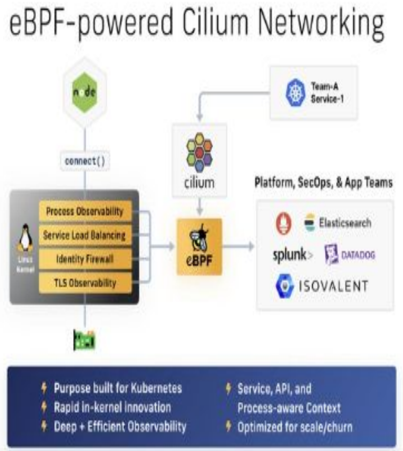
Kyle Wiggers @kyle_l_wiggers / 6:30 PM GMT+5:30 • September 14, 2022



Isovalent Raises $40M in Series B Funding

USA

Published on September 7, 2022


eBPF-powered Cilium Networking
node
Team-A Service-1
connect()
cilium
Platform, SecOps, & App Teams
Process Observability
Service Load Balancing
Identity Firewall
TLS Observability
Linux Kernel
eBPF
Elasticsearch
splunk>
DATADOG
ISOVALENT
Purpose built for Kubernetes
Rapid in-kernel innovation
Deep + Efficient Observability
Service, API, and Process-aware Context
Optimized for scale/churn

Isovalent, a Mountain View, CA-based company behind open source technologies Cilium and eBPF, closed a $40M Series B funding round.

The round was led by Thomvest Ventures with participation from M12 (Microsoft's Venture Fund) and Grafana Labs, which joined Google and Cisco as existing strategic investors in the company, as well as Andreessen Horowitz, Mango Capital, and Mirae Asset Capital.

The company intends to use the funds to expand its team both to

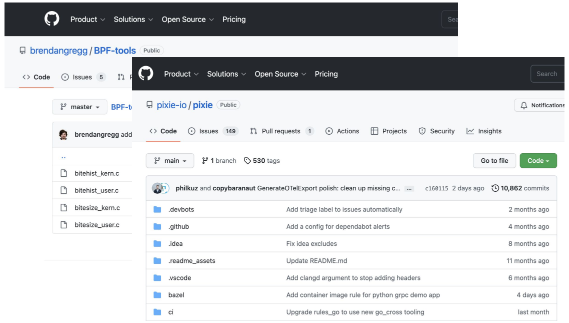
OPENED Tool for Managing eBPF Heterogeneity

Microservices Observatory (microserviceobservatory.github.io)

Theophilus A. Benson Brown University tab@cs.brown.edu
Palanivel Kodeswaran IBM Research palani.kodeswaran@in.ibm.com
Sayandeep Sen IBM Research sayandes@in.ibm.com

eBPF Programs are Monoliths

One Off Programs



brendangregg / BPF-tools Public

Code · Issues 5

master · BPF-t

brendangregg add

..
bitehist_kern.c
bitehist_user.c
bitesize_kern.c
bitesize_user.c

pixie-io / pixie Public

Code · Issues 149 · Pull requests 1 · Actions · Projects · Security · Insights

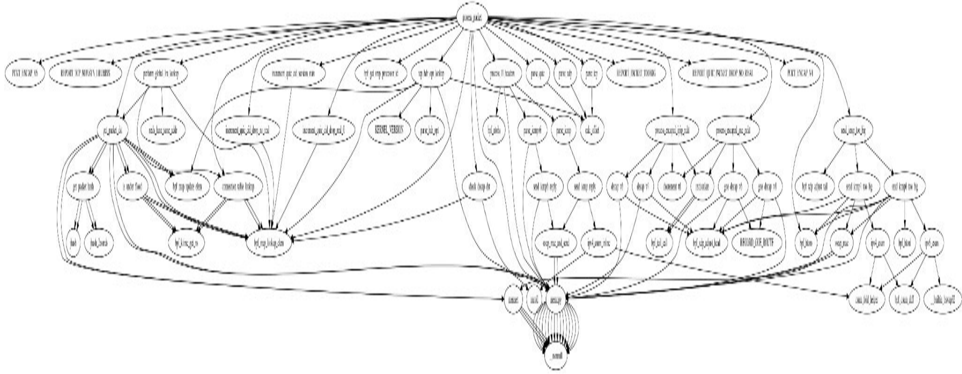
main · 1 branch · 530 tags · Go to file · Code

philkuz and copybaranaut GenerateOTelExport polish: clean up missing c... … c160115 2 days ago 10,862 commits

