NA DORAZ III

kam až sahají meze síťového subsystému Linuxového jádra

Jan Kučera
jan.kucera@cesnet.cz

Dominik Tran
tran@cesnet.cz

Jan Viktorin
viktorin@highpps.net

NA DORAZ III

využití eBPF/XDP pro optimalizaci výkonu síťového subsystému Linuxového jádra

Jan Kučera
jan.kucera@cesnet.cz

Dominik Tran
tran@cesnet.cz

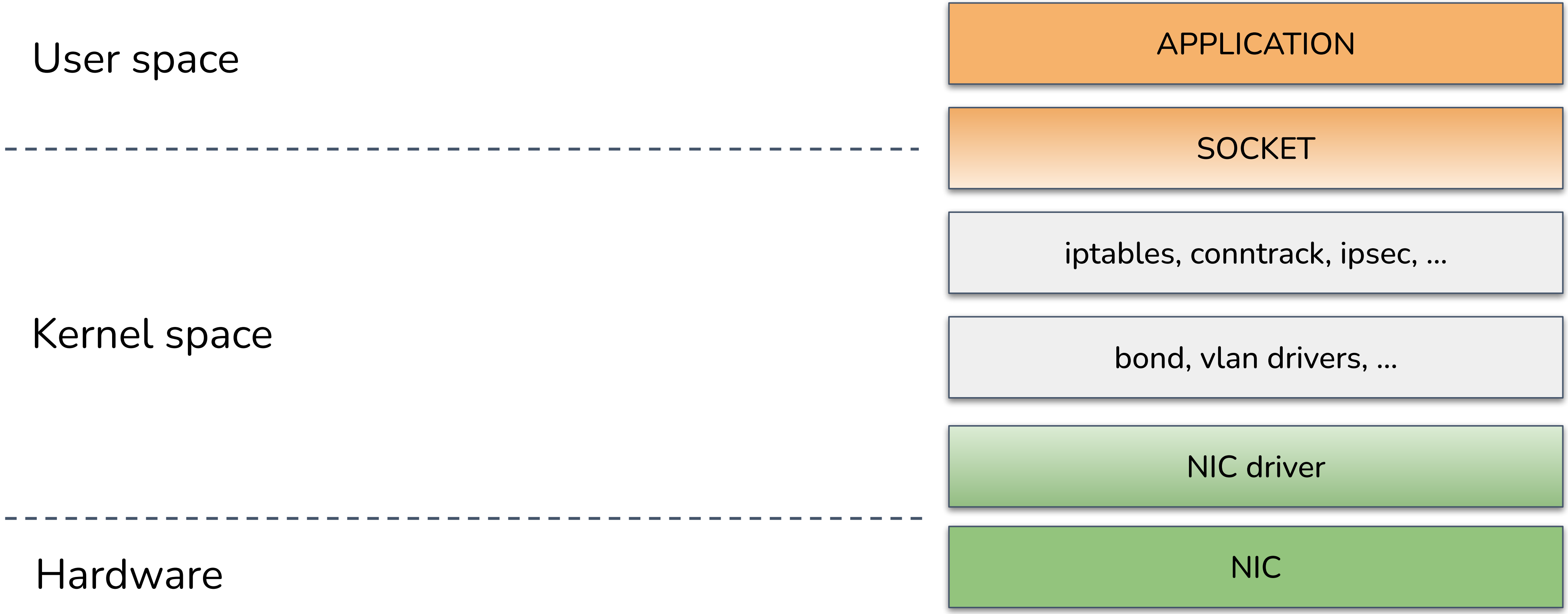
Jan Viktorin
viktorin@highpps.net

Ochrana serveru před útokem

- Ochrana koncového systému před DDoS útokem
- Webový server, HTTP služba na TCP portech 80, 443
- Metoda “ustát toho co nejvíce” s existujícími HW zdroji
- Z Linuxového systému vymáčknout maximální výkon
- Nejlevnější způsob ochrany = vhodná konfigurace systému

- Podrobněji k obecné ochraně viz dřívější přednáška: https://indico.csnog.eu/event/7/contributions/83/

