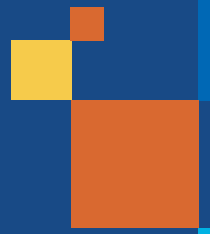
P4 Spring Workshop 2022

Architecture for multi-Terabit programmable networking functions

Petr Kastovsky, petr.kastovsky@intel.com

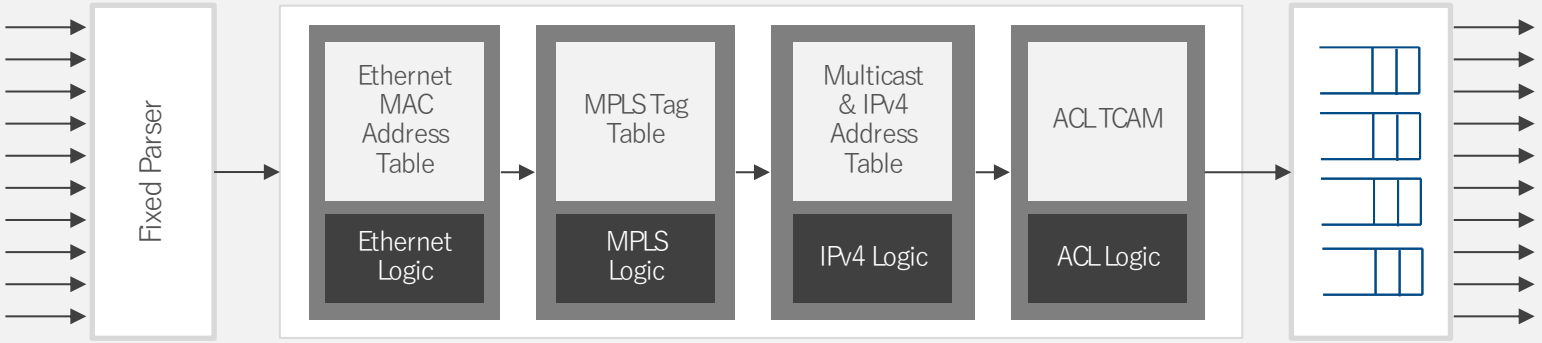
Georgios Nikolaidis, georgios.nikolaidis@intel.com

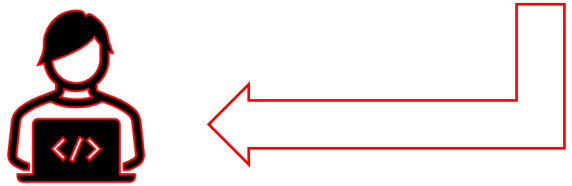


intel®

Fixed-function ASIC:

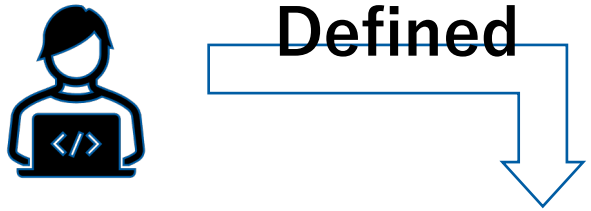
Features and table-sizes are **hard-coded**, not optimized, and often **wasted**.





