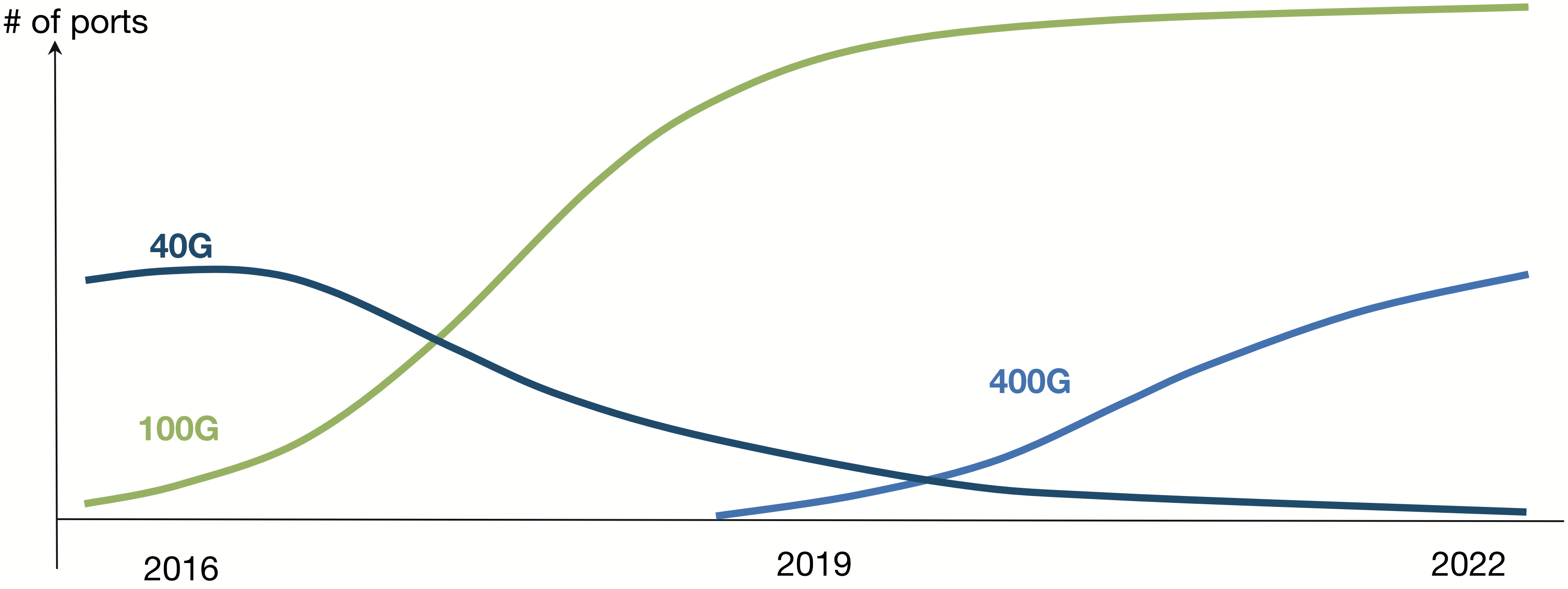
Merchant Silicon for Service Providers

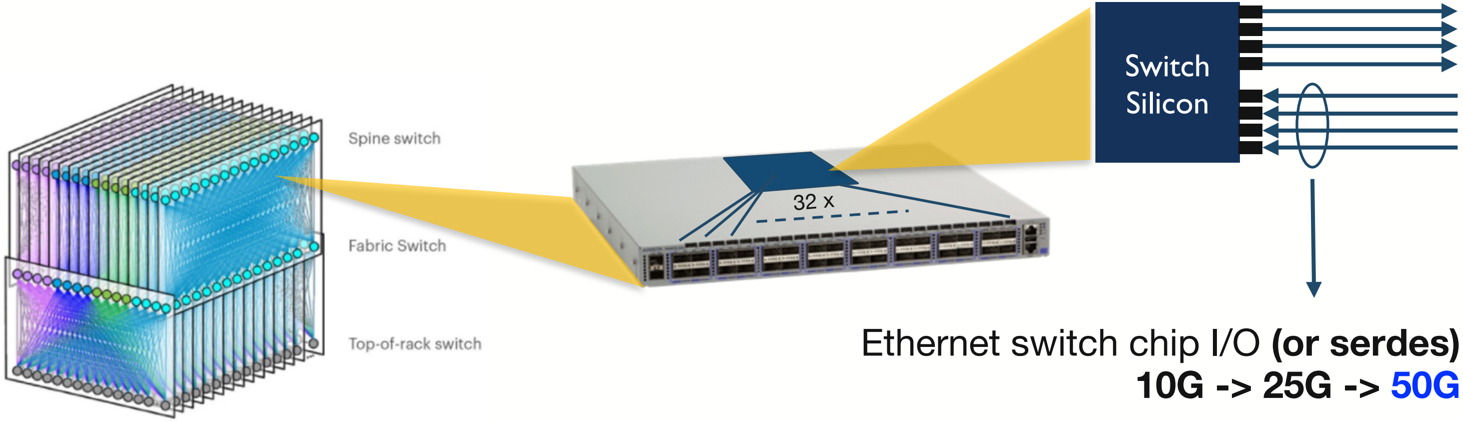
“The easiest way to go faster is to go faster”

