Arquitectura de los switches Nexus de Cisco serie 9500

Informe técnico

Noviembre de 2013

Contenido

Introducción a los switches Nexus serie 9500

Los switches Nexus de Cisco serie 9500 son una familia de switches modulares que proporcionan una conectividad de alto rendimiento, alta densidad y baja latencia de 1, 10, 40 y (en el futuro) 100 Gigabit Ethernet líder en el sector. Los switches Nexus serie 9500 pueden funcionar tanto en modo de infraestructura centrada en aplicaciones (ACI) como en el modo NX-OS clásico. Cuando funcionan en modo ACI, los switches Nexus serie 9500 sientan las bases de la transformadora arquitectura ACI para la solución de fabric de red automatizada y completamente integrada que impulsa un perfil de red de aplicaciones. Cuando se ejecutan en el modo NX-OS clásico, los switches Nexus serie 9500 son los mejores de su categoría gracias a sus capas de agregación y acceso al Data Center de elevado rendimiento y gran escalabilidad, con funcionalidades de automatización y programabilidad mejoradas. Este informe técnico se centra en la arquitectura de hardware habitual de los switches Nexus serie 9500 y la implementación de reenvío de paquetes en el modo NX-OS clásico.

El switch Nexus 9508 de 8 ranuras (figura 1) es la primera plataforma disponible de la familia; a esta le seguirán las plataformas de 4 ranuras y de 16 ranuras. El switch Nexus de Cisco serie 9508 admite hasta 1152 puertos de 10 GE o 288 puertos de 40 GE. El switch Nexus de Cisco serie 9516 doblará las densidades de puertos. Además, los switches Nexus serie 9500 proporcionan una elevada densidad de puertos para las conectividades de 1G SFP/1GBase-T y 10G SFP+/10GBaseT. Gracias a sus variados diseños de chasis, distintos tipos de tarjeta de línea y velocidades de puerto de panel frontal flexibles, el switch Nexus de Cisco serie 9500 ofrece excelentes soluciones de redes para todos los Data Centers de procesos de extremada importancia, independientemente de si son grandes, medianos o pequeños.

Figura 1. Switch Nexus de Cisco serie 9508

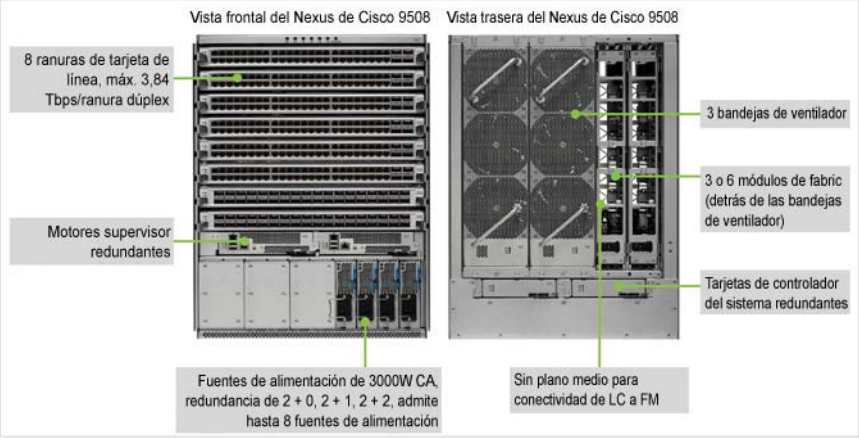


Tabla 1. Características de reenvío y chasis del switch Nexus de Cisco serie 9500

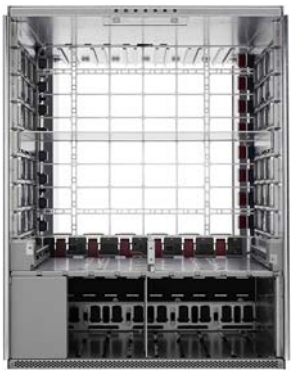
Métrica	NEXUS 9504	NEXUS 9508	NEXUS 9516
Altura	7 RU	13 RU	20 RU
Ranuras de supervisor	2	2	2
Ranuras de módulo de fabric	6	6	6
Ranuras de tarjeta de línea	4	8	16
Ancho de banda de fabric máximo por ranura (Tbps)	3,84 Tb/s	3,84 Tb/s	3,84 Tb/s
Ancho de banda de fabric máximo por sistema (Tbps)	15 Tb/s	30 Tb/s	60 Tb/s
Máximo de puertos de 1/10/40	192/576/144	384/1152/288	768/2304/576
Rendimiento de reenvío máximo por tarjeta de línea (Tbps)	2,88 Tb/s	2,88 Tb/s	2,88 Tb/s
Rendimiento de reenvío máximo por sistema (Tbps)	11,52 Tb/s	23,04 Tb/s	46,08 Tb/s

Métrica	NEXUS 9504	NEXUS 9508	NEXUS 9516
Flujo de aire	Parte frontal a la trasera	Parte frontal a la trasera	Parte frontal a la trasera
Fuentes de alimentación	4 PSU de 3 KW CA	8 PSU de 3 KW CA	8 PSU de 3 KW CA
Bandejas de ventilador	3	3	3

Los switches Nexus de Cisco serie 9500 cuentan con una arquitectura modular que se compone de chasis de switch, supervisores, controladores de sistema, módulos de fabric, tarjetas de línea, fuentes de alimentación y bandejas de ventilador. De todo esto, los supervisores, los controladores de sistema, las tarjetas de línea y las fuentes de alimentación son los componentes habituales que se pueden compartir en toda la familia de productos Nexus 9500.

El chasis del switch Nexus de Cisco serie 9500 incluye un innovador diseño sin plano medio (figura 2). El plano medio se suele usar en plataformas modulares con el fin de proporcionar conectividad entre las tarjetas de línea y los módulos de fabric. Dado que es una parte extra de hardware dentro del chasis del switch, obstruye el flujo de aire de refrigeración. De ahí que sea necesario aplicar métodos adicionales con objeto de crear una vía para el flujo de aire (por ejemplo, cortes en el plano medio o redirección del flujo) que dan como resultado una baja eficacia de la refrigeración. Nexus serie 9500 es la primera plataforma de switch del sector que elimina la necesidad del plano medio en el chasis. Con un preciso mecanismo de alineación, las tarjetas de línea y los módulos de fabric de los switches Nexus serie 9500 se conectan directamente entre sí con clavijas de conexión. Las tarjetas de línea y los módulos de fabric tienen orientaciones ortogonales en el chasis para que cada módulo de fabric se conecte con todas las tarjetas de línea y viceversa. Sin un plano medio que bloquee la vía del flujo de aire, el diseño del chasis ofrece una mayor eficacia de la refrigeración. Esto también permite un diseño de chasis compacto sin necesidad de grandes ventiladores de refrigeración.

Figura 2. Diseño del chasis sin plano medio de Nexus 9500



El diseño de chasis sin plano medio simplifica en gran medida la implementación de la plataforma de switches y la actualización del hardware. En algunos casos en los que se introducen nuevos componentes, como tarjetas de línea o módulos de fabric nuevos, se requiere una actualización del plano medio. Esto agrega complejidad y más interrupciones del servicio al proceso de actualización del hardware. El switch Nexus de Cisco serie 9500 reduce la necesidad de la instalación o actualización del plano medio. Otra ventaja de quitar el plano medio es que el tiempo medio de las reparaciones es significativamente menor. Con un plano medio, si se dobla una clavija de este, se deberá detener el funcionamiento de todo el switch y desmontarse para sustituir el plano medio. Con el 9500, los componentes dañados se pueden sustituir sin dejar fuera de servicio al resto de componentes del chasis.

