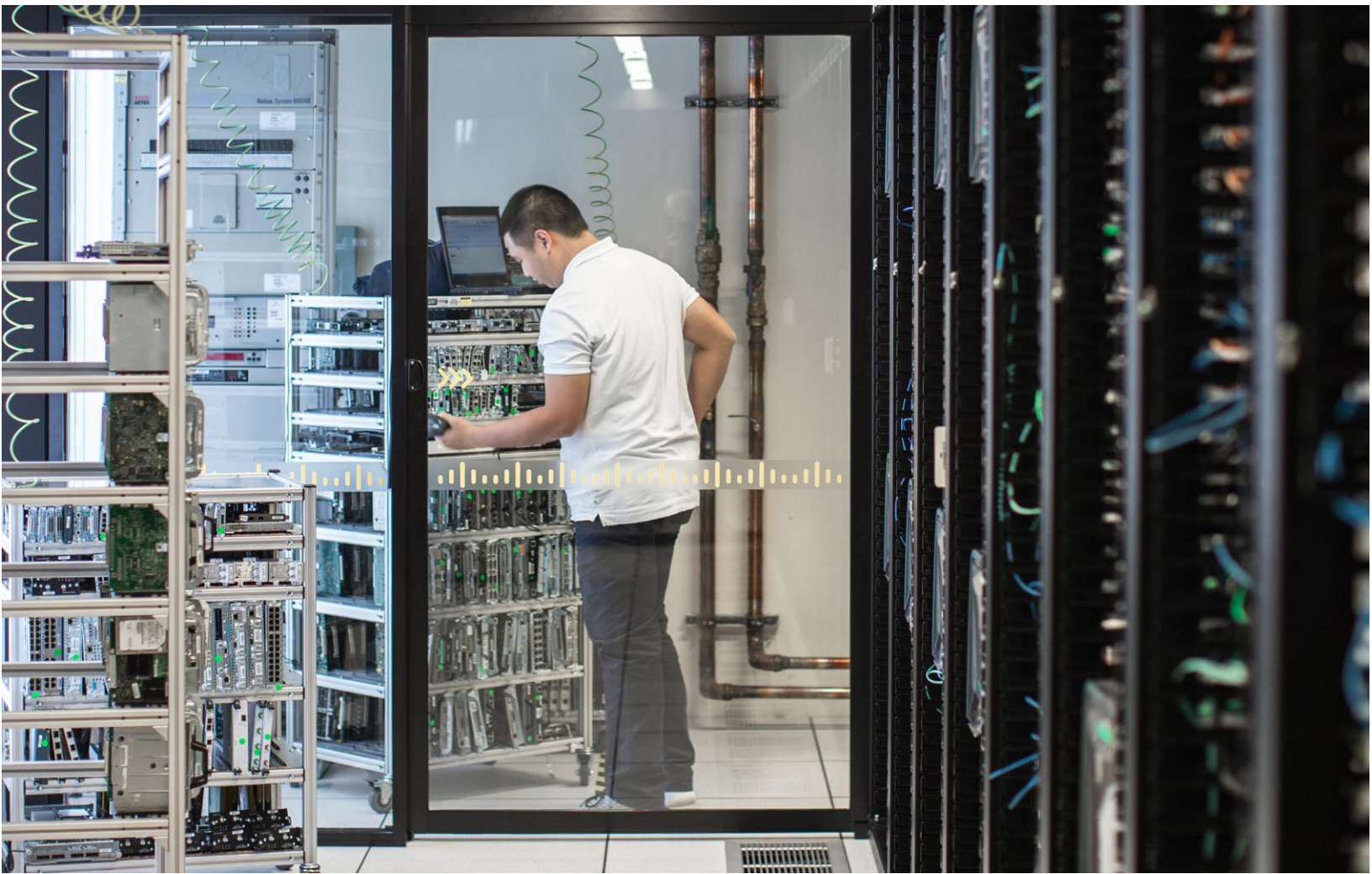
Whitepaper
Cisco - openbaar

Architectuur van Cisco Nexus 9300-EX-platformswitches



Inhoud

Inleiding

In 2016 begon de verschuiving in datacenterswitching naar nieuwe capaciteit en nieuwe mogelijkheden met de introductie van 25, 50 en 100 Gigabit Ethernet-connectiviteit. Deze nieuwe Ethernet-connectiviteit is een aanvulling op de vorige 10 en 40 Gigabit Ethernet-standaarden, met vergelijkbare kosten en energiezuinigheid, en vertegenwoordigt een toename in de capaciteit van grofweg 250%.

Cisco brengt een aantal nieuwe switchingproducten voor het datacenter uit om onze klanten te helpen beter presterende en meer kosteneffectieve datacenternetwerken te bouwen om grotere toepassingswerklasten en verschillende typen connectiviteit onder te brengen. De nieuwe switches van Cisco® ondersteunen zowel bestaande als nieuwe standaard Ethernet-snelheden, waaronder 1, 10 en 40 Gbps en 25, 50 en 100 Gbps, en zijn daarmee geschikt voor de netwerkinfrastructuur van bestaande en next-generation datacenters.

In dit document komt de hardwarearchitectuur aan bod van de nieuwe switchplatforms in de productreeks van Cisco Nexus® 9000 Series-switches, en dan name de Cisco Nexus 9300-EX-platformswitches. Deze switches vormen de volgende generatie van vaste Cisco Nexus 9000 Series-switches. Het nieuwe platform is gebaseerd op de Cisco Cloud Scale ASIC en ondersteunt kosteneffectieve implementaties op cloudschaal, meer endpoints en cloudservices met wire-rate beveiliging en telemetrie. Het platform is gebouwd op moderne systeemarchitectuur die is ontworpen voor hoge prestaties om aan de veranderende behoeften van zeer schaalbare datacenters en groeiende ondernemingen te voldoen. Cisco Nexus 9300-EX-platformswitches bieden diverse interfaceopties om bestaande datacenters transparant te migreren van snelheden van 100 Mbps, 1 Gbps, en 10 Gbps naar 25 Gbps op de server, en van 10 en 40 Gbps naar 50 en 100 Gbps in de aggregatielaag.

Het platform kan uitgebreide telemetriegegevens van Cisco Tetration Analytics™ verzamelen op lijnsnelheid over alle poorten zonder enige latentie aan de pakketten toe te voegen of switchprestaties negatief te beïnvloeden. Deze telemetriegegevens worden standaard elke 100 milliseconden geëxporteerd, direct vanaf de ASIC (Application-Specific Integrated Circuit) van de switch. Deze gegevens bestaan uit drie typen informatie:

- Stroominformatie: dit omvat informatie over endpoints, protocollen, poorten, wanneer de stroom begon, hoe lang de stroom actief was, enzovoort.
- Variaties tussen pakketten: dit betreft informatie over variaties tussen pakketten binnen de stroom. Voorbeelden hiervan zijn variaties in de TTL- (Time To Live), IP- en TCP-vlaggen en de payloadlengte.
- Contextdetails: dit betreft contextinformatie die afkomstig is van buiten de pakketheader. Dit omvat variaties in de benutting van de buffer, verloren gegane pakketten binnen een stroom, koppeling met tunnel-endpoints, enzovoort.

Het Cisco Tetration Analytics-platform gebruikt deze telemetriegegevens en past onbeheerde machine learning en gedragsanalyse toe om in real time uitstekende, diepgaande zichtbaarheid van alles binnen uw datacenter bieden. Cisco Tetration Analytics kan via algoritmische benaderingen diepgaand inzicht bieden in toepassingen en interacties, waardoor sterk vereenvoudigde bedrijfsprocessen, een zero-trust model en migratie van toepassingen naar elke programmeerbare infrastructuur mogelijk worden. Ga voor meer informatie naar https://www.cisco.com/go/tetration.

Cisco biedt twee gebruiksmodi voor Cisco Nexus 9000 Series-switches. Organisatie kunnen Cisco NX-OS-software gebruiken om de switches in standaardomgevingen voor Cisco Nexus-switches te implementeren (NX-OS-modus). Daarnaast kunnen organisaties kunnen een hardware-infrastructuur gebruiken die het Cisco Application Centric Infrastructure-platform (Cisco ACI™) kan ondersteunen om optimaal te profiteren van een geautomatiseerde, op beleid gebaseerde benadering van systeembeheer (ACI-modus).

Overzicht van het Cisco Nexus 9300-EX-platform

Het Cisco Nexus 9300-EX-platform bestaat uit switches met een vaste configuratie op basis van Cisco's nieuwe Cloud Scale ASIC.

De eerste lichting van het Cisco Nexus 9300-EX-platform omvat de volgende switchmodellen: Cisco Nexus 93180YC-EX, 93108TC-EX en 93180LC-EX (afbeelding 1). Tabel 1 bevat een overzicht van de modellen van het Cisco Nexus 9300-EX-platform.

