
TOMORROW starts here.
cisco
Cisco live!

Troubleshooting Cisco Nexus 7000 Series Switches

BRKDCT-3144

Omar Yassin – CCIE Data Center

Expert Engineer - Data Center Business Unit

Agenda

- Before You Get Started
  - Traditional Versus NX-OS Troubleshooting Approach
  - Nexus 7000 Built-in Troubleshooting Tools
  - Architecture Overview
- Troubleshooting
  - CPU
  - Control Plane
  - Memory Utilization
  - ACL
  - vPC
  - Unicast Layer 2 and Layer 3 Forwarding and ARP
  - Multicast Layer 2 and Layer 3 Forwarding

Before You Get Started

Troubleshooting Mind Map

What ..

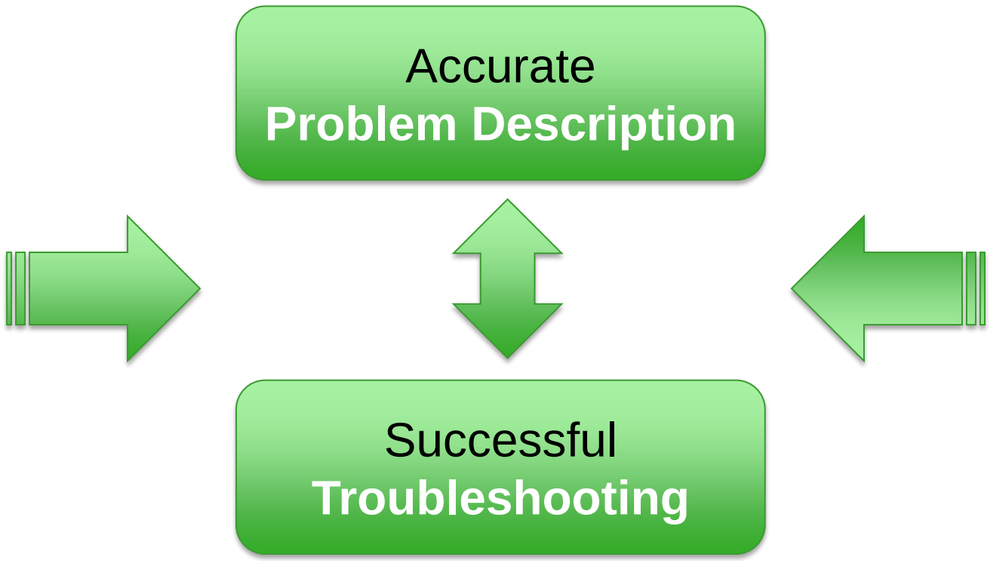
is broken or not functioning as expected

Why ..

is it broken and is there a workaround

When ..

the functionality broke or started to misbehave



Platform..

knowledge: Hardware, Software, Features, and Troubleshooting capabilities

Topology ..

knowledge and data path through topology

Interaction ..

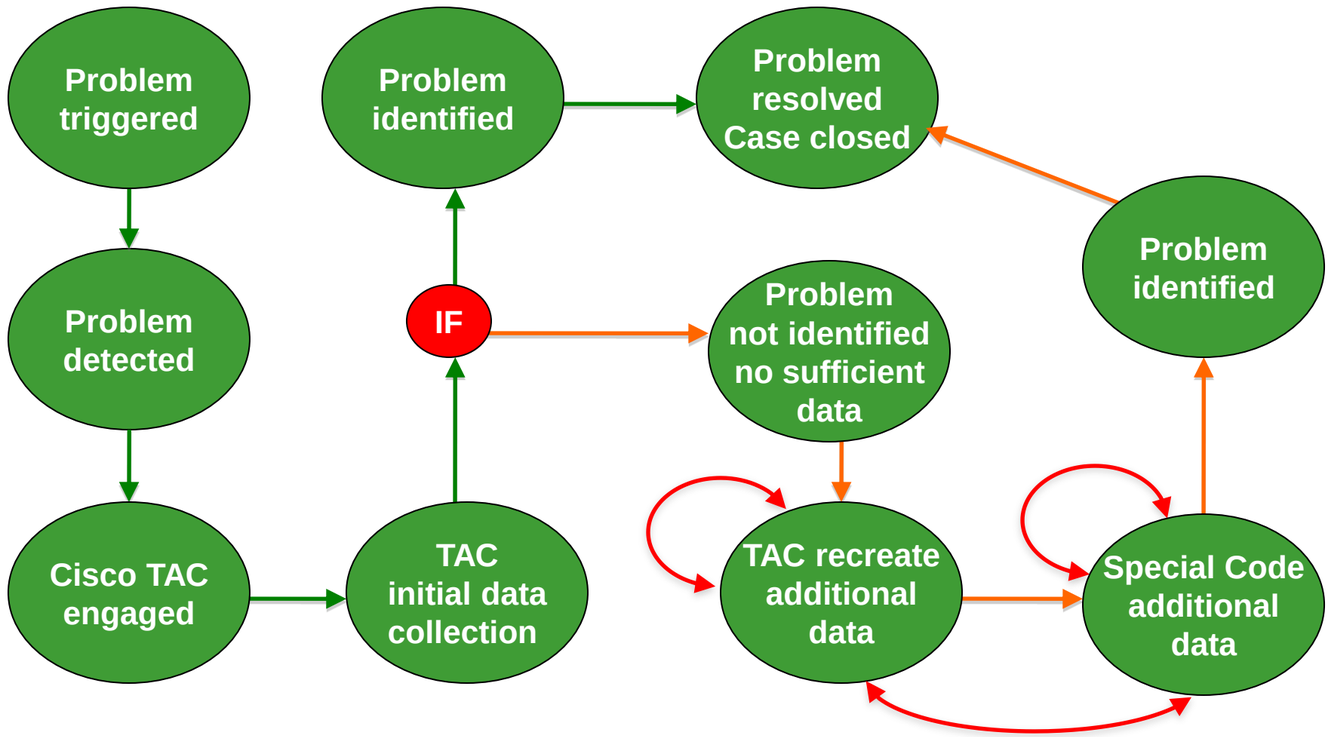
between Cisco devices and other vendors equipment's