This is how you process packets …

New World of Network Design: Software-Defined

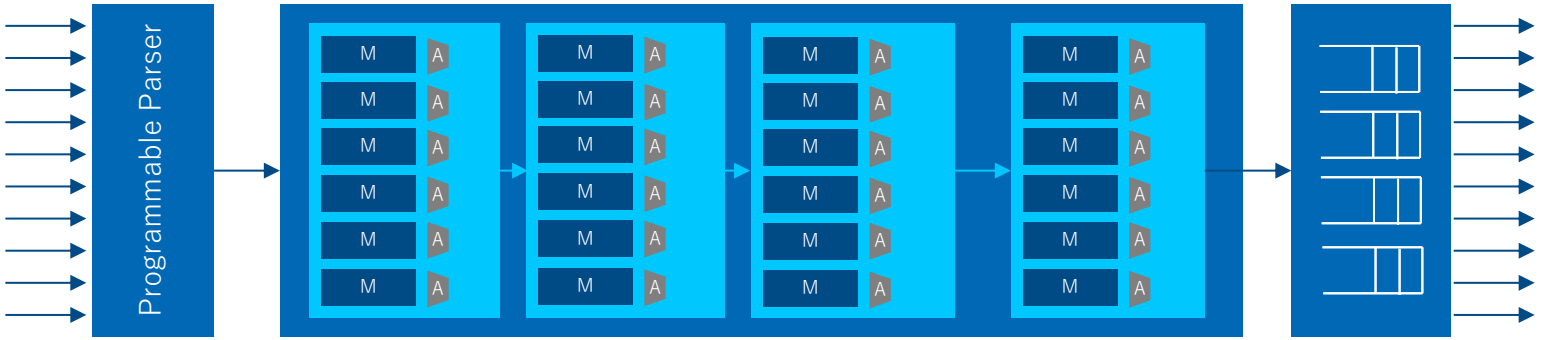


This is how I want to process packets …

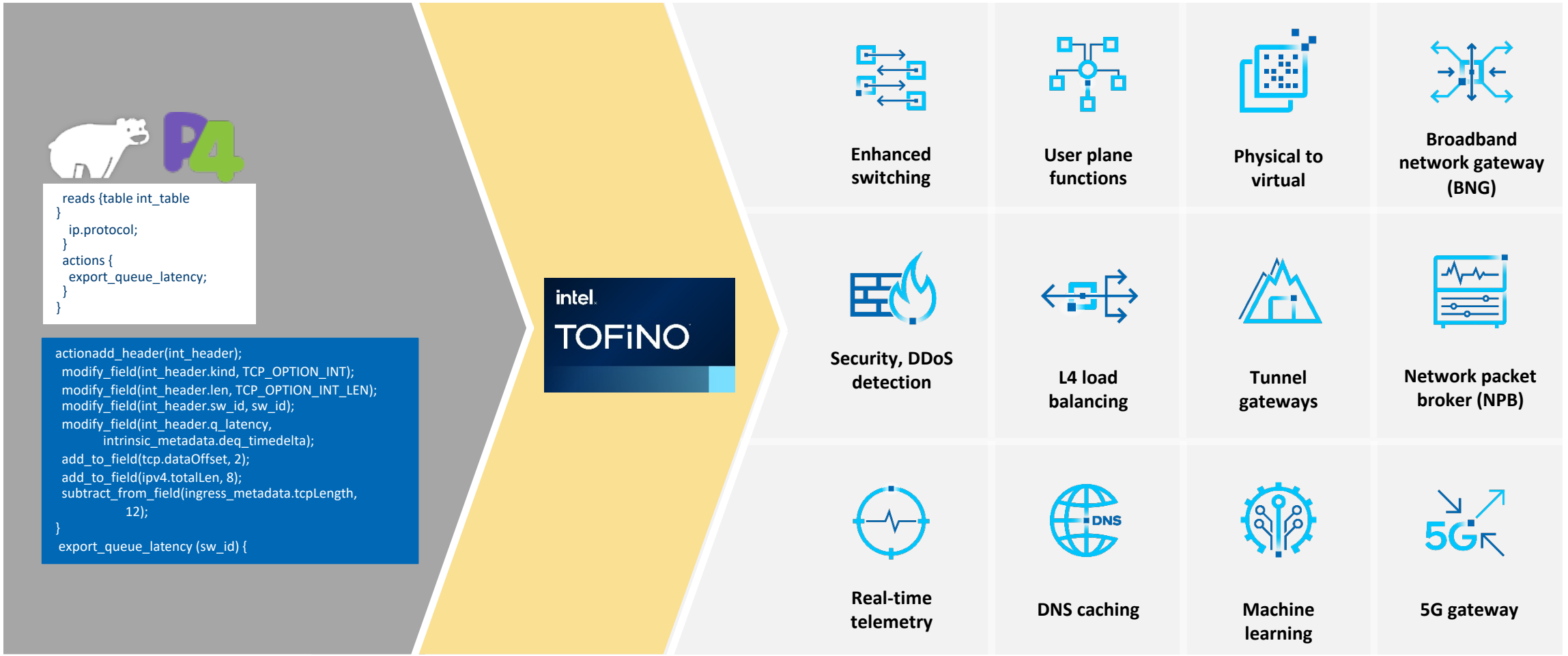
Programmable ASIC:

Software determines headers, table sizes, and packet-processing functions.

Pipeline is fully workload-optimized.



