OpenFlow

OpenFlow

- Su origen en proyectos de investigación en la Universidad de Stanford
- En 2011 se funda el consorcio ONF
  - *Open Networking Foundation*
  - https://www.opennetworking.org
  - Más de 140 empresas (fabricantes, operadoras, ISPs, startups, etc)
- OpenFlow es un protocolo “southbound”
- No hace “nada” sin una aplicación que lo emplee

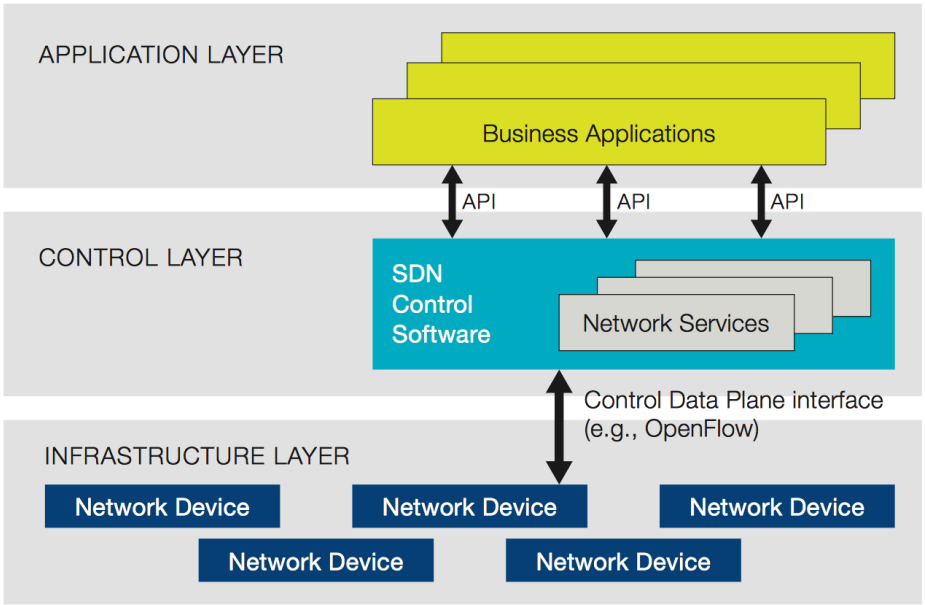
OPEN NETWORKING FOUNDATION

Martin Casado

ONF y SDN

- *"The aim of SDN is to provide open interfaces that enable the development of software that can control the connectivity provided by a set of network resources and the flow of network traffic though them, along with possible inspection and modification of traffic that may be performed in the network."*
- *"In the SDN architecture, the control and data planes are decoupled, network intelligence and state are logically centralized, and the underlying network infrastructure is abstracted from the applications. "*

OpenFlow



https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf

OpenFlow

- Dos tipos de conmutadores:
  - *OpenFlow-only*: solo soportan el modo de funcionamiento OpenFlow
  - *OpenFlow-hybrid*: también soportan funcionamiento "normal" (conmutación L2, conmutación L3, VLANs, ACLs, etc)
  - Los híbridos deberán tener alguna forma de clasificar si los paquetes pasan por procesado "normal" u OpenFlow

Ejemplo: HP 2920-24G

Key Features

- High-performance Gigabit Ethernet access switch
- Four optional 10GbE (SFP+ and/or 10GBASE-T) ports
- Stacking capability with a total of four switches
- L2 and L3 plus static and RIP routing, PoE, and PoE+ support
- Limited Lifetime Warranty 2.0, sFlow, ACLs, OpenFlow, and rate limiting

Product overview

The HP 2920 Switch Series consists of five switches: the 2920-24G and 2920-24G-PoE+ switches with 24 10/100/1000 ports and the 2920-48G, 2920-48G-PoE+, and 2920-48G 740W PoE+ switches with 48 10/100/1000 ports. Each switch has four dual-personality ports for 10/100/1000 or SFP connectivity.

In addition, the 2920 Switch Series supports up to four optional 10 Gigabit Ethernet (SFP+ and/ or 10GBASE-T) ports, as well as a two-port stacking module. These options provide you with flexible and easy-to-deploy uplinks and stacking.

Together with static and routing-information-protocol (RIP) routing, robust security and management, enterprise-class features, Limited Lifetime Warranty 2.0, and software updates included, the 2920 Switch Series is a comprehensive, cost-effective, and scalable solution for building high-performance networks. These switches can be deployed at the enterprise edge, in remote branch offices, and in converged networks.

Features and Benefits

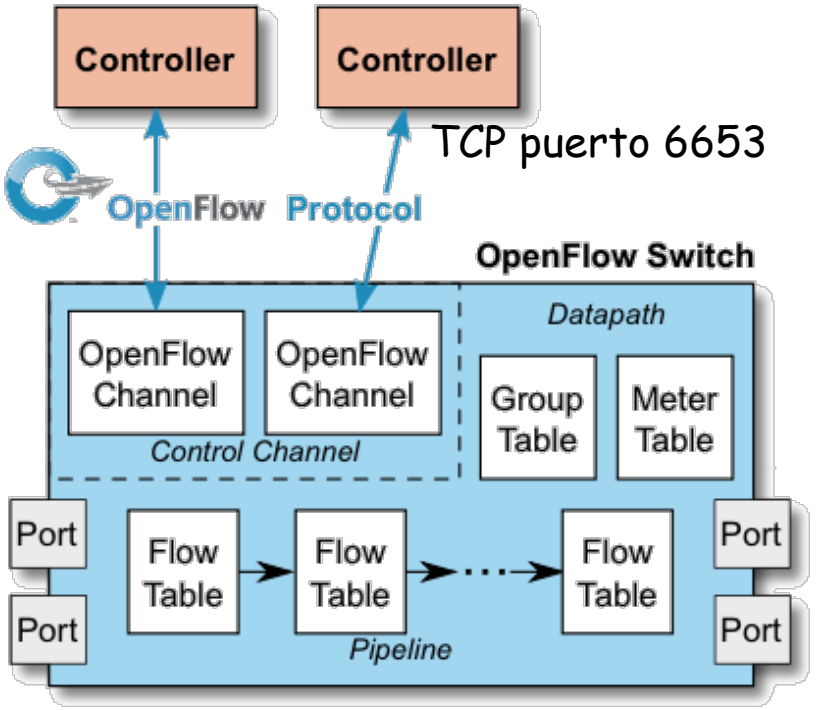
Software-defined networking

- **OpenFlow**
  supports OpenFlow 1.0 and 1.3 specifications to enable SDN by allowing separation of the data (packet forwarding) and control (routing decision) paths

http://h20195.www2.hp.com/v2/GetDocument.aspx?docname=c04111401&doctype=quickspecs&doclang=EN_US&searchquery=&cc=es&lc=es

Flow Tables

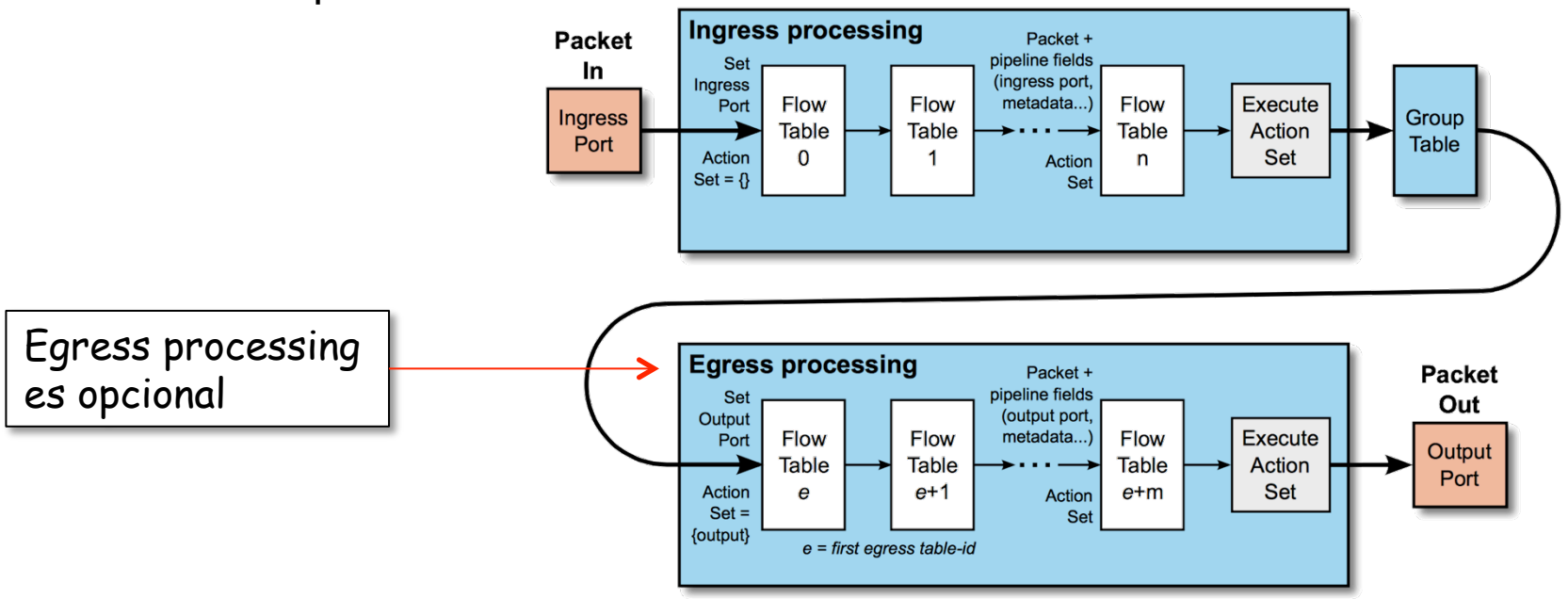
- Contienen la información sobre los campos a comprobar (*match fields*) en los paquetes y qué hacer con ellos
- El controlador puede añadir, modificar y borrar entradas empleando OF
- Las "acciones" son las operaciones en caso de que el paquete verifique la entrada en la tabla
- Puede reenviar el paquete, mandárselo al controlador, pasarlo a otra tabla, actualizar contadores, etc



https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-switch-v1.5.0.noipr.pdf