Afbeelding 1. Cisco Nexus 9300-EX-platformswitches

In navolging van de naamgevingsconventies voor de Cisco Nexus 9000 Series geven de letters in de productnamen voor het Cisco Nexus 9300-EX-platform de ondersteunde poortsnelheden of aanvullende hardwaremogelijkheden aan:

- Q: native 40 Gbps-poorten op voorpaneel
- Y: native 25 Gbps-poorten op voorpaneel
- C: native 100 Gbps-poorten op voorpaneel
- L: native 50 Gbps-poorten op voorpaneel
- T: 100M, 1GT en 10GT
- X (na het koppelteken): mogelijkheden voor Cisco NetFlow en gegevensanalyses

Tabel 1. Cisco Nexus 9300-EX-platformswitches (NX-OS-modus of leafswitches voor ACI-modus)

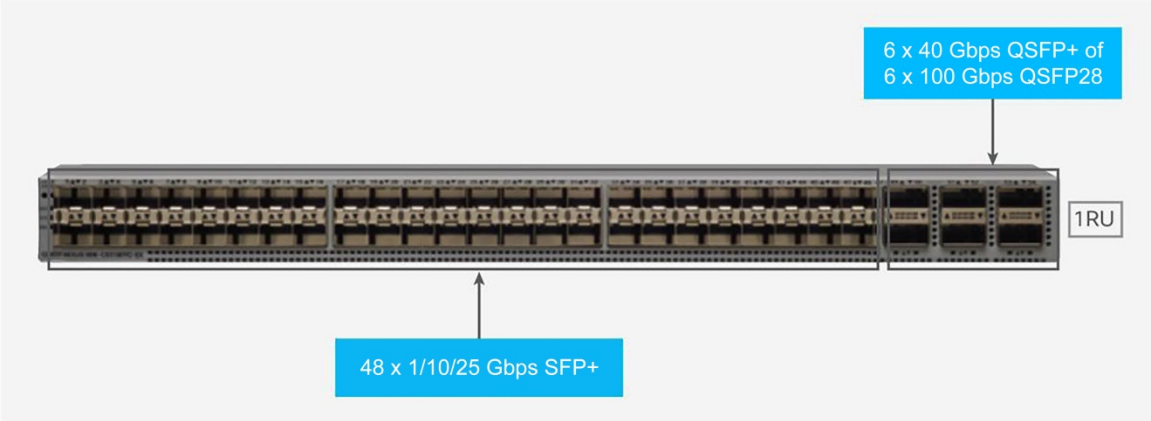
Model	Beschrijving	Cisco ASIC
Cisco Nexus 93180YC-EX	48 x 10/25 Gbps glasvezelpoorten en 6 x 40/100 Gbps QSFP28-poorten (Quad Small Form-Factor Pluggable)	Cloud Scale leaf-and-spine engine (LSE) ASIC
Cisco Nexus 93108TC-EX	48 x 10GBASE-T-poorten en 6 x 40/100 Gbps QSFP28-poorten	Cloud Scale LSE ASIC
Cisco Nexus 93180LC-EX	24 x 40/50 Gbps Enhanced QSFP-poorten (QSFP+) en 6 x 40/100 Gbps QSFP28-poorten	Cloud Scale LSE ASIC

De switches Cisco Nexus 93180YC-EX, 93108TC-EX en 93180LC-EX gebruiken allemaal dezelfde componenten, zoals CPU, systeemgeheugen en SSD-stations (Solid State Drive). Zie de gegevensbladen voor gedetailleerde informatie over de hardware-eigenschappen van Cisco Nexus 9300-EX-platformswitches.

Architectuur van de Cisco Nexus 93180YC-EX-switch

De Cisco Nexus 93180YC-EX-switch (afbeelding 2) is een 1RU-switch (1 rackunit) met een latentie van minder dan 1 microseconde die 3,6 Tbps (terabits per seconde) aan bandbreedte en meer dan 2,6 bpps (miljard pakketten per seconde) ondersteunt. De 48 downlink-poorten op de 93180YC-EX kunnen worden geconfigureerd als 1, 10, of 25 Gbps-poorten om implementatieflexibiliteit en investeringsbescherming te bieden. De uplink kan tot zes 40 en 100 Gbps-poorten ondersteunen, of een combinatie van 10, 25, 40, 50 en 100 Gbps-connectiviteit om flexibele migratieopties te bieden. Alle poorten zijn verbonden met de Cloud Scale LSE ASIC.

Afbeelding 2. Cisco Nexus 93180YC-EX-switch



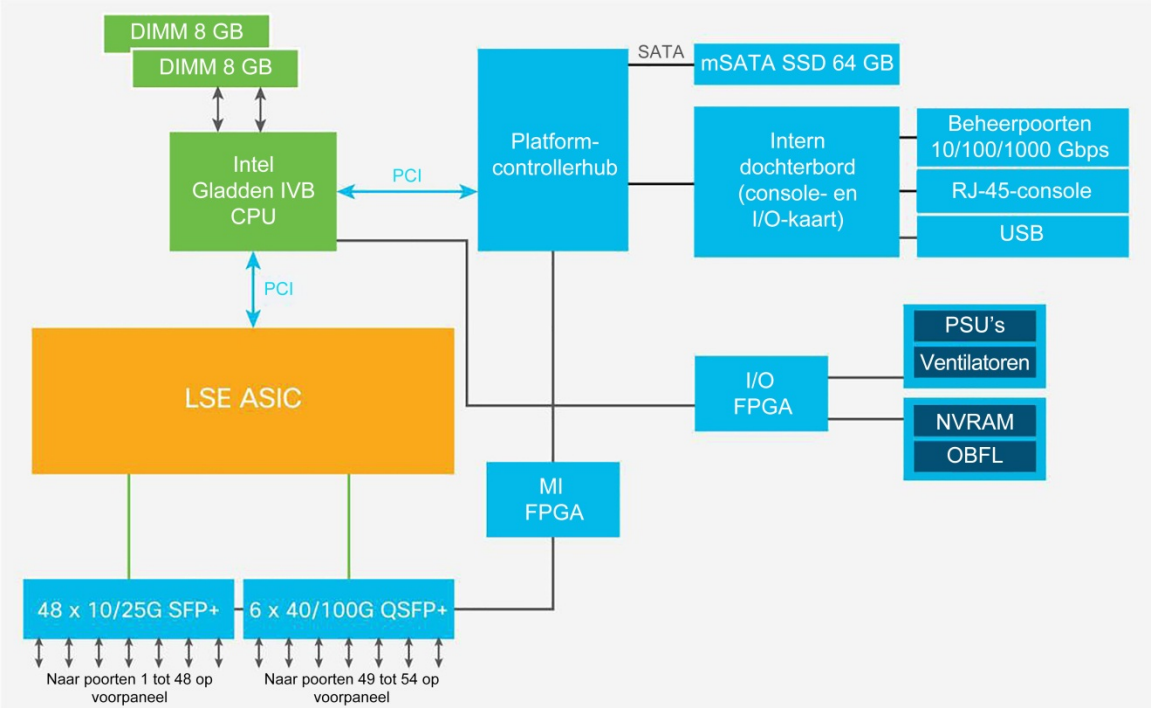
In afbeelding 3 wordt de hardwarearchitectuur van de Cisco Nexus 93180YC-EX getoond.

De 93180YC-EX is uitgerust met een Intel® Ivy Bridge Gladden-CPU met 4 cores en 24 GB aan systeemgeheugen. De CPU is via PCIe-verbindingen (PCI Express) verbonden met de controllerhub. De controllerhub biedt standaardinterfaces (SATA, USB, Ethernet, enzovoort) voor de opslag-, voedings-, koelings- en beheer-I/O-componenten. De 93180YC-EX is uitgerust met een 64 GB mSATA SSD-drive.

De console en het I/O-dochterbord omvatten een RG-45 seriële consolepoortverbinding en dual-media Ethernet-beheerpoorten die 10/100/1000BASE-T of 1 Gbps SFP (voor glasvezelverbindingen) ondersteunen. Slechts een van de twee beheerpoorten kan op enig moment actief zijn. De switch zal automatisch de poort met een actieve koppelingsstatus selecteren. Als beide koppelingen verbonden zijn, krijgt de BASE-T-interface prioriteit. De console en I/O-kaart bevatten een USB 2.0-poort.

De dataplane-doorstuurcomponenten op de 93180YC-EX omvatten een enkele multi-slice LSE ASIC. De LSE ASIC heeft directe verbindingen naar 48 poorten op het voorpaneel in 1/10/25 Gbps-modus: poorten 1 tot en met 48. Verder heeft de ASIC directe verbindingen naar 6 uplinkpoorten in 40/100 Gbps-modus: poorten 49 tot en met 54.

