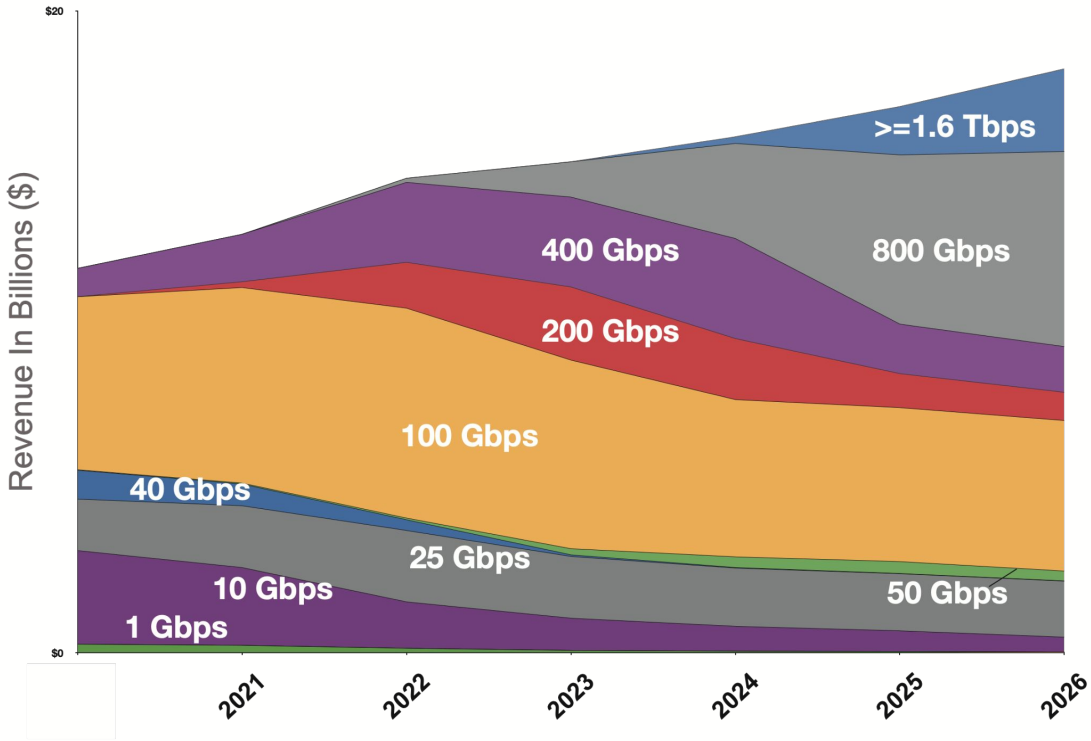
Rise of the Merchant Silicon

Next gen ASICs for the peering fabric

Ethernet Switch Revenue Forecast



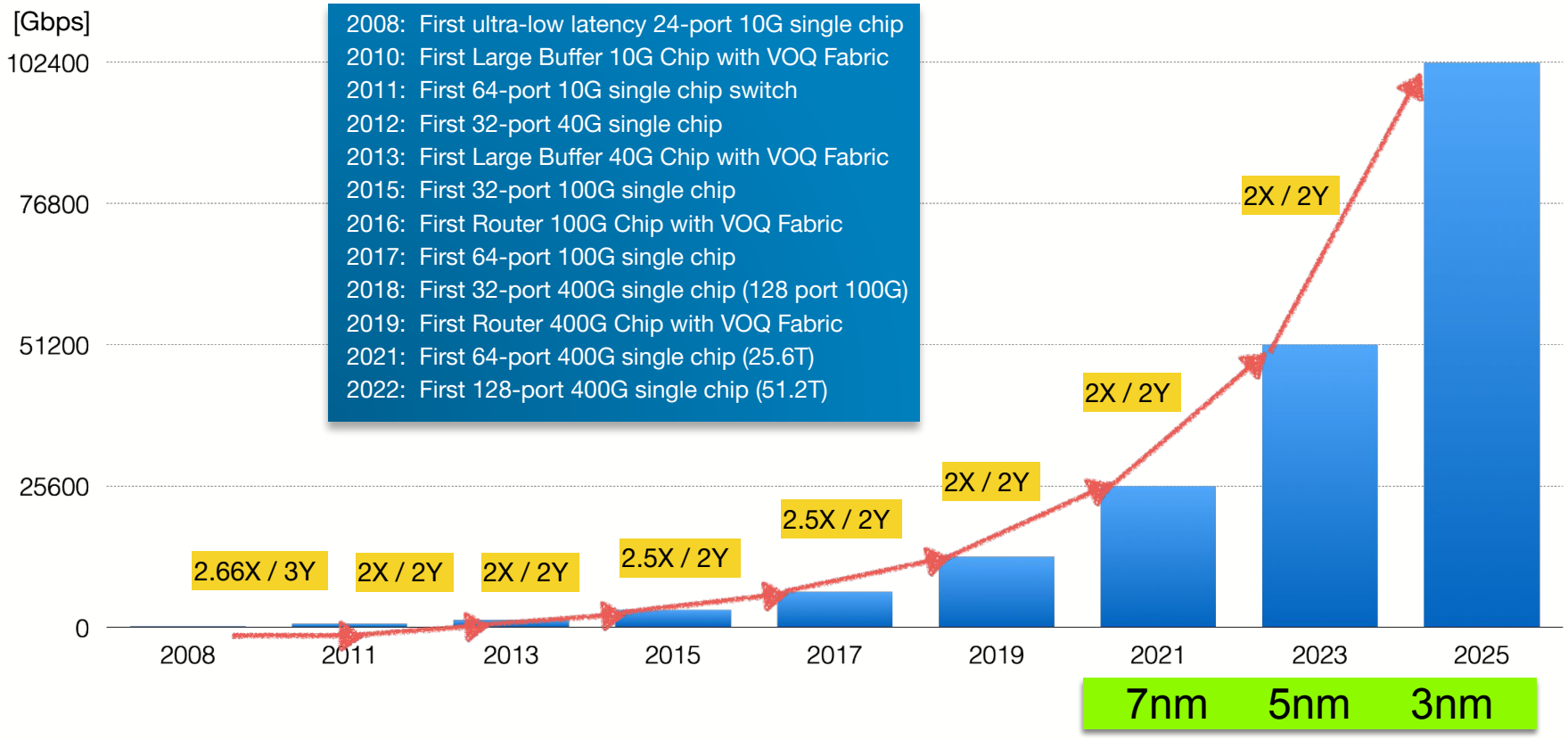
Source: Dell'Oro March 2022 - Long Term Ethernet Switch Forecast