Before You Get Started

Traditional Approach

Facts

- The decisions and actions taken when an issue is triggered depend on the device's debugging capabilities and troubleshooting tools.
- The initial data available for TAC to analyze is usually limited.

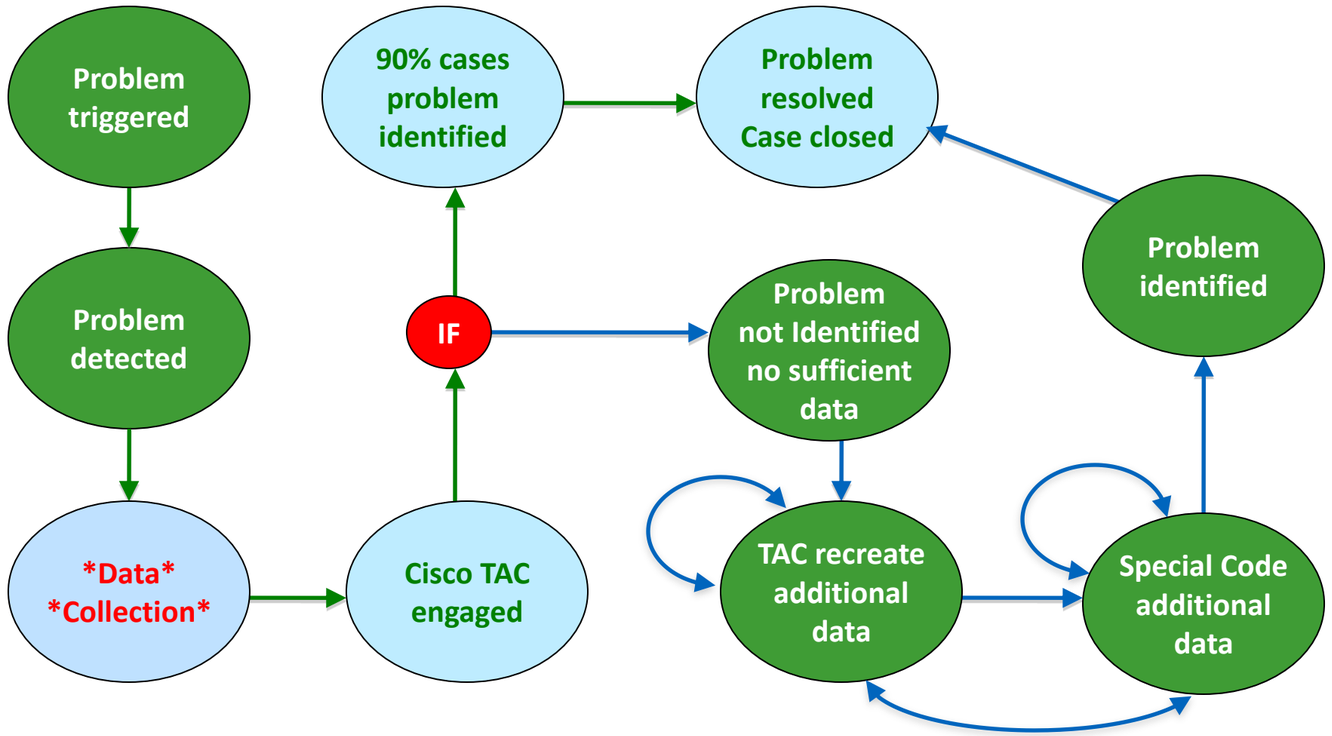


Before You Get Started

NX-OS Approach

Facts

- NX-OS debugging and troubleshooting tools are very rich, and allow engineers to accurately assess the situation
- Customized ‘show techs” make the collection of related information accurate and quick.



Before You Get Started

Traditional Versus NX-OS Troubleshooting Approach (Cont.)

Suggestions

- Identify detection and trigger time as accurately as possible to set ‘good’ start up point for collected data search and analysis
- Minimize delta time between trigger/detection time and data collection time
- Try to recall all activities before trigger/detection time
- Get proficient as much as possible with built-in tool box
- Get familiar with specific feature troubleshooting cli, feature show tech-support output for on-the-fly troubleshooting and analysis

Remember ..