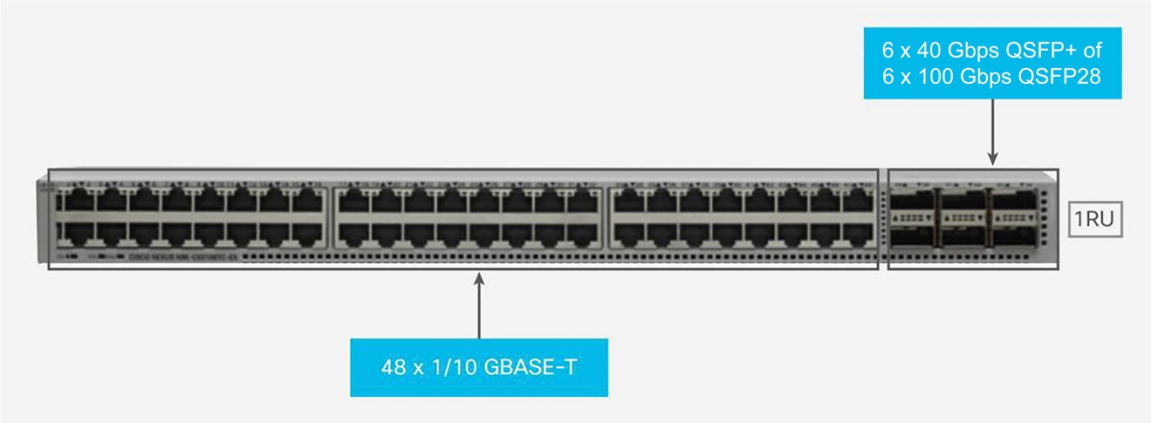
Afbeelding 3. Hardwarearchitectuur van de Cisco Nexus 93180YC-switch



Architectuur van de Cisco Nexus 93108TC-EX-switch

De Cisco Nexus 93108TC-EX-switch (afbeelding 4) is een 1RU-switch die 2,16 Tbps (terabits per seconde) aan bandbreedte en meer dan 1,5 bpps ondersteunt. De achtenveertig 10GBASE-T downlink-poorten op de 93108TC-EX kunnen worden geconfigureerd om als 100 Mbps-, 1 Gbps- of 10 Gbps-poorten te werken. De uplink kan tot zes 40 en 100 Gbps-poorten ondersteunen, of een combinatie van 10, 25, 40, 50 en 100 Gbps-connectiviteit om flexibele migratieopties te bieden.

Afbeelding 4. Cisco Nexus 93108TC-EX-switch



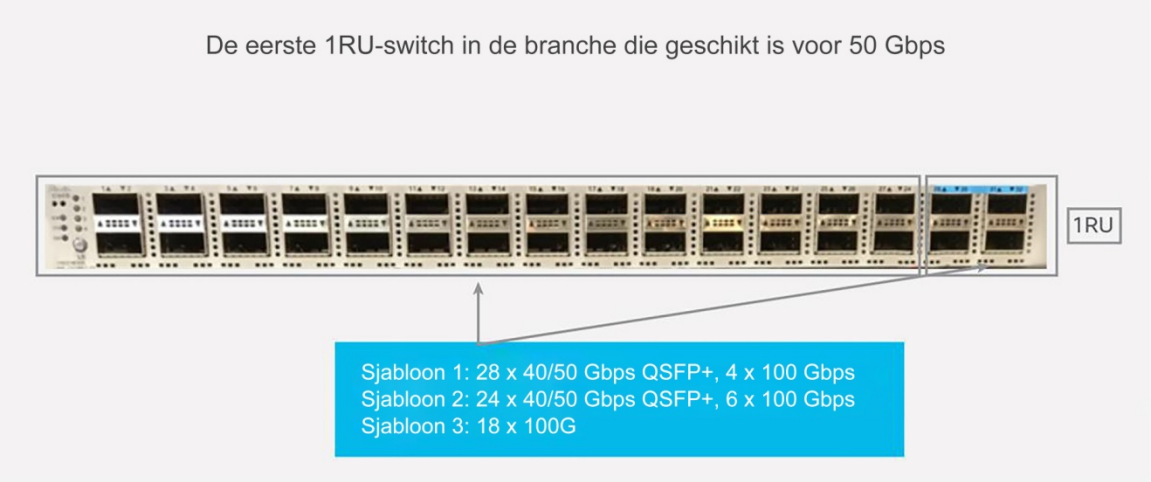
Behalve het verschil in de voorpaneel-poortconfiguratie is de hardwarearchitectuur van de 93108TC-EX vergelijkbaar met die van de 93180YC-EX.

Architectuur van de Cisco Nexus 93180LC-EX-switch

De Cisco Nexus 93180LC-EX-switch is de eerste 1RU-switch in de branche die geschikt is voor 50 Gbps. De switch ondersteunt 3,6 Tbps aan bandbreedte en meer dan
2.8 2,8 bpps verdeeld over maximaal 32 vaste poorten van 40 en 50 Gbps QSFP+ of maximaal 18 vaste poorten van 100 Gbps (afbeelding 5). Omdat 28 poorten van 40 Gbps via Gearbox zijn verbonden, biedt de switch flexibiliteit waarmee voor elke twee QSFP-connectoren een poort kan worden uitgeschakeld zodat de tweede in een andere modus kan werken met sjablonen, bijvoorbeeld een poortmodus met 18 x 100 Gbps, met 4 x 100 Gbps en 28 x 40 Gbps, of met 6 x 100 Gbps en 24 x 40 Gbps. De 100 Gbps-poort ondersteunt ook een breakoutmodus met 2 x 50 Gbps, 4 x 25 Gbps of 4 x 10 Gbps. Bekijk de softwarereleasenotities voor ondersteunde sjablonen.

Met uitzondering van het verschil in de configuratie van poorten op het voorpaneel is de hardwarearchitectuur van de 93180LC-EX vergelijkbaar met die van de 93180YC-EX.

Afbeelding 5. Cisco Nexus 93180LC-EX-switch



Cisco Cloud Scale ASIC's in Cisco Nexus 9300-EX-platform

De Cisco Nexus 9300-EX-platformswitches worden gebouwd met Cloud Scale ASIC LSE van Cisco. De Cloud Scale ASIC's worden vervaardigd met 16nm-technologie, terwijl ASIC-chips van derden worden vervaardigd met 28nm-technologie. Bij vervaardiging met 16 nm is er ruimte voor meer transistors in dezelfde maat matrijs als wordt gebruikt voor chips van derden. Hierdoor kon Cisco een enkele SoC-ASIC (Switch-on-a-Chip) vervaardigen die de volgende voordelen biedt:

- Hogere bandbreedte en grotere poortdichtheid tegen een lagere prijs: Cisco Nexus 9300-EX-switches bieden 10 en 25 Gbps-poorten met verschillende snelheden voor de prijs van 10 Gbps-poorten, en 40 en 100 Gbps-poorten voor de prijs van 40 Gbps-poorten. De switches bieden ook hogere bandbreedte en grotere poortdichtheid per rackunit met lagere kosten per poort.
- Grotere buffer: Cisco Nexus 9300-EX-switches hebben grotere interne buffers (40 MB in plaats van 16 MB) plus verschillende verbeterde wachtrij- en verkeersbeheerfuncties die de meeste switches met chips van derden niet bieden.
- Grotere tabel: Cisco Nexus 9300-EX-switches ondersteunen maximaal 256.000 MAC-adresvermeldingen in hardware en 256.000 IP-hostvermeldingen – veel meer dan bij switches met chips van derden.
- Diepgaande zichtbaarheid en telemetriegegevens: de Cloud Scale ASIC's van Cisco maken zichtbaarheid mogelijk van elk pakket en elke stroom op lijnsnelheid, zonder negatieve gevolgen voor de CPU van Cisco Nexus 9300-EX-switches.

Architectuur van de Cisco Cloud Scale LSE ASIC

Cisco biedt drie typen van de Cloud Scale ASIC's: Cisco ACI Spine Engine 2 (ASE2), ASE3 en LSE. De architectuur is vergelijkbaar, maar zij verschillen in poortdichtheid, buffervermogen, schaalbaarheid van doorsturen en sommige functies. De LSE ASIC is een superset van ASE2 en ASE3 en ondersteunt

leafswitch- en FEX-functies (Fabric Extender) van Cisco ACI. Net als de andere Cloud Scale ASIC's gebruikt de LSE een multi-slice SOC-ontwerp. De Cisco Nexus 9300-EX-platformswitches worden gebouwd met de LSE ASIC.

Elke ASIC heeft drie hoofdcomponenten:

- Slice-componenten: de slices vormen de switching-subsystemen. Deze omvatten multimodus MAC-adressen, pakketparser, controller voor doorstuurlookups, I/O-pakketbuffering, bufferaccounting, uitgangswachtrijen, planning en uitgangscomponenten voor herschrijven.
- I/O-componenten: de I/O-componenten bestaan uit SerDes-blokken (Serializer/Deserializer) met hoge snelheid. Deze variëren afhankelijk van het totale aantal poorten. Deze bepalen de totale bandbreedtecapaciteit van de ASIC's.
- Algemene componenten: deze bestaan uit de PCIe Gen 2-controller (Generation 2) voor registratie en EDMA-toegang (Enhanced Direct Memory Access) en een reeks point-to-multipoint draden om alle slices op elkaar aan te sluiten. Componenten omvatten ook de centrale statistiekmodules en modules voor het genereren van core- en MAC-adresklokken.

De LSE ASIC is samengesteld met twee slices, waarbij elke slice maximaal 800 Gbps bandbreedte ondersteunt voor in totaal 1,6 Tbps bij 1,25 GHz. Elke slice heeft 40 poorten, onafhankelijk van de snelheid of het type van de poorten. De beschikbare poortsnelheden zijn 1, 10, 25, 40, 50 en 100 Gbps.