800G and 1600G expected to be 45% of Market in 2026

400G and below expected to be 55% of Market in 2026

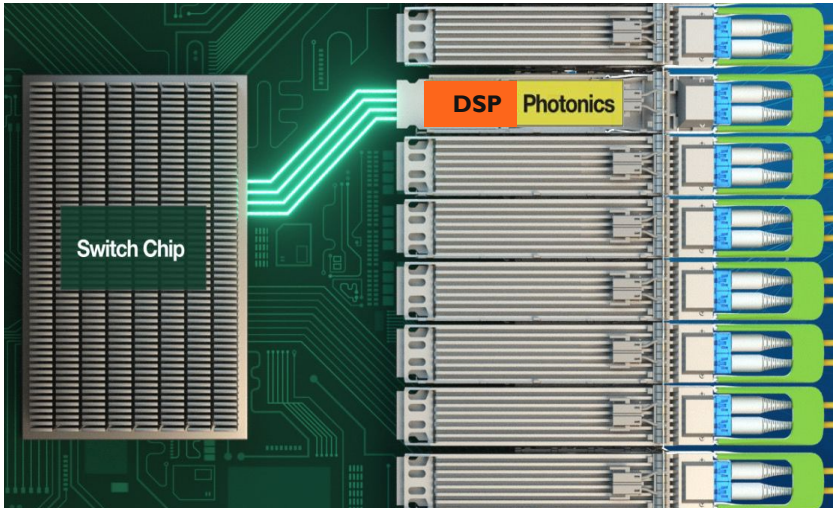
ARISTA

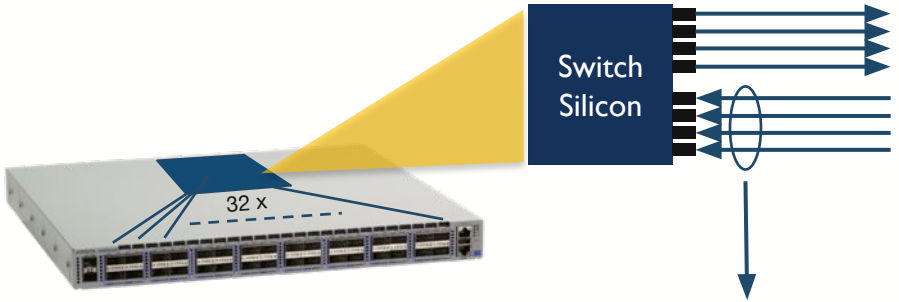
Single-chip Switch Bandwidth & Serdes Speeds



SERDES Speeds are Key to Scaling networks

- Serdes (or **Ser**ializer-**Des**erializers) refer to the technology used for high-speed chip I/O
- Serdes speeds place a fundamental limit on datacenter bandwidth
- The easiest way to go faster is (for serdes speeds) to go Faster

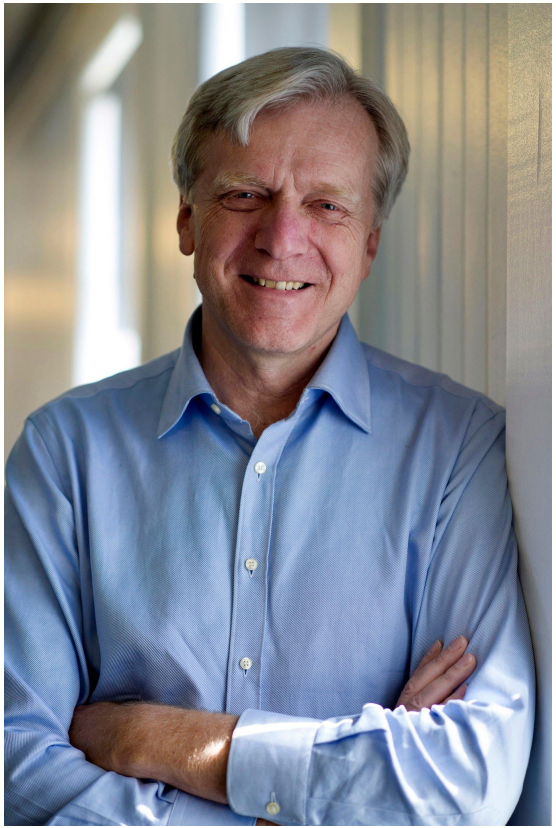




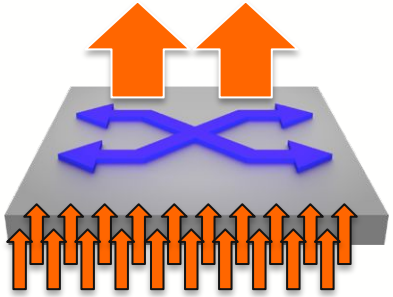
Ethernet switch chip I/O **(or serdes) speeds**:
10G -> 25G -> 50G -> 100G