Síťový subsystém Linuxového jádra

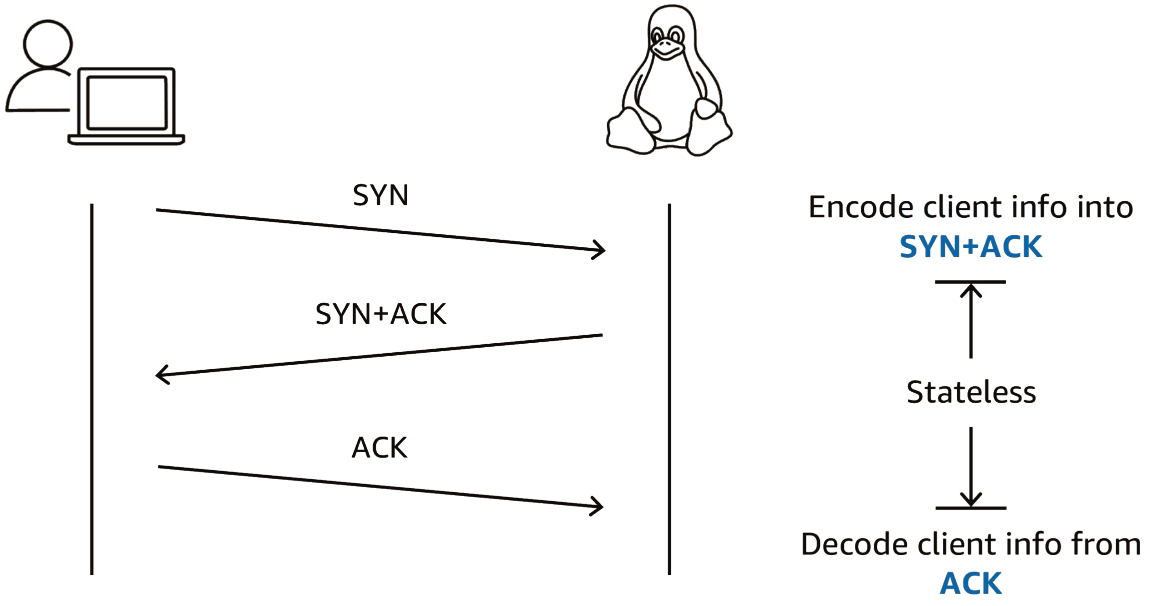


Provoz neidentifikovatelný zdrojem

- Nelze jednoduše rozlišit zda zdrojová IP adresa odpovídá legitimnímu zdroji nebo je podvržená
  - Nelze použít IP blacklisting, rate limiting per IP

- Typicky **úvodní SYN paket**
  - Ustavení TCP spojení (handshake)
  - Nelze na něj neodpovědět
  - TCP SYN Flood útok

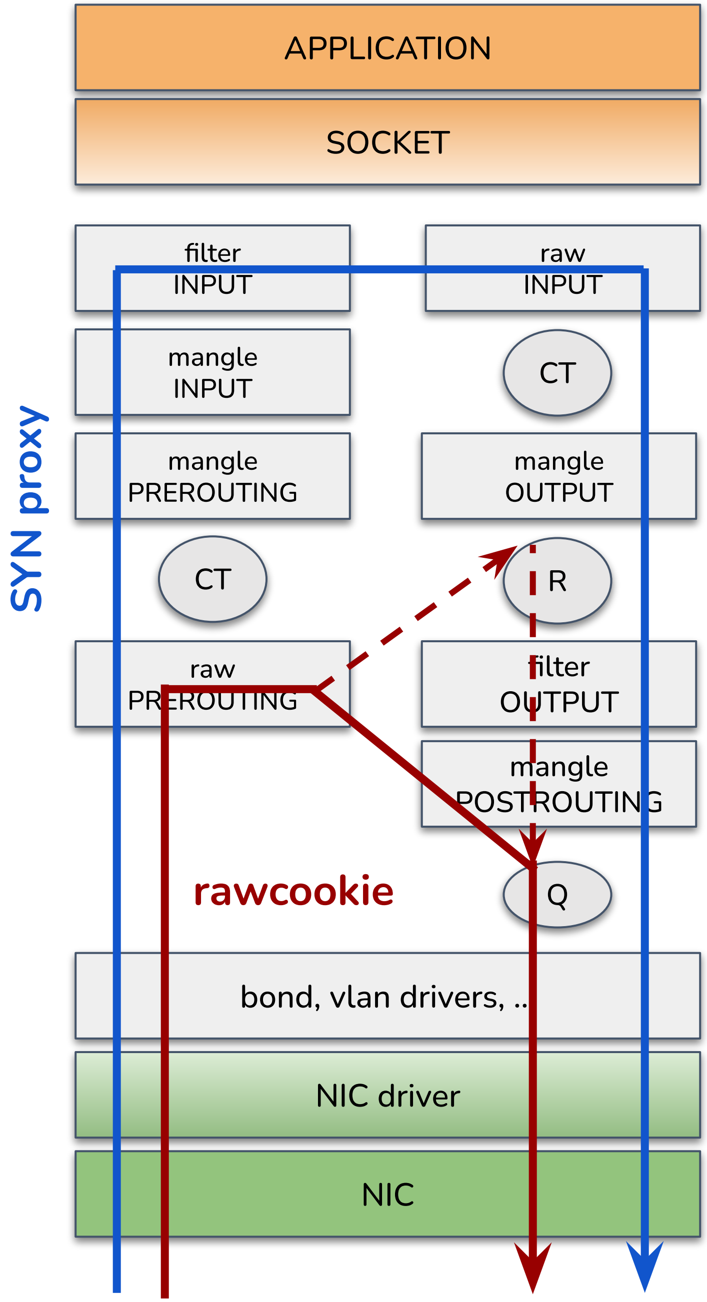
- Řešením **SYN cookies**
  - Přenesení stavu do hlavičky paketu
  - Server nemusí udržovat stav
  - Podpora v Linux Kernel



