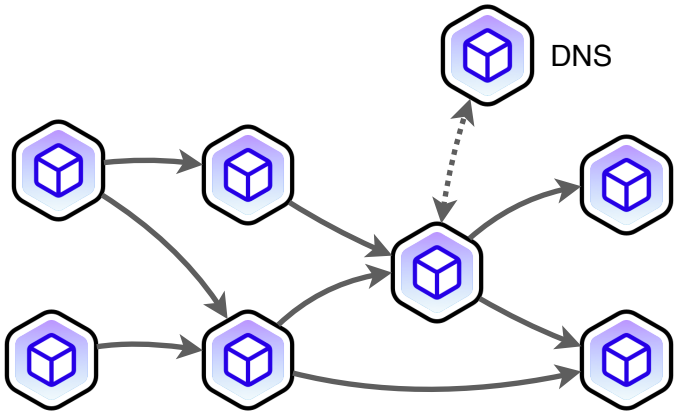
Is Your Kubernetes Cluster's DNS Working?

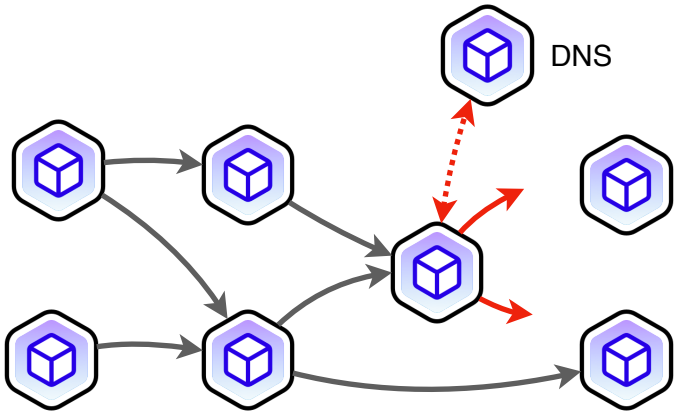
Jonathan Perry, Flowmill


DNS IS FINE.

Why care?

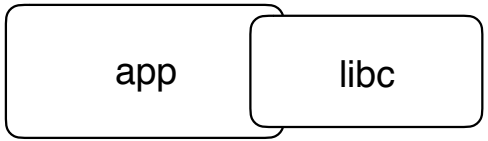


Why care?

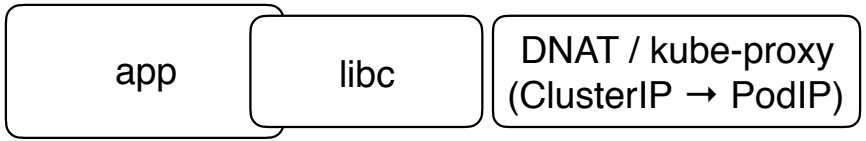


How does it work?

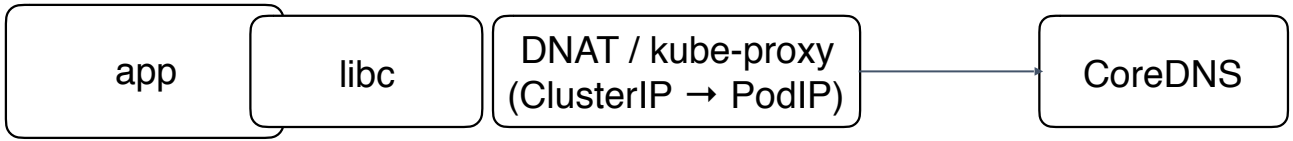
How does it work?



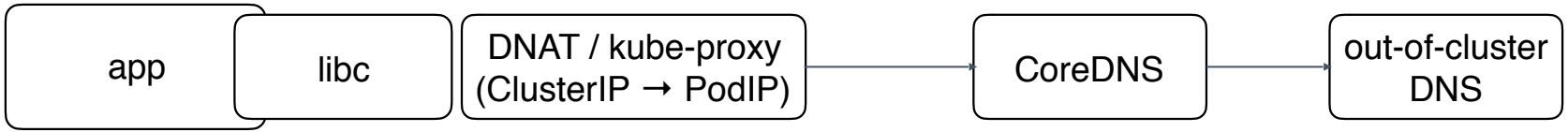
How does it work?



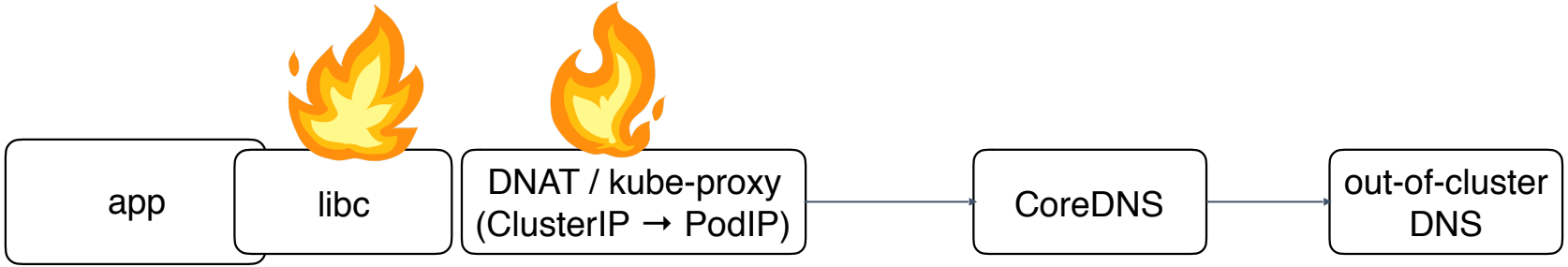
How does it work?

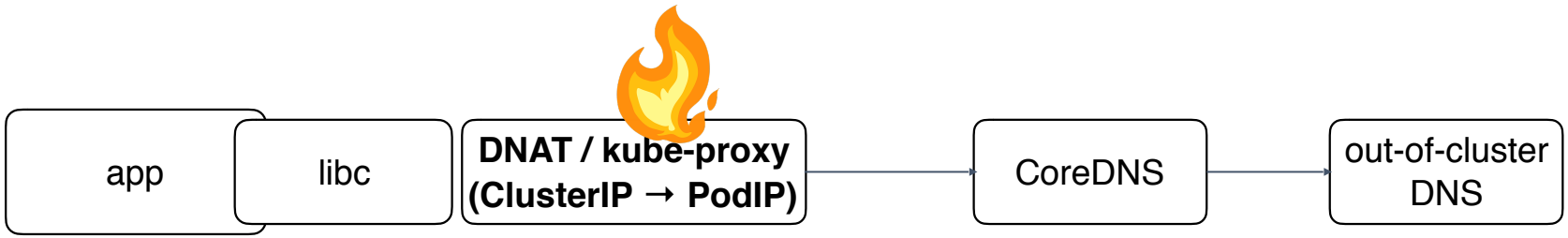


How does it work?

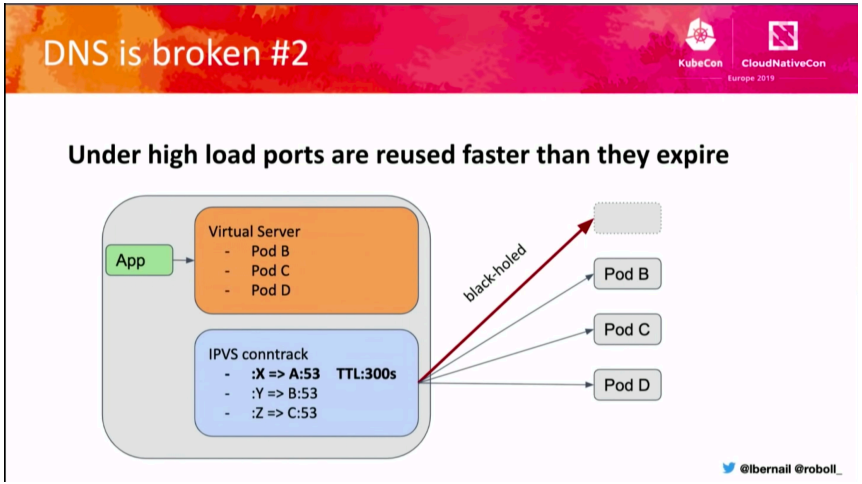


How does it work?