OpenFlow pipeline

- Debe tener al menos una tabla aunque pueden ser más (desde 1.1, permite procesado de etiquetas MPLS)
- Hay procesado a la entrada del paquete (al menos una tabla)
- Si se decide reenviarlo pasa por tablas de salida (desde 1.5)
- Las tablas se comprueban en orden
- Si el paquete verifica una regla se ejecuta la acción que indique
- Si no verifica ninguna es un *"table miss"* y hay una acción por defecto en la tabla para este caso



https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-switch-v1.5.0.noipr.pdf

Acciones

- Incluimos aquí la acción por defecto para el caso de *"table miss"*
- La acción puede ser pasar a otra tabla posterior (no anterior)
- Puede ser hacer inundación
- O reenviar por un puerto en concreto
- O puede ser reenviar el paquete al controlador (dentro de un mensaje OF)
- O pasar el paquete a un reenvío tradicional si es un conmutador híbrido
- O modificar campos de cabeceras del paquete (una modificación afecta a las comprobaciones en egress tables)
- etc

Entradas en las tablas

- *Match Fields:*
  - Puede valer ANY (comodín) o soportarse bitmasks
  - Hasta la versión 1.1 se miraban ciertos campos:
    - Puerto de entrada, metadatos provenientes de tabla anterior
    - Direcciones MAC origen y destino, Ethertype, VLAN ID, PCP
    - Etiqueta MPLS, TC
    - Direcciones IP origen y destino, protocolo, ToS
    - Puertos origen y destino TCP/UDP/SCTP
    - Tipo y código ICMP
  - Otros que se han ido añadiendo:
    - Bits ECN
    - Flags TCP
    - Código de opción de ARP, direcciones MAC e IP en el mensaje ARP
    - Direcciones IPv6, flow label IPv6, tipo y código ICMPv6
    - Etc
  - (…)

Match Fields	Priority	Counters	Instructions	Timeouts	Cookie	Flags

Entradas en las tablas

- Prioridad:
  - Pueden verificarse varias entradas de la tabla
  - En ese caso se selecciona solo la de mayor prioridad
- (…)

Match Fields	Priority	Counters	Instructions	Timeouts	Cookie	Flags

Entradas en las tablas

- Contadores:
  - Se actualizan cuando la entrada es seleccionada
- (…)

Counter	**Bits**	
Per Flow Table		
Reference Count (active entries)	32	*Required*
Packet Lookups	64	*Optional*
Packet Matches	64	*Optional*
Per Flow Entry		
Received Packets	64	*Optional*
Received Bytes	64	*Optional*
Duration (seconds)	32	*Required*
Duration (nanoseconds)	32	*Optional*
Per Port		
Received Packets	64	*Required*
Transmitted Packets	64	*Required*
Received Bytes	64	*Optional*
Transmitted Bytes	64	*Optional*
Receive Drops	64	*Optional*
Transmit Drops	64	*Optional*
Receive Errors	64	*Optional*
Transmit Errors	64	*Optional*
Receive Frame Alignment Errors	64	*Optional*
Receive Overrun Errors	64	*Optional*
Receive CRC Errors	64	*Optional*
Collisions	64	*Optional*
Duration (seconds)	32	*Required*
Duration (nanoseconds)	32	*Optional*
Per Queue		
Transmit Packets	64	*Required*
Transmit Bytes	64	*Optional*
Transmit Overrun Errors	64	*Optional*
Duration (seconds)	32	*Required*
Duration (nanoseconds)	32	*Optional*
Per Group		
Reference Count (flow entries)	32	*Optional*
Packet Count	64	*Optional*
Byte Count	64	*Optional*
Duration (seconds)	32	*Required*
Duration (nanoseconds)	32	*Optional*
Per Group Bucket		
Packet Count	64	*Optional*
Byte Count	64	*Optional*
Per Meter		
Flow Count	32	*Optional*
Input Packet Count	64	*Optional*
Input Byte Count	64	*Optional*
Duration (seconds)	32	*Required*
Duration (nanoseconds)	32	*Optional*
Per Meter Band		
In Band Packet Count	64	*Optional*
In Band Byte Count	64	*Optional*

Entradas en las tablas

- *Instructions:*
  - Cambio al paquete, acciones, etc, cuando se selecciona la entrada
  - Las hay de implementacion requerida y opcional
  - Ejemplos:
    - Enviar a un puerto de salida, descartar, asignar cola en el puerto out
    - Añadir/retirar etiquetas (MPLS, VLAN, PBB)
    - Modificar valor de un campo de cabecera
- (…)

Match Fields	Priority	Counters	Instructions	Timeouts	Cookie	Flags

Entradas en las tablas

- *Timeouts:*
  - Máximo tiempo inactiva antes de expirar
- (…)

Match Fields	Priority	Counters	Instructions	Timeouts	Cookie	Flags

Entradas en las tablas

- *Cookie*:
  - Ahí el controlador puede guardar un valor
  - El switch no lo emplea para nada
- (…)

Match Fields	Priority	Counters	Instructions	Timeouts	Cookie	Flags

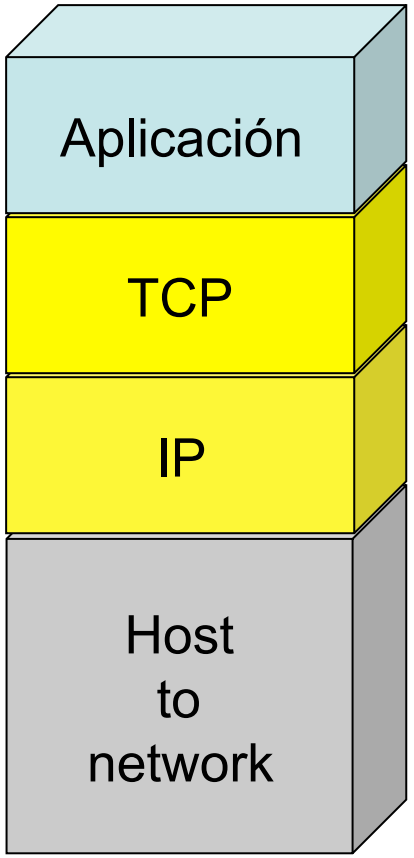
Entradas en las tablas

- *Flags*:
  - Diferentes opciones
  - Ejemplo:
    - Que envíe un mensaje al controlador al eliminarse o expirar una entrada
    - Que no lleve contadores de bytes o de paquetes

Match Fields	Priority	Counters	Instructions	Timeouts	Cookie	Flags

El protocolo

- TCP (puerto 6653), opcionalmente empleando TLS
- Hay mensajes:
  - De controlador a conmutador (…)
  - Asíncronos (desde el conmutador)
  - Simétricos



El protocolo

- TCP (puerto 6653), opcionalmente empleando TLS
- Hay mensajes:
  - De controlador a conmutador
    - Petición de capacidades
    - Establecer o preguntar por configuración o estado
    - Entregarle un paquete para enviar por un puerto
  - Asíncronos (desde el conmutador) (…)
  - Simétricos

El protocolo

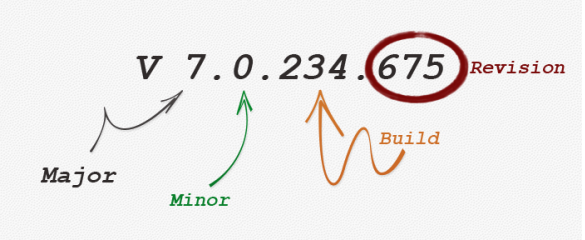
- TCP (puerto 6653), opcionalmente empleando TLS
- Hay mensajes:
  - De controlador a conmutador
  - Asíncronos (desde el conmutador)
    - Envío al controlador de un paquete recibido
    - Notificación de entrada en tabla eliminada
    - Notificación de cambio de estado de un puerto
  - Simétricos (…)

El protocolo

- TCP (puerto 6653), opcionalmente empleando TLS
- Hay mensajes:
  - De controlador a conmutador
  - Asíncronos (desde el conmutador)
  - Simétricos
    - Hello, al establecer la conexión
    - Echo, para comprobar que el otro extremo está vivo y tal vez para medir latencia o bw
    - Error