ARISTA

“The easiest way to go faster is to go faster”

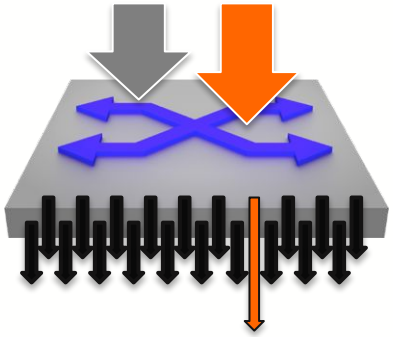


When Buffers Matter in Networks

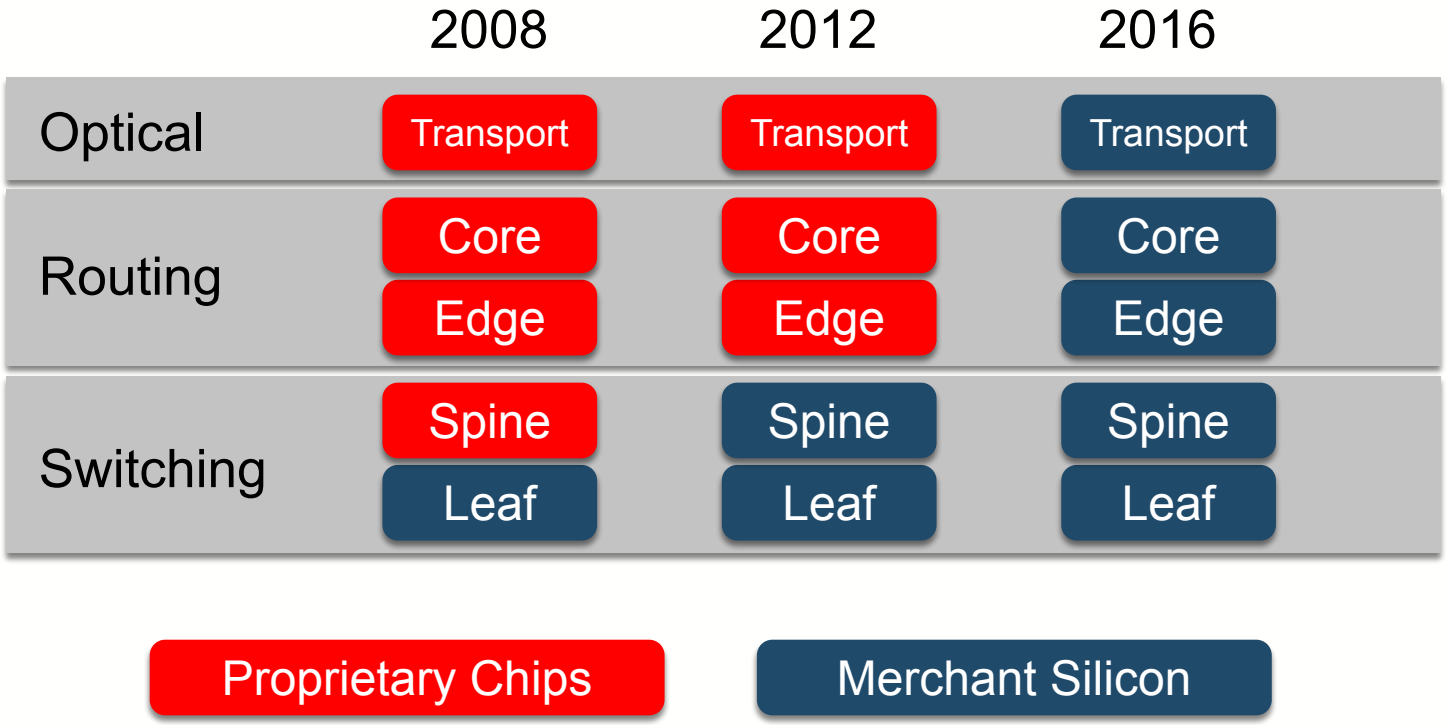


Incast (Many to Fewer)

Speed Change (Faster to Slower)



The Evolution of Merchant Silicon



ARISTA

Process Technology Improvements (TSMC)

Process Node	7nm	5nm	3nm
Relative Density	1	1.5	2.25
Speed @ IsoPower	1	1.15	1.4
Power @ IsoSpeed	1	0.8	0.6
Volume Manufacturing	2019	2021	2023

Each process generation enables more throughput, better Power Efficiency, more buffers, bigger routing tables, etc

