2021 NCTU P4 workshop

Server Load Balancer Accelerator (SLBA) P4-based Solution

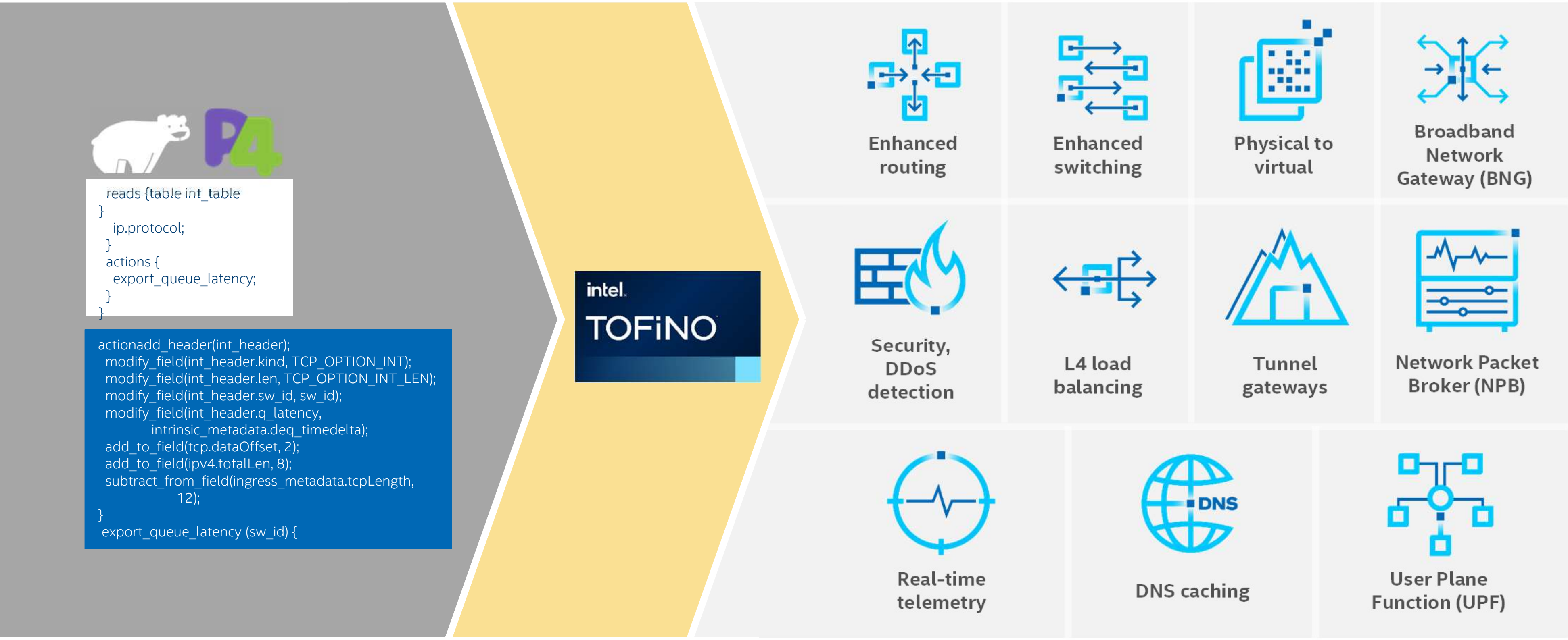
December 21, 2021

Petr.Kastovsky@intel.com, product line manager

Notices & Disclaimers

- Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.
- Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.
- Your costs and results may vary.
- Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.
- Intel technologies may require enabled hardware, software or service activation.
-

Programmability drives applications



```
  reads {table int_table
}
   ip.protocol;
  }
  actions {
   export_queue_latency;
  }
}
```

```
actionadd_header(int_header);
  modify_field(int_header.kind, TCP_OPTION_INT);
  modify_field(int_header.len, TCP_OPTION_INT_LEN);
  modify_field(int_header.sw_id, sw_id);
  modify_field(int_header.q_latency,
            intrinsic_metadata.deq_timedelta);
  add_to_field(tcp.dataOffset, 2);
  add_to_field(ipv4.totalLen, 8);
  subtract_from_field(ingress_metadata.tcpLength,
                  12);
}
 export_queue_latency (sw_id) {
```

Enhanced routing

Enhanced switching

Physical to virtual

Broadband Network Gateway (BNG)

Security, DDoS detection

L4 load balancing

Tunnel gateways

Network Packet Broker (NPB)

Real-time telemetry

DNS caching

User Plane Function (UPF)