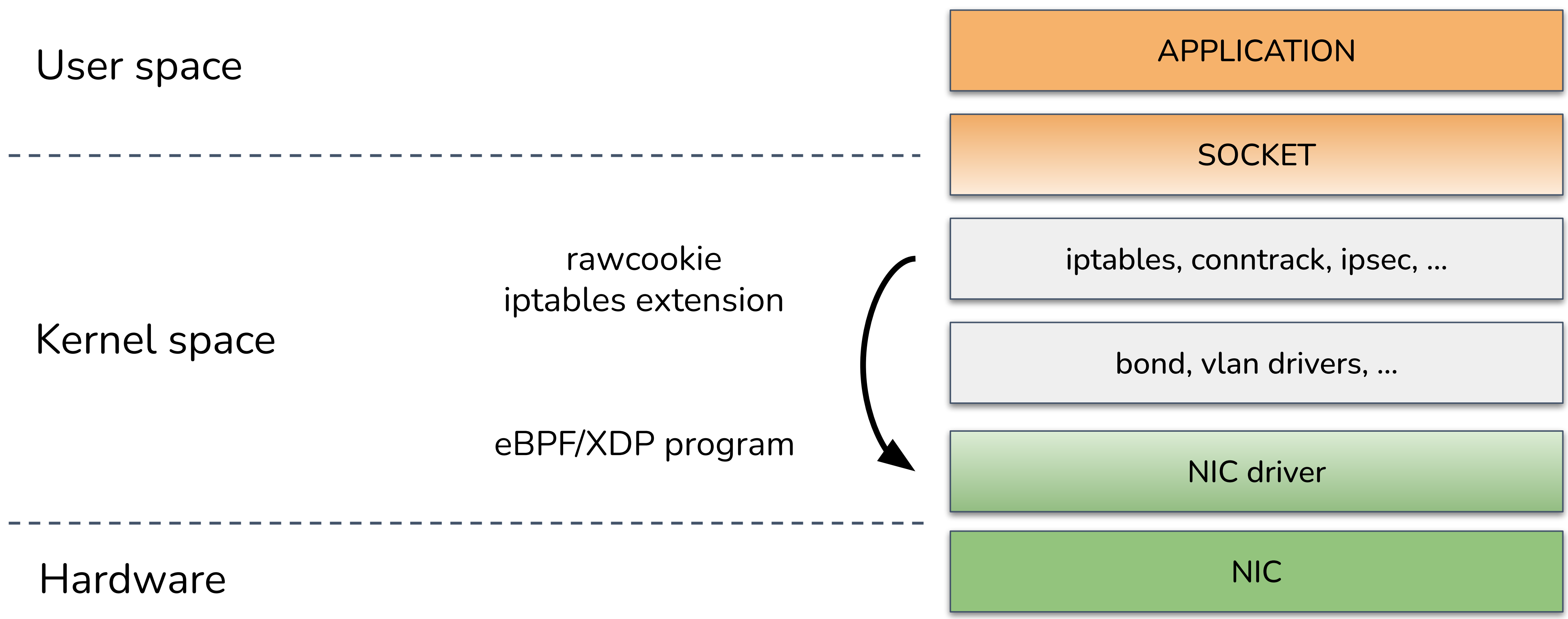
Převzato ze "SYN Proxy at Scale with BPF": https://lpc.events/event/17/contributions/1645/.

Modul rawcookie

- Nativní SYN cookies a SYN proxy nepoužitelné
- Rozšiřující modul pro iptables
  - Posouvá reakci na SYN paket a odpověď SYN+ACK do nižší vrstvy iptables (raw table)
  - Obchází connection tracking (conntrack)
  - Obchází routing, SYN+ACK odešle na MAC adresu směrovače odkud přišel původní SYN
- Přibližně 2x rychlejší než SYN proxy
- Řádově vyšší jednotky Mp/s
- https://github.com/netx-as/xt_RAWCOOKIE