Cost of silicon design

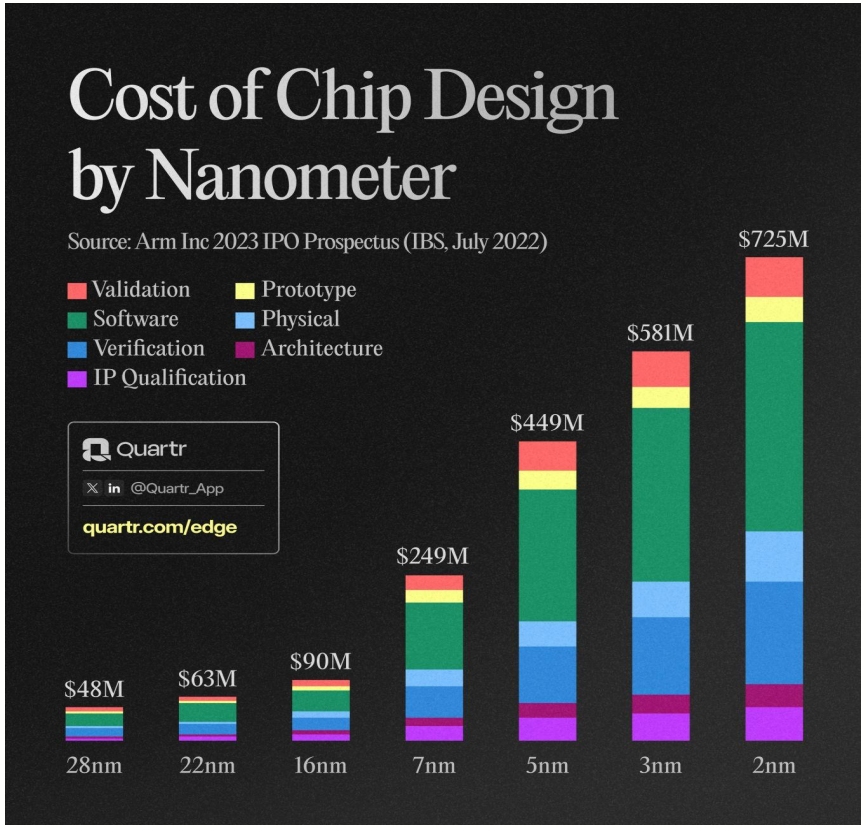
Cost of design (aka tape out) has risen significantly

- 28nm: 48M$
- 22nm: 63M$ +33%
- 16nm: 90M$ +43%
- 7nm: 249M$ +176%
- 5nm: 449M$ +80%
- 3nm: 581M$ +30%
- 2nm: 725M$ +25% (new: ~ 2nd half of 2025)

This favors merchant silicon because of higher volume

This also leads to:

- Lower cost / better economy of scale
- Faster product cycles
- Faster innovation (because of shorter cycles)
- Lower power consumption
- Better scale
- More features (not always but often)



Choices in Switching Silicon

All chip makers have access to the same technology

- same fabs and processes
- same memories, TCAMs, serdes
- same clock rate

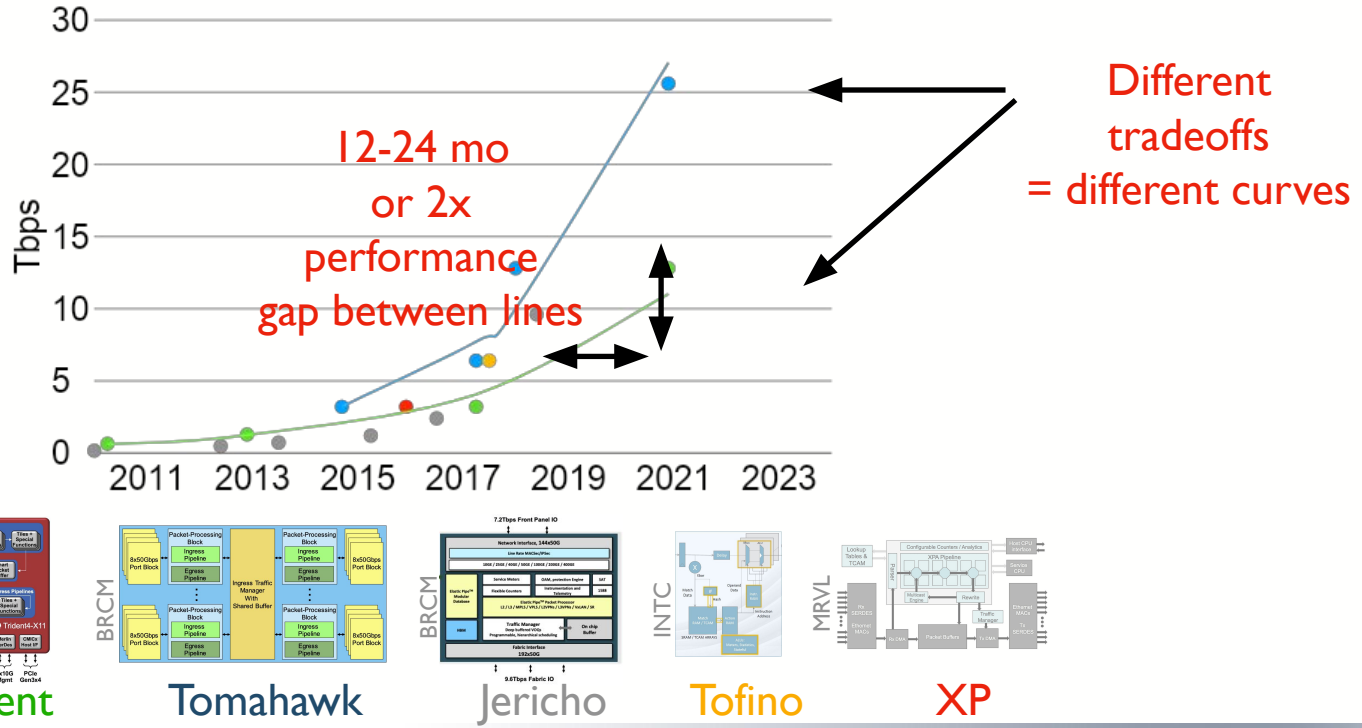
Differences arise *primarily* because of