40G - 100G - 400G Switch Port Transition



SERDES Speeds are Key to Scaling networks

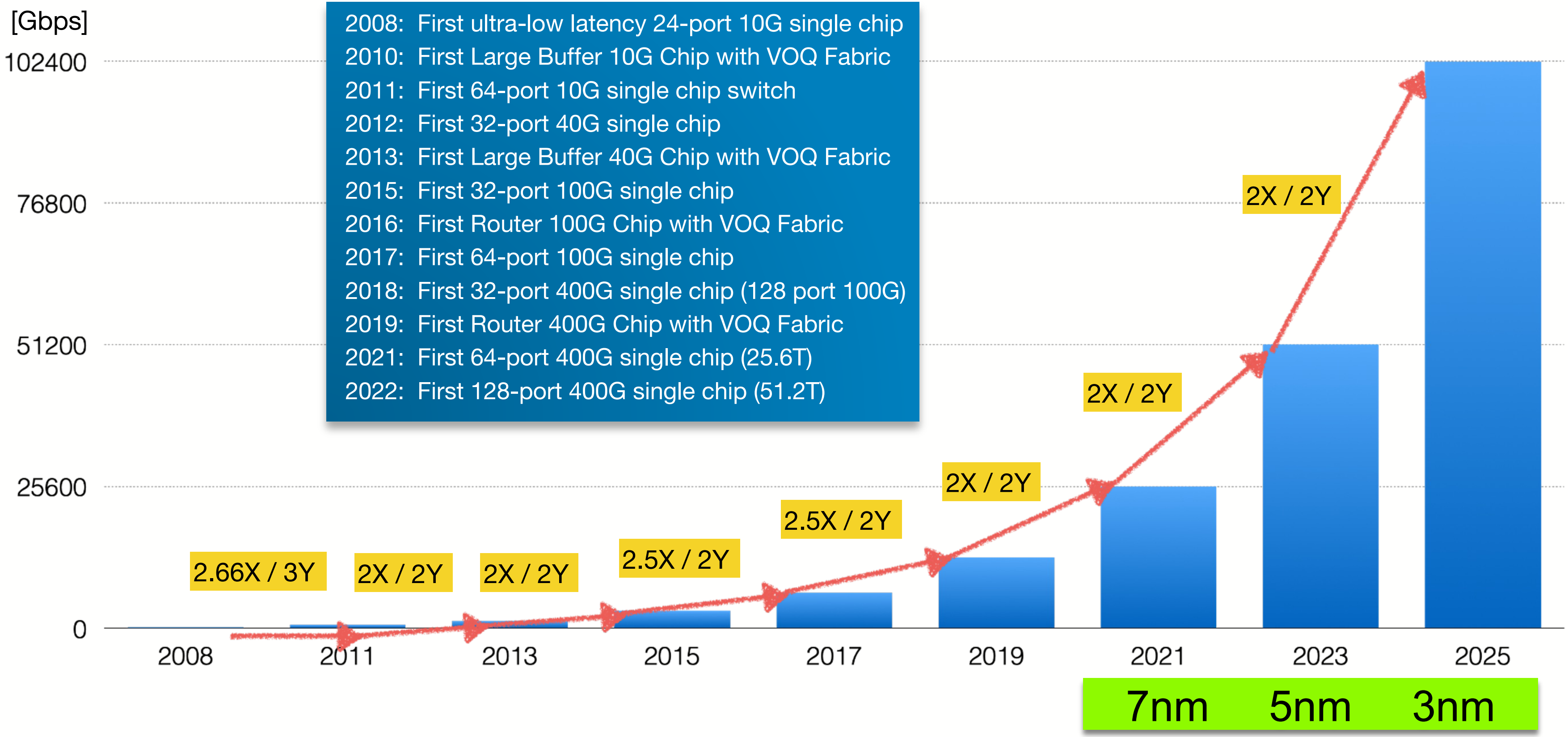
- Serdes (or **Ser**ializer-**Des**erializers) refer to the technology used for high-speed chip I/O
- Serdes speeds place a fundamental limit on datacenter bandwidth
- The easiest way to go faster is (for serdes speeds) to go Faster



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

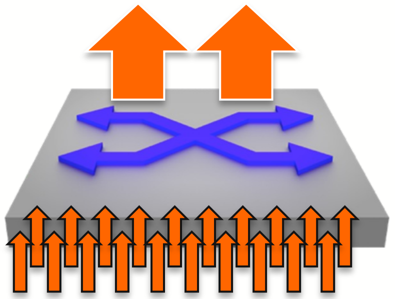
Facebook F16 data center network topology.
https://engineering.fb.com/data-center-engineering/f16-minipack/

Single-chip Switch Bandwidth & Serdes Speeds



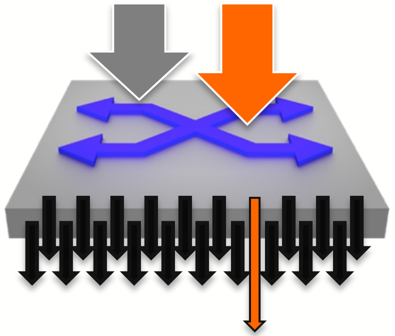
2008: First ultra-low latency 24-port 10G single chip
2010: First Large Buffer 10G Chip with VOQ Fabric
2011: First 64-port 10G single chip switch
2012: First 32-port 40G single chip
2013: First Large Buffer 40G Chip with VOQ Fabric
2015: First 32-port 100G single chip
2016: First Router 100G Chip with VOQ Fabric
2017: First 64-port 100G single chip
2018: First 32-port 400G single chip (128 port 100G)
2019: First Router 400G Chip with VOQ Fabric
2021: First 64-port 400G single chip (25.6T)
2022: First 128-port 400G single chip (51.2T)

When Buffers Matter in Provider Networks



Incast (Many to Fewer)

Speed Change (Faster to Slower)



Merchant Silicon Maturation

	2008	2014	2020
Optical	🟥 Transport	🟥 Transport	🟦 Transport
Routing	🟥 Core	🟥 Core	🟦 Core
	🟥 Edge	🟥 Edge	🟦 Edge
Switching	🟥 Spine	🟦 Spine	🟦 Spine
	🟦 Leaf	🟦 Leaf	🟦 Leaf

🟥 Proprietary Chips

🟦 Merchant Silicon

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

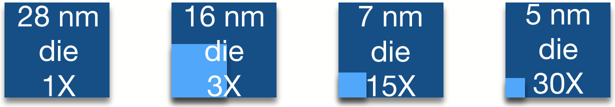
Choices in Switching Silicon

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

Domain-Specific Products for Different Networks

Trident

Enterprise application stacks

RoCEv2, EVPN, VXLAN

Rich Telemetry for deep visibility

Compute TOR for 10/25/50/100G

Flexible traffic management

128 x 100G in 4RU

Tomahawk

Cloud application stacks

Highest Switch performance

Lowest Latency

Scale Out & High Radix

High density 400G Fixed Spines

128 x 200G in 4RU

Jericho

Switch with **deep buffers**

EVPN, MPLS, SR

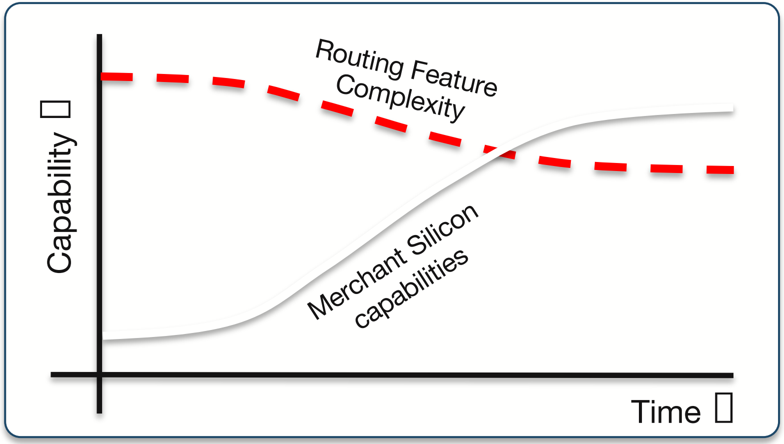
High capacity routing scale

Metro & DCI with MACSec & ZR

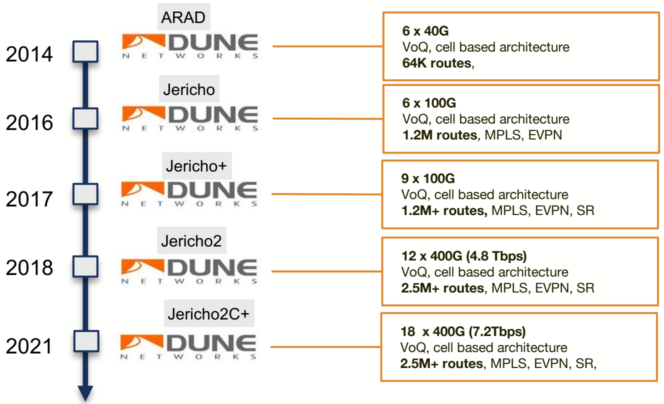
Fixed and Modular form factors

Arista: Bringing Merchant Silicon to the Routing Market

- Look at the routing market
  - The domain of the Network vendor’s own in-house ASIC
  - Due to complexity of functionality and table scale requirements