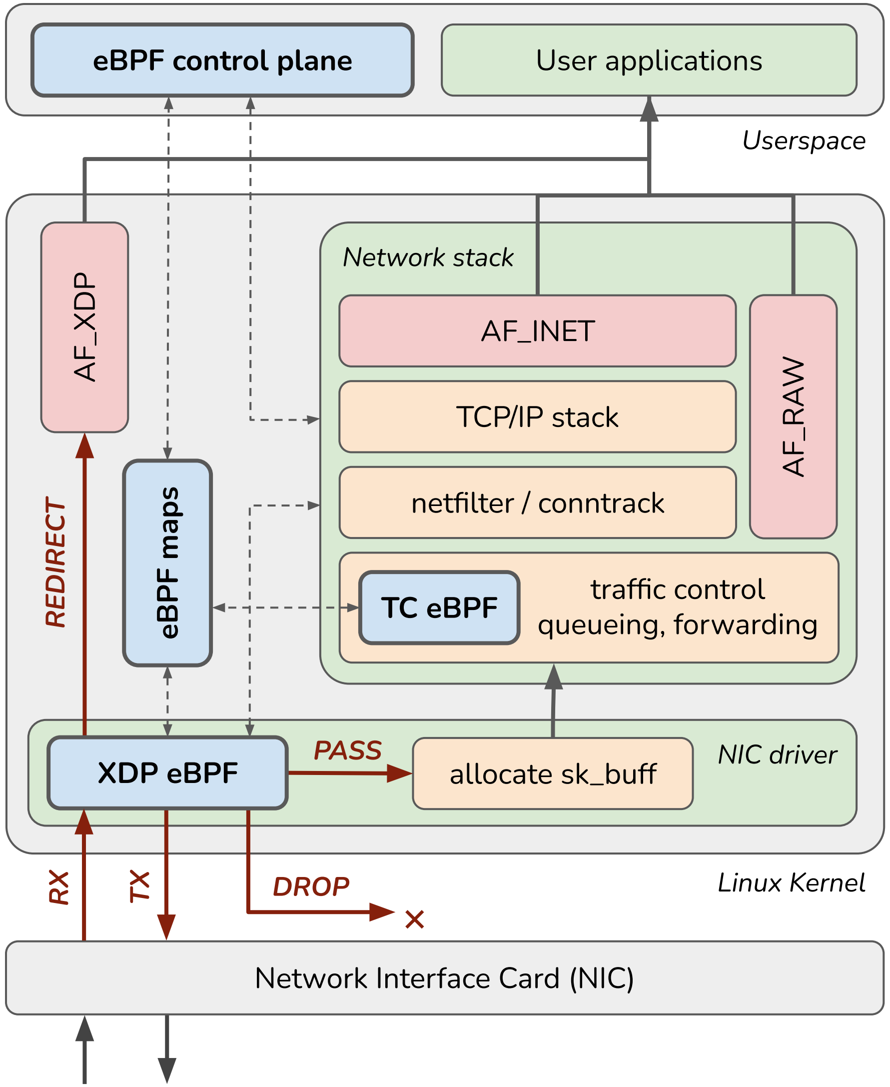


Využití eBPF/XDP pro akceleraci



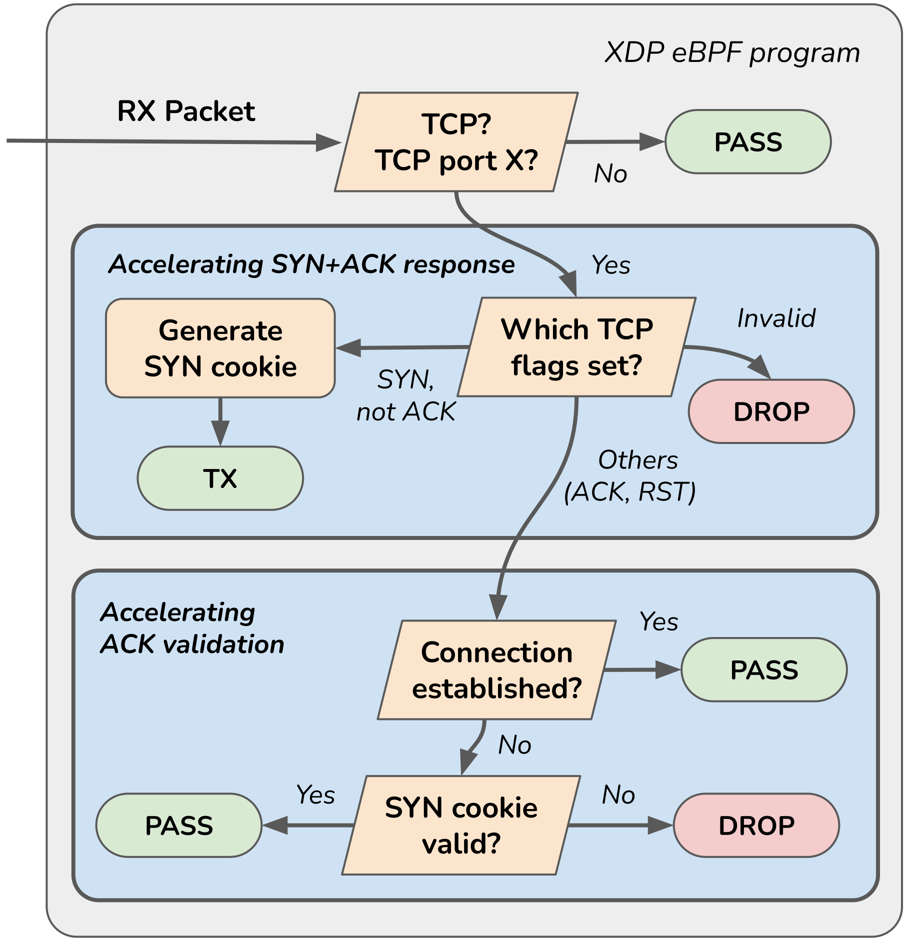
eBPF/XDP akcelerace

- Extended Berkeley Packet Filter (eBPF)
  - Vložení programu na vybraná místa v Kernelu
  - XDP (eXpress Data Path) hook – nejnižší vrstva
  - TC (Traffic Control) hook – po alokaci sk_buff
- Užitečné eBPF helpers
  - *bpf_tcp_gen/check_syncookie()* – v5.4
  - *bpf_skb/xdp_ct_lookup()* – v5.18
  - *bpf_tcp_raw_gen/check_syncookie_*()* – v6.0
- Podpora VLAN stripping
  - *bpf_xdp_metadata_rx_vlan_tag()* – v6.8



Program xdpcookie

- XDP implementace SYN cookie
  - Akcelerace SYN+ACK odpovědi
  - Akcelerace ACK validace (ACK flood)
- Vychází z `/tools/testing/selftests/bpf`
- Rozšířená funkcionalita
  - Podpora VLAN (stripping, filtrace)
  - Zapnutí / vypnutí výpočtu L3/L4 checksums
  - Fragmentace bufferu “xdp.frags” (MTU > 4kB)
  - Instalovatelné jako DEB balíček
- https://github.com/highpps/xdpcookie



Příklad použití xdpcookie

- Konfigurace SYN cookies proxy
- Zavedení / odstranění programu

```
# xdpcookie --attach --iface eth0 --vlan 85 --port 80 --port 443
    --mss4 1460 --wscale 7 --ttl 64 --checkack --checksum --calcsum
# xdpcookie --detach --iface eth0
```

- Konfigurace TX checksum offload

```
# ethtool --set-priv-flags eth0 tx_xdp_hw_checksum on
```

- Výpis statistik

```
# xdpcookie --iface eth0
```

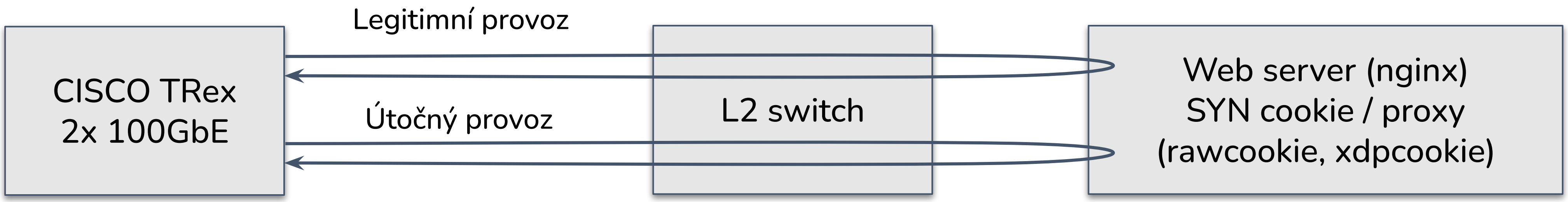
```
SYN-ACK responses generated: 5014945704
CPU00 responses: 121486873
CPU01 responses: 132257478
...
```