Además de la magnífica eficacia de la refrigeración, el switch Nexus de Cisco serie 9500 también es líder gracias a su elevada eficiencia energética. Sus fuentes de alimentación han obtenido la certificación 80 PLUS Platinum estándar del sector por su alta eficacia energética. Las tarjetas de línea y los módulos de fabric del Nexus 9500 se han diseñado con un número mínimo de ASIC, lo que reduce la cantidad de bolsas de calor en un módulo. Como resultado de estas innovaciones, se logra el consumo energético más bajo posible y sin igual por puerto:

Consumo de electricidad por puerto	Puerto de 10 Gbps	Puerto de 40 Gbps
Vatios por puerto	3,85 W/puerto	15,4 W/puerto

Plano de control escalable en los switches Nexus de Cisco serie 9500

El motor supervisor Cisco Nexus 9500 proporciona un plano de control escalable para los switches Nexus de Cisco serie 9500. El controlador de sistemas descarga la conectividad de los componentes internos y las funciones de gestión del motor supervisor. La separación de las tareas de gestión interna del motor supervisor aumenta la fiabilidad del plano de control del switch. Proporciona una mejor modularidad y resistencia de todo el sistema de switches.

Motor supervisor

Cisco Nexus serie 9500 admite motores supervisores de media anchura redundantes que son responsables de las funciones del plano de control. El software del switch, NX-OS mejorado, se ejecuta en los módulos supervisores. Los módulos supervisores redundantes realizan funciones activas y en espera compatibles con stateful switch over en caso de que se produzca un fallo en el hardware del módulo supervisor y y la actualización de software en funcionamiento (ISSU), lo que posibilita la actualización o el mantenimiento del software sin que ello afecte a los servicios de producción.

El complejo de CPU del supervisor de Nexus 9500 está basado en la plataforma Romley de Intel con procesadores de 4 núcleos Sandy Bridge Xeon. El tamaño predeterminado de la memoria del sistema es de 16 GB, actualizable sobre el terreno hasta 48 GB. Incluye una unidad SSD integrada de 64 GB para ofrecer almacenamiento adicional no volátil incorporado. La CPU de alta velocidad de varios núcleos y la amplia memoria constituyen las bases de un plano de control rápido y fiable para el sistema de switches. Los protocolos del plano de control se beneficiarán de la amplia capacidad informática y lograrán un inicio veloz y una convergencia instantánea ante los cambios de estado de la red. Además, la DRAM de gran tamaño y ampliable, y la CPU de varios núcleos ofrecen suficientes recursos y capacidad informática para admitir contenedores basados grupos de control de Linux en los que se pueden instalar y ejecutar aplicaciones de terceros en un entorno bien contenido. La unidad SSD integrada proporciona almacenamiento extra para registros, archivos de imagen y aplicaciones de terceros.

Figura 3. Motor supervisor Cisco Nexus serie 9500

Módulo supervisor	
Procesador	Romley, 1,8 GHz, 4 núcleos
Memoria del sistema	16 GB, actualizable a 48 GB
Puertos serie RS-232	Uno (RJ-45)
Puertos de gestión de 10/100/1000	Uno (RJ-45)
Interfaz USB 2.0	Dos
Almacenamiento SSD	64 GB

El motor supervisor tiene un puerto de consola serie (RJ-45) y un puerto de gestión de 10/100/1000 Ethernet (RJ-45) para una gestión fuera de banda. Se dispone de dos interfaces USB 2.0 mediante una unidad de almacenamiento flash USB externa para imágenes, syslog, transferencia de archivos de configuración y otros usos. Un puerto de entrada de reloj de pulso por segundo (PPS) en el módulo supervisor posibilita una sincronización integrada precisa.

La comunicación entre los módulos de fabric y de supervisor o las tarjetas de línea utilizan un canal fuera de banda Ethernet (EOBC, por sus siglas en inglés) o canal de protocolo Ethernet (EPC, por sus siglas en inglés). Ambos canales tienen un hub central en los controladores del sistema que ofrece rutas redundantes a los controladores del sistema.

Controladores del sistema

Los controladores del sistema de Cisco Nexus serie 9500 se usan para descargar las funciones de gestión y switching sin ruta de datos internas de los motores supervisor. Además, proporcionan la vía para acceder a las fuentes de alimentación y las bandejas de ventilador.

Los controladores de sistemas son los switches centrales de comunicación interna. Cada uno aloja dos rutas de comunicación de gestión y control principales, el canal fuera de banda Ethernet (EOBC, por sus siglas en inglés) y el canal de protocolo Ethernet (EPC, por sus siglas en inglés), entre los motores supervisores, las tarjetas de línea y los módulos de fabric.

Toda la comunicación de gestión dentro del sistema entre los módulos tiene lugar a través del canal EOBC. El canal EOBC se proporciona mediante un conjunto de chips de switch en los controladores del sistema que interconectan todos los módulos juntos, incluidos los motores supervisores, los módulos de fabric y las tarjetas de línea.

El canal EPC gestiona la comunicación por protocolo del plano de datos dentro del sistema. Esta vía de comunicación la proporciona otro conjunto de chips de switch Ethernet redundante en los controladores del sistema. A diferencia del canal EOBC, el switch EPC solo conecta los módulos de fabric con los motores supervisores. Si es necesario enviar los paquetes de protocolo a los supervisores, las tarjetas de línea utilizarán la ruta de datos interna para transferir los paquetes a los módulos de fabric. Seguidamente, los módulos de fabric redirigirán el paquete a través del canal EPC a los motores supervisores.

El controlador del sistema también se comunica con las unidades de fuentes de alimentación y los controladores de ventiladores, y los gestiona, mediante el bus de administración del sistema (SMB) redundante.

Cisco Nexus serie 9500 admite los controladores del sistema redundantes. Cuando hay dos controladores del sistema presentes en un chasis, un proceso arbitrario seleccionará el controlador del sistema activo. El otro asumirá una función secundaria o en espera para ofrecer redundancia.

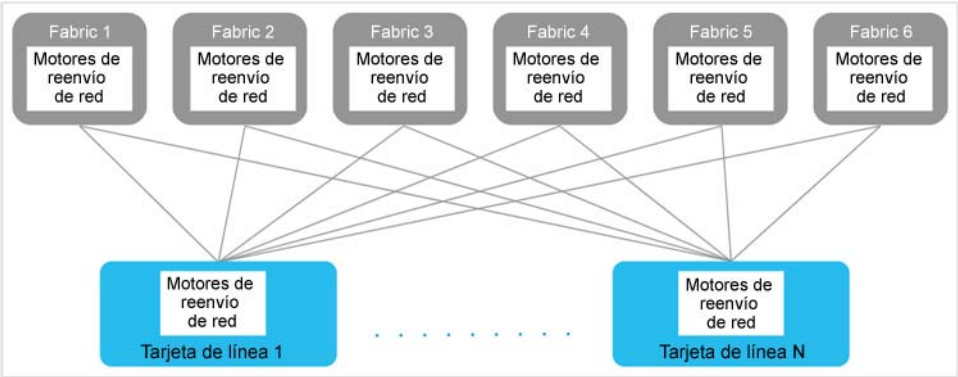
Figura 4. Controlador del sistema Cisco Nexus serie 9500

Plano de datos distribuido sin bloqueos en los switches Nexus de Cisco serie 9500

Mientras el plano de control de switch se ejecuta de forma centralizada en los motores supervisores, las funciones de búsqueda y reenvío de paquetes en el plano de datos se lleva a cabo de una forma de elevada distribución en la que participan las tarjetas de línea y los módulos de fabric.

