White paper
Cisco public

The bridge to possible

Nexus 9000: SPAN-on-Drop

Contents

Introduction

Packet drops are very common in networks and may cause severe issues to the applications. While TCP-based applications rely on retransmission of lost packets for UDP-based applications, packet drops need to be dealt at the application layer. In either scenario, the response time of the application gets affected.

Packet drops are always undesired, but at the same time they are not always avoidable. However, having the information about the traffic/packets that are being dropped gives network and application administrators the capability to adjust traffic flows so that packet drops can at least be minimized.

Packet drops can be classified into two major categories:

- Buffer drops
- Forwarding drops

Buffer drops

Buffer drops happen when a device does not have enough buffer credits to accept the packets to be sent out on an egress port.

When a switch has more packets to be sent out than the capacity of the egress port, it starts buffering them. If the ingress flows are constantly pushing more traffic, the switch will eventually run out of all of the buffer space allocated to the egress port, and subsequent packets will be dropped. This type of drop is also known as tail drop, because the packets at the tail of the buffer are dropped due to their failure to be admitted to the buffer.

Another common reason for buffer drops is when using traffic policing that is configured for certain QoS classes. Policing is commonly used to rate-limit the traffic for a certain class, and any traffic exceeding the defined threshold (CIR) is simply dropped.

Forwarding drops

Forwarding drops happen during the forwarding lookup processes, when the switch decides the fate of the packet by looking at its header. Some of the examples for forwarding drops are FIB miss, VLAN miss, MTU mismatch, ACL drops, ingress policer drops, etc.

SPAN-on-Drop (SOD)

The SPAN-on-drop feature enables spanning of packets that would otherwise be dropped due to forwarding issues or the unavailability of buffers in queue spaces. The feature can be used to identify victims of packet loss due to congestion or forwarding drops. If it is detected that a packet is going to be dropped, the system copies those packets in a separate buffer, marks those packets as SPAN-on-drop, and then sends them to the specified SPAN-on-drop destination.

The SPAN-on-drop feature is supported over an ERSPAN session. The administrator configures an ERSPAN source session on the node where they want to SPAN any drops. The SPAN copies of the drops are then encapsulated over IP-GRE ERSPAN header and sent across the tunnel. The receiver device can be another networking device that supports ERSPAN decapsulation where the user will configure an ERSPAN termination session.

Note: This document focuses on SPAN-on-drop feature for the Cisco Nexus® 9000 (Cloud Scale family) series of switches.

Nexus 9000 Cloud Scale ASIC architecture

The Cisco Nexus 9000 Cloud Scale family of ASICs are based on SOC (switch on chip) architecture. Each ASIC is divided into one or more slices. A slice has the following properties:

- Self-contained forwarding complex controls a subset of ports on a single ASIC.
- Ingress and egress functions are separated.
- Ingress of each slice is connected to egress of all other slices.