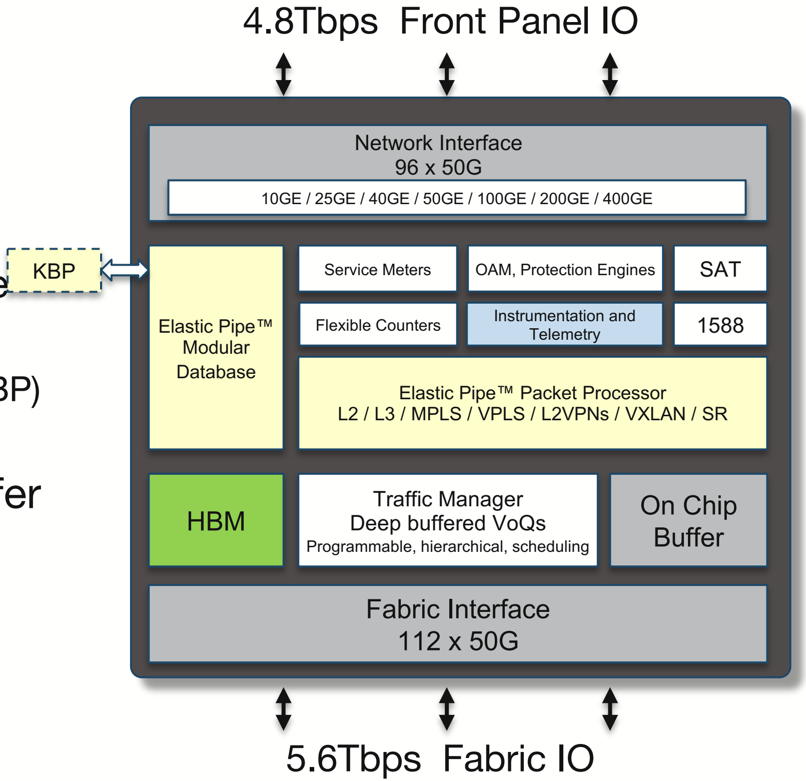
- Lines are blurring with latest Merchant silicon
  - Jericho chipset design for routing deployments
  - Market leading performance and 100G/400G density
  - Internet scale, multiple encap, Deep label stack, VoQ



2014	ARAD DUNE NETWORKS	**6 x 40G** VoQ, cell based architecture **64K routes**,
2016	Jericho DUNE NETWORKS	**6 x 100G** VoQ, cell based architecture **1.2M routes**, MPLS, EVPN
2017	Jericho+ DUNE NETWORKS	**9 x 100G** VoQ, cell based architecture **1.2M+ routes**, MPLS, EVPN, SR
2018	Jericho2 DUNE NETWORKS	**12 x 400G (4.8 Tbps)** VoQ, cell based architecture **2.5M+ routes**, MPLS, EVPN, SR
2021	Jericho2C+ DUNE NETWORKS	**18 x 400G (7.2Tbps)** VoQ, cell based architecture **2.5M+ routes**, MPLS, EVPN, SR,

10Tbps - Jericho2

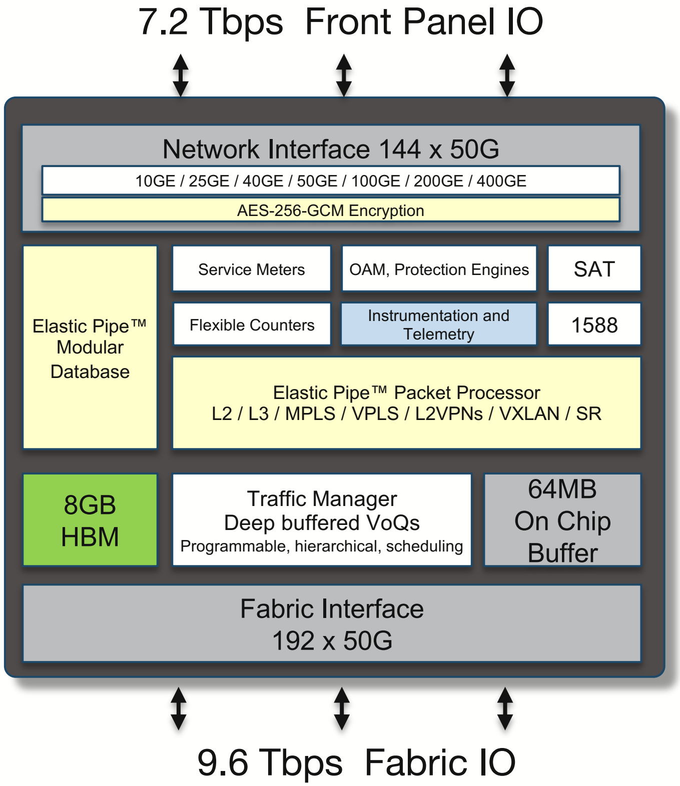
- 10Tbps of High Performance with rich features
  - Total of 208 PAM-4 50G Serdes
  - 4.8Tbps Network I/O and 2Bpps packet processing
  - Flexible Network Interfaces - 10G to 400G
- Flexible Lookup Tables and Programmable Pipeline
  - Fungible on chip tables allow multiple use case profiles
  - Off-chip expandability with External table expansion (KBP)
  - Flexible Pipeline allows reconfiguration of forwarding
- Hierarchical Traffic Management with Deep Buffer
  - 8GB High Bandwidth Memory (HBM)
  - 32MB On Chip Buffer
- Network Instrumentation and Telemetry
  - Hardware Accelerator
  - Monitor of large numbers of sessions



4.8Tbps Front Panel IO
Network Interface
96 x 50G
10GE / 25GE / 40GE / 50GE / 100GE / 200GE / 400GE
KBP
Elastic Pipe™ Modular Database
Service Meters
OAM, Protection Engines
SAT
Flexible Counters
Instrumentation and Telemetry
1588
Elastic Pipe™ Packet Processor
L2 / L3 / MPLS / VPLS / L2VPNs / VXLAN / SR
HBM
Traffic Manager
Deep buffered VoQs
Programmable, hierarchical, scheduling
On Chip Buffer
Fabric Interface
112 x 50G
5.6Tbps Fabric IO

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


7.2 Tbps Front Panel IO
Network Interface 144 x 50G
10GE / 25GE / 40GE / 50GE / 100GE / 200GE / 400GE
AES-256-GCM Encryption
Elastic Pipe™ Modular Database
Service Meters
OAM, Protection Engines
SAT
Flexible Counters
Instrumentation and Telemetry
1588
Elastic Pipe™ Packet Processor
L2 / L3 / MPLS / VPLS / L2VPNs / VXLAN / SR
8GB HBM
Traffic Manager
Deep buffered VoQs
Programmable, hierarchical, scheduling
64MB On Chip Buffer
Fabric Interface
192 x 50G
9.6 Tbps Fabric IO

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

	KAPS tbd		BIG KAPS			
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
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

