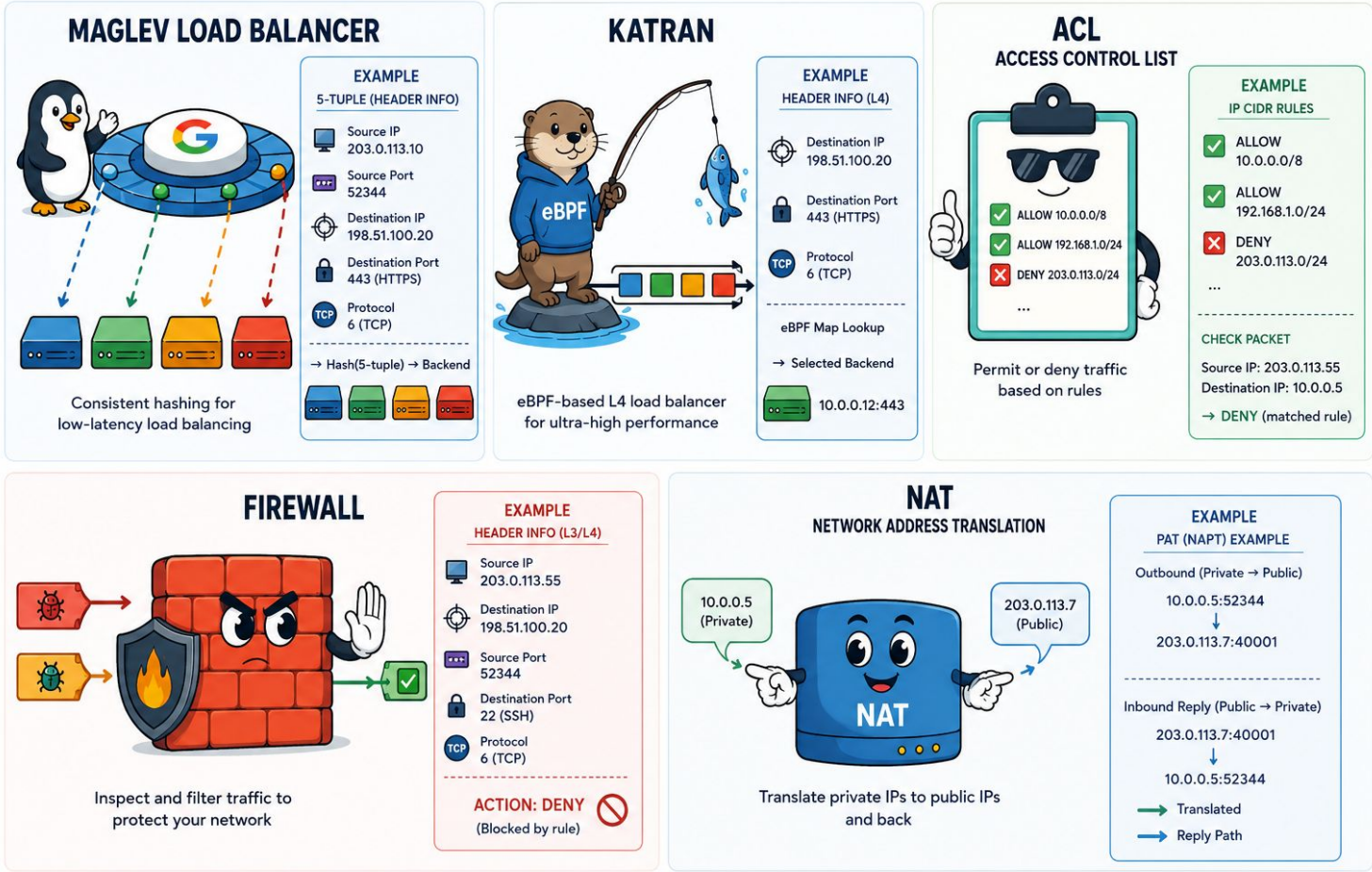
From Protocol Diversity to Hardware Offloading: Rethinking Packet Parsing Costs

COSENERS 26

Filippo Carloni, Gregorio Procissi, Giuseppe Bianchi, Aurojit Panda, Gianni Antichi

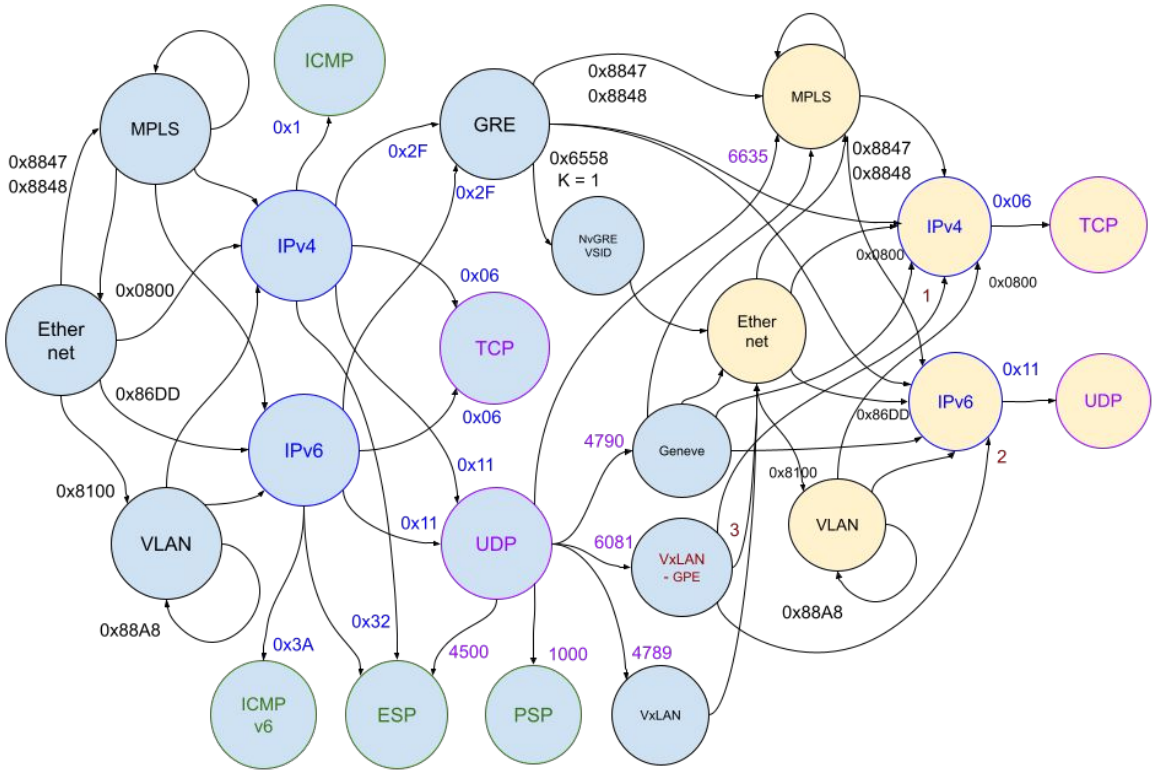
Networking Functions Need to Know Packet Fields



This Means That NFs Need a Packet Parser

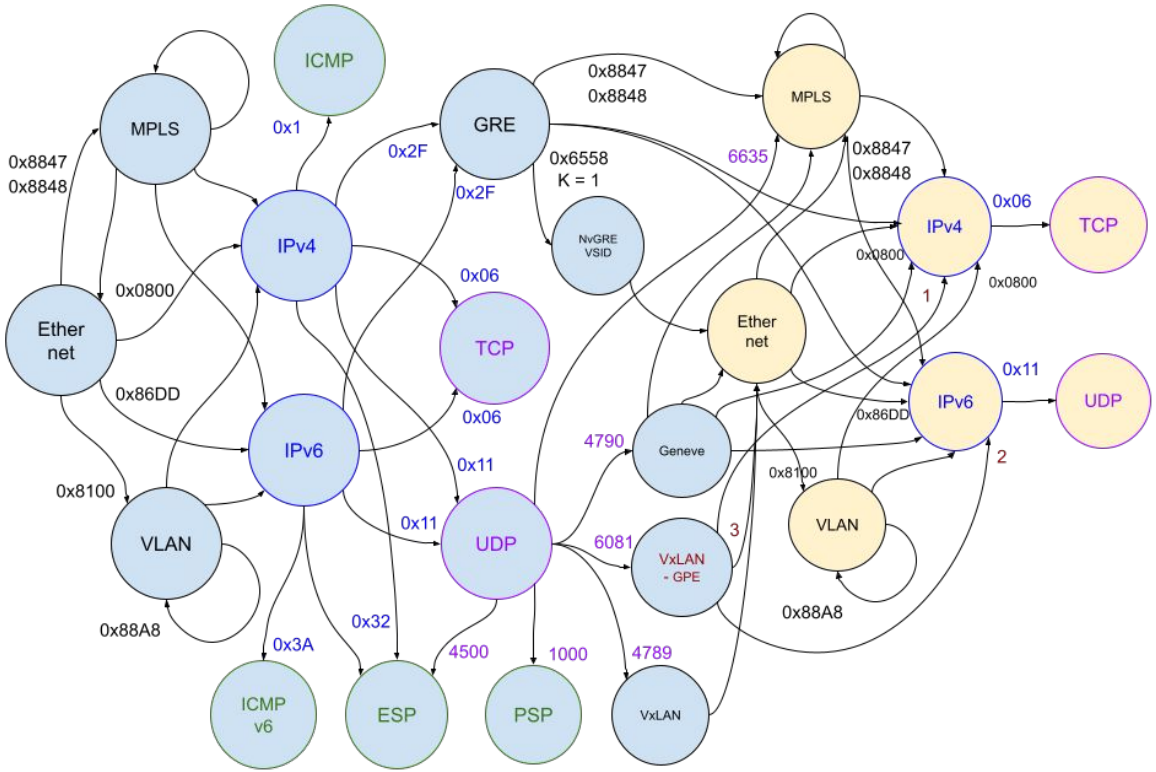
```
parse(pkt):
    off = 14                              # Ethernet
    et  = pkt.ethertype
    if   et == 0x0800:                    # ── IPv4 ──
        proto = pkt.ipv4.proto
        off  += pkt.ipv4.ihl * 4
    elif et == 0x86DD:                    # ── IPv6 ──
        proto = pkt.ipv6.next_hdr
        off  += 40
    else:
        return UNSUPPORTED
    if   proto == 6:   l4 = read_tcp(pkt, off)   # ── TCP ──
    elif proto == 17:  l4 = read_udp(pkt, off)   # ── UDP ──
    else:              return PARTIAL
    return (et, proto, l4)

  return go(f, seed, [])
}
```