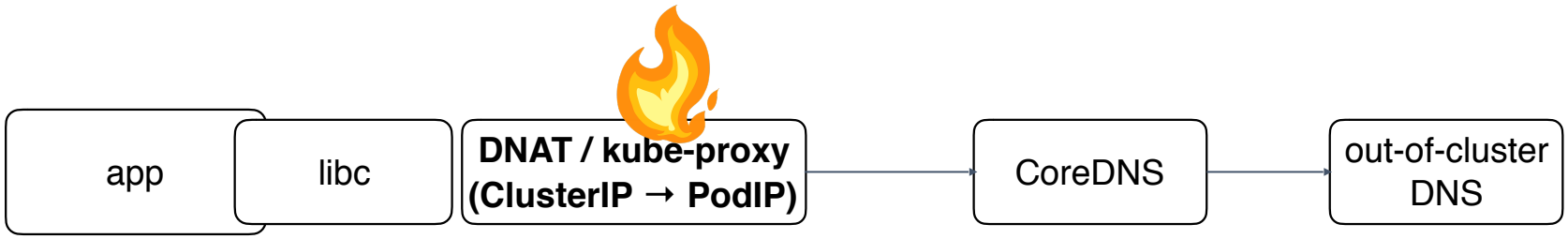




“High load” black hole (Datadog):



- Even if kube-proxy removed a dead pod, conntrack can still have mappings to it
- Under load, source port number is reused -> routed to dead pod



Kernel race: Parallel DNS packets

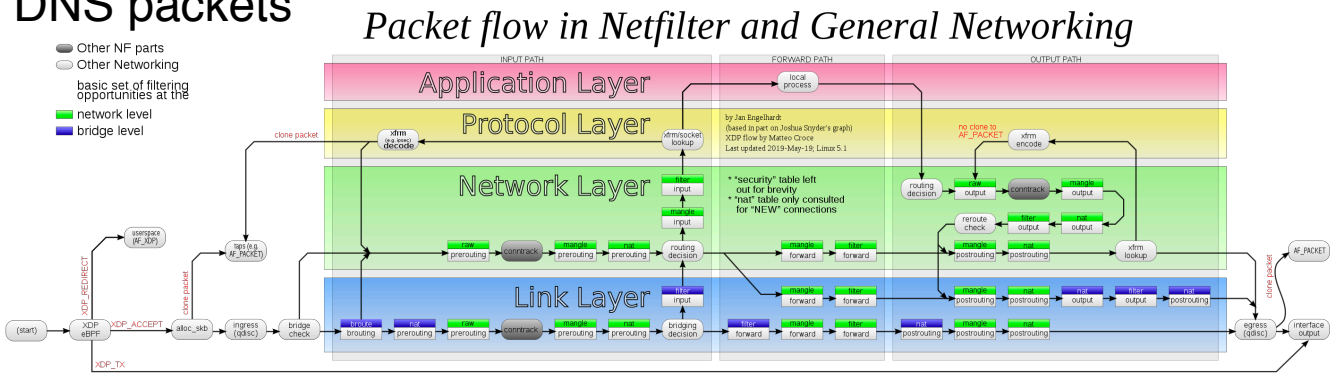
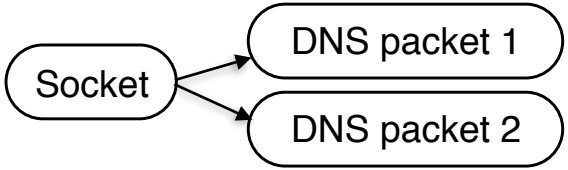
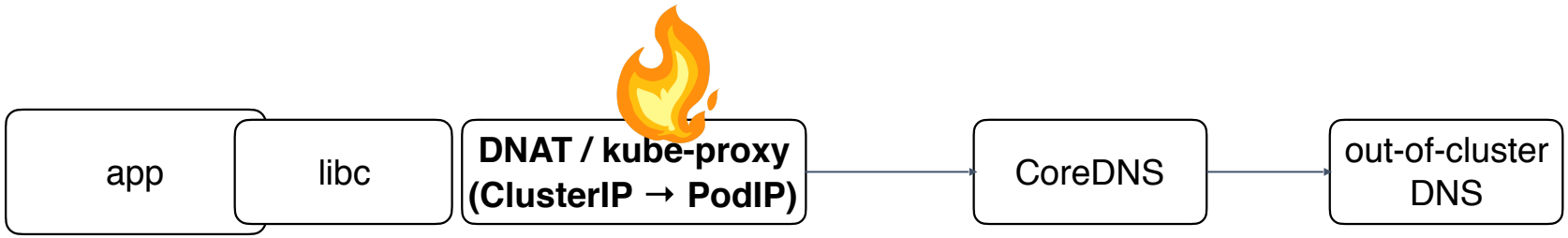


Diagram by Jan Engelhardt

[1] https://github.com/kubernetes/kubernetes/issues/56903
[2] https://tech.xing.com/a-reason-for-unexplained-connection-timeouts-on-kubernetes-docker-abd041cf7e02
[3] https://blog.quentin-machu.fr/2018/06/24/5-15s-dns-lookups-on-kubernetes/
[4] https://github.com/weaveworks/weave/issues/3287#issuecomment-387178077
Conntrack-hooks NAT-hooks priority-consts



FLOWMILL



Kernel race: Parallel DNS packets

Packet flow in Netfilter and General Networking

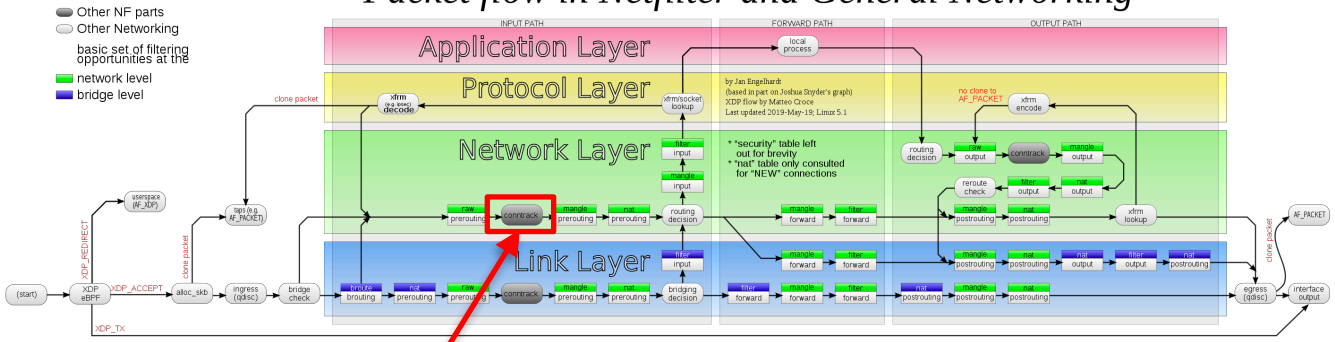
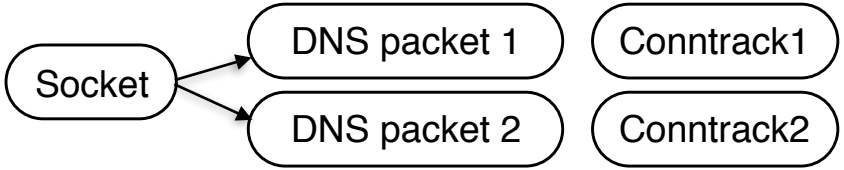


Diagram by Jan Engelhardt

[1] https://github.com/kubernetes/kubernetes/issues/56903
[2] https://tech.xing.com/a-reason-for-unexplained-connection-timeouts-on-kubernetes-docker-abd041cf7e02
[3] https://blog.quentin-machu.fr/2018/06/24/5-15s-dns-lookups-on-kubernetes/
[4] https://github.com/weaveworks/weave/issues/3287#issuecomment-387178077
Conntrack-hooks NAT-hooks priority-consts

Allocate conntrack entry



FLOWMILL

app | libc | **DNAT / kube-proxy (ClusterIP → PodIP)** → CoreDNS → out-of-cluster DNS

Kernel race: Parallel DNS packets

Packet flow in Netfilter and General Networking

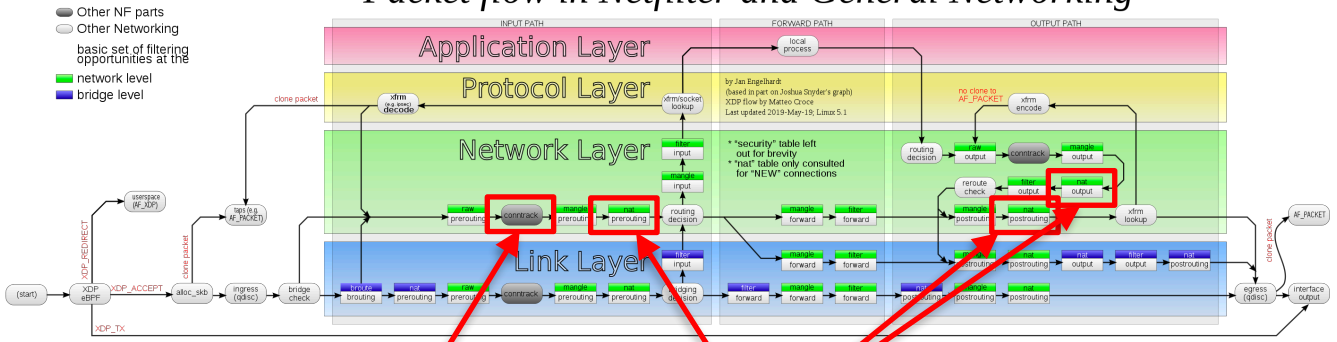


Diagram by Jan Engelhardt

[1] https://github.com/kubernetes/kubernetes/issues/56903
[2] https://tech.xing.com/a-reason-for-unexplained-connection-timeouts-on-kubernetes-docker-abd041cf7e02
[3] https://blog.quentin-machu.fr/2018/06/24/5-15s-dns-lookups-on-kubernetes/
[4] https://github.com/weaveworks/weave/issues/3287#issuecomment-387178077
Conntrack-hooks NAT-hooks priority-consts

Allocate conntrack entry

NAT

Socket → DNS packet 1 → Conntrack1 → DST: CoreDNS A

Socket → DNS packet 2 → Conntrack2 → DST: CoreDNS B

FLOWMILL

app | libc | **DNAT / kube-proxy (ClusterIP → PodIP)** → CoreDNS → out-of-cluster DNS