.devbots	Add triage label to issues automatically	2 months ago
.github	Add a config for dependabot alerts	4 months ago
.idea	Fix idea excludes	8 months ago
.readme_assets	Update README.md	11 months ago
.vscode	Add clangd argument to stop adding headers	6 months ago
bazel	Add container image rule for python grpc demo app	4 days ago
ci	Upgrade rules_go to use new go_cross tooling	last month

Observability

Complex



Network Functions

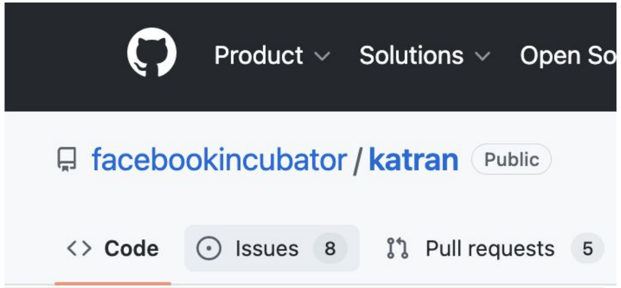
*Code from a Katran function

Implications of Monolith on Developer Productivity

Developing a new program

Find sub functionality on GitHub

Extracting and reusing functionality is non-trival

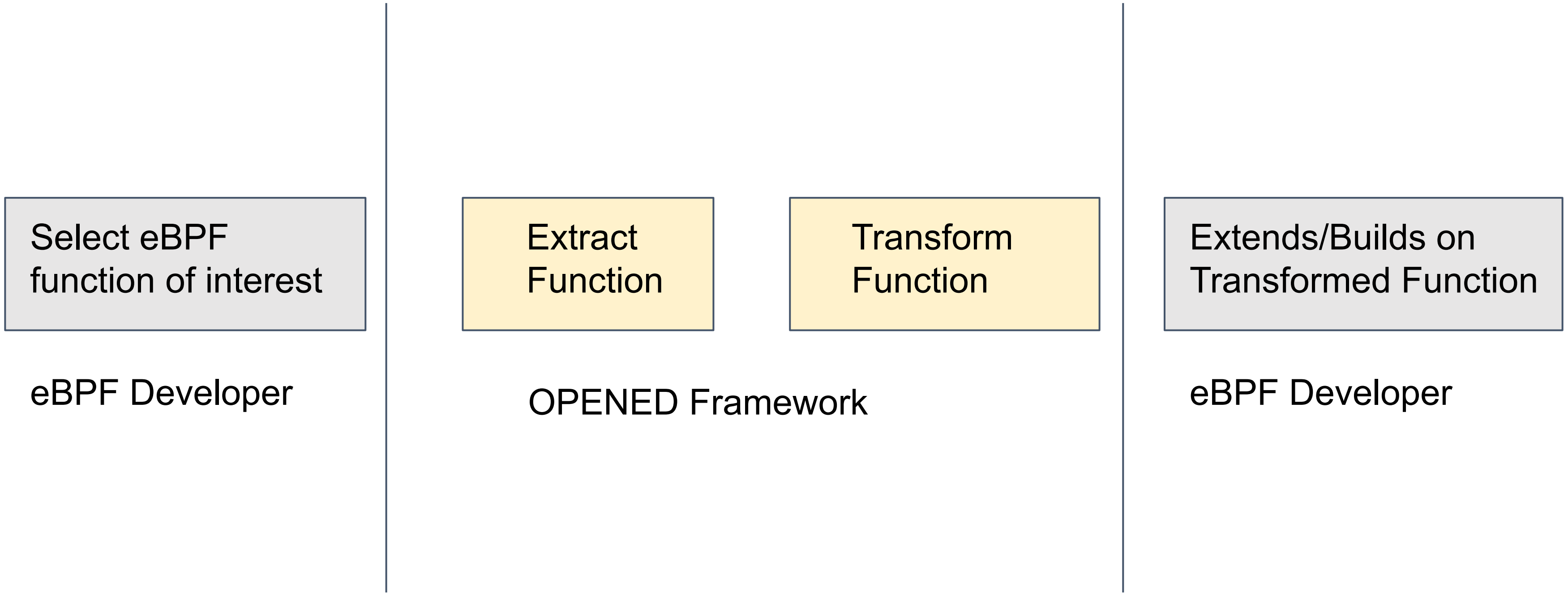


Step 1: Extract lines

Step 2: Identify + Extract Deps

Surprise Step 3: Rewrite for your target hookpoint

The OPENED Vision



OPENED Vision: Reduce time to new functionality development

- Automated extraction of relevant code

- Automated transformation of code
  - Enable moving code between hook-points
  - Enable moving code between programs

- Developer-first automation
  - Extraction + Transformation guided by developer choices