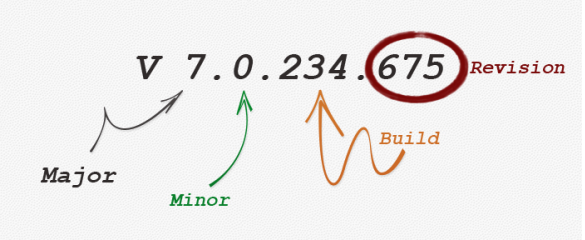
Versiones

- https://www.opennetworking.org/sdn-resources/technical-library
- Versión 1.5.1 Abril de 2015
- Probablemente OF 1.0 sea lo más implementado en hardware
- Las siguientes versiones han ido introduciendo mejoras, más flexibilidad, pero también haciéndolo más complejo
- OF 1.1
  - Múltiples tablas
  - Soporte de acciones para MPLS (soporta multi-etiqueta)
  - Acciones sobre el TTL
  - Soporte de VLANs en QinQ
  - Soporte para agrupar puertos de cara a acciones
- (…)



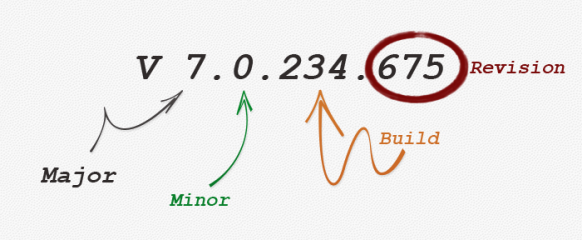
Versiones

- https://www.opennetworking.org/sdn-resources/technical-library
- Versión 1.5.1 Abril de 2015
- Probablemente OF 1.0 sea lo más implementado en hardware
- Las siguientes versiones han ido introduciendo mejoras, más flexibilidad, pero también haciéndolo más complejo
- OF 1.1
- OF 1.2
  - Soporte de campos de IPv6, ICMPv6, ND
  - Mejora la extensibilidad de las reglas de *match*
- (…)



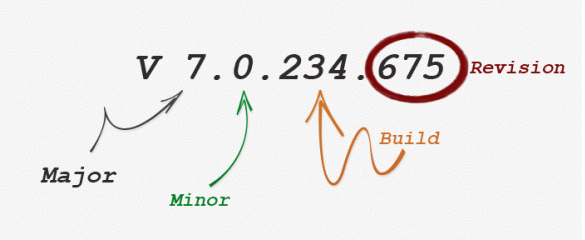
Versiones

- https://www.opennetworking.org/sdn-resources/technical-library
- Versión 1.5.1 Abril de 2015
- Probablemente OF 1.0 sea lo más implementado en hardware
- Las siguientes versiones han ido introduciendo mejoras, más flexibilidad, pero también haciéndolo más complejo
- OF 1.1
- OF 1.2
- OF 1.3.x
  - *Meters* por flujo (limitadores para QoS)
  - Soporte de PBB
- (…)



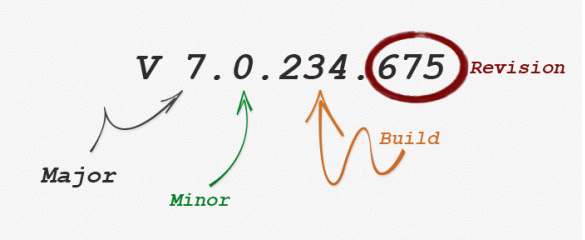
Versiones

- https://www.opennetworking.org/sdn-resources/technical-library
- Versión 1.5.1 Abril de 2015
- Probablemente OF 1.0 sea lo más implementado en hardware
- Las siguientes versiones han ido introduciendo mejoras, más flexibilidad, pero también haciéndolo más complejo
- OF 1.1
- OF 1.2
- OF 1.3.x
- OF 1.4
  - Mayor extensibilidad
  - Soporte de puertos ópticos (frecuencias, potencia, etc)
- (…)



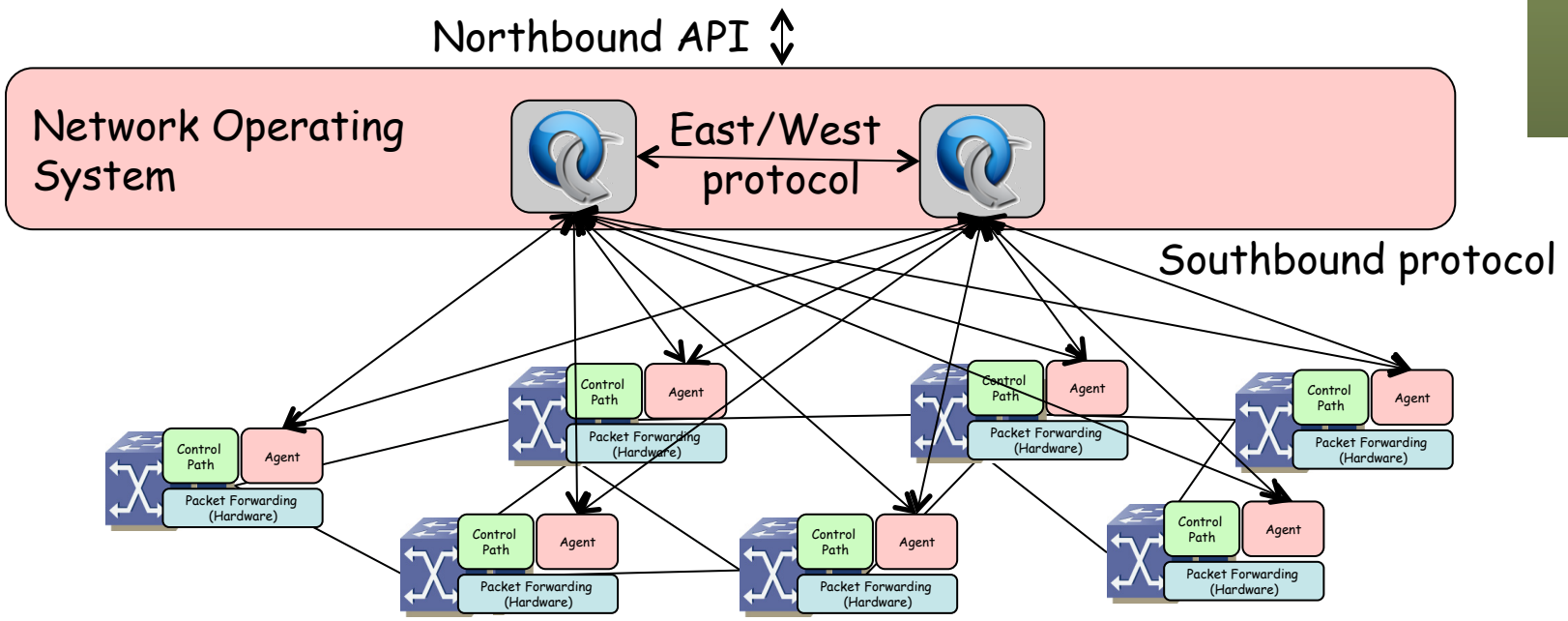
Versiones

- https://www.opennetworking.org/sdn-resources/technical-library
- Versión 1.5.1 Abril de 2015
- Probablemente OF 1.0 sea lo más implementado en hardware
- Las siguientes versiones han ido introduciendo mejoras, más flexibilidad, pero también haciéndolo más complejo
- OF 1.1
- OF 1.2
- OF 1.3.x
- OF 1.4
- OF 1.5
  - *Egress tables*
  - Soporte para más que Ethernet
  - Flags TCP



APIs

- OpenFlow es un *Southbound API*
- El ONF asocia OpenFlow a SDN pero una SDN no necesita emplear necesariamente OpenFlow
- Podríamos considerar OF a día de hoy el API south estándar
- No hay *Northbound API* estandarizada, ni *de facto*
- No hay *East/West API* estandarizada

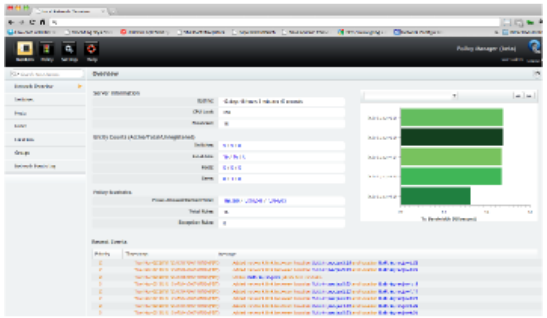


Controladores

- NOX
  - http://www.noxrepo.org
  - Desarrollado por Nicira, cedido el código en 2008
  - Ofrece un API C++ para OF 1.0
  - Muchos otros heredan de su código
  - Incluye componentes de ejemplo para descubrir la topología, implementar un puente transparente y un switch distribuido
  - Open Source
- POX
  - Hereda de NOX
  - Permite el desarrollo en Python
  - Open Source
- Beacon
  - https://openflow.stanford.edu/display/Beacon/Home
  - Java (desarrollo con eclipse)
  - Open Source