Kernel race: Parallel DNS packets

Packet flow in Netfilter and General Networking

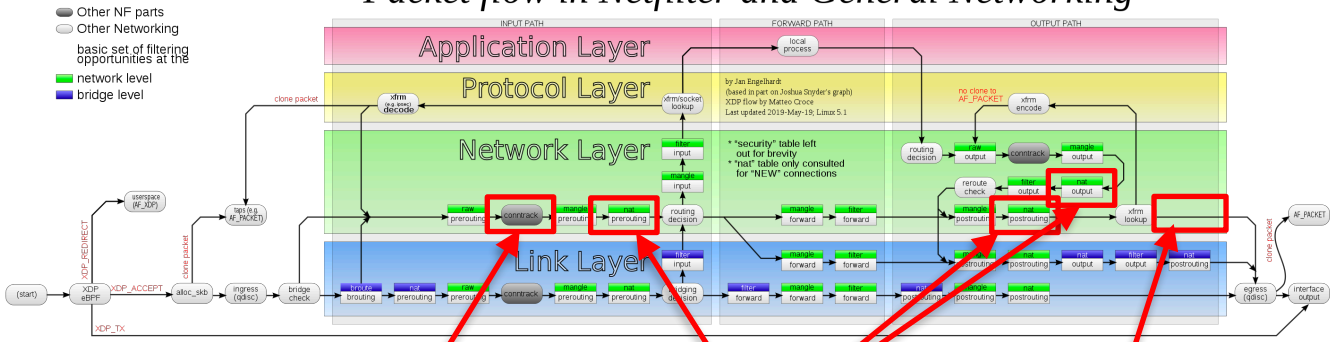


Diagram by Jan Engelhardt

[1] https://github.com/kubernetes/kubernetes/issues/56903
[2] https://tech.xing.com/a-reason-for-unexplained-connection-timeouts-on-kubernetes-docker-abd041cf7e02
[3] https://blog.quentin-machu.fr/2018/06/24/5-15s-dns-lookups-on-kubernetes/
[4] https://github.com/weaveworks/weave/issues/3287#issuecomment-387178077
Conntrack-hooks NAT-hooks priority-consts

Allocate conntrack entry

NAT

"confirm" conntrack

Socket → DNS packet 1 → Conntrack1 → DST: CoreDNS A

Socket → DNS packet 2 → Conntrack2 → DST: CoreDNS B

FLOWMILL

app

libc

DNAT / kube-proxy (ClusterIP → PodIP)

CoreDNS

out-of-cluster DNS

Kernel race: Parallel DNS packets

Packet flow in Netfilter and General Networking

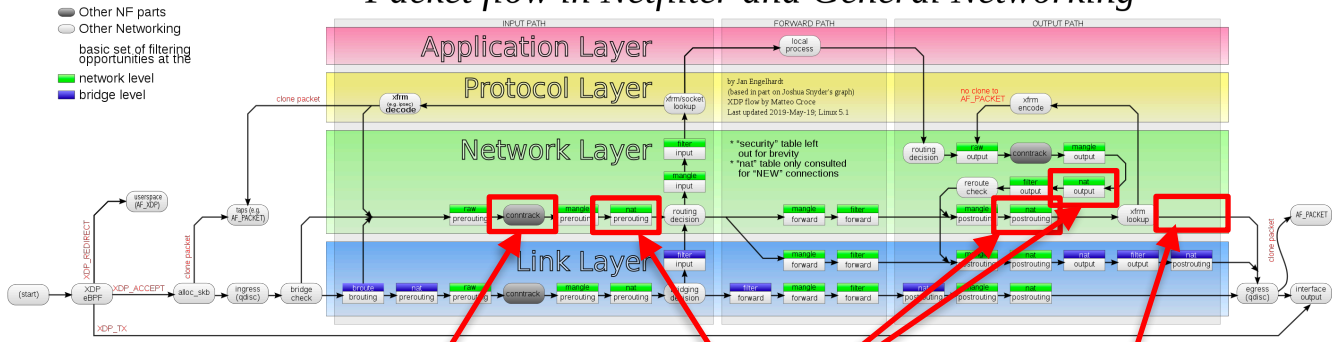


Diagram by Jan Engelhardt

[1] https://github.com/kubernetes/kubernetes/issues/56903
[2] https://tech.xing.com/a-reason-for-unexplained-connection-timeouts-on-kubernetes-docker-abd041cf7e02
[3] https://blog.quentin-machu.fr/2018/06/24/5-15s-dns-lookups-on-kubernetes/
[4] https://github.com/weaveworks/weave/issues/3287#issuecomment-387178077
Conntrack-hooks NAT-hooks priority-consts

Allocate conntrack entry

NAT

"confirm" conntrack

Socket

DNS packet 1 — Conntrack1 — DST: CoreDNS A

DNS packet 2 — Conntrack2 — DST: CoreDNS B

ACCEPT

FLOWMILL

app | libc | **DNAT / kube-proxy (ClusterIP → PodIP)** → CoreDNS → out-of-cluster DNS

Kernel race: Parallel DNS packets

Packet flow in Netfilter and General Networking

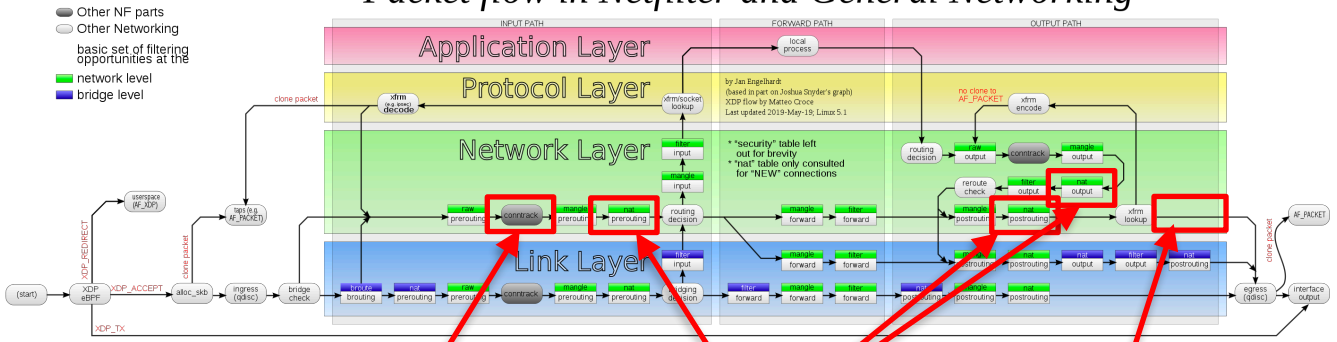


Diagram by Jan Engelhardt

[1] https://github.com/kubernetes/kubernetes/issues/56903
[2] https://tech.xing.com/a-reason-for-unexplained-connection-timeouts-on-kubernetes-docker-abd041cf7e02
[3] https://blog.quentin-machu.fr/2018/06/24/5-15s-dns-lookups-on-kubernetes/
[4] https://github.com/weaveworks/weave/issues/3287#issuecomment-387178077
Conntrack-hooks NAT-hooks priority-consts

Allocate conntrack entry

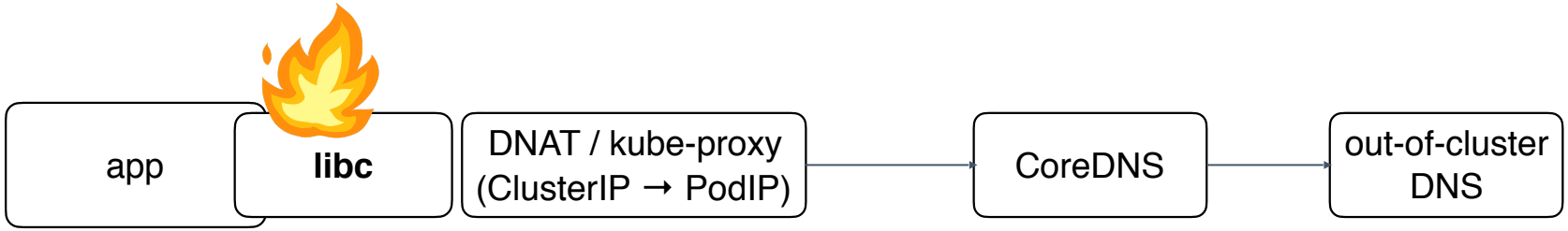
NAT

"confirm" conntrack

Socket → DNS packet 1 → Conntrack1 → DST: CoreDNS A → ACCEPT

Socket → DNS packet 2 → Conntrack2 → DST: CoreDNS B → DROP

FLOWMILL



Alpine/musl:

Queries are performed in parallel:

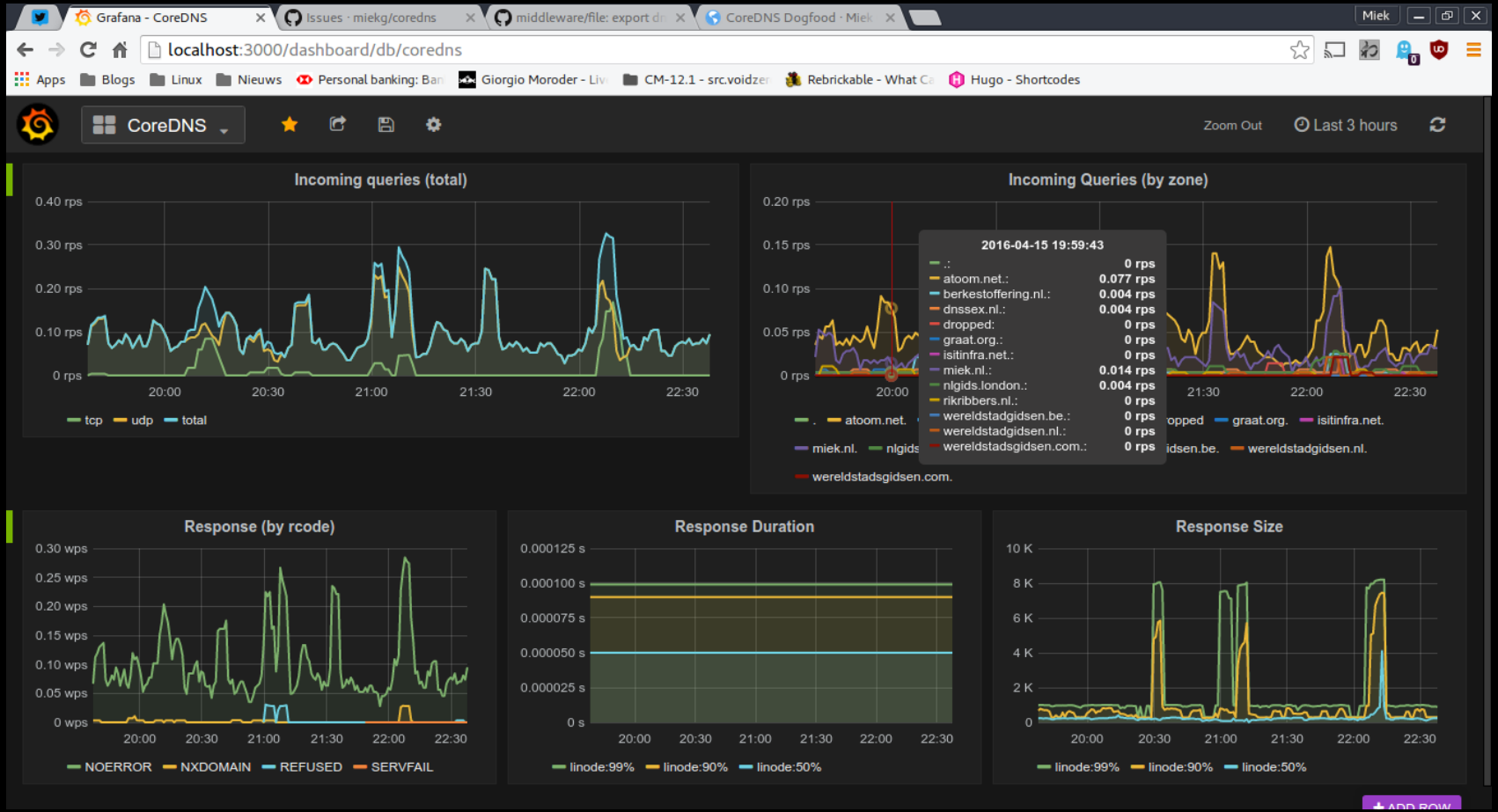
- Triggers the kernel race (no flag for sequential)

NODATA (success + no records):

- Stricter standard adherence than glibc
- Apparently broke k8s DNS for rancher-dns (since fixed), Cloudflare, github issue

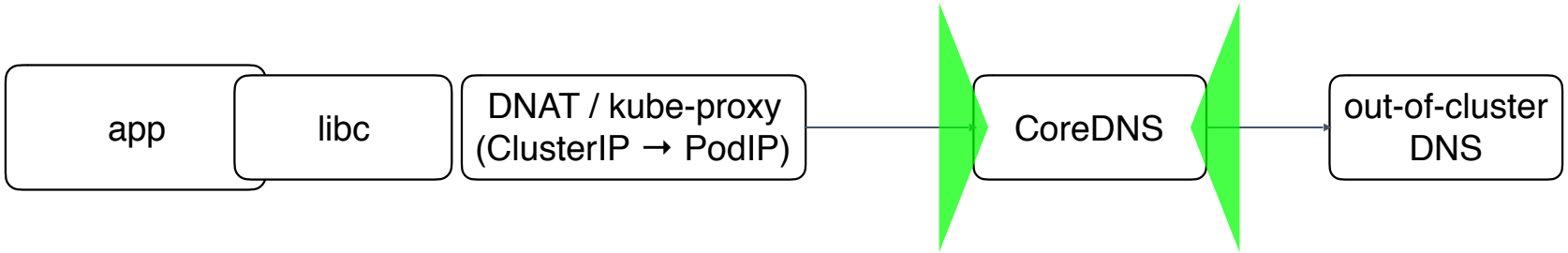
[1] Alpine caveats https://github.com/gliderlabs/docker-alpine/blob/master/docs/caveats.md

[2] Discussion https://github.com/gliderlabs/docker-alpine/issues/8



Credit: Miek Gieben (CoreDNS author) https://miek.nl/2016/april/15/coredns-dogfood-part-2/

CoreDNS metrics

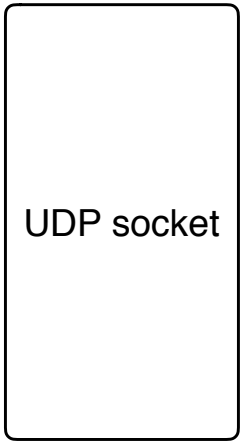


Great!

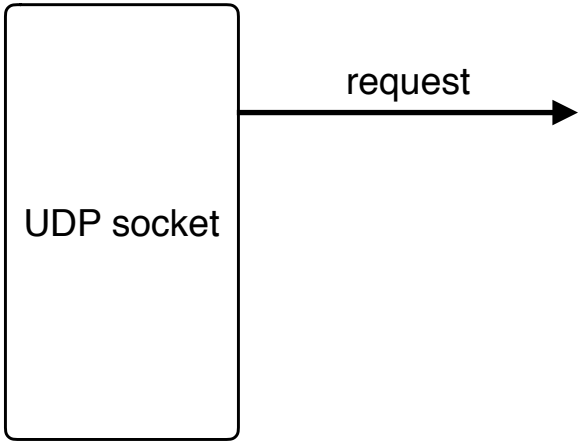
Caveats:

- needs to be available when DNS is down
- missing visibility around app

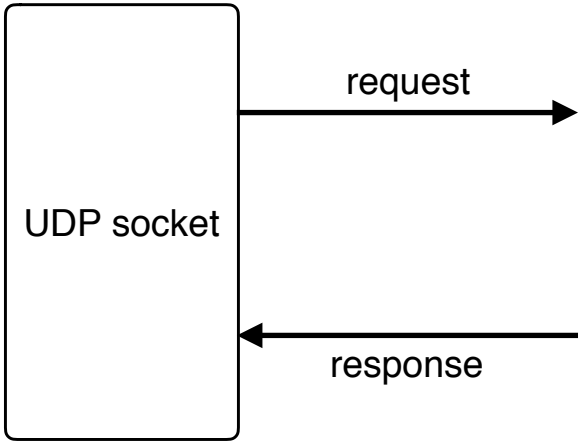
Can get visibility from the OS



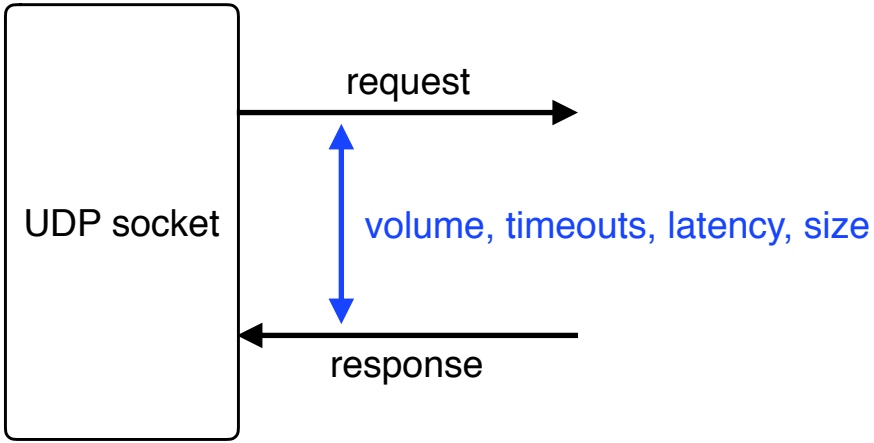
Can get visibility from the OS



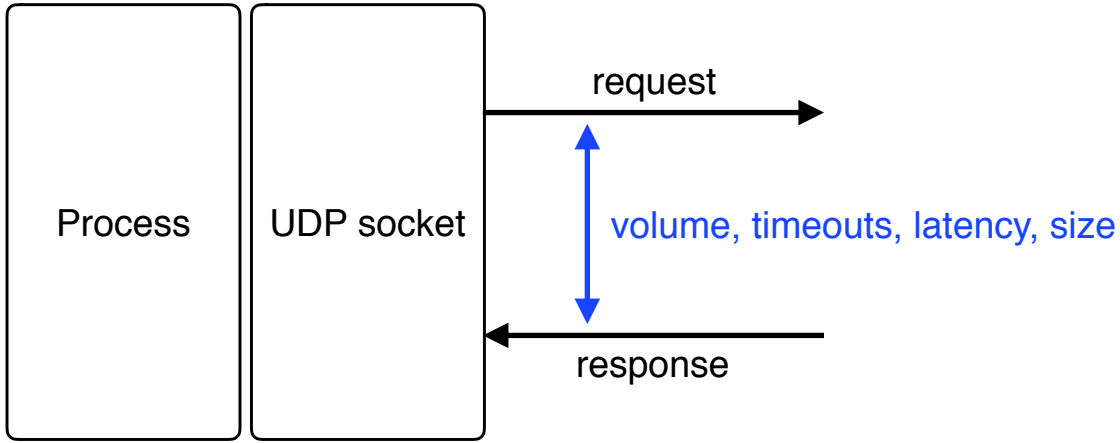
Can get visibility from the OS



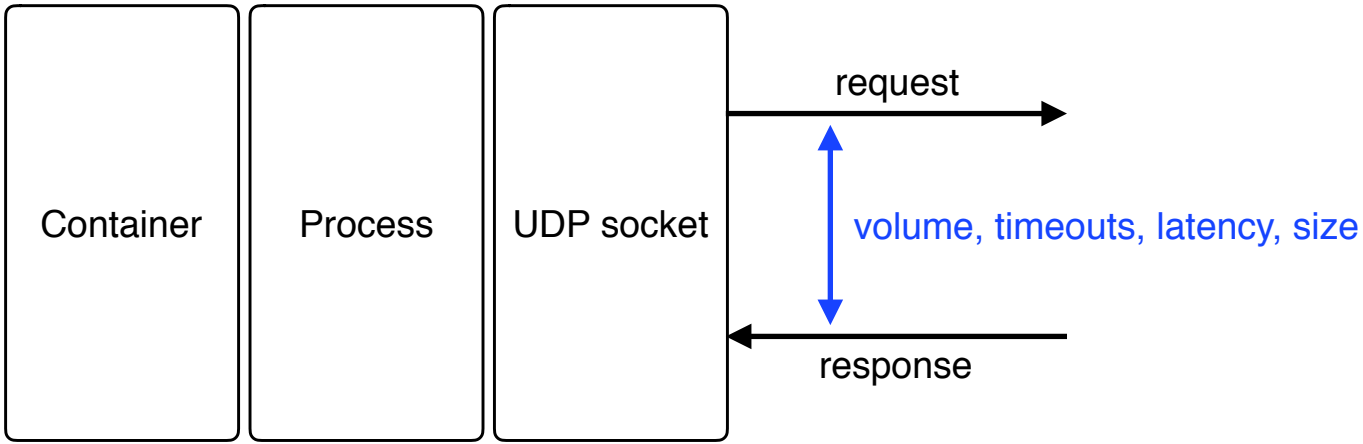
Can get visibility from the OS



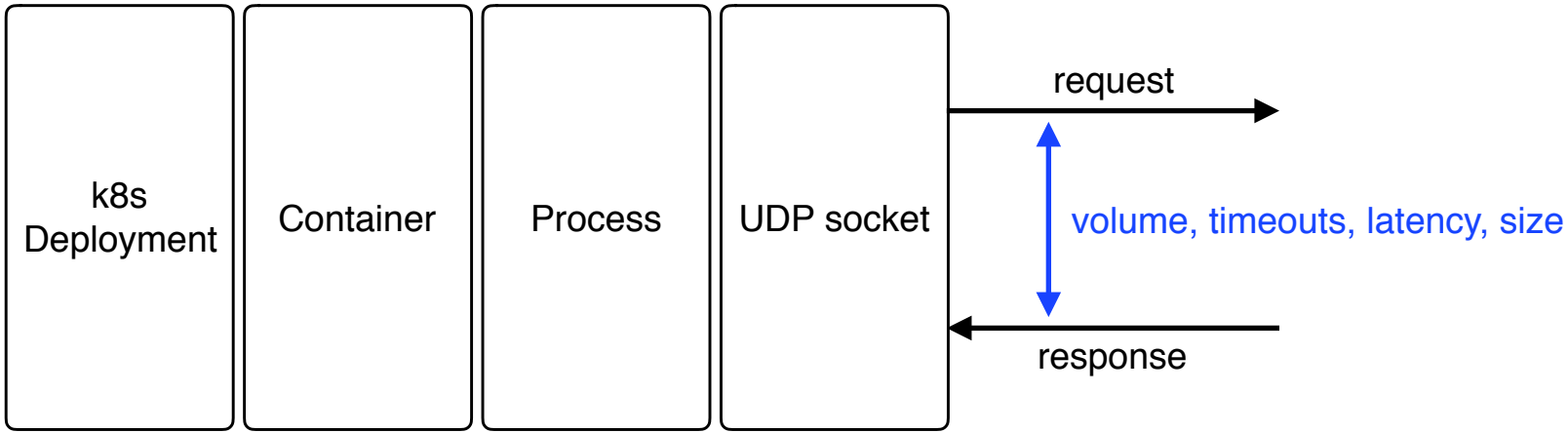
Can get visibility from the OS



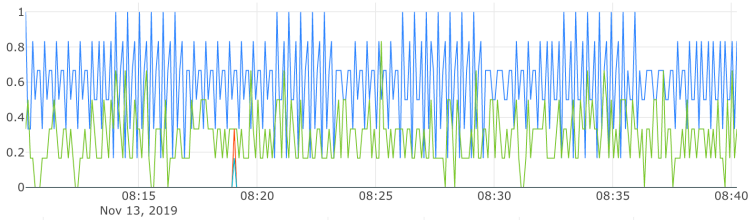
Can get visibility from the OS



Can get visibility from the OS

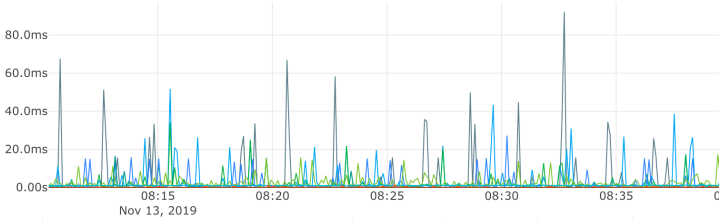


DNS Timeouts (DNS Timeouts)



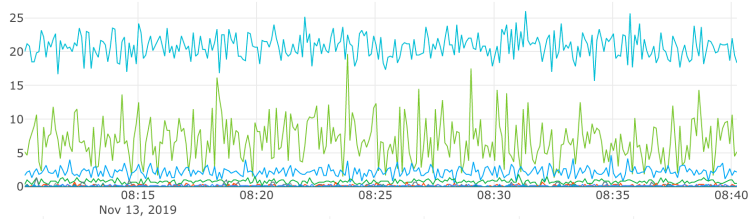
From Service	To Service	Max	Average	↓ Total
misconfiguredservice	DNS	1	0.6	1.1k
checkoutservice	DNS	0.8	0.3	520
kube-dns	DNS	0.3	0	2
loadgenerator	DNS	0.2	0	1
prometheus-to-sd	DNS	0	0	0
cartservice	DNS	0	0	0
coredns	DNS	0	0	0
ecs-agent	DNS	0	0	0
kernel-collector	DNS	0	0	0

DNS Latency (DNS Latency)



From Service	To Service	Min	Max	↓ Average
prometheus-to-sd	DNS	0s	92.1ms	3.32ms
coredns	DNS	151µs	51.8ms	3ms
checkoutservice	DNS	181µs	17.4ms	2.39ms
flowmill-k8s-collector	DNS	0s	33.9ms	1.77ms
misconfiguredservice	DNS	0s	27.2ms	1.36ms
loadgenerator	DNS	799µs	1.5ms	1.05ms
cartservice	DNS	0s	1.45ms	132µs
kernel-collector	DNS	0s	8.66ms	64.7µs
ecs-agent	DNS	0s	16.6ms	64.7µs

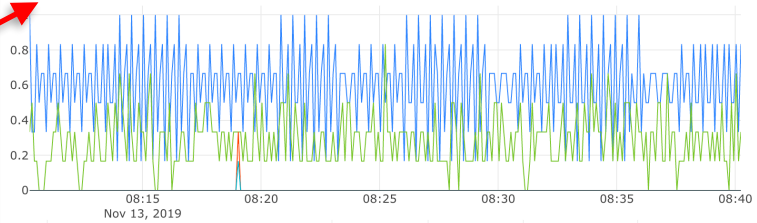
DNS Responses (DNS Responses)



From Service	To Service	Max	Average	↓ Total
loadgenerator	DNS	26	20.8	37.5k
checkoutservice	DNS	19.7	6.9	12.4k
coredns	DNS	4.7	2.2	4k
flowmill-k8s-collector	DNS	1.8	0.8	1.4k
cartservice	DNS	0.7	0.1	267
misconfiguredservice	DNS	0.3	0.1	236
prometheus-to-sd	DNS	1.3	0.1	148
ecs-agent	DNS	0.3	0	19
fluentd-gcp-v3.1.1	DNS	0.3	0	12

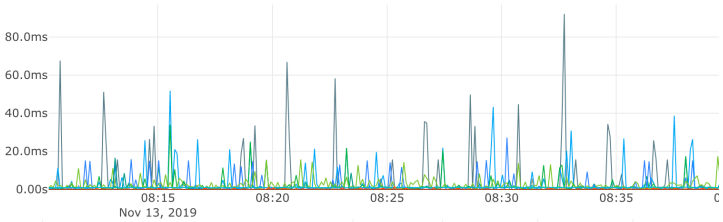
Timeouts

DNS Timeouts (DNS Timeouts)



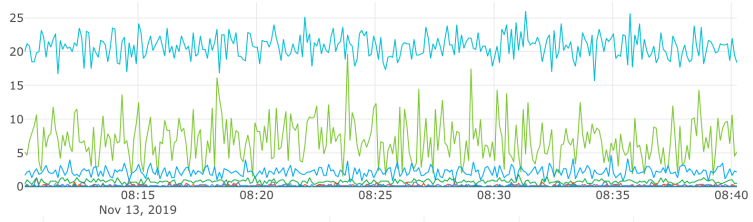
From Service	To Service	Max	Average	Total
misconfiguredservice	DNS	1	0.6	1.1k
checkoutservice	DNS	0.8	0.3	520
kube-dns	DNS	0.3	0	2
loadgenerator	DNS	0.2	0	1
prometheus-to-sd	DNS	0	0	0
cartservice	DNS	0	0	0
coredns	DNS	0	0	0
ecs-agent	DNS	0	0	0
kernel-collector	DNS	0	0	0

DNS Latency (DNS Latency)