Protocol Alternatives Can Rapidly Increase …



https://networking-docs.nvidia.com/doca/sdk/doca-target-architecture

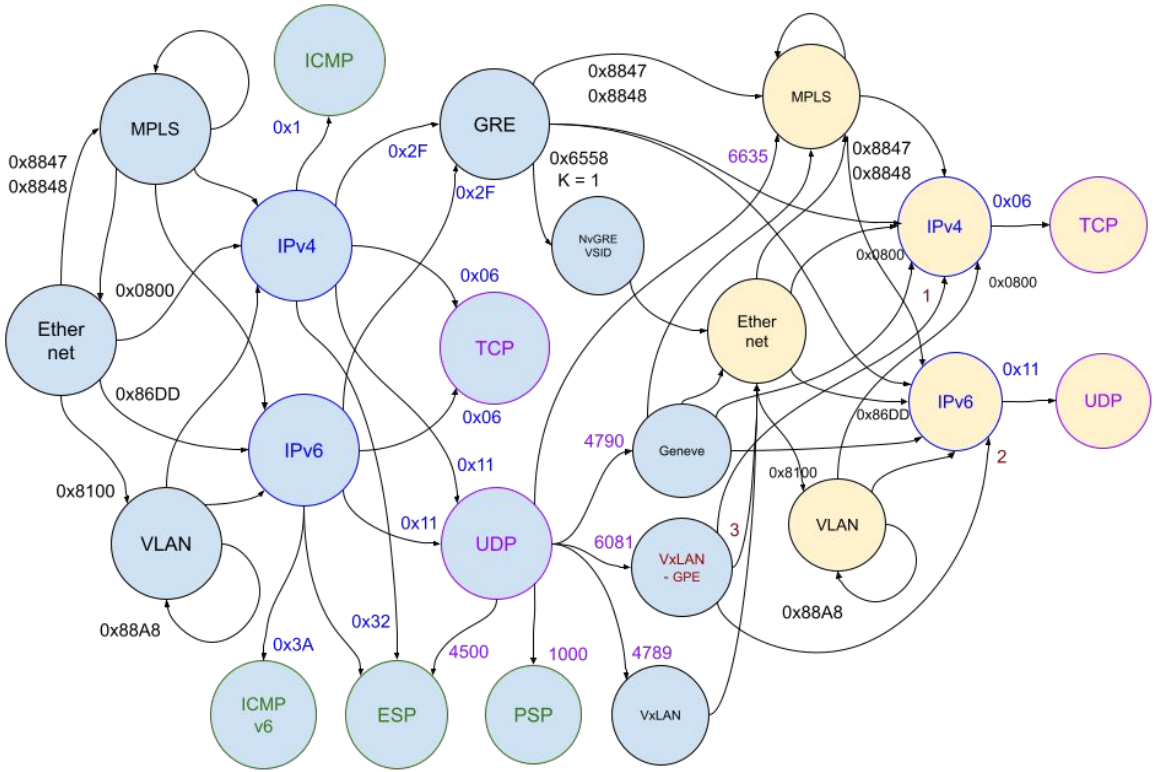
Protocol Alternatives Can Rapidly Increase …



Many Protocols

https://networking-docs.nvidia.com/doca/sdk/doca-target-architecture

Protocol Alternatives Can Rapidly Increase …

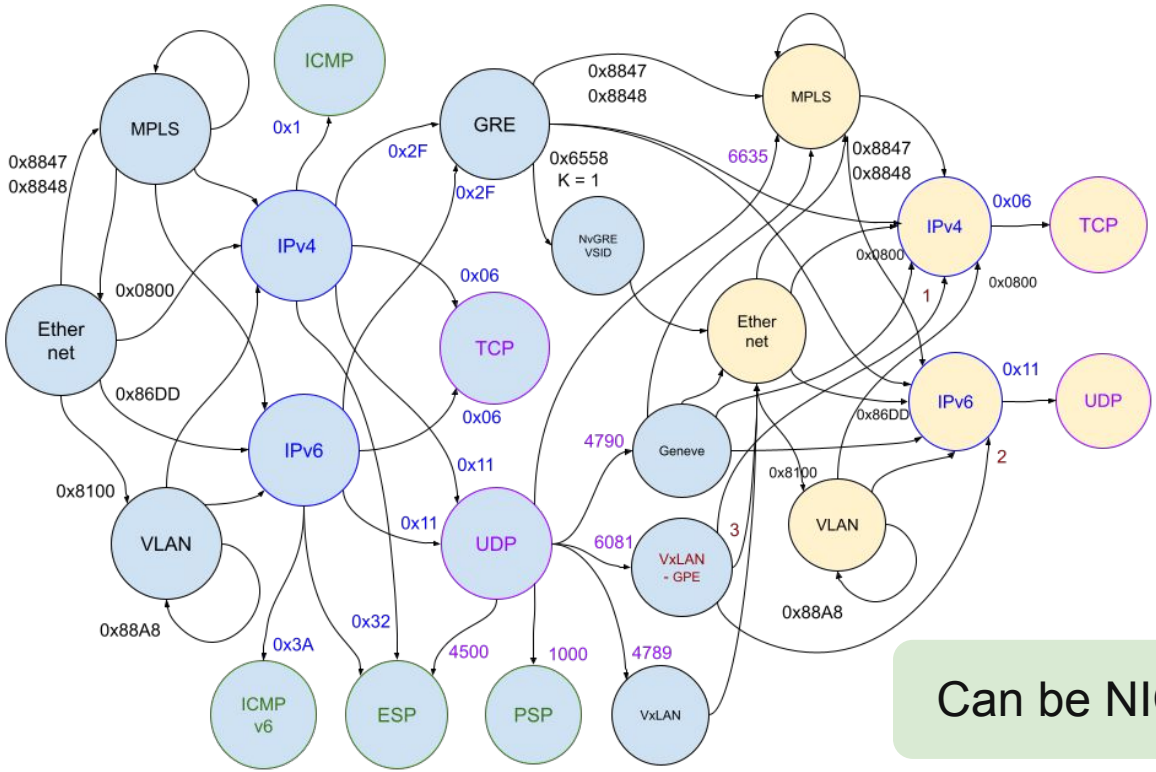


Many Protocols

Complex SW Parsing

https://networking-docs.nvidia.com/doca/sdk/doca-target-architecture

Protocol Alternatives Can Rapidly Increase …



Many Protocols

Complex SW Parsing

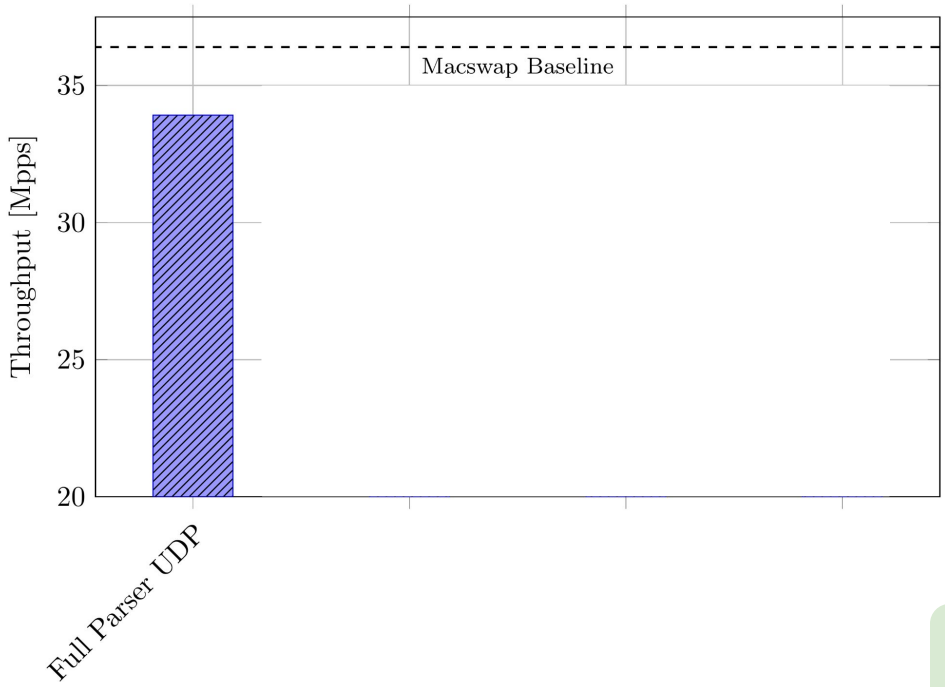
Can be NIC standard features helpful?