- design tradeoffs for different use cases
- process shifts (28nm -> 16nm -> 7nm -> 5nm)
- faster innovation cycles

There is no fundamental advantage to proprietary silicon

Merchant Silicon Trajectory

10 years at Arista, across chip families



ARISTA

2024: Next Generation Silicon for Networks

Tofino

- very high density
- very high performance
- very low power per Gbps
- Highly flexible and programmable
- intermediate programming complexity
- deterministic, low latency

Tomahawk5

- 2x higher performance
- Up to 51.2 Tbps
- 165MB Buffer
- Scale Out & High Radix

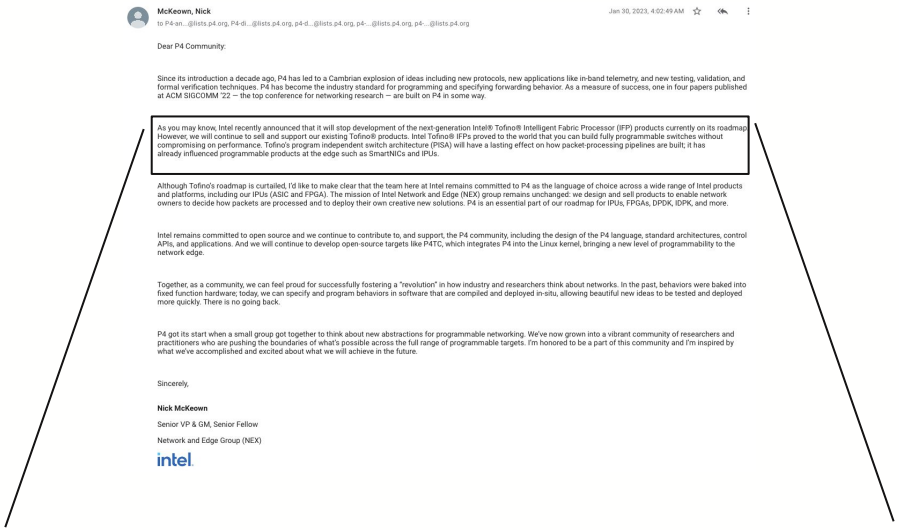
Trident4

- 4x higher performance
- Up to 12.8 Tbps
- 132MB Buffer
- Programmable Pipeline

Jericho2C+

- 50% higher performance
- 7.2 Tbps
- 2.7 Bpps
- Deep Buffers
- Extensible

The last Tofino?



McKeown, Nick Jan 30, 2023, 4:02:49 AM

to P4-an...@lists.p4.org, P4-di...@lists.p4.org, p4-d...@lists.p4.org, p4-...@lists.p4.org, p4-...@lists.p4.org

Dear P4 Community:

Since its introduction a decade ago, P4 has led to a Cambrian explosion of ideas including new protocols, new applications like in-band telemetry, and new testing, validation, and formal verification techniques. P4 has become the industry standard for programming and specifying forwarding behavior. As a measure of success, one in four papers published at ACM SIGCOMM '22 – the top conference for networking research – are built on P4 in some way.

As you may know, Intel recently announced that it will stop development of the next-generation Intel® Tofino® Intelligent Fabric Processor (IFP) products currently on its roadmap. However, we will continue to sell and support our existing Tofino® products. Intel Tofino® IFPs proved to the world that you can build fully programmable switches without compromising on performance. Tofino's program independent switch architecture (PISA) will have a lasting effect on how packet-processing pipelines are built; it has already influenced programmable products at the edge such as SmartNICs and IPUs.

Although Tofino's roadmap is curtailed, I'd like to make clear that the team here at Intel remains committed to P4 as the language of choice across a wide range of Intel products and platforms, including our IPUs (ASIC and FPGA). The mission of Intel Network and Edge (NEX) group remains unchanged: we design and sell products to enable network owners to decide how packets are processed and to deploy their own creative new solutions. P4 is an essential part of our roadmap for IPUs, FPGAs, DPDK, IDPK, and more.

Intel remains committed to open source and we continue to contribute to, and support, the P4 community, including the design of the P4 language, standard architectures, control APIs, and applications. And we will continue to develop open-source targets like P4TC, which integrates P4 into the Linux kernel, bringing a new level of programmability to the network edge.

Together, as a community, we can feel proud for successfully fostering a "revolution" in how industry and researchers think about networks. In the past, behaviors were baked into fixed function hardware; today, we can specify and program behaviors in software that are compiled and deployed in-situ, allowing beautiful new ideas to be tested and deployed more quickly. There is no going back.

P4 got its start when a small group got together to think about new abstractions for programmable networking. We've now grown into a vibrant community of researchers and practitioners who are pushing the boundaries of what's possible across the full range of programmable targets. I'm honored to be a part of this community and I'm inspired by what we've accomplished and excited about what we will achieve in the future.

Sincerely,

Nick McKeown

Senior VP & GM, Senior Fellow

Network and Edge Group (NEX)

intel.

"As you may know, Intel recently announced that it will stop development of the next-generation Intel® Tofino® Intelligent Fabric Processor (IFP) products currently on its roadmap."

