Talking IPv6 to IPv4 Without NAT

Alok Tiagi, Keerti Lakshminarayan, Kyle Anderson

(Netflix)

**LINUX PLUMBERS CONFERENCE** > September 20-24, 2021

WHY

- Netflix uses 100k+ instances in AWS EC2
- We use a network architecture with non-overlapping CIDRs
- Testing the scaling limits of AWS
- Alternatives to the flat IPv4 network exist but..

- **Move to IPv6**

WHY

- Dual Stack doesn't help
- How do we move from v4 to v6 transparently?

IPv6 vs IPv4 Flows Percentage Observed from EC2 Instances

IPv6 ■ IPv4
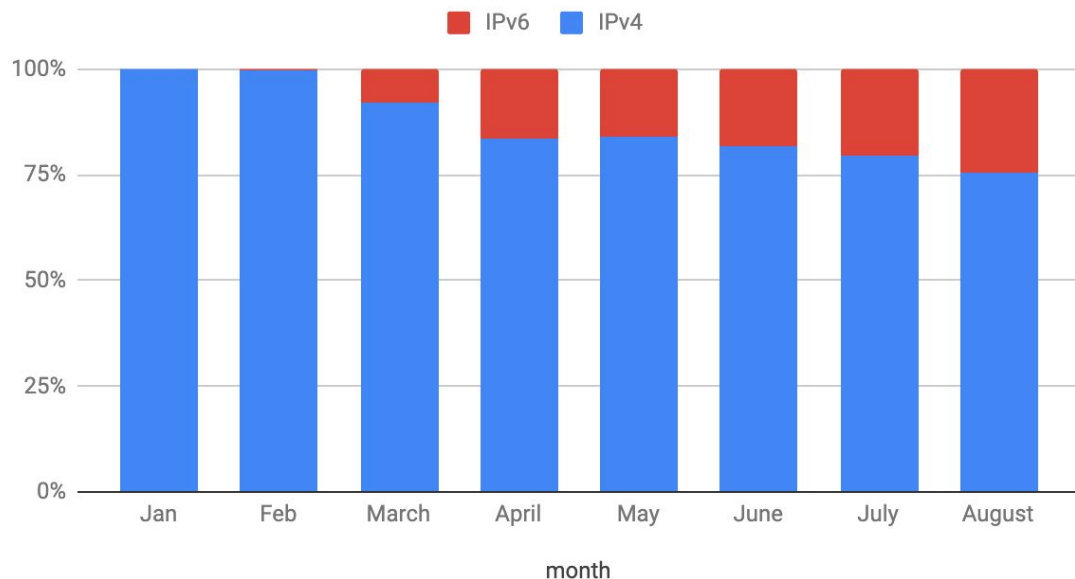
How do we make this all Red?

Current state of the art

NAT (Network Address Translation)

- TAYGA: http://www.litech.org/tayga/
- NAT on dedicated hardware

Per Packet Processing! 🙁

Is there a more optimal solution?

(Spoiler: TSA)

Titus Syscall Agent

Powerful and efficient mechanism to intercept system calls.

- How does it work?
- TSA Architecture
- Demos
- Future
- Q&A

Titus = Netflix's Multi-tenant Container Platform

How does TSA's Syscall Interception work?