https://networking-docs.nvidia.com/doca/sdk/doca-target-architecture

Do SW NFs Need Help? The Load Balancer Example

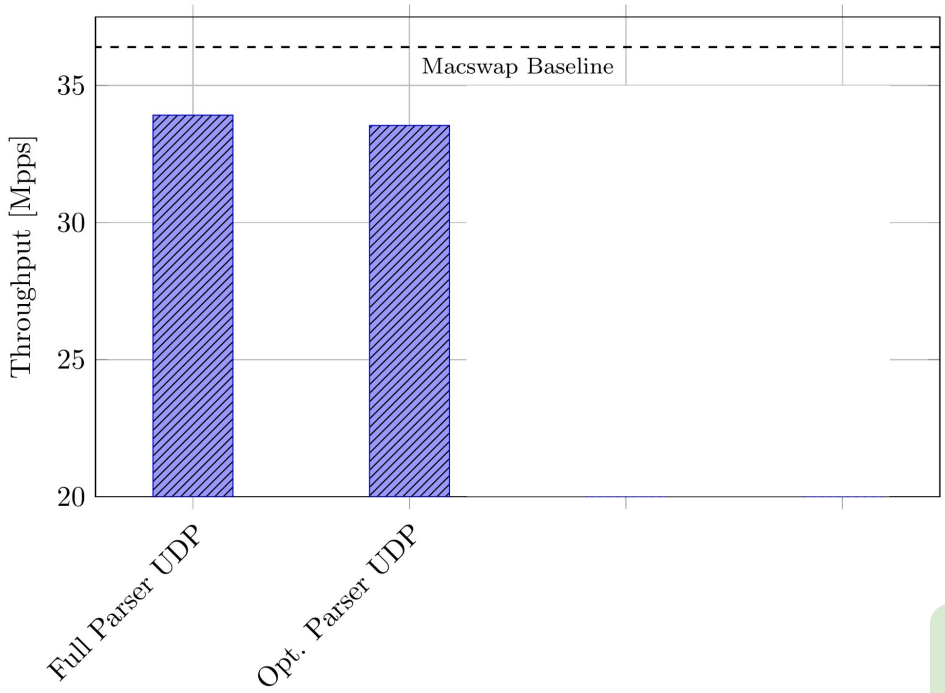
Setting a maximum performance bar

Do SW NFs Need Help? The Load Balancer Example



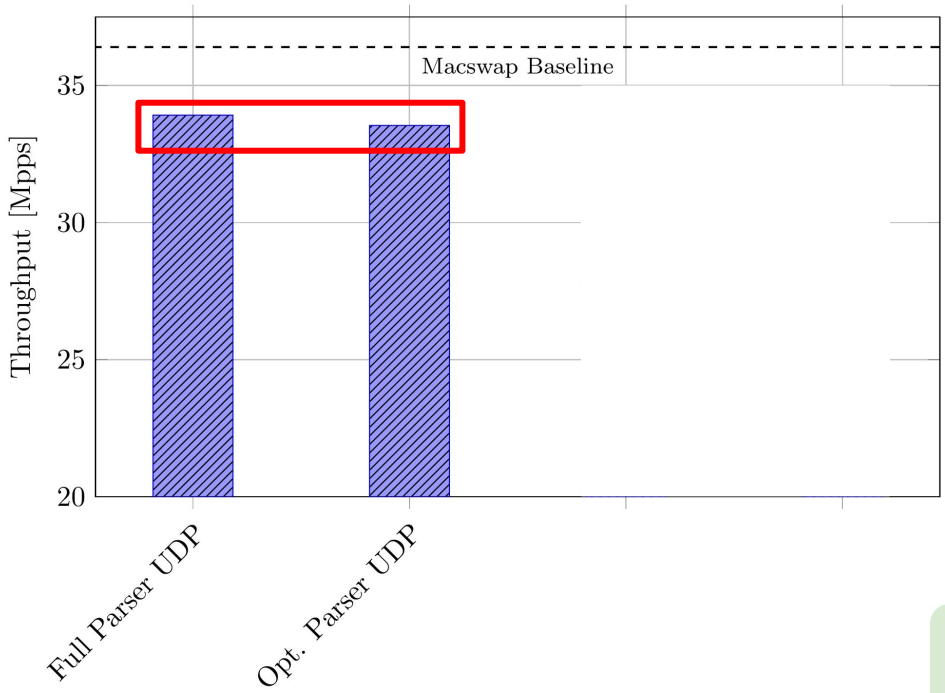
Parsing multiple protocols even if not actually used

Do SW NFs Need Help? The Load Balancer Example



Parsing only needed protocols

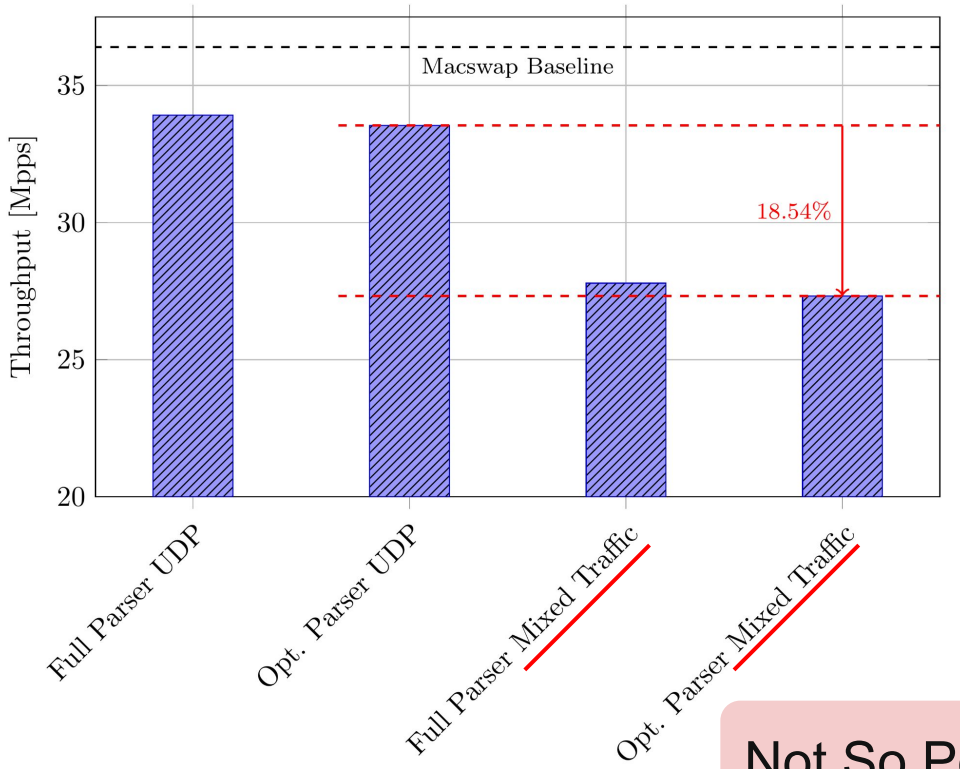
Do SW NFs Need Help? The Load Balancer Example



Thank you HW/SW optimizations!

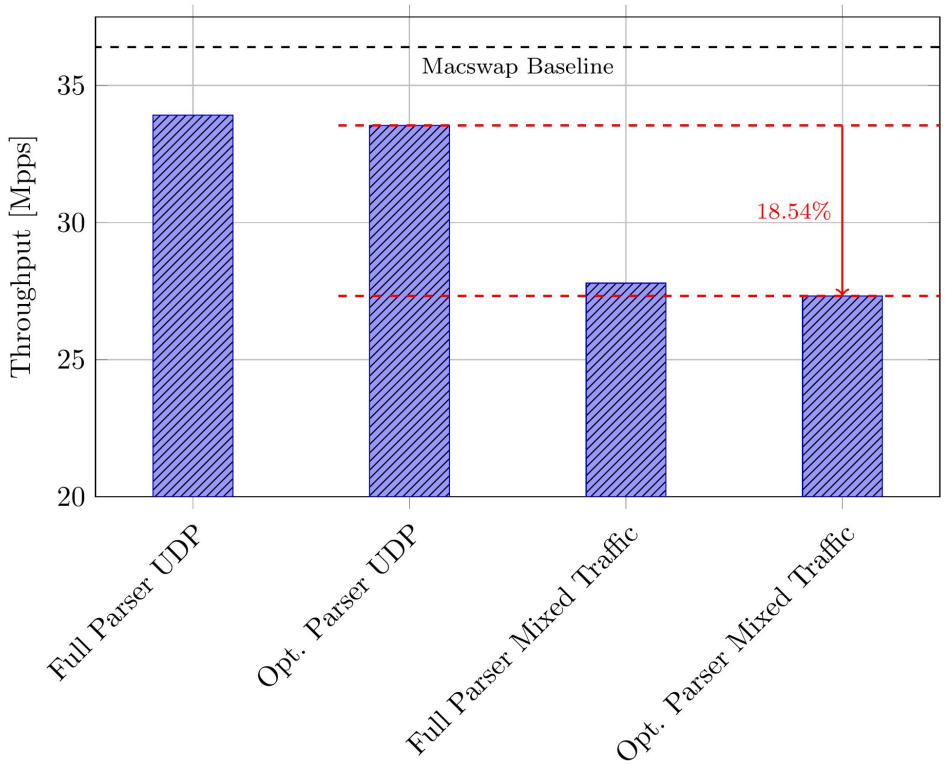
https://blog.cloudflare.com/branch-predictor/
https://gcc.gnu.org/onlinedocs/gcc-4.3.4/gcc/Optimize-Options.html

Do SW NFs Need Help? The Load Balancer Example



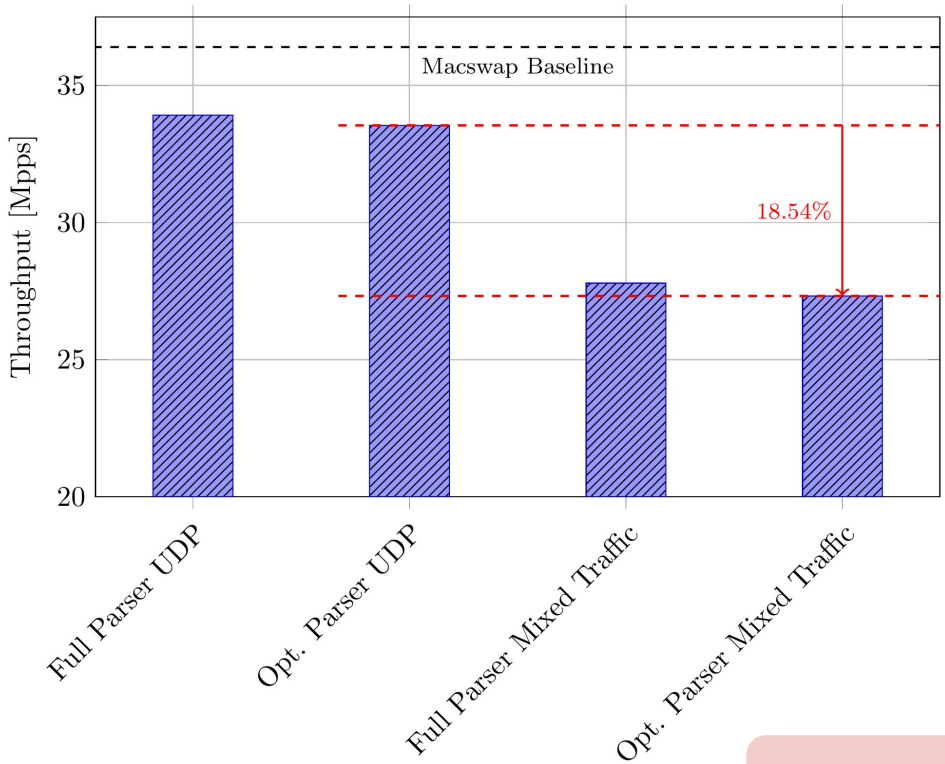
Not So Performant … What is happening?