Tabel 2 bevat een overzicht van de poortdichtheid in de LSE ASIC. De 10 Gbps-poorten ondersteunen tevens 1 Gbps. De LSE ASIC werkt bij zestien van de in totaal achttien 100 Gbps-poorten op lijnsnelheid voor alle pakketgrootten. Als de minimale pakketgrootte groter is dan 72 bytes, kan de ASIC op alle achttien 100 Gbps-poorten op lijnsnelheid werken.

Tabel 2. Poorteigenschappen van de LSE ASIC

ASIC	1 en 10 Gigabit Ethernet-poorten	25 Gigabit Ethernet-poorten	40 Gigabit Ethernet-poorten	100 Gigabit Ethernet-poorten
LSE	80	72	36	18

LSE-doorstuurtabel

LSE ASIC's gebruiken een gedeelde hashtabel die de UFT (Unified Forwarding Table) wordt genoemd, om Layer 2- en Layer 3-doorstuurgegevens op te slaan. De grootte van de UFT op LSE ASIC's is 544.000 vermeldingen. De UFT is gepartitioneerd in diverse regio's ter ondersteuning van MAC-adressen, IP-hostadressen, LPM-vermeldingen (Longest-Prefix Match) voor IP-adressen en multicast lookups. De UFT wordt ook gebruikt voor next-hop en aangrenzingsgegevens en RPF-controlevermeldingen (Reverse Path Forwarding) voor multicast verkeer.

De UFT bestaat intern uit meerdere tiles. Elke tile kan onafhankelijk worden geprogrammeerd voor een bepaalde doorstuurtabelfunctie. Dankzij dit programmeerbaar delen van geheugen wordt flexibiliteit geboden voor diverse implementatiescenario's en wordt de efficiëntie van de benutting van geheugenresources vergroot.

Naast de UFT hebben de ASIC's een TCAM (Ternary Content-Addressable Memory) voor 12.000 vermeldingen dat kan worden gebruikt voor doorstuurlookupgegevens.

Dankzij het programmeerbare gedeelde hashtabelgeheugen kan de inrichting van de doorstuurtabel voor verschillende doorstuurfuncties op het Cisco Nexus 9300-EX-platform in hardware worden geconfigureerd voor diverse implementatiescenario's in het datacenternetwerk. Het besturingssysteem van de switch (NX-OS) kan softwarebesturing boven op de flexibele hardware plaatsen ter ondersteuning van gevalideerde algemene doorstuurtabelprofielen.

Tabel 3 bevat een overzicht van het door NX-OS ingestelde sjabloonprofiel voor de doorstuurschaal. Raadpleeg de whitepaper over gevalideerde schaalbaarheid voor uw specifieke NX-OS-release voor aanvullende profielen.

Tabel 3. ASIC-tabelcapaciteit

Tabel	Sjabloon 1	Sjabloon 2
LPM IPv4-routes	512.000*	768.000*
LPM IPv6-routes (/64)	512.000*	768.000*
LPM IPv6-routes (/65 tot /127)	2.000	2.000

Tabel	Sjabloon 1	Sjabloon 2
IPv4-hostroutes	512.000*	768.000*
IPv6-hostroutes	24.000	24.000
Multicast	16.000	16.000
MAC-adressen	96.000	16.000

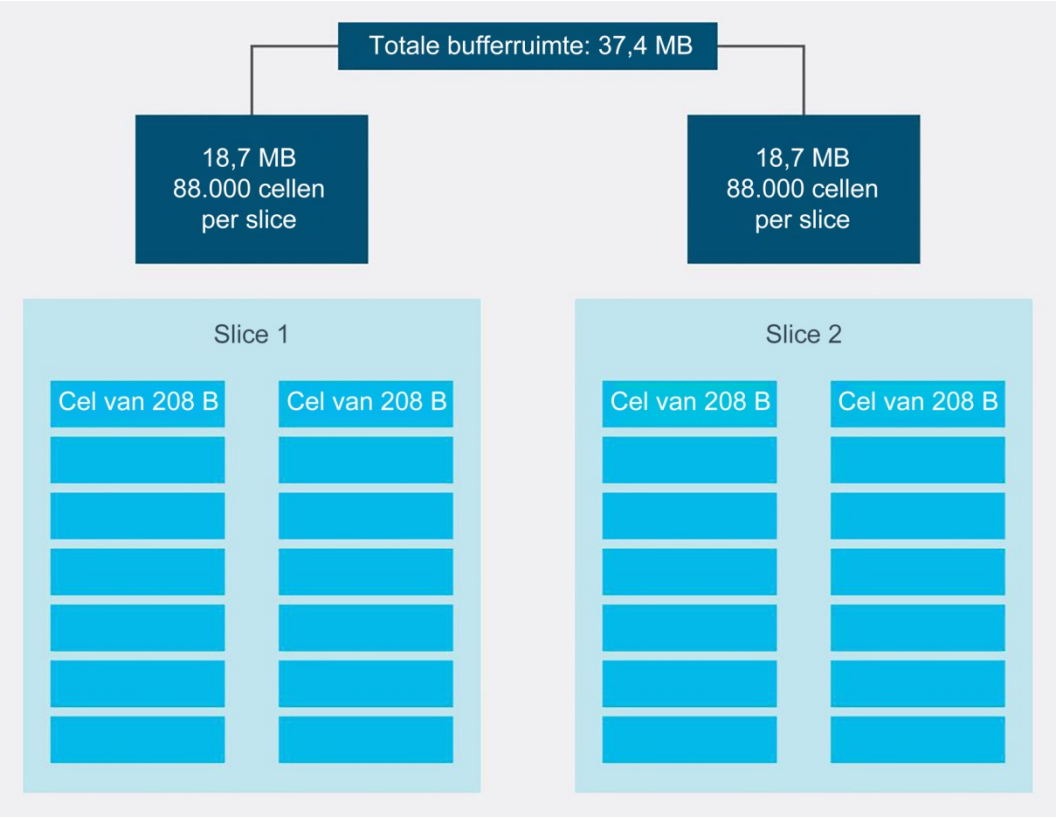
* Gedeelde vermeldingen

Architectuur van de LSE-buffer

De slices in de LSE ASIC's fungeren als switchingsubsystemen. Elke slice heeft een eigen buffergeheugen die door alle poorten op die slice wordt gedeeld. Alleen poorten in die slice kunnen de gedeelde bufferruimte gebruiken.

Om de buffergeheugenresources efficiënt te gebruiken, is het ruwe geheugen opgedeeld in cellen van 208 byte en worden meerdere cellen aan elkaar gekoppeld om het gehele pakket op te slaan. Elke cel kan een geheel pakket of een deel van een pakket bevatten (afbeelding 6).

Afbeelding 6. LSE ASIC-buffer



Tabel 4 bevat een overzicht van de hoeveelheid bufferruimte in de LSE ASIC.

Tabel 4. LSE ASIC-buffercapaciteit

ASIC	Aantal 100 Gigabit Ethernet-poorten	Aantal slices	Aantal buffercellen per slice	Bufferomvang per slice	Totale bufferomvang
LSE	18	2	88.000	18,7 MB	37,4 MB

LSE ASIC's ondersteunen 18 serviceklassen (CoS): 16 door de gebruiker gedefinieerde CoS's, 1 SPAN-CoS (Cisco Switched Port Analyzer) CoS en 1 CPU-CoS. Via de software kan de buffer worden opgedeeld in maximaal vier poolgroepen. Drop en no-drop klassen (mogelijk gemaakt via Priority Flow Control [PFC]) hebben bijvoorbeeld verschillende poolgroepen, en CPU- en SPAN-klassen hebben andere poolgroepen dan door de gebruiker gedefinieerde klassen. Een bepaald aantal cellen wordt toegewezen aan elke poolgroep en deze worden niet met andere poolgroepen gedeeld. Op die manier worden voor elke poolgroep bufferresources gegarandeerd voor de verkeerstypen waarvoor die groep verantwoordelijk is.

Buffertoewijzing

Het bulkgeheugen van de pakketbuffer kan met software statisch worden opgedeeld in ingangs- en uitgangsverwerking via de switchconfiguratie. Het Cisco Nexus 9300-EX-platform maakt standaard gebruik van wachtrijen met op klasse gebaseerd uitgaand verkeer en daarom zijn de meeste buffercellen toegewezen aan de uitgaande wachtrij. Als PFC echter is ingeschakeld, maakt de switch gebruik van inkomende wachtrijen voor de no-drop klassen om onderbrekingsbewerkingen te verwerken. Bij deze configuratie zijn meer buffercellen toegewezen aan de inkomende wachtrij. Deze configuratiegebaseerde bufferopdeling tussen inkomende en uitgaande wachtrijen vergroot de effectieve bufferresources voor de wachtrijstrategie die op de switch wordt geïmplementeerd.