Existing approaches

- LD_PRELOAD
- Kernel Modules
- Strace/Ptrace/Sysdig
- BPF
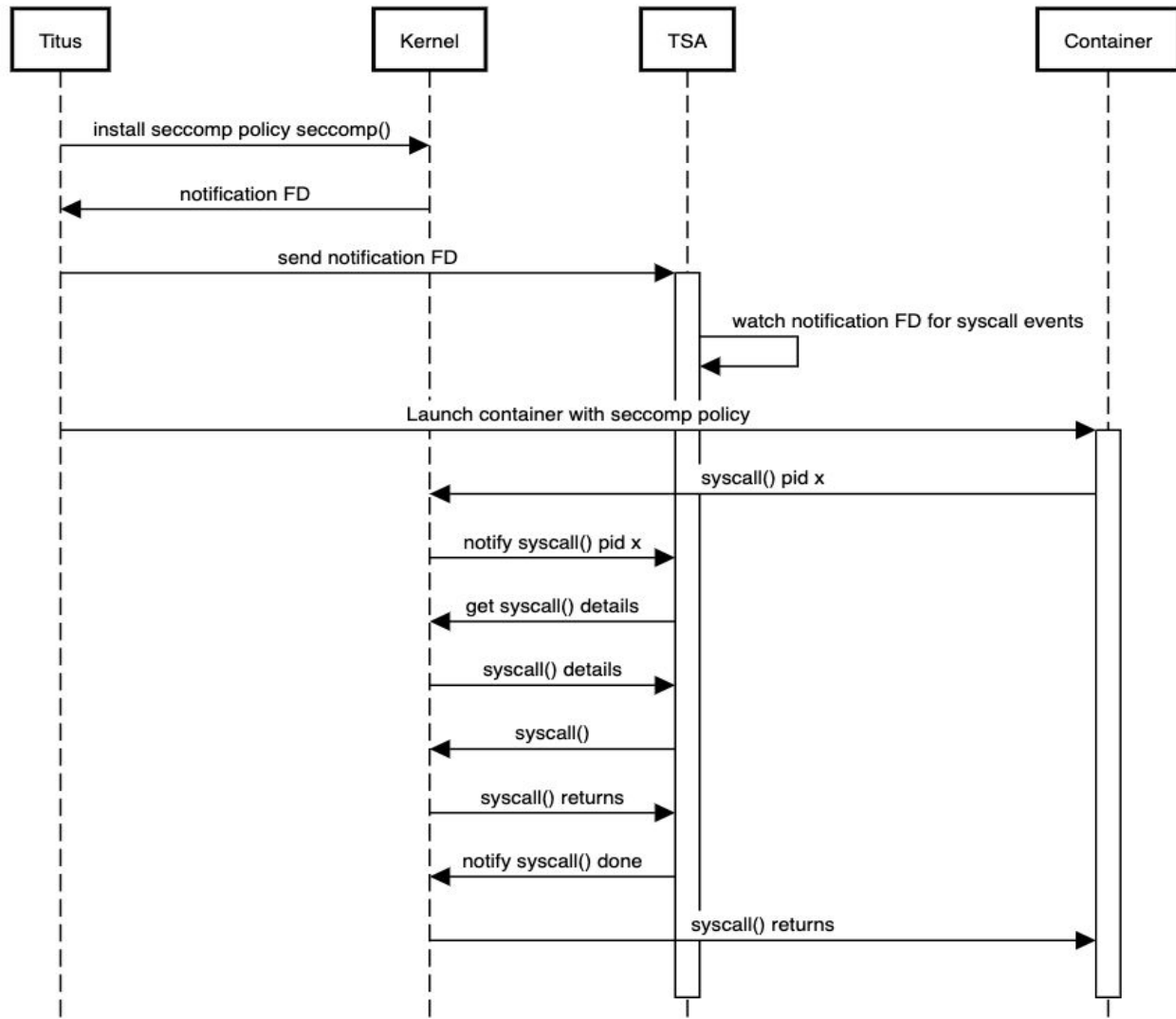- **Seccomp**

What's new?

- Seccomp Notifier (LPC'17)

- Seccomp Notifier + add_fd support (2020)