- Internal data logs have limited size, adjust them ahead of time for relevant features you have deployed
- Even max-ed log size may not prevent data wrap up
- Use configuration rollback or other configuration backup method while troubleshooting and making configuration changes
- Forensic data survives reload or switchover via ‘Onboard logging’, ‘accounting-log’, ‘nvram’

Agenda

- Before You Get Started
  - Traditional Versus NX-OS Troubleshooting Approach
  - Nexus 7000 Built-in Troubleshooting Tools
  - Architecture Overview
- Troubleshooting
  - CPU
  - Control Plane
  - Memory Utilization
  - ACL
  - vPC
  - Unicast Layer 2 and Layer 3 Forwarding and ARP
  - Multicast Layer 2 and Layer 3 Forwarding

Built-In Troubleshooting Tools

Make Troubleshooting Easier and more Effective — Almost Fun to Do ☺

Powerful show cli

Standard CLI:

- Platform independent (PI) and dependent (PD) output
- Hardware keyword indicates platform hardware specific output

Engineering CLI

- Internal keyword
- No XML or SNMP support

Event-history logging

- Extensive feature and software component event-history logging
- Permanent engineering debugs output of process Finite State Machine (FSM)

Logflash logging

- Extensive system activity logging to dedicated logflash with filtering to display only ‘what I want to see

Built-In Troubleshooting Tools

Make Troubleshooting Easier and more Effective — Almost Fun to Do ☺

Onboard & Accounting

Onboard logging, accounting log logging (config and exec)

- Forensic data surviving reload and switchover
- Hardware component events and manipulation activity
- Use it to ‘recall’ all activity around ‘trigger and detection’ time

GOLD system

- A diagnostic framework to detect hardware failures while the system is online and operational
- Test types:
  - Bootup
  - Health Monitoring
  - On-demand
  - Scheduled

Standard tools

- Ping, Traceroute
- Span, Netflow, XML, EEM
- Build in Linux tools e.g. grep, egrep, last, less, sed, wc, sort, diff, redirect, exclude, include, pipe etc

Built-In Troubleshooting Tools

Make Troubleshooting Easier and more Effective — Almost Fun to Do ☺

Debugs

- Traditional feature related debugs e.g.

debug ip packet protocol igmp , debug ipv6 icmp, debug icmp

- NX-OS debugs with debug-filter, e.g.

debug-filter ip packet direction inbound

ASIC info

- Easy to read asic counters and registers
- Software copy not clear-on-read, must use clear cli to clear them
- Comprehensive per module, ASIC, port, counter category filtering

ELAM & Ethanalyzer

- Embedded Logic Analyzer Module (ELAM capture) provides detailed frame's internal header info
- Built-in wireshark analyzer capturing mgmt interface and CPU traffic. The output can be redirected to a text file with no performance impact

Agenda

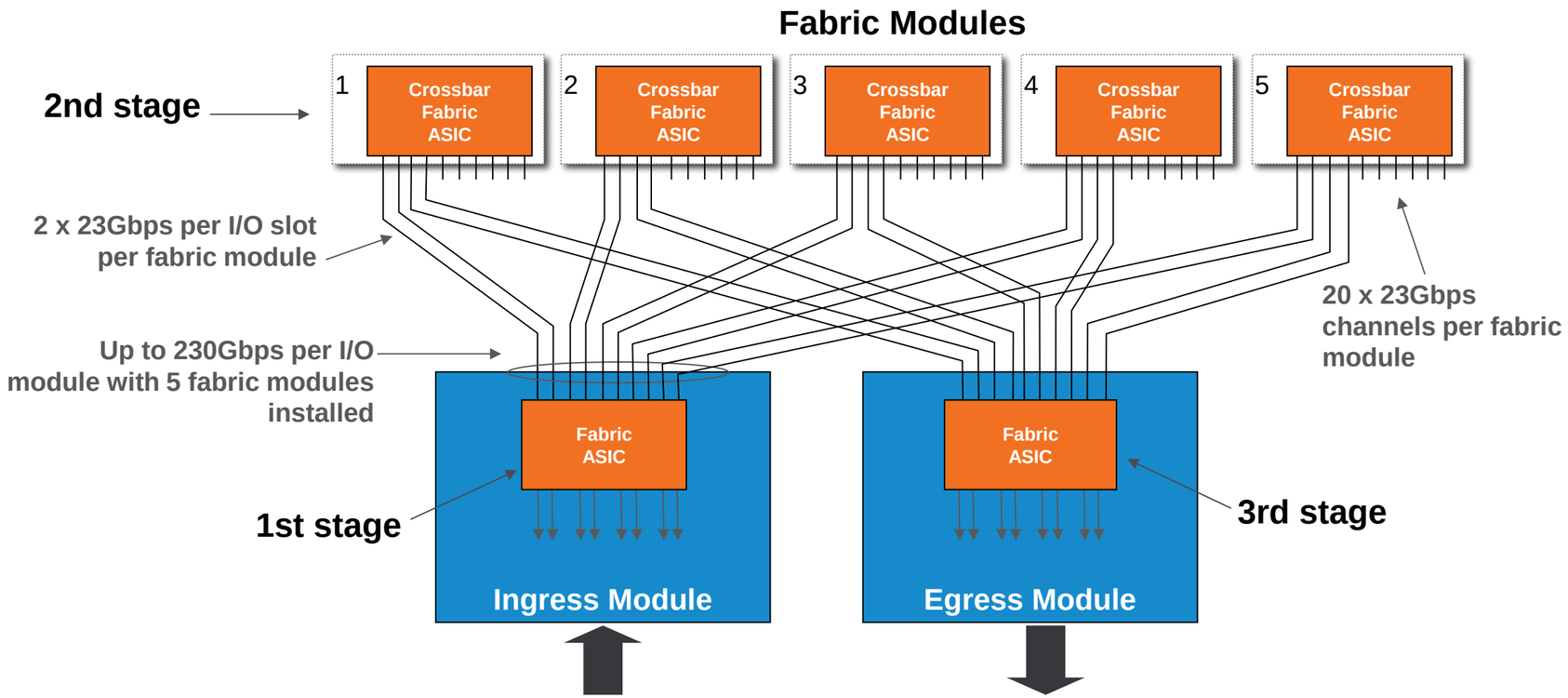
- Before You Get Started
  - Traditional Versus NX-OS Troubleshooting Approach
  - Nexus 7000 Built-in Troubleshooting Tools
  - Architecture Overview
- Troubleshooting
  - CPU
  - Control Plane
  - Memory Utilization
  - ACL
  - vPC
  - Unicast Layer 2 and Layer 3 Forwarding and ARP
  - Multicast Layer 2 and Layer 3 Forwarding