Innovate at the pace of software with new P4 Applications



Alibaba Cloud Gateway XGW

- **Business challenge:** network traffic during a shopping festival can draw millions of shopper visits creating terabits per second of data traffic
- Scale and growth rate of Alibaba Cloud today makes horizontal scaling (adding servers) unsustainable
  - Increased CAPEX (hardware acquisition costs) and OPEX (maintenance, management, troubleshooting)
  - Heavy-hitter flows overloading single CPU core, need for headroom further increasing CAPEX & OPEX
- Technical challenge:
  - Limited total on-chip memory capacity storing the O(1M)VXLAN routing table and O(1M) VM-NC mapping table ***while stateful tables and ultra-large in table entries, such as SNAT require O(100M) entries***

Source: Tian Pan, Nianbing Yu, Chenhao Jia, Jianwen Pi, Liang Xu, Yisong Qiao, Zhiguo Li, Kun Liu, Jie Lu, Jianyuan Lu, Enge Song, Jiao Zhang, Tao Huang, Shunmin Zhu. 2021. Sailfish: Accelerating Cloud-Scale Multi-Tenant Multi-Service Gateways with Programmable Switches. In ACM SIGCOMM 2021Conference (SIGCOMM '21), August 23–28, 2021, Virtual Event, USA. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3452296.3472889

Tofino Expanded Architecture: Intro

Complementing Tofino by FPGAs to enable 100x increase in table and buffer capacity

eXtra large tables

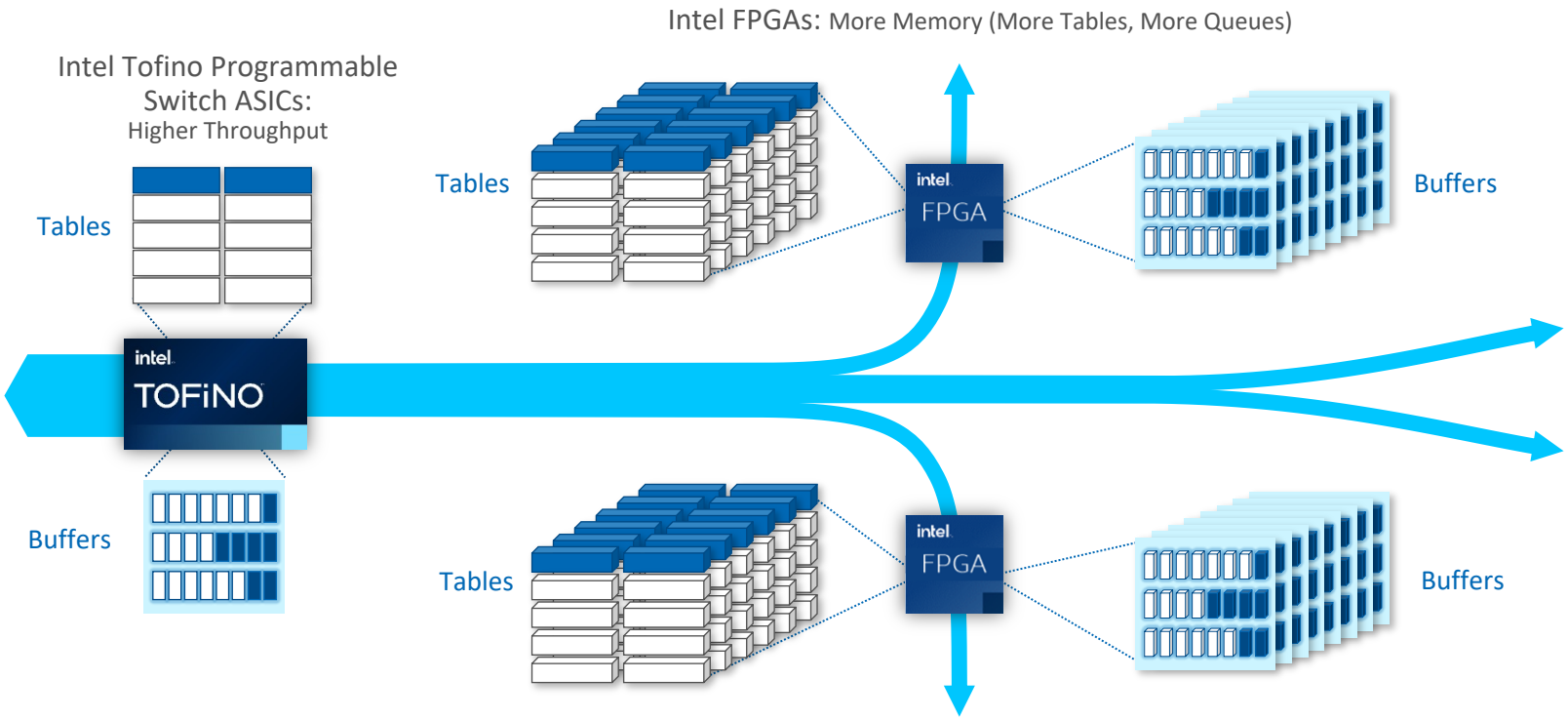
up to 100s of millions of entries

- CSP: cloud gateway (L4 LB, firewall, VxLAN, NAT)
- CoSP: carrier grade NAT, NPB, IPv6 NAT, 5G metro router etc.