Controladores

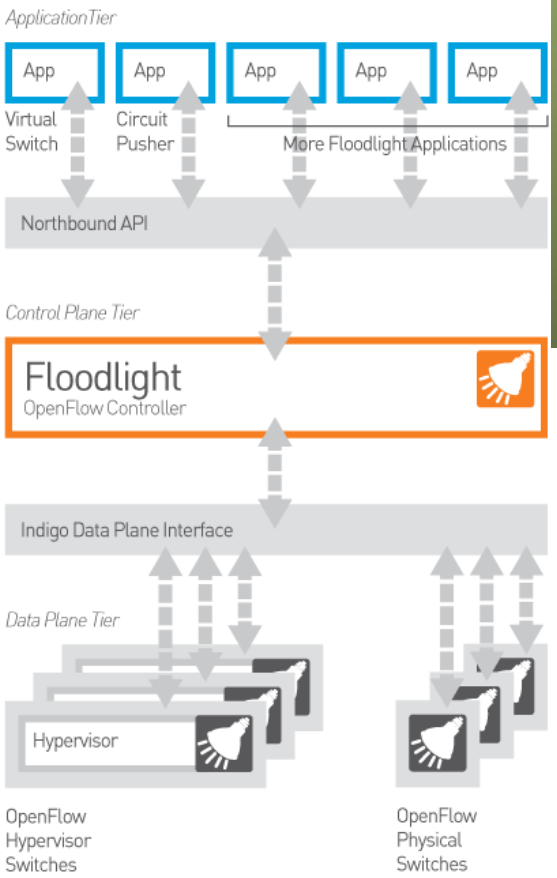
- SNAC
  - http://www.openflowhub.org/display/Snac/SNAC+Home
  - Incluye GUI web
  - Incluye un lenguaje de definición de políticas
  - Open Source



- FloodLight
  - http://www.projectfloodlight.org/floodlight/
  - Basado en Java (basado en Beacon)
  - Apoyado por Big Switch Networks
  - Lo emplean para construir su controlador
  - Open Source

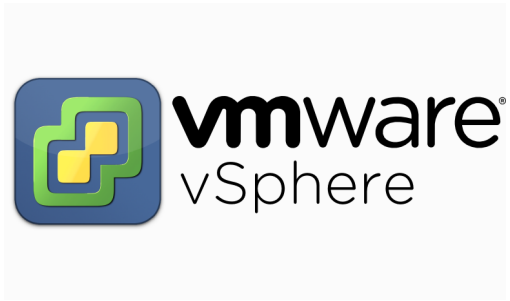
Guido Appenzeller

big switch networks



VMware

- Controlador propietario
- vCenter Server controla los VDS (Virtual Distributed Switches)
- Otros componentes: vSphere, vCloud Director, vCloud Networking and Security, vCloud Automation Center, vCenter Site Recovery Manager, vCenter Operations Management Suite, vFabric Application Director for Provisioning
- Máximos vSphere 6.0:
  - 1024 VMs por host
  - 10 vNICs por VM
  - 1000 hosts por VDS
  - 1016 puertos de VDS activos por host
  - 60.000 puertos por VDS
  - 1000 hosts, 10.000 VMs en funcionamiento y 128 VDS por vCenter
  - 65.536 direcciones MAC por vCenter
  - 4/8 operaciones vMotion simultáneas por host por NIC 1/10Gbps
  - 16 VDS por host
  - etc



http://www.vmware.com/pdf/vsphere6/r60/vsphere-60-configuration-maximums.pdf

Nicira

- Fundada en 2007
- Miembro fundador del ONF
- En 2011 empieza a distribuir su NVP (*Network Virtualization Platform*)
- Es un controlador para OVS (Open vSwitch)
- No emplea solo OF sino OVSDB (Open vSwitch DataBase Management Protocol)
- Adquirida en 2013 por VMware (por unos 1260 millones de $)

Martin Casado

OvS
Open vSwitch

Otro software

- Frameworks
  - Onix, Trema, Maestro, Ryu
  - Indigo (para añadir OF a switches)
- FlowVisor:
  - https://github.com/OPENNETWORKINGLAB/flowvisor/wiki
  - Actúa como un proxy entre los switches y los controladores OF
  - Permite repartir recursos de la red entre varios controladores
- ONOS
  - http://onosproject.org/
  - Open Network Operating System
- Avior, Oflops, Cbench, Twister, FortNOX, LINC, Pantou, Of13softwitch, Cisco OnePK, Plexxi, etc etc etc
- ¡Se abrió la veda al software!

