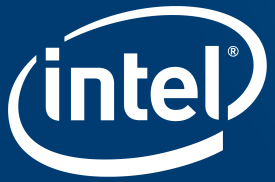


BAREFOOT
NETWORKS | an Intel company

DOMAIN SPECIFC ARCHITECTURE COMES TO NETWORKING

Prem Jonnalagadda

CASTOR Software Days | October 15, 2019 | KTH, Stockholm, Sweden.

Notices and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

No product or component can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks .

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/benchmarks .

Intel® Advanced Vector Extensions (Intel® AVX) provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

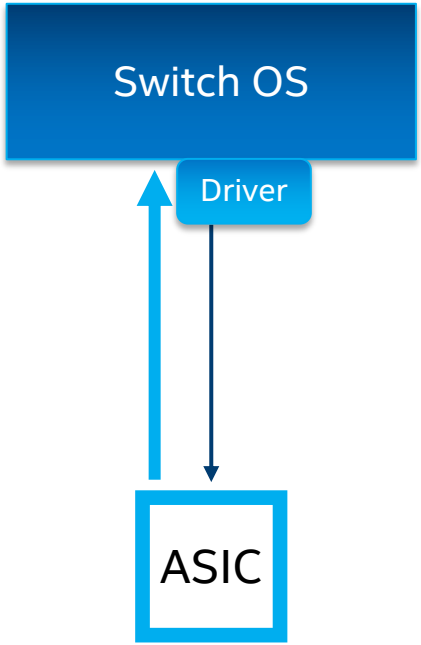
Intel, the Intel logo, Intel Nervana, Intel Optane, Intel Xeon, Intel Xeon Phi, Stratix and the Stratix logo are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as property of others.

Problem: Network Systems are Built “Bottom-up”

- SDN tackled control plane
- Disaggregation added flexibility
- Data planes have not kept up!

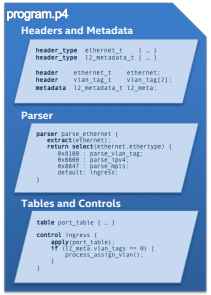
“This is how I process packets …”



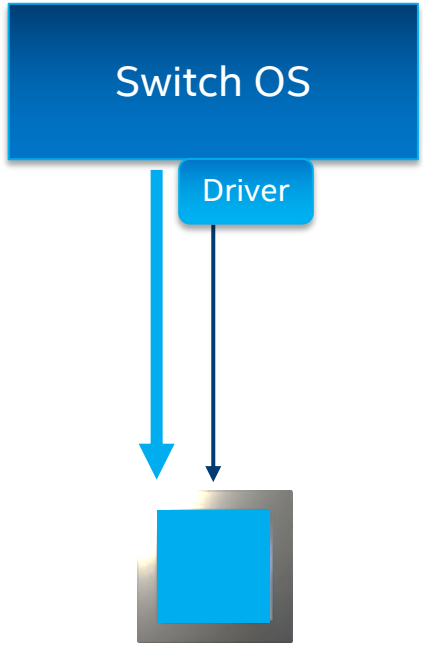
Fixed-function switch

Network Systems need to be Programmed “Top-down”

“This is precisely how you must process packets”

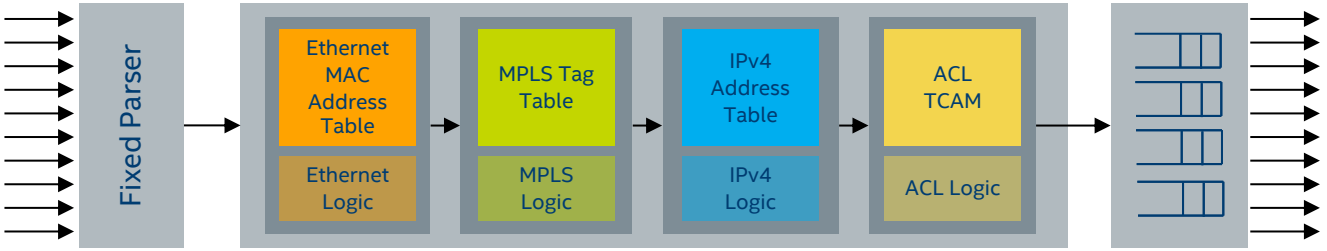


CONSEQUENCE:
Vendor-driven replaced by user-driven

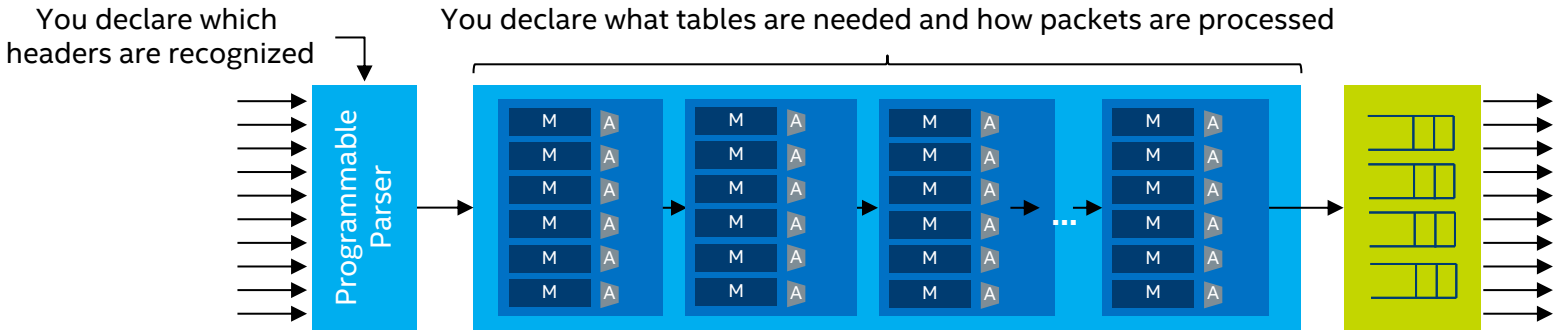


Programmable switch

Fixed vs. Programmable Packet Processing

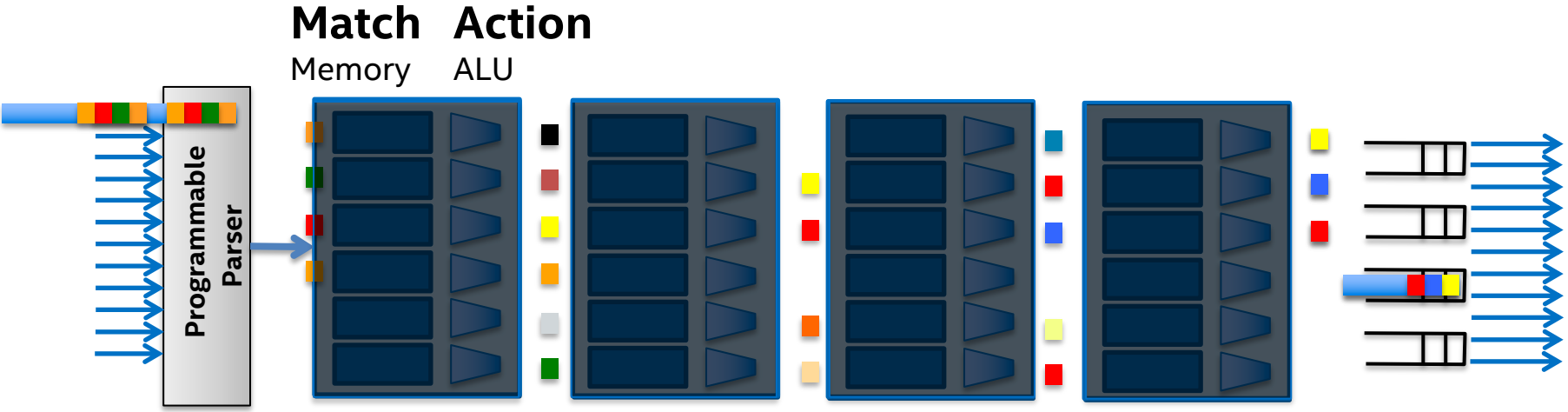


Fixed Pipeline: features and table-sizes are baked in at design time



Programmable Pipeline: all stages identical, customer-defined match-action logic

Protocol Independent Switch Architecture (PISA)