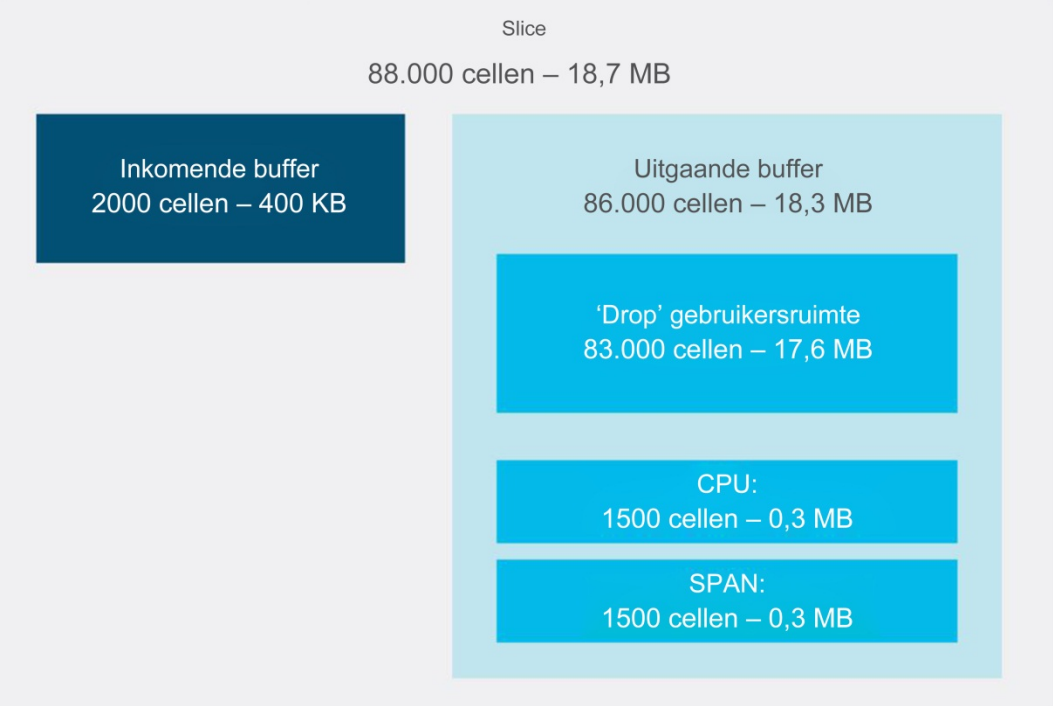
In afbeelding 7 wordt de standaardbuffertoewijzing per slice op de LSE getoond. U kunt zien dat de meeste buffercellen worden toegewezen aan de uitgaande poolgroepen, met uitzondering van een minimale buffertoewijzing voor de inkomende buffer.

Er worden drie uitgaande bufferpoolgroepen gebruikt:

- Door gebruiker gedefinieerde klassen
- CPU
- SPAN

Binnen de poolgroep voor door de gebruiker gedefinieerde klassen kunnen maximaal 16 pools worden gemaakt en onderhouden: twee voor elke CoS (in elke klasse een voor unicastverkeer en een voor multicastverkeer).

Afbeelding 7. Standaardbuffertoewijzingen op LSE ASIC's

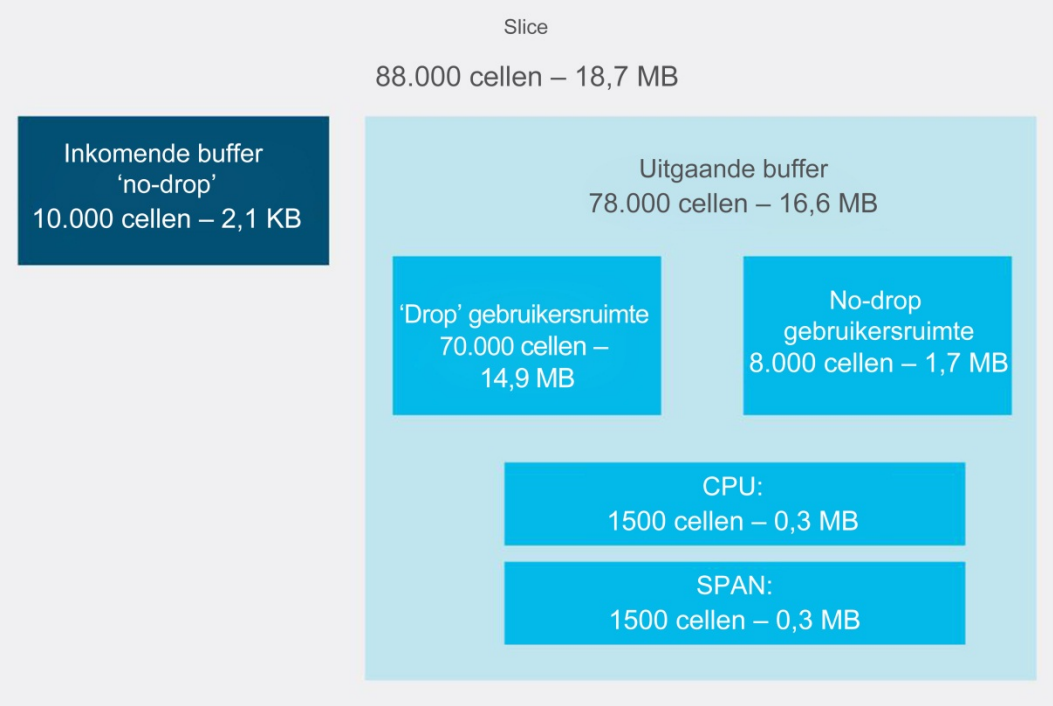


De LSE ondersteunt PFC. PFC biedt lossless semantiek voor verkeer in de no-drop klassen door het onderbrekingsmechanisme per klasse en per poort te gebruiken voor de upstream apparaten. LSE ASIC's voeren onderbrekingsbewerkingen uit met behulp van een buffer voor inkomend verkeer en kunnen maximaal drie no-drop klassen ondersteunen. In een ontwerp met veel poorten is het gebruik van een buffer voor inkomend verkeer efficiënter bij het uitvoeren van onderbrekingsbewerkingen, omdat de buffergrootte alleen nodig is voor de onderbrekingslatenties voor de ingangspoort. Als de onderbrekingsbuffer op de uitgangspoort wordt geïmplementeerd, moet het gedeelde geheugen in het slechtste geval de som van alle poorten op de switch verwerken.

Wanneer PFC is ingeschakeld op het Cisco Nexus 9300-EX-platform, wijst de switch een bepaalde hoeveelheid van de buffer toe aan de inkomende wachtrij op elke ASIC-slice. Deze buffer voor inkomend verkeer wordt op alle poorten in de slice gedeeld en opgedeeld per pool en per poort. Een pool is een intern mechanisme en de softwareconfiguratie bepaald de toewijzing van klassen aan pools.

In afbeelding 8 wordt de buffertoewijzing op de ASIC's getoond wanneer PFC is ingeschakeld. Er is een groter aantal buffercellen gereserveerd voor inkomende no-drop wachtrijen.

Afbeelding 8. Buffertoewijzing op LSE ASIC's met PFC



Intelligent bufferbeheer

LSE ASIC's hebben ingebouwde functies voor intelligent bufferbeheer, met name Approximate Fair Drop (AFD) en Dynamic Packet Prioritization (DPP), voor actief wachtrijbeheer. De functies voor intelligent bufferbeheer voegen per-flow controle toe aan de bestaande mechanismen voor het vermijden en beheren van stremmingen om zo betere toepassingsprestaties te garanderen.

AFD (Approximate Fair Drop)

AFD is een stroombewust early-drop mechanisme dat stremming van het netwerk doorgeeft aan TCP. Voorafgaand aan AFD was WRED (Weighted Random Early Discard) de belangrijkste technologie om stremming door te geven, en het stond ook wel bekend als AQM (Active Queue Management). WRED past een bufferdrempelwaarde voor vroeg afwijzen toe op elke op basis van klasse gewogen wachtrij, maar heeft geen bewustzijn van stromen binnen een klasse. Daarom moet WRED alle verkeersstromen gelijk behandelen en worden pakketten willekeurig voor alle stromen afgewezen. Dit willekeurige proces van afwijzen kan nadelig uitpakken voor kortdurende kleine stromen ('mouse flows'), die gevoeliger zijn voor pakketverlies, terwijl langdurende, grote stromen ('elephant flows') potentieel nog steeds het grootste deel van de buffer opeisen. Daardoor wordt de stroomvoltooiingstijd voor mouse flows aanzienlijk langer en worden de elephant flows onderling ook niet rechtvaardig verwerkt.

AFD houdt daarentegen rekening met informatie over stroomgrootten en gegevensaankomstsnelheden voordat een beslissing over negeren wordt genomen. Zo kan het algoritme de mouse flows die gevoelig zijn voor pakketverlies beschermen en een rechtvaardige verdeling bieden aan concurrerende elephant flows.

Met een ETRAP (Elephant Trap) kan AFD onderscheid wordt gemaakt tussen korte mouse flows en lange elephant flows binnen een bepaalde verkeersklasse en alleen de elephant flows onderwerpen aan de AFD-functie voor vroeg afwijzen. Een stroom kan worden gedefinieerd met verschillende parameters, maar meestal wordt het '5-tuple 'gebruikt. AFD gebruikt een hashtabel om alle actieve stromen te volgen en hun byte-tellingen bij binnenkomst te meten. Een door de gebruiker te configureren ETRAP-drempelwaarde die op de byte-telling is gebaseerd, wordt geïmplementeerd om te beslissen of een stroom een mouse flow of een elephant flow is. Een stroom is een mouse flow wanneer deze tijdens zijn levenscyclus minder bytes overdraagt dan de ETRAP-drempelwaarde. Nadat de byte-telling van een stroom de ETRAP-drempelwaarde overstijgt, wordt de stroom beschouwd als elephant flow en wordt deze verplaatst naar de elephant flowtabel om verder te volgen en is deze onderwerp van AFD-beslissingen om weg te plaatsen.

