White Paper

Data Center | Software-Defined Networking

September 2022

Let Your Networks Soar with Intel® Tofino™ Expandable Architecture

Achieve multi-terabit traffic rates with Intel® Tofino™ Expandable Architecture, which combines fully programmable Intel® Tofino™ Intelligent Fabric Processors (IFPs) and Intel® CPUs with accelerators such as Intel® FPGAs and Intel® IPUs

Author

Petr Kastovsky
Product Marketing Engineer
Switch and Fabric Group

Table of Contents

Contributors

Bert Klaps
Senior Systems Engineer
Programmable Solutions Group

John Eastwood
Product Marketing Analyst
Switch and Fabric Group

Executive Summary

To help solve today's networking challenges, Intel developed Intel® Tofino™ Expandable Architecture. This new offering lets network owners and operators take advantage of the P4 programmability of Intel® Tofino™ Intelligent Fabric Processors (Intel® Tofino™ IFPs) for packet processing, while employing the power of Intel® Xeon® processors and the acceleration capabilities of Intel® FPGAs and/or Intel® Infrastructure Processing Units (Intel® IPUs) to extend and augment Intel Tofino IFP functionality.

Intel Tofino Expandable Architecture increases table and buffer capacity by two orders of magnitude[1]—compared to a stand-alone data center-focused switch application-specific integrated circuit (ASIC)—and supports network transformation through the ability to add new types of network functions.

Massive performance and scale, flexibility, open standards and reduced total cost of ownership are all critical factors in today's constantly growing and evolving networks. Intel Tofino Expandable Architecture can help provide all of these—today and into the future.

Acronyms

ASIC	application-specific integrated circuit
FPGA	field programmable gate array
IFP	Intelligent Fabric Processor
IPU	Infrastructure Processing Unit
SLBA	server load-balancing acceleration

Networking Challenges

A perfect storm of digitization trends creates an urgent need for cloud service providers (CSPs), communications service providers (CoSPs) and even enterprises to transform their networks to handle more traffic, scale easily and support new networking functions. In fact, nearly everything is going digital at an astounding rate:

- **Data explosion.** The total amount of data created, captured copied and consumed globally is forecasted to nearly triple from 64.2 zettabytes (ZB) in 2020 to 180 ZB in 2025.[2] Web-scale services are accessed by billions of users worldwide every day, such as at Meta, Google and Twitter.[3]
- **Burgeoning mobile and work-from-home workers.** The COVID-19 pandemic accelerated the shift from work-on-premises to work-from-anywhere.[4] Whether it's the local coffee shop, the home office or an airport lobby, workers expect to be able to connect to the network at any time.
- **Increasing adoption of hybrid and multicloud.** 94% of enterprises now use a cloud service, which means they rely on networks to run their workloads.[5]
- **Smartphone and 5G growth.** The global monthly average usage per smartphone was 12 GB at the end of 2021 and is forecast to reach 40 GB by the end of 2027.[6] This trend is partially driven by online video consumption, which in 2022 is expected to make up more than 82% of all consumer internet traffic—15 times higher than it was in 2017.[7]
- **Customer expectations.** Today's consumers of networking services—whether on Instagram or accessing a corporate virtual private network (VPN)—expect fast, reliable connectivity.[8] Buffering, jitter and dropped calls are unacceptable and can lead to customer churn and worker frustration.

The need for more and faster bandwidth is clear. But simply scaling horizontally by deploying more servers is unsustainable. For one thing, the costs for server procurement, management and maintenance are high. Furthermore, traffic speed tends to increase faster than per-core performance—leading to oversubscription and extra headroom when sizing systems, which further increases costs.

An additional challenge for network owners and operators is that today's network requirements cannot be met using traditional fixed-function networking devices, for two main reasons. First, device design and features are determined by device architects, not by network owners and operators. Device design is not evolving as fast as the network itself. Second, hard-coded table and buffer sizes may not support the session scale required by modern network traffic flows. In short, fixed-function network devices limit network capabilities and scalability.

To surmount these significant challenges, networks must be able to scale to meet growing traffic volume as well as become smarter to support network functions virtualization (NFV), cloud-native microservices, distributed workloads and data.