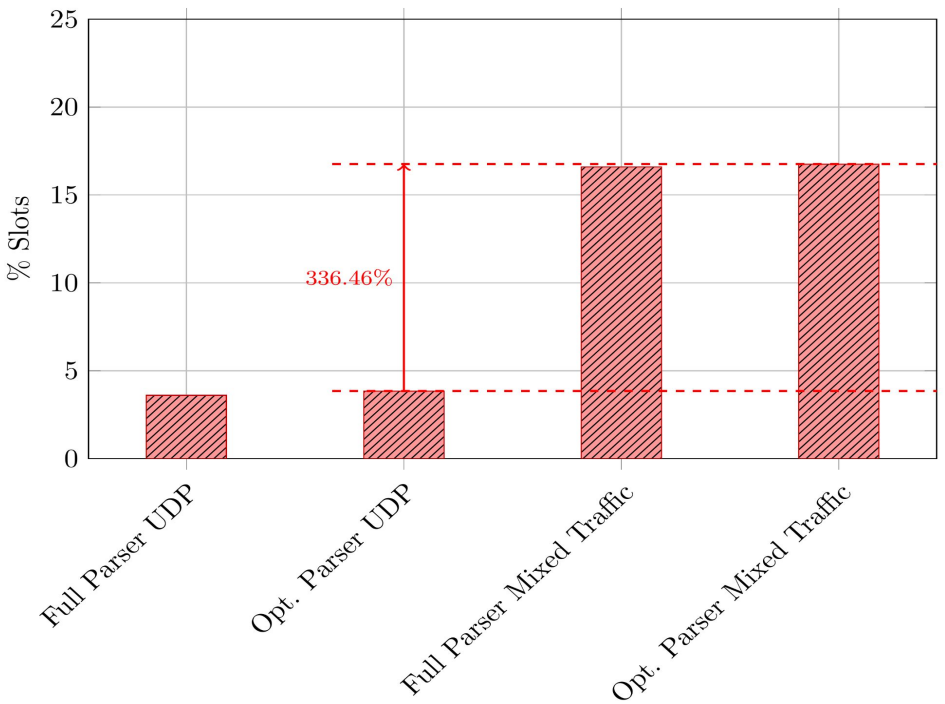
Understanding Protocol Variability Performance Impact



Yasin, Ahmad. "A top-down method for performance analysis and counters architecture." 2014 IEEE ISPASS
https://www.intel.com/content/www/us/en/docs/vtune-profiler/cookbook/2023-0/top-down-microarchitecture-analysis-method.html
https://perfwiki.github.io/main/top-down-analysis/

Understanding Protocol Variability Performance Impact





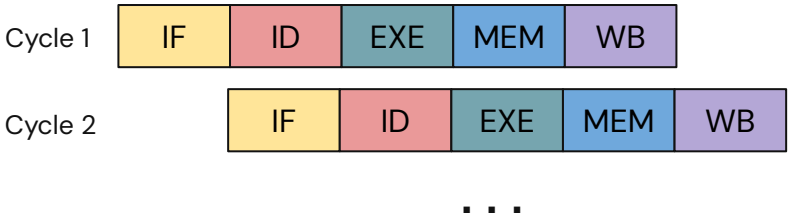
Yasin, Ahmad. "A top-down method for performance analysis and counters ar
https://www.intel.com/content/www/us/en/docs/vtune-profiler/cookbook/20
https://perfwiki.github.io/main/top-down-analysis/

Branch misspredictions are increasing …

Why CPUs Need Branch Predictions?

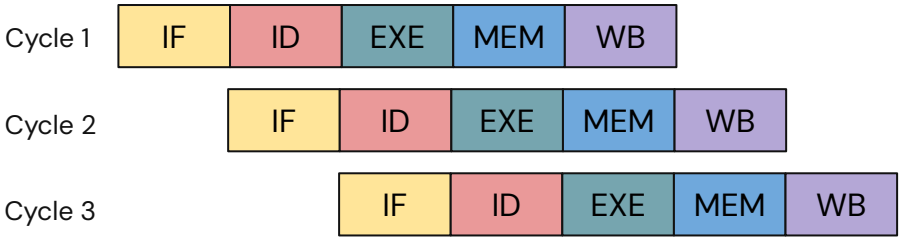
Common 5-stage Pipeline

Why CPUs Need Branch Predictions?



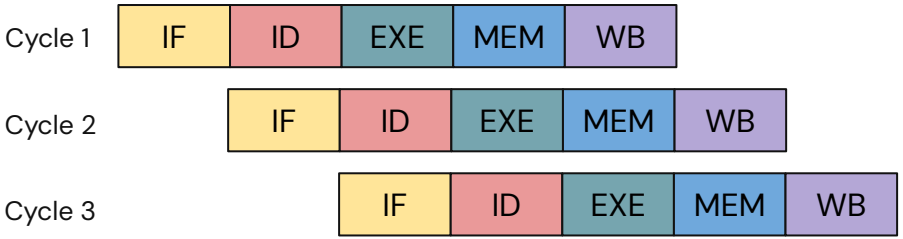
The goal is a new instruction each cycle

Why CPUs Need Branch Predictions?




```
if (condition)
```

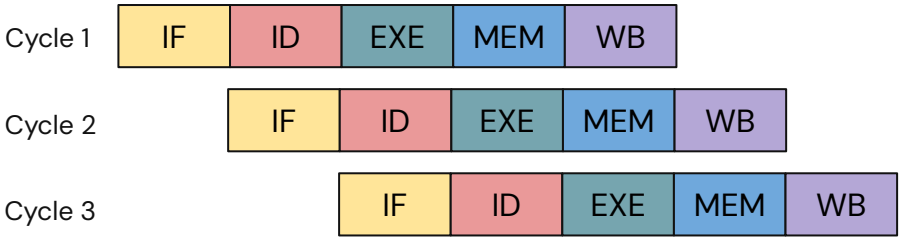
Why CPUs Need Branch Predictions?




```
if (condition)
```

Which is the next instruction?

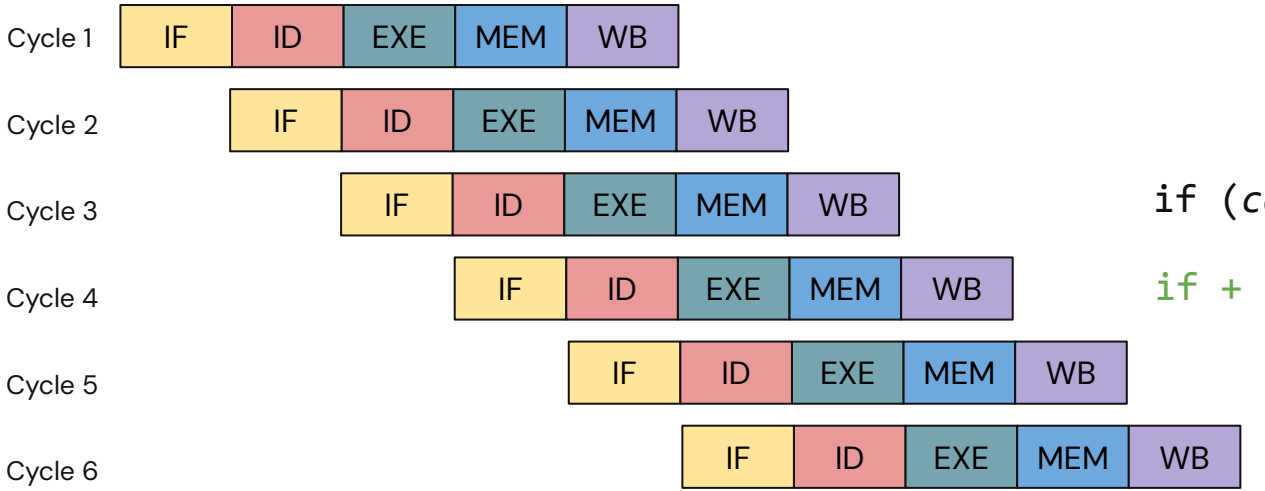
Why CPUs Need Branch Predictions?




```
if (condition)
```

Which is the next instruction?

