upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática

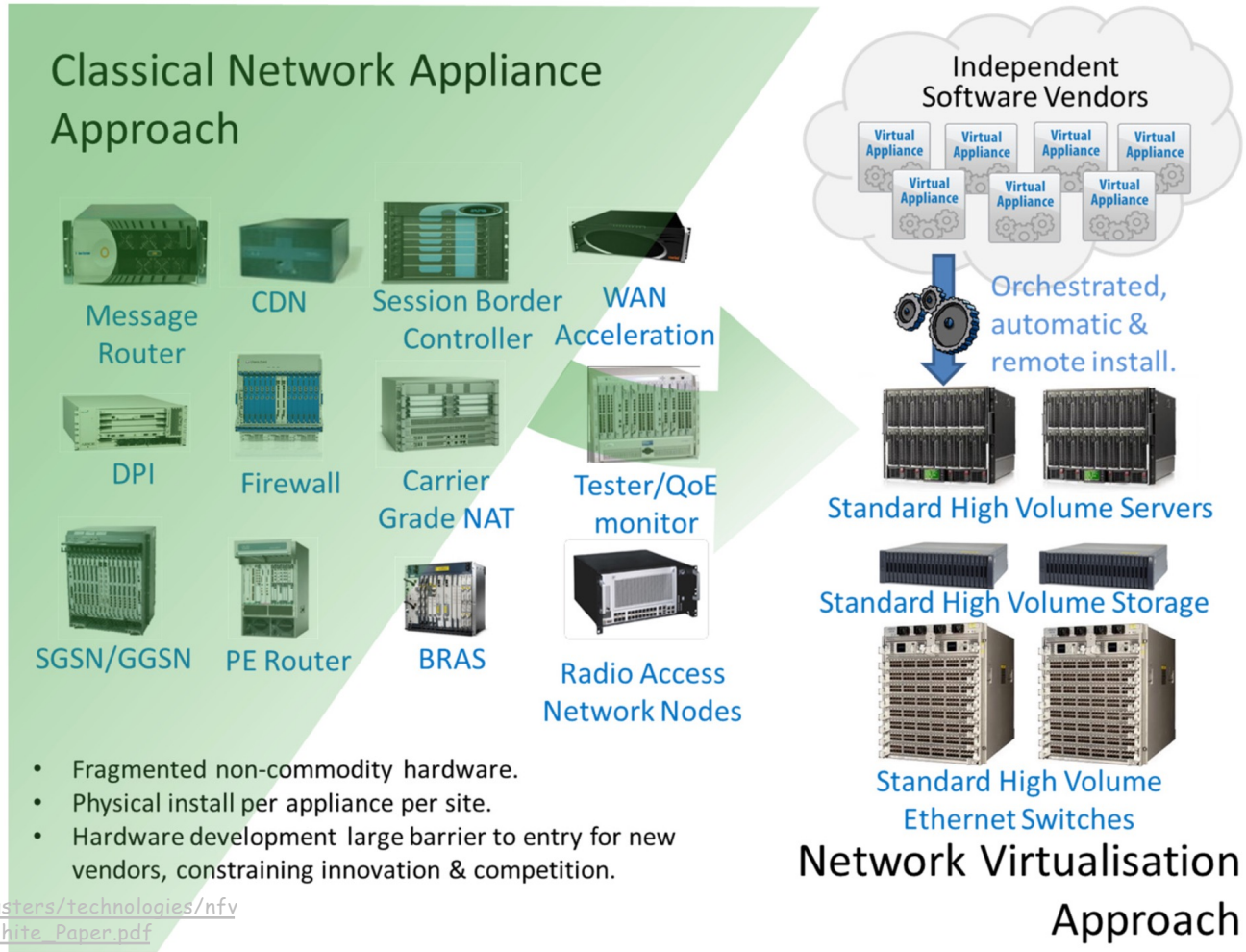
NFV

El problema

- Problema de las operadoras
- Gran cantidad de *appliances*
- Desplegar un nuevo servicio requiere espacio y alimentación para ese nuevo hardware
- Nuevas habilidades de la gente para diseñar, integrar y operar el servicio con ese nuevo hardware
- Ese hardware alcanza su límite de vida con rapidez, lo cual requiere políticas de remplazo que no crean nuevo beneficio
- Los operadores declaran no estar incrementando sus beneficios pero aumentan sus costes (más tráfico, más servicios)

NFV

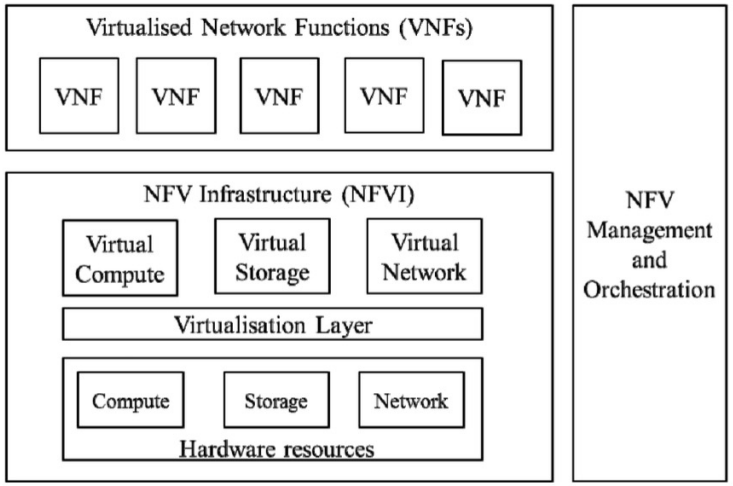
- *Network Functions Virtualisation* (complementario a SDN)
- Se busca mover de hardware dedicado a máquinas virtuales
- Un ISG (*Industry Specification Group*) de ETSI desde finales de 2012
- Hoy más de 200 compañías



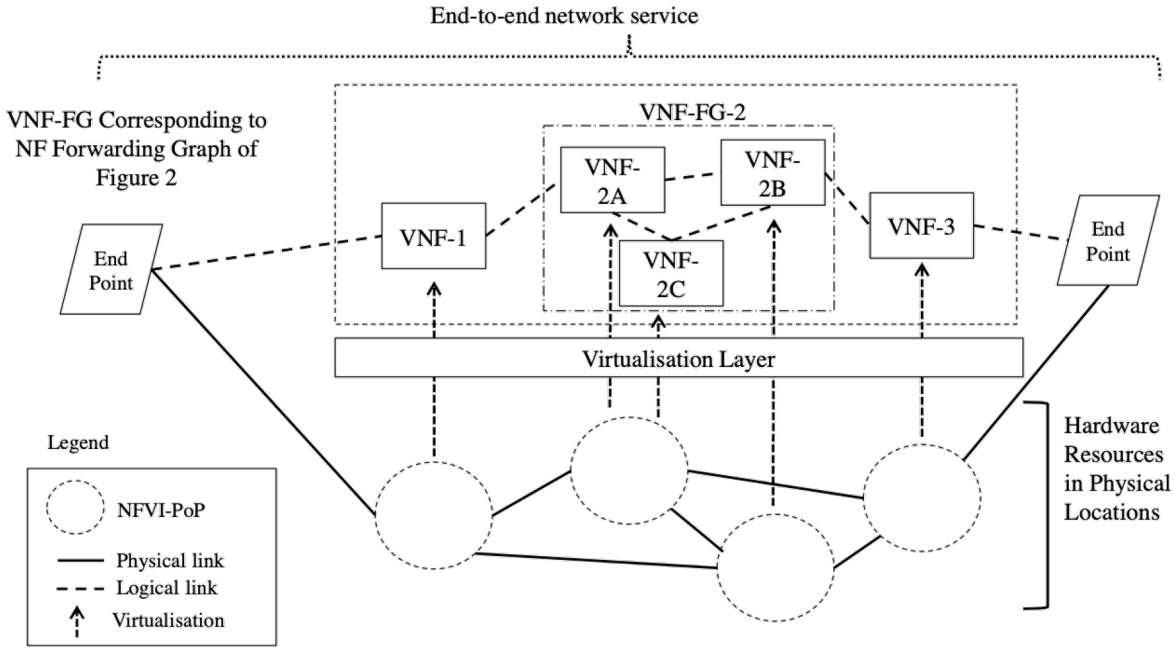
http://www.etsi.org/technologies-clusters/technologies/nfv
https://portal.etsi.org/NFV/NFV_White_Paper.pdf

Arquitectura

- NF = Network Function
- VNF = Virtualised Network Function (implementación de NF)
- NFVI = Network Functions Virtualisation Infrastructure (donde se despliegan VNFs)



Network Function (NF): functional block within a network infrastructure that has well-defined external interfaces and well-defined functional behaviour

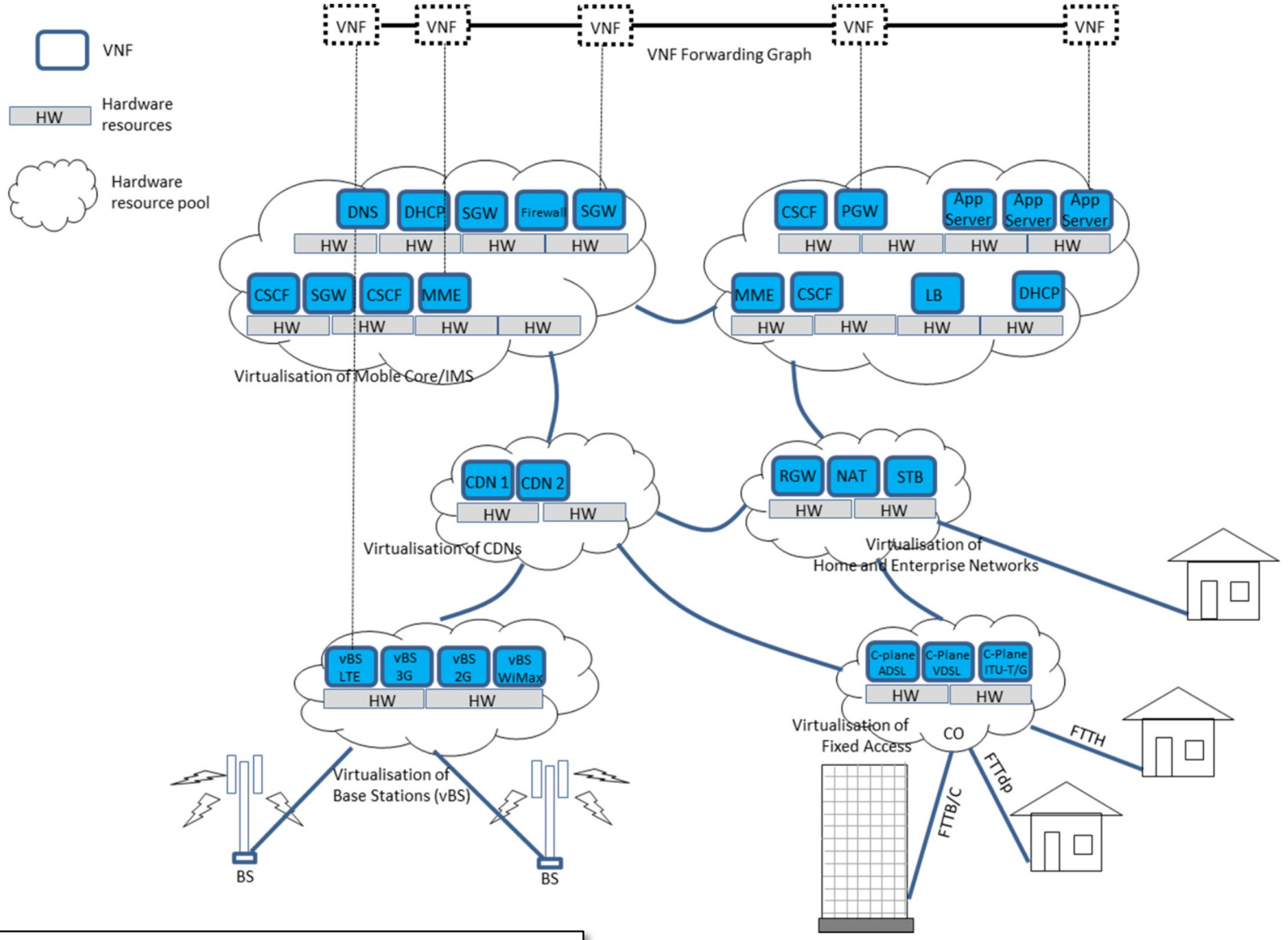


Use cases

- Switching elements: BNG, CG-NAT, routers.
- Mobile network nodes: HLR/HSS, MME, SGSN, GGSN/PDN-GW, RNC, Node B, eNode B.
- Functions contained in home routers and set top boxes to create virtualised home environments.
- Tunnelling gateway elements: IPSec/SSL VPN gateways.
- Traffic analysis: DPI, QoE measurement.
- Service Assurance, SLA monitoring, Test and Diagnostics.
- NGN signalling: SBCs, IMS.
- Converged and network-wide functions: AAA servers, policy control and charging platforms.
- Application-level optimisation: CDNs, Cache Servers, Load Balancers, Application Accelerators.
- Security functions: Firewalls, virus scanners, intrusion detection systems, spam protection.