Built-In Troubleshooting Tools

System Architecture — Multistage Switch Fabric

Facts

- Nexus 7000 implements 3-stage switch fabric
- Stages 1 and 3 on I/O modules
- Stage 2 on xbar modules

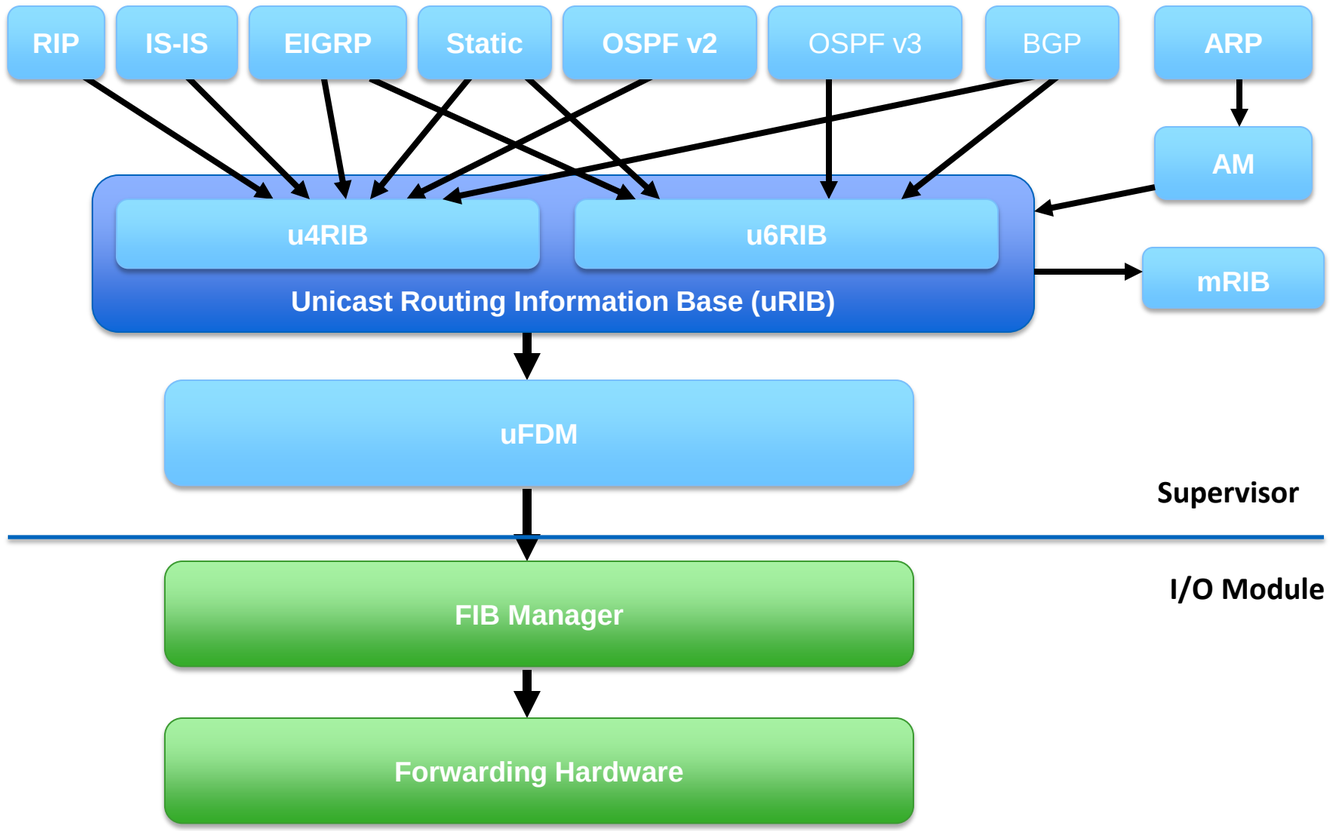


Built-In Troubleshooting Tools

Architecture — Unicast Routing Software Architecture

Facts

- uRIB digests all routing related information and builds the final routing table.
- Unicast Forwarding Distribution Module (UFDM) distributes forwarding information to Modules.
- FIB programs forwarding info on Modules.