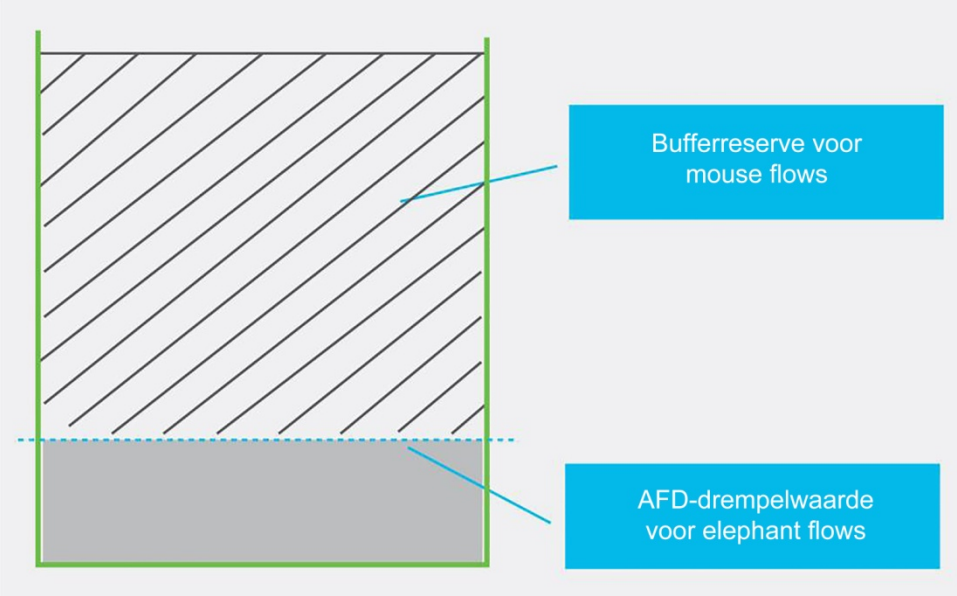
Daarnaast heeft AFD de intelligentie om rechtvaardige afwijzingen toe te passen onder elephant flows op basis van hun gegevensaankomsttijd wanneer de AFD-bufferdrempelwaarde voor vroeg afwijzen wordt overschreden. Het algoritme heeft twee hoofdelementen.

Eén element is snelheidsmeting: ETRAP meet de aankomsttijd van elke stroom in de tabel met elephant flows op de inkomende poort, en de gemeten aankomsttijd wordt meegegeven in de pakketheader wanneer pakketten intern worden doorgestuurd naar de uitgaande poort.

Het andere hoofdelement van AFD is de berekening van een rechtvaardige snelheid: het AFD-algoritme berekent dynamisch een rechtvaardige snelheid per stroom op een uitgaande poort met een feedbackmechanisme op basis van de bezetting van de wachtrij van de uitgaande poort. Wanneer een pakket van een elephant flow de wachtrij van de uitgaande poort binnenkomt, dan vergelijkt het AFD-algoritme de gemeten aankomsttijd van de stroom met de berekende rechtvaardige verhouding. Wanneer de aankomstsnelheid van een elephant flow lager is dan de rechtvaardige snelheid per stroom, dan wordt het pakket niet afgewezen. Wanneer de aankomstsnelheid echter hoger is dan de berekende rechtvaardige snelheid per stroom op de uitgaande poort, dan wordt hetzelfde aantal pakketten afgewezen uit die stroom als waarmee de stroom de rechtvaardige snelheid overstijgt. De kans op afwijzing wordt berekend met de rechtvaardige snelheid en de gemeten stroomsnelheid. Daardoor bereiken alle elephant flows de rechtvaardige snelheid. De AFD-parameters voor de uitgaande wachtrij worden geconfigureerd met profielen. Het profiel kan, net als bij WRED, worden geconfigureerd zodat een pakket met ECN (Explicit Congestion Notification) wordt aangemerkt in plaats van dat het wordt afgewezen.

In afbeelding 9 wordt het algehele effect van AFD getoond. Door alleen op elephant flows het algoritme voor vroeg afwijzen toe te passen, kan AFD ongewenst afwijzen van pakketten van mouse flows voorkomen en genoeg bufferruimte vrijhouden om pieken aan te kunnen in de vorm van grote aantallen gelijktijdige mouse flows (inkomend en microburst verkeer). Bij lange elephant flows past het AFD-algoritme rechtvaardig vroeg afwijzen toe op basis van de gegevensaankomstsnelheid.

Afbeelding 9. Stroomgebaseerd vroeg afwijzen met AFD



Dynamic Packet Prioritization (DPP)

DPP kan significante latentievoordelen bieden voor korte kleine stromen tijdens stremming van het netwerk door automatisch prioriteit te geven aan de eerste paar pakketten uit elke stroom.

Wanneer een verkeersstroom een uitgaande wachtrij kruist, dan wordt de pakkettelling gemeten en gecontroleerd ten opzichte van een prioriteringsdrempelwaarde gebaseerd op een door de gebruiker instelbare pakkettelling. Wanneer het aantal ontvangen pakketten in een stroom onder de prioriteringsdrempelwaarde ligt, dan krijgen de pakketten prioriteit om de rest van de wachtrij te passeren. Wanneer de pakkettelling van de stroom hoger is dan de drempelwaarde, dan krijgen de overtollige pakketten in de stroom niet langer prioriteit. Omdat korte kleine stromen, zoals microburst stromen, uit zeer weinig pakketten per stroom bestaan, zullen ze de drempelwaarde niet overstijgen. Zo krijgt de gehele kleine stroom prioriteit.

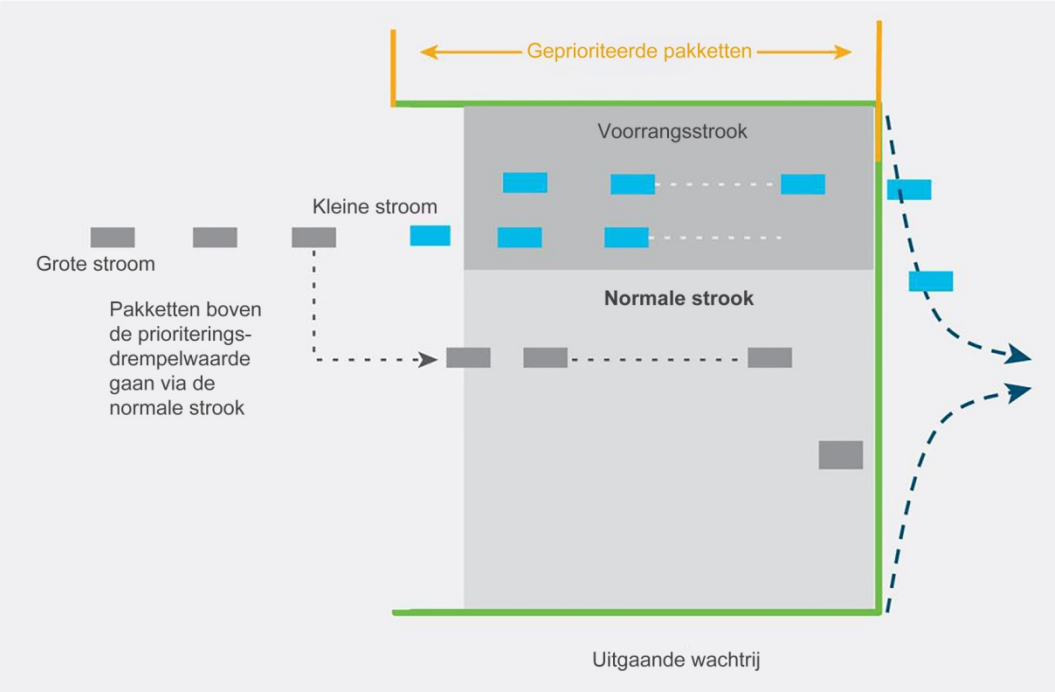
Voor lange grote stromen doorloopt de rest van de stroom, nadat de eerste paar pakketten zijn toegestaan door de drempelwaarde, het normale wachtrijproces.

Zoals getoond in afbeelding 10 creëert DPP in feite een voorrangsstrook voor korte kleine stromen, en laat het lange grote stromen op de normale strook. Deze aanpak zorgt dat kleine stromen prioriteit kunnen hebben in zowel de switch als het netwerk om het aantal afwijzingen en de latentie omlaag te brengen.

Omdat kleine stromen in de meeste datacentertoepassingen gevoeliger zijn voor pakketverlies en lange latentie dan lange grote stromen, verbetert het prioriteren van kleine stromen de algemene toepassingsprestaties.