Tofino™ X: Intel® Tofino™ Intelligent Fabric Processor with Intel® FPGA

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

eXtra large tables

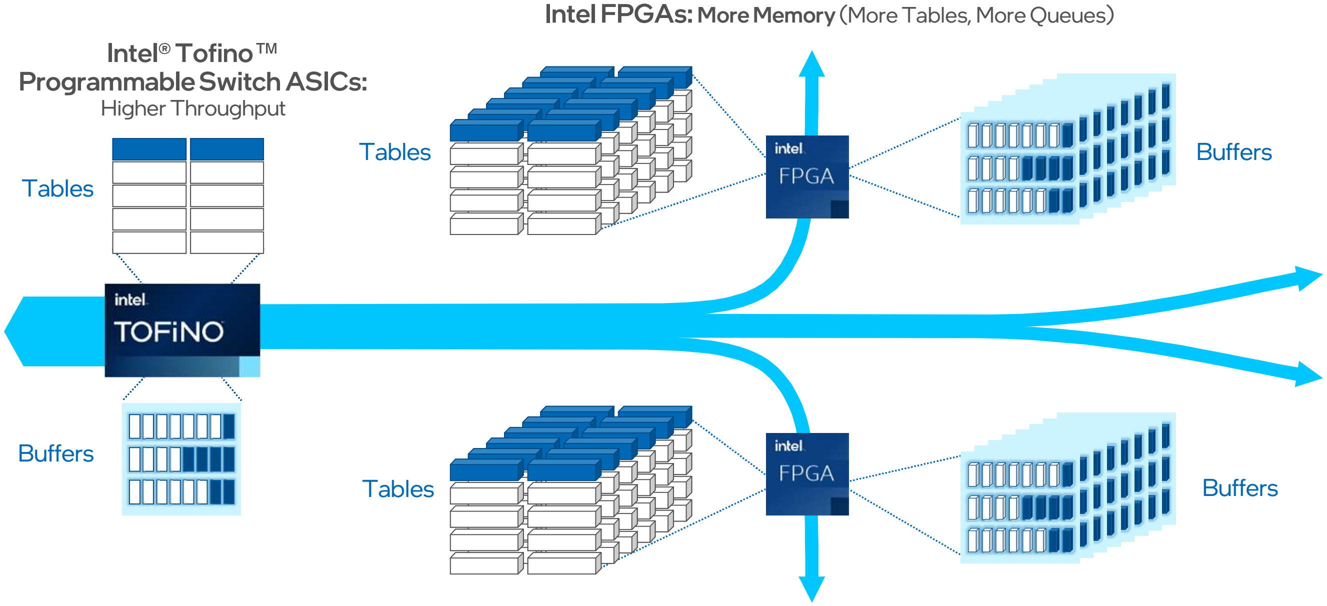
up to 100s of millions of entries

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

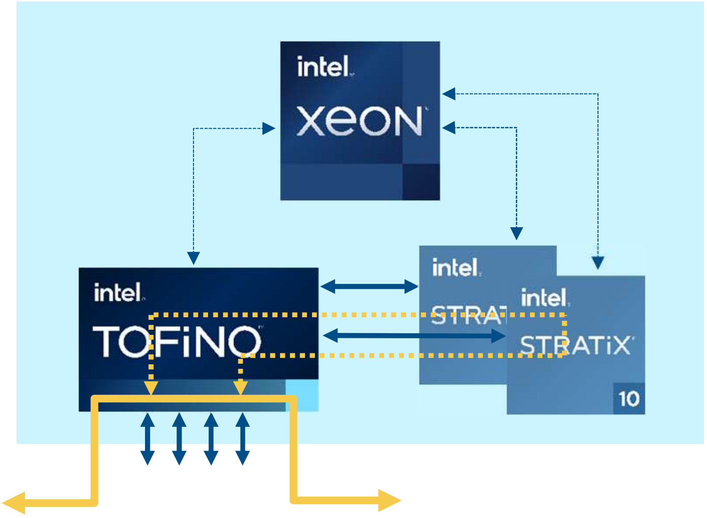
eXtra large buffers

up to 10s of GBs of buffers

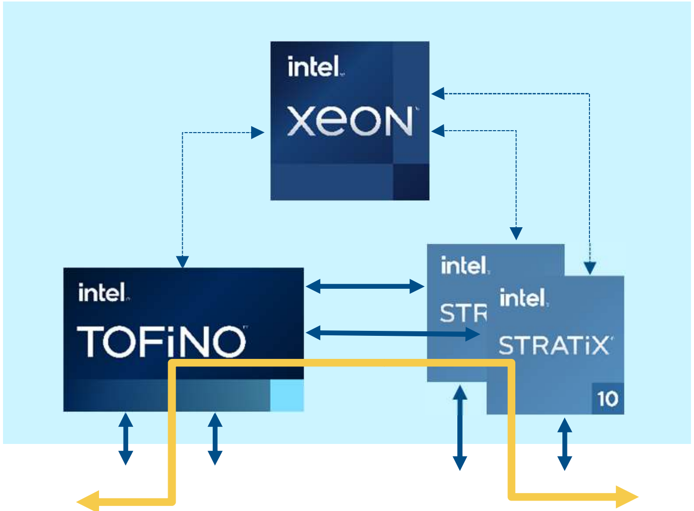
- CoSP: telco gateway BNG/5G UPF/AGF, 5G metro router, NFV acceleration



