past, present & future of

High Speed Packet Filtering on Linux

CLOUDFLARE®

Gilberto Bertin

$ whoami

- System engineer at Cloudflare
  - DDoS mitigation team
- Enjoy messing with networking and low level things

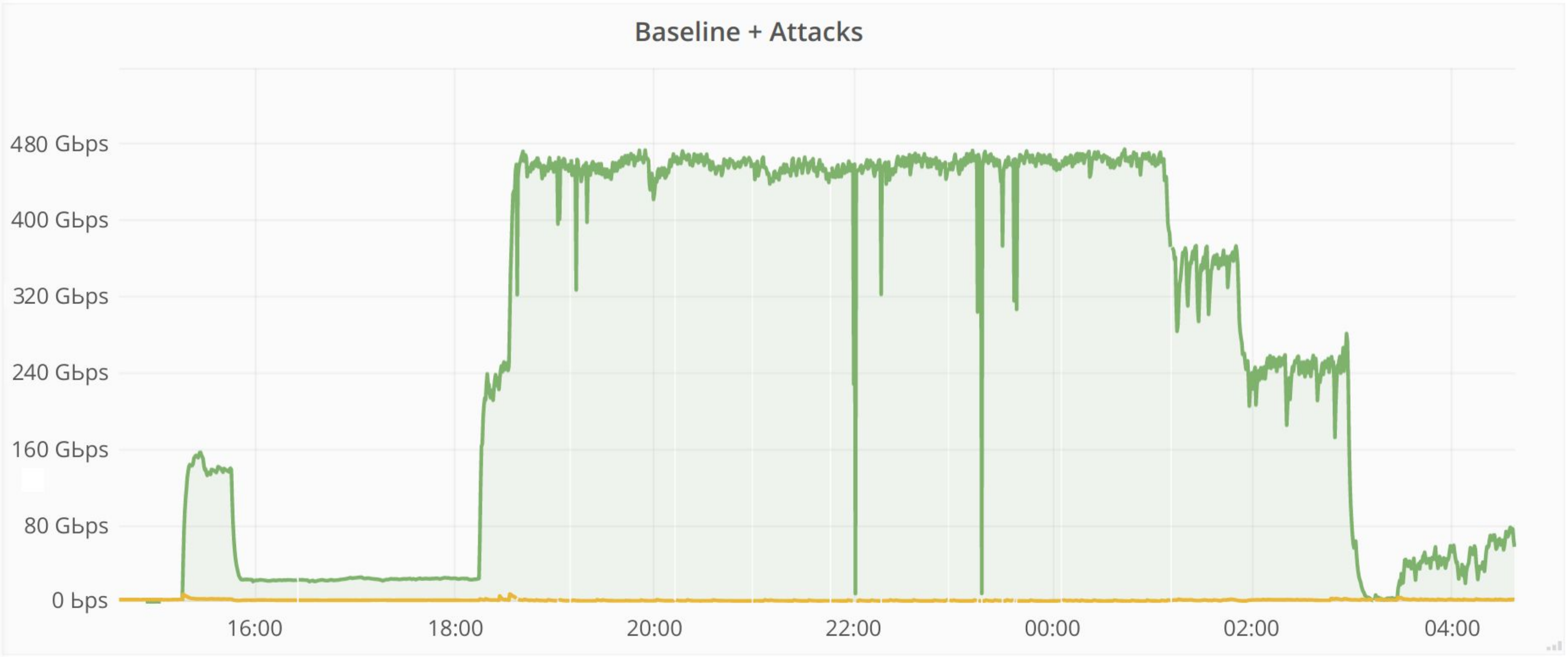
Cloudflare

- 115 PoPs
- 6M+ domains
- 4.8M HTTP requests per second
- 1.2M DNS queries per second

Everyday we have to mitigate hundreds of different DDoS attacks

Size of the attacks

- On a normal day:
    - 50-100Mpps
    - 50-250Gbps
- Recorded peaks:
    - 250Mpps
    - 480Gbps



Baseline + Attacks
480 Gbps
400 Gbps
320 Gbps
240 Gbps
160 Gbps
80 Gbps
0 bps
16:00
18:00
20:00
22:00
00:00
02:00
04:00