Tanto las tarjetas de línea como los módulos de fabric del Cisco Nexus serie 9500 están equipados con varios motores de reenvío de red (NFE, por sus siglas en inglés) que llevan a cabo las funciones de reenvío, procesamiento y búsqueda de paquetes. Los switches Nexus serie 9500 han sido ideados para una arquitectura sin bloqueos y un rendimiento con velocidad de línea completa en todos los puertos, con independencia del tamaño del paquete. Dado que muchas aplicaciones de Data Center modernas usan paquetes de tamaño reducido, es esencial permitir un rendimiento de velocidad de línea para incluso el paquete más pequeño de 64 bytes. Con el fin de lograr este nivel de capacidad de reenvío, las tarjetas de línea y los módulos de fabric de Nexus serie 9500 disponen de una arquitectura con el número requerido de NFE. Se usan hasta 24 puertos de 40 GE en cada NFE para garantizar un rendimiento de velocidad de línea. Entre los 24 puertos de 40 GE, 12 puertos de 40 GE, que llegan a los 42 GE para dar respuesta a los bits extra del encabezado de trama interna, se usan para la conectividad interna hacia los módulos del fabric. Los otros 12 puertos se usan como interfaces del panel frontal para admitir puertos de datos de usuarios de 1, 10, 40 y (en el futuro) 100 GE.

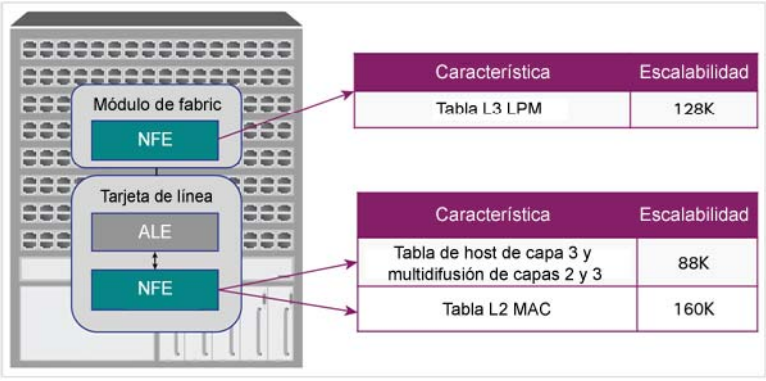
Figura 5. Plano de datos distribuido de los switches Nexus serie 9500



Los motores de reenvío de redes usan una combinación de espacio de tabla TCAM dedicado y memoria de tabla de hash compartida, conocida como tabla de reenvío unificada (UFT, por sus siglas en inglés) para almacenar la información de reenvío de la capa 2 y la capa 3. La UFT se puede separar de forma flexible en tres tablas de reenvíos, la tabla de direcciones MAC, la tabla de host IP y la tabla LPM. Este enfoque de uso compartido de la memoria programable ofrece flexibilidad para ocuparse de las distintas situaciones de implementación y aumenta la eficiencia de la utilización de recursos de la memoria.

Para maximizar la escalabilidad de reenvío del sistema, los switches Nexus serie 9500 están diseñados para usar las tablas UFT en tarjetas de línea y módulos de fabric para distintas funciones de búsqueda de reenvío. La UFT en tarjetas de línea almacena la tabla de MAC de capa 2 y la tabla de host de capa 3. Por tanto, las tarjetas de línea son responsables de la búsqueda de switching de capa 2 y la búsqueda de routing de capa 3. La UFT en módulos de fabric aloja la tabla LPM de capa 3 y realiza la búsqueda de routing de LPM de capa 3. Tanto las tarjetas de línea como los módulos de fabric cuentan con tablas multidifusión y toman parte en búsquedas multidifusión distribuidas y replicación de paquetes. La multidifusión comparte el mismo recurso de tabla con las entradas de host de capa 3 en las tarjetas de línea. En la figura 6 se muestra la escalabilidad de reenvío de todo el sistema de los switches Nexus serie 9500.

Figura 6. Escalabilidad de reenvío de todo el sistema Nexus 9500

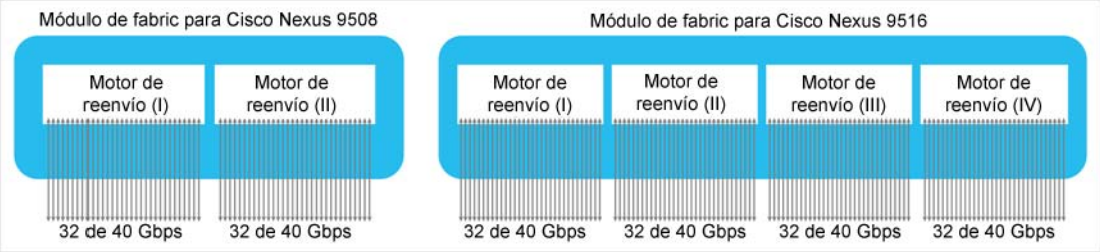


Módulo de fabric Nexus serie 9500

Un switch Nexus serie 9500 puede tener hasta seis módulos de fabric que funcionen en modo activo. Cada módulo de fabric está formado por varios motores de reenvío de red (NFE), 2 para un switch Nexus 9508 y 4 para un switch Nexus 9516 (figura 7).

Para un switch Nexus 9508, pueden haber disponibles hasta doce NFE en sus modelos de fabric. Esto proporciona el ancho de banda de ruta de datos y la capacidad de reenvío de paquetes necesarios para lograr una verdadera arquitectura sin bloqueos. De ahí que el Nexus 9508 sea compatible con un verdadero rendimiento de velocidad de línea, independientemente del tamaño del paquete, en todas las tarjetas de línea.

Figura 7. Módulo de fabric Nexus serie 9500



El módulo de fabric de los switches Nexus serie 9500 realiza estas importantes funciones en la arquitectura de chasis modular:

- Proporcionar una conectividad de reenvío de datos sin bloqueos de alta velocidad para las tarjetas de línea. Todos los enlaces en los motores de reenvío de red son rutas de datos activas. Cada módulo de fabric puede aportar hasta 8 enlaces de 40 Gbps para cada ranura de tarjeta de línea. Un chasis de Nexus 9500 implementado con 6 módulos de fabric puede proporcionar potencialmente 48 rutas de fabric de 40 Gbps a cada ranura de tarjeta de línea. Esto equivale a un ancho de banda dúplex completo de 3,84 Tbps por ranura.
- Realizar búsquedas de routing de correspondencia de prefijo más extenso (LPM, por sus siglas en inglés) distribuidas para tráfico IPv4 e IPv6. La información de reenvío de LPM se almacena en los módulos de fabric de un switch Nexus serie 9500. Es compatible con prefijos de IPv4 de hasta 128K o prefijos IPv6 de 32K.
- Realizar búsquedas multidifusión distribuidas y replicación de paquetes para enviar copias de paquetes multidifusión para recibir NFE de salida.