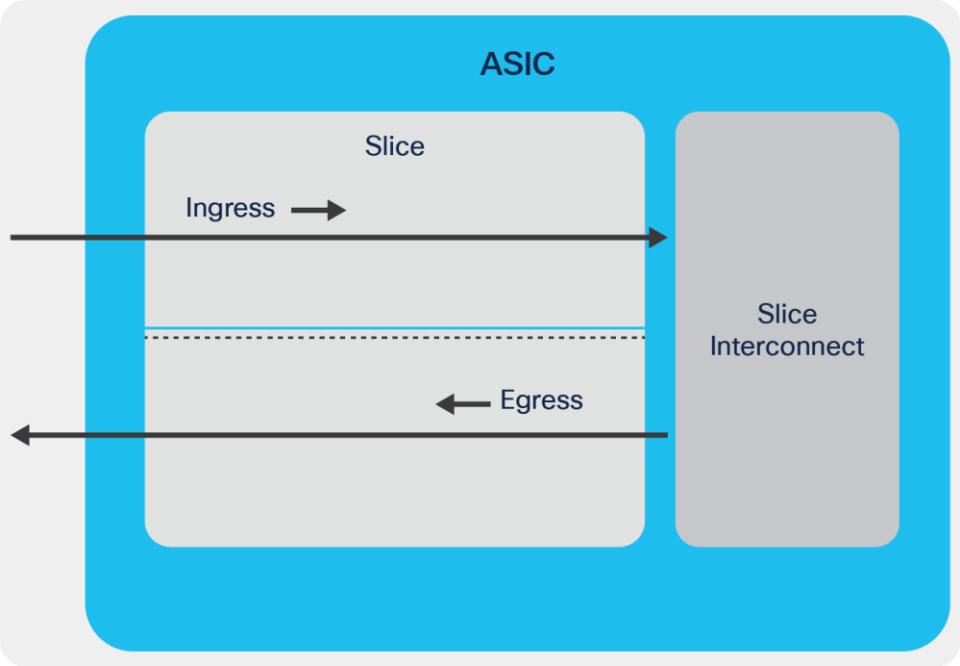

Figure 1.
Cisco® Cloud Scale ASIC

As shown in the diagram above, the forwarding pipeline of a slice is separated into ingress and egress pipelines.

The ingress pipeline of each slice is connected to the egress pipeline of all other slices (including itself) that are present on the same ASIC. This slice interconnect (shown in the figure below) is a nonblocking any-to-any interconnection.

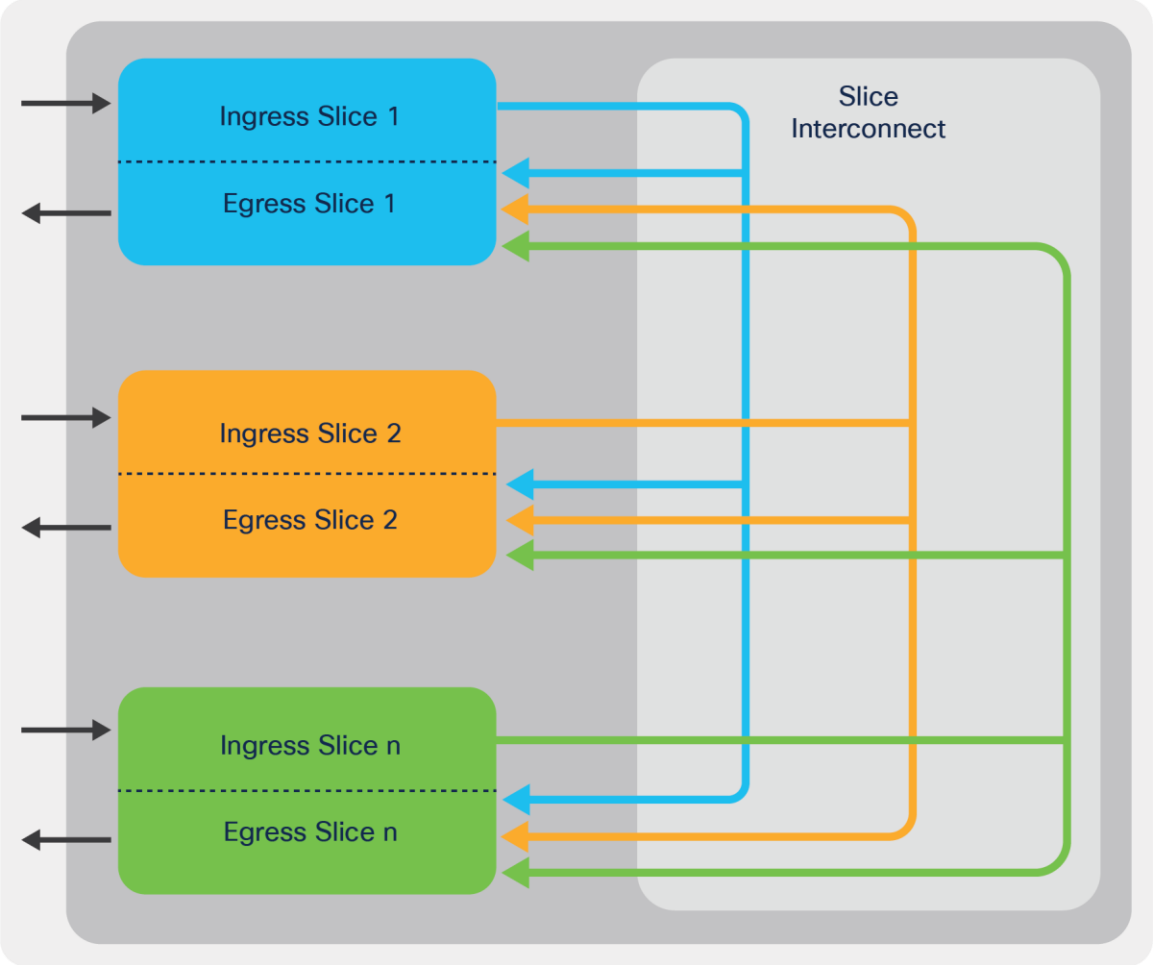

Figure 2.
Slice interconnect

Slice forwarding path

The forwarding functions of a slice can be divided into two parts: ingress forwarding and egress forwarding.