Intel® Tofino™ X Architecture Implementations



FPGA Look-Aside to Switch ASIC



FPGA Inline with Switch ASIC

⟷	100GbE data connections	⟷	Data-plane traffic path
⟷	10GbE control connections	⟷	Look-aside traffic path

Intel® Tofino™ X Hardware Form factors

Switch + FPGA SmartNIC

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

Switch server

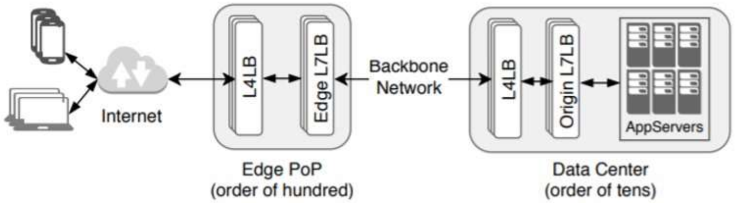
- Integrated platform
- Tofino, FPGA, and CPU in one box

Chassis platform

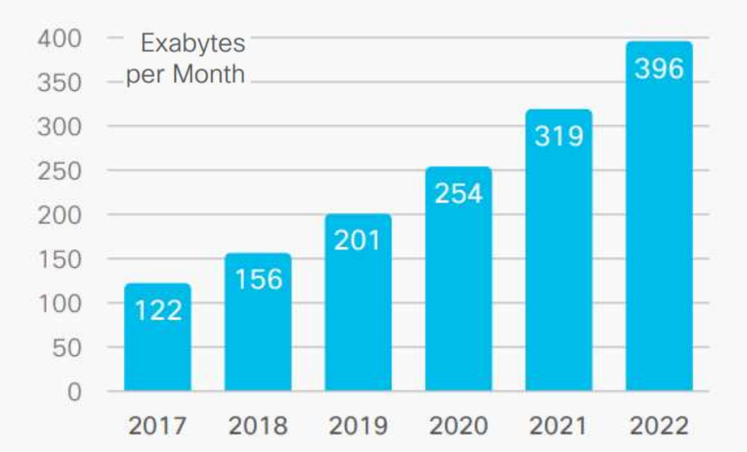
- Modules with Tofino, FPGA, and CPU

Application: L4 Server Load Balancing

- Load balancing is a key service in a data center to guarantee efficient utilization of the DC resources (compute & storage)
- Data center traffic is continuously growing, today DCs need to handle 10s or even 100s of Tbps of traffic
- How to handle the traffic growth while reducing TCO?