Stroom prioriteren kan worden gebruikt in combinatie met het AFD-algoritme om rechtvaardig af te wijzen bij de langdurende grote stromen en de kleine stromen prioriteit te geven met voldoende bufferruimte om grote aantallen kleine stromen tegelijk mogelijk te maken (inkomend en microburst verkeer). Deze aanpak verkort de gemiddelde wachtrijlengte zonder het aantal time-outs voor kleine stromen te laten toenemen, waardoor de prestaties sterk verbeteren.

Afbeelding 10. Dynamic Packet Prioritization (DPP)



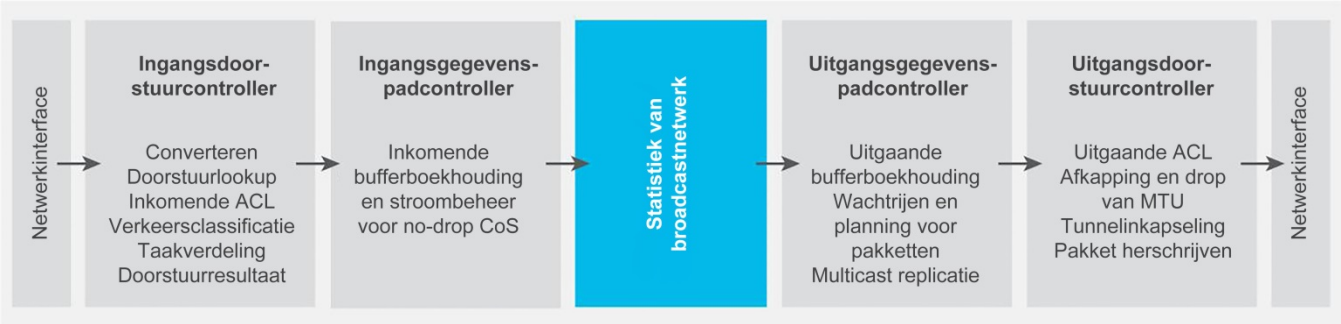
Doorsturen van unicast pakketten op het Cisco Nexus 9300-EX-platform

Doorstuurpipelines op LSE ASIC's

Het doorsturen van unicast pakketten wordt op het Cisco Nexus 9300-EX-platform uitgevoerd door de LSE ASIC van de doorstuur-engine. De LSE ASIC heeft twee slices. Elke slice vertegenwoordigt een switching subsysteem met zowel een inkomende doorstuurpipeline als een uitgaande doorstuurpipeline. De inkomende doorstuurpipeline op elke slice bestaat uit een I/O-component, ingangsdoorstuurcontroller en ingangsgegevenspadcontroller. De uitgaande doorstuurpipeline bestaat uit de uitgangsgegevenspadcontroller, uitgangsdoorstuurpadcontroller en een andere I/O-component. Alle slices zijn verbonden met een broadcastnetwerk dat point-to-multipoint verbindingen biedt vanuit elke slice, zodat tussen slices connectiviteit in alle richtingen mogelijk is. Het broadcastnetwerk biedt voldoende bandbreedte om doorsturen op volle lijnsnelheid tussen alle slices tegelijk mogelijk te maken.

Afbeelding 11 toont de doorstuurpipelines op een Cisco Nexus 9300-EX-platformswitch. Wanneer een pakket een Cisco Nexus 9300-EX-platformswitch binnenkomt, gaat dit door de inkomende pipeline van de slice waarop de inkomende poort huist, passeert dit het interne ASIC-broadcastnetwerk om op de uitgaande slice te komen en gaat dit vervolgens door de uitgaande pipeline van de uitgaande slice.

Afbeelding 11. Doorstuurpipelines op het Cisco Nexus 9300-EX-platform



Inkomende pipeline: ingangsdoorstuurcontroller

De ingangsdoorstuurcontroller ontvangt het pakket van het MAC-adres van de inkomende poort, parseert de pakketheaders en voert een reeks lookups uit om te bepalen of het pakket moet worden geaccepteerd en hoe het moet worden doorgestuurd naar de beoogde bestemming. De controller genereert ook instructies voor het gegevenspad om het pakket op te slaan en in de wachtrij te zetten. Omdat de volgende generatie ASIC-switches van Cisco uit cut-through switches bestaat, worden inkomende doorstuurlookups uitgevoerd terwijl het pakket wordt opgeslagen in het bufferblok voor onderbrekingen. De ingangsdoorstuurcontroller voert meerdere taken uit in de volgorde zoals getoond in afbeelding 11:

- Pakketheader parseren
- Layer 2-lookup
- Layer 3-lookup
- Inkomende ACL-verwerking (Access Control List)
- Classificatie van inkomend verkeer
- Genereren van doorstuurresultaten

Pakketheader parseren

Wanneer een pakket binnenkomt via een poort op het voorpaneel, gaat die door de inkomende pipeline en wordt als eerste stap de pakketheader geparseerd. De flexibele pakketparser converteert de eerste 128 bytes van het pakket om informatie te extraheren en op te slaan, zoals de Layer 2-header, EtherType-waarde, Layer 3-header en het TCP/IP-protocol. Deze informatie wordt gebruikt voor het latere lookups van pakketten en voor proceslogica.

Layer 2- en Layer 3-doorstuurlookups

Wanneer het pakket door de inkomende pipeline gaat, wordt het onderworpen aan Layer 2-switching- en Layer 3-routing-lookups. Eerst bekijkt het doorstuurproces de DMAC (Destination MAC-adres) van het pakket om te bepalen of het pakkent moet worden geswitched (Layer 2) of gerouteerd (Layer 3). Als de DMAC overeenkomt met het eigen router-MAC-adres van de switch, dan wordt het pakket overgedragen aan de logica voor Layer 3-routing-lookups. Als de DMAC niet bij de switch hoort, dan wordt een Layer 2-switching-lookup uitgevoerd op basis van de DMAC en het VLAN-id. Als een overeenkomst wordt gevonden in de MAC-adrestabel, wordt het pakket doorgestuurd naar de uitgaande poort. Als er geen resultaat is voor de combinatie van DMAC en VLAN, wordt het pakket doorgestuurd naar alle poorten in hetzelfde VLAN.

Binnen de logica voor Layer 3-lookups wordt de DIP (Destination IP-adres) gebruikt voor zoekopdrachten in de Layer 3-hosttabel. Deze tabel slaat doorstuurvermeldingen op voor direct verbonden hosts en geleerde /32 hostroutes. Wanneer de DIP overeenkomt met een vermelding in de hosttabel, dan geeft de vermelding de bestemmingspoort aan, het MAC-adres van de volgende hop en het uitgaande VLAN. Wanneer er geen resultaat voor de DIP wordt gevonden in de hosttabel, dan wordt een LPM-lookup uitgevoerd in de LPM-routingtabel.

Inkomende ACL-verwerking

Naast verwerking van doorstuurlookups wordt het pakket onderworpen aan inkomende ACL-verwerking. Het ACL-TCAM wordt gecontroleerd op resultaten voor inkomende ACL's. Elke ASIC heeft een tabel met inkomende ACL-TCAM's met 4000 vermeldingen per slice om interne ACL's in het systeem en door de gebruiker gedefinieerde inkomende ACL's te ondersteunen. Deze ACL's omvatten Port ACL's (PACL's), Routed ACL's (RACL's) en VLAN ACL's (VAC's). ACL-vermeldingen worden gelokaliseerd naar de slice en alleen geprogrammeerd wanneer dat nodig is. Deze aanpak maakt optimaal gebruik van het ACL-TCAM in de Cisco Nexus 9300-EX-platformswitch.

Classificatie van inkomend verkeer

Cisco Nexus 9300-EX-platformswitches ondersteunen classificatie van inkomend verkeer. Op een inkomende interface kan verkeer worden geclassificeerd op basis van het adresveld, de IEEE 802.1q-CoS, en IP-voorrang of DSCP (Differentiated Services Code Point) in de pakketheader. Het geclassificeerde verkeer kan worden toegewezen aan één van de acht QoS-groepen (Quality-of-Service). De QoS-groepen identificeren intern de verkeersklassen die worden gebruikt voor volgende QoS-processen wanneer pakketten het systeem doorlopen.

Aanmaken van doorstuurresultaten inkomend verkeer

De laatste stap in de inkomende doorstuurpipeline is om alle doorstuurmetagegevens die eerder in de pipeline zijn gegenereerd, te verzamelen en door te geven aan de blokken verderop in het gegevenspad. Een interne header van 64 bytes wordt samen met het inkomende pakket in de pakketbuffer opgeslagen. Deze interne header bevat 16 bytes aan iETH-headergegevens (intern communicatieprotocol), die wordt toegevoegd bovenop het pakket wanneer het pakket wordt overgedragen naar de uitgangsgegevenspadcontroller door het broadcastnetwerk. Deze iETH-header van 16 bytes wordt verwijderd wanneer het pakket de poort op het voorpaneel verlaat. De overige 48 bytes interne headerruimte worden alleen gebruikt om metagegevens van de ingangsdoorstuurwachtrij door te geven aan de uitgangsdoorstuurwachtrij en worden verwerkt door de uitgangsdoorstuur-engine.