Sestava použitá pro měření

- 2x Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz
- 10 jader, hyperthreading (20 vláken)
- Max performance profile, vypnuté šetřící režimy (C states)
- Systém Debian 12.8 + Debian Backports repozitář
- Kernel 6.10.11+bpo-amd64 #1 SMP x86_64 GNU/Linux
- 2x 100GbE Mellanox ConnectX-5 Ex, mlx5_core 24.07-0.6.1.0
- Firmware 16.35.3006 (DEL0000000004)
- Agregace portů (bond driver)

Metodika měření

- Generátor provozu – CISCO TRex – 2x 100GbE

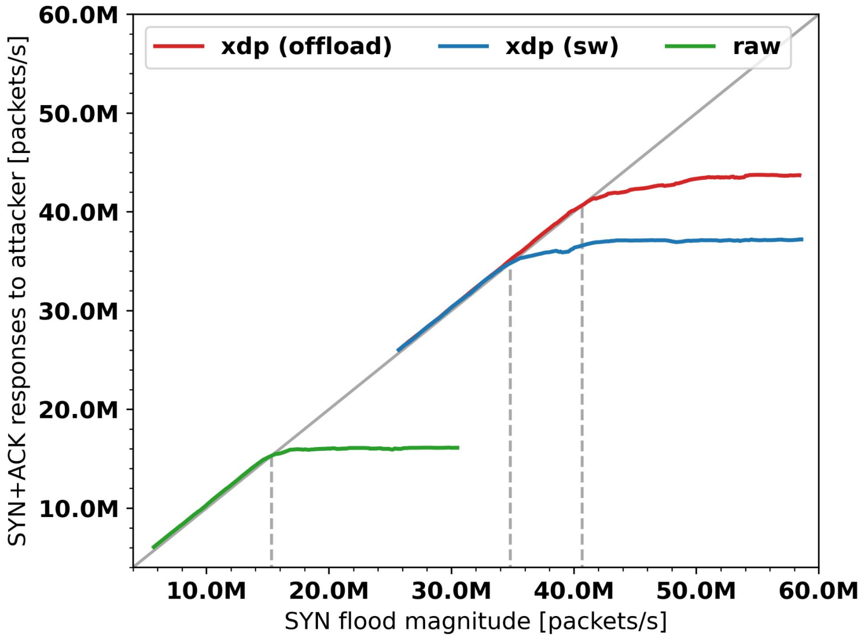


- Port A: legitimní provoz
  - ~ 10k TCP spojení / s
  - 16k unikátních adres
- Port B: útočný provoz – SYN Flood
  - 5M – 60M paketů / s
  - 64k unikátních adres

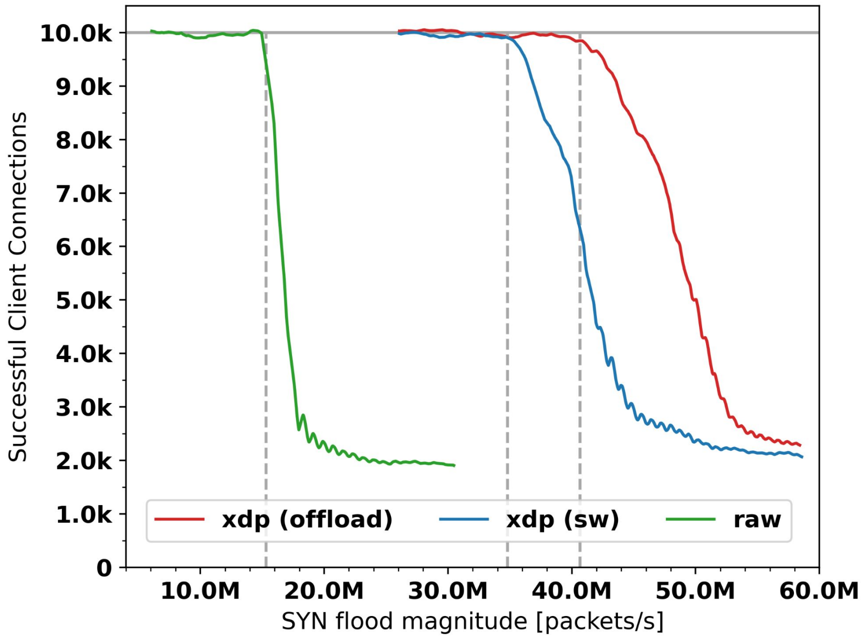
- Metriky:
  - Packet rate SYN+ACK odpovědí na útočníka / úspěšná spojení klienta
  - Počet retransmisí / latence odpovědí z pohledu klienta

Výsledky měření (1)