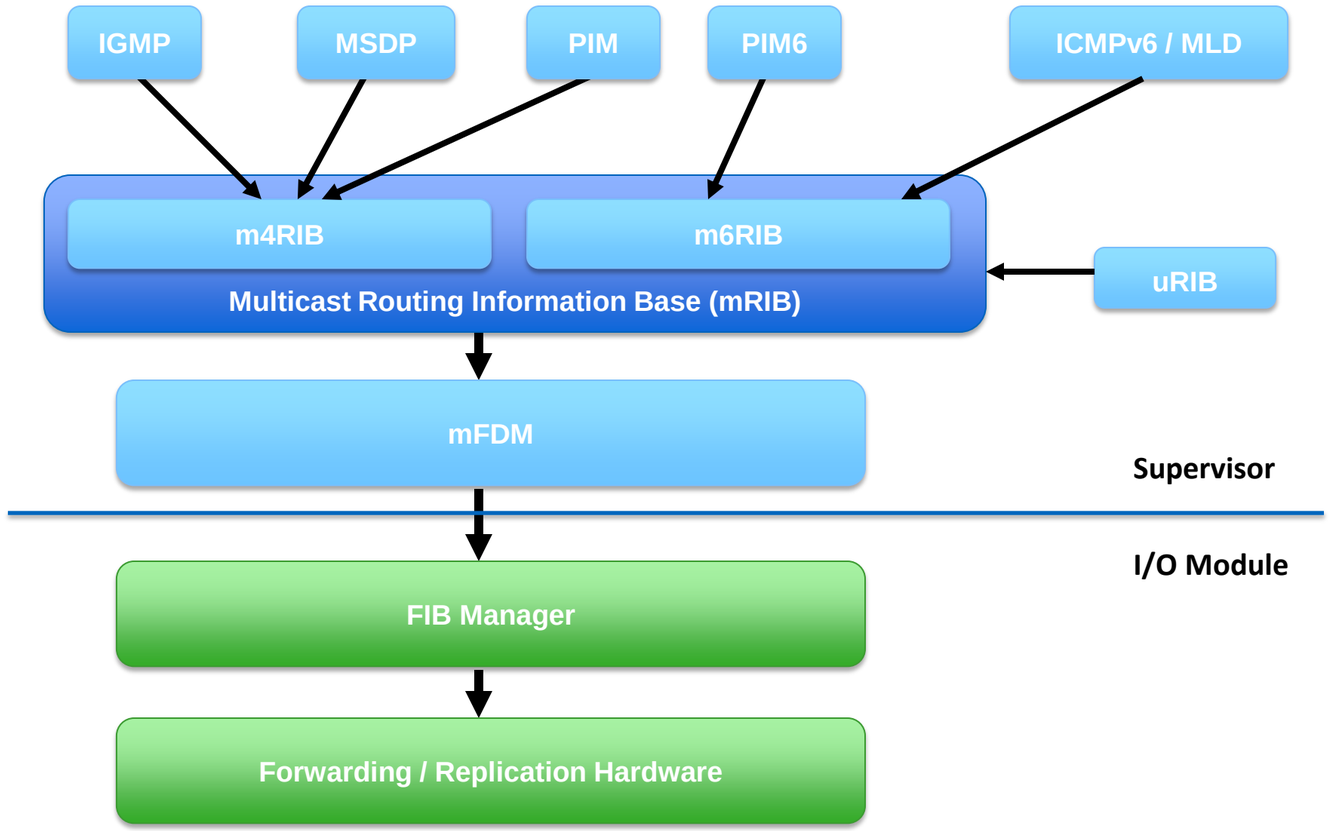


Built-In Troubleshooting Tools

Architecture — Multicast Routing Software Architecture

Facts

- mRIB adds routes, OIFs and handles updates when RPF changes
- mFDM distributes forwarding information to Modules.
- FIB programs forwarding info and MET tables on Modules