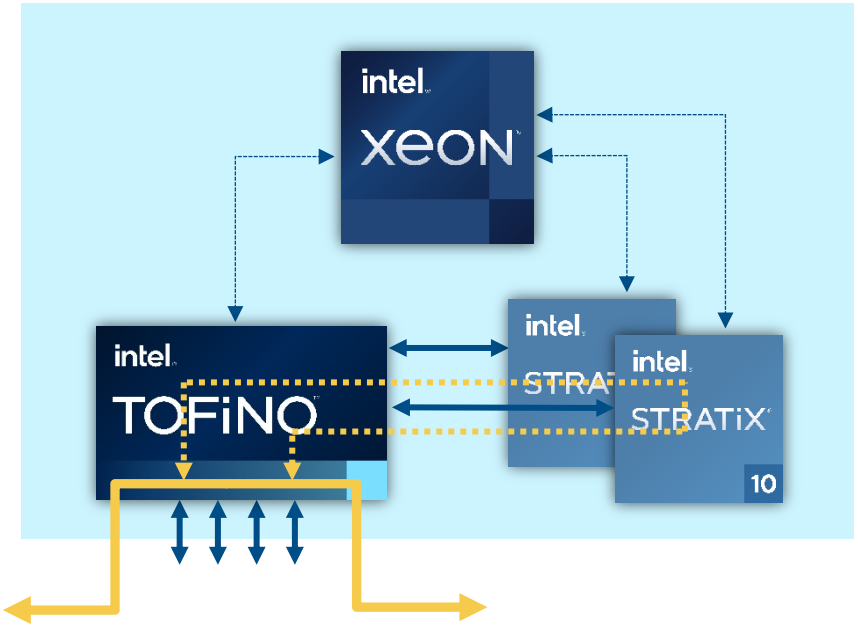
eXtra large buffers

up to 10s of GBs of buffers

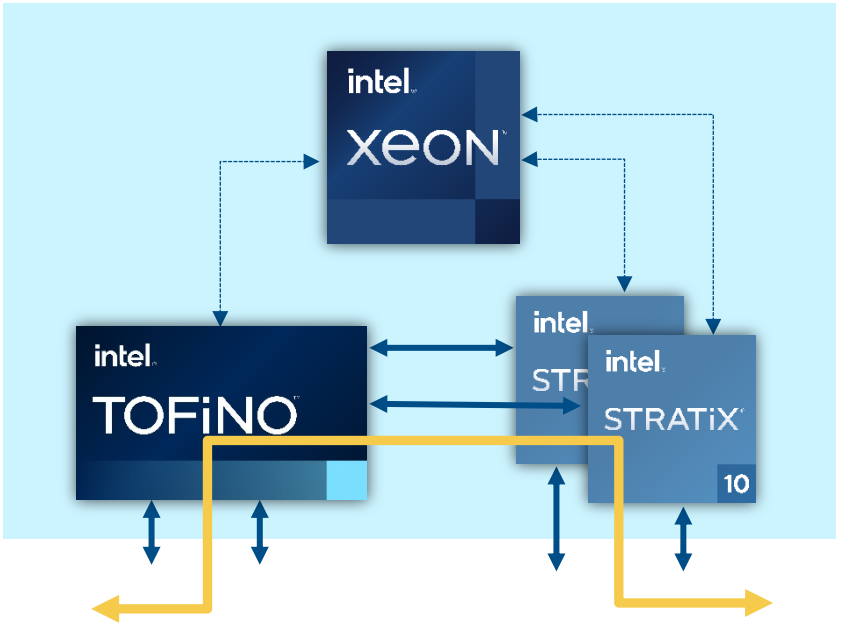
- CoSP: telco gateway BNG/5G UPF/AGF, 5G metro router, NFV acceleration