- Packet rate SYN+ACK odpovědí



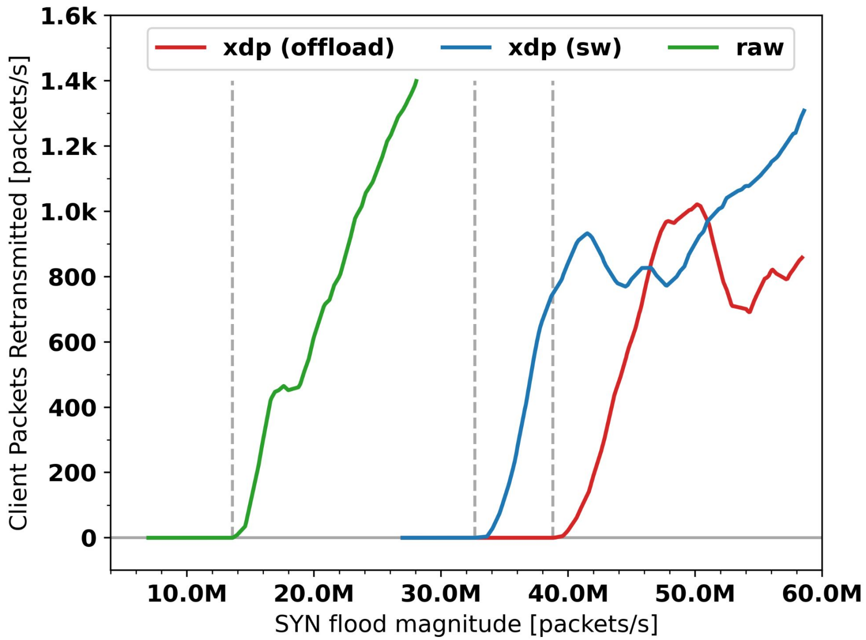
- Úspěšná klientská spojení



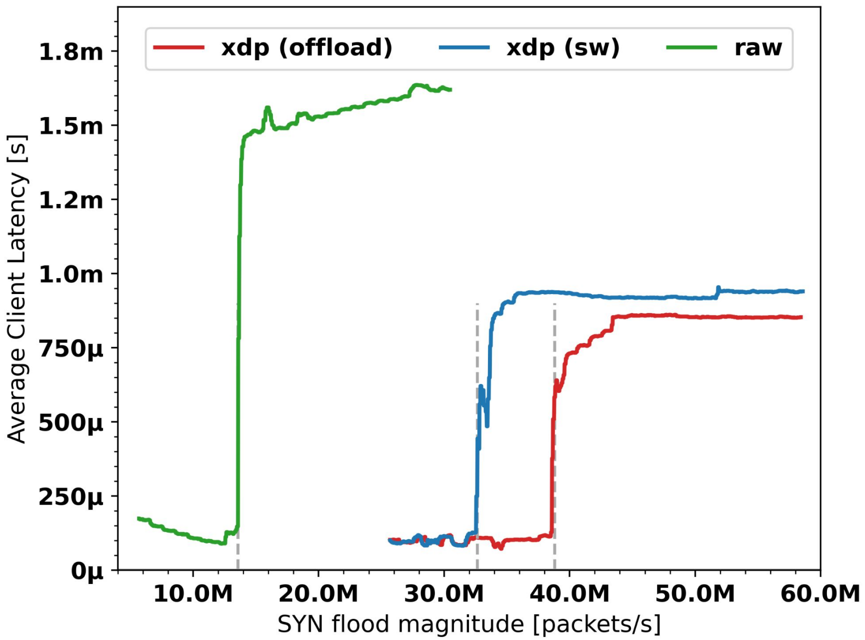
- rawcookie: **15,2 Mp/s**
- xdpcookie: **34,7 Mp/s** (sw checksum) → **40,6 Mp/s** (checksum offload)

Výsledky měření (2)

- Počet klientských retransmisí



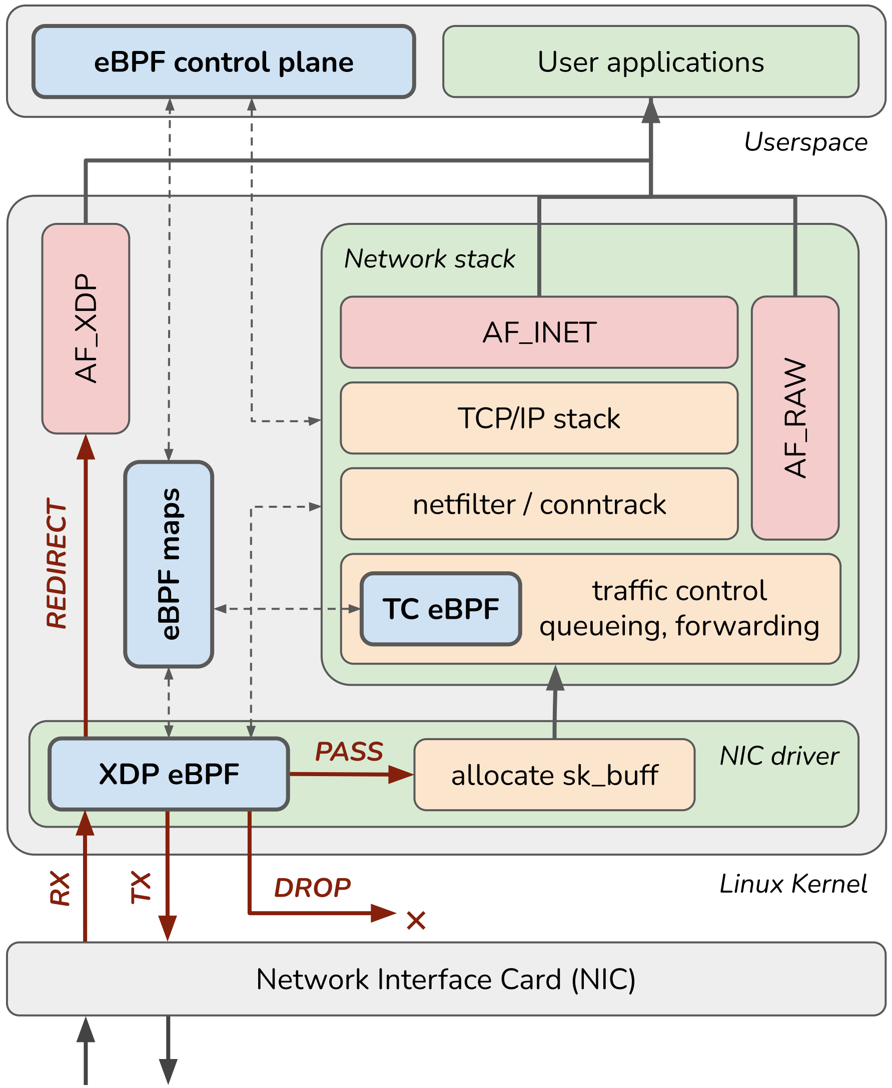
- Latence SYN+ACK odpovědí



- rawcookie: **13,5 Mp/s**
- xdpcookie: **32,6 Mp/s** (sw checksum) → **38,8 Mp/s** (checksum offload)