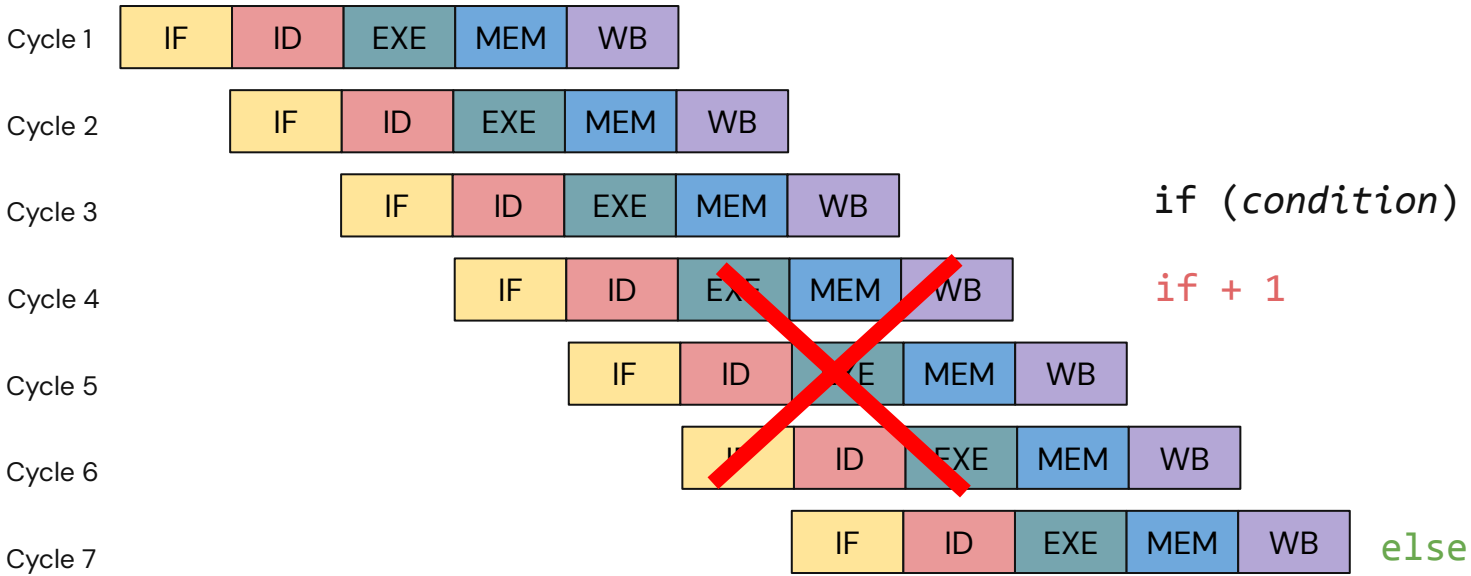
Why CPUs Need Branch Predictions?




```
if (condition)

if + 1
```

If the prediction is correct the CPU saved cycle

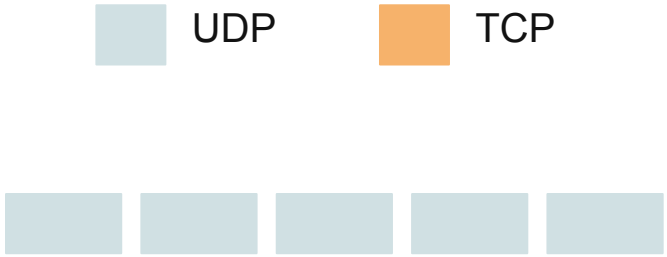
Why CPUs Need Branch Predictions?

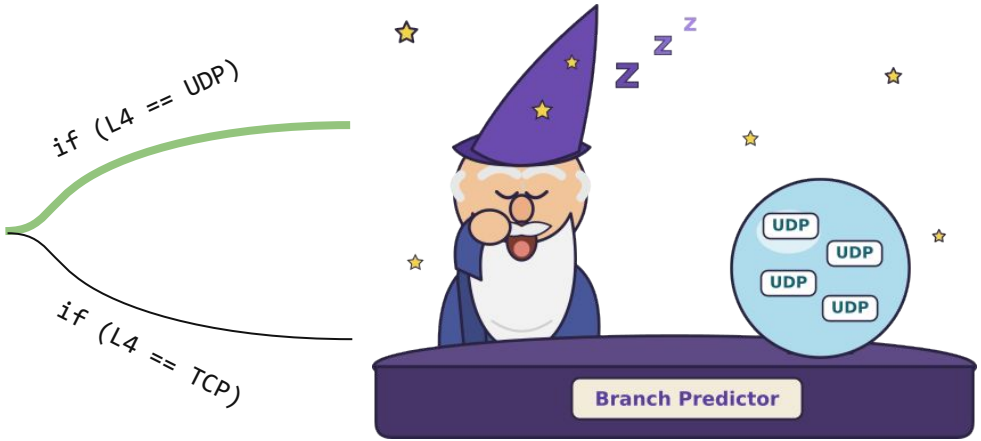


If the prediction is wrong the CPU wasted work

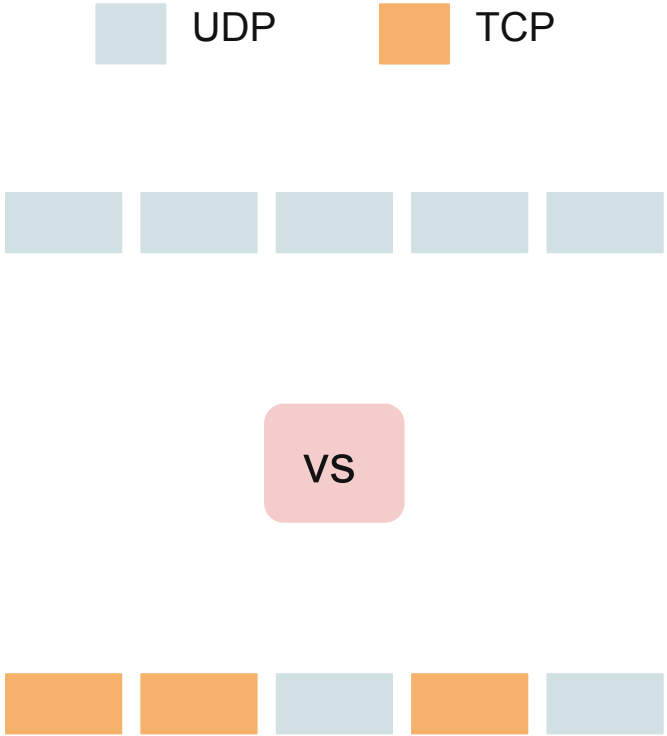
https://blog.cloudflare.com/branch-predictor/
D. A. Patterson and J. L. Hennessy, Computer Organization and Design RISC-V Edition: The Hardware Software Interface, 2nd ed. Morgan Kaufmann, 2021.
Kocher, Paul, et al. "Spectre attacks: Exploiting speculative execution." Communications of the ACM 63.7 (2020)

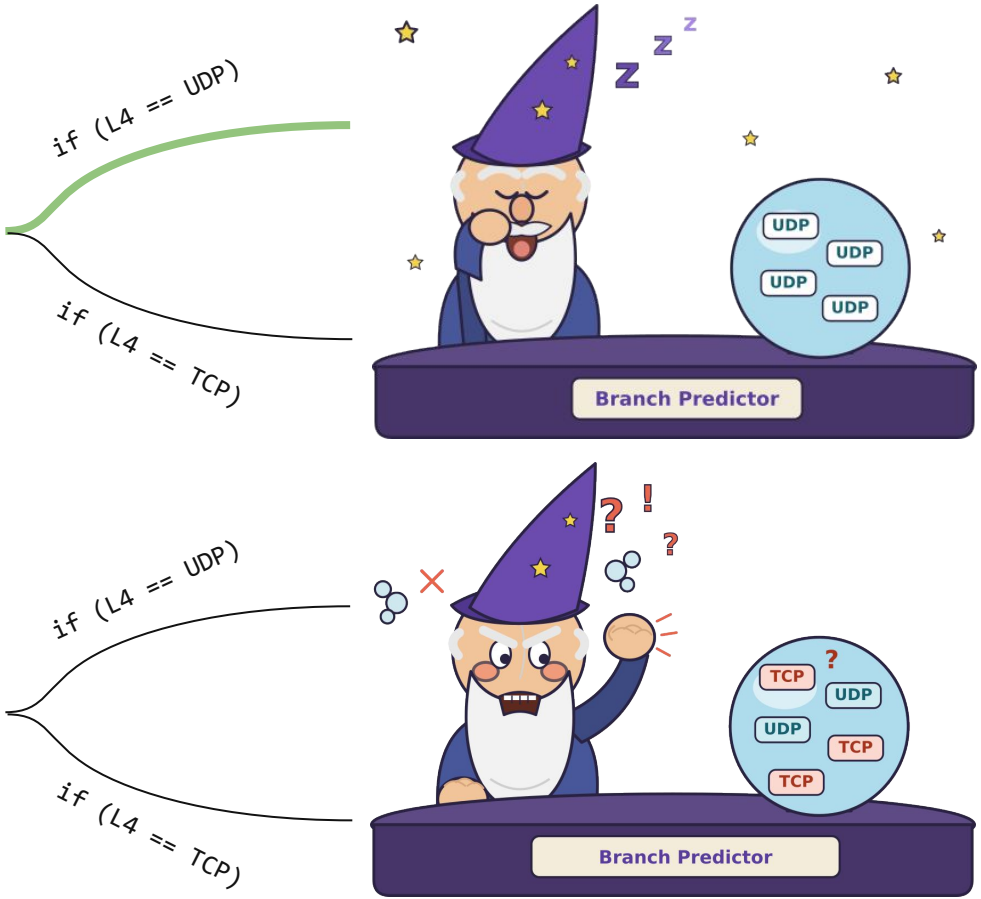
Understanding Protocol Variability Performance Impact





Understanding Protocol Variability Performance Impact





Understanding Protocol Variability Performance Impact

5.8 3rd Gen Intel® Xeon® Scalable Processor with Intel® SST-BF + VIRTIO 1.1 + DPCLS Intel® AVX-512 + Intel® AVX-512 Vector ICE PMD + DPIF/MFEX Patches

This benchmark configuration builds on the previous configuration by including the DPIF and MFEX optimized patches using Intel® AVX-512. This means that the datapath interface (DPIF, or datapath glue code) is accelerated using Intel® AVX-512 to perform tasks like batching of packets. The MFEX (miniflow extract, or packet parsing code) is optimized using a "mask and match" approach, where specific traffic patterns are optimized. By combining the width of the Intel® AVX-512 SIMD registers and incoming packet headers, it is possible to probe for a single packet type in parallel. This avoids the branchy protocol-by-protocol parsing that is implemented in scalar code, and significantly reduces the instruction count to parse the specific packet type.