Public announcement by Nick McKeown (Intel Senior VP & GM)
https://groups.google.com/a/lists.p4.org/g/p4-announce/c/frXi_jjmawE

ARISTA

2024: Next Generation Silicon for Networks

Tofino

- very high density
- very high performance
- very low power per Gbps
- Highly flexible and programmable
- intermediate programming complexity
- deterministic, low latency

Tomahawk5

- 2x higher performance
- Up to 51.2 Tbps
- 165MB Buffer
- Scale Out & High Radix

Trident4

- 4x higher performance
- Up to 12.8 Tbps
- 132MB Buffer
- Programmable Pipeline

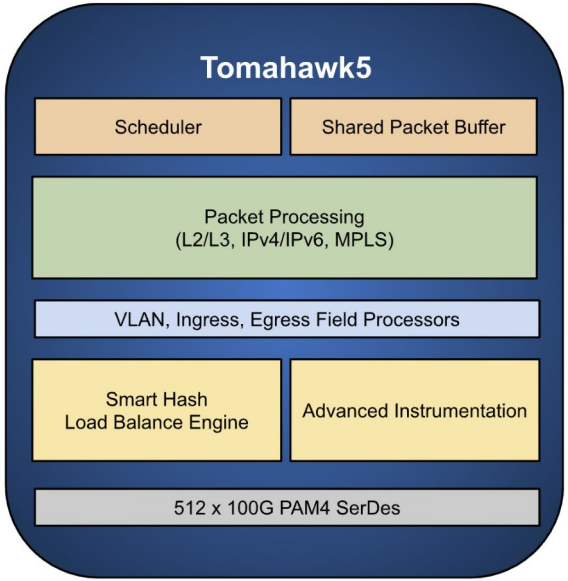
Jericho2C+

- 50% higher performance
- 7.2 Tbps
- 2.7 Bpps
- Deep Buffers
- Extensible

Tomahawk

Tomahawk5 for AI Networks

- 2X performance
  - 51.2Tbps: 512 x 100G PAM4
  - Powerful new SerDes help in supporting LPO optics
- Efficient and Scalable Architecture
  - IPv4/v6, VxLAN and advanced instrumentation
- High Radix with flexible port speeds
  - Up to 320 front panel ports at 10G to 800G speeds
- Cloud optimized pipeline and unified packet buffer
  - 165MB shared buffer
  - Absorbs bursts 10x better



Tomahawk Evolution

	7060X4 12.8T (TH3)	7060X5 25.6 (TH4)	7060X6 51.2 (TH5)
Max I/O	256 x 50G 32 x 400G	512 x 50G 64 x 400G	64 x 800G 128 x 400G 256 x 200G 512 x 100G
Max Throughput	12.8Tbps	25.6Tbps	51.2Tbps
Logical Ports	144	256	320
Buffer	64MB	114MB	165MB
L2 MAC	8K	128K	
L3 Hosts	8K	Shared with ALPM	
IPv4/IPv6 LPM (ALPM)	770K/500K		
Tunnel TCAM	256		512
True Egress Mirror	No		Yes
VXLAN	Not Supported	Yes	Yes
LPO Support	Not Supported	Not Supported	Yes

2X I/O Increase

Increased Buffer

Consistent Scale

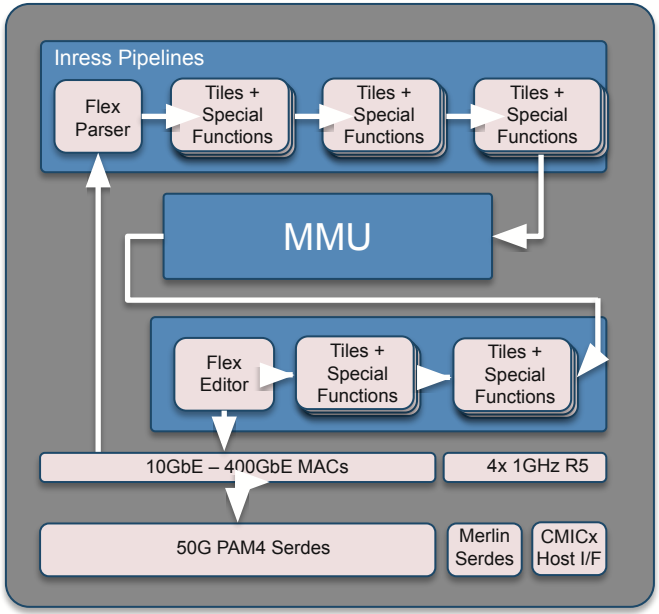
Advanced Traffic Management

Trident

Arista 7050X4 X.11(12.8T) vs X.9(8.0T) – Comparison

	7050X3 3.2T (TD3)	7050X4 12.8T (TD4)	7050X4 8T (TD4)
Max I/O	128 x 25G 32 x 100G	256 x 50G 32 x 400G	**160 x 50G** **20 x 400G**
Max Throughput	3.2Tbps	12.8Tbps	**8.0Tbps**
Logical Ports	128	144	**72**
Buffer	32MB	132MB (Hybrid-Shared)	**82MB** **(Fully Shared)**
Latency	800ns	900ns	
L2 MAC	288K	128K	
L3 Hosts	168K	320K	
IPv4/IPv6 LPM (ALPM)	384K/192K	800K/500K	
MACsec & IPSec	No	No	**Yes (4.8T)**
Exact Match Rules	128K	256K*	
Counters	114K	256K	
ACLs	7K	11K + 2K egress	
VXLAN, uRPF, VLAN Translation	Yes	Yes	

