From Theory to Production: eBPF's Role in Next-Generation Systems

Angelo Tulumello
PhD School @ TMA ‘24 - May 21st, Dresden
Introduction

- eBPF is fueling next-generation ICT systems
- We will see *three* Networking applications which use eBPF at its foundations

- *Katran*
  - Meta
  - L4 load balancer

- *Isovalent*
  - K8s Networking

- *Cloudflare*
  - Cloudflare
  - DDoS mitigation
Katran: a L4 load balancer
What is a Load Balancer?

● A device which distributes user requests to different servers
  ○ for scalability and fault tolerance → a fundamental component in Data Center networks

● You can do it with many different techniques...
  ○ with DNS, ECMP, etc.

● ... and at different levels
  ○ L4 and L7 load balancing
Meta Load Balancing in the past

- Two layers of load balancing → L4 and L7
  - with different machines specialized for L4 and L7
Meta Load Balancing in the past

- Two layers of load balancing → L4 and L7
  - with different machines specialized for L4 and L7
- Before eBPF, L4 load balancing in Linux was IPVS
  - A Linux Kernel module for load balancing built on top of Netfilter

Problems:

- Did not scale well, especially with high number of new connections
- Low flexibility → adding functionality to IPVS means changing the kernel (with all the correlated problems we have seen so far)
- High CPU usage → machines dedicated to L4 load balancing could do only that
eBPF/XDP to the rescue

- At some point (2018), Meta engineers replaced IPVS with an eBPF program in the XDP hook
Katran: eBPF based L4 Load Balancer

katran:
1. if service is configured
2. if belongs to existing session
3. calculate hash
4. pick real
5. update sessions table
6. encapsulate packet
7. and send to "real"

https://github.com/facebookincubator/katran
Katran: eBPF based L4 Load Balancer

- At some point (2018), Meta engineers replaced IPVS with an eBPF program in the XDP hook

Features:
- Fast → thanks to the processing at XDP level
- Scalability → scales linearly with the number of cores
- Custom Load Balancing strategy → modified Maglev hashing for efficient balancing and possibility to configure unequal weights
Katran: performance

- Able to process 3x packets with 7x less CPU usage
- Positive side benefit → Now the Load Balancing servers can host other applications
  - For example they can host L4 + L7 proxy in the same machine
L4Drop: DDoS mitigation
What is a DDoS attack?

- Denial of Service: an attacker attempts to exhaust a system’s resources
  - Flooding a system with high-capacity traffic, exhausting the capacity of the links
  - Application DoS → use flaws in a specific application (e.g. using special HTTP requests)

- In Distributed DoS (DDoS), the attacker controls a number of machines (botnet) that simultaneously flood the target
What is a DDoS attack?

- Denial of Service, an attacker attempts to exhaust a system's resources
  - Flooding a system with high-speed traffic, exhausting the capacity of the links
  - Application DoS → use flaws in a specific application (e.g. using special HTTP requests)
- In Distributed DoS (DDoS), the attacker controls a number of machines (botnet) that simultaneously flood the target
What is a DDoS attack?

- **Denial of Service**: an attacker attempts to exhaust a system’s resources
  - Flooding a system with high-capacity traffic, exhausting the capacity of the links
  - Application DoS → use flaws in a specific application (e.g. using special HTTP requests)

- In **Distributed DoS (DDoS)**, the attacker controls a number of machines (botnet) that simultaneously flood the target

- *Very hard to prevent* → in fact we talk about *mitigation*

- **How**?
  - must discriminate malicious traffic over legitimate traffic
  - at traffic speeds that can reach *several Terabits per second*
## Effects of downtime

<table>
<thead>
<tr>
<th>COMPANY</th>
<th>ESTIMATED ANNUAL ECOMMERCE REVENUE</th>
<th>REVENUE LOSS PER HOUR</th>
<th>REVENUE LOSS PER MINUTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon.com</td>
<td>$115,879,000,000.00</td>
<td>$13,219,128.00</td>
<td>$220,318.80</td>
</tr>
<tr>
<td>WalMart.com</td>
<td>$21,443,900,000.00</td>
<td>$2,446,272.00</td>
<td>$40,771.20</td>
</tr>
</tbody>
</table>
Cloudflare DDoS mitigation