Programmable Switch Approach

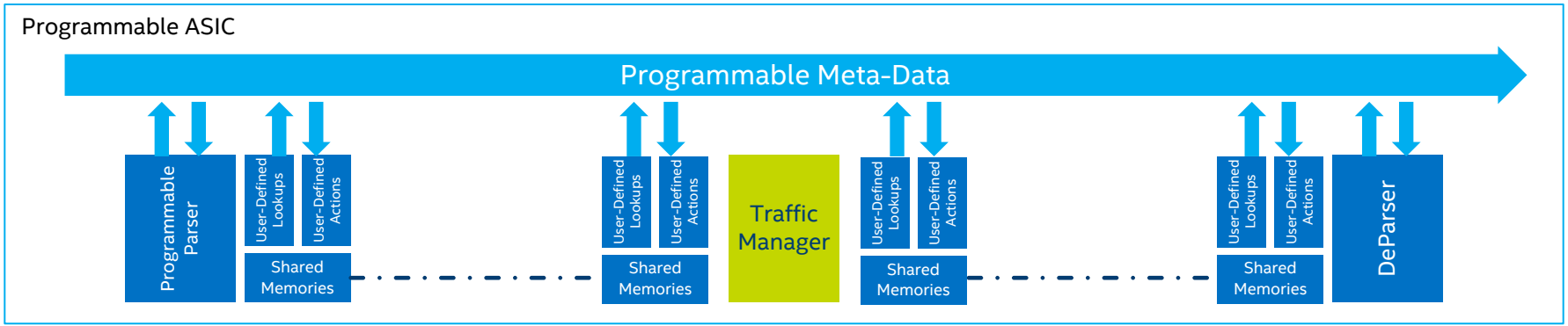
1 Protocol Authoring

2 Compile

3 Configure

my_switch.P4

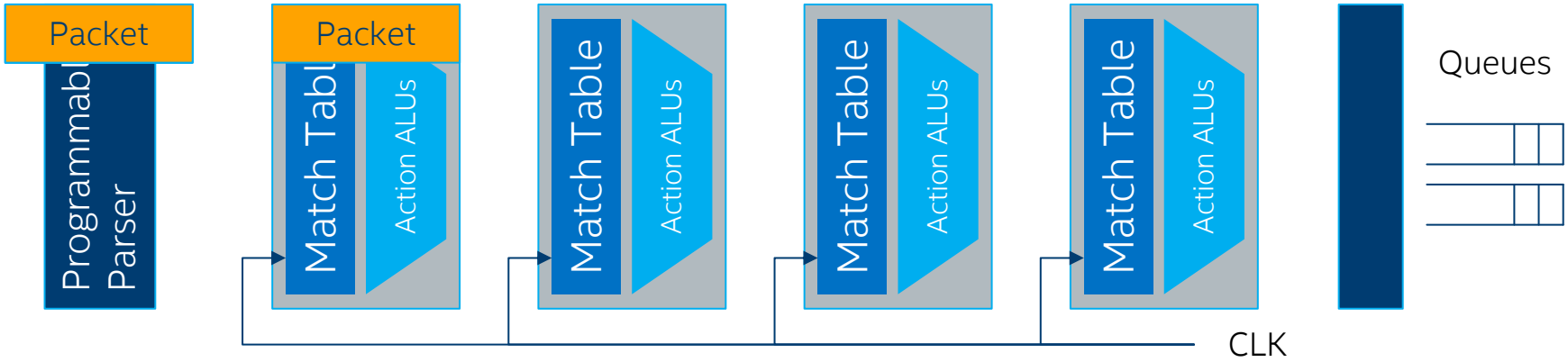
Auto Generated Run-time API



Device Does Not Understand Any Protocols Until it Gets Programmed

Logical Data-plane View
(your P4* program)

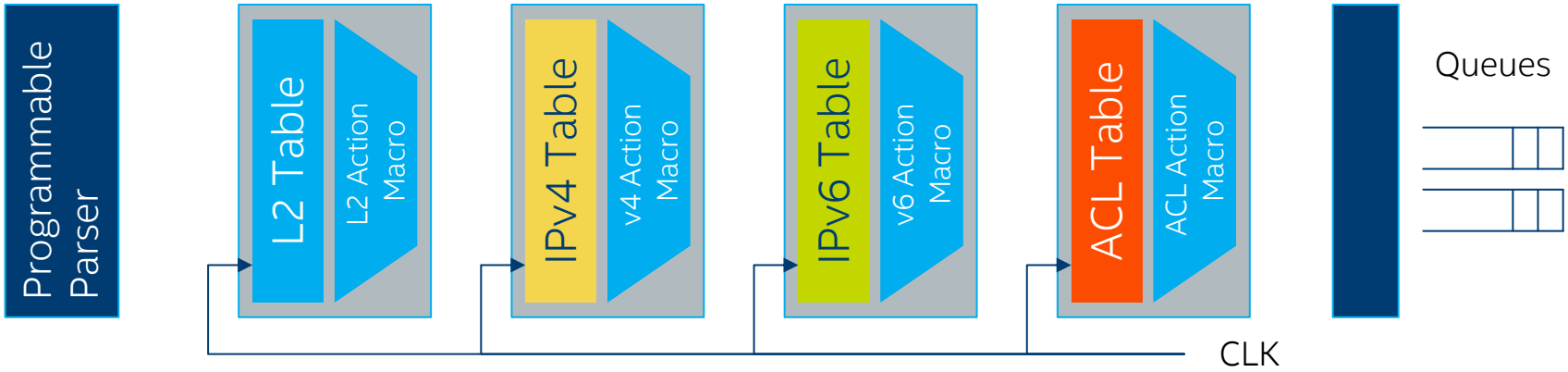
Switch Pipeline



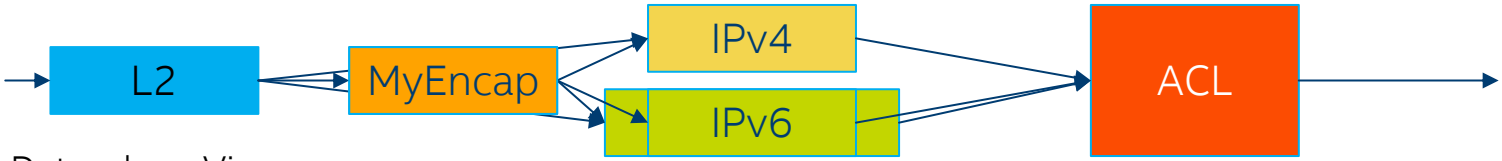
Mapping Logical Data-Plane Design to Physical Resources