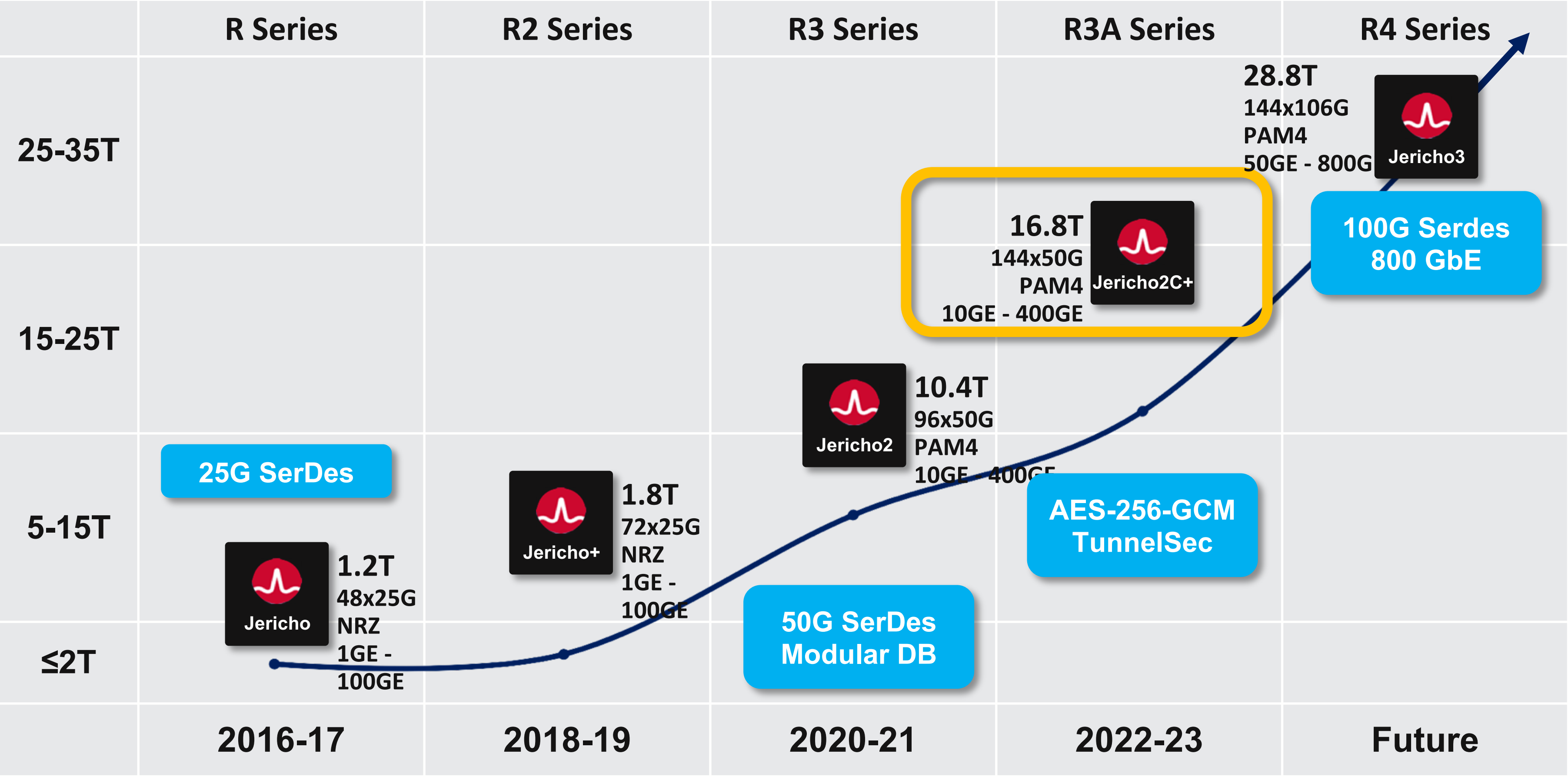
Jericho2 hardware resources are fungible. Values shown are unidimensional maxima for default profiles

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



R Series

R2 Series

R3 Series

R3A Series

R4 Series

25-35T

15-25T

5-15T

≤2T

25G SerDes

Jericho

1.2T
48x25G
NRZ
1GE -
100GE

Jericho+

1.8T
72x25G
NRZ
1GE -
100GE

Jericho2

10.4T
96x50G
PAM4
10GE - 400GE

50G SerDes
Modular DB

16.8T
144x50G
PAM4
10GE - 400GE

Jericho2C+

AES-256-GCM
TunnelSec

28.8T
144x106G
PAM4
50GE - 800G

Jericho3

100G Serdes
800 GbE

2016-17

2018-19

2020-21

2022-23

Future

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

1. Requires a high-performance switch SERDES

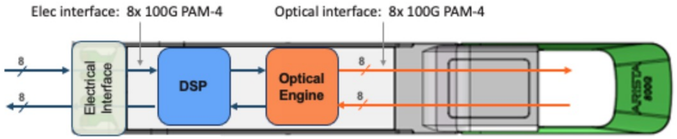
   And very careful signal integrity design

1. Achieves power savings similar to direct drive CPO

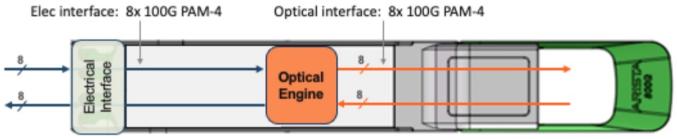
   While retaining the many advantages of pluggable optics modules

   Opportunity to cut optics module power by 50% and system power by up to 25%

Traditional DSP Based 800G Optical Modules

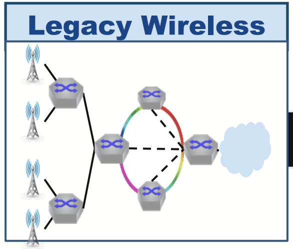

Elec interface: 8x 100G PAM-4
Optical interface: 8x 100G PAM-4
8
8
Electrical Interface
DSP
Optical Engine
8
8
ARISTA 8000

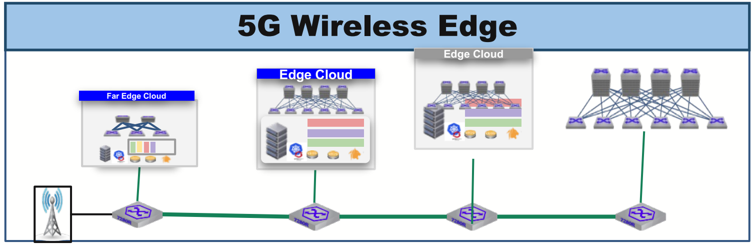
Linear Drive 800G Optical Modules


Elec interface: 8x 100G PAM-4
Optical interface: 8x 100G PAM-4
8
8
Electrical Interface
Optical Engine
8
8
ARISTA 8000