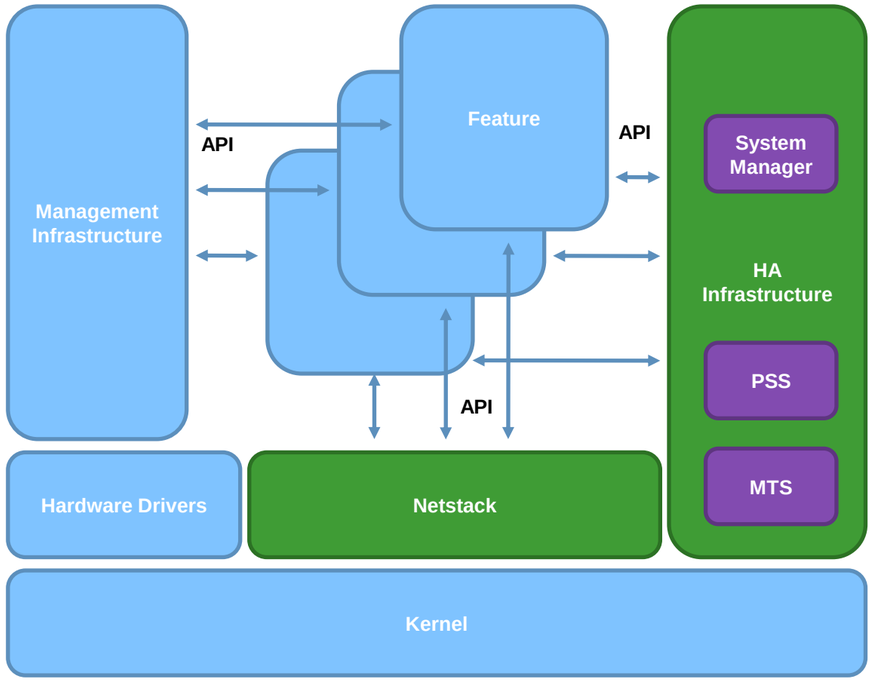


Built-In Troubleshooting Tools

Architecture — NXOS Software Architecture

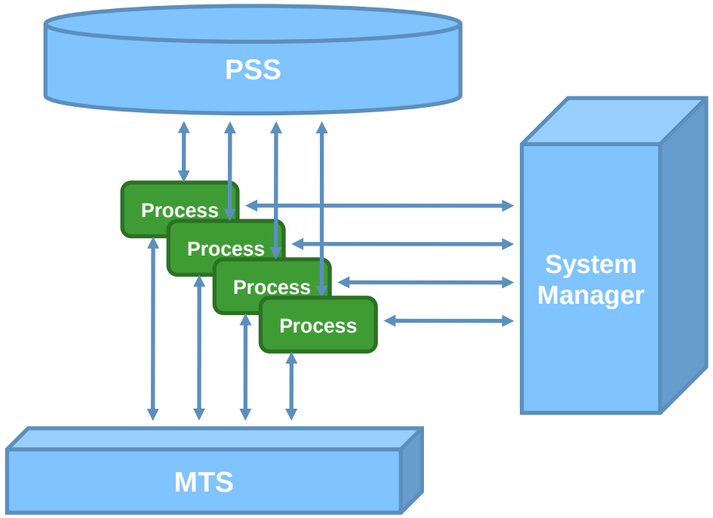
Facts

- NX-OS runs on top of a modified Linux Kernel.
- Rich CLI, debugs, and event-history logs reveal detailed operation information for each software component.



Supervisor Software Architecture

Processes interaction



Agenda

- Before You Get Started
  - Traditional Versus NX-OS Troubleshooting Approach
  - Nexus 7000 Built-in Troubleshooting Tools
  - Architecture Overview
- Troubleshooting
  - CPU
  - Control Plane
  - Memory Utilization
  - ACL
  - vPC
  - Unicast Layer 2 and Layer 3 Forwarding and ARP
  - Multicast Layer 2 and Layer 3 Forwarding

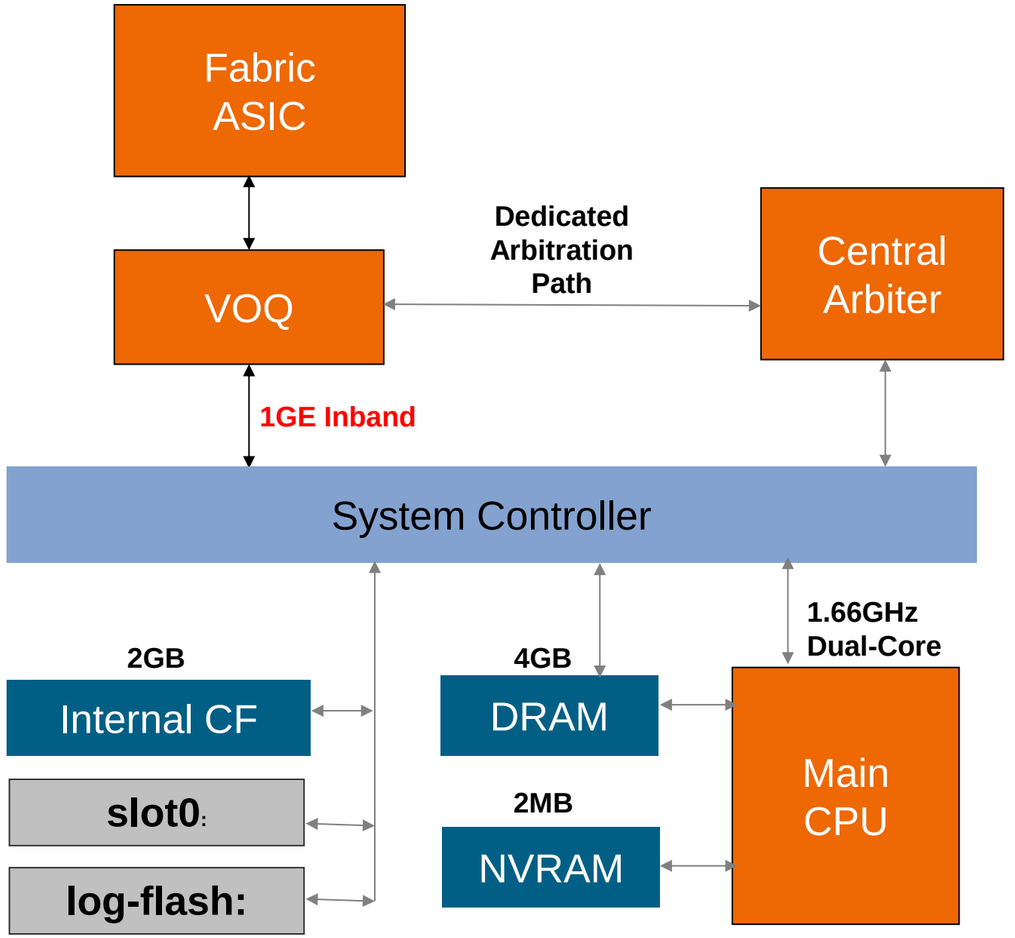
Troubleshooting

CPU — Is there a Problem?

Should I Panic?

High CPU utilization is not automatically problem indication!

- NEXUS 7000 is dual core linux based system with robust preemptive scheduler (one functional unit for both rp and sp)
- Strict control-plane and data-plane separation
- Scheduler assures fair access to CPU for all processes
- Lower level processes (drivers) run in FIFO or non-preemptive mode



Troubleshooting

CPU — Causes

Process

Misbehaving process(s)

- Consume CPU cycles which impact normally-functioning processes
- Delay or prevent CPU from processing control traffic
- Usually triggered by a software bug, but it might be a product of a network event

Traffic

Unexpected traffic

- Excessive CPU bound traffic, control-plane churn
- Acess-list processing, hardware programming
- Possible typical data center traffic (arp, ipv6 nd, etc)

Troubleshooting

CPU — Supervisor, General Health Check

```
N7k-3-VDC3# show system resources
Load average:   1 minute: 0.64    5 minutes: 1.08    15 minutes: 1.30
Processes    :   3912 total, 2 running
CPU states   :   4.5% user,    5.0% kernel,    90.5% idle
Memory usage:   4115232K total,    3434268K used,     680964K free
```