Iptables

Iptables is great

- Well known CLI
- Lots of tools and libraries to interface with it
- Concept of tables and chains
- Integrates well with Linux
  - IPSET
  - accounting
- BPF matches support (xt_bpf)

Handling SYN floods with Iptables, BPF and p0f

```
$ ./bpfgen p0f -- '4:64:0:*:mss*10,6:mss,sok,ts,nop,ws:df,id+:0'
56,0 0 0 0,48 0 0 8,37 52 0 64,37 0 51 29,48 0 0 0,84 0 0 15,21 0 48 5,48 0 0
9,21 0 46 6,40 0 0 6,69 44 0 8191,177 0 0 0,72 0 0 14,2 0 0 8,72 0 0 22,36 0 0
10,7 0 0 0,96 0 0 8,29 0 36 0,177 0 0 0,80 0 0 39,21 0 33 6,80 0 0 12,116 0 0
4,21 0 30 10,80 0 0 20,21 0 28 2,80 0 0 24,21 0 26 4,80 0 0 26,21 0 24 8,80 0
0 36,21 0 22 1,80 0 0 37,21 0 20 3,48 0 0 6,69 0 18 64,69 17 0 128,40 0 0 2,2
0 0 1,48 0 0 0,84 0 0 15,36 0 0 4,7 0 0 0,96 0 0 1,28 0 0 0,2 0 0 5,177 0 0
0,80 0 0 12,116 0 0 4,36 0 0 4,7 0 0 0,96 0 0 5,29 1 0 0,6 0 0 65536,6 0 0 0,

$ BPF=(bpfgen p0f -- '4:64:0:*:mss*10,6:mss,sok,ts,nop,ws:df,id+:0')
# iptables -A INPUT -d 1.2.3.4 -p tcp --dport 80 -m bpf --bytecode "${BPF}"
```

bpftools: https://github.com/cloudflare/bpftools

Iptables can’t handle big packet floods.

It can filter 2-3Mpps at most, leaving no CPU to the userspace applications.