* 128bit wide entries



Consistent features

Advanced Instrumentation – In-band telemetry for latency monitoring

Dynamic Load Balancing – Traffic awareness improves ECMP performance

Traffic Scheduling – Microburst and Elephant Flow Detection and Prioritization

High-Performance Shared-Buffer memory – Improves burst absorption

Increased Routing and ACLs – Larger IPv4/v6 Scale and robustness

Trident4: Efficient System Design

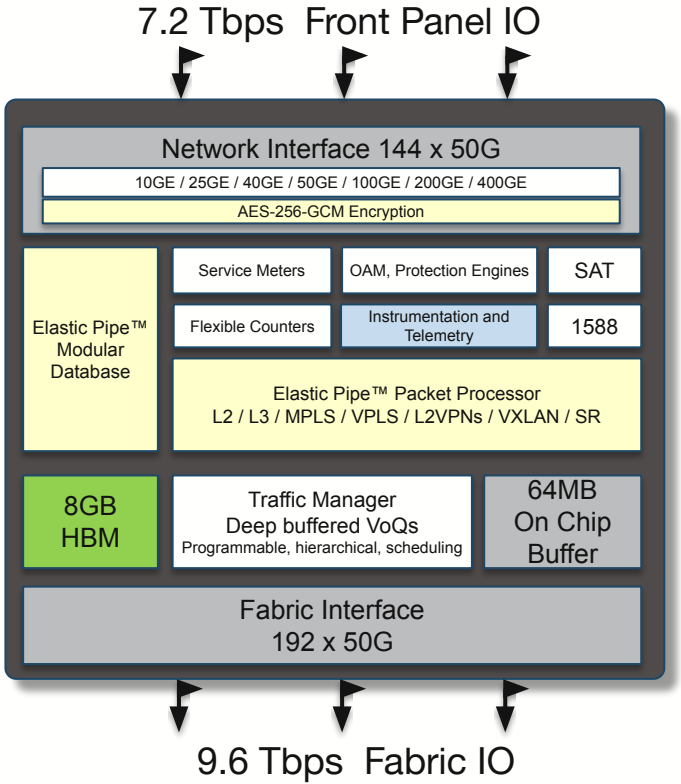
Low power consumption	**20-40% power savings**
Compact form factors	**SFP & QSFP based systems for rich and diverse connectivity**
Quad-Speed Investment Protection	**Backwards and forward compatibility for current and future connectivity**
Feature and architecture consistency	**Consistent feature and scale parity with VXLAN support**

Complete Portfolio - Uncompromised Features and Scale

Jericho

16.8 Tbps - Jericho2C+

- 16.8 Tbps of High Performance with rich features
  - Total of 336 PAM-4 50G SerDes
  - 7.2Tbps Network I/O and 2.7Bpps packet processing
  - Flexible Network Interfaces - 10G to 400G
  - Integrated TunnelSec Encryption (MACsec, IPsec, VXLANsec)
- Flexible Lookup Tables and Programmable Pipeline
  - Fungible on chip tables allow multiple use case profiles
  - Off-chip expandability with External table expansion (KBP)
  - Flexible Pipeline allows reconfiguration of forwarding
- Hierarchical Traffic Management with Deep Buffer
  - 8GB High Bandwidth Memory (HBM)
  - 64MB On Chip Buffer
- Network Instrumentation and Telemetry
  - Hardware Accelerator
  - Monitor of large numbers of sessions



Consistent System Resources: J2C+/J2/J2C/Q2C

	R3 Series		R3K Series			
Profile	**L3 (default)**	**Balanced**	**L3-XL (default)**	**L3-XXL**	**L3-XXXL**	**Balanced-XL**
ARP Entries	88k	80k	112k	112k	80k	96k
MAC Addresses	224k	224k	256k	192k	384k	256k
IPv4 Unicast Routes	1450k	800k	2250k	2850k	3950k	1850k
Additional IPv4 Unicast Routes with FlexRoute	+1,792k	+1,792k	+2,048k	+1,536k	+3,072k	+2,048k
IPv6 Unicast Routes	433-483k	250-267k	683-750k	833-950k	1100-1317k	567-617k
Multicast Routes	128k	128k	128k	128k	128k	128k
TCAM ACL Entries (Per chip)	24k	24k	24k	24k	24k	24k
Traffic Policy ACL IPv4 Prefixes	30k	30k	430k	296k	30k	430k
Traffic Policy ACL IPv6 Prefixes	10k	10k	150k	100k	10k	150k
ECMP	512-Way	512-Way	512-Way	512-Way	512-Way	512-Way