How many processes were scheduled to run in average per whole system in last 1, 5 and 15 minutes

```
N7k-3-VDC3# show processes cpu history

     1 2 111 11111211233 1 1 111 1 1 1 6 112 1 1 21132 1 111 123
    919275058862141899918384800583739174756080779143297264026770
100
 90
 80
 70
 60                                  #
 50                                  #
 40                    #             #
 30                   ###            #            ##         ##
 20    # # # ##   ######       #      # ###   # #  ##       ###
 10 ##################################################################
    0....5....1....1....2....2....3....3....4....4....5....5....
               0    5    0    5    0    5    0    5    0    5

                CPU% per second (last 60 seconds)
                       # = average CPU%
```

How much of CPU cycles are used by user configured processes and kernel processes
Output IS calibrated for 2 cores

CPU utilization 60 seconds ago

Troubleshooting

CPU — Identify the offending process(s)

```
N7K-3-VDC3# show system internal processes cpu
top - 14:01:06 up 21 days, 15:35,  4 users,  load average: 0.77, 0.73, 1.07
Tasks: 3257 total,   1 running, 422 sleeping,   0 stopped, 2834 zombie
Cpu(s):  5.8%us,  6.0%sy,  0.1%ni, 84.1%id,  0.4%wa,  0.1%hi,  3.4%si,
0.0%st
Mem:   4115232k total,  3875988k used,   239244k free,    82400k buffers
Swap:        0k total,        0k used,        0k free,  1817776k cached


PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
22683 root      20   0  182m  63m  14m S 93.7  1.6 636:17.84 netstack
29391 admin#03  20   0  5364 3312 1140 R 22.3  0.1   0:00.30 top
 3892 root      20   0  164m  54m  23m S  6.0  1.3   1095:00 netstack
 4149 root      20   0  111m  41m  19m S  4.5  1.0 994:43.26 stp
 3175 root      20   0 78100  19m  17m S  3.0  0.5 175:07.02 diagmgr
23028 root      20   0  101m  23m 9968 S  3.0  0.6 598:14.57 stp
 3181 root      20   0 77684 4564 3352 S  1.5  0.1   0:30.35 securityd
 3591 root      20   0  222m  13m 7132 S  1.5  0.3   0:09.61 igmp
 4753 root      20   0  162m  45m  16m S  1.5  1.1  34:59.22 netstack
    1 root      20   0  1988  612  532 S  0.0  0.0   0:16.32 init
```

Use X | no-more, where X is interval in seconds to get more snapshots

- Equivalent of Linux TOP monitoring tool output showing system processes across all vDCs
- Use it to cross check accuracy of ‘show system resources’ output
- Output is NOT calibrated for 2 cores so it would be expected to see 2 processes using 100% CPU
- Output show processes from all vDCs

Troubleshooting

CPU — Examine the offending process(s)

```
N7K-3-VDC3# show processes cpu | egrep "PID|--|ospf"
PID     Runtime(ms)   Invoked    uSecs  1Sec     Process
-----   -----------   --------   -----  ------   -----------
 9337           102         72   1418    0.0%   ospfv3
22916           118         62   1905   13.1%   ospf
```

```
N7K-3-VDC3# show system internal sysmgr service pid 22916
Service "__inst_001__ospf" ("ospf", 58):
        UUID = 0x41000119, PID = 22916, SAP = 320
        State: SRV_STATE_HANDSHAKED (entered at time Thu Mar  3 21:53:59 2012).
        Restart count: 1
        Time of last restart: Thu Mar  3 21:53:58 2011.
        The service never crashed since the last reboot.
        Tag = 6467
        Plugin ID: 1
```

```
N7K-3-VDC3# show system internal sysmgr service name ospfv3 tag 8893
Service "__inst_001__ospfv3" ("ospfv3", 59):
        UUID = 0x4100011A, PID = 9337, SAP = 328
        State: SRV_STATE_HANDSHAKED (entered at time Fri Mar 25 22:33:10 2012).
        Restart count: 2
        Time of last restart: Fri Mar 25 22:33:09 2011.
        The service never crashed since the last reboot.
        Tag = 8893
        Plugin ID: 1
```

PID – Process ID

Runtime – total non-idle time process has been actively using CPU

Invoked – number of times process has been context switched voluntary (finished job) and involuntary (scheduler interrupt)

uSecs - average amount of time process was running during a single context switch

Useful process level details

For testing purposes, process was manually restarted using ‘restart ospfv3 8893’ cli

Troubleshooting

CPU — Traffic Causes High CPU Utilization and Control-Plane Instability

Attackers

Typical “offending” datacenter

- ARP, ND (IPv6)
- DHCP traffic
- Glean traffic (no ARP or ND)
- Malicious traffic to 224.0.0.0/24 subnet
- Fragments or malicious L2 mcast or ‘other’ traffic

Remember:
misbehaving “expected” traffic, such as OSPF packets, might be a dangerous attacker as well

Defense

- CPU protection via CoPP policers
- CPU protection via L2/L3 hardware rate-limiters (RL)
- CoPP and RL default settings may need tweaking based on network requirement specifics
  - Both are configured/enabled per M1 I/O Module
  - Total inband traffic allowed is the sum across all M1 I/O Modules

Troubleshooting

CPU — Traffic Causes High CPU Utilization and Control-Plane Instability

Problem

OSPF neighbor's failing to come up.