https://portal.etsi.org/NFV/NFV_White_Paper.pdf

Ejemplos



VNF = Virtualised Network Function

https://portal.etsi.org/NFV/NFV_White_Paper2.pdf

Algunos beneficios

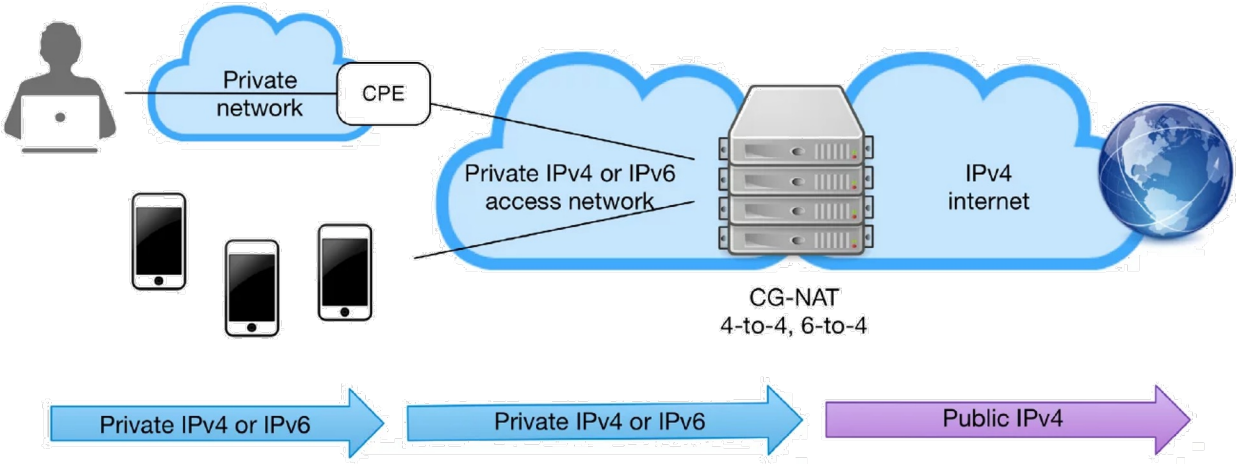
- Reducción de coste de equipos
- Reducción de consumo eléctrico
- Reducción de tiempo de despliegue de un nuevo servicio
- Posibilidad de tener servicios en producción, prueba y desarrollo en la misma infraestructura
- Escalado rápido del servicio
- Abre el mercado a desarrolladores de soft (no necesitan desarrollar hardware)
- Multi-tenancy
- Mejores habilidades existentes para la gestión de infraestructura IT de gran escala que de equipos de red
- Reducción de tiempos de reparación
- Reducción de tiempos de actualización de software
- Etc etc

Facilitadores

- *Cloud Computing*
  - Virtualización (hypervisores, vSwitch, smart NICs)
  - *Orchestration*
  - Open APIs

- Grandes volúmenes de servidores
  - Componentes estándar (por ejemplo x86), vendidos por millones (escala) e intercambiables (competencia)
  - En lugar de *appliances* que dependen de ASICs

Ejemplo: NFV CGNAT



Carrier Grade NAT - Performance and Features

Performance	Throughput per VM up to 370 Gbps
Modes	NAT44, NAT64
Routing	VRF, Static routing, BGP, BFD, OSPF, IS-IS, RIP
Application Layer Gateways	FTP, DNS, PPTP, IPSec, SIP, RTSP
Filtering	EIF Address Dependent Address and Port Dependent
Mapping	EIM (Endpoint Independent)
Logging	Syslog, NetFlow, IPFIX, RADIUS
Advanced Logging Features	Deterministic NAT, Ability to send logs to multiple syslog servers, Port Block Allocation (PBA)
AAA	TACACS+ Radius

Deployment Options

1

VNF (Virtual Network Function)

vCGNAT is purpose-built for virtualized and cloud environments. An integration with any VNF Manager and NFV Orchestration is always smooth and completed within a short time frame.

vCGNAT is officially Open Source MANO (OSM) compatible Virtual Network Function.

2

On top of x86 servers and hypervisor

The vCGNAT VM instance can be deployed on a standard x86 COTS server with a standard server operating system (CentOS/Ubuntu). For this option, it is required to use a Kernel-based Virtual Machine (KVM) hypervisor.

https://nfware.com/virtual-cgnat

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática

NFV

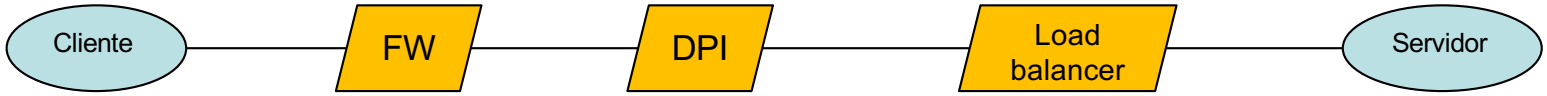
upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

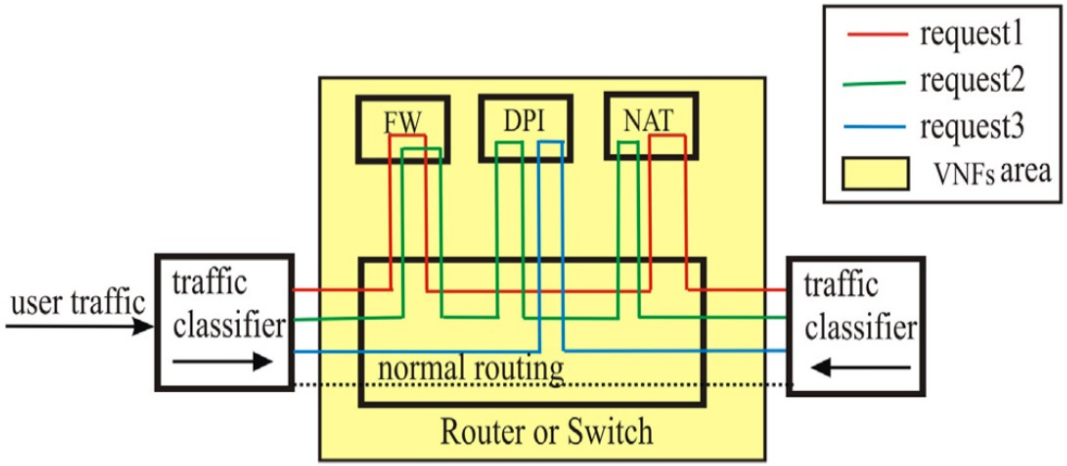
Redes de Nueva Generación
Área de Ingeniería Telemática

SFC

Service Function Chaining

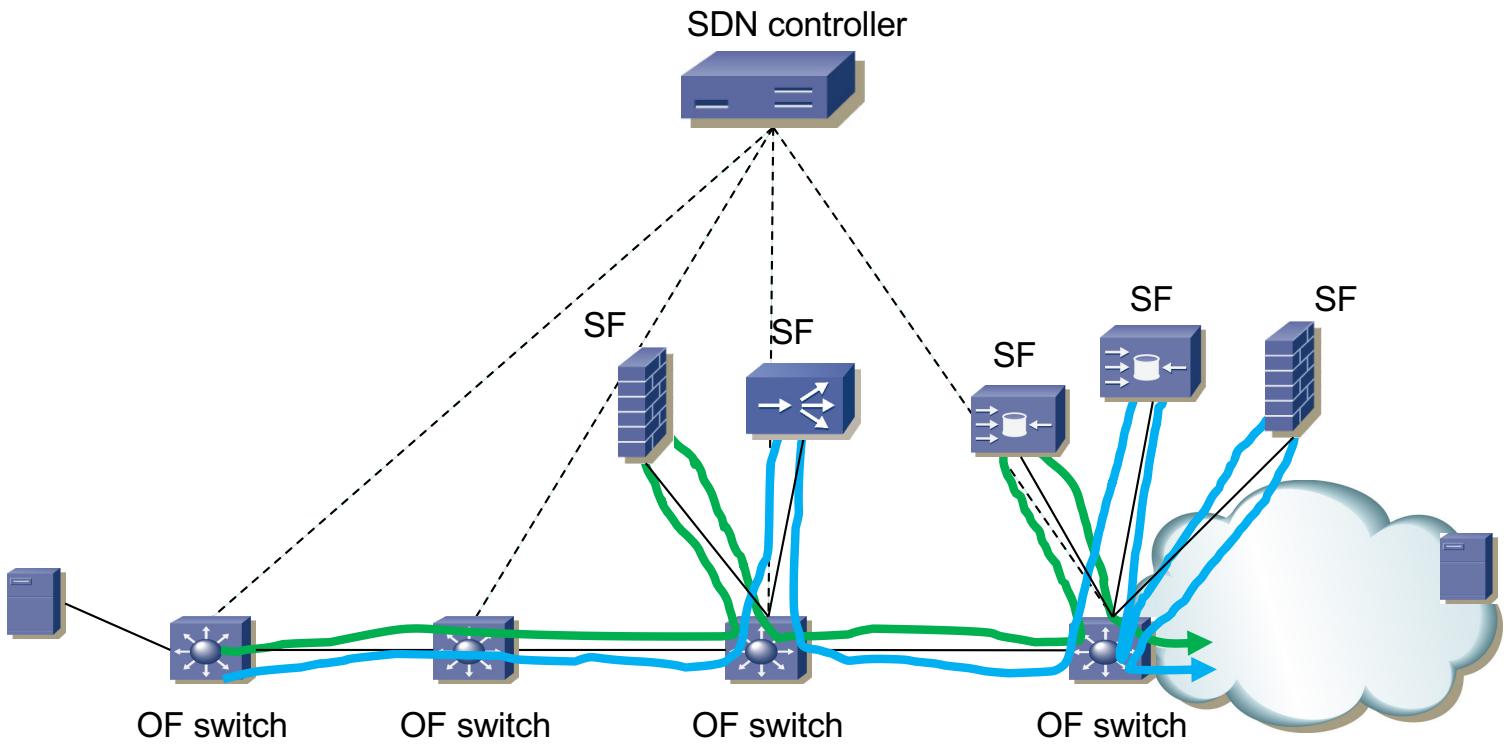
- RFC 7665: “Service Function Chaining (SFC) Architecture” (Ericsson, Cisco, 2015)
- Service functions: firewalls, load balancers, NATs, WAN accelerators, TCP optimizers, DPIs, etc
- SFC: lista ordenada de instancias de estas funciones de servicio
- Service Function Path (SFP)
- Es agnóstico a cómo se transporten los paquetes (por la *underlay*)