Logical Data-plane View
(your P4* program)

Switch Pipeline

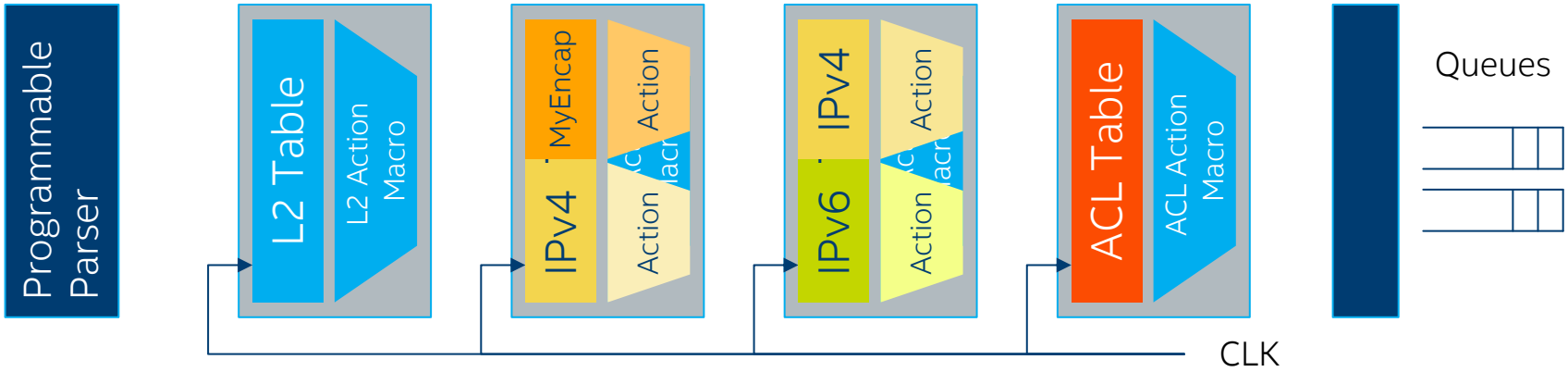


Re-Program in the Field

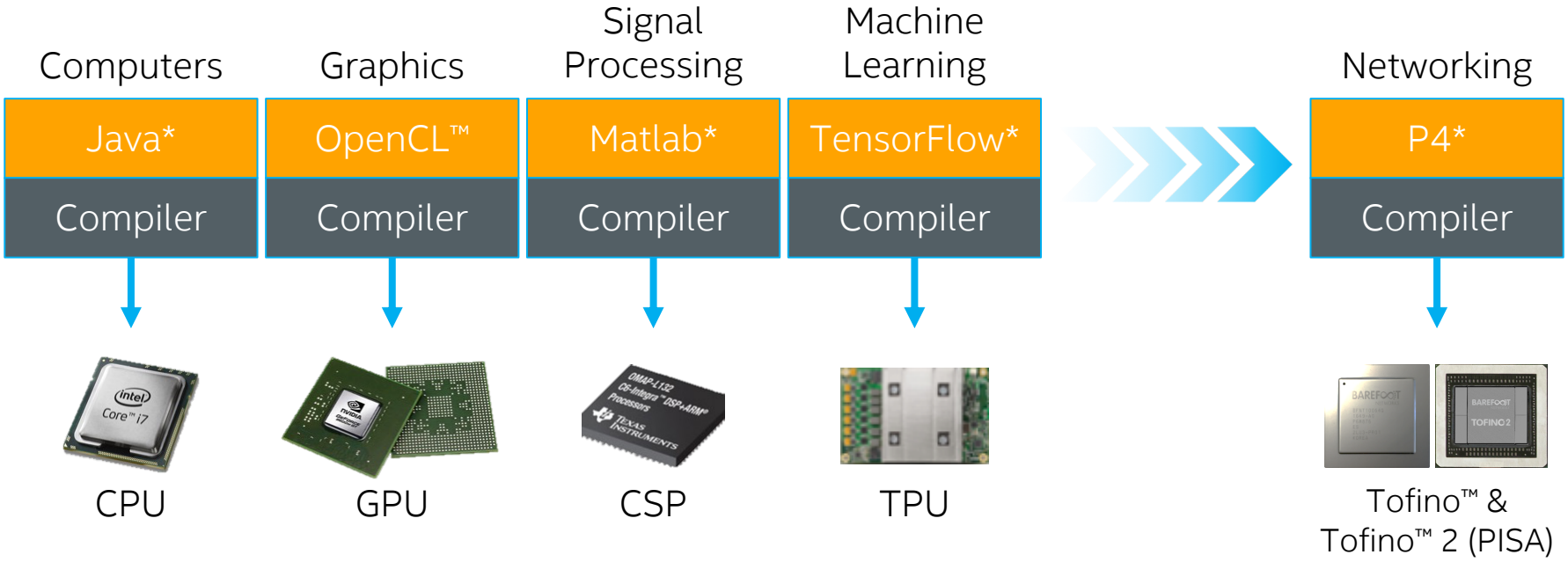


Logical Data-plane View
(your P4* program)

Switch Pipeline



General Industry Trend:
Rise of the Domain-Specific Architectures (DSAs)



P4

Programming Protocol-Independent Packet Processors

P4 Growth and Momentum

- Project under ONF
- No Membership Fee
- Apache* 2.0 License
- 100+ Members and Growing

P4* Paper

First P4 paper published

Dec 2013

Consortium Launched

P4.org goes live!

Sep 2014

First P4 Workshop

- 1st P4 event
- ~60 attendees
- **~15 member orgs**

June 2015

Growing Momentum

- First Dev. Day and bootcamp
- **~35 member orgs**
- **In-band Network Telemetry**

Q4 2015

P4 Products

- Broad range of products supporting P4-SmartNICs, FPGAs, NPUs, ASICs, Soft Switches
- **P4Runtime launched**

2016-17

Industry-wide adoption

- P4.org joins ONF and Linux Foundation*
- **~100 Member orgs**
- Growing developer community
- Members spanning OEM, ODM, Service Providers, Cloud, Enterprise, ...

1H 2018

P4 is DSL for Networking

- Global growth support and adoption
- **2000+ developers**
- 100+ papers and apps
- **5 active working groups**
- **3000+ commits**
- **200+ contributors**
- P4 workshops **US, UK, Netherlands, China, Japan, Korea, ...**

Today

Define Mac field

```
header ethernet_t {
   bit<48>    dstAddr;
   bit<48>    srcAddr;
   bit<16>    etherType;
}
```

Define table matching on mac field

```
table mac {
   key = {
      ingress_metadata.bd : exact;
      l2_metadata.lkp_mac_da : exact;
   }
   actions = {
      dmac_hit;
      dmac_miss;
      dmac_redirect_to_cpu;
   }
   default_action = dmac_miss;
   size =MAC_TABLE_SIZE;
}
```

Define table actions

```
action dmac_hit(bit<16> ifindex, bit<16> port_lag_index) {
   ingress_metadata.egress_ifindex = ifindex;
   ingress_metadata.egress_port_lag_index = port_lag_index;
   l2_metadata.same_if_check = l2_metadata.same_if_check, ^ ifindex;
}
```