Arquitectura de la tarjeta de línea de los switches Nexus serie 9500

Una tarjeta de línea de switches Nexus serie 9500 se puede categorizar en dos tipos: tarjetas de línea de agregación y tarjetas de línea de hoja preparadas para infraestructura centrada en aplicaciones (ACI). Las tarjetas de línea de agregación proporcionan una conectividad de alta densidad de 10GE/40GE en un switch Nexus 9500 que se ejecuta en el modo NX-OS clásico. Las tarjetas de línea de hoja preparadas para ACI pueden funcionar tanto en el modo NX-OS clásico como en el modo ACI.

Todas las tarjetas de línea de Nexus 9500 incluyen varios NFE para el reenvío y la búsqueda de paquetes. Además, las tarjetas de línea de hoja preparadas para ACI incluyen un conjunto de motores de hoja de aplicación (ALE, por sus siglas en inglés). Como su nombre indica, ALE realiza las funciones de nodo de hoja de ACI cuando el switch Nexus 9500 se implementa como nodo de hoja en una infraestructura de ACI. Cuando el switch Nexus 9500 funciona en el modo NX-OS clásico, el ALE en una tarjeta de línea de hoja preparado para ACI ofrece principalmente almacenamiento en búfer adicional y facilita algunas funciones de red, como el routing dentro de una superposición VxLAN.

Los NFE de una tarjeta de línea realizan búsquedas de switching de capa 2 y búsquedas de routing de capa 3. Las tarjetas de línea están equipadas con una cantidad variable de NFE para posibilitar el rendimiento de reenvío de velocidad de línea completa para todos los tamaños de paquetes de IP en todos los puertos de paneles frontales.

Además del rendimiento del plano de datos de velocidad de línea, las tarjetas de línea del switch Nexus serie 9500 también incluyen una CPU de doble núcleo integrada. Esta CPU se usa para descargar o acelerar algunas tareas del plano de control como programar los recursos de la tabla de hardware, recopilar y enviar los recuentos y estadísticas de las tarjetas de línea, y descargar la gestión del protocolo de BFD de los supervisores. Esto proporciona una mejora significativa del rendimiento del plano de control del sistema.

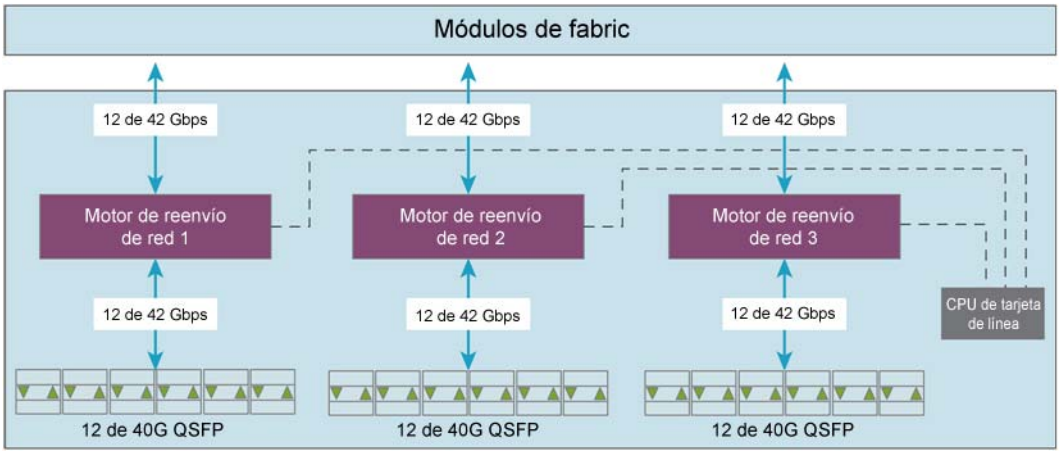
Tarjeta de línea de 36 puertos QSFP de 40 GE (N9K-X9636PQ)

N9K-X9636PQ (figura 8) es una tarjeta de línea de agregación que proporciona 36 puertos de panel frontal QSFP de 40 GE. Incluye tres motores de reenvío de red para el reenvío de paquetes, cada uno de ellos compatible con 12 puertos de panel frontal de 40 GE y 12 puertos internos a los módulos de fabric (que llegan a los 42 Gbps para responder a las sobrecargas internas). Los 36 puertos del panel frontal de 40 GE del N9K-X9636PQ admiten los 4 modos introductorios de 10 GE para que funcionen como 4 puertos de 10 GE individuales. Esto permite que la tarjeta de línea proporcione hasta 144 puertos SFP + de 10 GE.

Esta tarjeta de línea cuenta con un diseño sin PHY. Esto reduce la latencia de transporte de datos en el puerto 100 ns, reduce el consumo de electricidad del puerto y mejora la fiabilidad debido al menor número de componentes activos.

Las longitudes de seguimiento desde cada NFE a los 12 cables ópticos QSFP que admite se encuentran por debajo de las 7”, lo que reduce la necesidad de nuevos temporizadores. Esto simplifica aún más el diseño de la tarjeta de línea y reduce el número de componentes activos.

Figura 8. Tarjeta de línea de 36 puertos QSFP de 40 GE Nexus serie 9500



Tarjeta de línea de 48 puertos 1/10 G SFP+ (N9K-X9564PX)

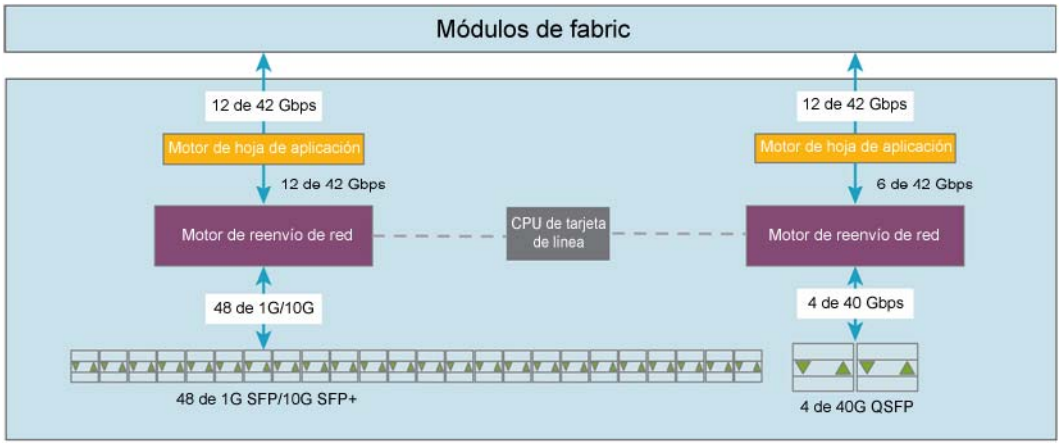
N9K-X9564PX (figura 9) es una tarjeta de línea de hoja preparada para ACI. Proporciona 48 puertos de 1 GE SPF/10 GE SPF+ y 4 puertos de 40 GE QSFP. Cada uno de los 4 puertos de 40 GE admite los 4 modos introductorios de 10 GE para que funcionen como 4 puertos de 10 GE individuales. Como resultado, la tarjeta de línea puede proporcionar un máximo de 64 puertos de 10 GE. La flexibilidad de la velocidad de puertos permite un diseño de agregación y un acceso a la red de gran sencillez y rentabilidad.