From Service	To Service	Min	Max	Average
prometheus-to-sd	DNS	0s	92.1ms	3.32ms
coredns	DNS	151µs	51.8ms	3ms
checkoutservice	DNS	181µs	17.4ms	2.39ms
flowmill-k8s-collector	DNS	0s	33.9ms	1.77ms
misconfiguredservice	DNS	0s	27.2ms	1.36ms
loadgenerator	DNS	799µs	1.5ms	1.05ms
cartservice	DNS	0s	1.45ms	132µs
kernel-collector	DNS	0s	8.66ms	64.7µs
ecs-agent	DNS	0s	16.6ms	64.7µs

DNS Responses (DNS Responses)

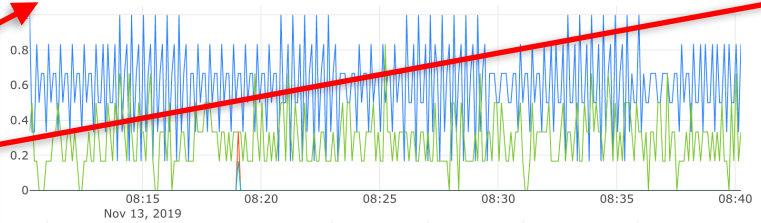


From Service	To Service	Max	Average	Total
loadgenerator	DNS	26	20.8	37.5k
checkoutservice	DNS	19.7	6.9	12.4k
coredns	DNS	4.7	2.2	4k
flowmill-k8s-collector	DNS	1.8	0.8	1.4k
cartservice	DNS	0.7	0.1	267
misconfiguredservice	DNS	0.3	0.1	236
prometheus-to-sd	DNS	1.3	0.1	148
ecs-agent	DNS	0.3	0	19
fluentd-gcp-v3.1.1	DNS	0.3	0	12

Timeouts

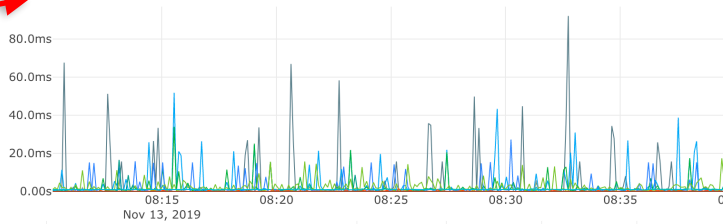
Latency

DNS Timeouts (DNS Timeouts)



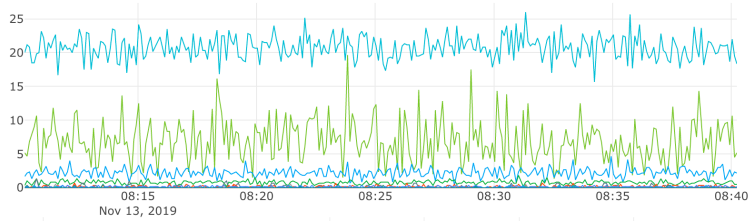
From Service	To Service	Max	Average	Total
misconfiguredservice	DNS	1	0.6	1.1k
checkoutservice	DNS	0.8	0.3	520
kube-dns	DNS	0.3	0	2
loadgenerator	DNS	0.2	0	1
prometheus-to-sd	DNS	0	0	0
cartservice	DNS	0	0	0
coredns	DNS	0	0	0
ecs-agent	DNS	0	0	0
kernel-collector	DNS	0	0	0

DNS Latency (DNS Latency)



From Service	To Service	Min	Max	Average
prometheus-to-sd	DNS	0s	92.1ms	3.32ms
coredns	DNS	151µs	51.8ms	3ms
checkoutservice	DNS	181µs	17.4ms	2.39ms
flowmill-k8s-collector	DNS	0s	33.9ms	1.77ms
misconfiguredservice	DNS	0s	27.2ms	1.36ms
loadgenerator	DNS	799µs	1.5ms	1.05ms
cartservice	DNS	0s	1.45ms	132µs
kernel-collector	DNS	0s	8.66ms	64.7µs
ecs-agent	DNS	0s	16.6ms	64.7µs

DNS Responses (DNS Responses)



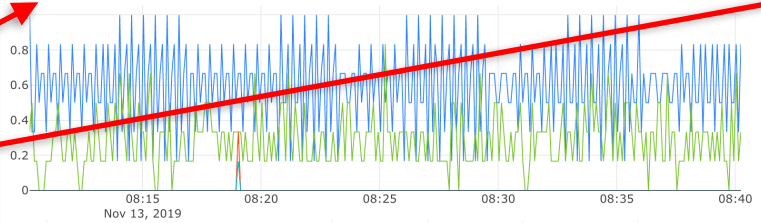
From Service	To Service	Max	Average	Total
loadgenerator	DNS	26	20.8	37.5k
checkoutservice	DNS	19.7	6.9	12.4k
coredns	DNS	4.7	2.2	4k
flowmill-k8s-collector	DNS	1.8	0.8	1.4k
cartservice	DNS	0.7	0.1	267
misconfiguredservice	DNS	0.3	0.1	236
prometheus-to-sd	DNS	1.3	0.1	148
ecs-agent	DNS	0.3	0	19
fluentd-gcp-v3.1.1	DNS	0.3	0	12

Timeouts

Latency

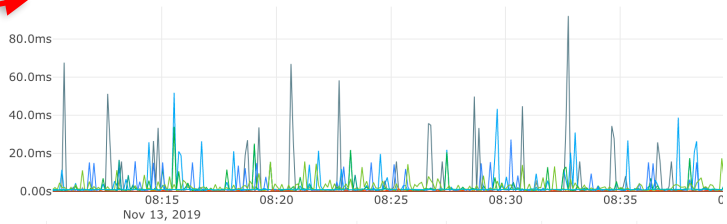
Volume

DNS Timeouts (DNS Timeouts)



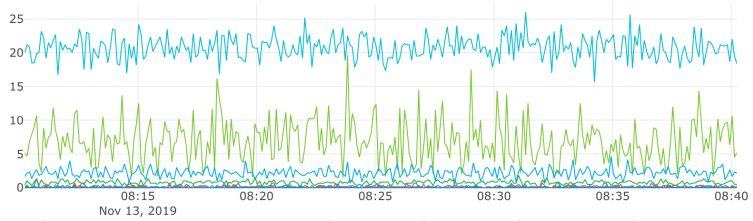
From Service	To Service	Max	Average	Total
misconfiguredservice	DNS	1	0.6	1.1k
checkoutservice	DNS	0.8	0.3	520
kube-dns	DNS	0.3	0	2
loadgenerator	DNS	0.2	0	1
prometheus-to-sd	DNS	0	0	0
cartservice	DNS	0	0	0
coredns	DNS	0	0	0
ecs-agent	DNS	0	0	0
kernel-collector	DNS	0	0	0

DNS Latency (DNS Latency)



From Service	To Service	Min	Max	Average
prometheus-to-sd	DNS	0s	92.1ms	3.32ms
coredns	DNS	151µs	51.8ms	3ms
checkoutservice	DNS	181µs	17.4ms	2.39ms
flowmill-k8s-collector	DNS	0s	33.9ms	1.77ms
misconfiguredservice	DNS	0s	27.2ms	1.36ms
loadgenerator	DNS	799µs	1.5ms	1.05ms
cartservice	DNS	0s	1.45ms	132µs
kernel-collector	DNS	0s	8.66ms	64.7µs
ecs-agent	DNS	0s	16.6ms	64.7µs

DNS Responses (DNS Responses)



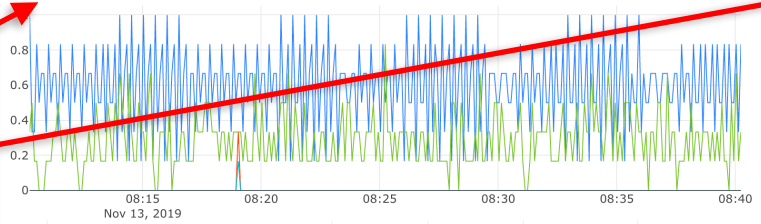
From Service	To Service	Max	Average	Total
loadgenerator	DNS	26	20.8	37.5k
checkoutservice	DNS	19.7	6.9	12.4k
coredns	DNS	4.7	2.2	4k
flowmill-k8s-collector	DNS	1.8	0.8	1.4k
cartservice	DNS	0.7	0.1	267
misconfiguredservice	DNS	0.3	0.1	236
prometheus-to-sd	DNS	1.3	0.1	148
ecs-agent	DNS	0.3	0	19
fluentd-gcp-v3.1.1	DNS	0.3	0	12

Timeouts

Latency

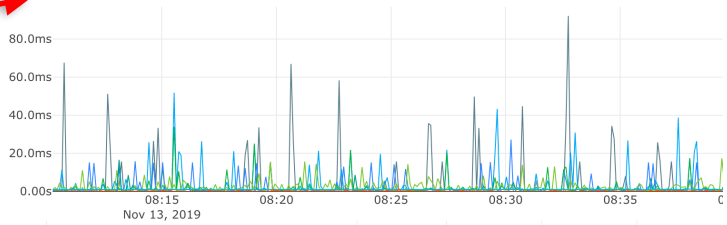
Volume

DNS Timeouts (DNS Timeouts)



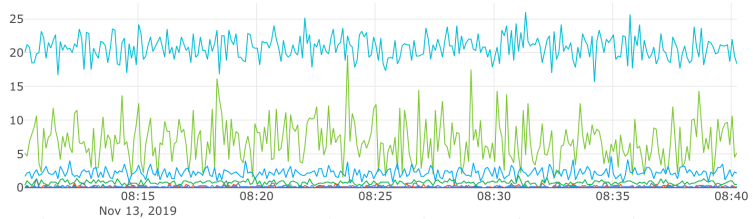
From Service	To Service	Max	Average	Total
misconfiguredservice	DNS	1	0.6	1.1k
checkoutservice	DNS	0.8	0.3	520
kube-dns	DNS	0.3	0	2
loadgenerator	DNS	0.2	0	1
prometheus-to-sd	DNS	0	0	0
cartservice	DNS	0	0	0
coredns	DNS	0	0	0
ecs-agent	DNS	0	0	0
kernel-collector	DNS	0	0	0