- CloudFlare is one of the most eminent Security Cloud Provider
- How DDoS mitigation is handled in Cloudflare?
Cloudflare DDoS mitigation

- CloudFlare is one of the most eminent Security Cloud Provider
- How DDoS mitigation is handled in Cloudflare?
Cloudflare DDoS mitigation

- CloudFlare is one of the most eminent Security Cloud Provider
- How DDoS mitigation is handled in Cloudflare?
Cloudflare DDoS mitigation

- CloudFlare is one of the most eminent Security Cloud Provider
- How DDoS mitigation is handled in Cloudflare?
Cloudflare DDoS mitigation

- CloudFlare is one of the most eminent Security Cloud Provider
- How DDoS mitigation is handled in Cloudflare?
Cloudflare DDoS mitigation

- CloudFlare is one of the most eminent Security Cloud Provider
- How DDoS mitigation is handled in Cloudflare?
Cloudflare DDoS mitigation

- CloudFlare is one of the most eminent Security Cloud Provider
- How DDoS mitigation is handled in Cloudflare?
Cloudflare DDoS mitigation

- CloudFlare is one of the most eminent Security Cloud Provider
- How DDoS mitigation is handled in Cloudflare?
eBPF to the rescue: L4 Drop

- Rewrite the FloodGate component in eBPF @XDP hook
- Gatebot automatically generates *eBPF programs* to drop packets
Results in production

Attack starts
Results in production

Gatebot detects it and dropping starts
Results in production

CPU utilization increases of about 15-20%
Results in production

CPU utilization increases of about 15-20%

CPU → 100%

with the iptables/Floodgate implementation
Cilium: Kubernetes Networking
With Cloud, networks changed a lot

- **VMs** 2003-2008
- **SDN** 2009
- **Docker** 2013
- **K8s - eBPF** 2014

**Apps**
- Metal Servers
- Virtual Machines
- Containers
- Kubernetes

**Net**
- Hardware Networking
- Software Defined Networking
- eBPF
Kubernetes (K8s)

- Kubernetes automates deploying, scaling, and orchestrating application containers across clusters of machines
  - abstracting the infrastructure below
- K8s pods exchange messages via virtual networking...
  - to communicate within the same node
- ... or via host networking
  - for inter-node communication or outside the cluster
- First K8s networking relied on iptables
  - eBPF was at its beginning...
  - it was (and still is, btw) the established tool for network programming in Linux
- As K8s evolved, the need for a flexible and scalable networking stack rose
Kubernetes (K8s)

- Kubernetes automates deploying, scaling, and orchestrating application containers across clusters of machines
  - abstracting the infrastructure below
- K8s pods exchange messages via *virtual networking*...
  - to communicate within the same node
- ... or via *host networking*
  - for inter-node communication or outside the cluster
- First K8s networking relied on iptables
  - eBPF was at its beginning...

- eBPF chosen as the way to connect K8s workloads
Cilium

- Open source project that addresses networking, security and visibility of container workloads, built on top of eBPF
Standard K8s Networking
Cilium
Cilium Features: Networking

- **Network connectivity to K8s workloads:**
  - Cilium is a Container Network Interface (CNI) plugin for K8s networking
  - re-implements routing, encapsulation, integration with external networks

- **Load balancing**
  - L7 Service load balancing → attaching to the socket `connect()`
    - including SSL termination
  - L4 Edge load balancing → XDP-based load balancing across K8s cluster

- **Connectivity between clusters**
  - no additional gateways or proxies

- **Integration with bare-metal servers**
  - seamless integration of bare-metal or VM machines as they were part of K8s cluster
Cilium Features: Security and Observability

- **Network Policies**
  - Security Policies built entirely in eBPF

- **Policy enforcement on API level**
  - Security Policies at API level, e.g. HTTP, Kafka, gRPC → enforce security policies tailored to the specific application

- **Flows visibility**
  - Auditing and logs on network flows at L3-L7

- **Troubleshooting**
  - Tracing systems with eBPF enables tracking of every functionality of Cilium