Software-Defined Networking and Full Programmability Enable a More Flexible Network

To enable this smarter, scalable network, network owners and operators are turning to software-defined networking (SDN) and programmable switch ASICs that use software to define the actual packet processing (i.e., supported protocols and headers, table sizes and packet-processing functions). These ASICs enable the pipeline to be fully workload-optimized.

The Intel® Tofino™ Intelligent Fabric Processor (Intel® Tofino™ IFP) is an example of just such a fully programmable switch. The Intel Tofino IFP is programmed using the open-source P4 language[9], and supports traditional switching use cases like leaf/spine and top-of-rack switching as well as a broad range of additional networking use cases across cloud, telco and enterprise (see Figure 1).

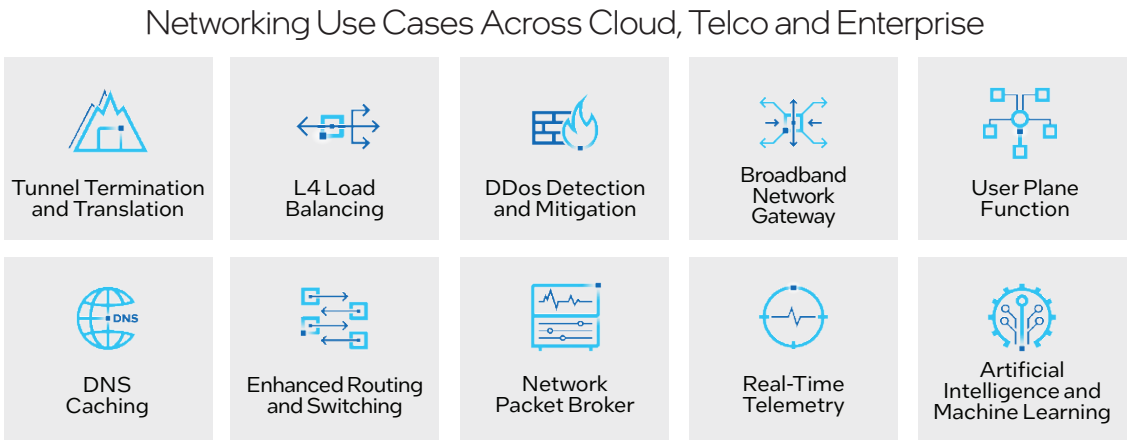

Figure 1. Intel® Tofino™ Intelligent Fabric Processors (Intel® Tofino™ IFPs) with full programmability can be used in many networking applications.

The Intel Tofino IFP is an intelligent, high-performance packet processor that enables Ethernet switches to deliver greater visibility and control across the network, provides the high bandwidth and flexibility required by the above use cases, and delivers up to 25.6 terabits per second (Tbps) of total switching capacity. But in certain cases, the Intel Tofino IFP can only do so much to help network owners and operators reach their scalability and programmability goals. To support extra-large session scale associated with some of the above use cases—like broadband network gateway (BNG)—or extra-large buffers required in CoSP networks, network owners and operators need something more.

Intel® Tofino™ Expandable Architecture Delivers Terabit-Level Speeds and Session Scale

Intel® Tofino™ Expandable Architecture combines Intel Tofino IFPs and Intel® CPUs with one or more accelerators like Intel® FPGAs and/or Intel® Infrastructure Processing Units (Intel® IPUs), as shown in Figure 2. This combination can handle multiple terabits of network traffic per single system, and supports a large number of sessions and/or very deep buffers. For example, Intel Tofino IFP on-chip memory can store tables with millions of entries. But the Intel Tofino Expandable Architecture can store up to hundreds of millions of table entries and up to tens of GBs of buffers.[10]

Intel Tofino Expandable Architecture is an excellent choice for use cases that need larger tables and deeper buffers than are available with standard switches.

These use cases include the following (among others):

- Cloud networking functions like layer 4 load balancing, firewalls, tunnel termination and translation and network address translation (NAT).
- Telco networking functions like carrier-grade network address translation (CG-NAT), high-end routing, network packet brokering or various user plane functions.
- Telco gateway use cases like BNG or access gateway function (AGF), which need extra-large buffers to perform advanced shaping and scheduling functions.