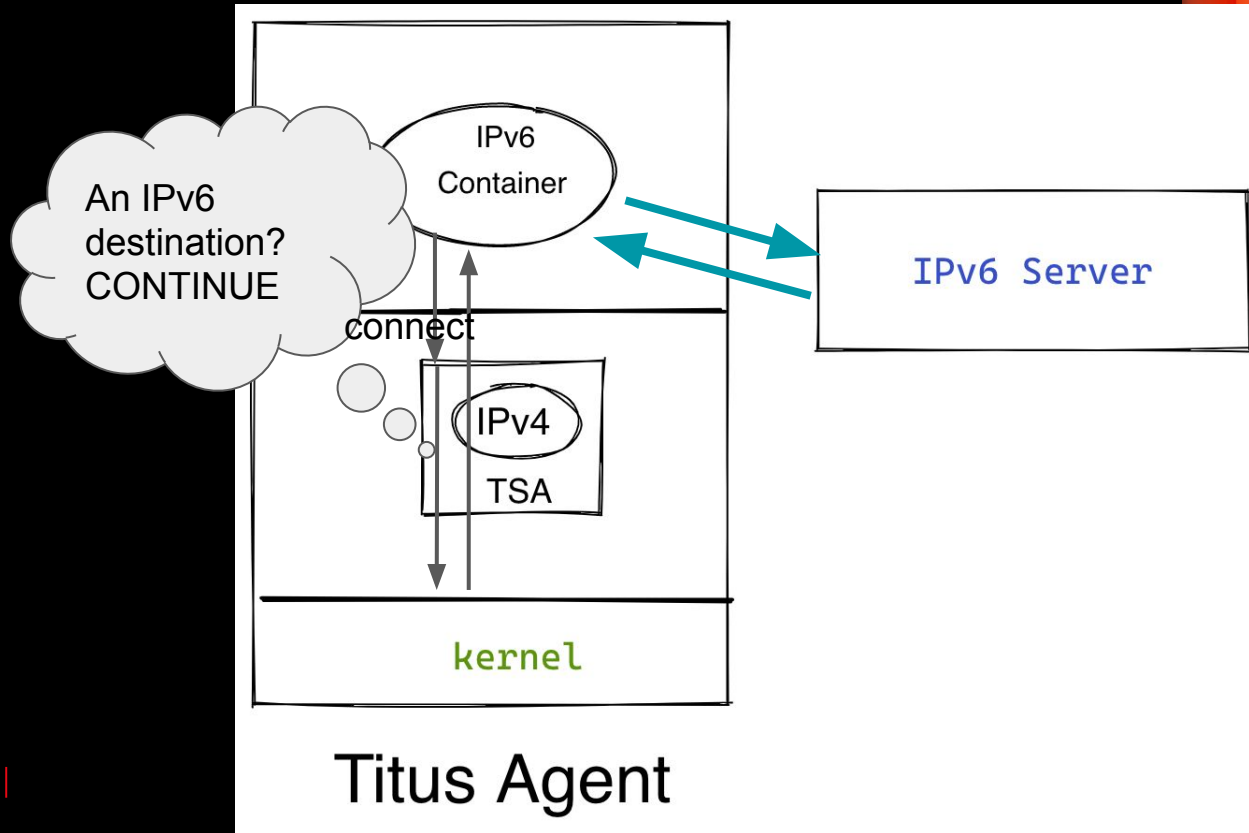Credit to sargun@sargun.me (happens to work on our Team!)
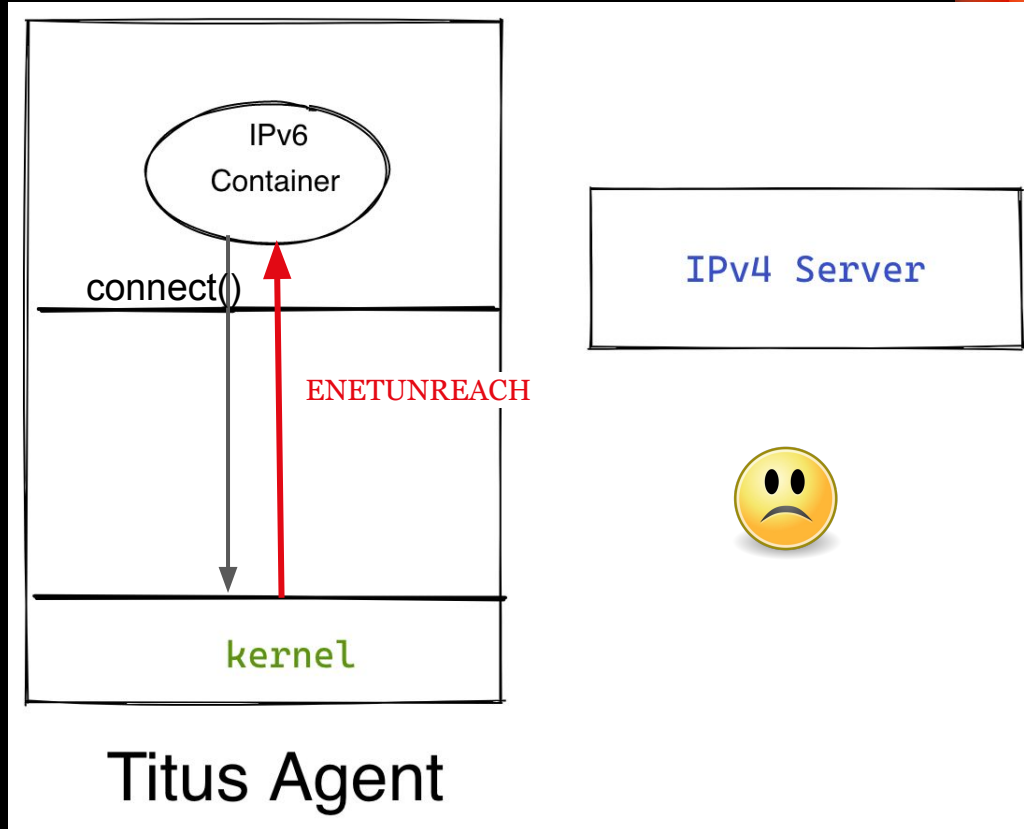