Note that the optimizations included in the benchmark configuration are not yet upstream in the OVS project; however, the patches used for this benchmark are publicly available here: https://patchwork.ozlabs.org/project/openvswitch/cover/20210212171718.2189798-1-harry.van.haaren@intel.com/.

Van Haaren, Harry

```
v9 Summary:
- Added AVX512 POC work for DPIF and MFEX in single patch at end
-- Note that the AVX512 MFEX is for Ether()/IP()/UDP() traffic.
-- A significant performance boost is possible with these optimizations.
```

Because of the branch misspredictions …

https://builders.intel.com/docs/networkbuilders/open-vswitch-optimized-deployment-benchmark-technology-guide-1645719787.pdf

Do We Need All of These Protocols?

Do We Need All of These Protocols?

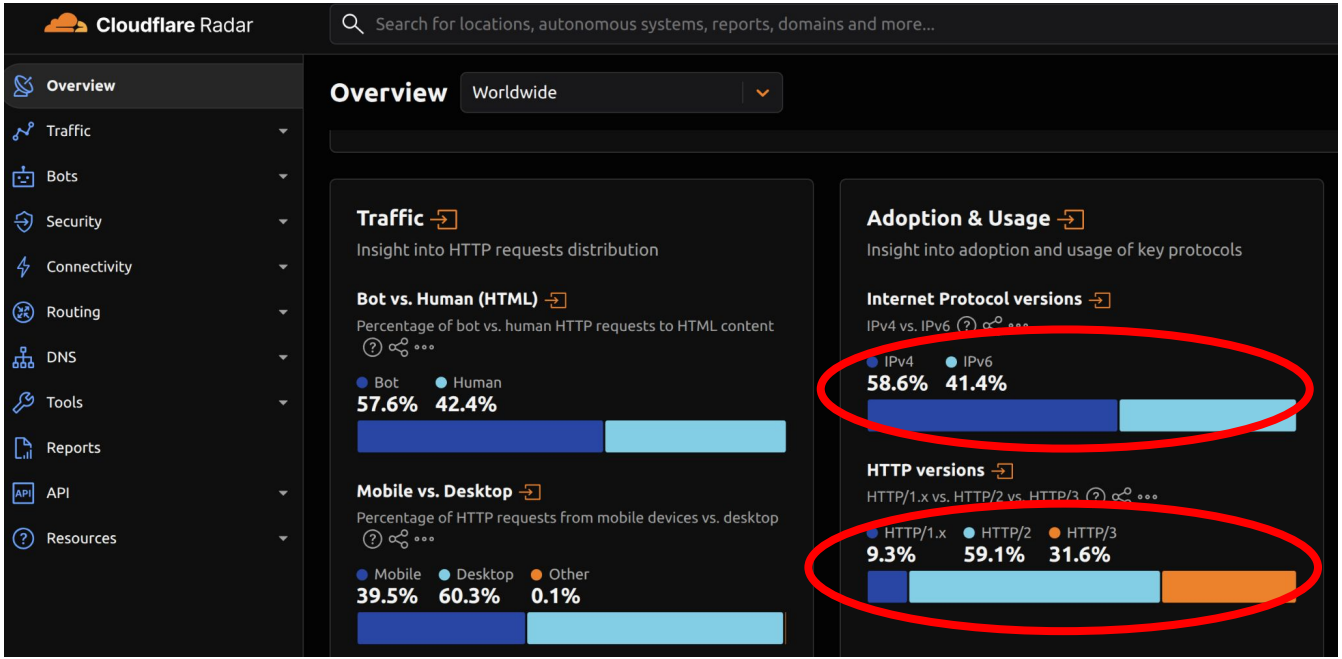
Not always but …

Do We Need All of These Protocols?

Not always but …

L3: IPv4 vs. IPv6
L4: UDP vs. TCP (QUIC?) vs …

Encapsulation: VLAN vs. VXLAN Vs. …



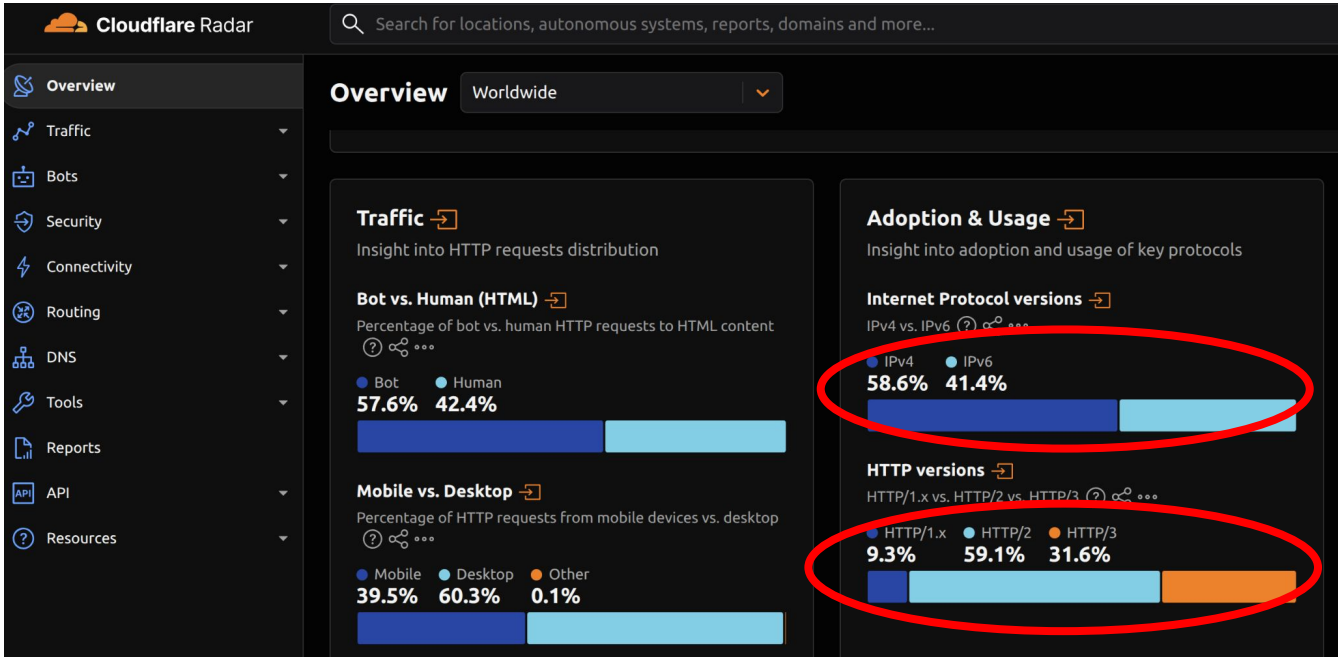
https://radar.cloudflare.com/

Do We Need All of These Protocols?

Not always but …

L3: IPv4 vs. IPv6
L4: UDP vs. TCP (QUIC?) vs …

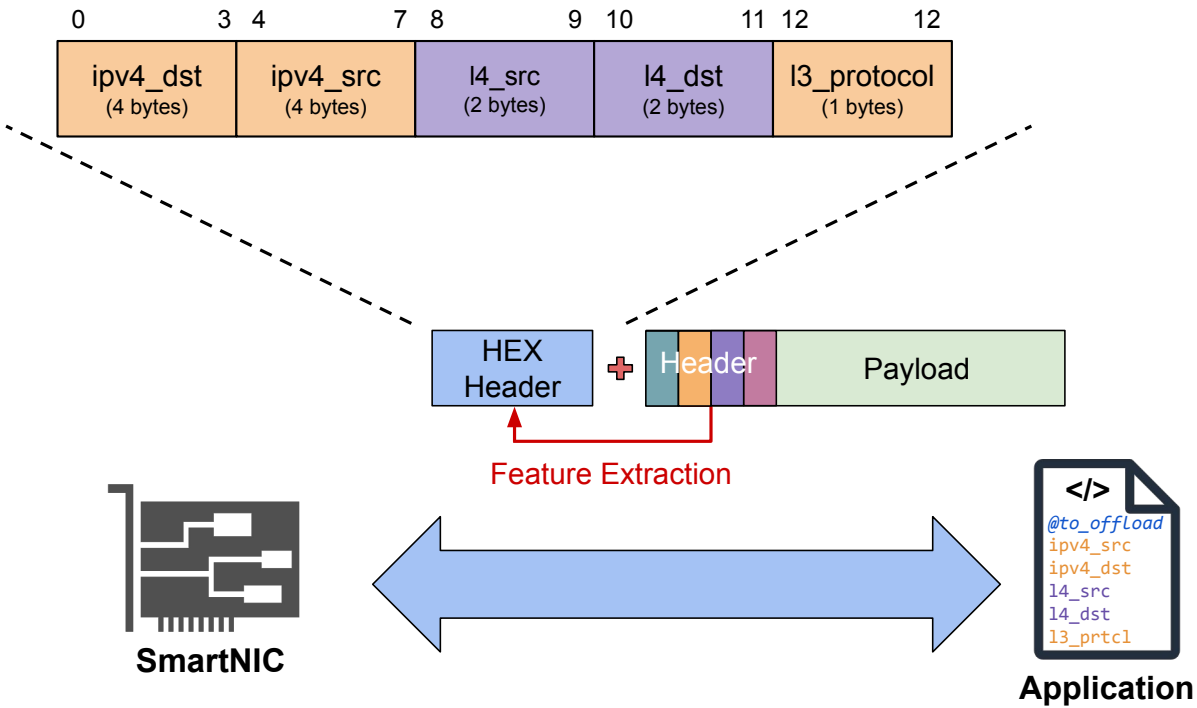
Encapsulation: VLAN vs. VXLAN Vs. …



UDP vs. TCP was enough to show the drop!

https://radar.cloudflare.com/

Our Proposal: NIC Could you Parse This For me?



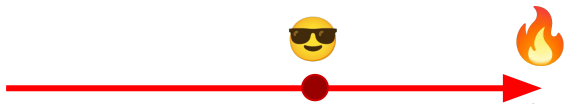
Parsing Offloading Recipe: What do We Need?

Generate a parser configuration

Parsing Offloading Recipe: What do We Need?