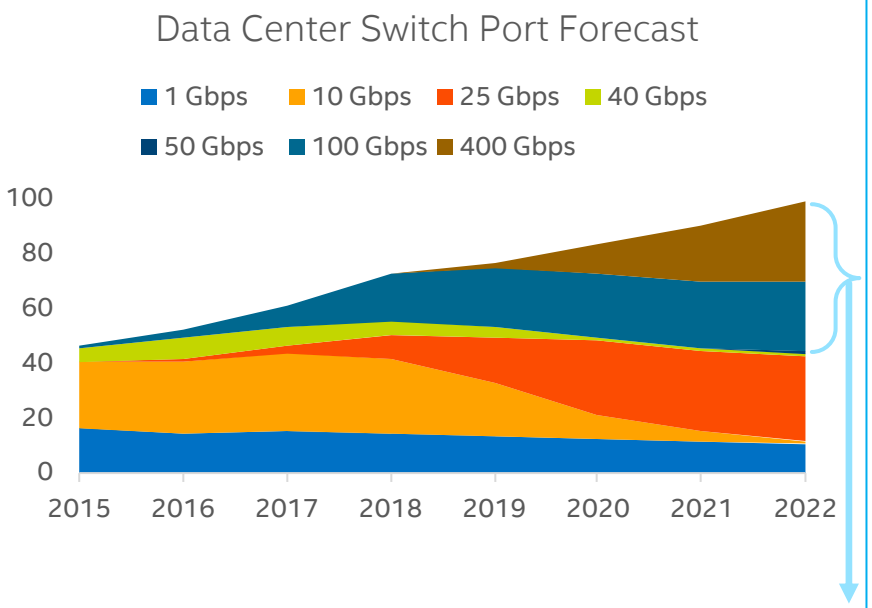
- Open source
- Reconfigurable
- Protocol independent
- Target independent
- Vendor independent

- Open Source
- p4 compiler -> p4 Runtime -> switch

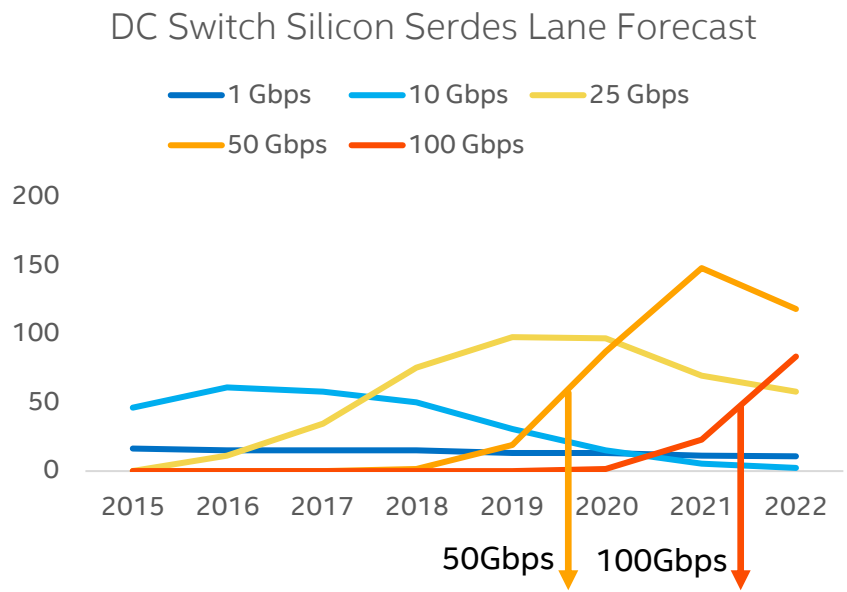
TOFINO AND TOFINO 2

P4-programmable Ethernet Switch ASICs

Hyperscale Spending to Drive Strong Growth



100G and 400G ports to comprise the bulk of data center switch spending



50G and 100G SerDes Silicon expected to dominate data center switch segment

Source: 650 Group 2018 Forecast

Barefoot Tofino™ 2

Leading with performance and programmability

Industry-leading Process Node
- 7nm technology
- Chiplet Architecture

Highest Bandwidth
- 12.8Tbps with 50G SerDes

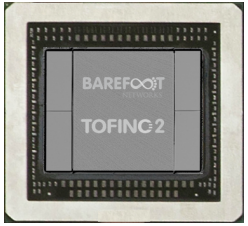
Highest Radix
- 256x10/25/50GE, 128x100GE, 32x400GE

Lower Power
- Up to 50% better performance per watt

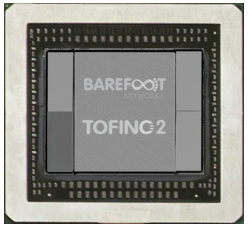
Modular Chip Architecture
- Disaggregated silicon with upgradability to 100G SerDes and Silicon Photonics

Field-proven PISA Architecture
- In production at several customers including Tier 1 OEMs and MSDCs

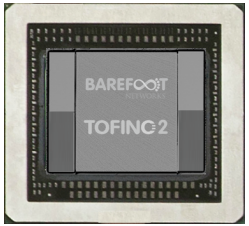
P4 Programmability
- Leverage 2000+ P4 developer community and thriving ecosystem