Architecture Implementations



FPGA Look-Aside to Switch ASIC



FPGA Inline with Switch ASIC

100GbE data connections

10GbE control connections

Data-plane traffic path

Look-aside traffic path

Hardware Form factors

Switch + FPGA SmartNIC

- Tofino-based switch
- FPGA SmartNIC cards in a separate server
  - Intel® FPGA PAC N3000
  - N5010 (LC)

intel N6000

N5010

Silicom Ltd Connectivity Solutions

Server switch

- Integrated platform
- Tofino, FPGA, and CPU in one box

Accton

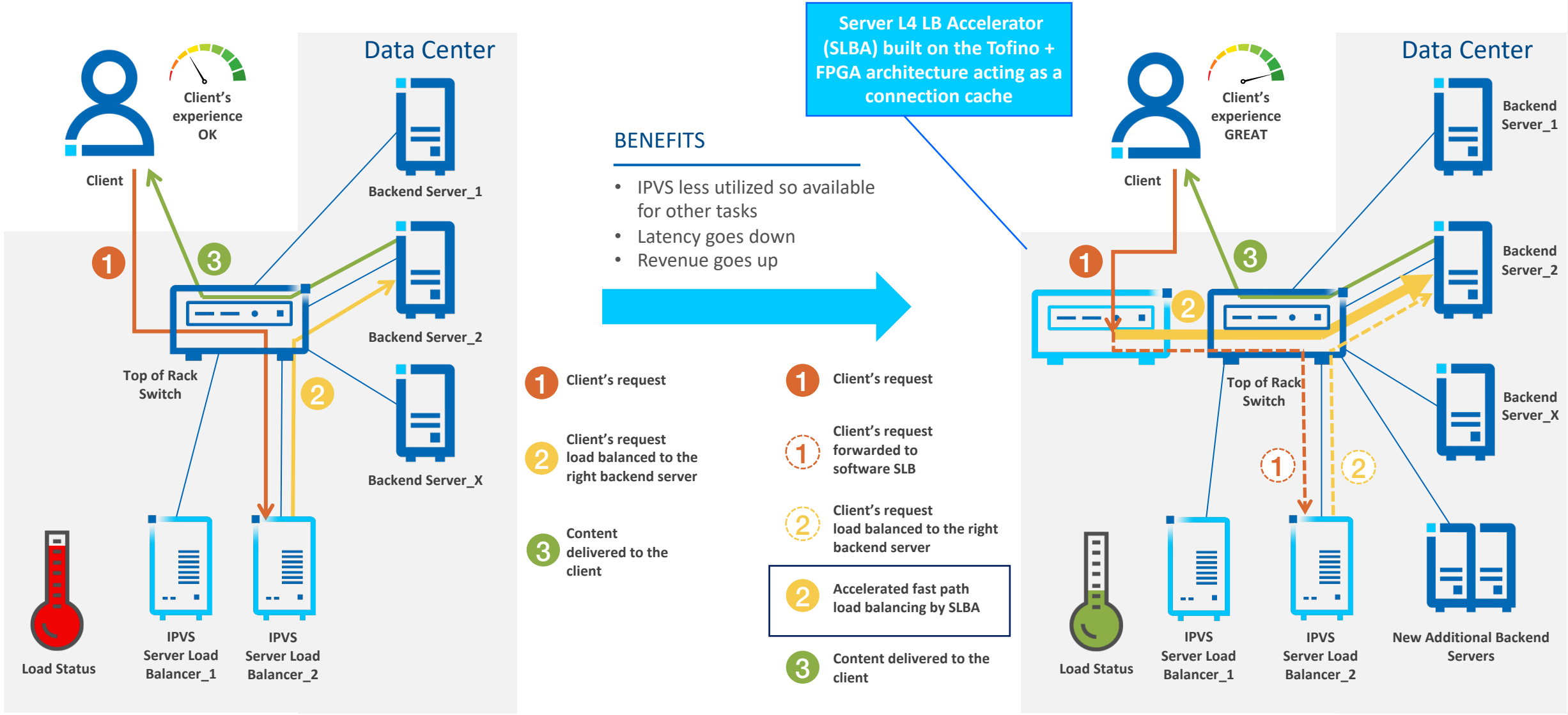