Entre los componentes clave de estas tarjetas de línea se incluyen dos NFE, dos ALE y una CPU de tarjeta de línea. Los dos NFE proporcionan los puertos del panel frontal. Un NFE tiene 48 puertos de 1/10 G y el otro tiene 4 puertos de 40 G. Los dos ALE proporcionan un espacio de búfer ampliado, gestión de paquetes adicional y la opción de usar la tarjeta de línea en modo ACI.

Para ofrecer el tipo de puerto y flexibilidad en cuanto a velocidad, los puertos del panel frontal de esta tarjeta de línea pueden funcionar a distintas velocidades. El desajuste en la velocidad de los puertos es uno de los principales motivos de la congestión de los puertos y el almacenamiento en búfer de paquetes. Por consiguiente, puede que esta tarjeta de línea requiera más espacio en el búfer del que pueden proporcionar sus NFE. Los dos ALE proporcionan hasta 40 MB de almacenamiento en búfer adicional cada uno. Dado que el ALE se sitúa entre los NFE y los módulos del fabric, puede almacenar en búfer el tráfico de tránsito entre ellos. El tráfico conmutado localmente de un puerto de 10G a uno de 1G en el mismo NFE también se puede redirigir al ALE ubicado en la interfaz norte para aprovechar el espacio de búfer ampliado.

Al igual que N9K-X9636PQ, esta tarjeta de línea también se puede beneficiar de un diseño sin PHY que favorece un menor consumo de electricidad, una menor latencia y una mayor fiabilidad.

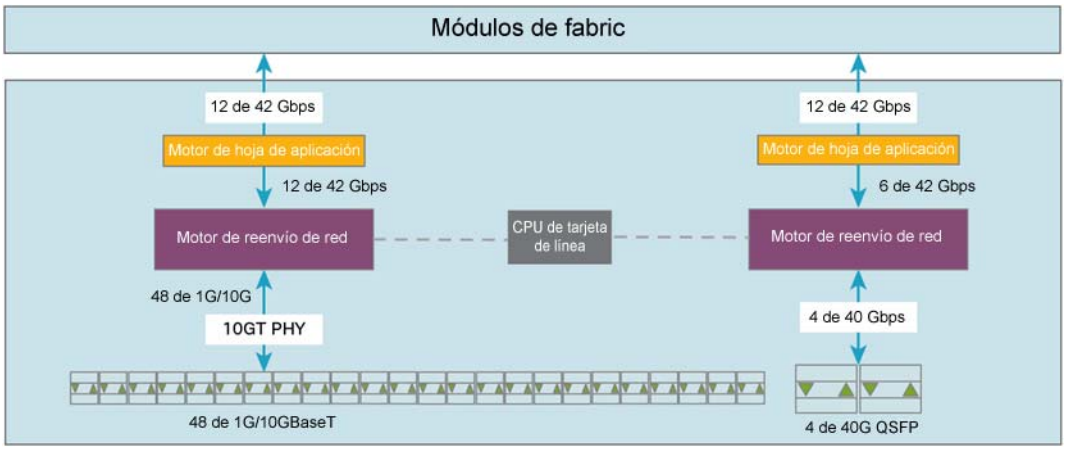
Figura 9. Tarjeta de línea Nexus serie 9500 de 48 puertos de 1/10 GE SPF+ y 4 puertos de 40 GE QSFP



Tarjeta de línea de 48 puertos 1/10 G BastT (N9K-X9564TX)

N9K-X9564TX (figura 10) es otra tarjeta de línea de hoja preparada para ACI. Proporciona 48 puertos de 1G/10GBaseT y 4 puertos de 40G QSFP. Cuenta con una arquitectura similar a N9K-X9564PX excepto en que los 48 puertos de 1G/10GBasteT se implementan con PHY de 10GT para convertirse en medios físicos de 1G/10GBaseT.

Figura 10. Tarjeta de línea Nexus serie 9500 de 48 puertos de 1/10 GBaseT y 4 puertos de 40 GE QSFP



Reenvío de paquetes unidifusión de Nexus serie 9500

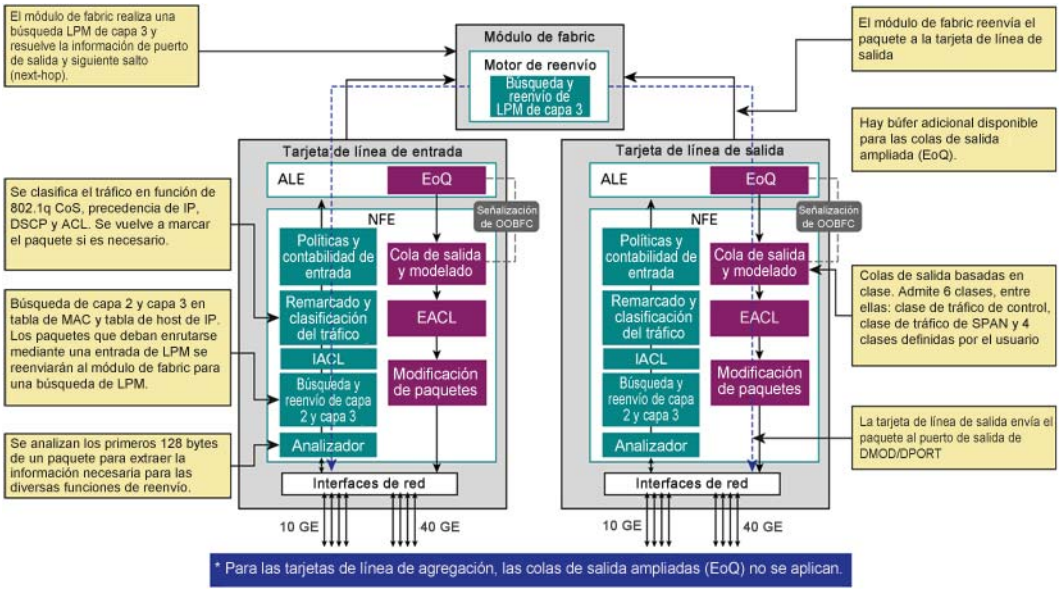
Como se ha mencionado anteriormente, tanto las tarjetas de línea como los módulos de fabric de los switches Nexus serie 9500 incluyen NFE que realizan funciones de reenvío y búsqueda de paquetes. Cada NFE tiene recursos de tabla de reenvío, entre los que se incluyen tablas TCAM y una tabla de hash programable conocida como tabla de reenvío unificada (UFT). Se puede asignar de forma flexible para entradas MAC de capa 2, entradas host de IP o entradas LPM. Esta flexibilidad, junto con la arquitectura de reenvío de datos completamente distribuida permite que los switches Nexus de Cisco serie 9500 optimicen la utilización de recursos de tabla en las tarjetas de línea y los módulos de fabric para maximizar la escalabilidad de reenvío de capa 2 y 3 del sistema. Además, ofrece la capacidad de implementar los switches Nexus 9500 en una amplia gama de tamaños de Data Center con diversos tipos de aplicaciones.