DNS Latency (DNS Latency)



From Service	To Service	Min	Max	Average
prometheus-to-sd	DNS	0s	92.1ms	3.32ms
coredns	DNS	151µs	51.8ms	3ms
checkoutservice	DNS	181µs	17.4ms	2.39ms
flowmill-k8s-collector	DNS	0s	33.9ms	1.77ms
misconfiguredservice	DNS	0s	27.2ms	1.36ms
loadgenerator	DNS	799µs	1.5ms	1.05ms
cartservice	DNS	0s	1.45ms	132µs
kernel-collector	DNS	0s	8.66ms	64.7µs
ecs-agent	DNS	0s	16.6ms	64.7µs

DNS Responses (DNS Responses)



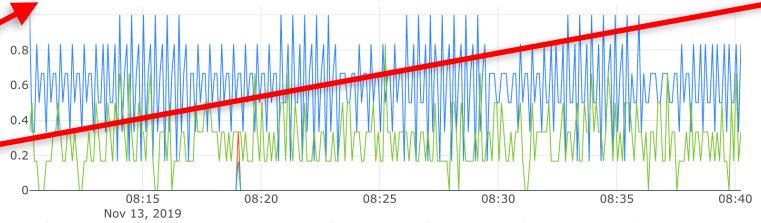
From Service	To Service	Max	Average	Total
loadgenerator	DNS	26	20.8	37.5k
checkoutservice	DNS	19.7	6.9	12.4k
coredns	DNS	4.7	2.2	4k
flowmill-k8s-collector	DNS	1.8	0.8	1.4k
cartservice	DNS	0.7	0.1	267
misconfiguredservice	DNS	0.3	0.1	236
prometheus-to-sd	DNS	1.3	0.1	148
ecs-agent	DNS	0.3	0	19
fluentd-gcp-v3.1.1	DNS	0.3	0	12

Timeouts

Latency

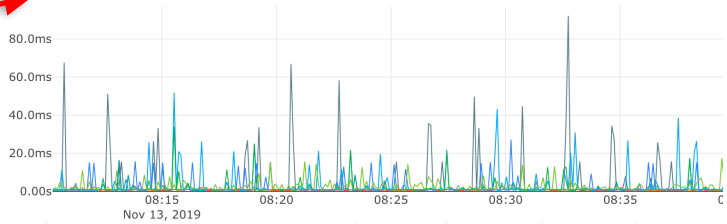
Volume

DNS Timeouts (DNS Timeouts)



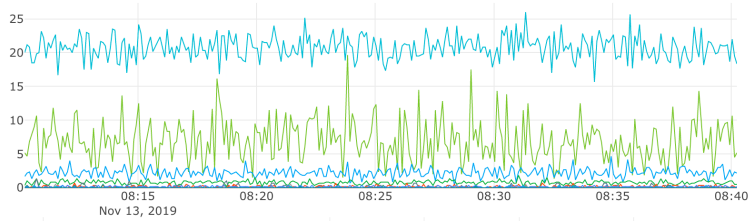
From Service	To Service	Max	Average	Total
misconfiguredservice	DNS	1	0.6	1.1k
checkoutservice	DNS	0.8	0.3	520
kube-dns	DNS	0.3	0	2
loadgenerator	DNS	0.2	0	1
prometheus-to-sd	DNS	0	0	0
cartservice	DNS	0	0	0
coredns	DNS	0	0	0
ecs-agent	DNS	0	0	0
kernel-collector	DNS	0	0	0

DNS Latency (DNS Latency)

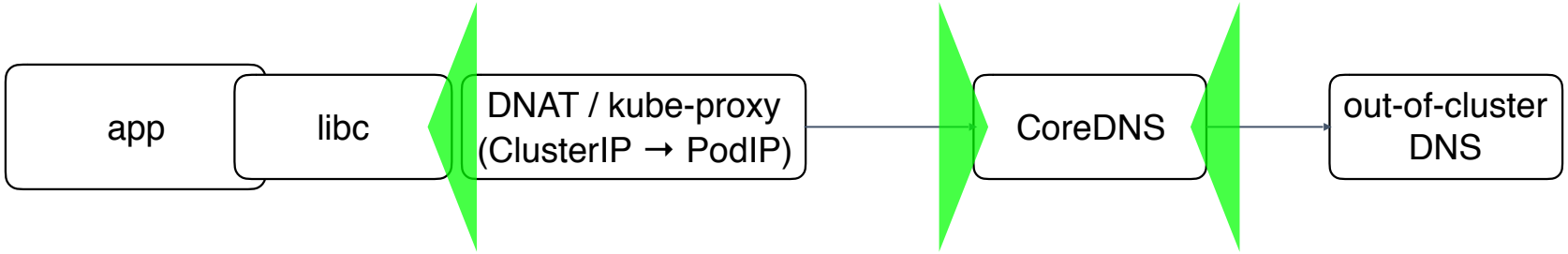


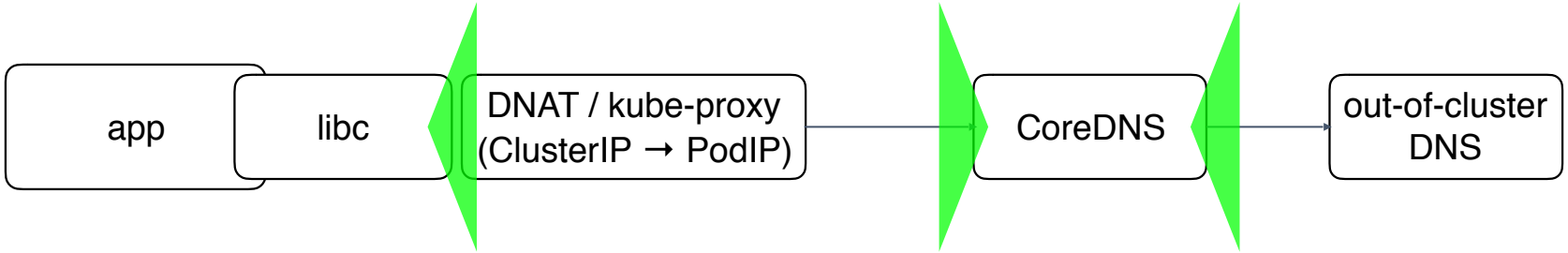
From Service	To Service	Min	Max	Average
prometheus-to-sd	DNS	0s	92.1ms	3.32ms
coredns	DNS	151µs	51.8ms	3ms
checkoutservice	DNS	181µs	17.4ms	2.39ms
flowmill-k8s-collector	DNS	0s	33.9ms	1.77ms
misconfiguredservice	DNS	0s	27.2ms	1.36ms
loadgenerator	DNS	799µs	1.5ms	1.05ms
cartservice	DNS	0s	1.45ms	132µs
kernel-collector	DNS	0s	8.66ms	64.7µs
ecs-agent	DNS	0s	16.6ms	64.7µs

DNS Responses (DNS Responses)



From Service	To Service	Max	Average	Total
loadgenerator	DNS	26	20.8	37.5k
checkoutservice	DNS	19.7	6.9	12.4k
coredns	DNS	4.7	2.2	4k
flowmill-k8s-collector	DNS	1.8	0.8	1.4k
cartservice	DNS	0.7	0.1	267
misconfiguredservice	DNS	0.3	0.1	236
prometheus-to-sd	DNS	1.3	0.1	148
ecs-agent	DNS	0.3	0	19
fluentd-gcp-v3.1.1	DNS	0.3	0	12