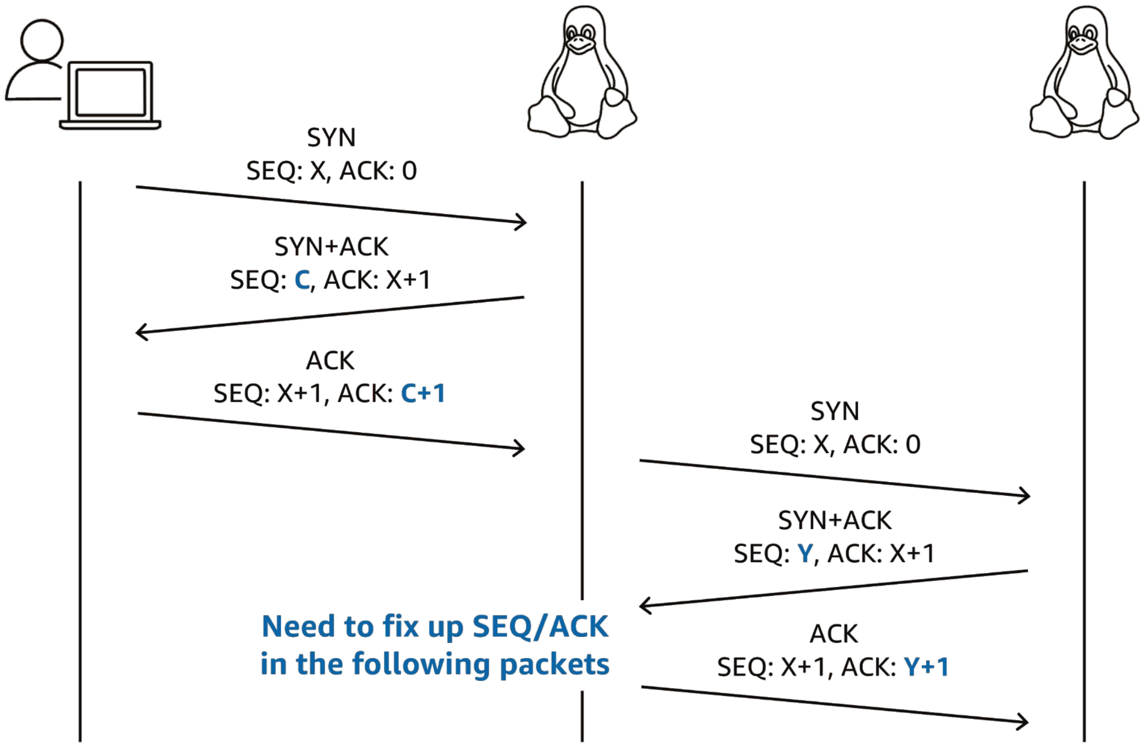
Shrnutí

- XDP akcelerace SYN cookies
  - Obchází connection tracking
  - Obchází routing (swap IP a MAC adres)
  - **Obchází alokaci sk_buff**
  - VLAN stripping, MTU >4kB
- Očekávat lze nižší desítky Mp/s
- Nárůst výkonu oproti RAW cookie
  - **2,3-2,4x** (SW checksum)
  - **2,7-2,9x** (checksum offload)
- **Jak dál?**



Jak dál: předřazení SYN cookie (NIC, extra stroj)

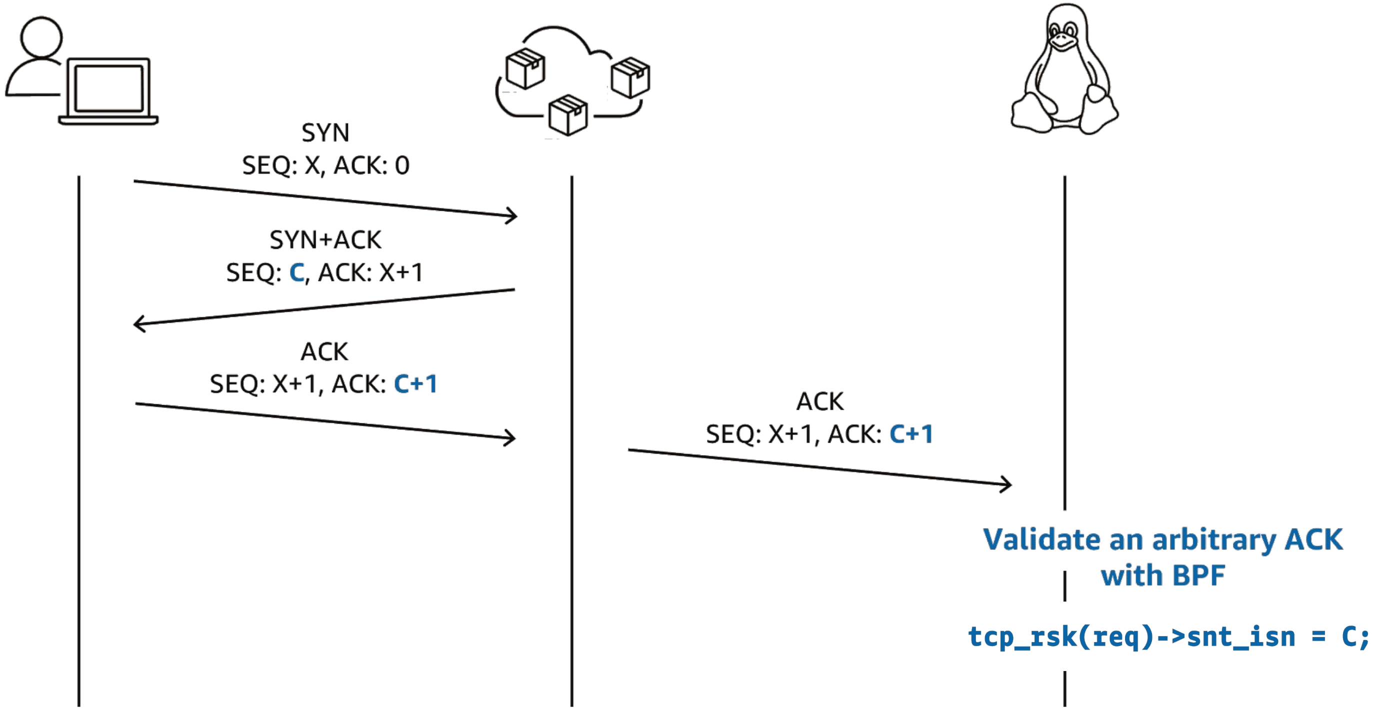
- Překlad sekvenčních čísel, proxy musí uchovávat stav