Maximum values dependent on shared resources / user configuration

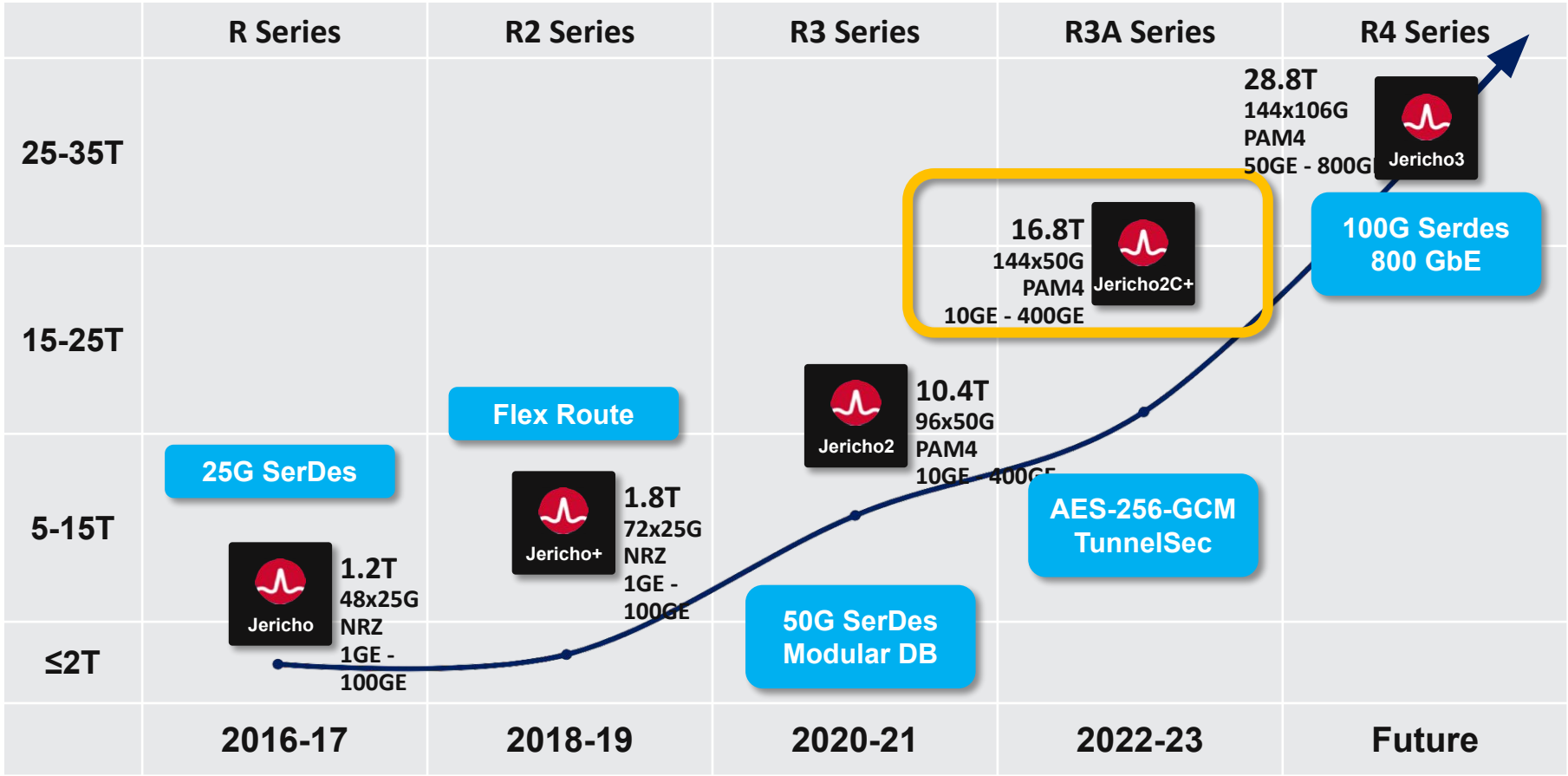
Jericho2 hardware resources are fungible. Values shown are unidimensional maxima for default profiles

Jericho2C+ - The Engine for 400Gbps

Jericho2c+ Lowest Cost, Power & RU per Gbps	Up to 50% Improvement from previous generation
Jericho2c+ 400Gbps Strong Encryption	MACsec, IPsec and VXLANsec at 10-400Gbps
Jericho2c+ Dense 400G ZR/ZR+ for WAN/DCI	Broad ZR/ZR+ Support with integrated Line System Ports
Jericho2c+ Rich DC and WAN Feature Set Large Scale Resources	Consistent Jericho2 Feature-set with dedicated 8GB HBM Deep Buffers
Jericho2c+ Flexible Product Choice	All Models Available in 3 Scale Configurations

Complete Portfolio - Uncompromised Features and Scale

Jericho based Portfolio

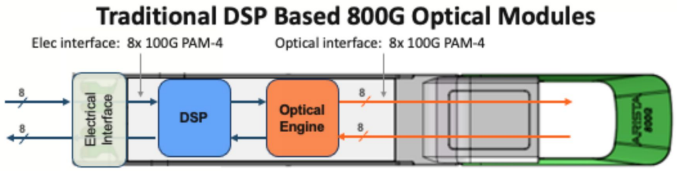


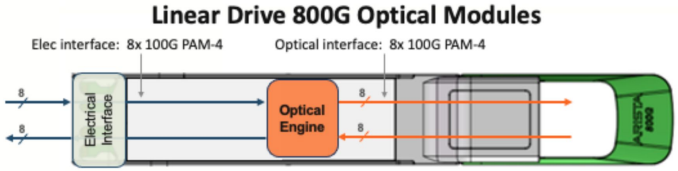
Rate Adapting 1G optics

- Support 1G-LX and 1G-SX on platforms that have a minimum port speed of 10G
  - e.g. J2 based platforms have a minimum port speed of 10G
- Connect to other devices that use CL37 (optical) autoneg when the used platform does NOT support CL37 autoneg
  - some platforms support 1/10/25G but don’t support CL37 autoneg

What are Linear Drive Optics Modules?

1. Linear Drive means no DSP or CDR in transceiver
   - Just a linear driver to provide required modulator voltage
2. Requires a high-performance switch SERDES
   - And very careful signal integrity design
3. Achieves power savings similar to direct drive CPO
   - While retaining the many advantages of pluggable optics modules
   - Opportunity to cut optics module power by 50% and system power by up to 25%





How is AI transforming the service provider market?

- AI networks need VoQ based deep buffer fabrics
- Perfect fit for Jericho chips
- next version (J3) already in
- made for high amount of 800G interfaces
- also beneficial for SPs as BW demands still >>
- also SPs benefit from enhanced telemetry functions

Summary

Tomahawk

- Low **latency**
- Shallow Buffers
- High **bandwidth**
- Limited features

Trident

- Low **latency**
- Shallow Buffers
- Datacenter feature set
- Optimal in compute leaf/spine

Jericho

- Deep **Buffer**
- **MPLS** capable
- Hairpinning
- Optimal for arbitrary topologies