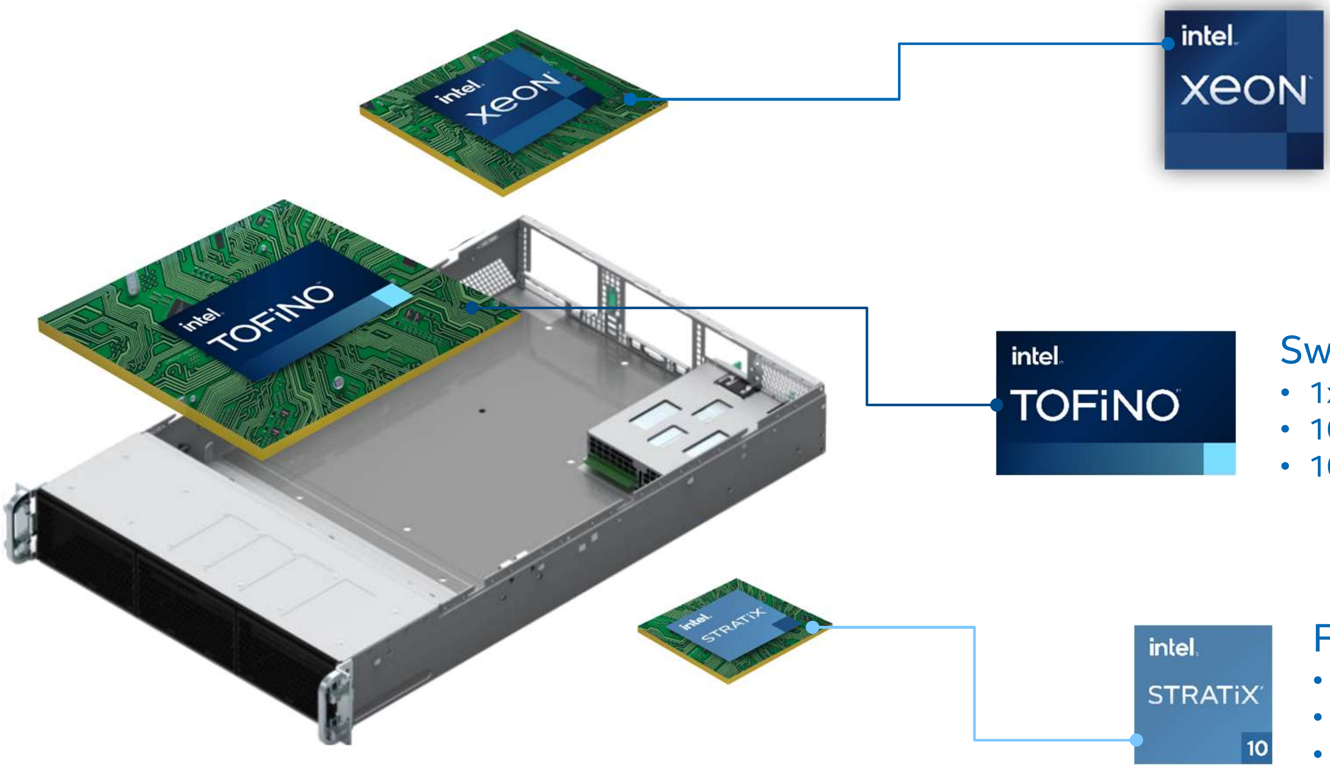


Source: Zero Downtime Release: Disruption-free Load Balancing of a Multi-Billion User Website



Source: Cisco VNI Global IP Traffic Forecast, 2017-2022

L4 Server Load-Balancer Accelerator Hardware Platform



CPU Module:

- 2x Intel® Xeon® Scalable Processors
- DDR4 Memory

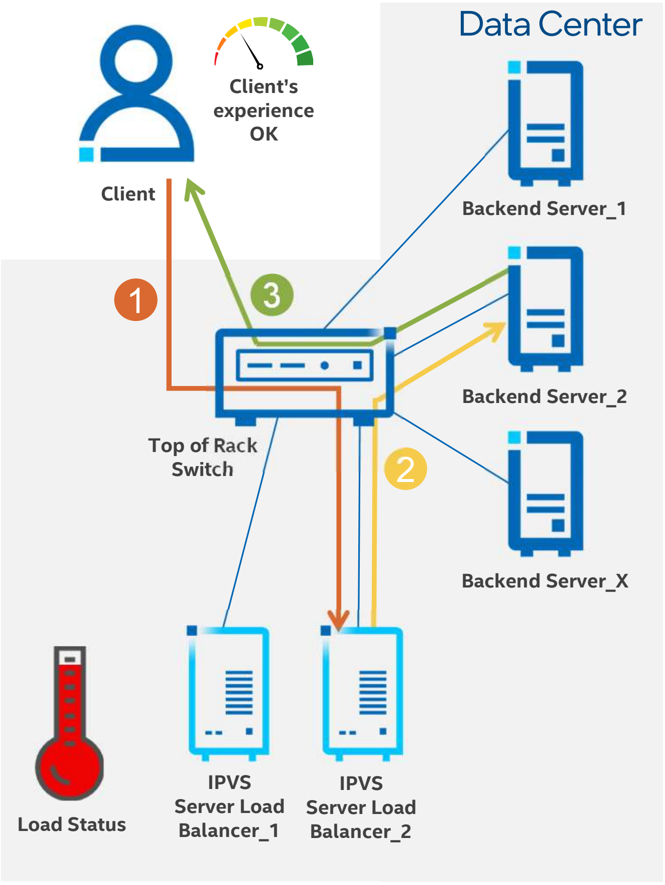
Switch Module:

- 1x Intel® Tofino™ Ethernet Switch ASIC
- 100GbE Ports for Data Plane
- 10GbE Ports for Control Plane

FPGA Module:

- Up to 4x Intel® Stratix® 10 FPGAs
- HBM2, DDR4 Dynamic Memory
- QDR Static Memory

Efficient High-Bandwidth Load Balancing



GOALS

- IPVS less utilized so available for other tasks
- Latency goes down
- Revenue goes up

1 Client's request

2 Client's request load balanced to the right backend server

3 Content delivered to the client

Server L4 LB Accelerator (SLBA) built on Tofino X switch server acting as connection cache