12.8 Tbps



8.0 Tb/s



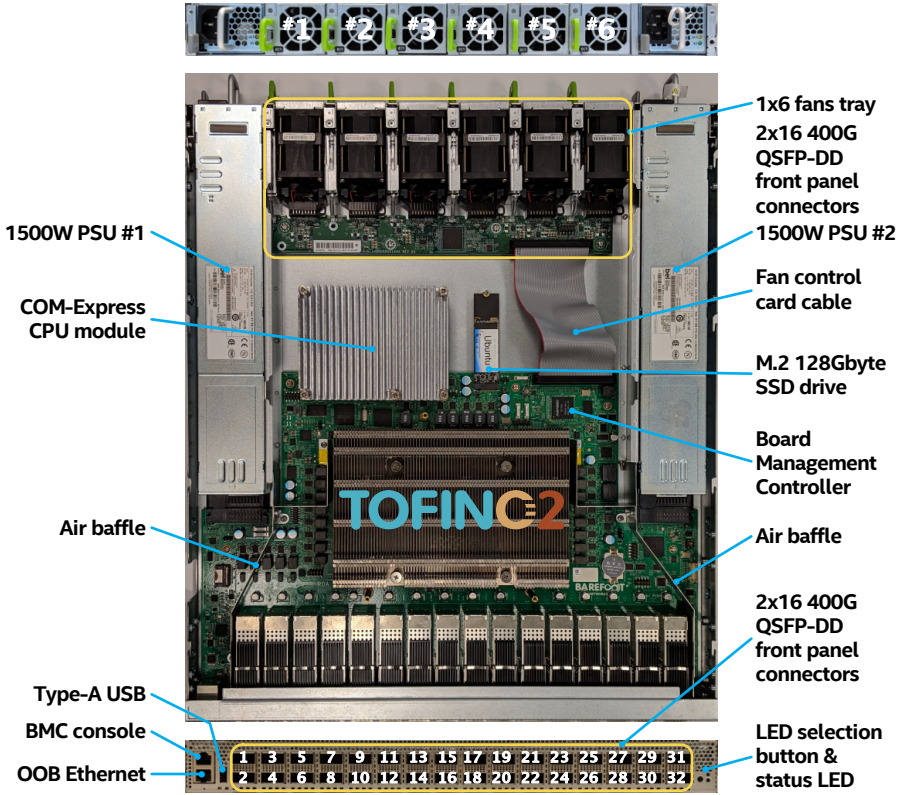
6.4 Tb/s

SAMPLING SINCE Q2 '19

Tofino 2 1RU Switch

System Specifications

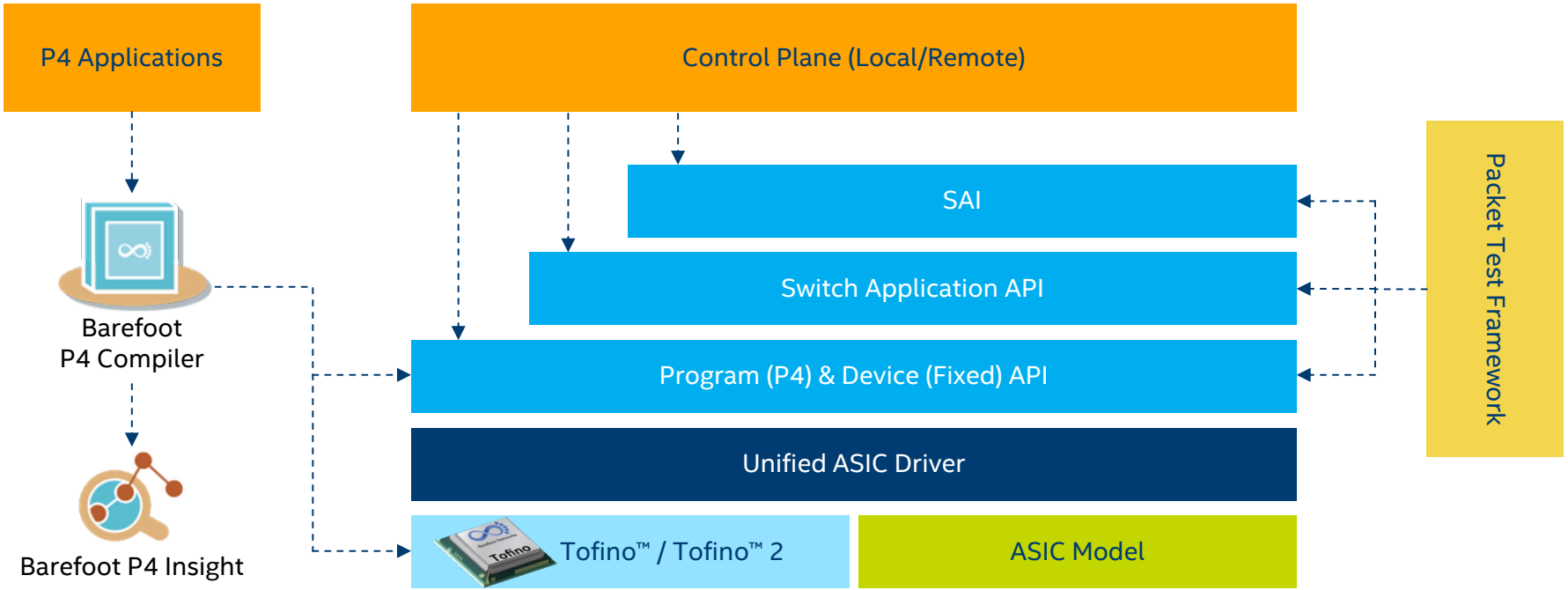
- 32 QSFP-DD ports using 2x1 stacked QSFP-DD cages
- Intel® Xeon® D processor COM-Express CPU module



BAREFOOT SOFTWARE

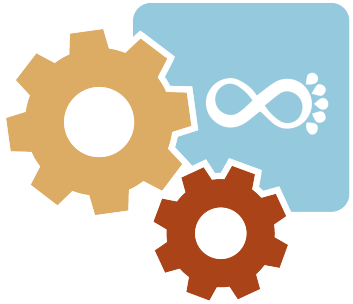
Software Development Environment including Compilers, Drivers and Debuggers.

Control Plane Integration with Programmable Data Plane



Barefoot P4 Compiler Benefits

Leveraging years of data plane programming experience



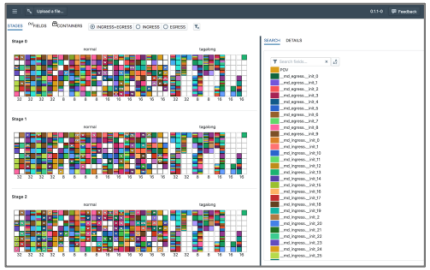
- Second generation
- Available to end-users with open-sourced front-end!
- Significant compilation time improvement (~10x)
- Improved hardware resource allocation
- P4-16 Support

Barefoot P4 Insight

Dynamic visualization of P4 program as mapped to Tofino™ / Tofino™ 2

Dashboard

Resource Utilization



PHV Utilization

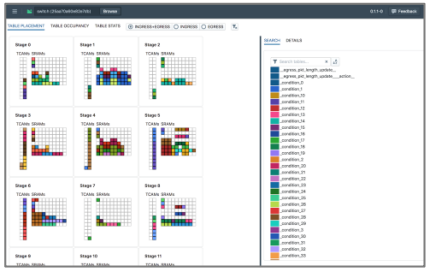


Table Placement

...

P4 Program

P4 Compiler

Tofino™ / Tofino™ 2

BAREFOOT DISAGGREGATED ECOSYSTEM

White box switches and Network Operating Systems

Barefoot Baremetal Switch Ecosystem

SWITCH OS / FABRIC SOLUTIONS

ipinfusion™ · SONiC · STRATUM · NoviFlow · stateless · kaloom

WHITE BOX HARDWARE (ODMS)

Edge-corE NETWORKS · STORDIS · 英業達 Inventec · 恒为科技 EmbedWay · ufiSpace · DNI · Celestica™ · WNC Wistron NeWeb Corp.

BAREFOOT ASICS

6.4 Tb/s

3.2 Tb/s

2.4 Tb/s

2.0 Tb/s

1.8 Tb/s

12.8 Tb/s

8.0 Tb/s

6.4 Tb/s

DEMO TIME!

WHAT CAN YOU DO WITH ALL OF THIS?

Innovation in networking like never before!

Beautiful New Ideas!

Graphics

OpenCL™

Compiler

GPU & GPGPU

- More...
- ML/DL - Inference
- ML/DL - Training
- Self-driving Vehicles
- HPC & Big Data
- Virtual Reality
- Augmented Reality
- Gaming/Professional Graphics
- 2D/3D Graphics
- Basic Graphics

Networking

P4

Compiler

Tofino™ (PISA)

- More...
- Storage/Memory Interconnect
- ML/DL Interconnect
- DNS Cache
- Network Packet Broker
- Firewall & DDos
- Layer 4 Load Balancer
- Telemetry Analytics
- Enterprise/Telco/SP Switching
- Cloud Switching