Who likes Sequence Diagrams???
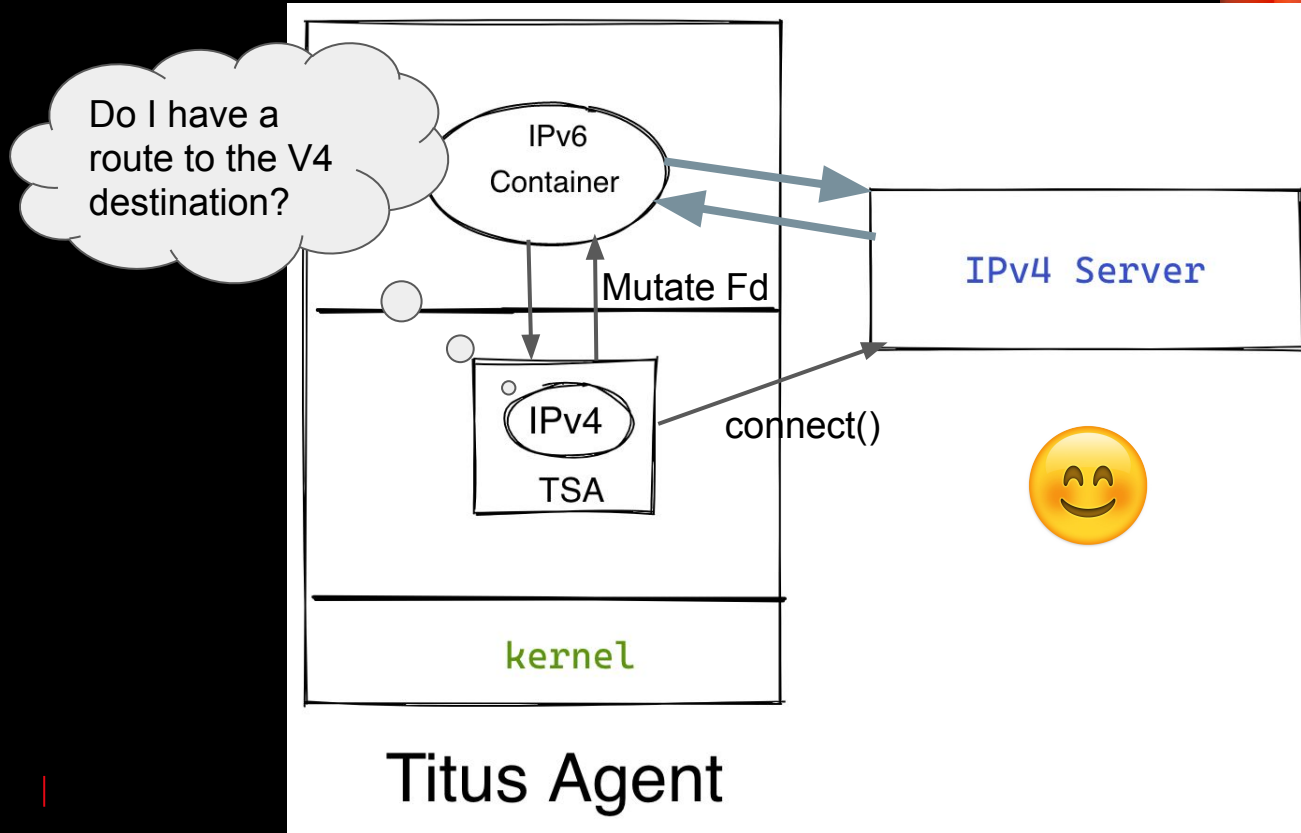
How does TSA intercept syscalls?