SP Access Networks Evolution

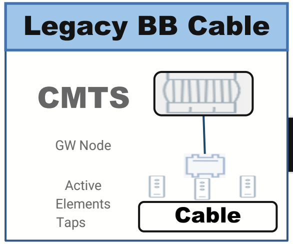
Closed vendor-centric

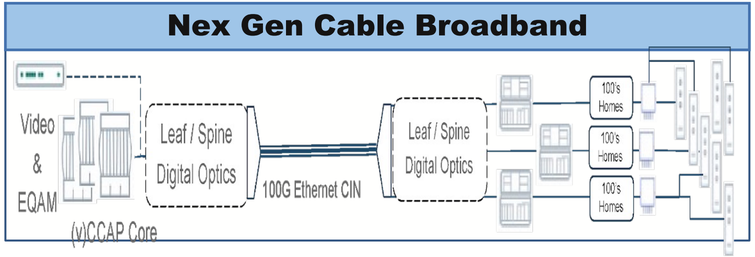

Legacy Wireless


5G Wireless Edge

Open software-centric

Capped scale

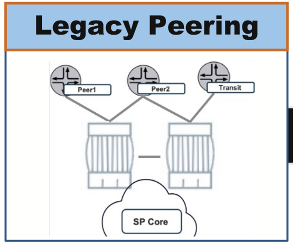

Legacy BB Cable

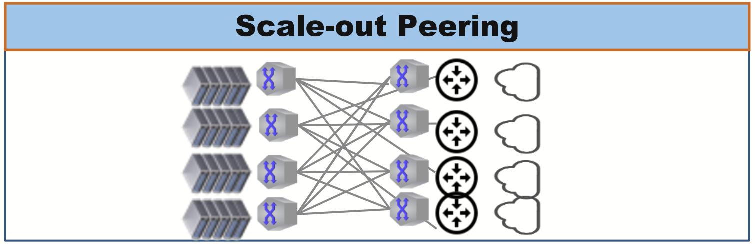

Nex Gen Cable Broadband

Elastic scale

Complex protocols

Open & simplified protocols


Legacy Peering

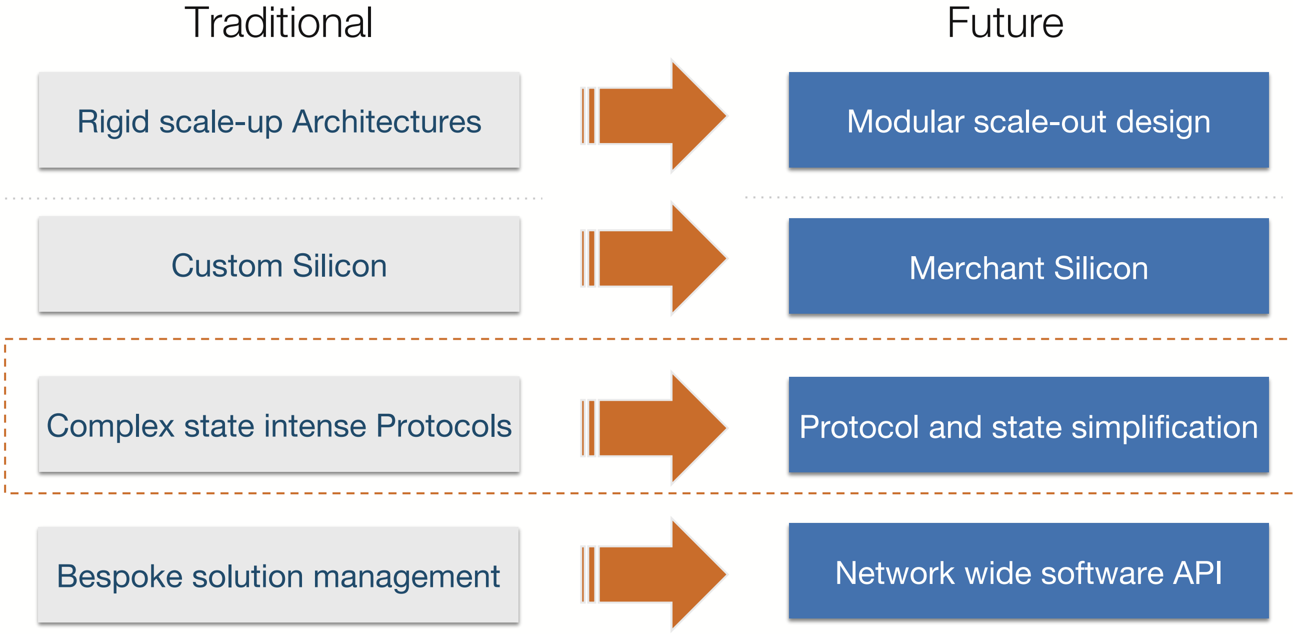

Scale-out Peering

Difficult to deploy & manage

Telemetry driven management

Simplify – The network Edge and core

- To accelerate the adoption of high performance Merchant Silicon
  - Lessons learnt from the Cloud on how to scale without linearly growing CapEx/OpEx costs



Traditional

Future

Rigid scale-up Architectures → Modular scale-out design

Providing elastic capacity, distributed state for better scale, feature and hardware evolution rather rip and replace
*- **evolution rather than revolution***

Custom Silicon → Merchant Silicon

*Merchant silicon for increased performance and reduction in **cost-per-bit***

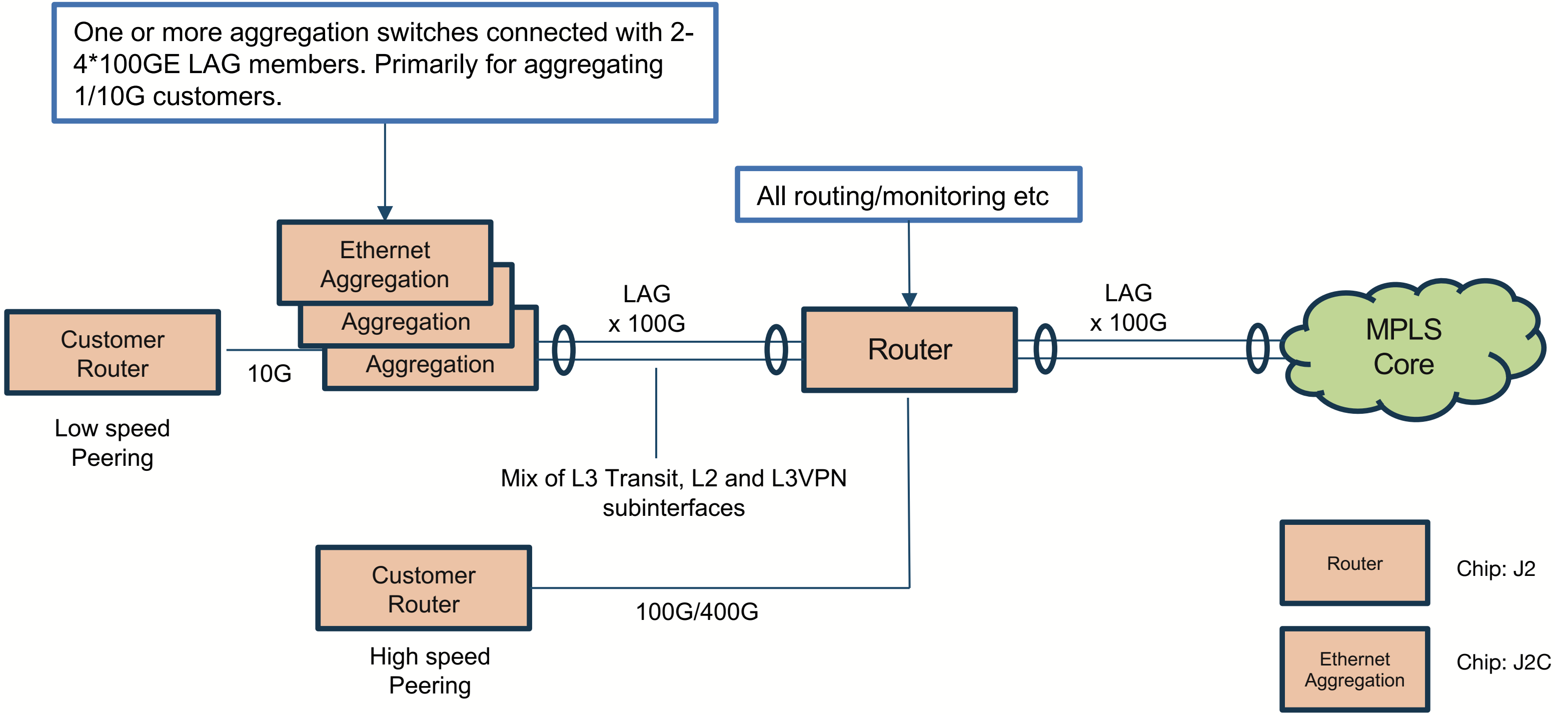
Complex state intense Protocols → Protocol and state simplification