app
libc
DNAT / kube-proxy
(ClusterIP → PodIP)
CoreDNS
out-of-cluster
DNS



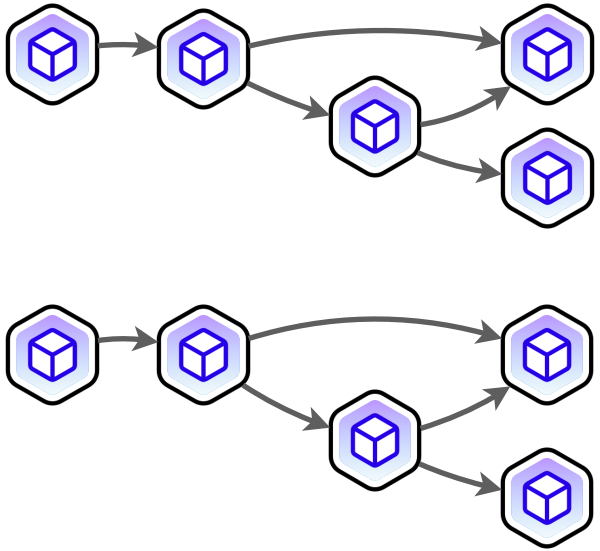
How? eBPF

- Linux bpf() system call, 3.18+, RHEL 7.6+
- Safe, High-Performance

Unofficial BPF mascot by Deirdré Straughan

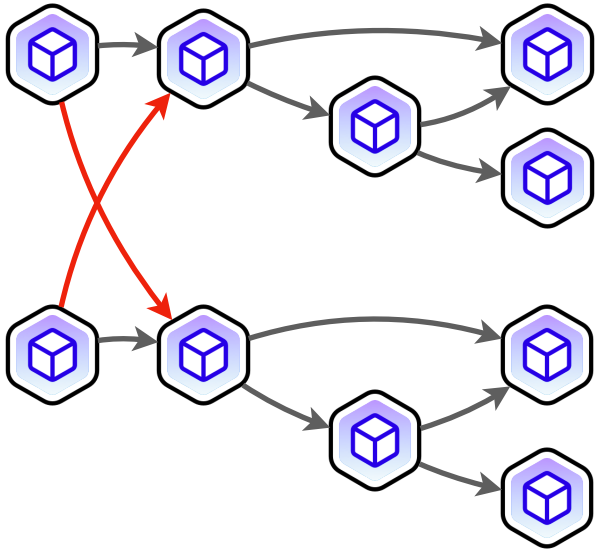
OS+network can provide more visibility

- **Map application & HA architecture**



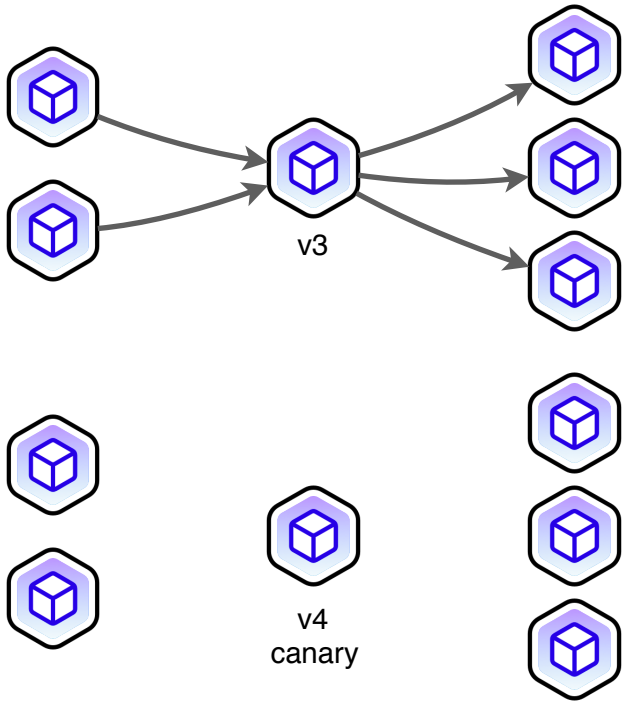
OS+network can provide more visibility

- **Map application & HA architecture**



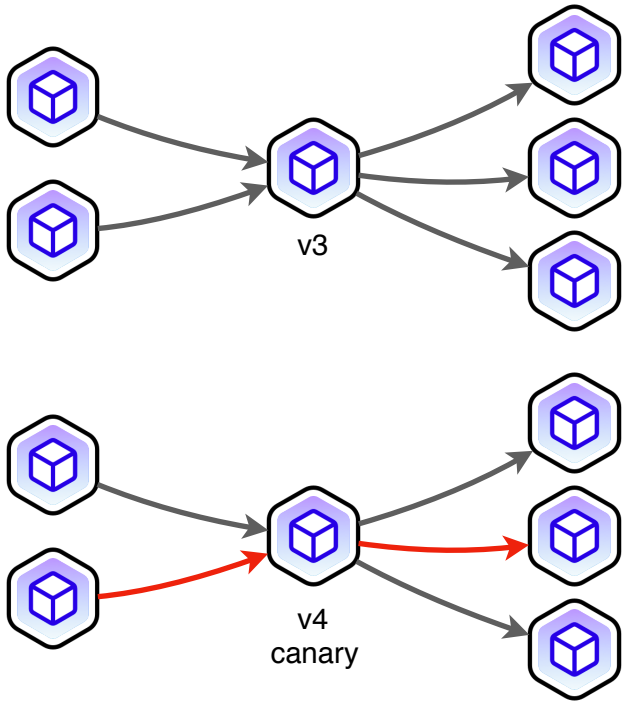
OS+network can provide more visibility

- Map application & HA architecture
- **Track SLOs between services**



OS+network can provide more visibility

- Map application & HA architecture
- **Track SLOs between services**

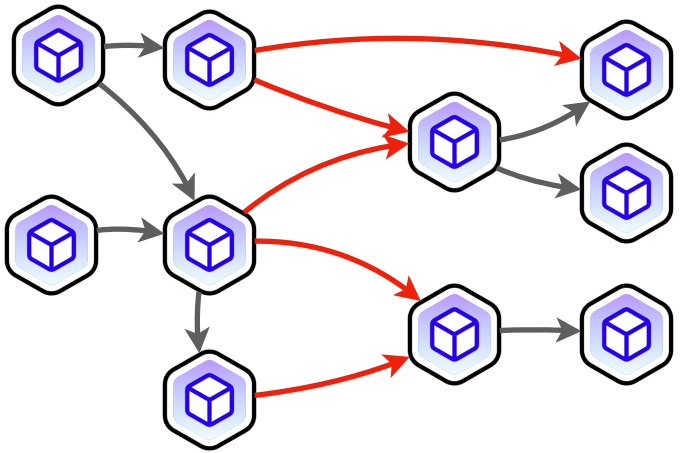


OS+network can provide more visibility

- Map application & HA architecture
- Track SLOs between services
- **Detect network issues**

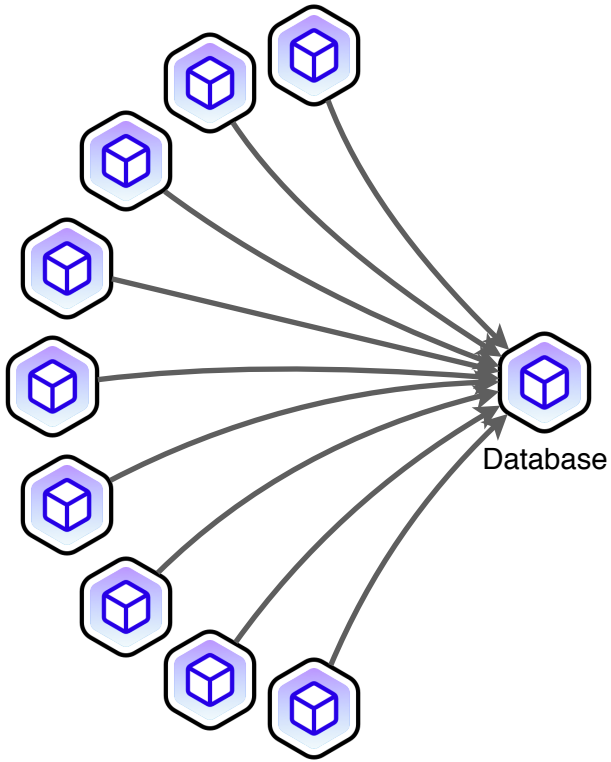
OS+network can provide more visibility

- Map application & HA architecture
- Track SLOs between services
- **Detect network issues**



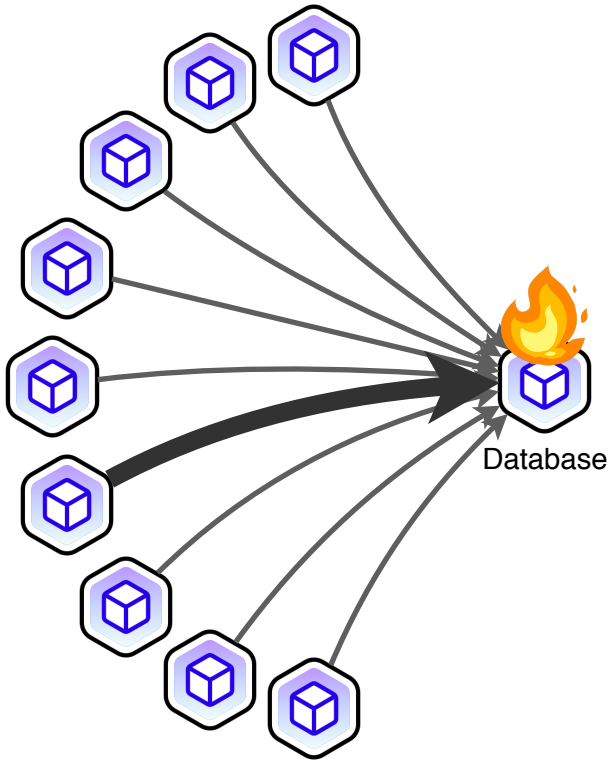
OS+network can provide more visibility

- Map application & HA architecture
- Track SLOs between services
- Detect network issues
- **Identify self-DDoS**



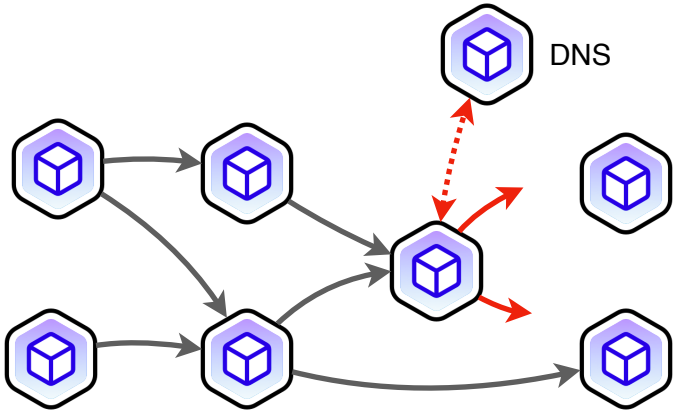
OS+network can provide more visibility

- Map application & HA architecture
- Track SLOs between services
- Detect network issues
- **Identify self-DDoS**



Take-aways:

- DNS can (and does) fail
- CoreDNS metrics are useful
- Can get per-service visibility with eBPF
- Network telemetry useful beyond DNS



Please
come say hi!