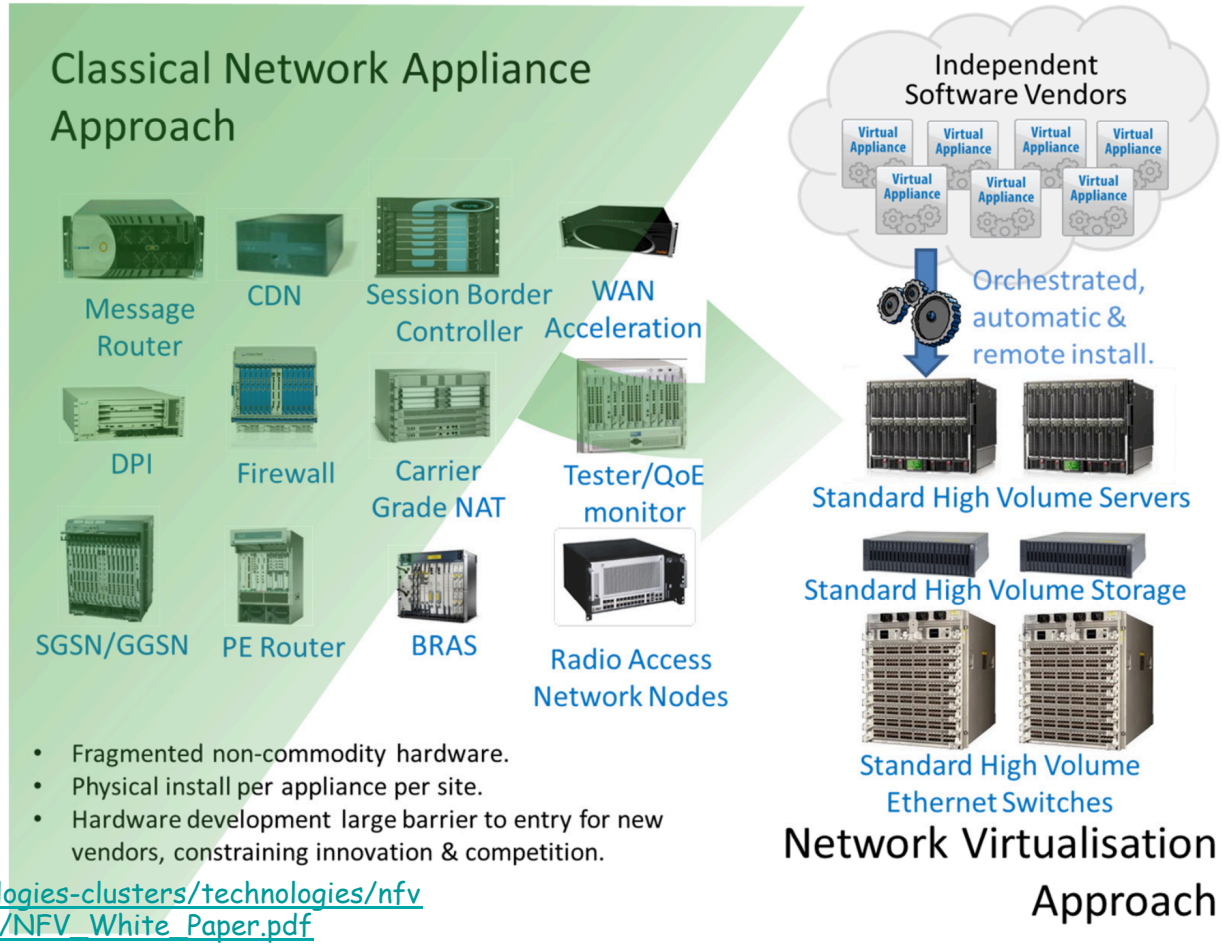
NFV

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



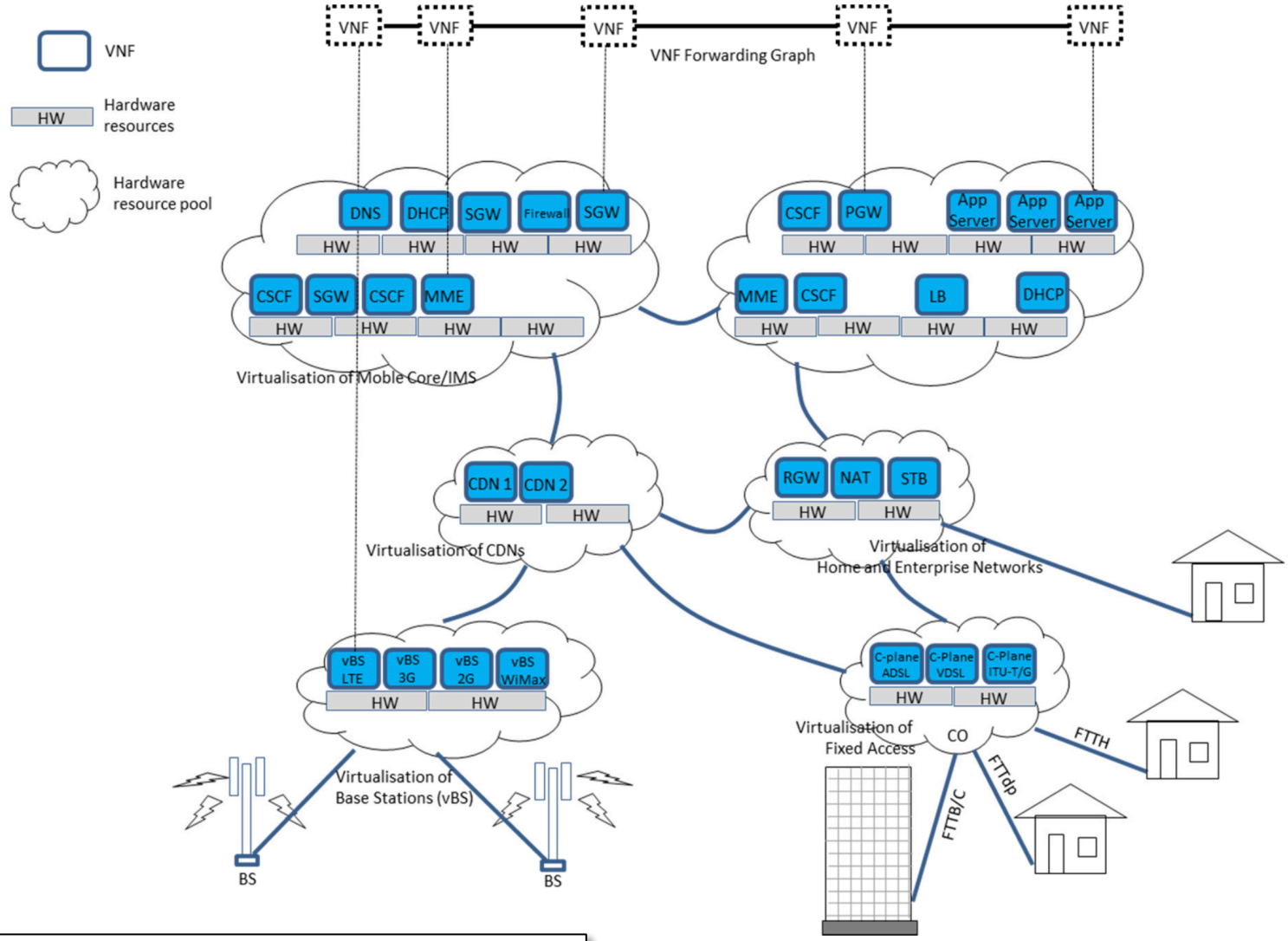
http://www.etsi.org/technologies-clusters/technologies/nfv
https://portal.etsi.org/NFV/NFV_White_Paper.pdf

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: B4N CG-NAT

WHAT IS B4N CG-NAT?

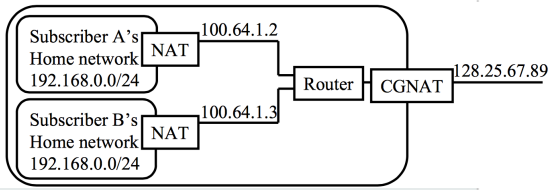
B4N CG-NAT is an NFV-based solution designed to provide high performance and transparent address and protocol translation. B4N CG-NAT helps Service Providers to mitigate IPv4 address exhaustion by using address and port translation in large scale and provides native integration within existent operators' infrastructure.

B4N CG-NAT is a fully virtualized and SDN ready solution that utilizes commodity x86 servers and provides carrier grade performance by using Intel® Data Plane Development Kit libraries.

CG-NAT solution provides maximum **500Gbps** throughput performance and fully compliant with **ETSI NFV ISG architecture**

B4N CG-NAT SPECIFICATIONS

	CONF.10	CONF.50	CONF.500	DISTRIBUTED
Max Throughput	**10 Gbps**	**50 Gbps**	**500 Gbps**	Unlimited
Connections per Second	200K per 10Gbps			Depends on OpenFlow switches performance, but not less than BOXED
Two-way sessions	10M per 10Gbps			
Resiliency	N+1. Active-Active, Active-Standby			
Supported protocols	NAT44 PCP			
Interfaces	REST API NETCONF			
Management	WEB CLI			
Supported hypervisors	LXC (Linux Containers) KVM VMware			
Logging	Local or external SYSLOG Server			



SCALABILITY
Simple extend capacity and performance by adding new B4N NAT VNFs and Distributed Switches, while maintaining existent network architecture

COMMODITY HARDWARE
Using commodity x86 servers instead of dedicated hardware devices

UNIFIED MANAGEMENT
Single point of management through powerful WEB-interface

AUTOMATION
B4N CG-NAT provides set of tools for automate service management

CONFORMANCE WITH REFERENCE ARCHITECTURE
Fully compliant with MANO Framework. B4N CG-NAT includes VNF-manager that can be integrated with Customer orchestration and management system.

SDN READY
B4N CG-NAT designed to be easy integrated with Customer SDN infrastructure

https://brain4net.com/solutions/cg-nat/

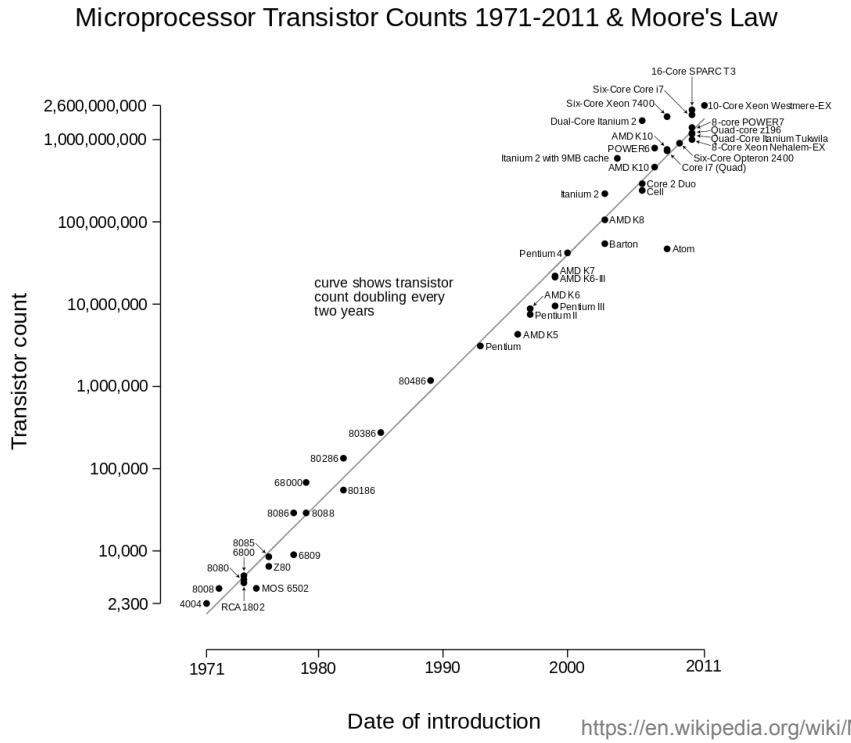
NFV

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

Networking hardware y el software

Evolución del hardware

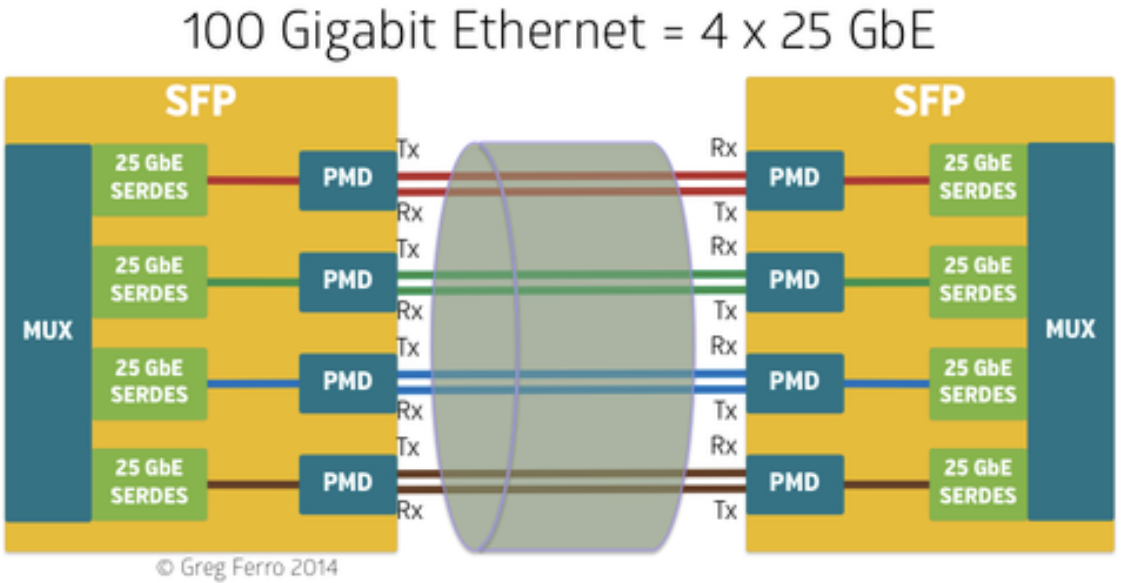
- 40 años de ley de Moore (x2 transistores cada 24 meses)
- Hoy en día ya estamos en 14nm
- Tick-tock
- Servidores ciclos de 2 años
- Networking reutilización 8-10 años