"Keep it **simple** and **consistent, regardless of the service, platform or site size**– to enable automation & performance at scale

Bespoke solution management → Network wide software API

*Service velocity via **closed loop automation, orchestration** & fine grained telemetry*

Example Deployment at the Service Edge

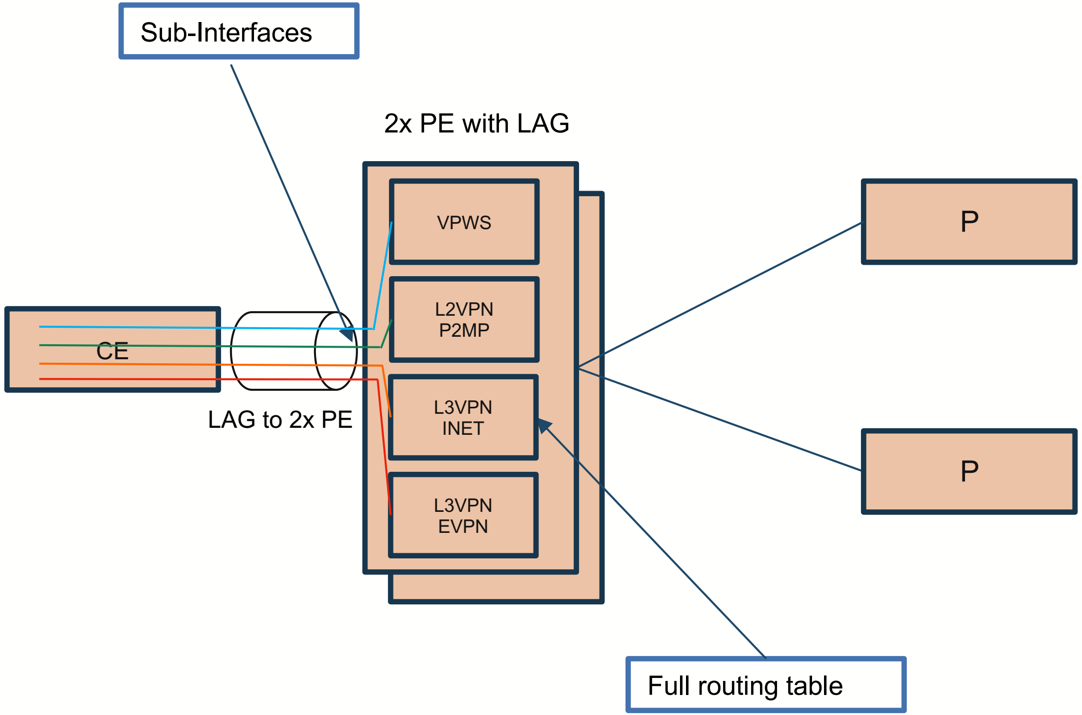


One or more aggregation switches connected with 2-4*100GE LAG members. Primarily for aggregating 1/10G customers.