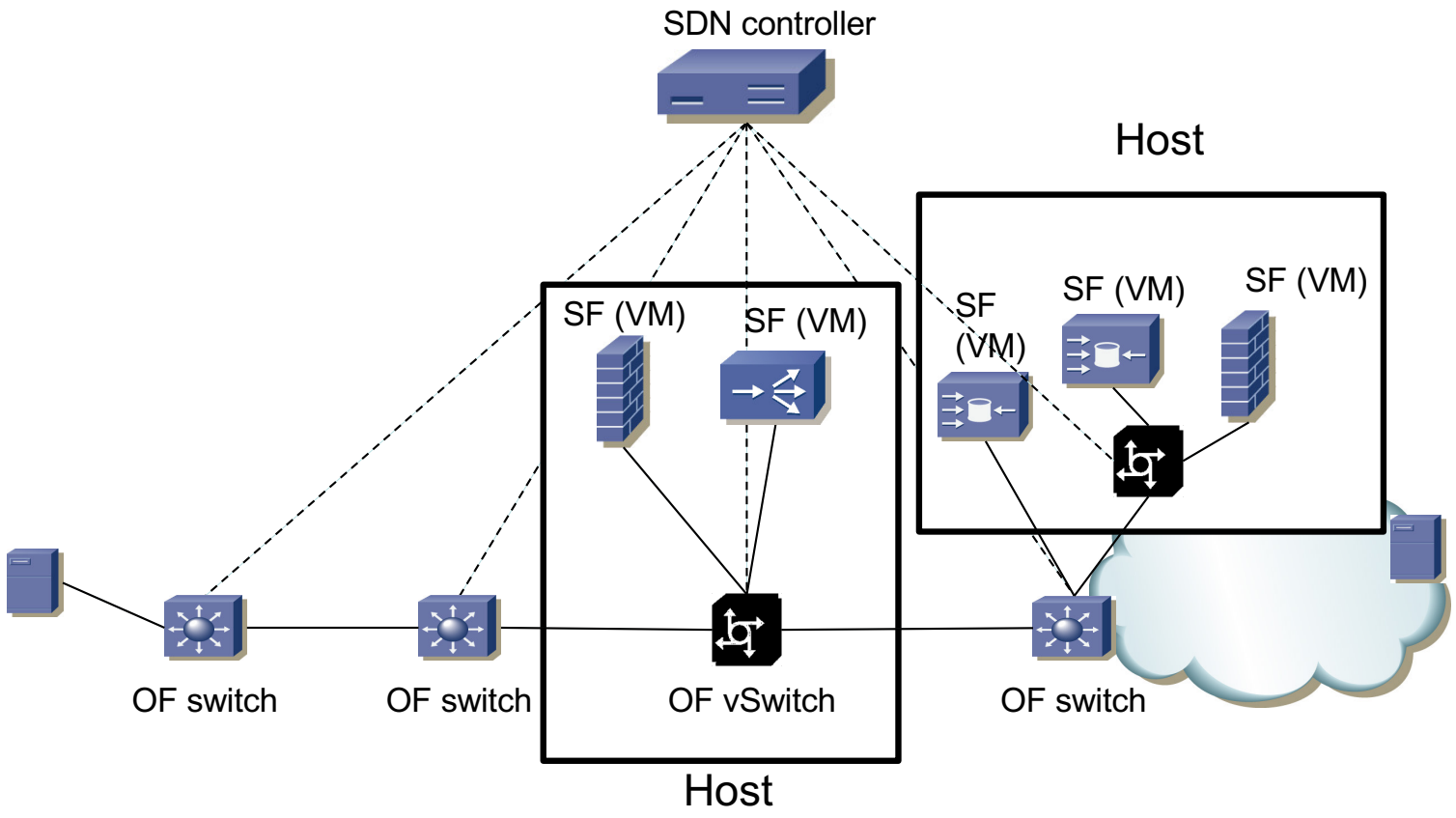
SFC + SDN

- SDN para dirigir los flujos



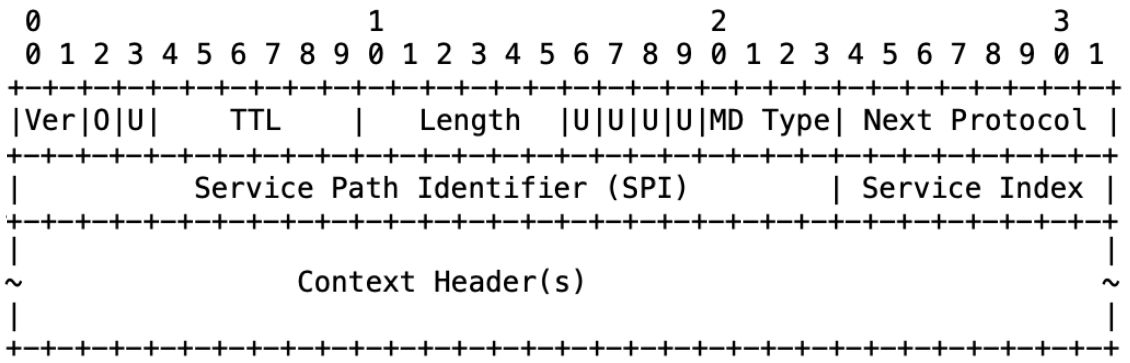
SFC + SDN + NFV

- NFV : La virtualización de los servicios
- SFC : La cadena de servicios por los que debe pasar el flujo
- SDN : El control de la red para llevar a cabo ese camino



NSH

- RFC 8300, “Network Service Header” (Cisco, Intel, 2018)
- TTL (*loop* prevention) indica el máximo número de saltos de servicio (SFFs)
- Next Protocol: IPv4 (0x1), IPv6 (0x2), Ethernet (0x3), NSH (0x4), MPLS (0x5)
- *Service Path Identifier:* identifica al SFP y con él a las SF por las que pasar
- *Service Index:* Da localización dentro del SFP (inicial 255 y lo decrementa cada SF o SFC Proxy)
- Transporta metadatos que pueden necesitar las SFs



SFF = Service Function Forwarder

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática

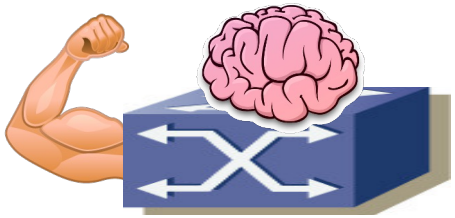
SFC

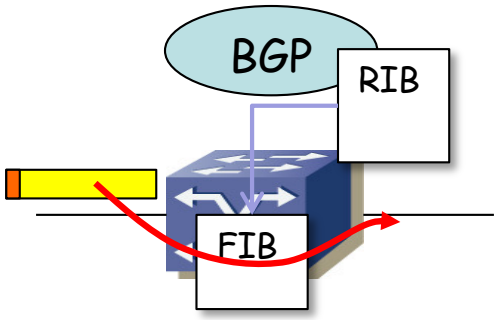
upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

Networking hardware y el software

Planos

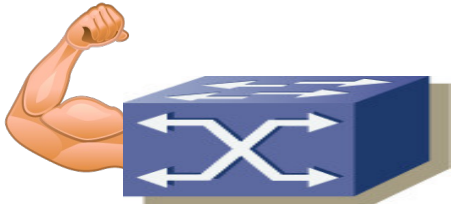
- Plano de control
- Plano de datos

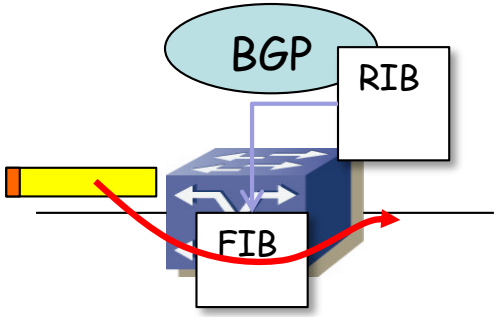




Planos

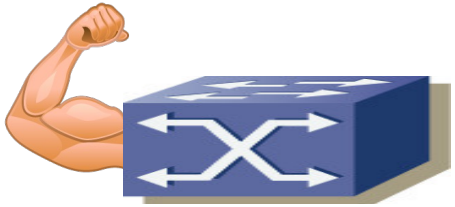
- Plano de control
  - Tradicionalmente atado al hardware
  - SDN lo independiza
- Plano de datos
  - (...)

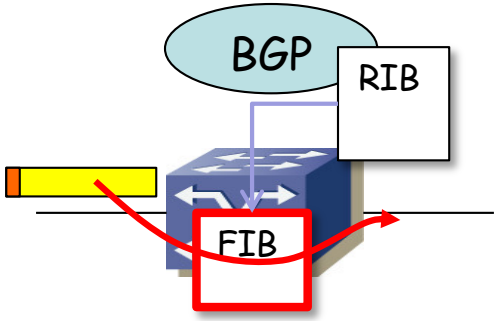




Planos

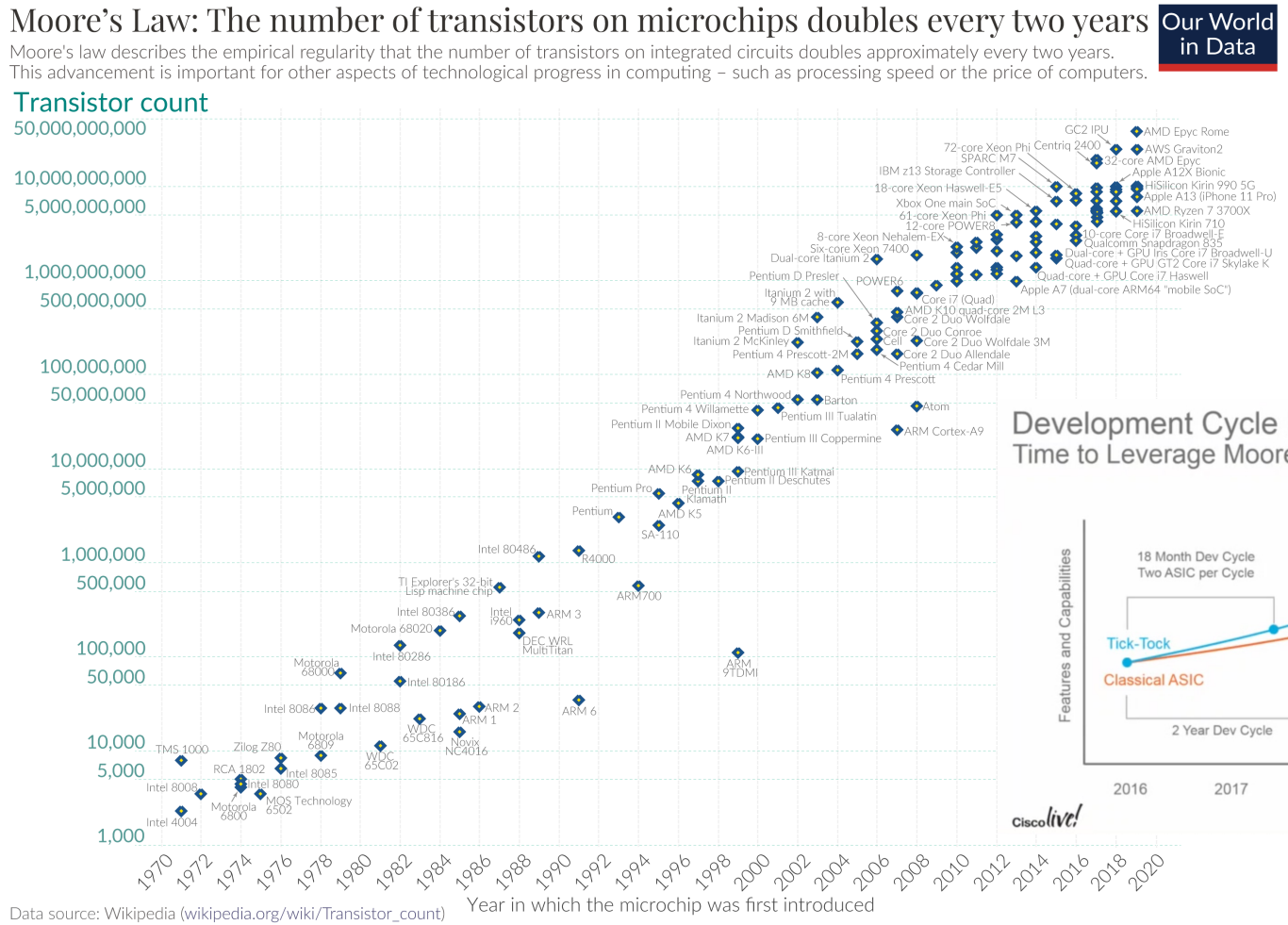
- Plano de control
  - Tradicionalmente atado al hardware
  - SDN lo independiza
- Plano de datos
  - En el propio diseño del hardware
  - (...)



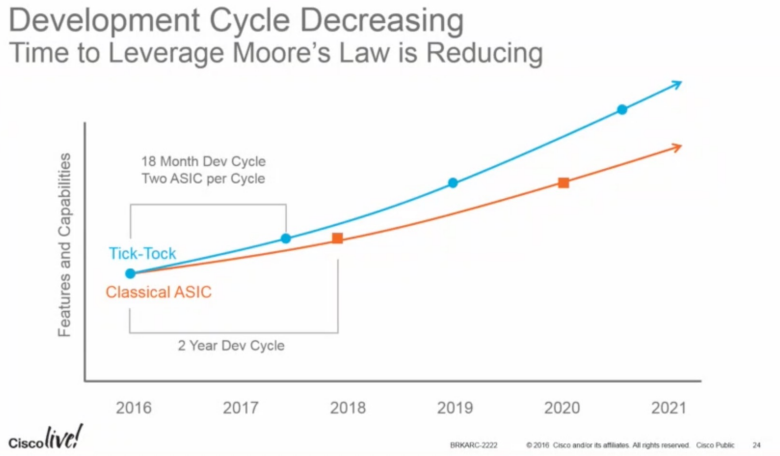


Evolución del hardware

- 40 años de ley de Moore (x2 transistores cada 24 meses)
- Hoy en día ya estamos empezando con la producción en 3 nm
- Tick-tock
- Servidores ciclos de 2 años, networking reutilización 8-10 años (!!)



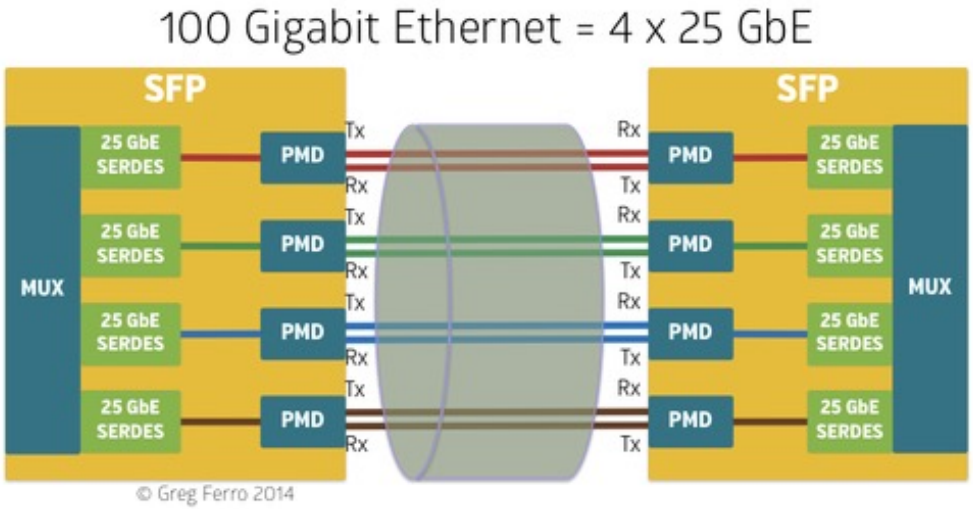
2023 Apple M2 Max: 67×10^9 transistores



Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)
OurWorldinData.org – Research and data to make progress against the world's largest problems.