Generate a parser configuration

NIC and NFs field communication (keep it simple)

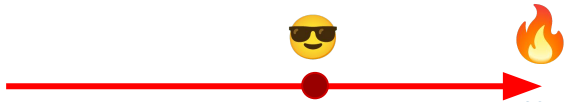


Parsing Offloading Recipe: What do We Need?

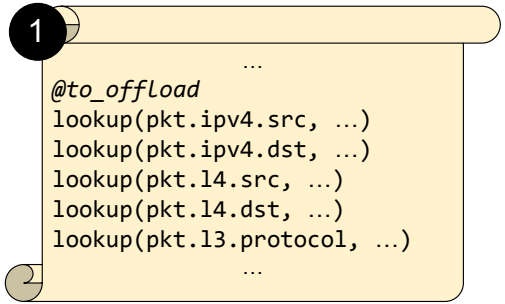
Understand which are the *fields of Interest*

Generate a parser configuration

NIC and NFs field communication (keep it simple)

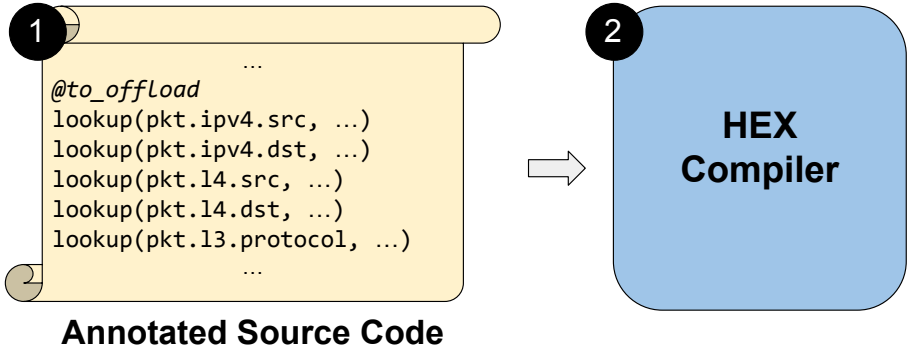


Our Proposal: A Compiler-Based Approach

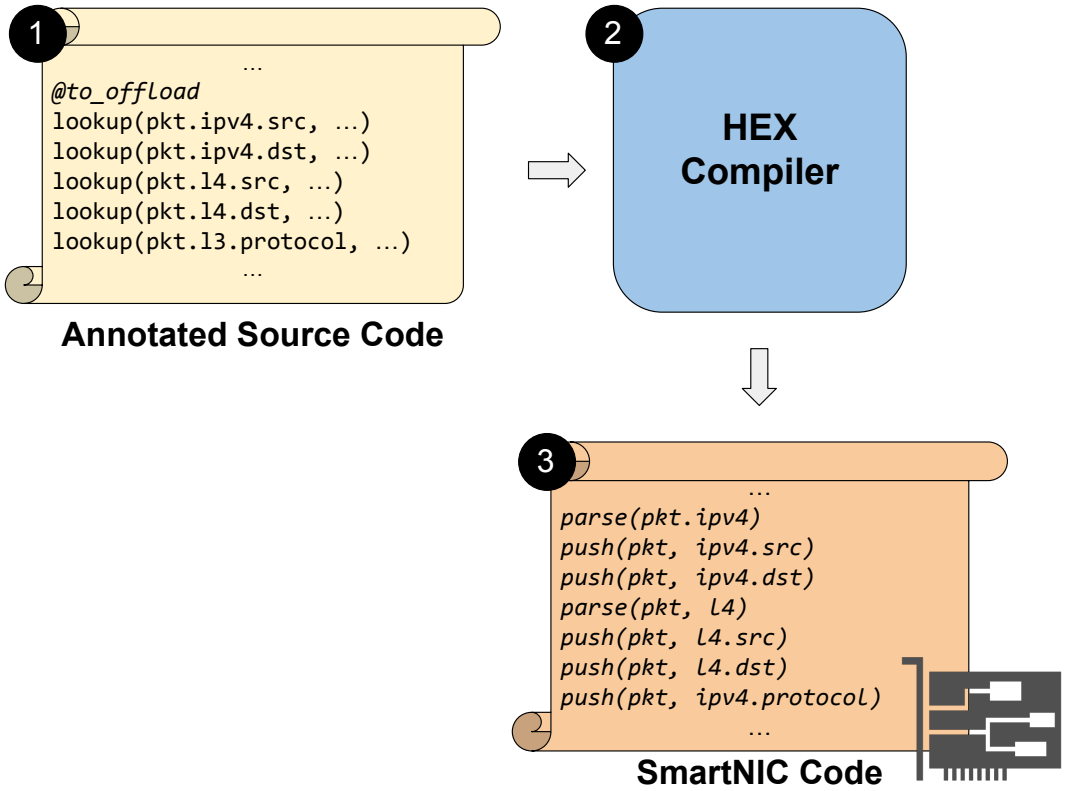


Annotated Source Code

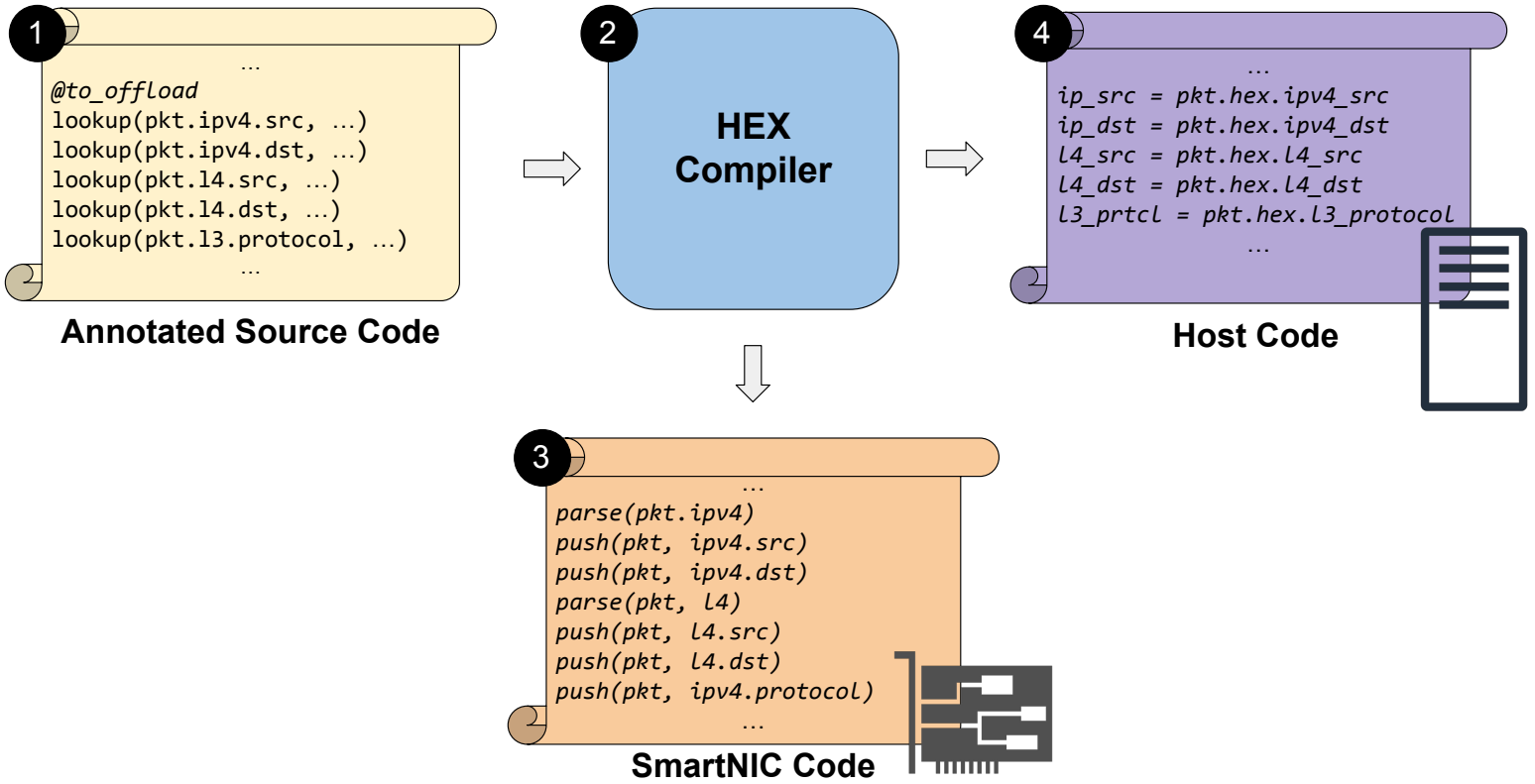
Our Proposal: A Compiler-Based Approach