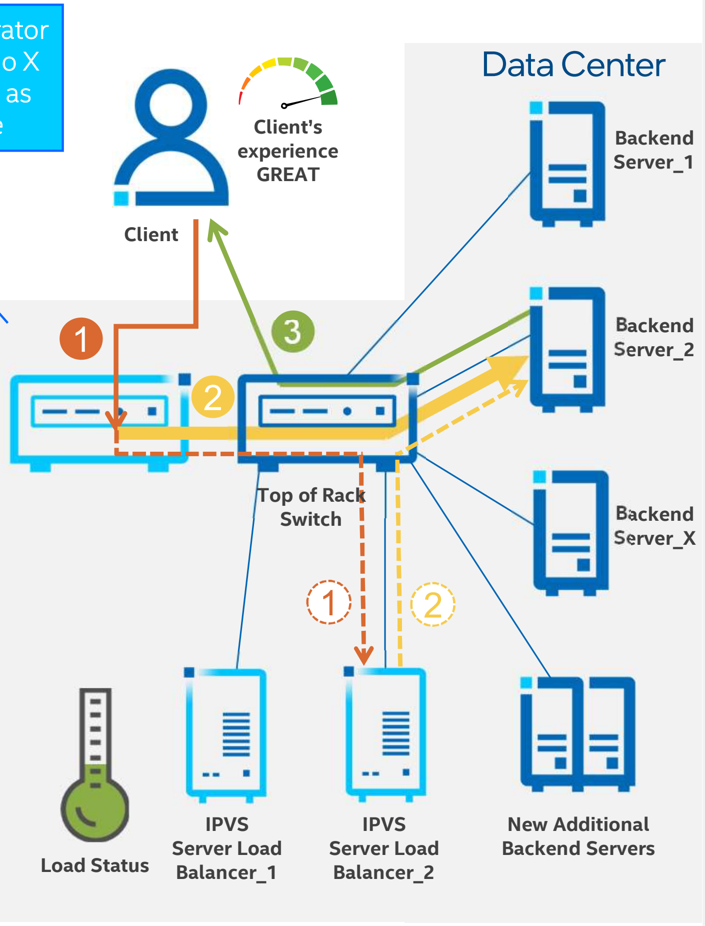
1 Client's request

1 Client's request forwarded to software SLB

2 Client's request load balanced to the right backend server

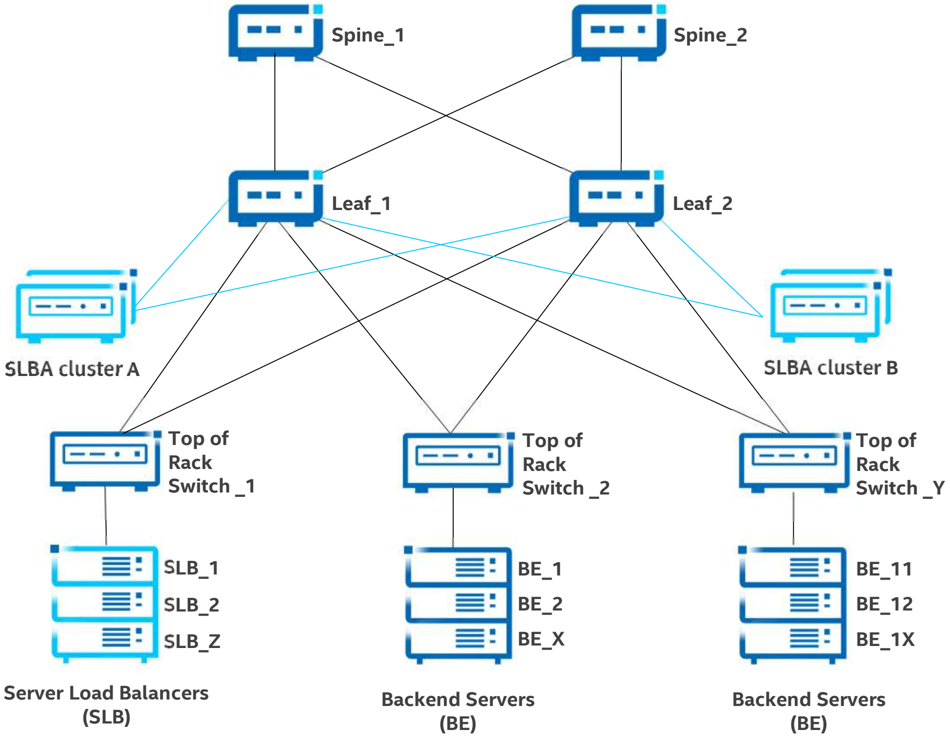
2 Accelerated fast path load balancing by SLBA