https://en.wikipedia.org/wiki/Moore's_law

Evolución

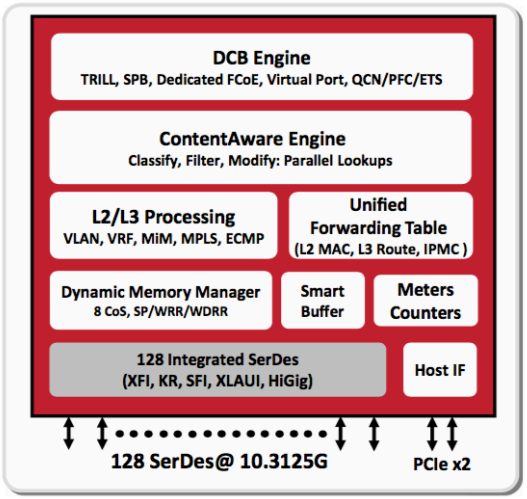
- Los fabricantes de equipos de red están adoptando los ritmos de producción de electrónica
- Empujados por pocos grandes clientes
- Por ejemplo: donde teníamos SerDes a 10Gbps los tendremos este año a 25Gbps, al mismo coste
- Esto permite interfaces 100GE donde antes teníamos 40GE, al mismo precio
- A día de hoy SoC (Switch on Chip) a 3.2Tbps

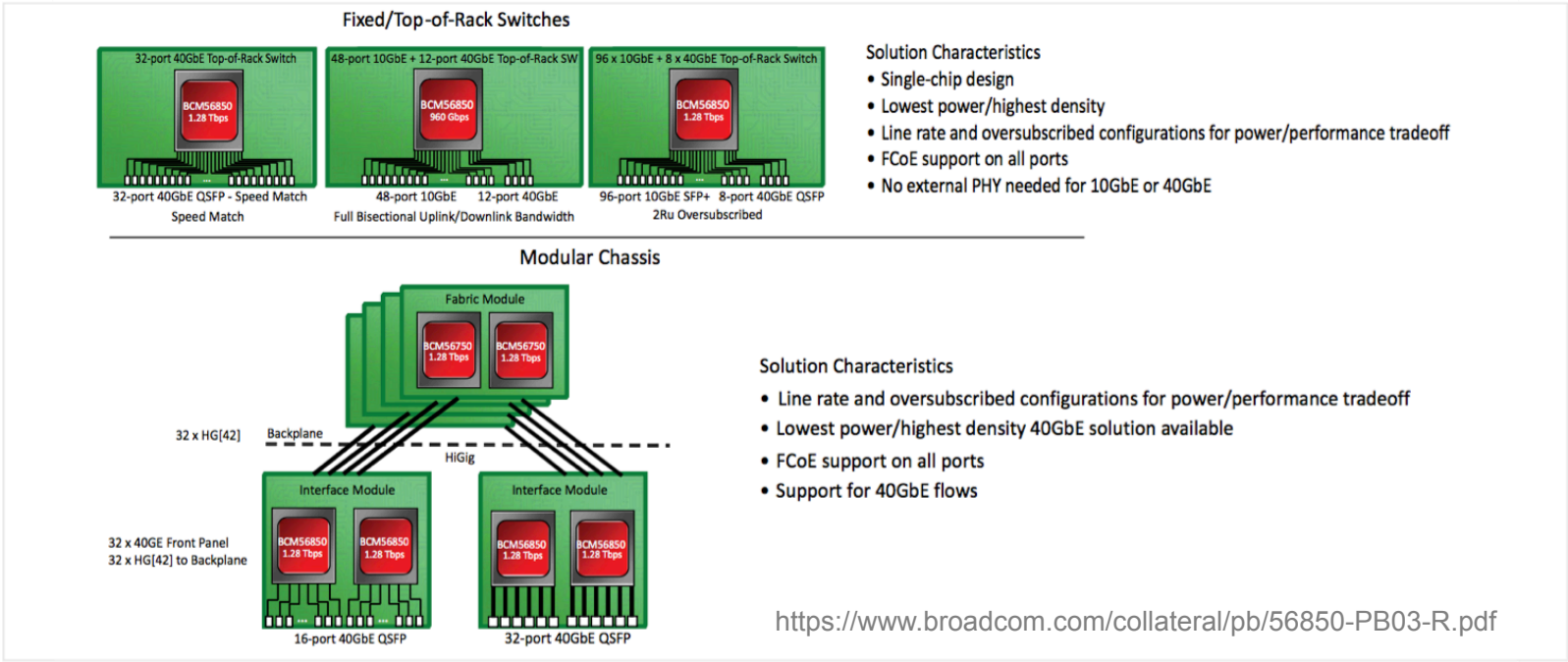


http://www.networkcomputing.com/data-centers/25-gbe-big-deal-will-arrive/1714647938

Broadcom Trident 2

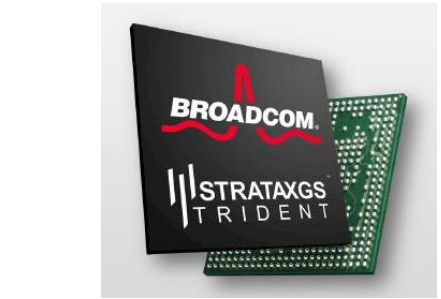
- 1.28 Tbps con puertos 10GE/40GE
- 128 SerDes 10GE (así que un máximo de 32 puertos 40GE en base a 4x10GE)
- Cut-through y Store&Forward
- VXLAN, NVGRE, 802.1Qbg EVR, 802.1BR
- Per VM traffic shaping
- DCB PFC, QCN y ETS. FCoE
- MPLS, VPLS, ISATAP, MAC-in-MAC, TRILL, SPB, Q-in-Q





Broadcom Tomahawk

- 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
- SerDes 25Gbps
- 10 colas por puerto
- Bridging de VXLAN a VLAN (no routing)
- NVGRE, MPLS, SPB
- Latencia de 300-500 ns



Trident 2 y Tomahawk

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

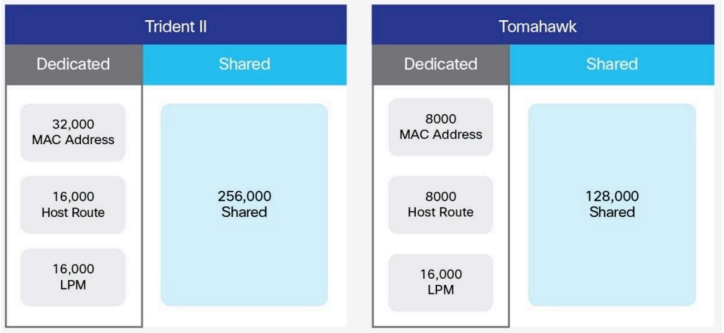


Table 1. Broadcom Trident 2 Forwarding Tables

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

Cisco ASE-2

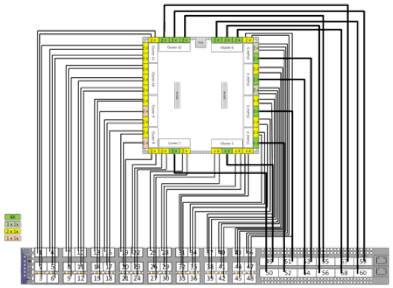
- ACI Spine Engine 2 (ACI = Applicacion Centric Infrastructure)
- 3.6Tbps (todos los tamaños de paquetes)
- 36x100GE, 72x40GE, 144x25GE
- 16K VRF, 32 SPAN, 64 mcast, 4K NAT
- Push/swap 5 etiquetas VPN
- DWRR con 16 colas por puerto
- WRED, ACN, AFD (Approximate Fair Dropping)

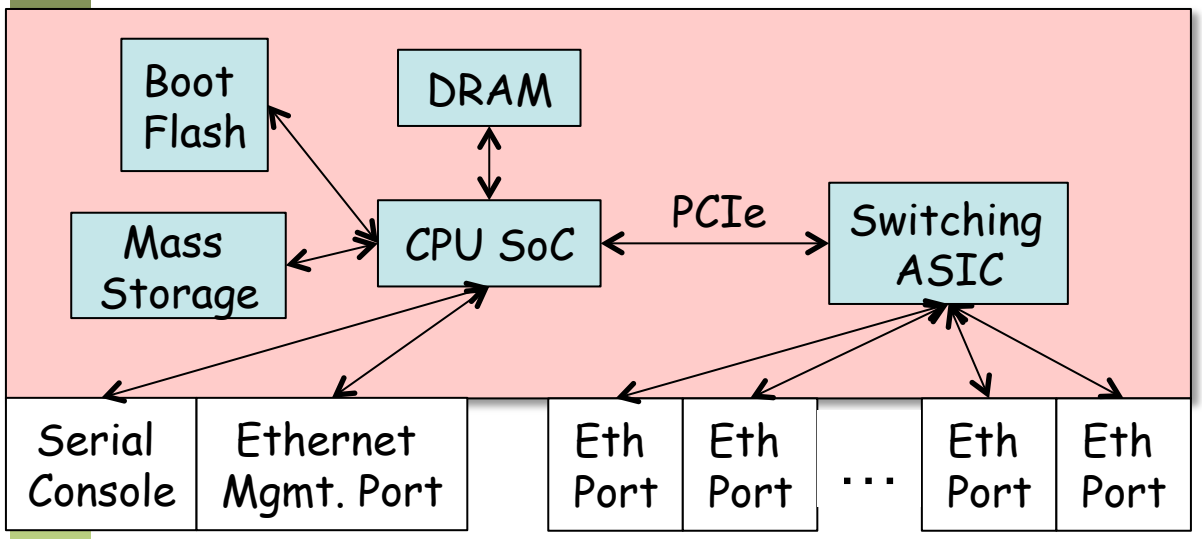
¿Evolución?