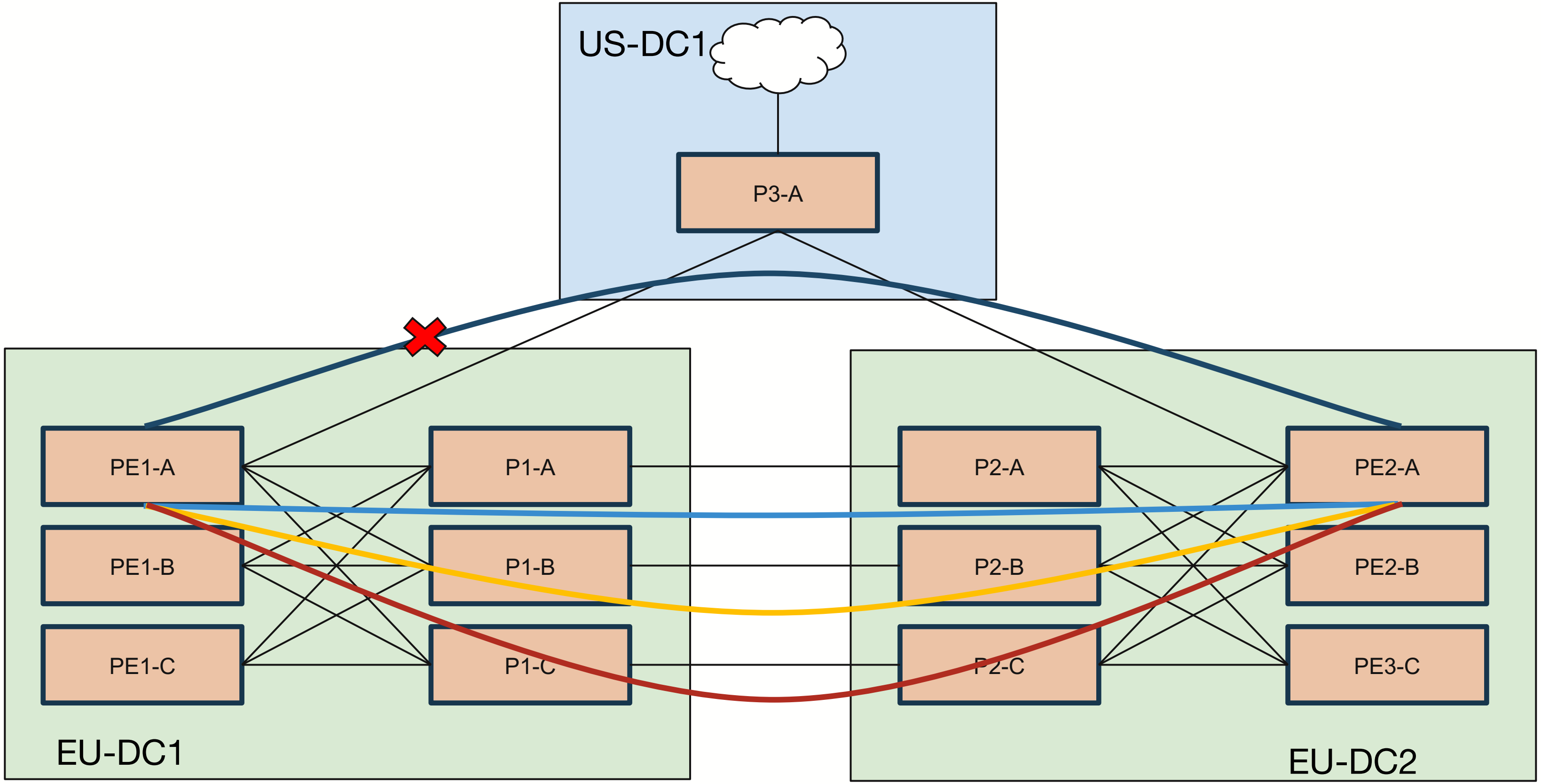
All routing/monitoring etc

Mix of L3 Transit, L2 and L3VPN subinterfaces

Example Deployment at the Service Edge



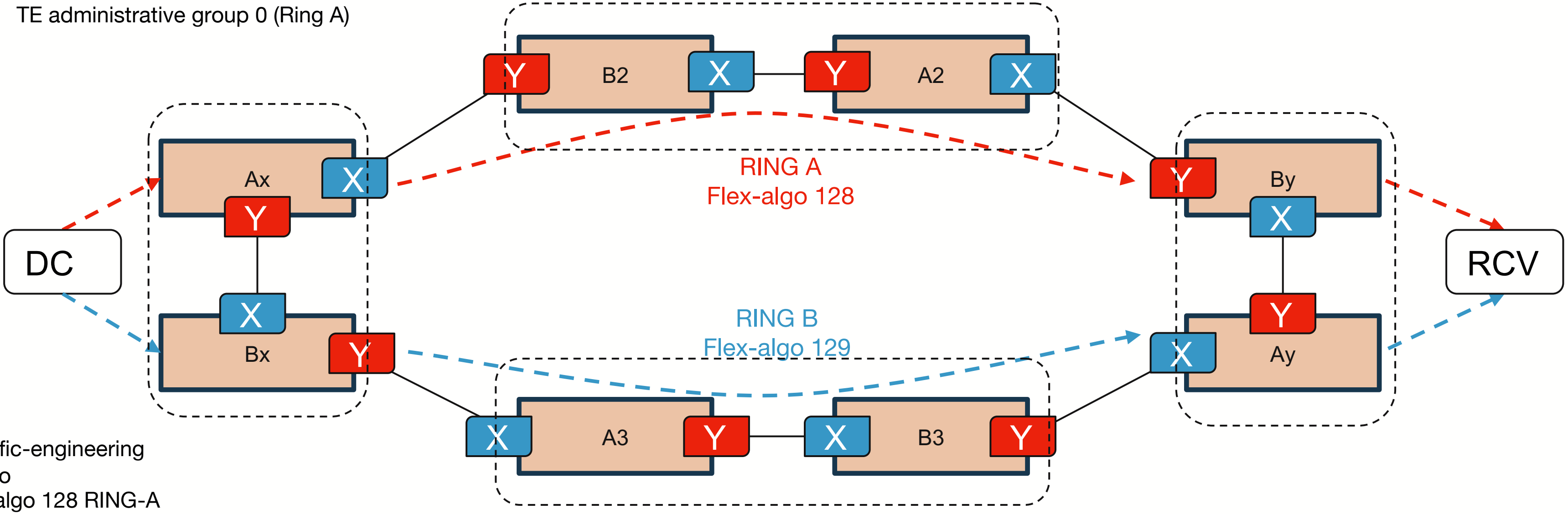
Example Deployment at the core



Example Deployment with Traffic-Engineering

X TE administrative group 1 (Ring B)

Y TE administrative group 0 (Ring A)

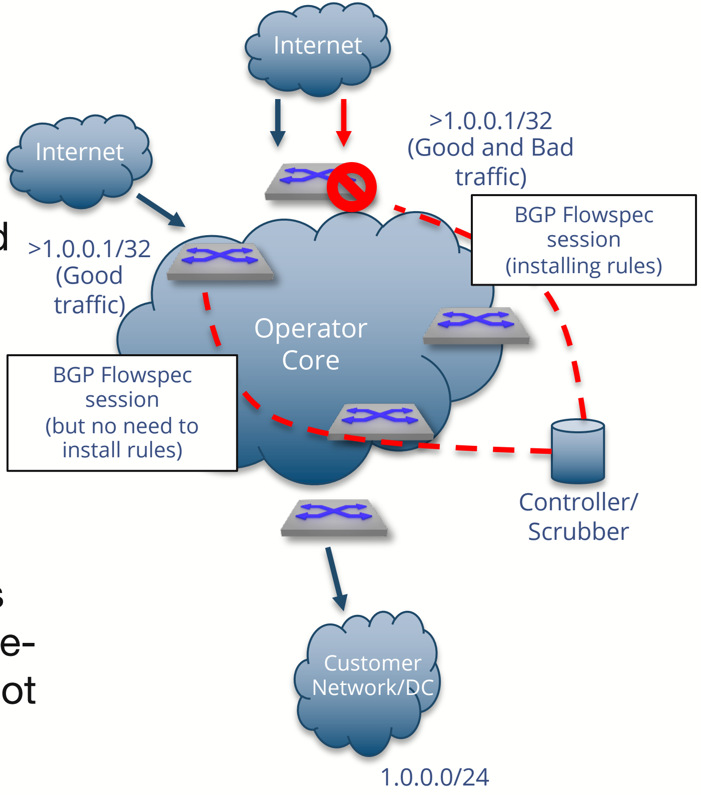


```
router traffic-engineering
   flex-algo
      flex-algo 128 RING-A
         administrative-group include all 0
         color 101
      !
      flex-algo 129 RING-B
         administrative-group include all 1
         color 201
```

Location/PoP

DDoS mitigation using BGP FlowSpec

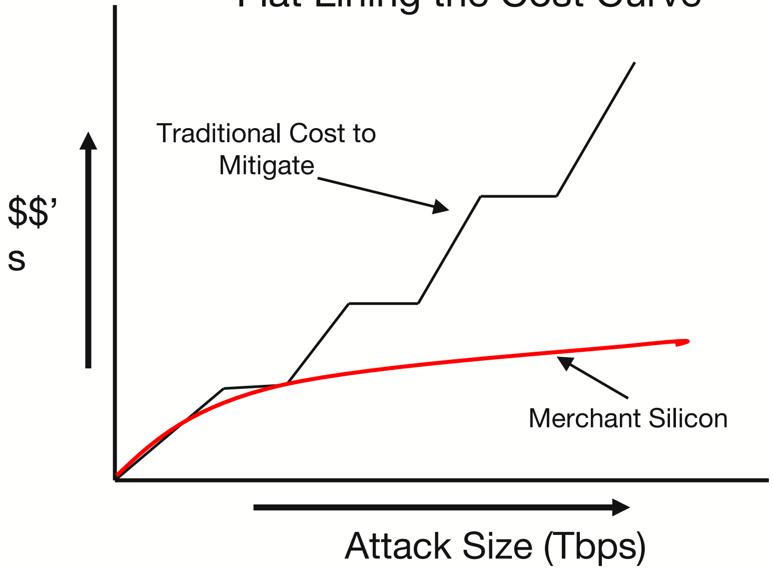
- Capacity
  - The Nodes capacity regards TCAM and other hardware dependency is easy controllable since Flowspec sessions are only where it make sense, and also controlled to each peer(s)
- Flowspec installed only where needed
  - Flowspec specific to Peer/Node/Customer interfaces where traffic enters => Flowspec installed only where it's needed
- Mix of actions
  - Drop, rate-limit, relay can be used based on demands
  - Example stages of them (start with relay, follow by rate-limit and perhaps in the end just drop in case shape not enough)