Evolución del hardware

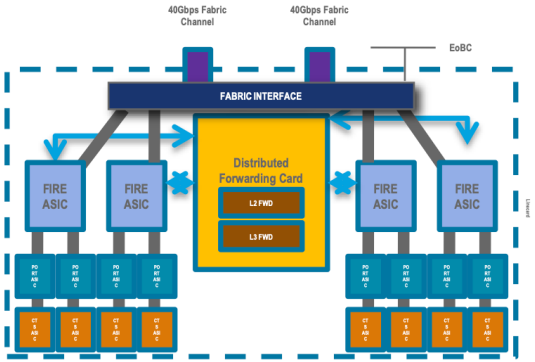
- Durante años ha estado desacoplada la evolución de hardware de computación del hardware de red
- Esto ha hecho que se moviera a software tareas de red (redes virtuales, VXLAN, etc) pues mejores CPUs eran más baratas que mejores switches
- Fabricantes de equipos de red están adoptando los ritmos de producción de electrónica
- Empujados por pocos grandes clientes
- Por ejemplo: donde teníamos SerDes a 10 Gb/s los tendremos posteriormente a 25 Gb/s, al mismo coste
- Esto permite interfaces 100GE (4x25) donde antes teníamos 40GE (4x10), al mismo precio
- Hoy en día SerDes a 112 Gb/s



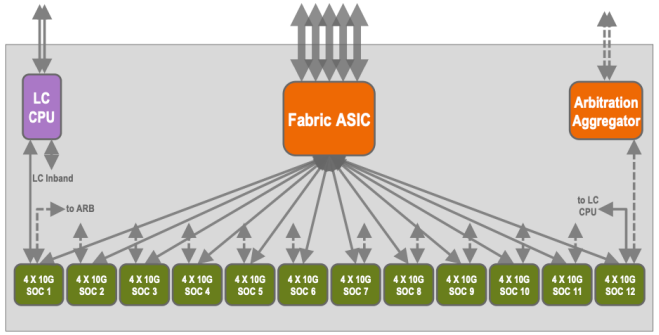
http://www.networkcomputing.com/data-centers/25-gbe-big-deal-will-arrive/1714647938

Evolución del hardware

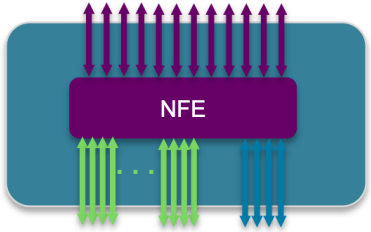
- Lo que era un switch modular puede ser ahora un SoC
- A día de hoy SoC (Switch on Chip) a varios Tbps
- En los últimos 20 años
  - Acceso a DRAM: x90
  - Número de transistores: x8.000
  - Capacidad en switch: x30.000
    (nº puertos x bandwidth)



32 x 10G Ports



48 x 10G Ports



64 x 10G Ports

Design Shifts Resulting from Increasing Gate Density and Bandwidth

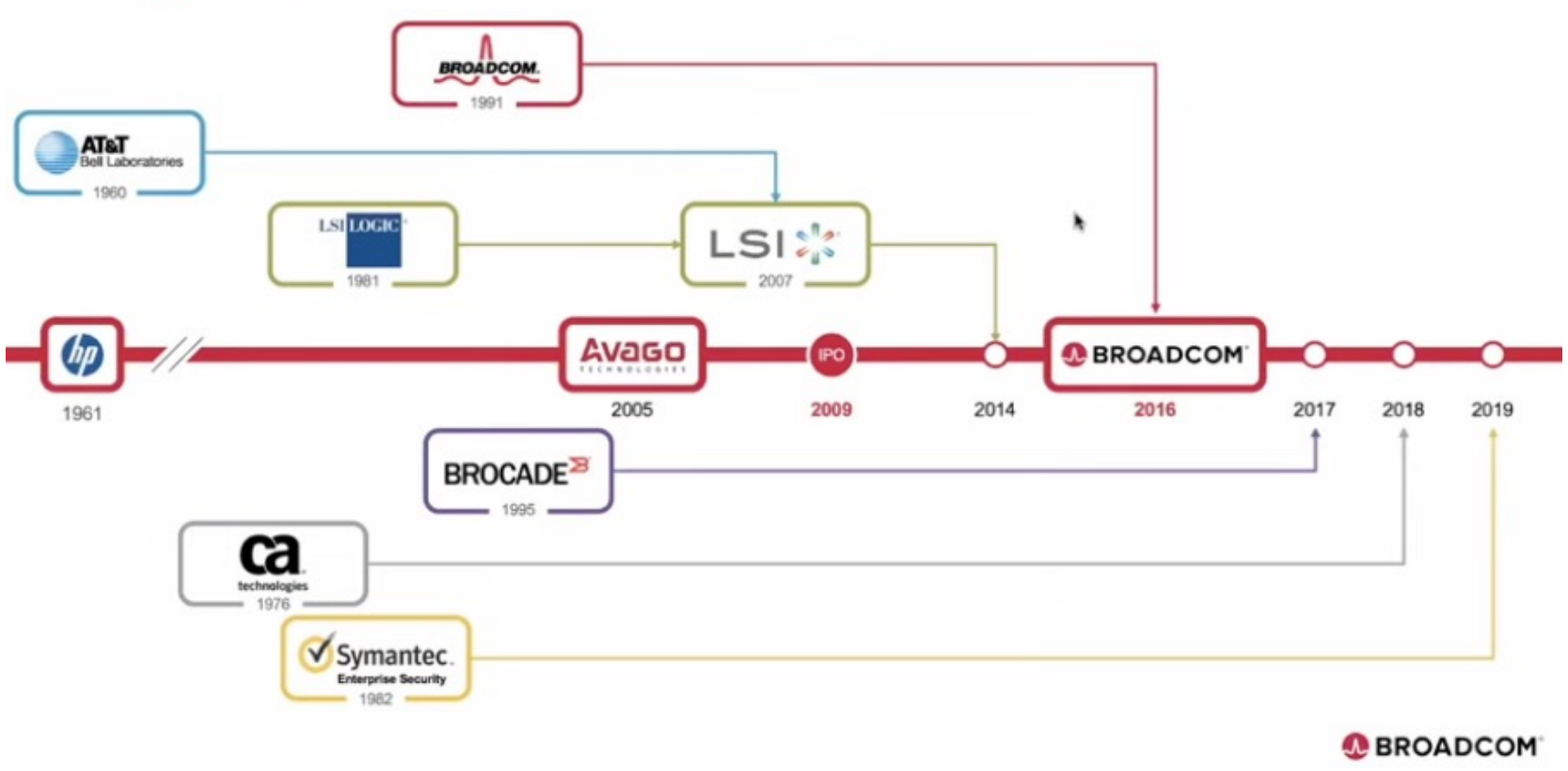
Merchant silicon

- ASICs creados por una compañía pero switches ensamblados por otra
- Broadcom
- Marvell
- Barefoot Networks (ahora Intel)
- Mellanox (ahora Nvidia)

MARVELL™

tsmc

Ejemplo: Broadcom

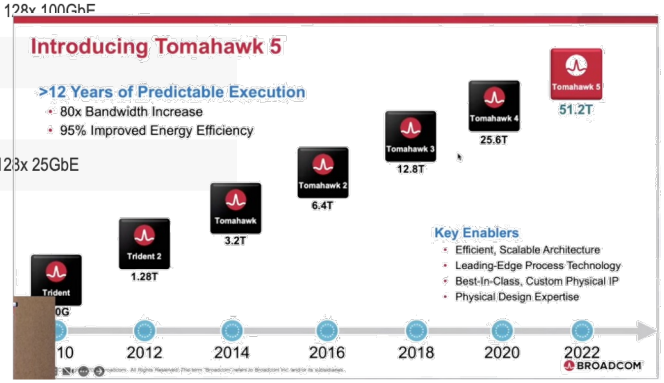


https://youtu.be/g7vcuV2_jSM?si=rkXuTUBihJLfbnoU

Broadcom

- Trident : Feature-rich (programmable for cloud Edge and Enterprise)
- Tomahawk : Hyperscale Fabrics
- Jericho : Scale-out, programmable, Deep buffered, Carrier-grade infrastructure

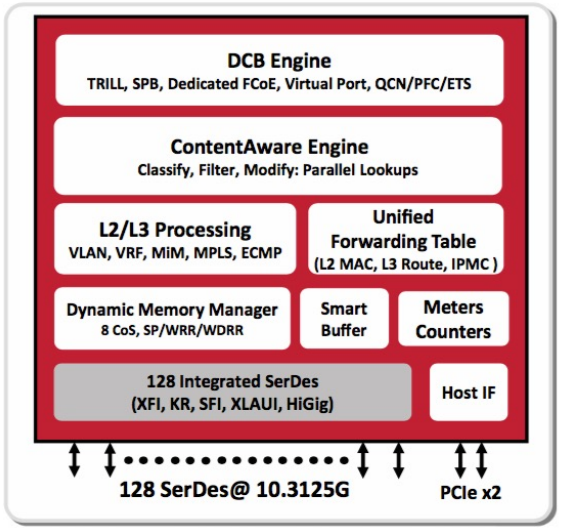
Product Line	Part Number	Bandwidth	Type	Sample I/O Configurations
StrataXGS® Switch Solutions	Trident 5 / BCM78800 Series			
StrataXGS® Switch Solutions	Trident 4 / BCM56690 Series			
StrataXGS® Switch Solutions	Wolfhound3+ / BCM53650	104 Gb/s	L3 Ethernet Switch	24x1GbE + 8x10GbE
StrataXGS® Switch Solutions	Trident4-X11C / BCM56890 Series	12.8 Tb/s	L3 Programmable Ethernet Switch	32x 400GbE, 64x 200GbE, 128x 100GbE
StrataXGS® Switch Solutions	BCM56080 Series		L3 Ethernet Switch	16x 25GbE, 20x 2.5GbE + 8x 10GbE
StrataXGS® Switch Solutions	Tomahawk 5 / BCM78900 Series	51.2 Tb/s	L3 Multilayer Switch	64 x 800GbE, 128 x 400GbE, 256 x 200GbE
StrataXGS® Switch Solutions	Tomahawk4 / BCM56990 Series	25.6 Tb/s	L3 Multilayer Switch	64 × 400GbE, 128 × 200GbE, or 256 × 100GbE
StrataXGS® Switch Solutions	Tomahawk3 / BCM56980 Series	12.8 Tb/s	L3 Ethernet Switch	32x 400GbE, 64x 200GbE, or 128x 100GbE
StrataXGS® Switch Solutions	Tomahawk2 / BCM56970 Series	6.4 Tb/s	L3 Ethernet Switch	64x 100GbE, 128x 40GbE
StrataXGS® Switch Solutions	Tomahawk / BCM56960 Series	3.2 Tb/s	L3 Ethernet Switch	32x 100GbE, 64x 40GbE, 128x 25GbE
StrataXGS® Switch Solutions	Trident4 / BCM56880 Series	12.8 Tb/s	L3 Programmable Ethernet Switch	32x 400GbE, 64x 200GbE, 128x 100GbE
StrataXGS® Switch Solutions	Trident2+ / BCM56860 Series	1.28 Tb/s	L3 Ethernet Switch	32x 40GbE, 104x 10GbE
StrataXGS® Switch Solutions	Trident2 / BCM56850 Series	1.28 Tb/s	L3 Ethernet Switch	32x 40GbE, 104x 10GbE
StrataXGS® Switch Solutions	Trident3-X7 / BCM56870 Series	3.2 Tb/s	L3 Programmable Ethernet Switch	32x 100GbE, 64x 40GbE, 128x 25GbE