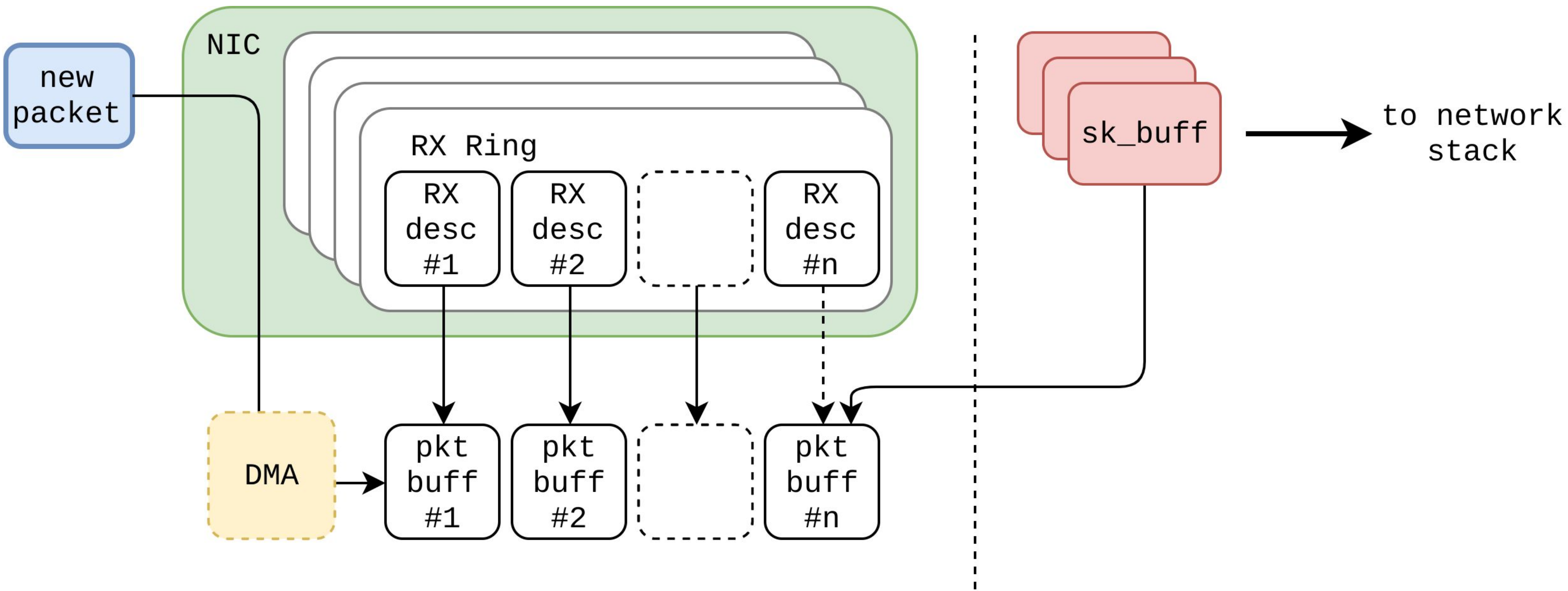
Linux alternatives

- Use raw/PREROUTING
- TC-bpf on ingress
- NFTABLES on ingress

We are not trying to squeeze
some more Mpps.
We want to use as little CPU as possible
to filter at line rate.

The path of a packet in the Linux Kernel

NIC and kernel packet buffers



NAPI polls the NIC

- net_rx_action() -> napi_poll()
- for each RX buffer that has a new packet:
  - dma_unmap() the packet buffer
  - build_skb()
  - netdev_alloc_frag() && dma_map() a new packet buffer
  - pass the skb up to the stack
  - free_skb()
  - free old packet page

```
net_rx_action() {
  e1000_clean [e1000]() {
    e1000_clean_rx_irq [e1000]() {
      build_skb() {
        __build_skb() {
          kmem_cache_alloc();
        }
      }
      _raw_spin_lock_irqsave();
      _raw_spin_unlock_irqrestore();
      skb_put();
      eth_type_trans();
      napi_gro_receive() {
        skb_gro_reset_offset();
        dev_gro_receive() {
          inet_gro_receive() {
            tcp4_gro_receive() {
              __skb_gro_checksum_complete() {
                skb_checksum() {
                  __skb_checksum() {
                    csum_partial() {
                      do_csum();
                    }
                  }
                }
              }
```

allocate skbs for the newly received packets

GRO processing

```
          tcp_gro_receive() {
            skb_gro_receive();
          }
        }
      }
    }
    kmem_cache_free() {
      ___cache_free();
    }
  }

  [ .. repeat ..]

  e1000_alloc_rx_buffers [e1000]() {
    netdev_alloc_frag() {
      __alloc_page_frag();
    }
    _raw_spin_lock_irqsave();
    _raw_spin_unlock_irqrestore();

  [ .. repeat ..]
  }
}
}
```

(repeat for each packet)

allocate new pages for packet buffers

```
napi_gro_flush() {
  napi_gro_complete() {
    inet_gro_complete() {
      tcp4_gro_complete() {
        tcp_gro_complete();
      }
    }
    netif_receive_skb_internal() {
      __netif_receive_skb() {
        __netif_receive_skb_core() {
          ip_rcv() {
            nf_hook_slow() {
              nf_iterate() {
                ipv4_conntrack_defrag [nf_defrag_ipv4]();
                ipv4_conntrack_in [nf_conntrack_ipv4]() {
                  nf_conntrack_in [nf_conntrack]() {
                    ipv4_get_l4proto [nf_conntrack_ipv4]();
                    __nf_ct_l4proto_find [nf_conntrack]();
                    tcp_error [nf_conntrack]() {
                      nf_ip_checksum();
                    }
                    nf_ct_get_tuple [nf_conntrack]() {
                      ipv4_pkt_to_tuple [nf_conntrack_ipv4]();
                      tcp_pkt_to_tuple [nf_conntrack]();
                    }
                    hash_conntrack_raw [nf_conntrack]();
```

process IP header

Iptables raw/conntrack

```
        __nf_conntrack_find_get [nf_conntrack]();
        tcp_get_timeouts [nf_conntrack]();
        tcp_packet [nf_conntrack]() {
          _raw_spin_lock_bh();
          nf_ct_seq_offset [nf_conntrack]();
          _raw_spin_unlock_bh() {
            __local_bh_enable_ip();
          }
          __nf_ct_refresh_acct [nf_conntrack]();
        }
      }
    }
  }
}
ip_rcv_finish() {
  tcp_v4_early_demux() {
    __inet_lookup_established() {
      inet_ehashfn();
    }
    ipv4_dst_check();
  }
  ip_local_deliver() {
    nf_hook_slow() {
      nf_iterate() {
        iptable_filter_hook [iptable_filter]() {
          ipt_do_table [ip_tables]() {
```