CASWELL Complete Your Solution

APS Networks

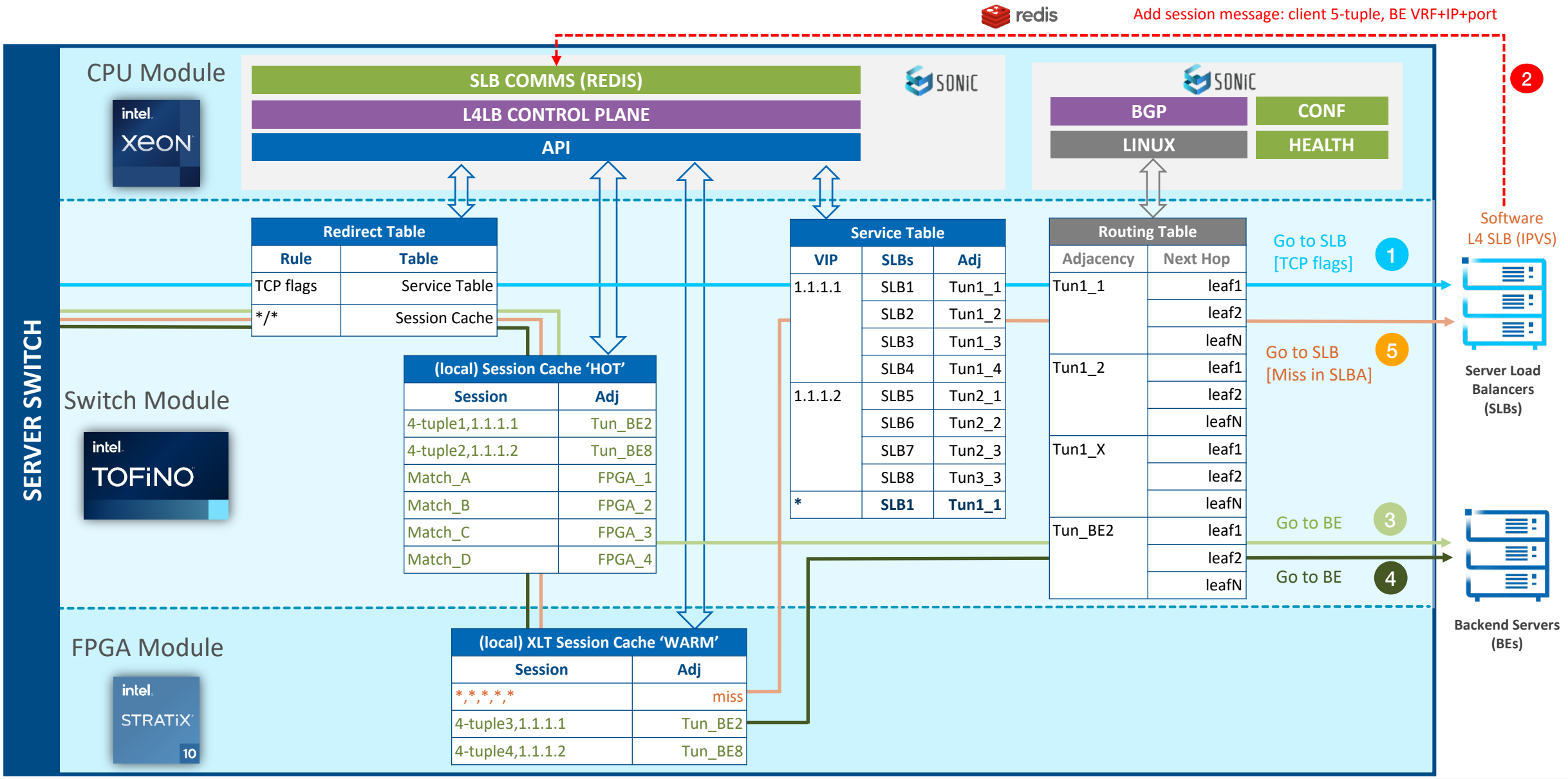
Chassis platform

- Modules with Tofino, FPGA, and CPU

Efficient High-Bandwidth Load Balancing

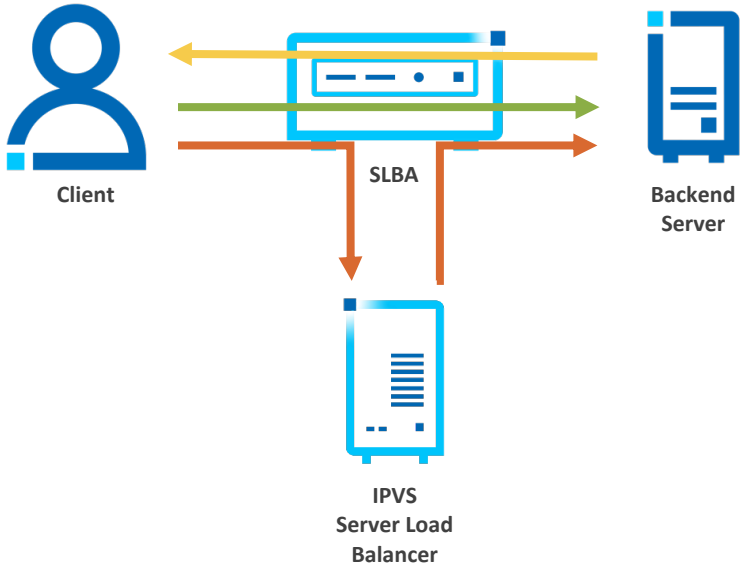


L4 Server Load-Balancer Accelerator Data-Plane Architecture



SLBA demo setup

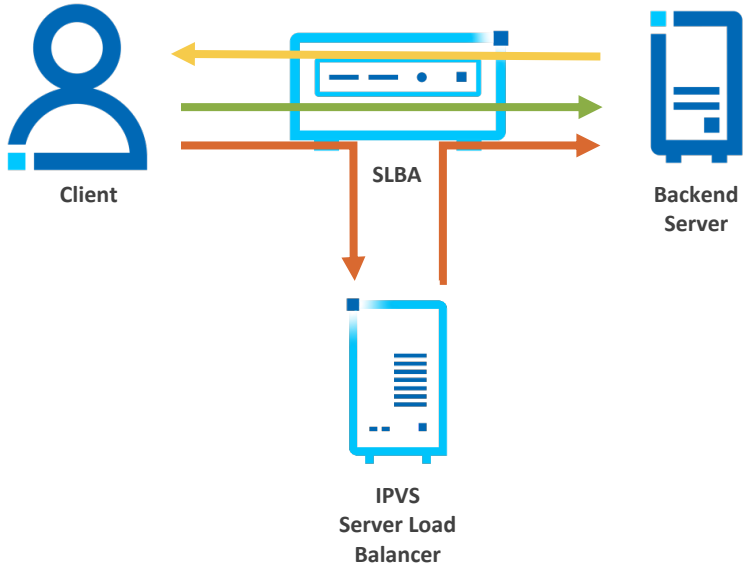
- One client, one SLB and one BE, 100GbE, 2x22 cores each
- Client initiates progressively 1000 iperf3 flows
- First run with SLBA disabled, second run with SLBA enabled