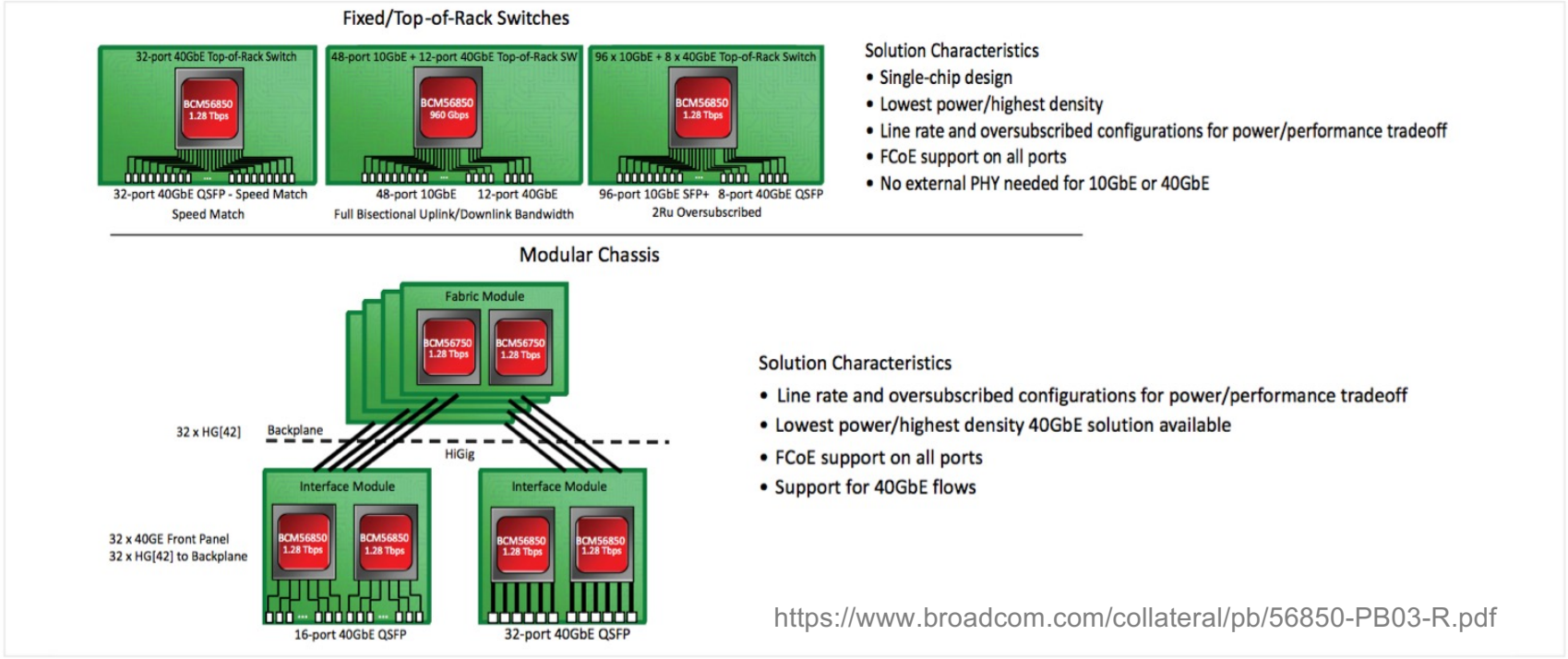


https://www.broadcom.com/products/ethernet-connectivity/switching/strataxgs

Broadcom Trident 2

- 1.28 Tbps con puertos 10GE/40GE
- 128 SerDes 10GE (así que un máximo de 32 puertos 40GE en base a 4x10GE)
- Cut-through y Store&Forward
- VXLAN, NVGRE, 802.1Qbg EVR, 802.1BR
- Per VM traffic shaping
- DCB PFC, QCN y ETS. FCoE
- MPLS, VPLS, ISATAP, MAC-in-MAC, TRILL, SPB, Q-in-Q



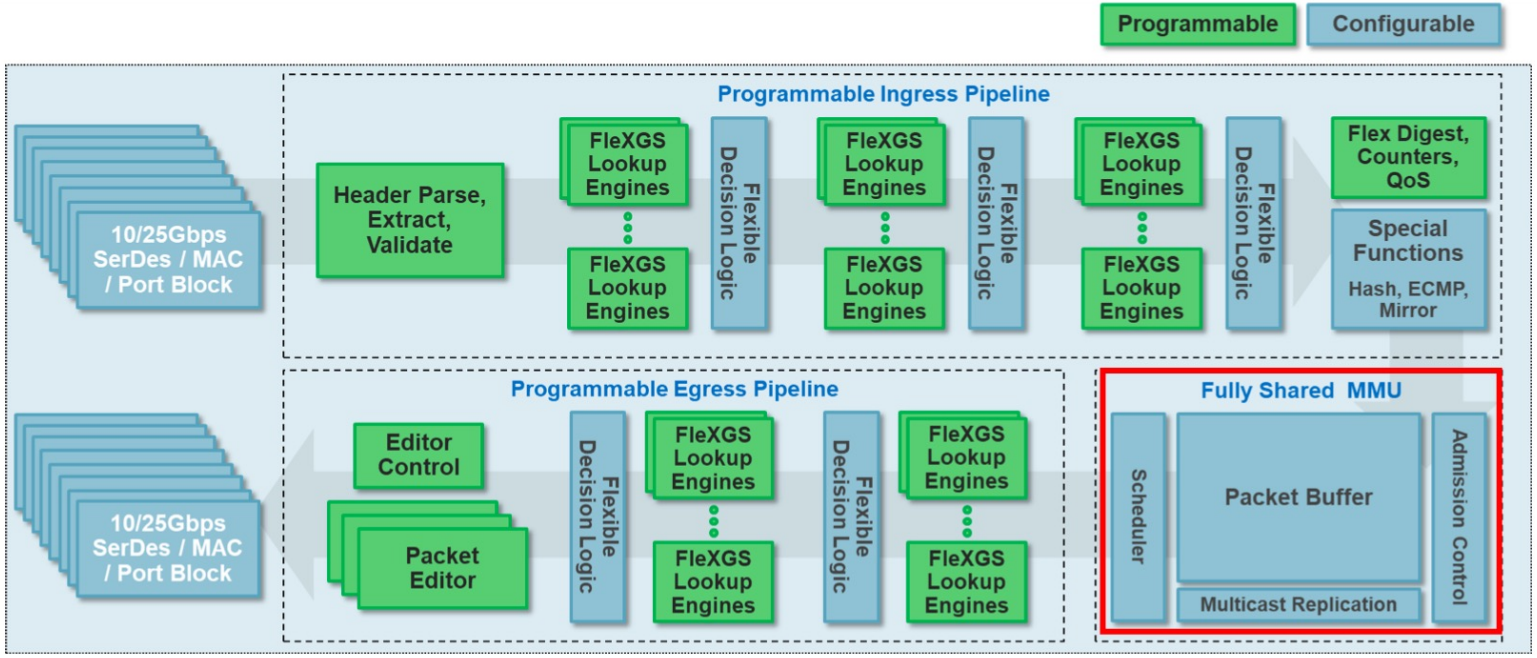


https://www.broadcom.com/collateral/pb/56850-PB03-R.pdf

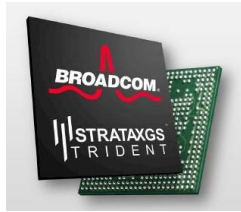
Broadcom Trident 3

- Conmutación a 3.2 Tbps para paquetes a partir de 250 bytes
- Para paquetes de 64 bytes da un throughput de 2 Tbps
- 32 x 100GE, cada uno divisible en 4x10GE, 4x25GE, 2x50GE o 1x40GE
- 32 MB fully shared packet buffer
- SerDes 25Gbps
- Support for new overlays and tunneling such as GENEVE, NSH, VXLAN, GPE, MPLS, MPLS over GRE/UDP, GUE, ILA and PPPoE

Figure 2. Broadcom Trident 3 Switch Silicon Internal Architecture Diagram



Source: Enterprise Strategy Group



https://docs.broadcom.com/doc/12395356

https://www.broadcom.com/products/ethernet-connectivity/switching/strataxgs/bcm56870-series

Trident 2 y Trident 3

- Memoria (SRAM, TCAM) particionable para diferentes usos del switch (muchas MACs, muchas rutas IPv4, etc)

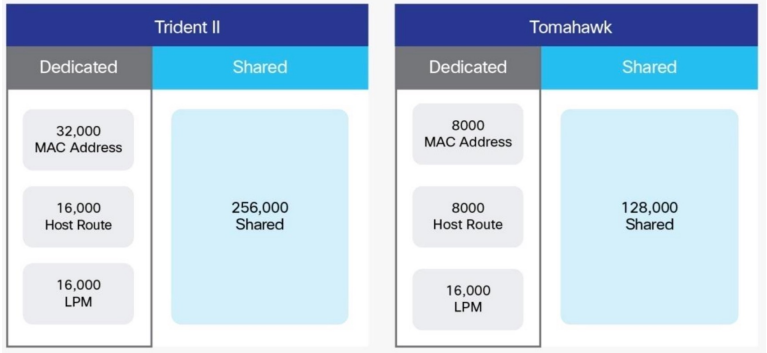


Table 1. Broadcom Trident 2 Forwarding Tables

Mode	Dedicated Layer 2	Shared Memory bank 1	Shared Memory bank 2	Shared Memory bank 3	Shared Memory bank 4	Host Route Dedicated	LPM Dedicated
Mode 0	32,000	Layer 2 (64,000)	Layer 2 (64,000)	Layer 2 (64,000)	Layer 2 (64,000)	Layer 3 (16,000)	16,000
Mode 1	32,000	Layer 2 (64,000)	Layer 2 (64,000)	Layer 2 (64,000)	Layer 3 (40,000)	Layer 3 (16,000)	16,000
Mode 2	32,000	Layer 2 (64,000)	Layer 2 (64,000)	Layer 3 (32,000)	Layer 3 (40,000)	Layer 3 (16,000)	16,000
Mode 3	32,000	Layer 2 (64,000)	Layer 3 (32,000)	Layer 3 (32,000)	Layer 3 (40,000)	Layer 3 (16,000)	16,000
Mode 4	32,000	LPM (32,000)	LPM (32,000)	LPM (32,000)	LPM (32,000)	Layer 3 (16,000)	16,000

Table 2. Broadcom Tomahawk Forwarding Tables

Mode	Dedicated Layer 2	Shared Memory bank 1	Shared Memory bank 2	Shared Memory bank 3	Shared Memory bank 4	Host Route Dedicated	LPM Dedicated
Mode 0	8000	Layer 2 (32,000)	Layer 2 (32,000)	Layer 2 (32,000)	Layer 2 (32,000)	Layer 3 (8000)	16,000
Mode 1	8000	Layer 2 (32,000)	Layer 2 (32,000)	Layer 2 (32,000)	Layer 3 (32,000)	Layer 3 (8000)	16,000
Mode 2	8000	Layer 2 (32,000)	Layer 2 (32,000)	Layer 3 (32,000)	Layer 3 (32,000)	Layer 3 (8000)	16,000
Mode 3	8000	Layer 2 (32,000)	Layer 3 (32,000)	Layer 3 (32,000)	Layer 3 (32,000)	Layer 3 (8000)	16,000
Mode 4	8000	LPM (32,000)	LPM (32,000)	LPM (32,000)	LPM (32,000)	Layer 3 (8000)	16,000

http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-736863.pdf

Tomahawk 4

- 2.5 años de desarrollo
- 25.6 Tb/s
- 7nm, 31.000 millones de transistores
- 8.000 pins
- 512 x 50G PAM-4 SerDes
- 4 cores ARM 1GHz (para telemetría)

	Broadcom Tomahawk 4 BCM56990	Broadcom Tomahawk 3 BCM56980
Bandwidth	25.6Tbps	12.8Tbps
Serdes	512x50Gbps PAM4	256x50Gbps PAM4
Network Ports	64x400GbE, 128x200GbE, 256x100GbE	32x400GbE, 64x200GbE, 128x100GbE
Host Interface	PCIe Gen3 x4	PCIe Gen3 x4
Buffer Memory	Unified, undisclosed	Unified, 64MB
IPv4 Addresses	>750K routes*	>750K routes
ECMP Members	64K*	64K
Latency (L3)	450ns*	450ns
IC Process	TSMC N7	TSMC 16FFC
Power (typ)	450W*	300W
Availability	Samples 4Q19	Production 3Q18

Table 1. Comparison of Tomahawk switch generations. Externally, Tomahawk 4 is nearly identical to its predecessor but has twice the port count. (Source: Broadcom, except *The Linley Group estimate)

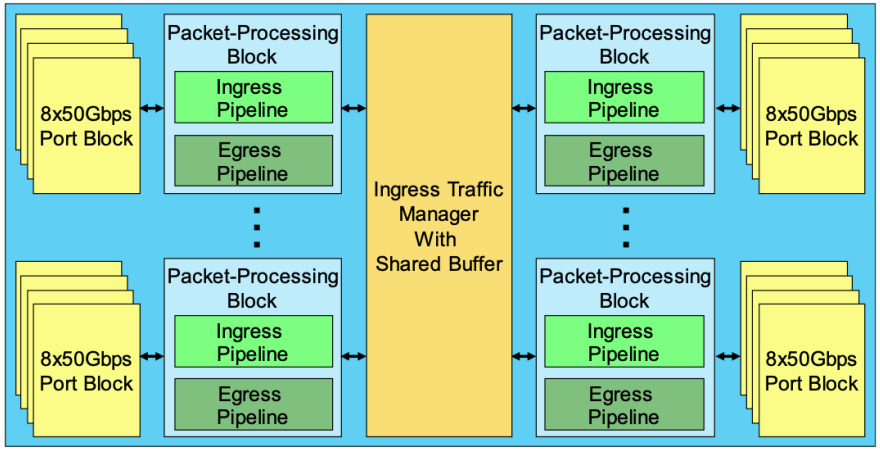


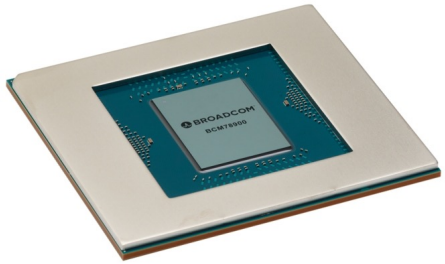
Figure 1. Tomahawk 4 switch chip. The top-level architecture carries over from Tomahawk 3, but the new chip instantiates 64 port blocks, each with eight serdes.