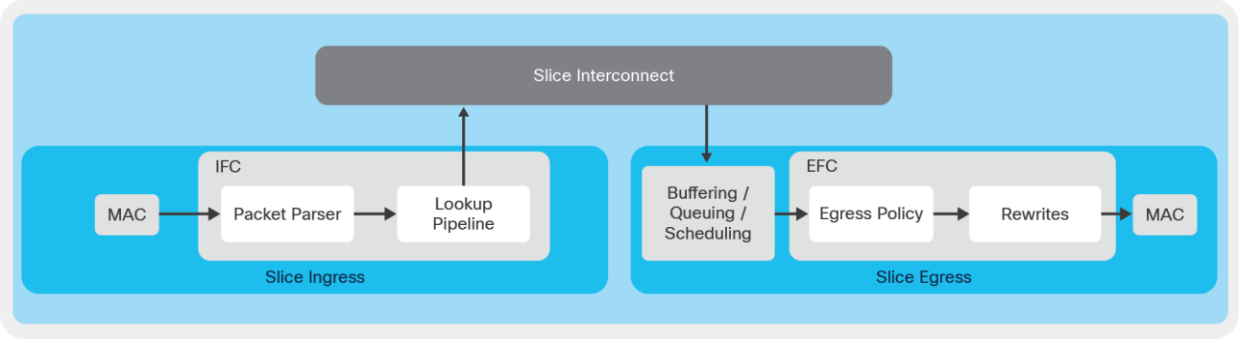

Figure 3.
Slice forwarding path

Some of the major operations are summarized as below:

Slice ingress forwarding

- Frames are received by the MAC block.
- Parser extracts the header information of the received frame and provides that information to the lookup pipeline for decision making.
- Lookup pipeline utilizes the packet header information to make forwarding decision with the help of various forwarding tables.

Once the lookup decision is made, the frame along with the lookup result is forwarded to the egress pipeline of the slice.

Slice egress forwarding

- Buffer block receives the frame through the slice interconnect, and, depending on the buffer occupancy, it admits or discards the packets.
- Egress policy block enforces the policies as applicable on the egress port.
- Rewrite block finally rewrites the header based on the lookup result provided by the slice ingress.
- Finally, the frame is transmitted out by the MAC block.

SPAN-on-drop operations

SPAN-on-drops can be categorized as forwarding (ingress) or buffer (egress) drops. The following drop reasons are supported by SPAN-on-drop:

Ingress drops:

- FIB miss
- MTU exceptions
- VLAN miss

Egress drops:

- Oversubscription drops
- Tail drops

Note:

If a tail drop happens on the same slice as an ERSPAN destination, the buffer manager converts drops to SPAN, and packets are sent to the ERSPAN destination.

If a tail drop happens on a different slice, span copies are recirculated and L3 forwarded in a second pass to the ERSPAN destination.

In both scenarios, the rewrite block on slice egress (on which the buffer drop is observed) is responsible for encapsulating the dropped packet with the ERSPAN header.

If the ERSPAN destination is reachable from same slice, it gets forwarded out from the egress pipeline.

If the ERSPAN destination is reachable from any other slice, this ERSPAN-encapsulated packet is fed back to the parser (the ingress pipeline) of the same slice and takes a normal forwarding path, but this time the forwarding decision is made using the ERSPAN header. This process is called recirculation (shown below).