Převzato ze "SYN Proxy at Scale with BPF": https://lpc.events/event/17/contributions/1645/.

Jak dál: vnucení sekvenčního čísla (ISN)

- Patch *bpf_sk_assign_tcp_reqsk()*
  https://lore.kernel.org/all/20240115205514.68364-1-kuniyu@amazon.com/

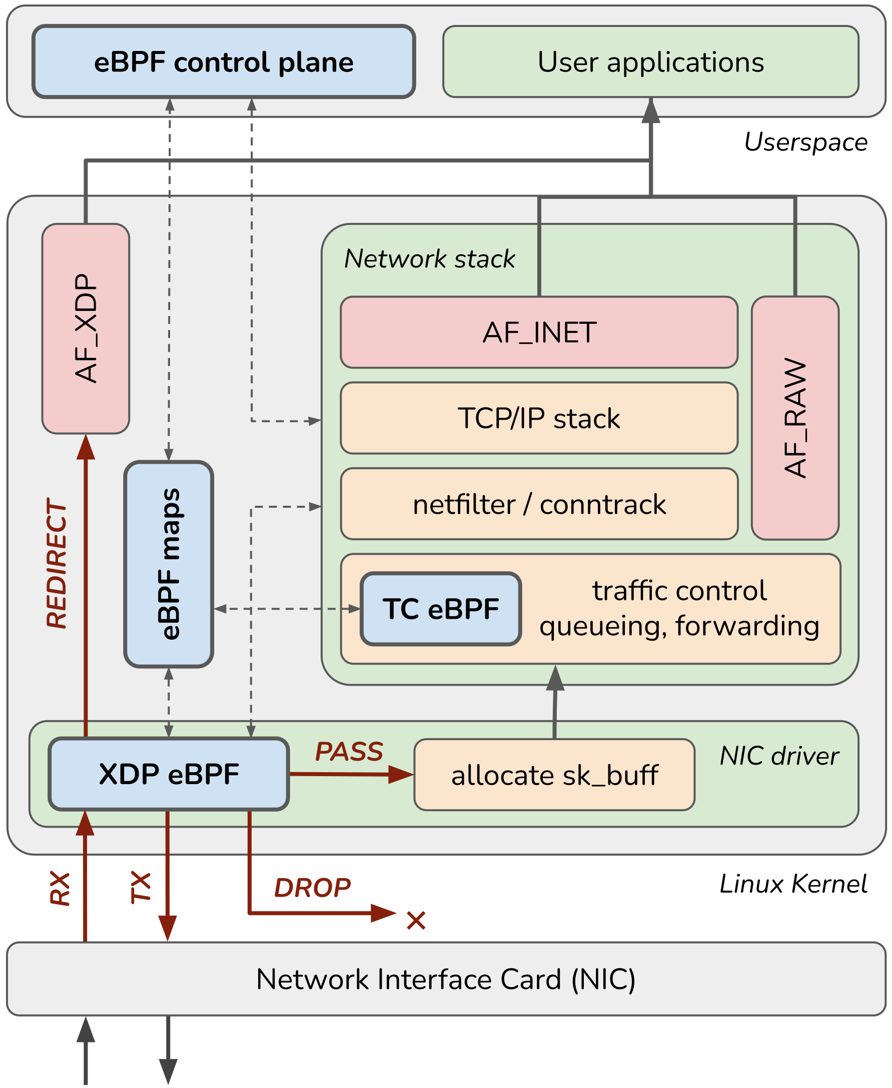


Převzato ze "SYN Proxy at Scale with BPF": https://lpc.events/event/17/contributions/1645/.

Pokračování příště ....

Shrnutí

- XDP akcelerace SYN cookies
  - Obchází connection tracking
  - Obchází routing (swap IP a MAC adres)
  - **Obchází alokaci sk_buff**
  - VLAN stripping, MTU >4kB
- Očekávat lze nižší desítky Mp/s
- Nárůst výkonu oproti RAW cookie
  - **2,3-2,4x** (SW checksum)
  - **2,7-2,9x** (checksum offload)
- **Jak dál zase příště ...**



Reference

- Modul rawcookie: https://github.com/netx-as/xt_RAWCOOKIE
- Program xdpcookie: https://github.com/highpps/xdpcookie
- Helper functions bpf_tcp_raw_gen_syncookie_ipv4/6():
  https://docs.ebpf.io/linux/helper-function/bpf_tcp_raw_gen_syncookie_ipv4/
- KFuncs bpf_xdp_metadata_rx_vlan_tag():
  https://docs.ebpf.io/linux/kfuncs/bpf_xdp_metadata_rx_vlan_tag/
- Accelerating synproxy with XDP:
  https://netdevconf.info/0x15/session.html?Accelerating-synproxy-with-XDP
- SYN Proxy at Scale with BPF: https://lpc.events/event/17/contributions/1645/