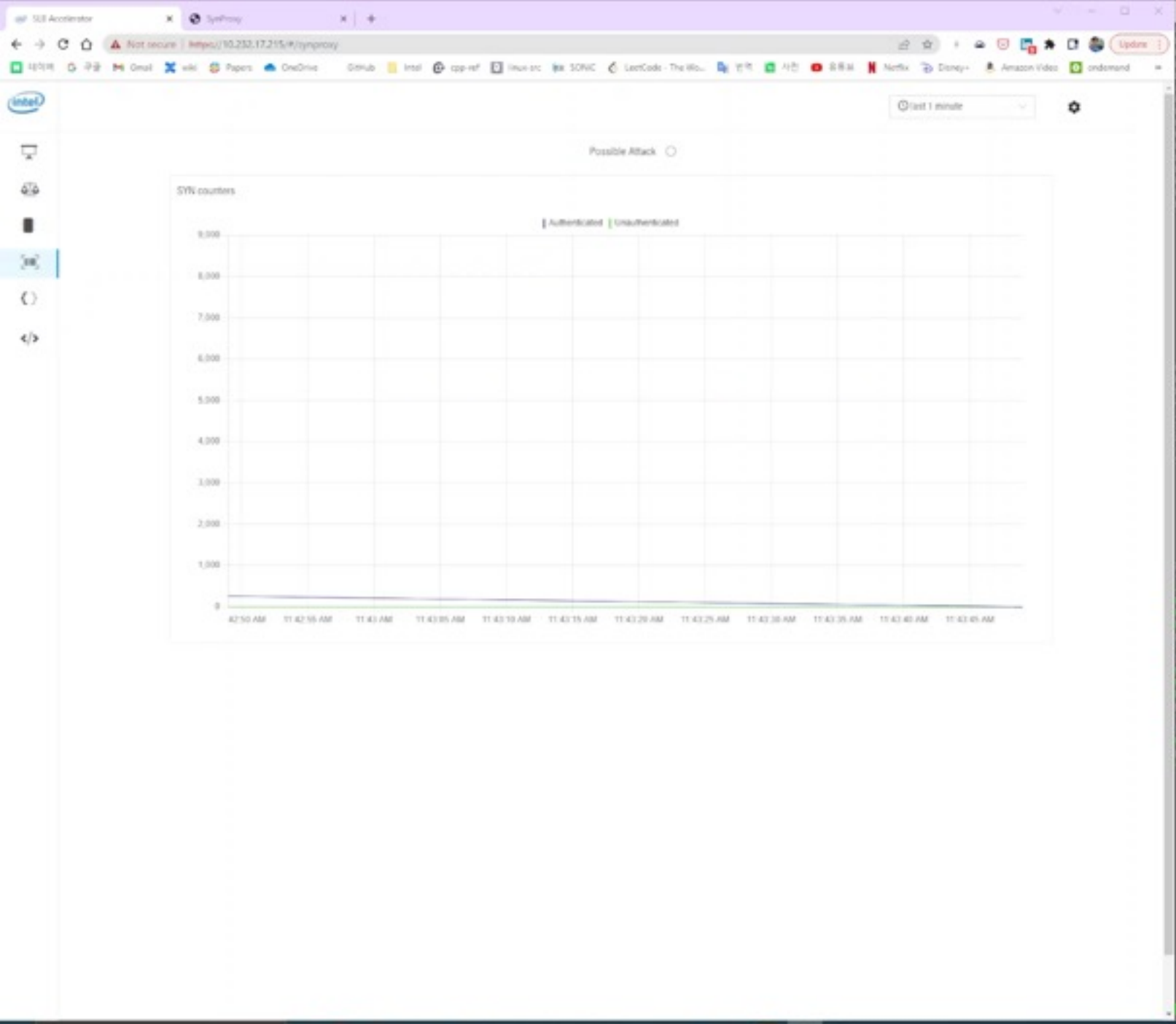



last 5 minutes
Tofino Hits (pkts)
0
Tofino Misses (pkts)
176786459
Tofino Connections
1
FPGA Hits (pkts)
105071926
FPGA Misses (pkts)
74939067
FPGA Connections
0
Tofino Cache Hit Rate (pkts/sec)
Tofino Cache Miss Rate (pkts/sec)
FPGA Cache Hit Rate (pkts/sec)
FPGA Cache Miss Rate (pkts/sec)
Tofino Connection Update Rate (conn/sec)
FPGA Connection Update Rate (conn/sec)
7:33 PM
7:34 PM
7:35 PM
7:36 PM
7:37 PM
Active VIPs
0
Active RIPs
0
Active CIPs
0
SLB Tx Rate (Kbps)
SLB Rx Rate (Kbps)
SLB CPU (%)
BE Tx Rate (Kbps)
BE Rx Rate (Kbps)
BE CPU (%)

SynProxy demo setup

- 8k legitimate requests
- 6k attacker requests periodically
- Initially SynFlood detection is off, so all connections are accepted
- We then turn on SynFlood detection and SynProxy mitigation, and attacker traffic is dropped






```
bf_switch:0>
bf_switch:0>
bf_switch:0>
bf_switch:0>
bf_switch:0> set service handle 1 attribute synproxy_mitigation false
service set success device 0: handle:0x8900000000001
bf_switch:0>
bf_switch:0>
bf_switch:0>
bf_switch:0>
bf_switch:0>
bf_switch:0>
bf_switch:0>
bf_switch:0>
bf_switch:0>
bf_switch:0>
bf_switch:0> show service handle 1
    service_handle: 0x8900000000001
    Attributes:
      internal_object: False