3 Content delivered to the client

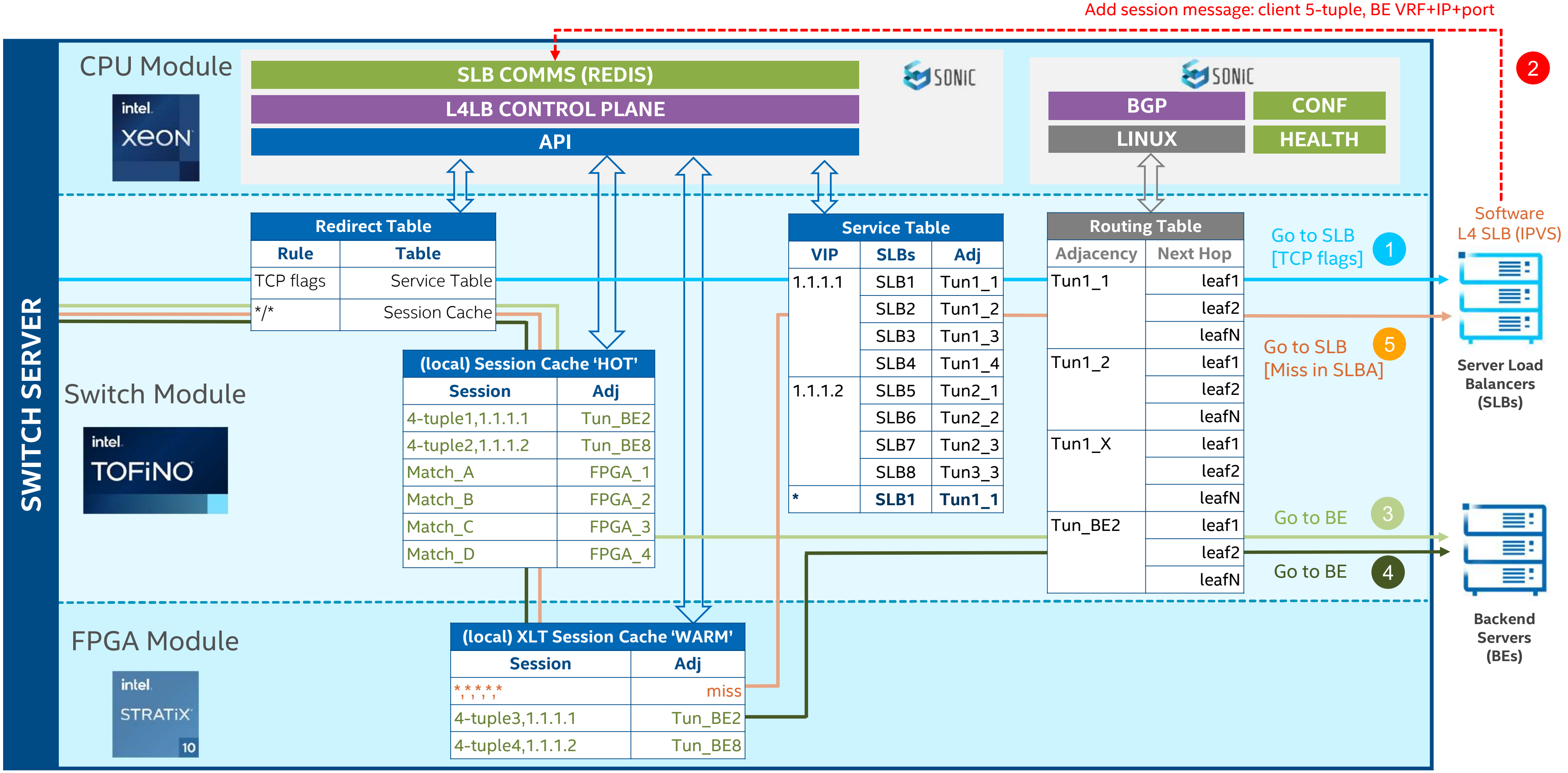

Data Center
Client
Client's experience GREAT
Backend Server_1
Backend Server_2
Backend Server_X
1
2
3
Top of Rack Switch
1
2
Load Status
IPVS Server Load Balancer_1
IPVS Server Load Balancer_2
New Additional Backend Servers

Disaggregated Control and Data Plane

- Server load-balancer accelerator can be connected to the leaf-spine Clos fabric
- Independent scaling of SLB Accelerator data plane (SLBA) and SLB software control plane (SLB servers) allows for redundancy and optimal ratio of data plane to control plane instances according to traffic patterns



L4 Server Load-Balancer Accelerator Data-Plane Architecture



Redirect Table

Rule	Table
TCP flags	Service Table
/	Session Cache

(local) Session Cache 'HOT'

Session	Adj
4-tuple1,1.1.1.1	Tun_BE2
4-tuple2,1.1.1.2	Tun_BE8
Match_A	FPGA_1
Match_B	FPGA_2
Match_C	FPGA_3
Match_D	FPGA_4

Service Table

VIP	SLBs	Adj
1.1.1.1	SLB1	Tun1_1
	SLB2	Tun1_2
	SLB3	Tun1_3
	SLB4	Tun1_4
1.1.1.2	SLB5	Tun2_1
	SLB6	Tun2_2
	SLB7	Tun2_3
	SLB8	Tun3_3
*****	**SLB1**	**Tun1_1**