Tomahawk4: Best-in-Class Instrumentation

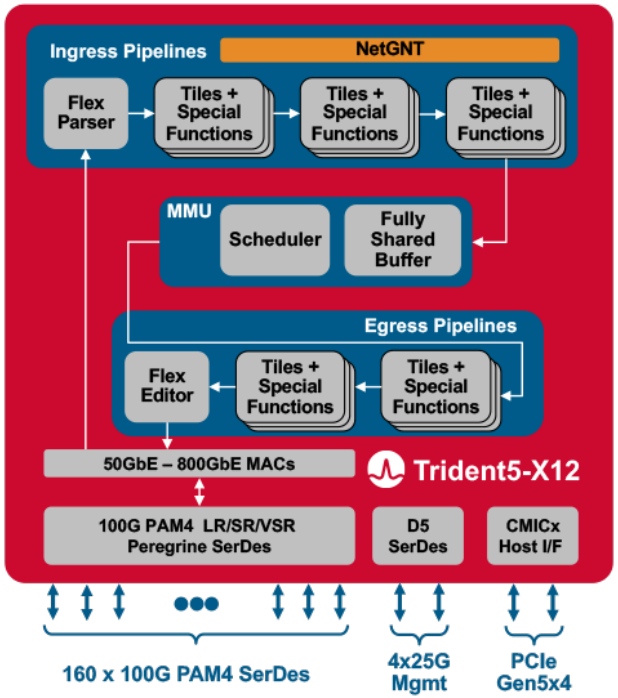
- In-band telemetry – packet tracing and latency monitoring
- Postcards
- Flight-Data Recorder – provides real-time SerDes link quality meters
- Visibility into all packet drops
- Flow and queue tracking
- Microburst and elephant-flow detection
- Programmable export formats
- ARM processors for statistics processing and summarization

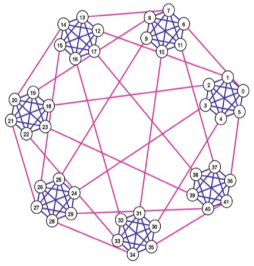
https://docs.broadcom.com/doc/12398014
https://youtu.be/B-COGMbaUg4

Tomahawk 5



- Up to 51.2 Tb/s on a single chip
- 64 × 800GbE, 128 × 400GbE, or 256 × 200GbE
- 64 integrated SerDes cores, each with eight integrated 106-Gb/s PAM4
- L2 and L3 switching, routing, and tunneling
- Support for Clos and non-Clos topologies such as torus, Dragonfly, Dragonfly+, and Megafly
- 9.352 pins
- Telemetría programable (6 cores ARM)





(a) "Canonical" Dragonfly with $a = 6$, $g = 7$, $h = 1$.

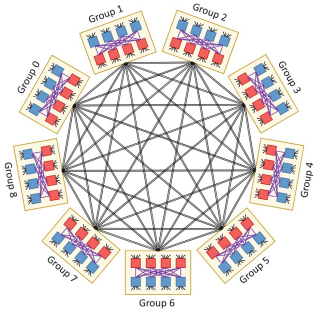


Fig. 4. Example Megafly topology.

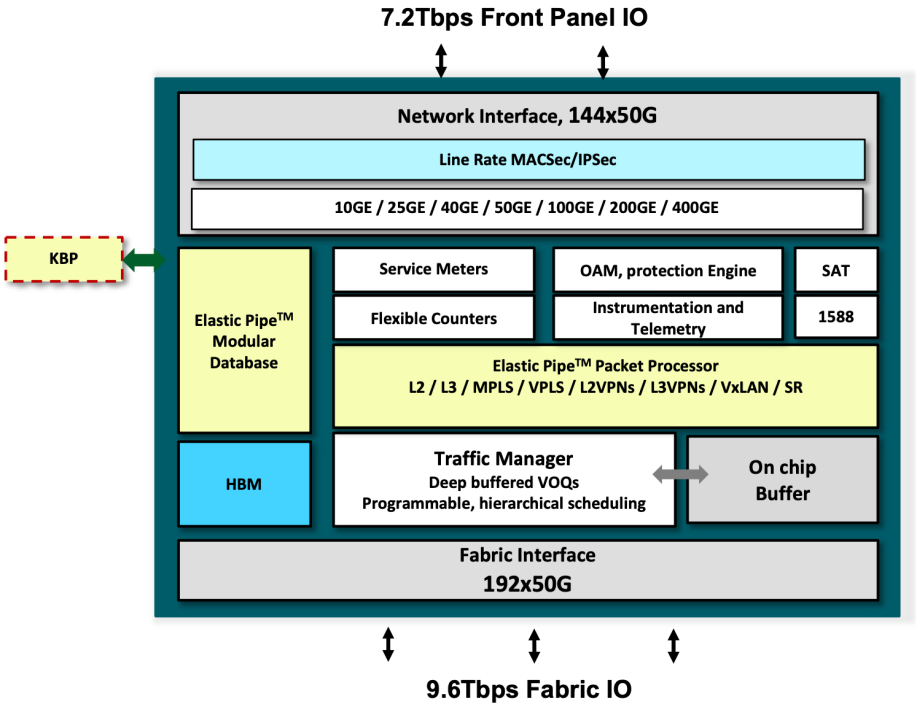
Teh, M.Y., Wilke, J.J., Bergman, K., Rumley, S. (2017). Design Space Exploration of the Dragonfly Topology. In: Kunkel, J., Yokota, R., Taufer, M., Shalf, J. (eds) High Performance Computing. ISC High Performance 2017. Lecture Notes in Computer Science(), vol 10524. Springer, Cham. https://doi.org/10.1007/978-3-319-67630-2_5

Flajslik, M., Borch, E., Parker, M.A. (2018). Megafly: A Topology for Exascale Systems. In: Yokota, R., Weiland, M., Keyes, D., Trinitis, C. (eds) High Performance Computing. ISC High Performance 2018. Lecture Notes in Computer Science(), vol 10876. Springer, Cham. https://doi.org/10.1007/978-3-319-92040-5_15

Jericho2

- Jericho2c+ : 14.4Tb/s
- Hasta 18 puertos 100GE
- High Bandwidth Memory (HBM): 8GB, en el mismo encapsulado
- Pipeline reprogramable (C++)
- Jericho2 + Ramon: permite construir single-stage system 900Tb/s

Figure 1: BCM88850 Block Diagram



https://docs.broadcom.com/doc/88850-PB1-PUB

Jericho2 : Ejemplos

- Arista 7280CR3-32P4

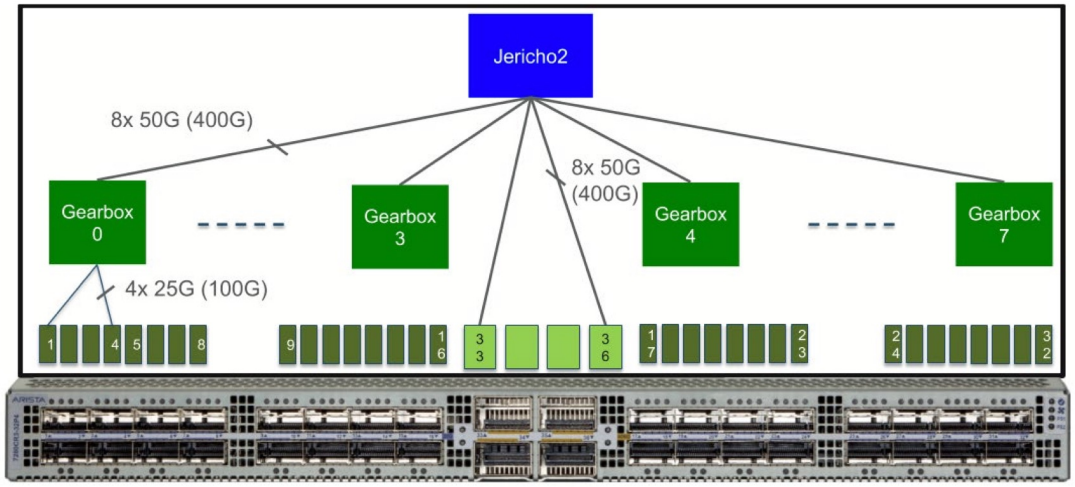


Figure 6: Arista 7280CR3-32P4 Switch Architecture

- 7280PR23-24

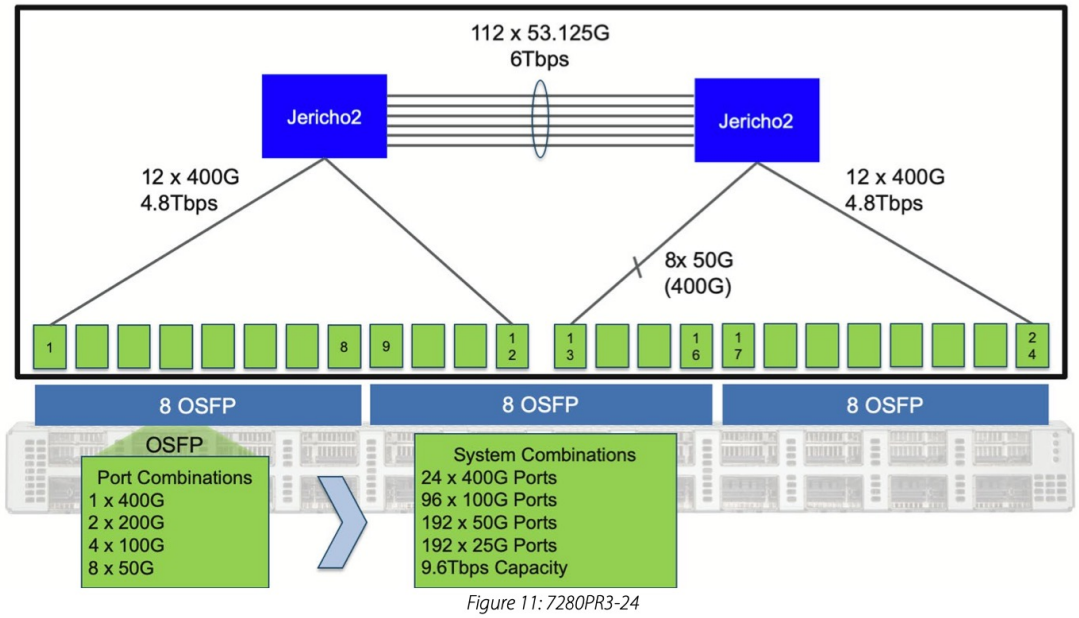
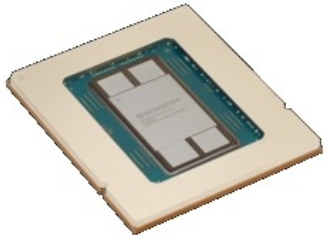


Figure 11: 7280PR3-24

https://www.arista.com/assets/data/pdf/Whitepapers/7280R3-Platform-Architecture-WP.pdf

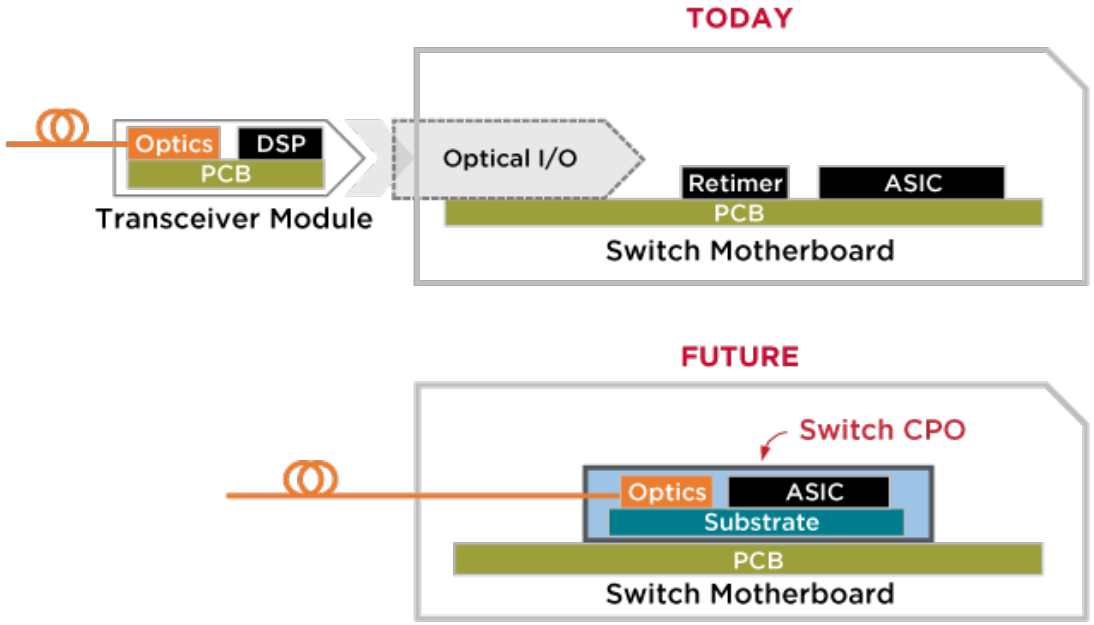
Jericho3AI

- 28.8 Tb/s
- 144 SerDes @ 106-Gb/s PAM4
- Up to 18 × 800GbE, 36 × 400GbE, or 72 × 200GbE
- Support for 25GE, 50GE, 100GE, 200GE, 400GE, 800GE Ethernet port interfaces
- *“Hierarchical traffic manager, scalable packet buffer memory and low latency forwarding”*
- *“The BCM88890, combined with the BCM88920 (Ramon3), is designed to meet the unique requirements for next-generation Artificial Intelligence / Machine Learning (AI/ML) routed networks.”*
- *“... using the BCM88920 two-stage fabric to create a scalable core platform that delivers up to 25,000 ports of 800GbE”*