Sequence diagram with participants: Titus, Kernel, TSA, Container

- Titus → Kernel: install seccomp policy seccomp()
- Kernel → Titus: notification FD
- Titus → TSA: send notification FD
- TSA → TSA: watch notification FD for syscall events
- Titus → Container: Launch container with seccomp policy
- Container → Kernel: syscall() pid x
- Kernel → TSA: notify syscall() pid x
- TSA → Kernel: get syscall() details
- Kernel → TSA: syscall() details
- TSA → Kernel: syscall()
- Kernel → TSA: syscall() returns
- TSA → Kernel: notify syscall() done
- Kernel → Container: syscall() returns

How does v6 only->v6 work?

V6 only -> v4 only (without TSA) ?

v6 only -> v4 only with TSA?

Demo: TSA

8.00

6.00

4.00

2.00

0.00

```
PDT: 14:00 - Saturday, May 15 2021
● connect: 3.91
● sendm_msg: 454m
● sendmsg: 7.23
● sendto: 5.05
Press + hold ctrl to view raw numbers
```

May15　May16　May17　May18　May19　May20　May21

| Hide All | Show All | Show Filtered Only | Step: 20m | | Filter: |
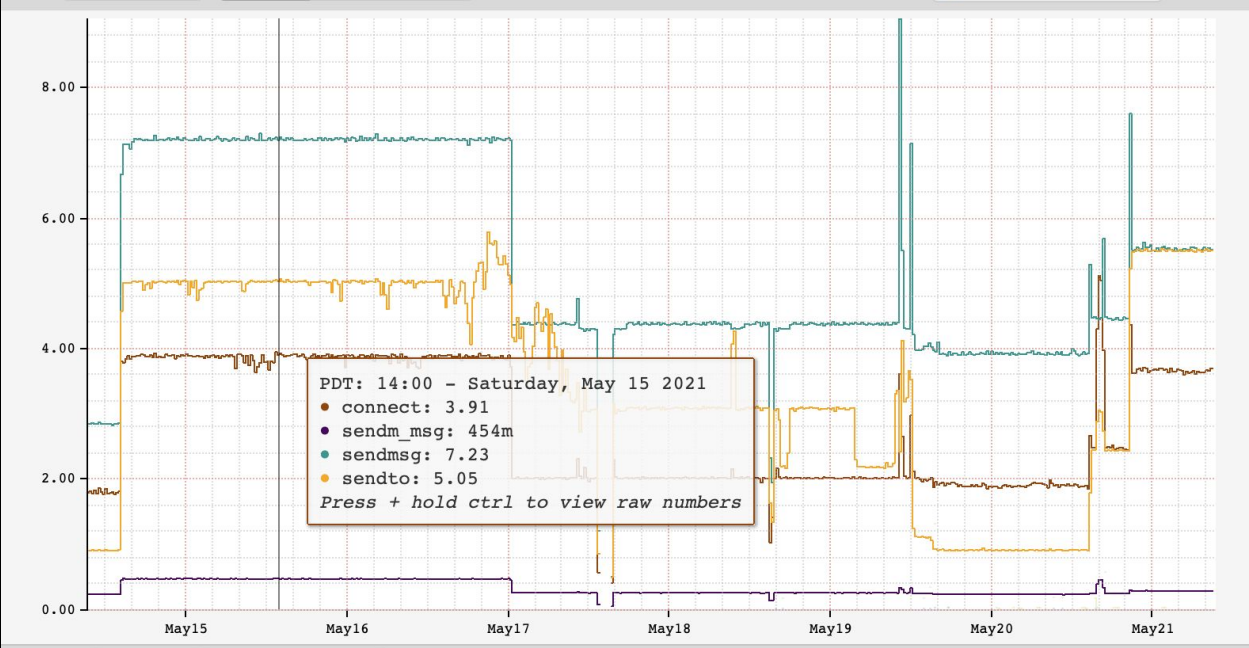
Min	Max	Average	Total	Name				Visibility
833μ	15.0m	3.15m	28.3m	🟦 bpf				☀️ 🌙 ⊘
387m	5.10	2.78	1.39k	🟫 connect				☀️ 🌙 ⊘
0.00	1.30	129m	3.34	🟩 perf_event_open				☀️ 🌙 ⊘
31.7m	458m	313m	156	🟪 sendm_msg				☀️ 🌙 ⊘
492m	9.04	5.33	2.65k	🟩 sendmsg				☀️ 🌙 ⊘
472m	5.78	3.50	1.74k	🟧 sendto				☀️ 🌙 ⊘

Future

❖ Kernel: "seccomp: epoll issues post addfd"
❖ TSA Transition on by default @Netflix
❖ LongTerm : IPv6 Mostly but TSA will get invoked for 3rd party IPv4 only servers
❖ TSA will continue to be used for advanced features like Perf, Bpftools

NETFLIX

Thank You.