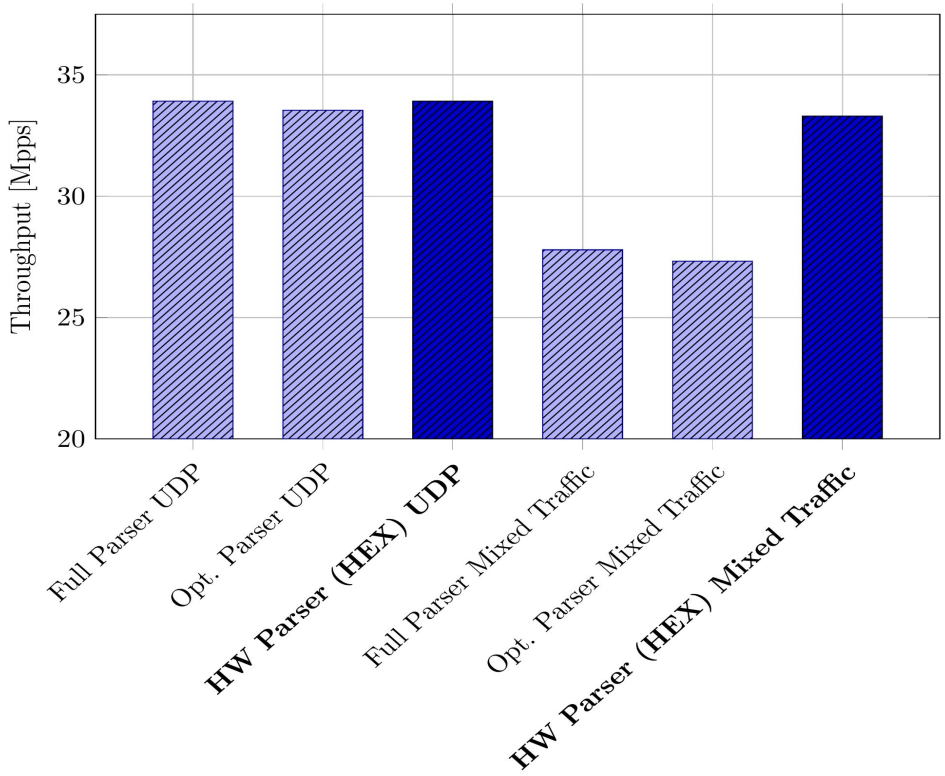
Our Proposal: A Compiler-Based Approach



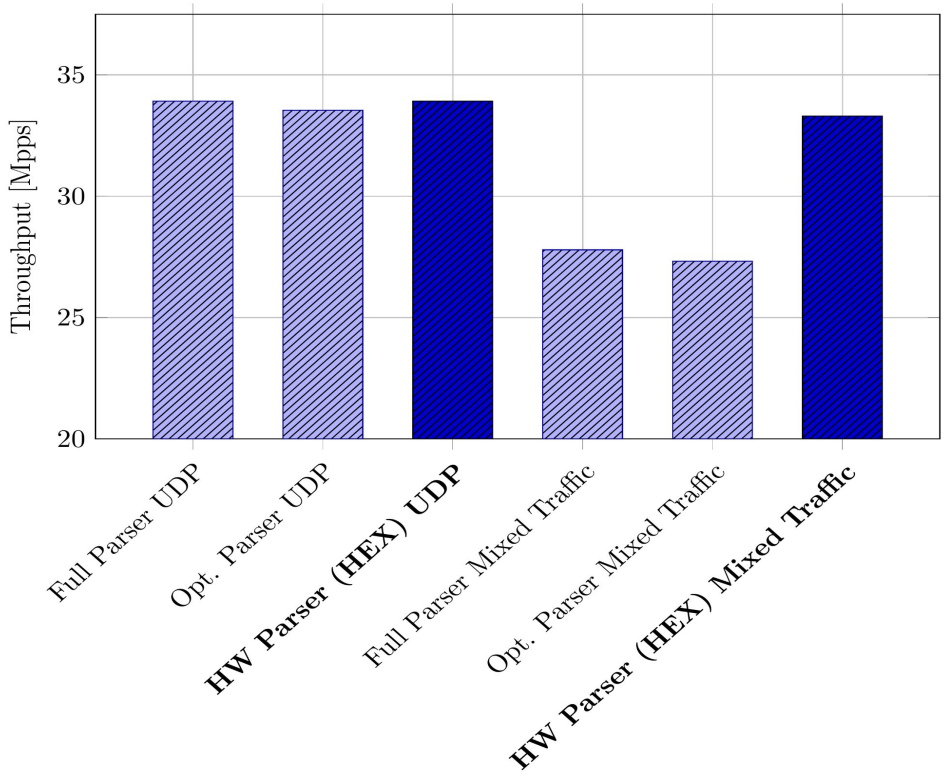
Our Proposal: A Compiler-Based Approach

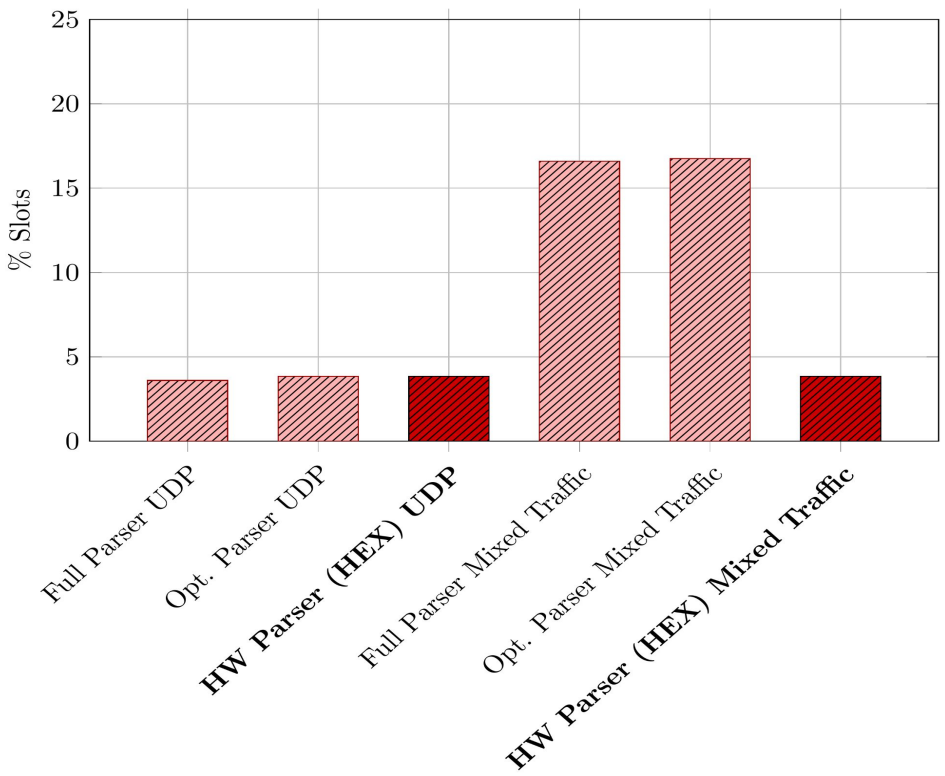


Taming Protocol Variability Challenges



Taming Protocol Variability Challenges

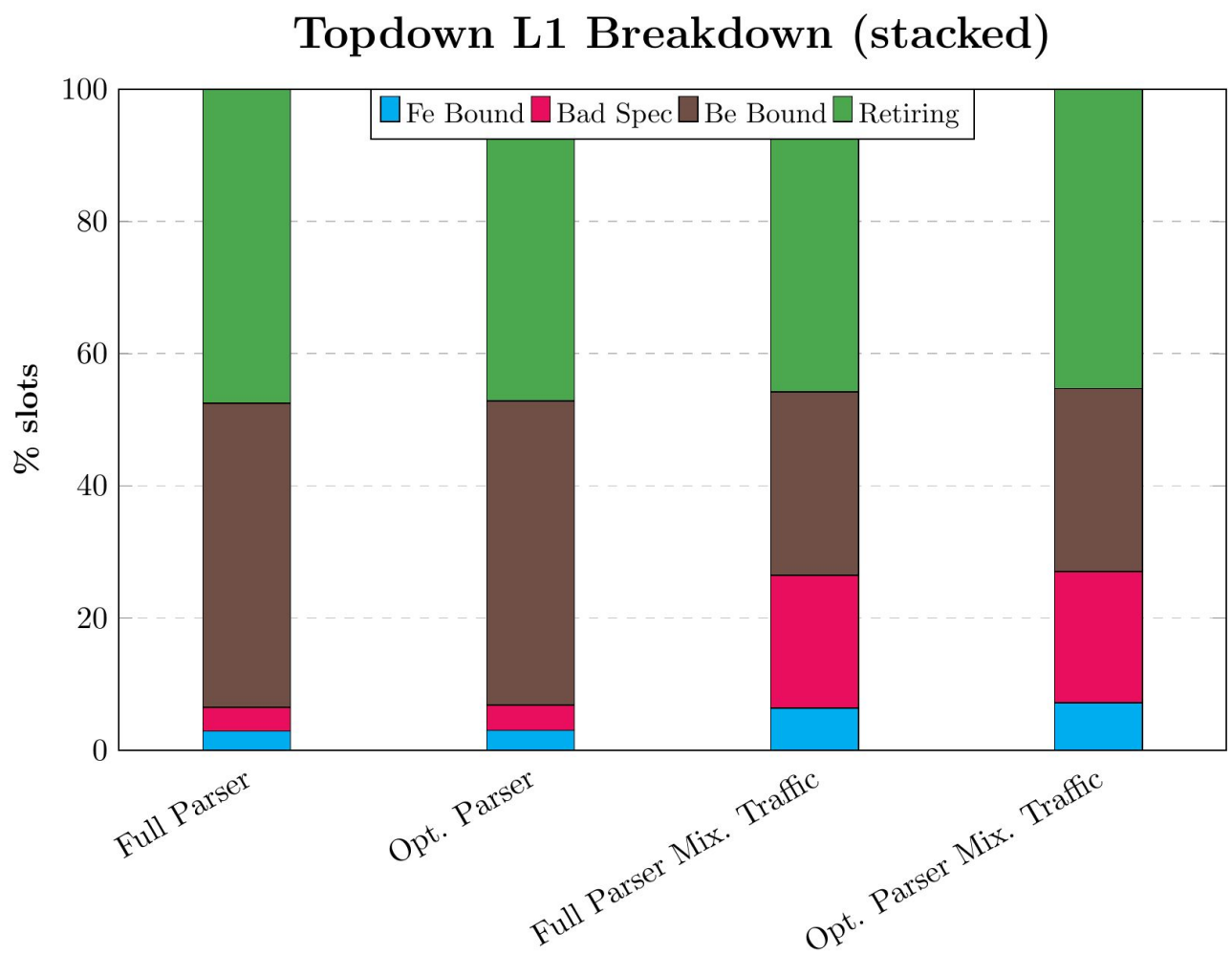




Taming Protocol Variability Challenges

Thanks

Backup


Topdown L1 Breakdown (stacked)
Fe Bound
Bad Spec
Be Bound
Retiring
% slots
100
80
60
40
20
0
Full Parser
Opt. Parser
Full Parser Mix. Traffic
Opt. Parser Mix. Traffic