Because it is based on open industry standards, Intel Tofino Expandable Architecture enables customers to choose from a wide variety of products to optimize total cost of ownership for large-scale use cases. For additional value, the Intel FPGAs and/or Intel IPUs that are part of the Intel Tofino Expandable Architecture can be programmed to enable other high-value use cases, such as floating-point computations and hierarchical quality of service (HQoS).

Multiple companies are already using Intel Tofino Expandable Architecture to transform and modernize their networks.

Intel® Tofino™ Expandable Architecture In Action

See how cloud service providers are scaling network bandwidth and functionality.

Baidu Intelligent Cloud is using Intel® Tofino™ Expandable Architecture in its cloud gateways to help meet terabit-level network traffic flows generated by cloud workloads like autonomous driving, vehicle-to-everything (V2X) communication, Internet of Things (IoT), e-commerce, online videos and games.[11]

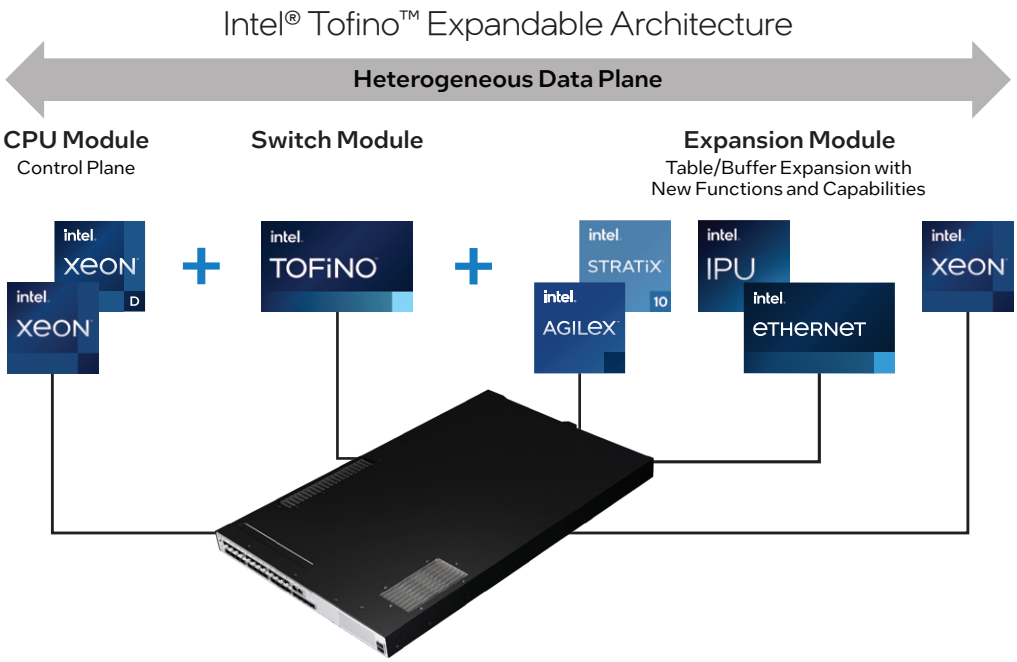

Figure 2. High-level view of Intel® Tofino™ Expandable Architecture.

Implementation Options

Intel Tofino Expandable Architecture can be implemented in different ways depending on the specific application and/or use case requirements. Below, two example implementations using FPGAs to enhance data plane capabilities are described in more detail (see Figure 3):

- **Look-aside architecture.** In this scenario, the Intel Tofino IFP provides all the front-panel connectivity, and it is up to the code running in the Intel Tofino IFP to decide what part of the network traffic is forwarded for additional processing by one or more Intel FPGAs.
- **Inline architecture.** In contrast to the look-aside option, the inline architecture connects some of the network-facing ports to the Intel Tofino IFP and some to the Intel FPGAs, so that the traffic flowing from one part of the network to another must pass through both the Intel Tofino IFP and the Intel FPGAs.