Figure 4. Slice recirculation path

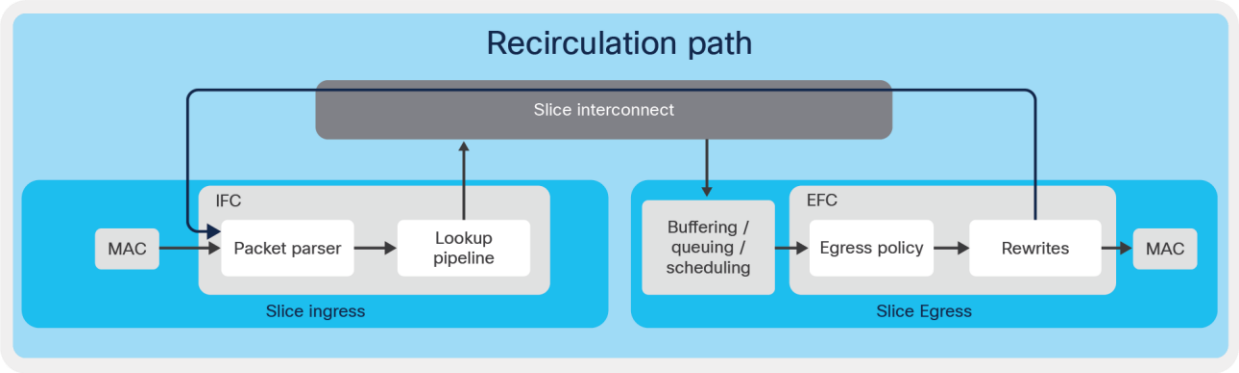

Figure 4.
Packet recirculation path

Each slice has two recirculation ports that have reserved source IDs; the ports have a bandwidth of 200 Gig each. For SPAN-on-drops, only one of the recirculation ports is used.

SPAN-to-drop support matrix

SPAN-to-drop is supported on Nexus 9000 Cloud Scale ToR (Top of Rack) and EoR (End of Row) platforms. Nexus 9000 EX/FX/FX2/FX3/GX series support only the forwarding drops, while Nexus 9000 GX2 series supports both forwarding drops and buffer drops.

Table 1. SPAN-on-drop support matrix

Support	EX/FX/FX2/FX3	GX	GX2
Forward drops support	Yes	Yes	Yes
Egress drops support	No	No	Yes
Buffer drops	N/A	Yes (Uses feedback mechanism)	Yes (Uses recirculation)
All buffer drops are spanned.	No	No	Yes

In the SPAN-on-drop (SOD) feedback mechanism, the first packet always gets dropped, which signals the system to initiate SOD, and subsequent packets are captured by span. In the feedback mechanism, once the first packet is set to be dropped and identified as a candidate for SOD, the system starts capturing all the subsequent packets pertaining to the same flow for SOD irrespective whether they are actually getting dropped or not.

However, in the recirculation mechanism, all the packets to be dropped are spanned.

SOD is supported on the Nexus Cloud Scale platform from Cisco NX-OS Release 6.x onwards.

SOD is supported on Cisco Nexus 9000 GX2 series switches from Cisco NX-OS Release 10.2 onwards.

Guidelines and limitations

- Only ERSPAN is supported as the destination for SPAN-on-drop.
- Local SAPN is not supported as an SOD destination.
- Recirculation bandwidth is limited to only 200 Gbps; this might affect the rate of SOD captures.
- SOD captures only the first 624 bytes of the frames, so for drops with a larger frame size, the original packets will be truncated to a size of 624 bytes + ERSPAN encapsulation.
- TTL, WRED, shaping, and policing drops are not supported by SOD.

Configuration

```
N9K(config)# monitor session 1 type erspan-source
N9K(config-erspan-src)# erspan-id 3
N9K(config-erspan-src)# vrf default
N9K(config-erspan-src)# destination ip 20.20.20.2
N9K(config-erspan-src)# source forward-drops rx
N9K(config-erspan-src)# no shut
N9K(config)# monitor erspan origin ip-address 20.20.20.1 global
```

Verification

```
N9K(config)# sh monitor session 1
   session 1
---------------
type              : erspan-source
state             : up
erspan-id         : 3
vrf-name          : default
acl-name          : acl-name not specified
ip-ttl            : 255
ip-dscp           : 0
destination-ip    : 20.20.20.2
origin-ip         : 20.20.20.1 (global)
source intf       :
    rx            :
    tx            :
    both          :
source VLANs      :
    rx            :
    tx            :
    both          :
filter VLANs      : filter not specified
source fwd drops  : high priority
marker-packet     : disabled
packet interval   : 100
packet sent       : 0
packet failed     : 0
egress-intf       : Eth1/2
source VSANs      :
    rx            :
N9K(config)#
```

Conclusion

The SPAN-on-drop feature can be helpful in troubleshooting application slowness issues caused by packet losses in the network. By using SOD, network administrator can gather information about the packets getting dropped and can determine drop reason and take corrective actions by optimize the network or application accordingly.

Additionally, data captured by SOD can be fed to telemetry tools to analyze congestions and drops in the network.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **https://www.cisco.com/go/offices**.

Printed in USA C11-2721953-00 01/22