Merchant Silicon Security Advantage for Service Providers

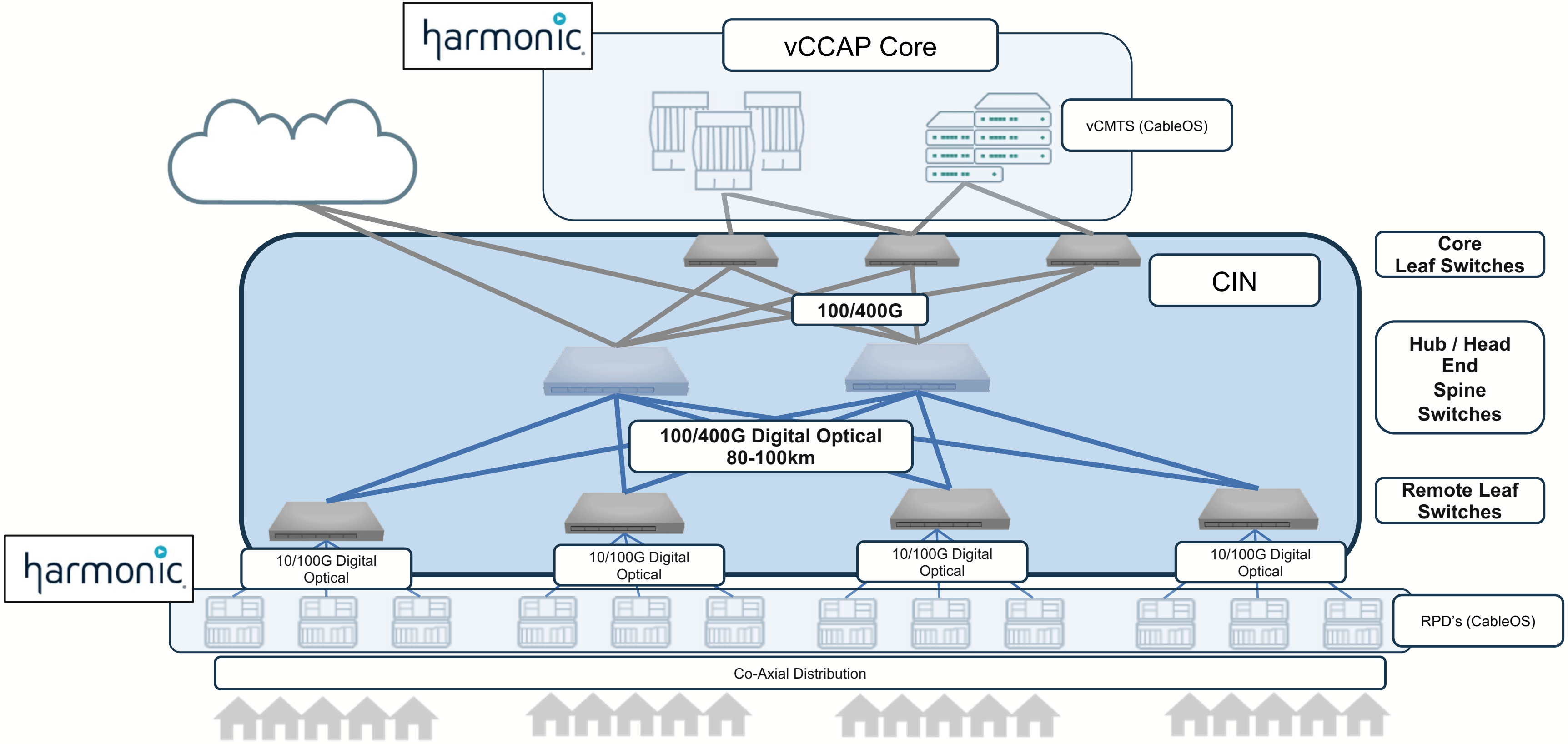
- Cost to Mitigate no longer directly proportional to the size of DDOS Attacks
- High Scale ACL Support
- Elastic Resources to expand or contract based on Volumetric Attacks
- Fine Grain Telemetry

Flat Lining the Cost Curve

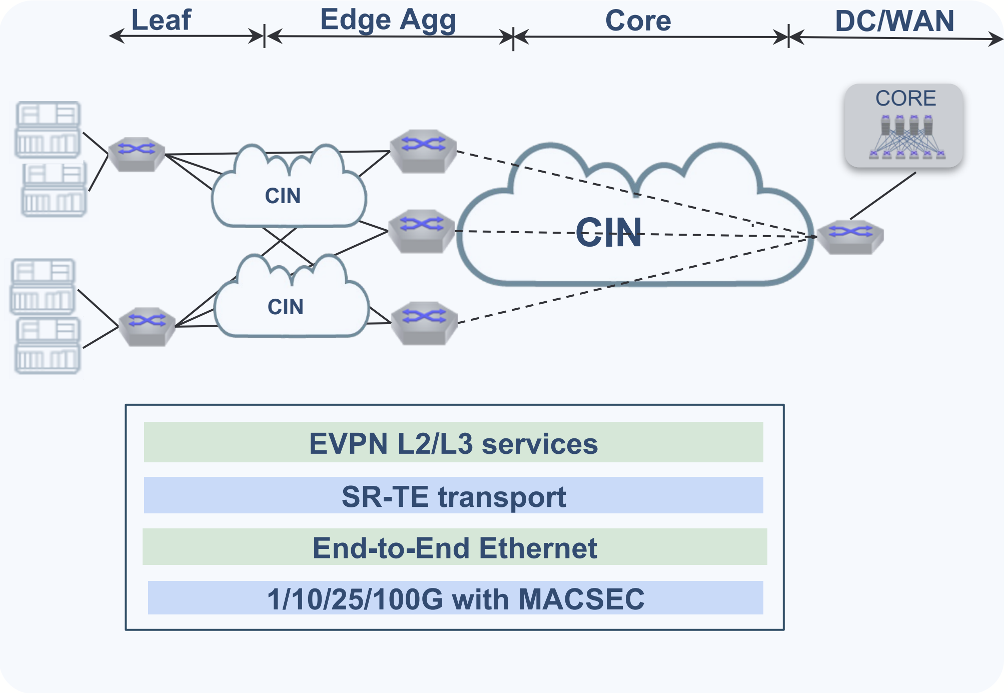


Merchant Silicon bringing Cost Effective DDOS Mitigation Solution

Common R-PHY CIN Topology



Towards Efficient and Simplified CIN Transport



EVPN L2/L3 services

SR-TE transport

End-to-End Ethernet

1/10/25/100G with MACSEC

- Simplified transport and services
  - SR/SR-TE transport
  - EVPN for L2/L3 services
  - Infrastructure slicing with Flex Algo
- CIN-Transport Optimized Routers
  - Compressed environmental footprint
  - Timing/Sync-E support
  - Ultra Low Latency components
  - 1G → 400G with variety of Optics
  - (r)ECMP tunable hashing
- Operational simplicity
  - Hitless software upgrade
  - Full programmability with fine-grained telemetry

Thank You

arista.com