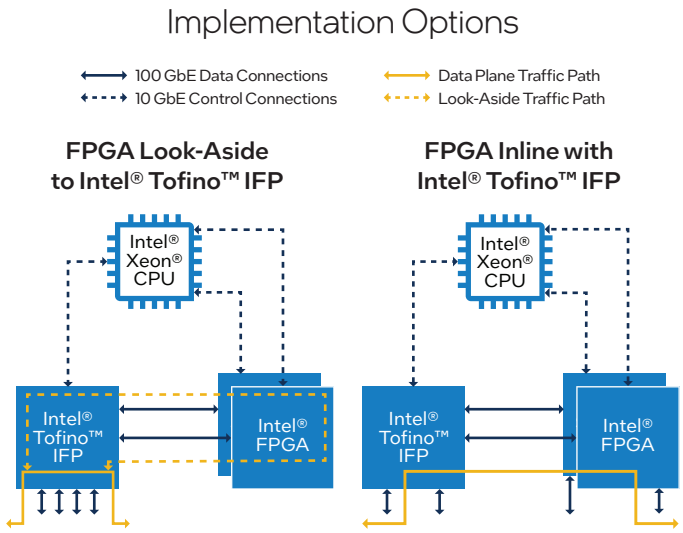

Figure 3. Implementation options for Intel® Tofino™ Expandable Architecture.

Hardware Form Factor Options

Network owners and operators can choose between three hardware form factors (see Figure 4). Each has its advantages and disadvantages; the choice is defined by available space, use case and other considerations.

- **Intel Tofino IFP plus Intel FPGA-based card and/or Intel IPU.** This form factor combines an off-the-shelf Intel Tofino IFP-based switch and an off-the-shelf Intel FPGA-based card and/or Intel IPU.
  - Advantage: Flexible form factor that allows for arbitrary connections between the Intel Tofino-based switch and the Intel FPGA-based cards and/or Intel IPUs using optical cables.
  - Disadvantage: Requires a server to host the Intel FPGA-based cards and/or Intel IPUs.
- **Server switch.** This form factor combines CPUs, Intel FPGAs and/or Intel IPUs, and Intel Tofino IFPs into a single rack-mount system.
  - Advantage: Integrated platform has a small footprint and is more cost-efficient than the disaggregated alternative of Intel Tofino IFP plus Intel FPGA-based card and/or Intel IPU.
  - Disadvantage: The internal architecture is predefined, so the system needs to be selected according to the use case requirements.
- **Chassis-based platform.** In this form factor, different compute sleds or blade servers can be used to achieve different ratios of compute, networking and storage.
  - Advantage: Scalable and configurable through sleds or blades.
  - Disadvantage: Restricted by the vendor ecosystem and chassis physical limitations.

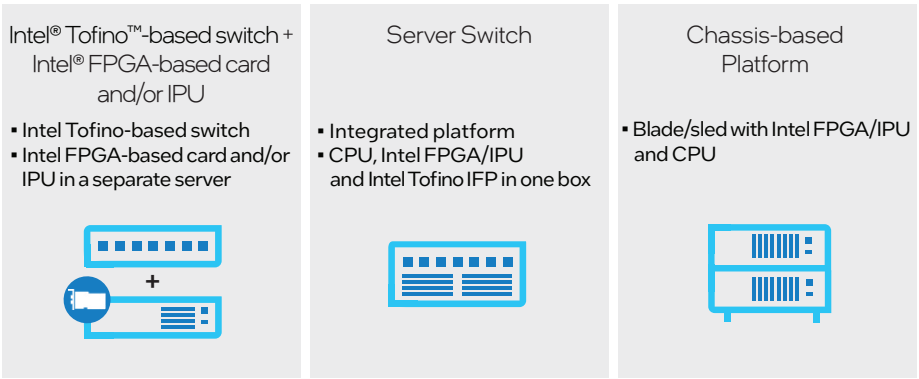

Figure 4. Intel® Tofino™ Expandable Architecture offers several form factors—choose the form factor that is best for your use case.