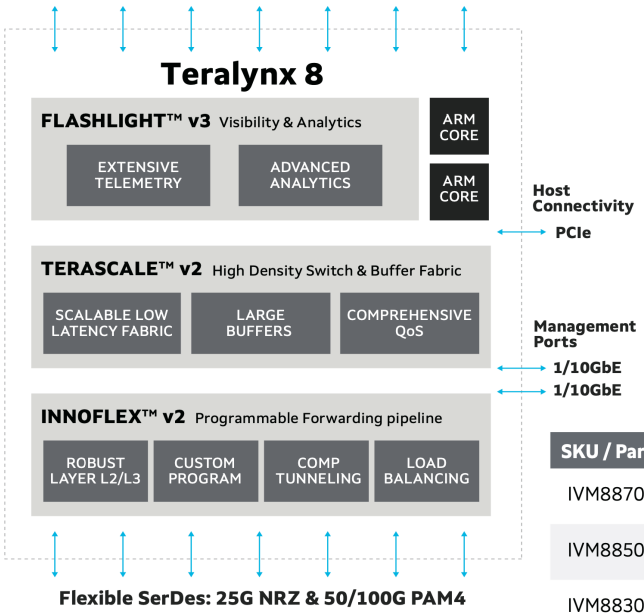
CPO

- Co-Packaged Optics



Marvell Teralynx8

Features	Benefits
· 25.6 Tbps throughput with 256x 112 Gbps SerDes	· Industry leading performance and scale enables customers to deploy fewer network switches and tiers dramatically reducing cost, power, latency & management
· Comprehensive IP forwarding and highly scalable/flexible layer 2 and 3 tables for IPv4, IPv6 and hybrid networks	· Proven, innovative & highly scalable architecture delivers 64 x 400Gbe, 128 x 200G and 256 x 100GbE ports
· Line-rate programmability to accommodate future networking protocols with software upgrades	· 100G LR SerDes enables higher scale IO with backward compatibility to 50G PAM4 & 10/25G NRZ
· Extensive tunneling capabilities such as IP-in-IP, GRE, MPLS, VXLAN and Geneve	· Breakthrough visibility and analytics capabilities enable predictive, faster & more accurate issue resolution, higher automation and self-healing autono-mous networks
· Very low latencies - cut-through and store-and-forward	· Superior power efficiency enables customers to design 1RU 32 x 800G switches for best power and cost per bit
· Advanced QoS/traffic management feature set such as DCB, RDMA/RoCE	· InnoFlex™ programmable forwarding pipeline enables support of custom & new standard protocols without requiring ASIC spins to future proof the network
· FLASHLIGHT™ v3 innovations delivers breakthrough visibility and telemetry addressing Cloud customer requirements	



SKU / Part Number	Capacity (Tbps)	# of SerDes @ Gbps	10/25/50G ports	100G ports	200G ports	400G ports
IVM88700	25.6	256@ 10/25/50/100	256	256	128	64
IVM88500	12.8	256 @ 10/25/50	128	128	64	32
IVM88300	8.0	160 @ 10/25/50	80	80	40	20

https://www.marvell.com/content/dam/marvell/en/public-collateral/switching/marvell-teralynx-8-product-brief.pdf

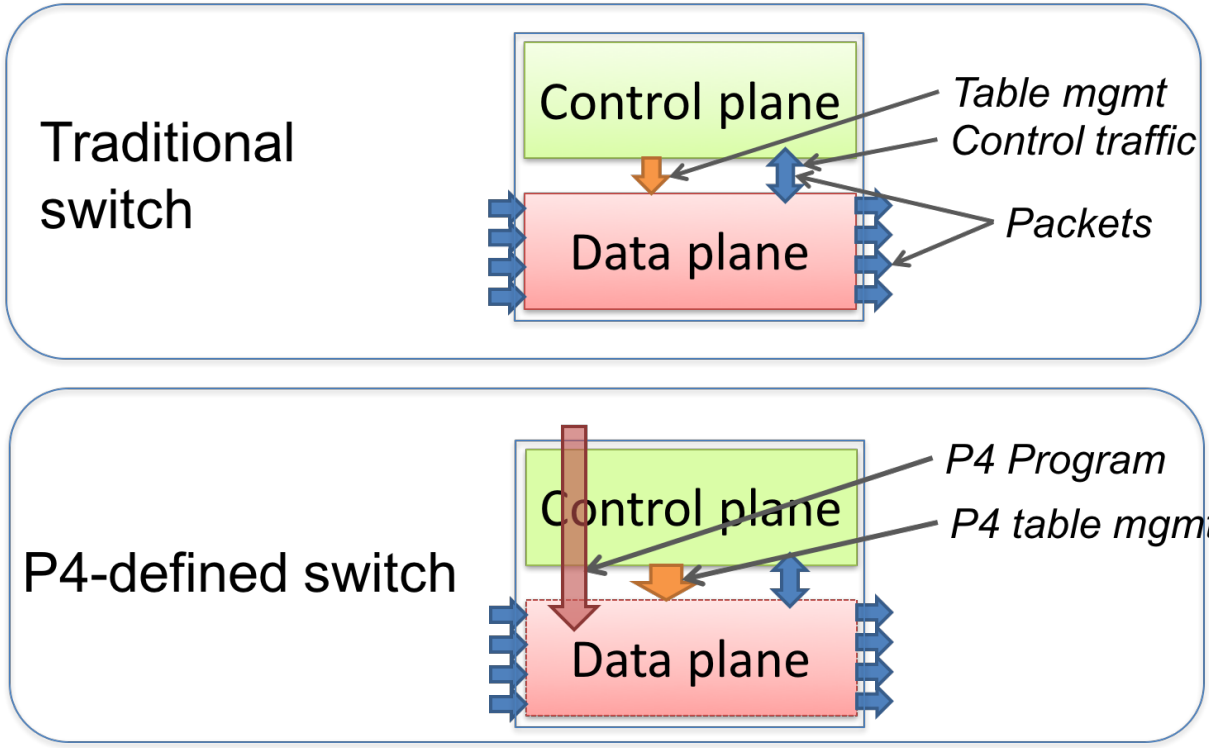
Intel Tofino

- Tofino 2
  - Hasta 12.8 Tb/s
  - Soporta puertos de 400Gb/s
- Tofino 3
  - Hasta 25.6 Tb/s (de 256x10GbE a 64x500GbE)
  - 10/25/40/50/100/200/400 GbE
  - 10.000 Mpkt/s. Cut-through
  - 192MB shared packet buffer, 160Mb SRAM, 8.2Mb TCAM
  - Hasta 256 puertos, hasta 128 egress queues por puerto
  - 112G PAM4 y 56G PAM4 SerDes
  - Habría estado en 2023 pero Intel ha parado el desarrollo
  - $200M el desarrollo de uno de estos
  - Ene 2023: Desarrollo detenido
- P4-programable (...)

https://www.intel.com/content/www/us/en/products/network-io/programmable-ethernet-switch/tofino-2-series.html
https://www.intel.es/content/www/es/es/products/network-io/programmable-ethernet-switch/tofino-3-brief.html
https://www.intel.com/content/dam/www/central-libraries/us/en/documents/2022-05/tofino-3-intelligent-fabric-processors-brief.pdf

Plano de datos

- En el propio diseño del hardware
- Nuevos ASICs son reprogramables
- Los objetos en el plano de datos no están fijos, se pueden modificar y con ellos lo que puede modificar el plano de control

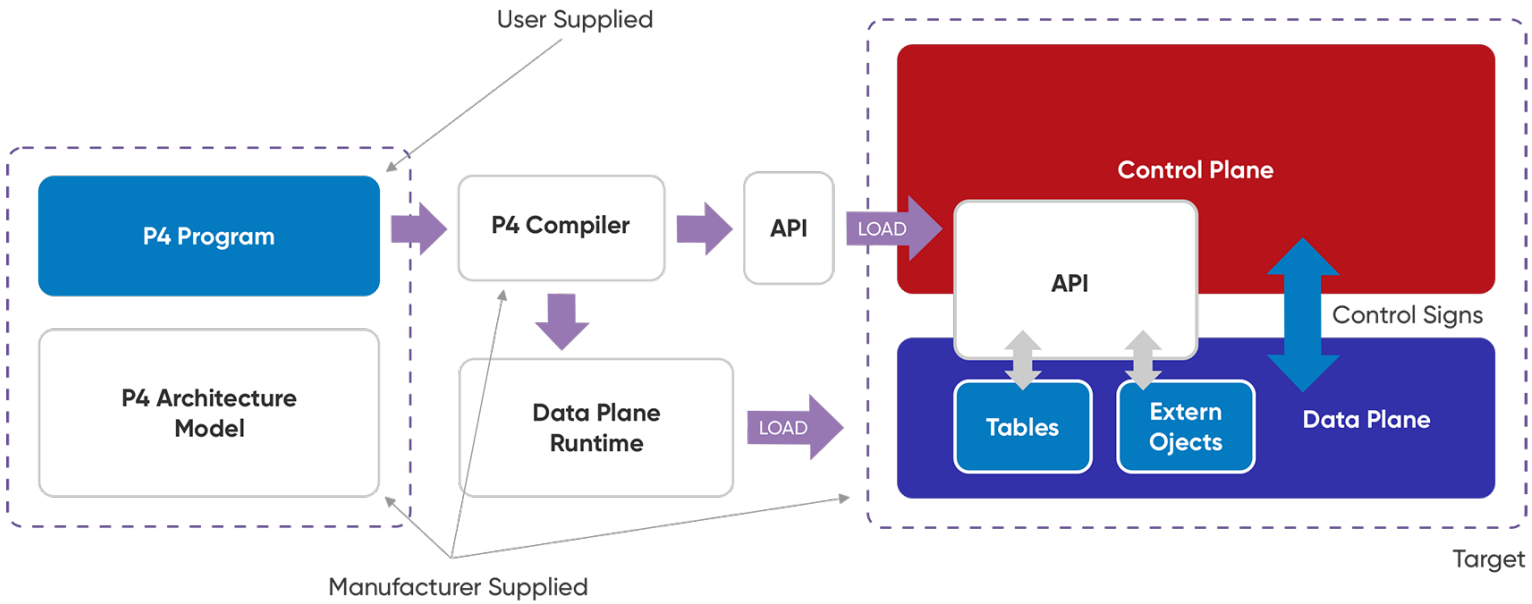


https://p4.org/p4-spec/docs/P4-16-v1.2.1.html

Ejemplo: P4

- Programming Protocol-independent Packet Processors

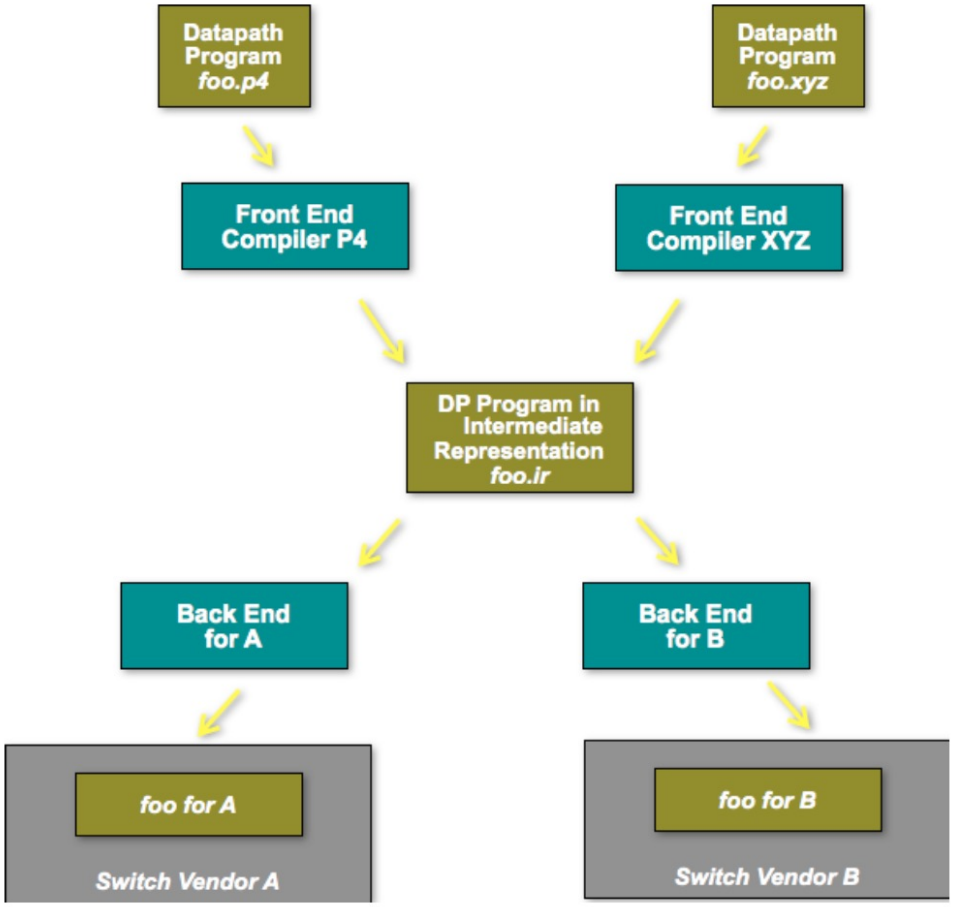
https://p4.org



https://p4.org/p4-spec/docs/P4-16-v1.2.1.html

P4

https://p4.org



https://opennetworking.org/wp-content/uploads/2013/05/TR-535_ONF_SDN_Evolution.pdf

```
parser TopParser(packet_in b, out Parsed_packet p) {
    Checksum16() ck;  // instantiate checksum unit

    state start {
        b.extract(p.ethernet);
        transition select(p.ethernet.etherType) {
            0x0800: parse_ipv4;
            // no default rule: all other packets rejected
        }
    }

    state parse_ipv4 {
        b.extract(p.ip);
        verify(p.ip.version == 4w4, error.IPv4IncorrectVersion);
        verify(p.ip.ihl == 4w5, error.IPv4OptionsNotSupported);
        ck.clear();
        ck.update(p.ip);
        // Verify that packet checksum is zero
        verify(ck.get() == 16w0, error.IPv4ChecksumError);
        transition accept;
    }
}
...
```

https://p4.org/p4-spec/docs/P4-16-v1.2.1.html

Netberg Aurora 710

- ONIE Pre-loaded
- Barefoot Tofino
- A unique development platform
- P4 programming language
- ONL-ready
- SDN-ready
- X86 Linux apps

https://netbergtw.com/products/aurora-710/

- Interfaces: 32x 100G QSFP28 Ports, 2x 10G SFP+ ports, Management (1000Base-T), mini-USB Console Port, and USB2.0 (Type A)
- Intel Xeon D-1527 CPU
- 8GB DRAM, 128GB m.2 SSD
- Barefoot Tofino DFN-T10-032D
- Switching Capacity: 3.2Tbps, 22MB Packet Buffer
- Jumbo Packet: 12K bytes