      vip: 10.10.10.101
      protocol: 6
      port: 80
      route_handle: route.15
      synproxy_enable: True
      synproxy_mitigation: False
      synproxy_nonce1: 1
      synproxy_nonce2: 2
      synproxy_time_mask: 4294950912
      synproxy_threshold: 13727
bf_switch:0>
bf_switch:0>
bf_switch:0>
```

```
101 --dst-port 80 --pps 5000 --size 10000
Initialization...
Starting the attack...
^CERROR: tcpreplay: execution failed (type)
Traceback (most recent call last):
  File "/usr/local/lib/python2.7/dist-packages/scapy/sendrecv.py", line 535, in sendpfast
    stdout, stderr = cmd.communicate()
  File "/usr/lib/python2.7/subprocess.py", line 463, in communicate
    return self._communicate(input)
  File "/usr/lib/python2.7/subprocess.py", line 1134, in _communicate
    stdout, stderr = self._communicate_with_poll(input)
  File "/usr/lib/python2.7/subprocess.py", line 1170, in _communicate_with_poll
    ready = poller.poll()
KeyboardInterrupt
^C
```

```
py --ip 10.10.10.101 --port 80 --no-stop 0
^C
Processed 138632 requests (7528.14977724 requests/sec.), 0 errors
py --ip 10.10.10.101 --port 80 --no-stop 0
^C
Processed 191913 requests (7559.51291702 requests/sec.), 0 errors
py --ip 10.10.10.101 --port 80 --no-stop 0
```


intel