Developing Application Software

Intel Tofino Expandable Architecture is a dynamic solution meant for a broad range of use cases. The combination of components working together allows consumers the flexibility to choose how they would like to program the solution (see Figure 5) and achieve its full potential. Approaches include the following:

- **Programming is fully owned and developed by the customer:** The Intel Tofino Expandable Architecture hardware can be programmed either by the customer (as Baidu has done—see the sidebar, "Intel® Tofino™ Expandable Architecture In Action") or by another ecosystem player. When programming FPGAs and IPUs, customers can use a variety of different languages and tools. The P4 compiler for Intel FPGAs is in the early stages of development and is available to select customers.
- **Customers take advantage of data plane building blocks:** Various data plane building blocks are available already or are going to be released soon. These components allow customers to develop the functionalities that they desire while providing pre-integrated functionalities to shorten the time to market.
- **Customers license an existing application:** For customers that are looking for a complete solution for a particular use case, they can license such an application from the respective vendor (either from Intel, an independent software vendor or systems integrator).

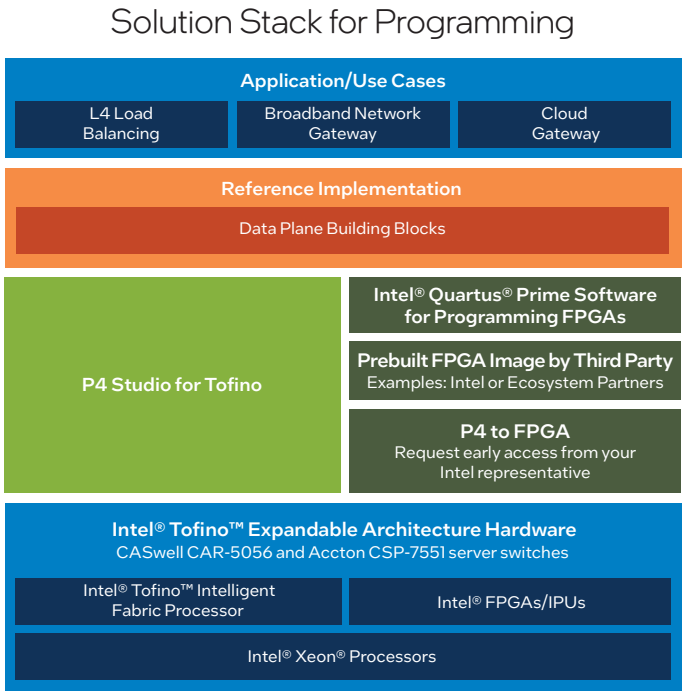

Figure 5. Solution stack for programming the components of Intel® Tofino™ Expandable Architecture

Sample Use Case: Layer 4 Server Load Balancing

Server load balancing is essential for efficient utilization of the data center resources—and the faster the load balancing happens, the fewer the network delays. This section presents a stateful server load-balancing acceleration use case using Intel Tofino Expandable Architecture. The combination of Intel Tofino IFP and Intel FPGAs provides both high-throughput as well as high-session scale: Intel Tofino Expandable Architecture delivers 3.2 Tbps of network bandwidth and supports 256 million sessions.

Traditional software-defined load balancing examines each client request and determines which backend server to forward the request to. That server then delivers the content to the client. Intel Tofino Expandable Architecture can improve performance of the load-balancing operation by introducing a session cache that runs on the server switch architecture described earlier.

Here's how the acceleration works: At the beginning of a session, the packet processing flow is similar to the traditional approach (client sends a request, which is forwarded to the software load balancer, etc.). But unlike the traditional software-only load-balancing scenario, the load-balancing decision is also stored in the session cache of the accelerator (in this case, a switch server containing an Intel Tofino IFP and an Intel FPGA), so that subsequent packets in the same session can be forwarded to the correct backend server directly by the accelerator. This approach significantly reduces the load on the software load balancers and improves the efficiency of the load-balancing service. Using Intel Tofino Expandable Architecture to accelerate server load balancing enables network operators to deploy more backend servers and potentially generate more revenue.

Let's examine the role of each component of the server switch in more detail:

- The CPU runs the control plane, which is based on the SONiC[12] network operating system where existing SONiC features are complemented by load-balancing-specific capabilities. The design features container-based, lightweight microservices and delivers fine-grained failure recovery and in-service upgrades with zero downtime.
- The Intel Tofino IFP performs most of the data plane processing and packet forwarding.
- The Intel FPGA provides the session table expansion as well as in-band data plane table management for extremely fast table updates.

The stateful server load-balancing acceleration (SLBA) API is based on Redis, which supplies a high-performance, in-memory key-value store.