Intel Barefoot Tofino

Description	A 64x100G ports P4 programmable Ethernet switch with maximum port bandwidth of 6.4 Tbps

https://www.intel.com/content/www/us/en/products/network-io/programmable-ethernet-switch/tofino-series/tofino.html

Netberg Aurora 710

- ECMP
- LAG, MC-LAG (L2)
- LLDP. LLDP extended MIB: lldpremtable, lldplocporttable, lldpremmanaddrtable, lldplocmanaddrtable, lldplocporttable, lldpLocalSystemData
- QoS - ECN, QoS - RDMA
- Priority Flow Control
- WRED
- COS
- SNMP
- Syslog, Sysdump
- NTP
- COPP
- DHCP Relay Agent
- SONiC to SONiC upgrade
- One Image
- VLAN, VLAN trunk
- ACL permit/deny
- IPv6
- Tunnel Decap
- Mirroring
- Post Speed Setting
- BGP Graceful restart helper
- BGP MP
- Fast Reload
- PFC WD
- RADIUS AAA, TACACS+
- LACP Fallback
- MTU Setting
- IPv6 ACL
- BGP
  BGP/Neighbor-down fib-accelerate
  BGP-EVPN support(type 5) (related HLD Fpmsyncd, Vxlanmgr,template)
- Port breakout
- Dynamic ACL Upgrade
- SWSS Unit Test Framework (best effort)
- ConfigDB Framework
- Critical Resource Monitoring
- MAC Aging
- IPv6 ACL
- BGP/Neighbor-down fib-accelerate
- PFC WD
- gRPC
- Dtel support
- Sensor transceiver monitoring
- Debian Kernel 4.9
- Warm Reboot
- Incremental Config (IP, LAG, Port shut/unshut)
- Asymmetric PFC
- PFC Watermark
- Routing Stack Graceful Restart
- Basic VRF and L3 VXLAN
- FRR as default routing stack
- Everflow enhancement
- Egress ACL bug fix and ACL CLI enhancement
- L3 RIF counter support
- PMon Refactoring
- MACSec

https://netbergtw.com/products/aurora-710/

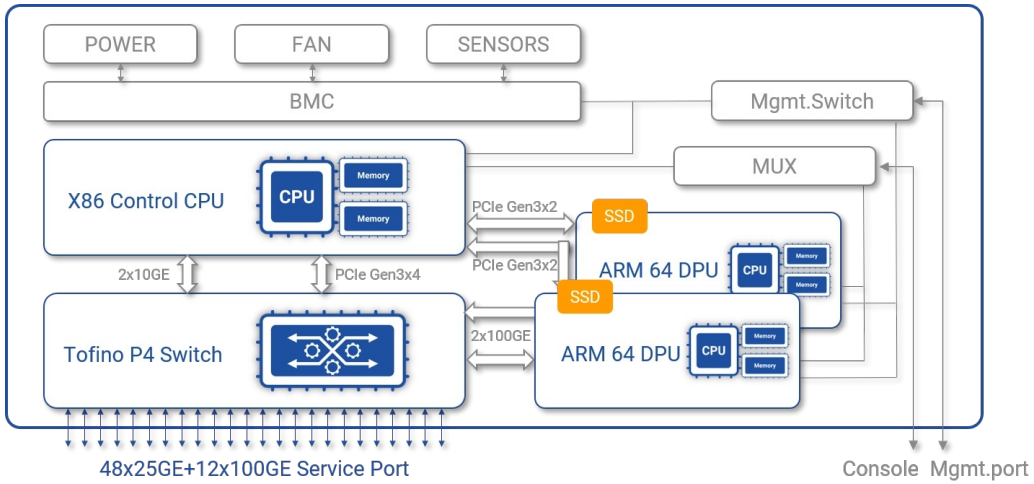
Intel Barefoot Tofino

Description	A 64x100G ports P4 programmable Ethernet switch with maximum port bandwidth of 6.4 Tbps

https://www.intel.com/content/www/us/en/products/network-io/programmable-ethernet-switch/tofino-series/tofino.html

Asterfusion X312P-T

- P4-programmable, 48 puertos 25GE + 12x100GBE
- 3.3 Tbps Intel Tofino ASIC + CPU Intel Broadwell
- + 2x Marvell Octeon TX CN9670 DPUs (24-core ARM64 1.8GHz)



DPU = Data Processing Unit (SoC con ARM multicore + NIC + engines para ML, seguridad, almacenamiento, etc)

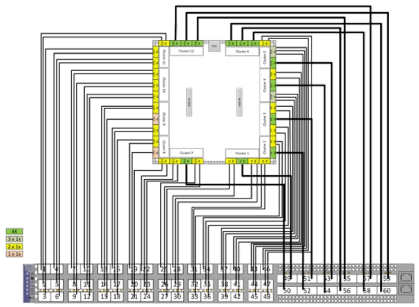
https://cloudswit.ch/product/48x25g12x100g-p4-switch/

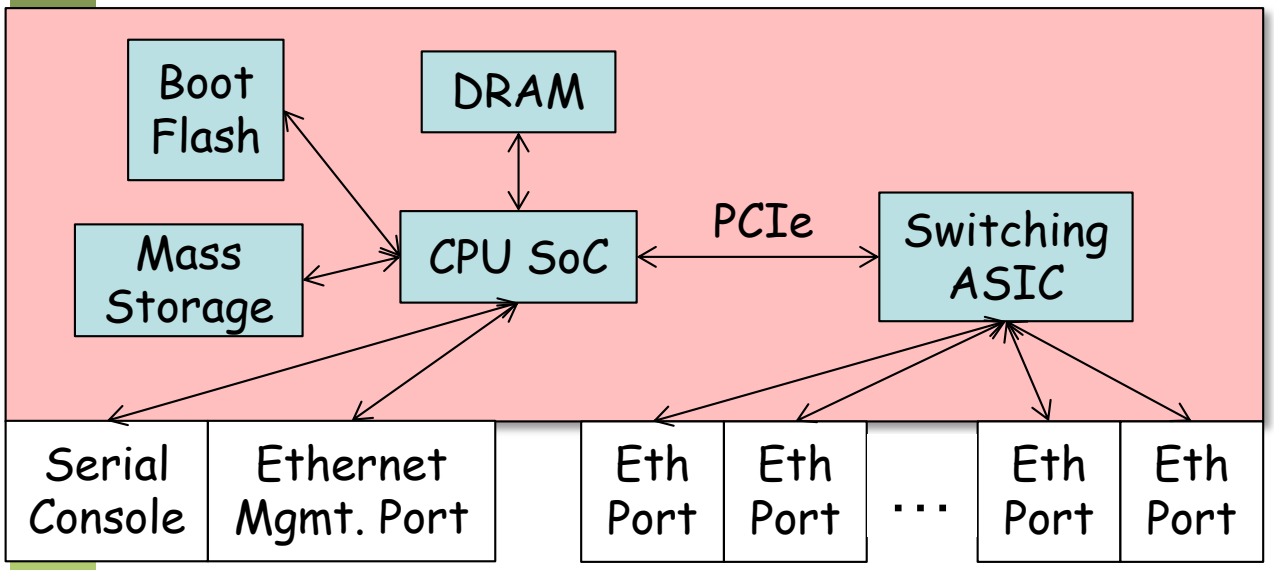
¿Evolución?

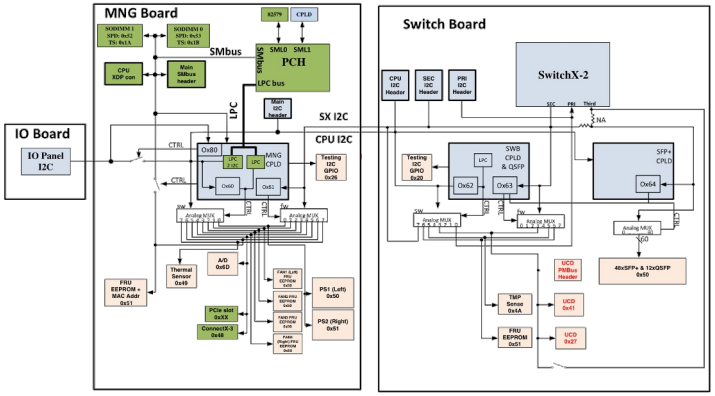
- La infraestructura se está simplificando
- Principalmente el hardware, controlable por software
- *White boxes* no solo servidores sino también switches
- También se venden ya switches *"Bare metal"* = solo el hardware

Bare metal switches

- Menores costes
- El mismo equipo un día es un switch, otro un firewall, otro un balanceador… dentro de las limitaciones del ASIC
- Ejemplo:
  - Open Compute Networking Project
  - http://www.opencompute.org/wiki/Networking
  - Especificaciones completas de conmutadores
- Fabricantes: Mellanox, Quanta, Penguin Computing, Edge-core, Acton, Dell, etc







¿Evolución?

- Para estos equipos sistemas operativos y gran cantidad de software, generalmente basados en linux, muchos de código abierto
- Ejemplos:
  - SONiC: https://azure.github.io/SONiC/
  - Open Network Install Environment (ONIE): http://onie.opencompute.org
  - Open Network Linux: http://opennetlinux.org
  - Big Switch's Switch Light OS
  - Pica8 PicOS
  - Cumulus Linux
- Es decir, igual que en el entorno de servidor, puedes cambiar el hardware, instalar el sistema operativo que quieras y desarrollar tus aplicaciones (…)

¿Evolución?

- Para diferenciarse, los proveedores desarrollan software propietario para ofrecer sus servicios
- Porque hoy en día ya es el software por lo que principalmente están cobrando los fabricantes “no-open”
- Muchos modelos ToR de fabricantes conocidos son switches bare-metal que han comprado, cambiado el software y el frontal

Software Defined X

- Software Defined Networking (SDN)
- Software Defined Infrastructure (SDI)
- Software Defined Data Center (SDDC)
- Software Defined Storage (SDS)
- Software Defined Radio (SDR)
- Software Defined WAN (SD-WAN)
- Software Defined Power (SDP)
- Software Defined Internet of Things (SDIoT)
- Software Defined Wireless LAN (SD-WLAN)
- Software Defined Contend Distribution Network (SD-CDN)
- etc

¿Software?

- ¿Amazon? (Libros antes, ahora cualquier cosa)
- ¿Netflix? (¿La industria del cine va a ser solo productora y no distribuidora?)
- ¿¡Tunes? ¿Spotify? ¿Pandora? (¿La industria de la música va a ser solo productora y no distribuidora?)
- ¿Zynga? ¿Rovio? (Fuerte competencia a los Electronic Arts y Nintendos de la década pasada)
- ¿Pixar? (¿Disney tuvo que adquirir una compañía de soft para seguir siendo relevante? ¿Quién compró a quién?)
- ¿Flickr? ¿Instagram? (¿A alguien le suena Kodak?)
- ¿Google? (¿Vemos publicidad en la TV?)
- ¿PayPal?
- ¿Groupon, Foursquare, Skype, LinkedIn?
- ¿Alguien es capaz de reparar hoy en día su coche sin las herramientas soft?
- ¿Wal-Mart? ¿FedEx? Su logística no es nada sin el soft
- ¿Salud?
- ¿Defensa?

„Software is eating the world.“

- Marc Andreessen
Cofundador de Netscape Comm. Corp.

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

Networking hardware y el software