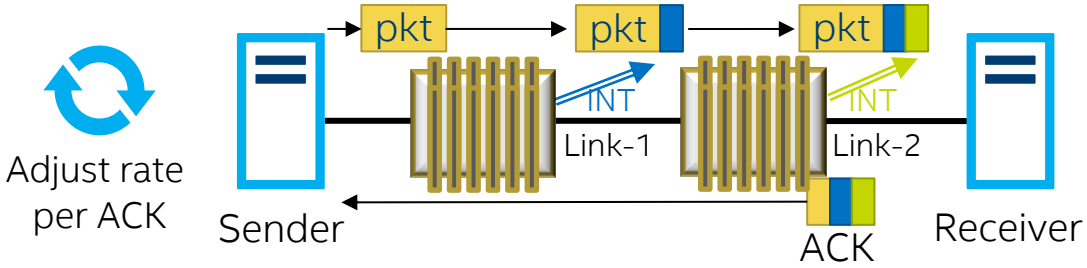
HIGH-PRECISION CONGESTION CONTROL

HPCC: INT-based High Precision Congestion Control

Published at SIGCOMM 2019, by Alibaba, Harvard, U of Cambridge, and MIT

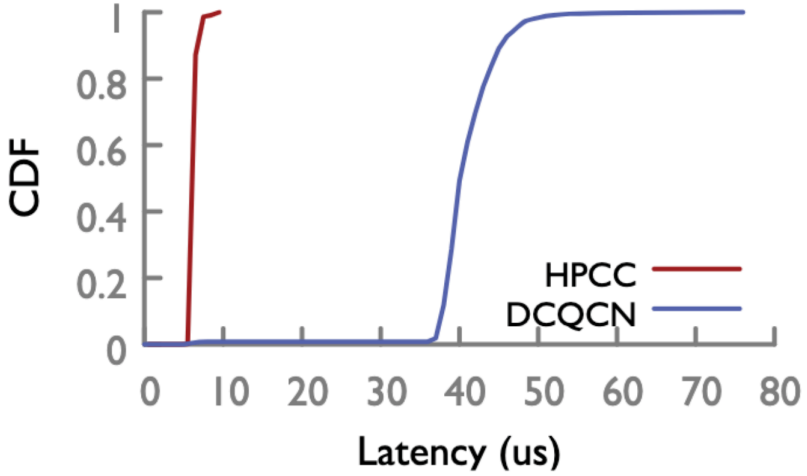
Using INT as explicit and precise feedback

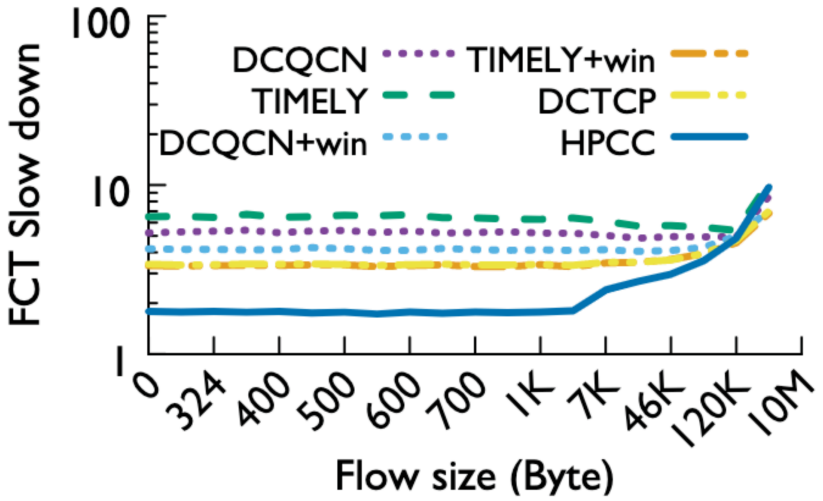
- Very fast convergence
- Near-zero queue
- Few parameters



Key Benefits of HPCC:

Very low latency and very high throughput at the same time

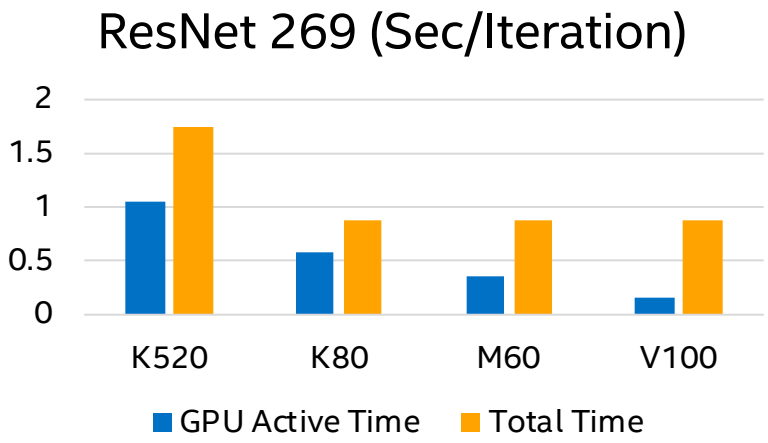




ML TRAINING ACCELERATION IN FABRIC

Accelerate Training in Machine Learning

- Training over huge data requires distributed processing
- With faster workers, sharing learned parameters becomes a bottleneck



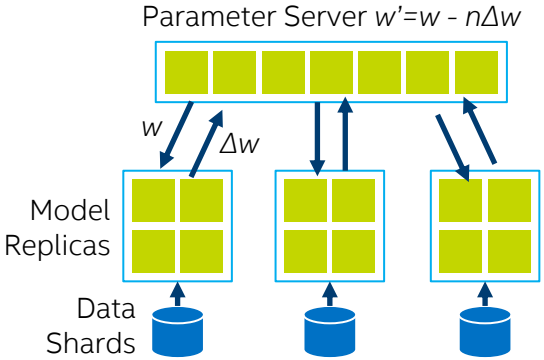
* Liang Luo et.al., Parameter Hub, SysML 2018

Distributed Deep Neural Networks

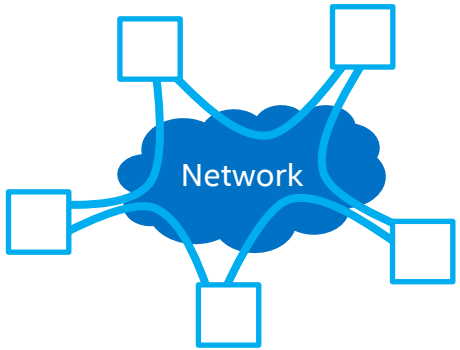
Workers repeat two steps:

1. Update local parameters based on data
2. Share parameters to compute aggregate values

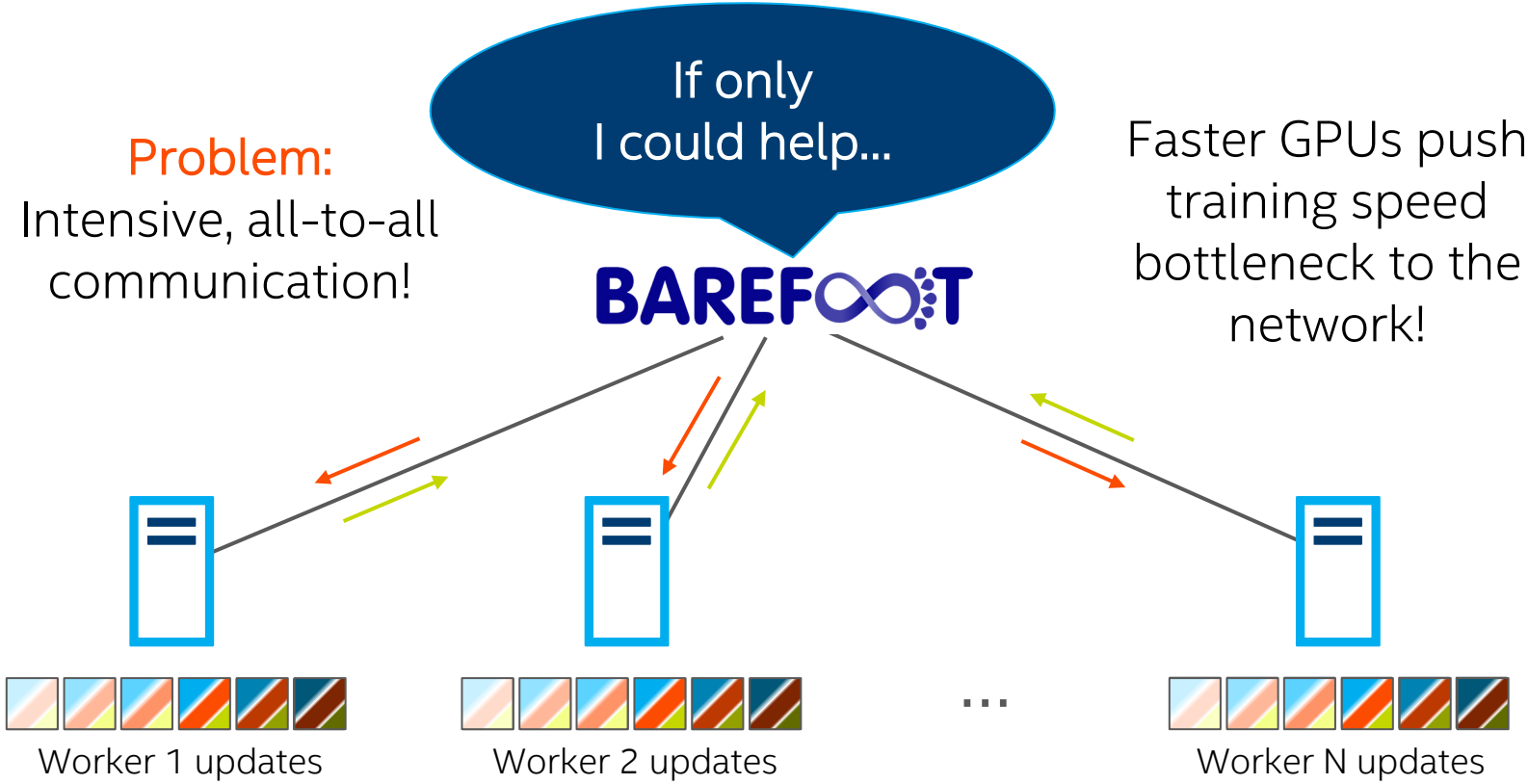
"Parameter server is the bottleneck"



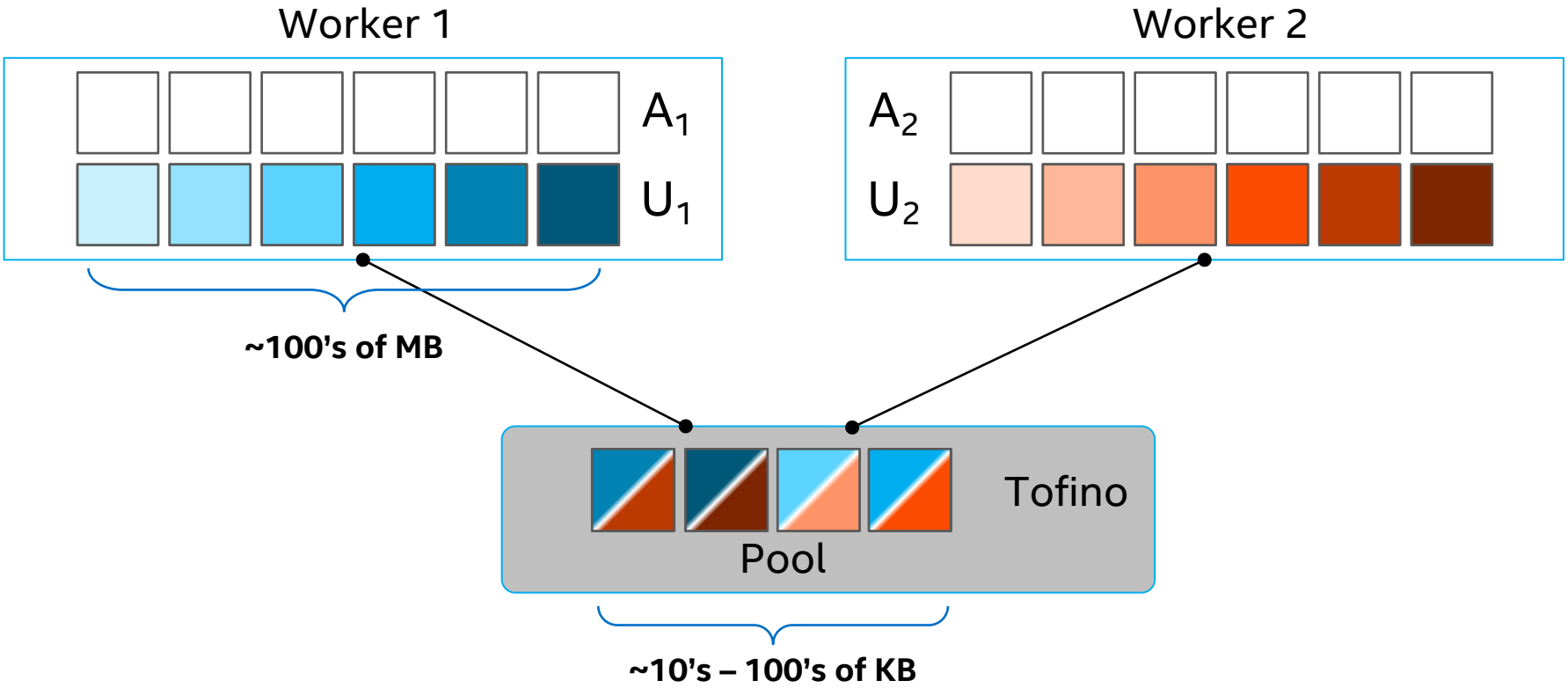
"Ring over workers increases latency"