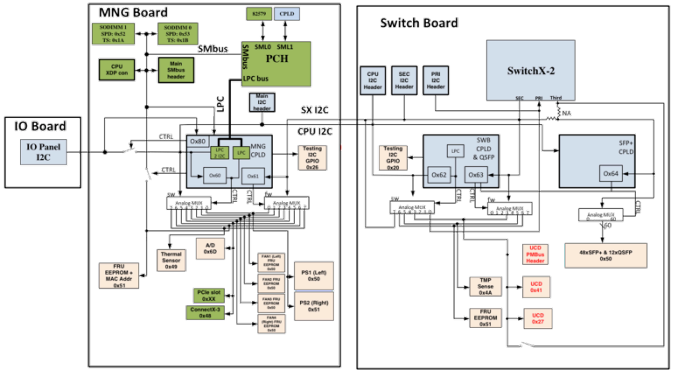
- La infraestructura se está simplificando
- Principalmente el hardware, controlable por software
- *White boxes* no solo servidores sino también switches
- También se venden ya switches *"Bare metal"* = solo el hardware

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:
  - 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

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática

Quanta Bare Metal Switch

A Powerful 10GBASE-T Top-of-Rack Switch for Data Center and Cloud Computing

- 10GBT
- ONIE Pre-loaded
- x86 CPU Board Support
- Ubuntu Compliant
- SDN Ready

Datacenter networks is facing a major paradigm shift toward the disaggregation of software and hardware. This move, combined with the benefits of software-defined networking (SDN) allows network administrators respond quickly to changing business requirements at a lower capital cost as well as reducing the network operations complexity.

QuantaMesh BMS products offer higher performance, increased availability, and better serviceability to meet datacenter installation environment. QuantaMesh T3048-LY9 supports 48 100/1000/10G Base-T ports and 6 40G QSFP+ ports in a compact rack unit size. By levering the new generation merchant silicon chips, T3048-LY9 is a high performance high density Ethernet switch with an affordable price for the deployment of data center infrastructure. Moreover, the CPU board design of T3048-LY9 provides option for 3rd party software choose. Finally, with ONIE (Open Network Installation Environment) pre-loaded, it provides the flexibility and allows choice of network operating system supported by different software vendors. This provides agile installation process and faster response for the changing business demand.

About QCT

QCT (Quanta Cloud Technology) is a global datacenter solution provider extending the power of hyperscale datacenter design in standard and open SKUs to all datacenter customers. Product lines include servers, storage, network switches, integrated rack systems and cloud solutions, all delivering hyperscale efficiency, scalability, reliability, manageability, serviceability and optimized performance for each workload. QCT offers a full spectrum of datacenter products and services from engineering, integration and optimization to global supply chain support, all under one roof. The parent of QCT is Quanta Computer Inc., a Fortune Global 500 technology engineering and manufacturing company.

Physical ports

- **Port configuration:** 48 100/1000/10GBASE-T and 6 QSFP+ ports
- **Management Port:** Out-of-band management port (RJ-45, 10/100/1000Base-T)
- **Console Port:** 1 (RJ-45)
- **USB:** 1 (v 2.0)

CPU Board 1

- **CPU:** Freescale P2020
- **Memory:** 2GB DDR3/ECC
- **Flash:** 128MB
- **Storage:** 8GB Micro SD

CPU Board 2

- **CPU:** Intel Rangeley
- **Memory:** 4GB DDR3/ECC
- **Storage:** 32GB SSD

Performance

- **MAC:** Unified Forwarding Table to dynamically allocate the L2/ L3 tables
- **Switching capacity:** 1.44Tbps
- **Maximum forwarding rate:** 1071Mpps
- **Latency:** <3us

High Availability

- **Redundant power supply:** 1+1
- **Hot-swappable fan tray:** N+1

Dell S4810-ON

The Dell Networking S4810-ON switch is the industry's first disaggregated hardware and software data center networking solution that empowers organizations to deploy modern workloads and applications designed for the open networking era.

Organizations that benefited from utilizing the disaggregation model with their data center server platforms can now leverage even greater benefits from Dell open networking platforms. Organizations can take advantage of this disaggregated networking model using industry-leading hardware and a choice of leading network operating systems to simplify data center fabric orchestration and automation and accelerate innovation.

These new offerings provide organizations the needed flexibility to transform their data centers and offer high-capacity network fabrics that are cost-effective, easy to deploy and provide a clear path to a software-defined data center.

The Dell S4810-ON supports the open source Open Network Install Environment (ONIE) for zero-touch installation of alternate network operating systems.

Ultra-low-latency, data center optimized

The Dell Networking S-Series S4810-ON is an ultra-low-latency 10/40GbE top-of-rack (ToR) switch purpose-built for applications in high-performance data center and computing environments. Leveraging a non-blocking switching architecture, the S4810 delivers line-rate L2 and L3 forwarding capacity with ultra-low-latency to maximize network performance. The compact S4810 design provides industry-leading density of 48 dual-speed 1/10GbE (SFP+) ports as well as four 40GbE QSFP+ uplinks to conserve valuable rack space and simplify the migration to 40Gbps in the data center core (Each 40GbE QSFP+ uplink can support four 10GbE ports with a breakout cable). In addition, the S4810 incorporates multiple architectural features that optimize data center network flexibility, efficiency and availability, including I/O panel to PSU airflow or PSU to I/O panel airflow for hot/cold aisle environments, and redundant, hot-swappable power supplies and fans.

Key applications

- Ultra-low-latency 10GbE switching in HPC, high-speed trading or other business-sensitive deployments that require the highest bandwidth and lowest latency
- High-density 10GbE ToR server aggregation in high-performance data center environments

Key features

- 1RU high-density 10/40GbE ToR switch with 48 dual-speed 1/10GbE (SFP+) ports and four 40GbE (QSFP+) uplinks (totaling 64 10GbE ports with breakout cables) with OS support
- 1.28Tbps (full-duplex) non-blocking switching fabric delivers line-rate performance under full load with sub 700ns latency
- I/O panel to PSU airflow or PSU to I/O panel airflow
- Supports the open source ONIE for zero-touch installation of alternate network operating systems
- Redundant, hot-swappable power supplies and fans
- Low power consumption

Performance

Switch fabric capacity:	1.28Tbps (full-duplex) 640Gbps (half-duplex)
Forwarding capacity:	960Mpps
Latency:	Sub 700ns
Packet buffer memory:	9MB
CPU memory:	2GB

Cisco Nexus 9300

The Cisco Nexus 9300 platform consists of fixed-port switches designed for top-of-rack (ToR) and middle-of-row (MoR) deployment in data centers that support enterprise applications, service provider hosting, and cloud computing environments. They are Layer 2 and 3 nonblocking 10 and 40 Gigabit Ethernet switches with up to 2.56 terabits per second (Tbps) of internal bandwidth.

Models

Table 1 summarizes the Cisco Nexus 9300 platform switch models.

Table 1. Cisco Nexus 9300 Platform Switches

Model	Description
Cisco Nexus 9332PQ Switch	32 x 40-Gbps Quad Enhanced Small Form-Factor Pluggable (QSFP+) ports
Cisco Nexus 9372PX-E Switch	48 x 1/10-Gbps SFP+ and 6 x 40-Gbps fixed QSFP+ ports
Cisco Nexus 9372TX-E Switch	48 x 1/10GBASE-T and 6 x 40-Gbps fixed QSFP+ ports
Cisco Nexus 9372PX Switch	48 x 1/10-Gbps SFP+ and 6 x 40-Gbps fixed QSFP+ ports
Cisco Nexus 9372TX Switch	48 x 1/10GBASE-T and 6 x 40-Gbps fixed QSFP+ ports
Cisco Nexus 9396PX Switch	48 x 1/10-Gbps SFP+ and up to 12 x 40-Gbps QSFP+ ports
Cisco Nexus 9396TX Switch	48 x 1/10GBASE-T and up to 12 x 40-Gbps QSFP+ ports
Cisco Nexus 93120TX Switch	96 x 1/10GBASE-T and 6 x 40-Gbps fixed QSFP+ ports
Cisco Nexus 93128TX Switch	96 x 1/10GBASE-T and up to 8 x 40-Gbps QSFP+ ports

Cisco Nexus 9300

All the Cisco Nexus 9300 platform switches use dual- core 2.5-GHz x86 CPUs with 64-GB solid-state disk (SSD) drives and 16 GB of memory for enhanced network performance.