Routing Table

Adjacency	Next Hop
Tun1_1	leaf1
	leaf2
	leafN
Tun1_2	leaf1
	leaf2
	leafN
Tun1_X	leaf1
	leaf2
	leafN
Tun_BE2	leaf1
	leaf2
	leafN

(local) XLT Session Cache 'WARM'

Session	Adj
,,*,*,*	miss
4-tuple3,1.1.1.1	Tun_BE2
4-tuple4,1.1.1.2	Tun_BE8

L4LB accelerator SW components

- Open-source software
- SONiC network operating system
  - Container based, lightweight micro-services
  - Fine-grained failure recovery and in-service upgrades with zero downtime
- L4LB accelerator API based on Redis
  - High-performance in-memory key-value store
  - Messages for adding services based on VIP + dest port and adding sessions mapping services to real backend servers

L4 Server Load Balancer Accelerator

Solution parameters	
Load balancing modes	Tunnelling (VxLAN, IPinIP), Direct return/routing[1],NAT[1], Full NAT[1]
Processing capacity	3.2Tbps[2]
Processing latency	<1us for hot cache hit, <2us for warm cache hit

[1] Supported. Not currently implemented.

[2] Total front panel capacity. Actual packet rate can be affected by corner-case traffic distributions.

Extra large table parameters	4x S10 GX FPGAs with 2xDDR4 per FPGA	4 S10 MX FPGAs with 8GBs of HBM2 per FPGA
Memory capacity	128 GBs	32 GBs
Table size[1]	256M session entries	128M session entries
Lookup rate[2]	Up to 600M lookups per second	Up to 4.8G lookups per second
Data path table update rate[3]	Up to 4M updates per second	Up to 32M updates per second
Cost[4]	$	$$

[1] Table size assumes 32B per entry and an optimization that trades off between capacity and lookup performance that leads to per-entry overhead.

[2] Lookup rate assumes even distribution of flows across the universe of possible entries.

[3] Achievable update rate under no or minimal load. The actual update rate during standard operation depends on the number of lookup requests processed by the extra-large table.

[4] Indicative comparison. Actual pricing depends on customer volume commitments.

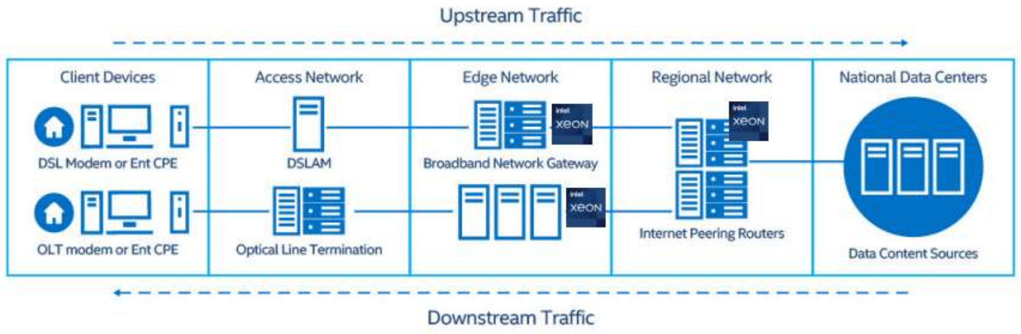
Related papers and resources

- P4 Practice at Baidu - Presentation for the 2021 P4 Workshop by Gang Cheng - YouTube
- Sailfish | Proceedings of the 2021 ACM SIGCOMM 2021 Conference
- Programmable network series (1): large-scale application and practice of programmable networks in Alibaba cloud (qq.com)