Aggregation is Communication-intensive

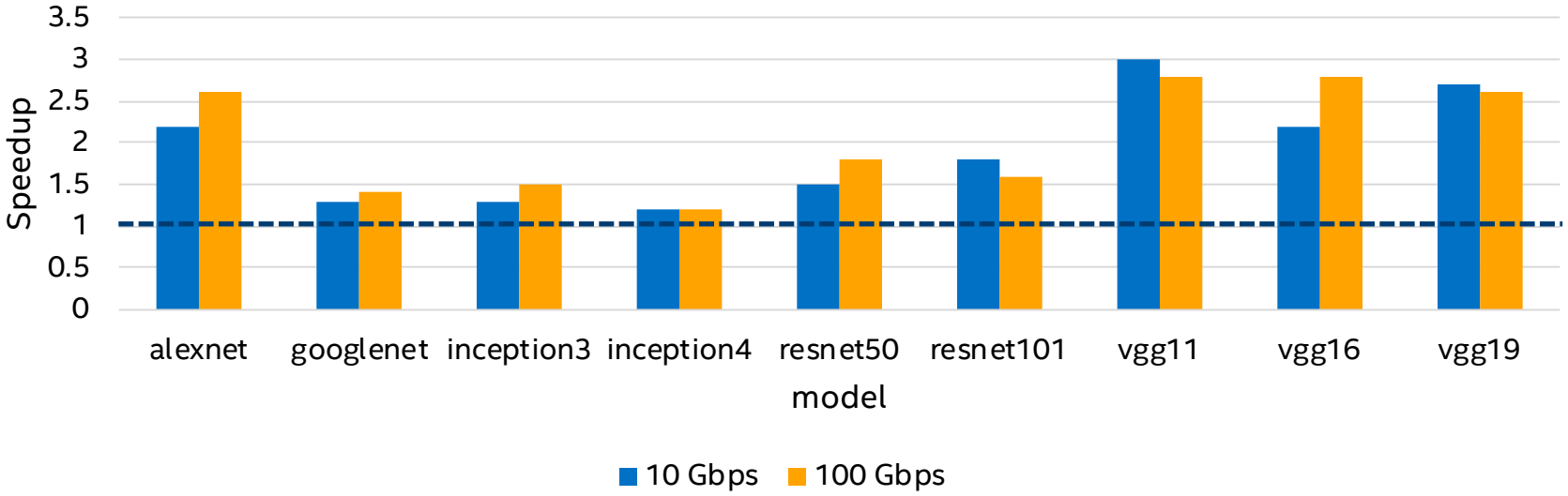


Streaming Aggregation with a Pool



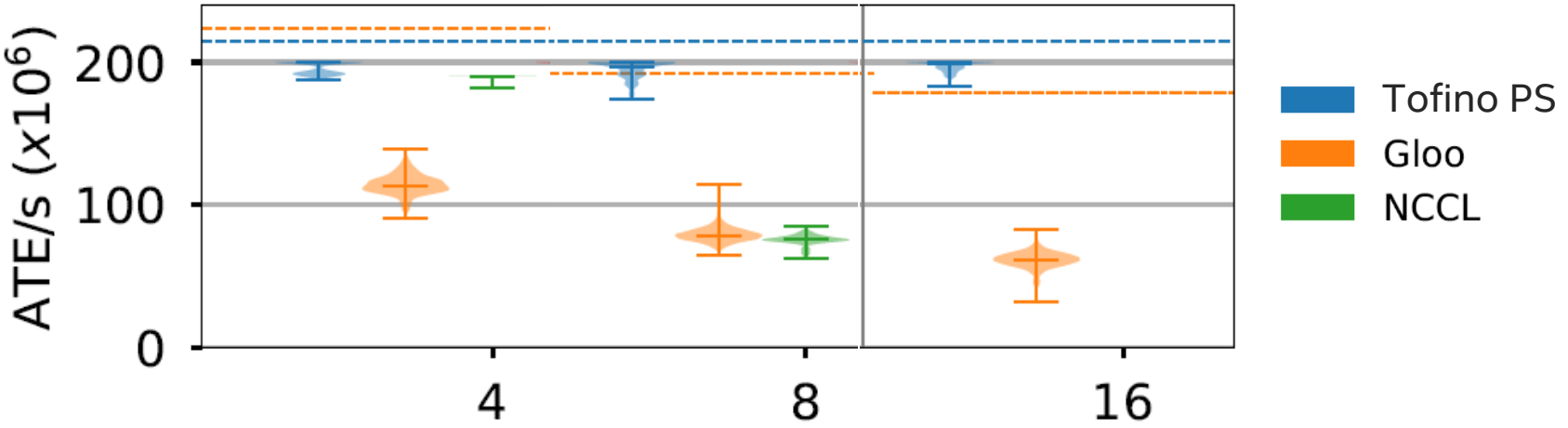
How Much Faster is Tofino™-based Weight Aggregation?

Tofino™-based aggregation provides a speedup from 20% to 300% compared to Tensorflow/NCCL (with direct GPU memory access)



How does Tofino™-based Aggregation Scale with the Number of Workers?

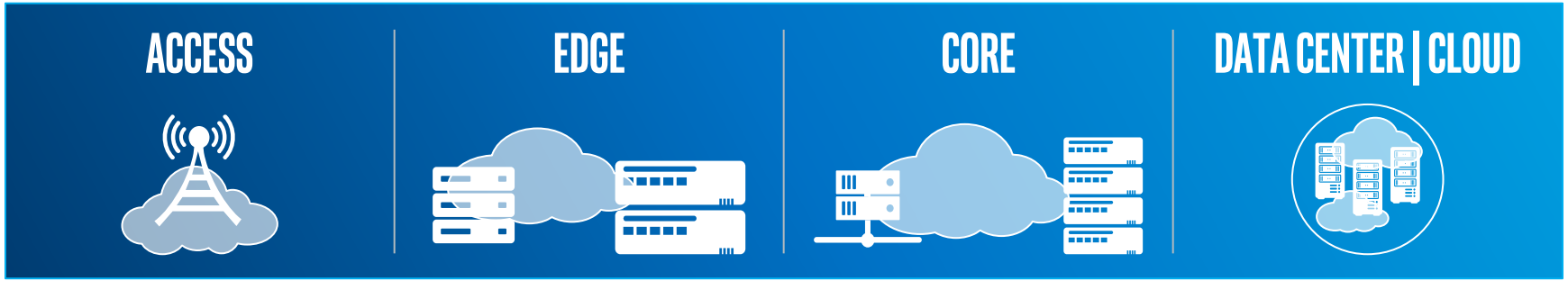
Tofino™ parameter server performance does ***not*** depend on the number of workers



When You Need Adapt Your Network…

1. Equipment vendor can just send you a software upgrade
2. New forwarding features take days to develop
3. By then, you don't have to figure out a hack to work around it
4. Eventually, when the upgrade is available,
   - It cleanly solves your problem, or
   - You don't need a complete hardware upgrade at huge expense.

End-to-End Fabric with Programmable Components



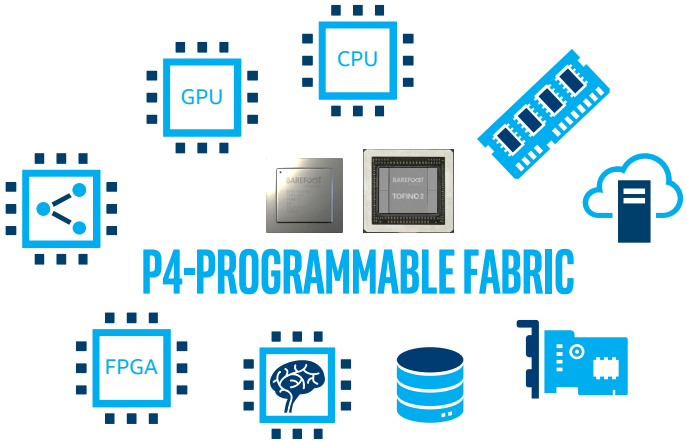
ACCELERATORS

FPGAS

MEMORY

CONNECTIVITY

SOFTWARE

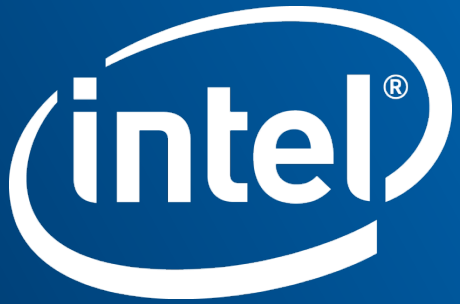


The Data-Centric World

Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

THANK YOU!

Prem.Jonnalagadda@Intel.com


intel®