(more conntrack)

routing decisions

Iptables INPUT chain

```
                          tcp_mt [xt_tcpudp]();
                          __local_bh_enable_ip();
                        }
                      }
                      ipv4_helper [nf_conntrack_ipv4]();
                      ipv4_confirm [nf_conntrack_ipv4]() {
                        nf_ct_deliver_cached_events [nf_conntrack]();
                      }
                    }
                  }
                  ip_local_deliver_finish() {
                    raw_local_deliver();
                    tcp_v4_rcv() {
                      [ .. ]
                    }
                  }
                }
              }
            }
          }
        }
      }
    }
  }
  __kfree_skb_flush();
}
```

l4 protocol handler

Iptables is not slow.

It's just executed **too late** in the stack.

Userspace Offload

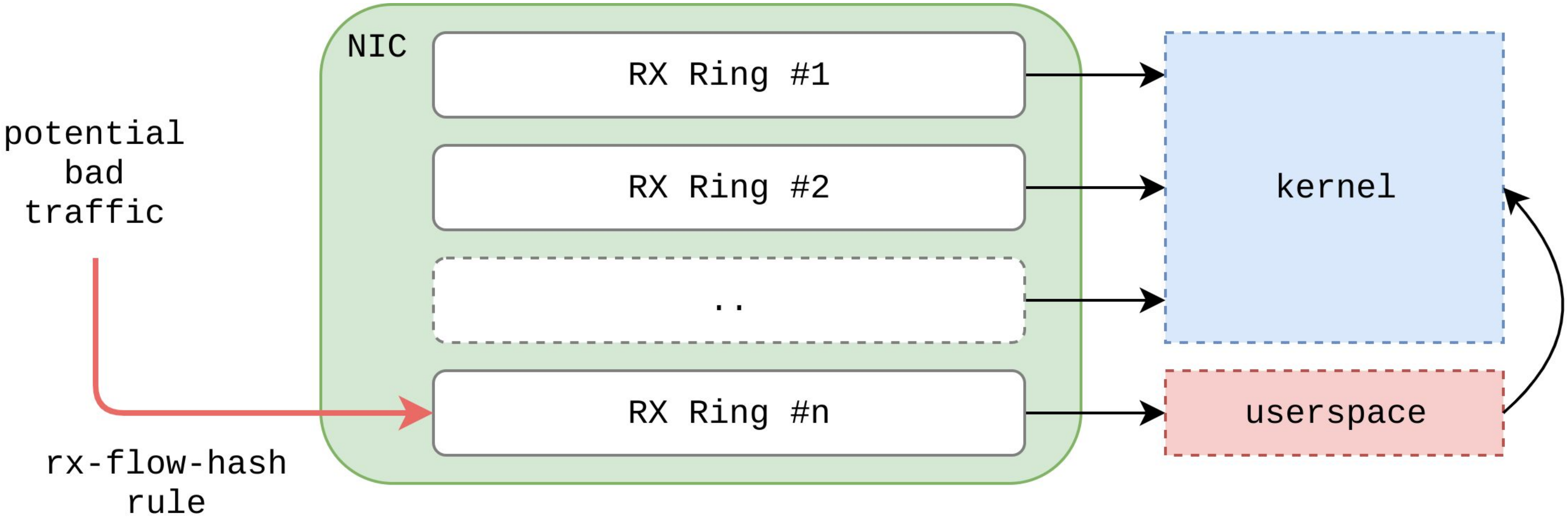
Kernel Bypass 101

- NIC RX/TX rings are mapped in and managed by userspace
- The network stack is completely bypassed

Kernel Bypass for packet filtering

- No sk_buff allocation
  - NIC uses circular preallocated buffer
- No kernel processing overhead
- Selectively offload traffic with flow-steering rules
- Inspect raw packets and
  - Reinject the legit ones
  - Drop the malicious one: no action required

Kernel Bypass for packet filtering



Netmap, EF_VI
PF_RING, DPDK

..