Application: Tbit Broadband Network Gateway

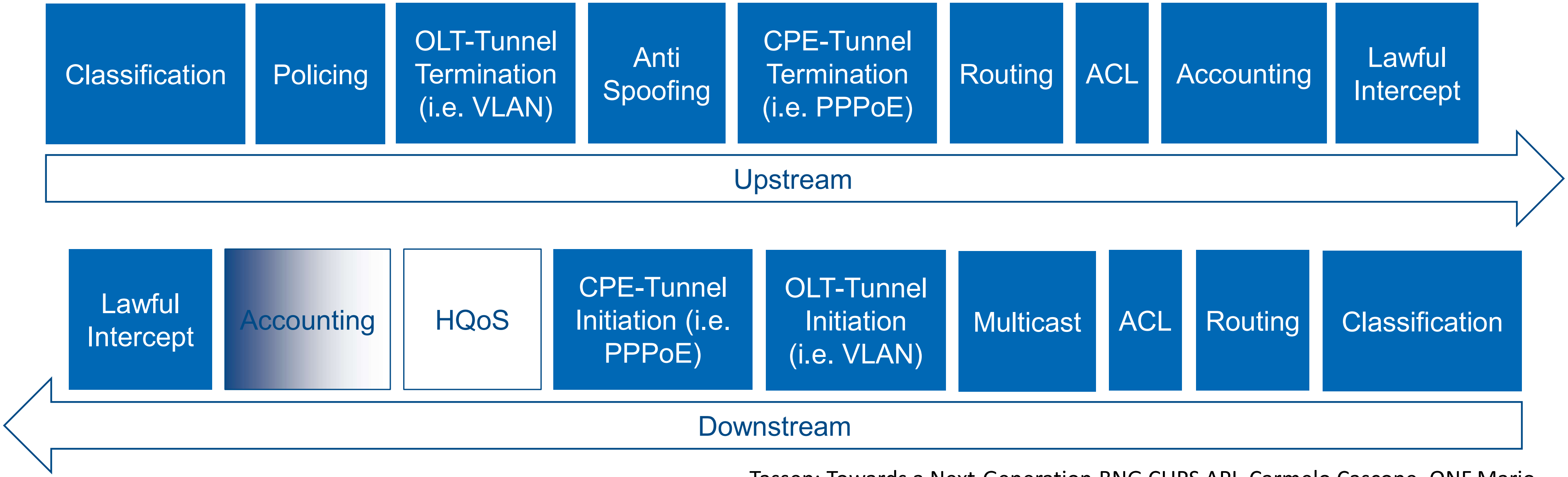
- Open Networking Foundation (ONF), Telecom Infra Project (TIP), Broadband Forum (BBF) & Open Compute Project (OCP) as “umbrella” projects
  - ONF Tassen: Towards a Next Generation BNG CUPS API
  - Telecom Infra Project - Open BNG technical requirements
- Key industry supporters
  - Deutsche Telecom, British Telecom, Telefonica, Vodafone, Telecom Italia

- Multiple presentations and papers by Deutsche Telecom et al:
  - Implementing a Programmable Service Edge - Update (ONF 2019)
  - OpenBNG: Central office network functions on programmable data plane hardware



Pipeline Overview (ONF/DT BNG)

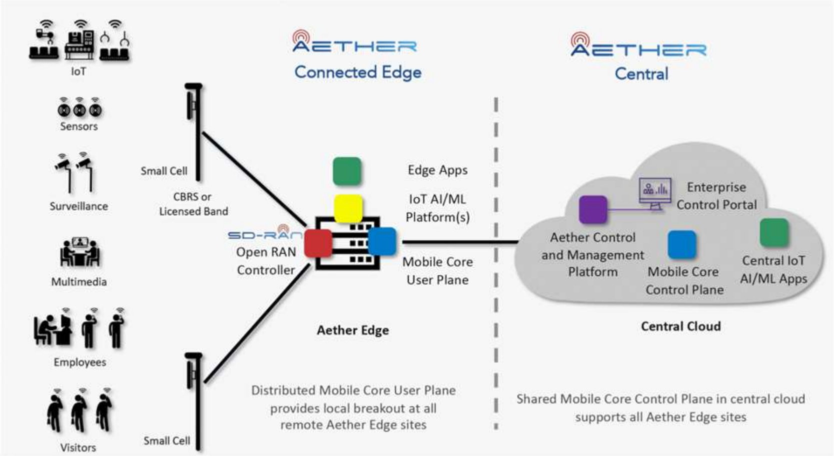
- Intel® FPGA (□): HQoS, Intel® Tofino™ IFP (■): Everything else



Tassen: Towards a Next-Generation BNG CUPS API, Carmelo Cascone, ONF Mario Kind, DT Craig Stevens, Dell, ONF Spotlight - Broadband, July 2020

PRONTO: DARPA funded $30M project

- Prontoproject.org PRONTO
- Leveraging ONF Aether project – open source 5G connected edge
- 5G UPF built on x86 + Tofino + FPGA


IoT
Sensors
Surveillance
Multimedia
Employees
Visitors
Small Cell
CBRS or Licensed Band
Small Cell
AETHER
Connected Edge
SD-RAN
Open RAN Controller
Edge Apps
IoT AI/ML Platform(s)
Mobile Core User Plane
Aether Edge
Distributed Mobile Core User Plane provides local breakout at all remote Aether Edge sites
AETHER
Central
Enterprise Control Portal
Aether Control and Management Platform
Mobile Core Control Plane
Central IoT AI/ML Apps
Central Cloud
Shared Mobile Core Control Plane in central cloud supports all Aether Edge sites

Recent papers

A P4-based 5G User Plane Function (princeton.edu)
User Plane Function Offloading in P4 switches for enhanced 5G Mobile Edge Computing (researchgate.net)

Innovation opportunities are limitless

- P4 programmability of Intel® Tofino™ IFP for packet processing
- Full flexibility of Intel® FPGA for extensions and augmentations of Tofino functionality

intel®