Cisco provides two modes of operation for the Cisco Nexus 9000 Series. Organizations can use Cisco® NX-OS Software to deploy the Cisco Nexus 9000 Series in standard Cisco Nexus switch environments. Organizations also can use a hardware infrastructure that is ready to support Cisco Application Centric Infrastructure (Cisco ACI™) to take full advantage of an automated, policy-based, systems management approach.

- Power-On Auto Provisioning (POAP) automates the process of upgrading software images and installing configuration files on Cisco Nexus switches that are being deployed in the network for the first time.
- Intelligent API (iAPI) provides operators with a way to manage the switch through remote procedure calls (RPCs; JavaScript Object Notation [JSON] or XML) over HTTP/HTTPS infrastructure.
- Patching allows NX-OS to be upgraded and patched without any interruption in switch operations.
- Line-rate overlay support provides Virtual Extensible LAN (VXLAN) bridging and routing at full line rate, facilitating and accelerating communication between virtual and physical servers as well as between multiple data centers in a campus environment.
- Network traffic monitoring with Cisco Nexus Data Broker builds simple, scalable, and cost-effective network taps or Cisco Switched Port Analyzer (SPAN) aggregation for network traffic monitoring and analysis.
- Cisco Intelligent Traffic Director allows customers to build a highly scalable and flexible solution for hardware-based Layer 4 load balancing and traffic steering.

Cisco Nexus 9300

Item	Cisco Nexus 9300 Platform
Maximum number of longest prefix match (LPM) routes	128,000*
Maximum number of IP host entries	208,000*
Maximum number of MAC address entries	96,000*
Number of multicast routes	• 32,000 (without virtual PortChannel [vPC]) • 32,000 (with vPC)
Number of Interior Gateway Management Protocol (IGMP) snooping groups	• 32,000 (without vPC) • 32,000 (with vPC)
Maximum number of Cisco Nexus 2000 Series Fabric Extenders per switch	16
Number of access control list (ACL) entries	• 4000 ingress • 1000 egress
Maximum number of VLANs	4096
Maximum number of Virtual Routing and Forwarding (VRF) instances	1000
Maximum number of links in a PortChannel	32
Maximum number of ECMP paths	64
Maximum number of PortChannels	528
Number of active SPAN sessions	4
Maximum number of Rapid per-VLAN Spanning Tree (RPVST) instances	507
Maximum number of Hot-Standby Router Protocol (HSRP) groups	490
Maximum number of Multiple Spanning Tree (MST) instances	64
Maximum number of VXLAN tunnel endpoints (VTEP)	256

* The actual maximum scale depends on the system forwarding mode.

HPE FlexFabric 5930 Switch

Key features

- Cut-through with ultra-low-latency and wire speed
- VXLAN VTEP OVSDB support for virtualized environments
- High-density 10GbE and 40GbE spine/ToR connectivity
- IPv6 support with full L2 and L3 features
- Convergence-ready with DCB, FCoE, and TRILL

http://www8.hp.com/h20195/v2/GetDocument.aspx?docname=c04111326

HPE FlexFabric 5930 Switch

Quality of Service (QoS)

- **Powerful QoS features**
  - **Flexible queue scheduling**
    including Strict Priority (SP), WRR, WDRR, WFQ, SP+WRR, SP+WDRR, SP+WFQ, Configurable Buffer, Time range, Queue Shaping, CAR with 8kbps granularity.
  - **Packet filtering and remarking:**
    packet filtering at L2 (Layer 2) through L4 (Layer 4); flow classification based on source MAC address, destination MAC address, source IP (IPv4/IPv6) address, destination IP (IPv4/IPv6) address, port, protocol, and VLAN.

cut-through and nonblocking architecture delivers low latency (~1 microsecond for 10GbE) for very demanding enterprise applications; the switch delivers high-performance switching capacity and wire-speed packet forwarding

- **Higher scalability**
  HPE Intelligent Resilient Fabric (IRF) technology simplifies the architecture of server access networks; up to nine HP 5930 switches can be combined to deliver unmatched scalability of virtualized access layer switches and flatter two-tier networks using IRF, which reduces cost and complexity
- **Advanced modular operating system**
  Comware v7 software's modular design and multiple processes bring native high stability, independent process monitoring, and restart; the OS also allows individual software modules to be upgraded for higher availability and supports enhanced serviceability functions like hitless software upgrades with single-chassis ISSU
- **TRILL, SPB and EVB/VEPA**
  TRansparent Interconnection of Lots of Links (TRILL) is supported including support of TRILL with IRF, TRILL ECMP up to 8 paths. Support for Shortest Path Bridging (IEEE 802.1aq) with ECMP up to 8 paths. Edge Virtual Bridging with Virtual Ethernet Port Aggregator (EVB/VEPA) provides connectivity into the virtual environment for a data center-ready environment

http://www8.hp.com/h20195/v2/GetDocument.aspx?docname=c04111326

HPE FlexFabric 5930 Switch

- **Data Center Bridging (DCB) protocols**
  provides support for IEEE 802.1Qbb Priority Flow Control (PFC), Data Center Bridging Exchange (DCBX), IEEE 802.1Qaz Enhanced Transmission Selection (ETS), Explicit Congestion Notification (ECN) for converged FCoE, iSCSI and RoCE environments.
- **FCoE support**
  provides support T11 standards-compliant FC-BB-5 Fibre Channel over Ethernet (FCoE), including FCoE Initialization Protocol (FIP), FCP, Fiber Channel enhanced port types VE, TE and VF, NPV, NPIV, Fabric Name Server, RSCN, Login Services, and name-server zoning, Per-VSAN Fabric Services, FSPF, Standard Zoning and Fiber Channel Ping.
- **Jumbo frames**
  with frame sizes of up to 10,000 bytes on Gigabit Ethernet and 10-Gigabit ports, allows high-performance remote backup and disaster-recovery services to be enabled
- **VXLAN Support**
  VXLAN Layer 2 Gateway support for up to 4k tunnels
- **Dynamic VXLAN configuration**
  OVSDB support for dynamic VXLAN configuration

http://www8.hp.com/h20195/v2/GetDocument.aspx?docname=c04111326

HPE FlexFabric 5930 Switch

Layer 3 routing

- **Virtual Router Redundancy Protocol (VRRP) and VRRP Extended**
  allow quick failover of router ports
- **Policy-based routing**
  makes routing decisions based on policies set by the network administrator
- **Equal-Cost Multipath (ECMP)**
  enables multiple equal-cost links in a routing environment to increase link redundancy and scale bandwidth
- **Layer 3 IPv4 routing**
  provides routing of IPv4 at media speed; supports static routes, RIP and RIPv2, OSPF, BGP, and IS-IS
- **Open shortest path first (OSPF)**
  delivers faster convergence; uses this link-state routing Interior Gateway Protocol (IGP), which supports ECMP, NSSA, and MD5 authentication for increased security and graceful restart for faster failure recovery
- **Border Gateway Protocol 4 (BGP-4)**
  delivers an implementation of the Exterior Gateway Protocol (EGP) utilizing path vectors; uses TCP for enhanced reliability for the route discovery process; reduces bandwidth consumption by advertising only incremental updates; supports extensive policies for increased flexibility; scales to very large networks
- **Intermediate system to intermediate system (IS-IS)**
  uses a path vector Interior Gateway Protocol (IGP), which is defined by the ISO organization for IS-IS routing and extended by IETF RFC 1195 to operate in both TCP/IP and the OSI reference model (Integrated IS-IS)
- **Static IPv6 routing**
  provides simple manually configured IPv6 routing
- **Dual IP stack**
  maintains separate stacks for IPv4 and IPv6 to ease the transition from an IPv4-only network to an IPv6-only network design

http://www8.hp.com/h20195/v2/GetDocument.aspx?docname=c04111326

HPE FlexFabric 5930 Switch

- **Routing Information Protocol next generation (RIPng)**
  extends RIPv2 to support IPv6 addressing
- **OSPFv3**
  provides OSPF support for IPv6
- **BGP+**
  extends BGP-4 to support Multiprotocol BGP (MBGP), including support for IPv6 addressing
- **IS-IS for IPv6**
  extends IS-IS to support IPv6 addressing
- **IPv6 tunneling**
  allows IPv6 packets to traverse IPv4-only networks by encapsulating the IPv6 packet into a standard IPv4 packet; supports manually configured, 6to4, and Intra-Site Automatic Tunnel Addressing Protocol (ISATAP) tunnels; is an important element for the transition from IPv4 to IPv6
- **Policy routing**
  allows custom filters for increased performance and security; supports ACLs, IP prefix, AS paths, community lists, and aggregate policies
- **Bidirectional Forwarding Detection (BFD)**
  enables link connectivity monitoring and reduces network convergence time for RIP, OSPF, BGP, IS-IS, VRRP, MPLS, and IRF
- **Multicast Routing PIM Dense and Sparse modes**
  provides robust support of multicast protocols
- **Layer 3 IPv6 routing**
  provides routing of IPv6 at media speed; supports static routing, RIPng, OSPFv3, BGP4+ for IPv6, and IS-ISv6

http://www8.hp.com/h20195/v2/GetDocument.aspx?docname=c04111326

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)
- etc