	Tarjeta de línea	Módulo de fabric
Tabla L2 MAC	160K	-
Tabla de host de capa 3	88K	-
Tabla LPM	-	128K

La arquitectura de reenvío de plano de datos de los switches Nexus de Cisco serie 9500 incluye un operador de comparación de entrada en el NFE de entrada, el reenvío de módulos de fabric y el operador de comparación de salida del NFE de salida. Los operadores de comparación de entrada y de salida se pueden ejecutar en la misma tarjeta de línea o incluso en el mismo NFE, si los puertos de entrada y de salida están en el mismo NFE.

Un NFE se compone de un operador de comparación de procesamiento de entrada, un gestor de búfer para poner en cola y programar, y un operador de comparación de procesamiento de salida. El operador de comparación de entrada realiza un análisis de encabezado de paquete, terminación de túnel, detección de VRF, búsqueda de capa 2 y 3 basada en la información del encabezado de paquete analizado y procesamiento de ACL de entrada. El gestor de búfer es el responsable de todas las funciones de poner en cola y programar. El operador de comparación de salida gestiona todos los ACL de salida y modificación de paquetes. Todas las búsquedas como las tablas de capa 2, 3 o ACL se realizan en el operador de comparación de entrada. Tanto los operadores de comparación de entrada como los de salida tienen varias fases para permitir un procesamiento paralelo de los paquetes.

Figura 11. Reenvío de paquetes unidifusión de Nexus 9500



1. Operador de comparación de procesamiento de entrada

Análisis del encabezado de paquete

Cuando un paquete entra por un puerto del panel frontal, pasa por el operador de comparación de entrada del motor de reenvío de red de la tarjeta de línea. El primer paso es el análisis del encabezado de paquete. El analizador flexible analiza los primeros 128 bytes del paquete para extraer y guardar información como el encabezado de capa 2, Ethertype, encabezado de capa 3 y los protocolos TCP IP. Esto se usa para la lógica de procesamiento y la búsqueda de paquetes subsiguientes.

Búsqueda de MAC de capa 2 y de host de capa 3

A medida que el paquete pasa por el operador de comparación de entrada, se le realizan búsquedas de switching de capa 2 y de routing de capa 3. Primero, el NFE examina la dirección de destino MAC (DMAC) del paquete para determinar si este necesita switching de capa 2 o routing de capa 3. Si el DMAC coincide con la dirección MAC del router del switch, el paquete pasa a la lógica de búsqueda de routing de capa 3. Si el DMAC no pertenece al switch, se lleva a cabo una búsqueda de switching de capa 2 basada en la ID de VLAN y DMAC. Si se encuentra una coincidencia en la tabla de direcciones MAC, el paquete se envía hacia el puerto de salida. Si no se encuentra ninguna coincidencia para la combinación de DMAC y VLAN, el paquete se reenvía a todos los puertos del mismo VLAN.

Como parte de la lógica de switching de capa 2, el NFE también realiza una búsqueda de dirección MAC de origen (SMAC) para lograr un aprendizaje basado en hardware. El SMAC junto con la ID de VLAN se usan para buscar en la tabla de direcciones MAC. Si no hay ninguna coincidencia, esta nueva dirección se aprende y se asocia al puerto de entrada del paquete. Si se encuentra una coincidencia, no se realiza ninguna acción de aprendizaje. El NFE también admite el estado obsoleto asistido por hardware. Las entradas que no se usen durante un período de tiempo ampliado (tiempo que tarda en quedarse obsoleto configurable) se eliminarán automáticamente.

Dentro de la lógica de búsqueda de capa 3 del NFE de la tarjeta de línea, la dirección IP de destino (DIP) se utiliza para buscar en la tabla de host de capa 3. Esta tabla almacena las entradas de reenvío para hosts conectados directamente o rutas de host aprendidas de 32. Si la DIP coincide con una entrada de la tabla de host, la entrada indica el puerto de destino, la dirección MAC de siguiente salto (next-hop) y la VLAN de salida. Si no hay ninguna coincidencia con la DIP en la tabla de host, el paquete se reenviará al módulo de fabric en el que se realice la búsqueda de coincidencia de prefijo más extenso (LPM) en la tabla de routing de LPM.

Al realizar switching de capa 2 y routing de host de capa 3, si el puerto de salida es local del NFE, el NFE reenviará los paquetes de forma local sin enviarlos a los módulos de fabric. En el caso de la tarjeta de línea de hoja preparada para ACI, si el puerto de entrada tiene una mayor velocidad que el puerto de salida, los paquetes se redirigirán al motor de hoja de aplicación (ALE) para el almacenamiento en búfer adicional con el fin de compensar el desajuste de las velocidades de los puertos.

Procesamiento de ACL de entrada

Además de las búsquedas de reenvío, se somete al paquete a un procesamiento de ACL de entrada. Se comprueba el TCAM de ACL en busca de coincidencias de ACL de entrada. Cada NFE tiene una tabla de TCAM de ACL de entrada compuesta por entradas de 4K para admitir ACL internos del sistema y ACL de entrada definidos por el sistema. Entre estos ACL se incluyen ACL de puertos, ACL enrutados y ACL de VLAN. Las entradas de ACL se ubican en el NFE y solo se programan cuando son necesarias. Esto permite una utilización máxima del TCAM de ACL en un switch Nexus 9500.

Clasificación del tráfico de entrada

Los switches Nexus serie 9500 admiten la clasificación del tráfico de entrada. En una interfaz de entrada, el tráfico se puede clasificar en función de los campos de dirección, 802.1q CoS y precedencia de la IP o DSCP en el encabezado del paquete. El tráfico clasificado se puede asignar a uno de los cuatro grupos de calidad del servicio. Los grupos de calidad del servicio funcionan como una identificación interna en las clases de tráfico que se usan para los procesos subsiguientes de QoS a medida que los paquetes pasan a través del sistema.

Admisión, colas y políticas de entrada

El gestor de búfer realiza las funciones de contabilidad y admisión de entrada del tráfico en el operador de comparación de procesamiento de entrada. Cada NFE incluye un búfer de 12 MB formado por 60 000 celdas de 208 bytes. Este recurso de búfer lo comparten de forma dinámica el tráfico de entrada y el de salida. El mecanismo de control de admisión de entrada decide si un paquete se debe admitir en la memoria. Esta decisión depende de la cantidad de memoria de búfer disponible y de la cantidad de búfer que ya usen el puerto de entrada y la clase de tráfico.

Los switches Nexus serie 9500 admiten las políticas de entrada basadas en clase. Las políticas se pueden definir usando un mecanismo de una velocidad y dos colores o un mecanismo de dos velocidades y tres colores.

2. Búsqueda de LPM de módulo de fabric

Cuando se reenvía un paquete al módulo de fabric, este realizará distintas acciones en función de los resultados de la búsqueda en la tarjeta de línea de entrada. En los casos en que el paquete sea conmutado de capa 2 o enrutado de host de capa 3, la tarjeta de línea de entrada habrá resuelto la información del puerto de salida, la MAC de siguiente salto (next-hop) y la VLAN de salida. El módulo de fabric simplemente reenviará el paquete a la tarjeta de línea de salida. Si el paquete necesita una búsqueda de LPM, el módulo de fabric buscará en la tabla de LPM y usará la mejor coincidencia para la dirección IP de destino (DIP) para reenviar el paquete. Si no hay ninguna coincidencia para la DIP, el paquete se descartará. La tabla de reenvío unificada (UFT) del motor de reenvío de red del módulo de fabric cuenta con un tamaño de LPM para 128 000 entradas.