Kernel Bypass for packet filtering

```
while(1) {
  // poll RX ring, return a packet if available
  u_char *pkt = get_packet();

  for (int i = 0, i < rules_n; i++)
    if (run_bpf(pkt, rule[i]) == DROP)
      continue;

  reinject_packet(pkt)
}
```

An order of magnitude faster than Iptables.

6-8 Mpps on a **single core**

Kernel Bypass for packet filtering - disadvantages

- Legit traffic has to be reinjected (can be expensive)
- One or more cores have to be reserved
- Kernel space/user space context switches

XDP

Express Data Path

XDP

- New alternative to Iptables or Userspace offload
- Filter packets as soon as they are received
- Using an eBPF program
- Which returns an action (XDP_PASS, XDP_DROP, ..)
- (It's even possible to modify the content of a packet, push additional headers and retransmit it)

```
net_rx_action() {
  e1000_clean [e1000]() {
    e1000_clean_rx_irq [e1000]() {
      build_skb() {
        __build_skb() {
          kmem_cache_alloc();
        }
      }
      _raw_spin_lock_irqsave();
      _raw_spin_unlock_irqrestore();
      skb_put();
      eth_type_trans();
      napi_gro_receive() {
        skb_gro_reset_offset();
        dev_gro_receive() {
          inet_gro_receive() {
            tcp4_gro_receive() {
              __skb_gro_checksum_complete() {
                skb_checksum() {
                  __skb_checksum() {
                    csum_partial() {
                      do_csum();
                    }
                  }
                }
              }
```

BPF_PRG_RUN()

Just before allocating skbs

e1000 RX path with XDP

```
act = e1000_call_bpf(prog, page_address(p), length);

switch (act) {

/* .. */

case XDP_DROP:
default:
    /* re-use mapped page. keep buffer_info->dma
     * as-is, so that e1000_alloc_jumbo_rx_buffers
     * only needs to put it back into rx ring
     */
    total_rx_bytes += length;
    total_rx_packets++;
    goto next_desc;
}
```

XDP vs Userspace offload

- Same advantages as userspace offload:
  - No kernel processing
  - No sk_buff allocation/deallocation cost
  - No DMA map/unmap cost
- But well integrated in the Linux kernel:
  - eBPF
  - no need to reinject packets

eBPF

extended Berkeley
Packet Filter

eBPF

- Programmable in-kernel VM
  - Extension of classical BPF
  - Close to a real CPU
    - JIT on many arch (x86_64, ARM64, PPC64)
  - Safe memory access guarantees
  - Time bounded execution (no backward jumps)
  - Shared maps with userspace
- LLVM eBPF backend:
  - .c -> .o

Simple XDP example

```
SEC("xdp1")
int xdp_prog1(struct xdp_md *ctx)
{
	void *data     = (void *)(long)ctx->data;
	void *data_end = (void *)(long)ctx->data_end;

	struct ethhdr *eth = (struct ethhdr *)data;
	if (eth + 1 > (struct ethhdr *)data_end)
		return XDP_ABORTED;
	if (eth->h_proto != htons(ETH_P_IP))
		return XDP_PASS;

	struct iphdr *iph = (struct iphdr *)(eth + 1);
	if (iph + 1 > (struct iphdr *)data_end)
		return XDP_ABORTED;
	// if (iph->..
	//	return XDP_PASS;

	return XDP_DROP;
}
```

access packet buffer begin and end

access ethernet header

make sure we are not reading past the buffer

Simple XDP example

```
SEC("xdp1")
int xdp_prog1(struct xdp_md *ctx)
{
	void *data     = (void *)(long)ctx->data;
	void *data_end = (void *)(long)ctx->data_end;

	struct ethhdr *eth = (struct ethhdr *)data;
	if (eth + 1 > (struct ethhdr *)data_end)
		return XDP_ABORTED;
	if (eth->h_proto != htons(ETH_P_IP))
		return XDP_PASS;

	struct iphdr *iph = (struct iphdr *)(eth + 1);
	if (iph + 1 > (struct iphdr *)data_end)
		return XDP_ABORTED;
	// if (iph->..
	//	return XDP_PASS;

	return XDP_DROP;
}
```