The SLBA can be connected to a leaf/spine CLOS fabric, and independent scaling of the SLBA data plane and the server load-balancer software control plane allows for redundancy and an optimal ratio of data plane to control plane instances according to traffic patterns.

When 1 + 1 Is Greater Than 2

The key to the acceleration in this SLBA use case is a two-level cache. The hot cache is on the Intel Tofino IFP, while the warm cache is on the Intel FPGA. Figure 6 illustrates how the two-level cache works in the SLBA use case:

- When a client request arrives, the lookup happens first in the **hot cache**:
  - Hot cache hit: The packet is forwarded to the correct backend server.
  - Hot cache miss: A lookup request is created and sent to one of the Intel FPGAs (warm cache):
- **Warm cache** lookup:
  - Warm cache hit: The Intel Tofino IFP uses the lookup result to forward the packet to the correct backend server.
  - Warm cache miss: No rule for this session exists; the packet is forwarded to the control plane.

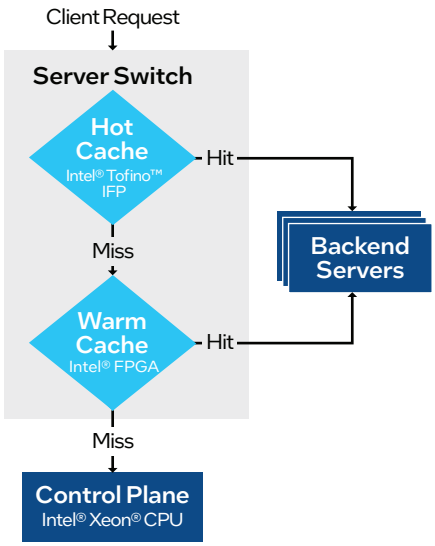

Figure 6. Decision flow for the layer 4 server load-balancing acceleration (SLBA) two-level cache.

Conclusion

Intel Tofino IFPs let network owners and operators customize their networks for specific workloads today, and into the future. To further enable network transformation that can support network functions that go beyond traditional switching, Intel Tofino Expandable Architecture offers four key benefits.

- Massive Performance and Scale
- Flexibility
- Open Standards Support
- Reduced Total Cost of Ownership

Intel Tofino Expandable Architecture offers limitless opportunities for innovation. Network owners and operators can take advantage of the P4 programmability of Intel Tofino IFPs for packet processing, while employing the power of Intel CPUs and the acceleration capabilities of Intel FPGAs and Intel IPUs to extend and augment Intel Tofino IFP functionality.

Learn More

You may find the following resources helpful:

- Intel® Tofino™ Expandable Architecture
- Intel® Tofino™ Intelligent Fabric Processors (Intel® Tofino™ IFPs)
- 3rd Generation Intel® Xeon® Scalable processors
- Intel® FPGAs
- P4 open source programming language
- SONiC operating system
- Architecture for Multi-Terabit Programmable Networking Functions video

For more information, contact your Intel representative and visit **intel.com/fabric**.

[1] A typical data center-focused switch offers hundreds of Mbits of table memory https://www.intel.com/content/dam/www/central-libraries/us/en/documents/tofino-product-family-brochure.pdf and hundreds of Mbytes of buffer capacity https://www.linleygroup.com/mpr/article.php?id=12304 Intel Stratix 10 MX FPGA comes with 8 GBytes or 16 GBytes of HBM2 memory https://cdrdv2.intel.com/v1/dl/getContent/652451 offering 50-100x more memory capacity.

[2] Statista, "Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025."

[3] Statista, "Worldwide visits to Google.com from December 2021 to May 2022."

[4] Science Daily, "Pandemic accelerated remote work, a trend likely to remain."

[5] Web Tribunal, "Cloud Adoption Statistics for 2022."

[6] Ericsson Mobility Report, "Mobile data traffic outlook."

[7] Invideo, "135 Video Marketing Statistics You Can't Ignore in 2022."

[8] Network Computing, "How Better Network Visibility Can Reduce Customer Churn."

[9] P4 Open Source Programming Language

[10] See footnote 1.

[11] Baidu Intelligent Cloud blog

[12] Open Compute Project – SONiC