3. Operador de comparación de procesamiento de salida

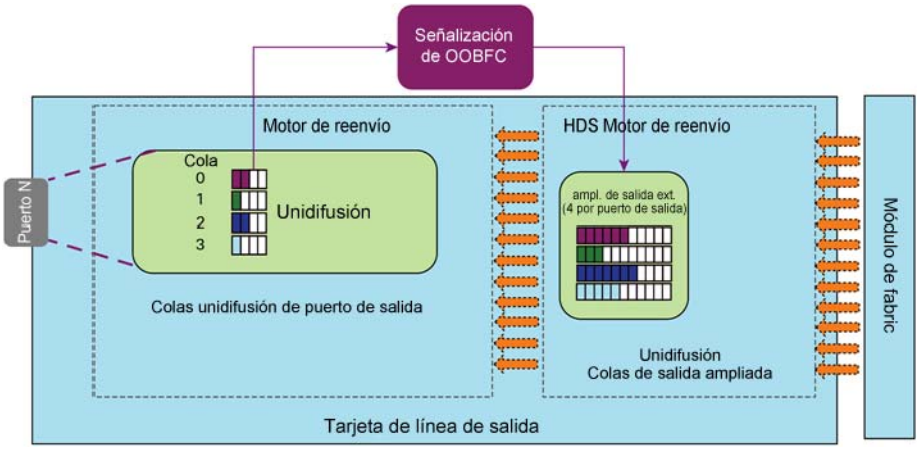
El operador de comparación de procesamiento de salida es relativamente sencillo, debido a que la mayoría de las búsquedas ya se ha realizado y la mayoría de las decisiones ya se ha tomado en el operador de comparación de entrada. Sin embargo, una importante función que se realiza en el operador de comparación de salida es el QoS de salida, que incluye la WRED/ECN, colas de salida y modelado.

Programar y poner en cola de salida

Los switches Nexus serie 9500, que siguen el principio de diseño de simplicidad y eficiencia, usan una sencilla arquitectura de colas de salida. En el caso de congestión en los puertos de salida, los paquetes se ponen en cola directamente en el búfer de la tarjeta de línea de salida. No hay colas de salida virtual (VoQ) en las tarjetas de línea de entrada. Esto simplifica enormemente la gestión del búfer del sistema y la implementación de las colas. Un switch Nexus 9500 puede admitir hasta seis clases de tráfico en la salida (cuatro definidas por el usuario que identifican las ID de grupo de calidad del servicio, una clase de tráfico de control de CPU y una clase de tráfico de SPAN). Cada clase definida por el usuario tiene una cola unidifusión y una cola multidifusión por puerto de salida. Los puertos locales de un NFE comparten el búfer de 12 MB. El software del switch incluye un mecanismo para medir y limitar el uso del búfer por puerto de salida. Esto garantiza que ningún puerto individual pueda consumir más que la parte que le corresponde de memoria del búfer y provocar escasez de búfer en los demás puertos.

Las tarjetas de línea preparadas para ACI disponen de un búfer adicional de 40 MB en cada uno de sus motores de hoja de ACI (ALE). 10 MB del búfer están asignados al tráfico asignado al fabric. Los 30 MB restantes están asignados al tráfico de salida de los módulos de fabric y el tráfico conmutado localmente del puerto de entrada de mayor velocidad al puerto de salida de menor velocidad. Este espacio de búfer de 30 MB se usa para las colas de salida ampliadas para el tráfico de unidifusión. El NFE comunica el estado de la cola de unidifusión al ALE mediante un canal de señalización de control de flujo fuera de banda (OOBFC). Cuando una cola de salida supera el umbral configurado, el NFE envía una señal de OOBFC para ordenar al ALE que deje de reenviar tráfico para esta cola y comience a poner en cola los paquetes en su propio búfer. Al recibir esta señal, el ALE comienza a formar la cola de salida ampliada para esta clase de tráfico en el puerto de salida proporcionado. Cuando la longitud de la cola de salida se reduce al umbral de reinicio configurado, el NFE envía otra señal de OOBFC para ordenar al ALE que reanude la transmisión del tráfico para esta cola concreta.

Figura 12. Cola de salida ampliada (EoQ) de Nexus 9500



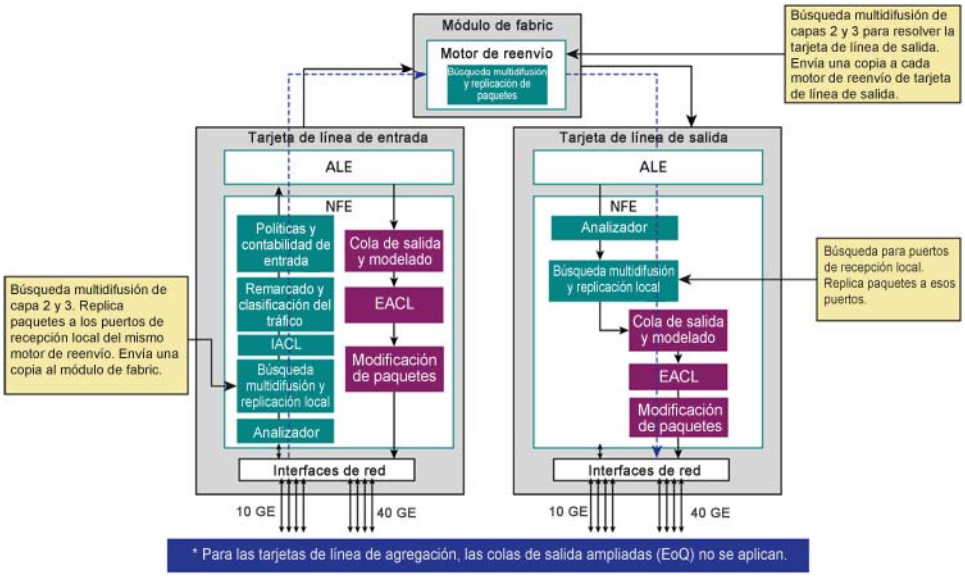
La arquitectura de cola de salida con colas de salida ampliadas, aunque sencilla, es un planteamiento muy eficaz para abordar la congestión de puertos con equidad. Garantiza que ningún puerto podrá provocar la escasez de memoria del búfer en ningún otro puerto.

Reenvío de paquetes multidifusión de Nexus serie 9500

Los paquetes multidifusión pasan a través de los mismos operadores de comparación de procesamiento de entrada y salida que los paquetes unidifusión. Sin embargo, una diferencia en el proceso de búsqueda y reenvío de paquetes es que los switches Nexus 9500 realizan una replicación y una búsqueda de multidifusión distribuida de tres fases. La tabla de routing multidifusión se almacena en todas las tarjetas de línea y los módulos de fabric. La NFE de entrada realiza la primera búsqueda para resolver los receptores locales. Si hay algún receptor local, el NFE crea una copia por puerto de recepción local. Además, el NFE de entrada envía una copia del paquete entrante al módulo de fabric. Al recibir el paquete, el módulo de fabric realiza la segunda búsqueda para encontrar las tarjetas de línea de salida. El módulo de fabric replica el paquete a cada NFE de salida.

El NFE de salida realiza la tercera búsqueda para resolver sus receptores locales y replica el paquete a esos puertos. Esta búsqueda y replicación multidifusión de varias fases es la forma más eficiente de replicar y reenviar el tráfico multidifusión.