check this is an IP packet

access IP header

make sure we are not reading past the buffer

XDP and maps

```
struct bpf_map_def SEC("maps") rxcnt = {
	.type = BPF_MAP_TYPE_PERCPU_ARRAY,
	.key_size = sizeof(unsigned int),
	.value_size = sizeof(long),
	.max_entries = 256,
};

SEC("xdp1")
int xdp_prog1(struct xdp_md *ctx)
{
	unsigned int key = 1;

// ..

	long *value = bpf_map_lookup_elem(&rxcnt, &key);
	if (value)
		*value += 1;

}
```

define a new map

get a ptr to the value indexed by "key"

update the value

Why not automatically generate XDP programs!

```
➜  p0f2ebpf git:(master) ./p0f2ebpf.py --ip 1.2.3.4 --port 1234
'4:64:0:*:mss*10,6:mss,sok,ts,nop,ws:df,id+:0'

static inline int match_p0f(struct xdp_md *ctx)
{
      void *data     = (void *)(long)ctx->data;
      void *data_end = (void *)(long)ctx->data_end;

      struct ethhdr *eth_hdr;
      struct iphdr  *ip_hdr;
      struct tcphdr *tcp_hdr;
      unsigned char *tcp_opts;

      eth_hdr = (struct ethhdr *)data;
      if (eth_hdr + 1 > (struct ethhdr *)data_end)
            return XDP_ABORTED;
      if_not (eth_hdr->h_proto == htons(ETH_P_IP))
            return XDP_PASS;
```

```
ip_hdr = (struct iphdr *)(eth_hdr + 1);
if (ip_hdr + 1 > (struct iphdr *)data_end)
      return XDP_ABORTED;
if_not (ip_hdr->version == 4)
      return XDP_PASS;
if_not (ip_hdr->daddr == htonl(0x1020304))
      return XDP_PASS;
if_not (ip_hdr->ttl <= 64)
      return XDP_PASS;
if_not (ip_hdr->ttl > 29)
      return XDP_PASS;
if_not (ip_hdr->ihl == 5)
      return XDP_PASS;
if_not ((ip_hdr->frag_off & IP_DF) != 0)
      return XDP_PASS;
if_not ((ip_hdr->frag_off & IP_MBZ) == 0)
      return XDP_PASS;

tcp_hdr = (struct tcphdr*)((unsigned char *)ip_hdr + ip_hdr->ihl * 4);
if (tcp_hdr + 1 > (struct tcphdr *)data_end)
      return XDP_ABORTED;
if_not (tcp_hdr->dest == htons(1234))
      return XDP_PASS;
if_not (tcp_hdr->doff == 10)
      return XDP_PASS;
if_not ((htons(ip_hdr->tot_len) - (ip_hdr->ihl * 4) - (tcp_hdr->doff * 4)) == 0)
      return XDP_PASS;
```

```
    tcp_opts = (unsigned char *)(tcp_hdr + 1);
    if (tcp_opts + (tcp_hdr->doff - 5) * 4 > (unsigned char *)data_end)
          return XDP_ABORTED;
    if_not (tcp_hdr->window == *(unsigned short *)(tcp_opts + 2) * 0xa)
          return XDP_PASS;
    if_not (*(unsigned char *)(tcp_opts + 19) == 6)
          return XDP_PASS;
    if_not (tcp_opts[0] == 2)
          return XDP_PASS;
    if_not (tcp_opts[4] == 4)
          return XDP_PASS;
    if_not (tcp_opts[6] == 8)
          return XDP_PASS;
    if_not (tcp_opts[16] == 1)
          return XDP_PASS;
    if_not (tcp_opts[17] == 3)
          return XDP_PASS;

    return XDP_DROP;
}
```

How to try it

- New technology, drivers are still being worked on
  - mlx4, mlx5, npf, qed, virtio_net, ixgbe (e1000)
- Generic XDP from Linux 4.12
- Compile and install a fresh kernel; cd samples/bpf/
  - Actual XDP programs:
    - xdp1_kern.c
    - xdp1_user.c
  - Helpers:
    - bpf_helpers.h
    - bpf_load.{c,h}
    - libbpf.h

Thanks!

Questions?

gilberto@cloudflare.com