Inkomende pipeline: ingangsgegevenspadcontroller

De ingangsgegevenspadcontroller voert inkomende boekhoudfuncties, toelatingsfuncties en stroombeheer uit voor de no-drop CoS. Het inkomende toelatingsbeheermechanisme bepaalt of een pakket in het geheugen moet worden toegelaten. Deze beslissing wordt gebaseerd op het beschikbare buffergeheugen en de hoeveelheid bufferruimte die al wordt gebruikt door de inkomende poort en verkeersklasse. De ingangsgegevenspadcontroller stuurt het pakket via het broadcastnetwerk door naar de uitgangsgegevenspadcontroller.

Broadcastnetwerk en centrale statistiekmodule

Het broadcastnetwerk is een set van point-to-multipoint draden die connectiviteit mogelijk maken tussen alle slices op de ASIC. De ingangsgegevenspadcontroller heeft een point-to-multipoint verbinding met de uitgangsgegevenspadcontroller op alle slices, waaronder de eigen slice. De centrale statistiekmodule is verbonden met het broadcastnetwerk. De centrale statistiekmodule biedt statistieken voor pakketten, bytes en atoomtellers.

Uitgaande pipeline: uitgangsgegevenspadcontroller

De uitgangsgegevenspadcontroller voert de uitgaande bufferboekhouding uit, wachtrijen en planning voor pakketten en multicast replicatie. Alle poorten delen dynamisch de uitgaande bufferresource. De details van dynamische buffertoewijzing staan eerder in dit document vermeld.

De uitgangsgegevenspadcontroller voert ook de shaping van pakketten uit. Volgens het ontwerpprincipe van eenvoud en efficiëntie, maakt het Cisco Nexus 9300-EX-platform gebruik van een eenvoudige uitgaande wachtrijarchitectuur. In het geval van stremming van de uitgaande poort worden pakketten direct in de buffer van de uitgaande slice in de wachtrij geplaatst. Er bevinden zich geen VoQ's (Virtual output Queues) op de inkomende slice. Deze aanpak vereenvoudigt bufferbeheer van het systeem en wachtrij-implementatie aanzienlijk.

Een Cisco Nexus 9300-EX-switch kan tot 10 uitgaande verkeersklassen ondersteunen, 8 gebruiker-bepaalde klassen die worden herkend aan QoS-groep-id's, een CPU-verkeersklasse en een SPAN verkeersklasse. Elke door de gebruiker gedefinieerde gebruiker klasse kan per uitgaande poort een unicast wachtrij en een multicast wachtrij hebben. Deze aanpak helpt te zorgen dat geen enkele poort meer kan verwerken dan zijn rechtvaardige deel van het buffergeheugen en dat geen buffertekort optreedt voor andere poorten.

Uitgaande pipeline: uitgangsdoorstuurcontroller

De uitgangsdoorstuurcontroller ontvangt het ingangspakket en bijbehorende metagegevens van de bufferbeheerder en is verantwoordelijk voor alle herschrijftaken voor pakketten en toepassing van het uitgaande beleid. De controller extraheert interne headergegevens en diverse pakketheadervelden uit het pakket, voert een aantal lookups uit en genereert de herschrijfinstructies.

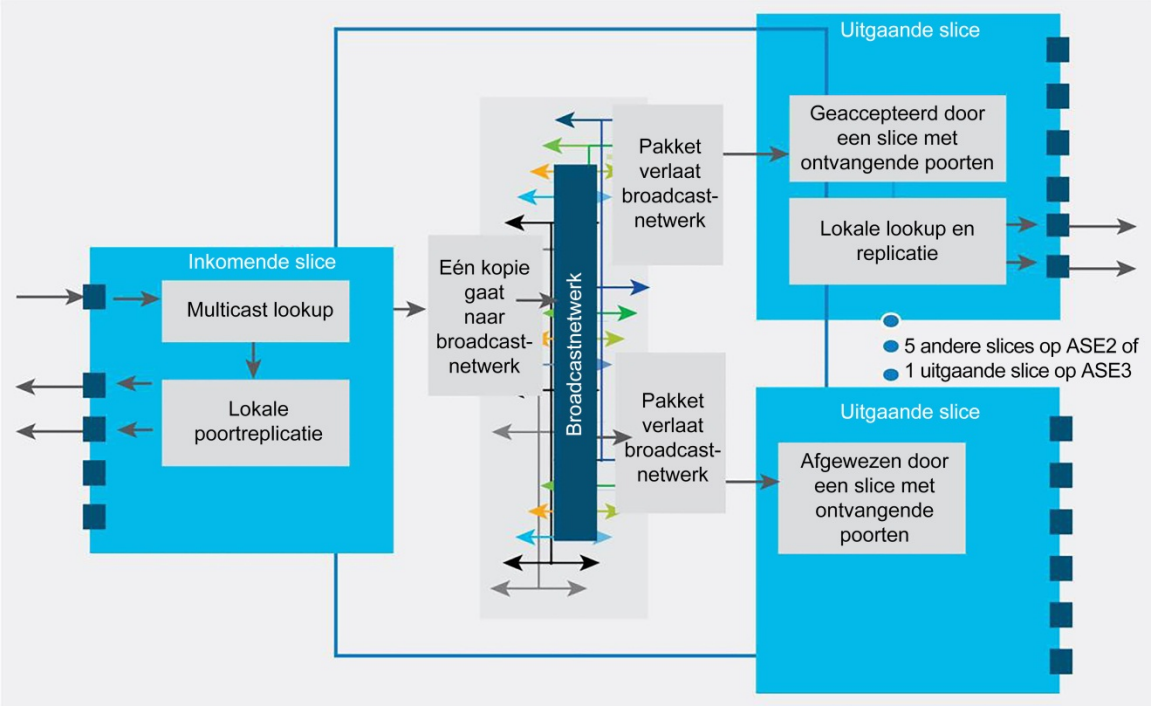
Doorsturen van multicast pakketten op het Cisco Nexus 9300-EX-platform

Multicast pakketten doorlopen op een Cisco Nexus 9300-EX-platformswitch dezelfde inkomende en uitgaande doorstuurpipelines als de unicast pakketten, behalve dat multicast doorstuurlookups gebruikmaken van multicast tabellen en dat multicast pakketten een replicatieproces van meerdere stappen doorlopen om te worden doorgestuurd naar meerdere bestemmingspoorten.

De LSE ASIC bestaat uit twee slices die onderling verbonden zijn door een niet-blokkerend intern broadcastnetwerk. Wanneer een multicast pakket aankomt bij een poort op het voorpaneel, dan voert de ASIC een doorstuurlookup uit. Deze lookup haalt lokale ontvangende poorten op dezelfde slice als de inkomende poort op en geeft een lijst van bedoelde ontvangende slices met ontvangende poorten in de multicast bestemmingsgroep. Het pakket wordt gerepliceerd op de lokale poorten en één kopie van het pakket wordt verzonden naar het interne broadcastnetwerk, waarbij de bitvector in de interne header is ingesteld om de beoogde ontvangende slices aan te duiden. Alleen de beoogde ontvangende slices zullen het pakket direct van de draad van het broadcastnetwerk ontvangen. De slices zonder ontvangende poorten voor deze groep zullen het pakket simpelweg afwijzen. De ontvangende slice voert vervolgens lokale Layer 3-replicatie of Layer 2-fanout-lookup en -replicatie om een kopie van het pakket door te sturen naar elk van de lokale ontvangende poorten.

Afbeelding 12 toont het multicast doorstuurproces.

Afbeelding 12. Multicast doorstuurproces



Conclusie

Cisco Nexus 9300-EX-platformswitches vormen de volgende generatie van vaste Cisco Nexus 9000 Series-switches. Het nieuwe platform is gebaseerd op de Cisco Cloud Scale ASIC en ondersteunt kosteneffectieve implementaties op cloudschaal, meer endpoints en cloudservices met wire-rate beveiliging en telemetrie. Het platform is gebouwd op een moderne systeemarchitectuur die is ontworpen voor hoge prestaties om aan de veranderende behoeften van zeer schaalbare datacenters en groeiende ondernemingen te voldoen. Cisco Nexus 9300-EX-platformswitches bieden diverse interfaceopties om bestaande datacenters transparant te migreren van snelheden van 100 Mbps, 1 Gbps, en 10 Gbps naar 25 Gbps op de server, en van 10 en 40 Gbps naar 50 en 100 Gbps in de aggregatielaag. Het platform kan uitgebreide telemetriegegevens van Cisco Tetration Analytics™ verzamelen op lijnsnelheid over alle poorten zonder enige latentie aan de pakketten toe te voegen of switchprestaties negatief te beïnvloeden.

Meer informatie

Klik hier voor meer informatie over Cisco Nexus 9000 Series-switches.
Of vraag een gesprek aan met een Cisco-expert.

Hoofdkantoor Amerika
Cisco Systems, Inc
San Jose, CA

Hoofdkantoor Zuidoost-Azië
Cisco Systems (USA) Pte, Ltd.
Singapore

Hoofdkantoor Europa
Cisco Systems International BV Amsterdam,
Nederland

Cisco beschikt wereldwijd over meer dan 200 kantoren. Adressen, telefoonnummers en faxnummers vindt u op de Cisco-website op **www.cisco.com/go/offices.**

Gedrukt in de VS C11-739134-00 06/17