Figura 13. Reenvío de paquetes multidifusión de Nexus 9500



Otra diferencia entre el reenvío de tráfico multidifusión y unidifusión es que el tráfico de multidifusión no tiene colas de salida ampliada. El motor de reenvío de red admite cuatro colas de multidifusión por puerto de salida. En presencia de los motores de hoja de ACI, pone en cola el tráfico de multidifusión independientemente de las colas de multidifusión del motor de reenvío de red. No hay señal de contrapresión para controlar las colas de multidifusión a través del canal de OOBFC.

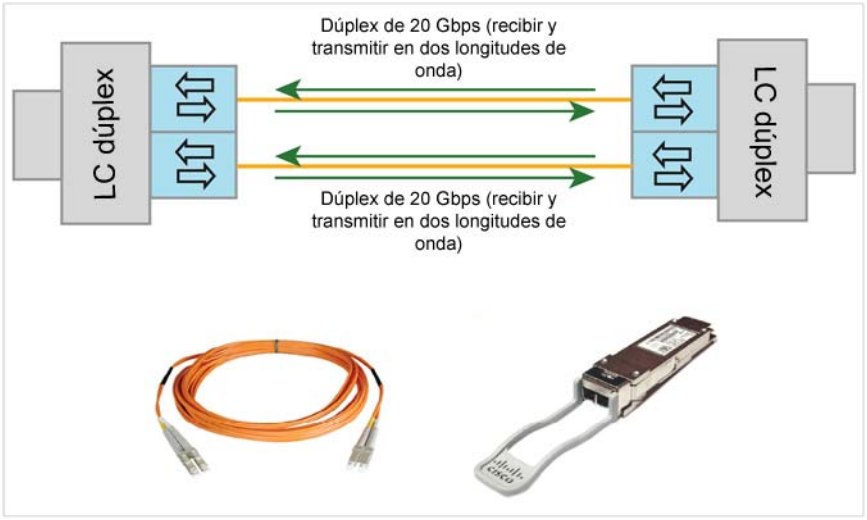
Tecnología Cisco QSFP BiDi para la migración a 40 Gbps

Los switches Nexus serie 9500, con una densidad de puertos y un rendimiento altos para la conectividad de 1/10/40 GE, son perfectos para la última generación de infraestructura de Data Center. Al mismo tiempo que ofrecen 1/10GE para acceso/hoja y enlaces de 40GE en la agregación/columna, ofrecen un ancho de banda más escalable para aplicaciones de Data Center.

Sin embargo, la migración de una red de Data Center existente de 10 GE a 40 GE implica la actualización de más de una plataforma de red. La migración de la infraestructura de cableado es uno de los mayores retos que entraña esta tarea. La infraestructura de cableado actual 10 GE usa 2 cables de fibra MMF para una conexión de 10 GE. Sin embargo, los transceptores de cables ópticos de 40 GE existentes de corto alcance, de SR4 o CSR4, tienen secciones de transmisor y receptor independientes, cada una con 4 hilos de fibra en paralelo. Como resultado, se requieren 8 hilos de fibra para una conexión de 40 GE dúplex. Estas diferencias imponen que el cambio de la infraestructura de 10 GE actual a 40 GE con los transceptores ópticos de 40 GE existentes requieran una reconstrucción o actualización de la infraestructura del cableado a gran escala. El impresionante coste y la potencial interrupción del servicio hacen que resulte muy difícil migrar un Data Center de producción existente a una infraestructura de 40 GE.

La tecnología del transceptor Cisco QSFP BiDi resuelve este problema, ya que proporciona la capacidad de transmitir 40G dúplex completo a dos hilos de fibra MMF con conectores LC. En otras palabras, el transceptor QSFP BiDi permite que la conectividad de 40 GE reutilice las fibras de 10G y la línea troncal de fibra existentes sin la necesidad de ampliación ni reconstrucción. Además, elimina las barreras que suponen los costes del cableado de 40 Gbps para realizar la migración de una conectividad de 10 Gbps a 40 Gbps en redes de Data Center.

Figura 14. Tecnología del transceptor bidireccional de Cisco



Conclusión

Los switches Nexus serie 9500 son los switches de Data Center líderes del sector y ofrecen la mayor densidad de puertos para la conectividad de 1, 10, 40 y, en el futuro, 100 GE, además de una velocidad de línea real sin precedentes y un rendimiento de reenvío de baja latencia. Los switches Nexus serie 9500 admiten la densidad de puertos de 10 GE y 40 GE líderes del sector. Gracias a su diseño de chasis y velocidades de puerto flexibles, los switches Nexus serie 9500 se adaptan a las implementaciones de Data Centers virtualizados, de varios clientes y en la nube a todos los niveles: pequeño, mediano o grande.

El diseño de chasis sin plano medio permite la máxima eficiencia de la refrigeración. La combinación de la tecnología Merchant Silicon y Custom Silicon permite que las tarjetas de línea tengan el menor número de ASIC, a la vez que ofrece un rendimiento récord. Con innovaciones como el flujo de aire de la parte frontal a la trasera y las fuentes de alimentación de gran eficiencia con la certificación 80 PLUS Platinum, los switches Nexus serie 9500 suponen un nuevo récord en el ámbito de la eficiencia energética, la fiabilidad y el rendimiento de los switches de Data Center.

Al separar la gestión interna del sistema del plano de control del switch, los switches Nexus serie 9500 alcanzan una estabilidad del plano de control sin precedentes. Los switches Nexus serie 9500, equipados con motor supervisor construido con la CPU de varios núcleos más vanguardista y con CPU de tarjeta de línea para descargar tareas de los motores supervisores, sientan las bases para crear un switch de Data Center totalmente fiable.

Cuando funcionan en el modo NX-OS clásico, los switches Nexus serie 9500 se ejecutan en una sola imagen para todos los switches de la familia, lo que simplifica enormemente la administración de la red. Cuando se ejecutan en el último kernel de Linux de 64 bits con modularidad de procesos real, alta resistencia de software y varias mejoras en la automatización y programabilidad, el NX-OS mejorado para los switches Nexus serie 9500 se convierte en la mejor solución para los Data Centers que buscan modernizarse y automatizar los modelos de funcionamiento y gestión de redes de los Data Center.

Gracias a las características únicas que se han mencionado anteriormente, los switches Nexus de Cisco serie 9500 son los switches de Data Center perfectos para permitir a las organizaciones construir Data Centers fiables, escalables, resistentes y automatizados.

Apéndice

Apéndice A: terminología

ACI: infraestructura centrada en aplicaciones (Application Centric Infrastructure)

NFE: motor de reenvío de red (Network Forwarding Engine)

ALE: motor de hoja de ACI (ACI Leaf Engine)

EoQ: cola de salida ampliada (Extended Output Queue)

OOBFC: control de flujo fuera de banda (Out-of-Band Flow Control)

Sede central en América
Cisco Systems, Inc.
San José. CA

Sede Central en Asia Pacífico
Cisco Systems (EE. UU.) Pte. Ltd.
Singapur

Sede Central en Europa
Cisco Systems International BV Amsterdam.
Países Bajos

Cisco cuenta con más de 200 oficinas en todo el mundo. Las direcciones, los números de teléfono y de fax están disponibles en el sitio web de Cisco: www.cisco.com/go/offices.

Impreso en EE. UU. C11-729987-00 11/13