- Syslog messages report OSPF neighbor failures
- CPU states show high utilization caused by OSPF and Netstack process.

```
N7K-1-VDC2# show system resources
Load average:   1 minute: 2.92   5 minutes: 2.38   15 minutes: 2.27
Processes    :   1267 total, 4 running
CPU states   :   34.0% user,   42.5% kernel,   23.5% idle
Memory usage:   4115232K total,   3638780K used,    476452K free

N7K-1-VDC2# show processes cpu sort

PID    Runtime(ms)   Invoked   uSecs  1Sec    Process
-----  -----------  --------  -----  ------  -----------
 3981          127       276    462   43.2%  ospf
 3841          267        78   3427   16.4%  netstack
 2941     34146488   7377876   4628    0.9%  platform
 3982          118       245    485    0.9%  ospfv3

2011 Mar 26 15:38:56.395 N7K-1-VDC2 %OSPF-5-NBRSTATE:  ospf-6467 [3981]  Process
6467, Nbr 192.251.19.22 on Vlan19 from INIT to DOWN, DEADTIME
2011 Mar 26 15:38:56.584 N7K-1-VDC2 %OSPF-5-NBRSTATE:  ospf-6467 [3981]  Process
6467, Nbr 192.251.19.22 on Vlan19 from DOWN to INIT, HELLORCVD
2011 Mar 26 15:39:33.865 N7K-1-VDC2 %OSPF-5-NBRSTATE: ospf-6467  [3981]  Process
6467, Nbr 192.251.19.22 on Vlan19 from INIT to DOWN, DEADTIME
2011 Mar 26 15:39:35.754 N7K-1-VDC2 %OSPF-5-NBRSTATE:  ospf-6467 [3981]  Process
6467, Nbr 192.251.19.22 on Vlan19 from DOWN to INIT, HELLORCVD
```

Troubleshooting

CPU Traffic — Inband stats

```
N7K-1# show hardware internal cpu-mac inband stats | egrep "  Rx|
Tx|counters|Throttle|Tick|rate|total|good|XOFF p|XON p"
RMON counters                              Rx                   Tx
total packets                       779905245           1421785114
good packets                        779905245           1421650279
good octets (hi)                            0                    0
good octets (low)              172303021767         192965708376
total octets (hi)                           0                    0
total octets (low)             172302724342         192974265660
XON packets                                 0                67627
XOFF packets                                0                67208
Interrupt counters
Error counters
Throttle statistics
Throttle interval ..... 2 * 100ms
Packet rate limit ..... 32000 pps
Tick counter ..... 12414130
Rx packet rate (current/max)  4993 / 20296 pps
Tx packet rate (current/max)  60 / 3474 pps
--snip--
```

The Challenge

how to identify offending traffic type and its source

Total number of frames received and send by CPU

Hard coded maximum limit, with larger packet size, this number may not be reached

How many times did throttling kicked in

CPU bound traffic current pps /maximum pps reached

Troubleshooting

CPU Traffic — Pktmgr debugs

```
N7K-1-VDC2# show system internal pktmgr interface vlan 64
Vlan64, ordinal: 117
  SUP-traffic statistics: (sent/received)
    Packets: 3771848 / 40687558
    Bytes: 304360445 / 36018498390
    Instant packet rate: 0 pps / 4951 pps
    -- snip --
```

Use this cli first without specific interface to identify the 'offending' traffic - the one with the highest rate. Alternatively, use 'show system internal pktmgr internal vdc inband' which identifies vDC interfaces and number of packet sent to the CPU

```
N7K-1-VDC2# debug pktmgr frame
2011 Mar 26 21:22:30.599670 netstack: In  Vlan 64 0x0800 992  7
0000.1301.1301 -> 0100.5e00.0005 Vlan64
```

debug-filter pktmgr vlan 64

Offending host mac

```
N7K-1-VDC2# show ip arp vlan 64 | i 0000.1301.1301

N7K-1-VDC2# show mac address-table address 0000.1301.1301 vlan 64
VLAN     MAC Address       Type       age     Secure NTFY Ports/SWID.SSID.LID
---------+-----------------+--------+---------+------+----+------------------
-
64       0000.1301.1301    dynamic    0          F     F  Eth2/9
```

No ARP entry??

Source Port

Troubleshooting

CPU Traffic — Other Capture methods

Debug the offending process

```
N7K-1-VDC2# debug-filter ip ospf interface vlan 64
N7K-1-VDC2# debug logfile offending_traffic
N7K-1-VDC2# show debug logfile offending_traffic

2011 Mar 26 23:33:25.992586 ospf: 6467 [3981]
(default) rcvd: prty:7 ver:2 t:HELLO len:44
rid:0.0.0.0 area:0.0.0.0 crc:0xfdd2 aut:0 aukid:0 from
192.253.64.254/Vlan64
2011 Mar 26 23:33:25.992780 ospf: 6467 [3981] Invalid
src address 192.253.64.254, should not be seen on
Vlan64
```

Ethanalyzer

- Ethanalyzer can be used to capture the traffic that is taking the inband interface to the CPU.
- Write the capture output to a pcap file and open it using wireshark for analysis
- If still more digging is needed, use a more specific trigger to narrow down the search offending host (s)

Agenda

- Before You Get Started
  - Traditional Versus NX-OS Troubleshooting Approach
  - Nexus 7000 Built-in Troubleshooting Tools
  - Architecture Overview
- Troubleshooting
  - CPU
  - Control Plane
  - Memory Utilization
  - ACL
  - vPC
  - Unicast Layer 2 and Layer 3 Forwarding and ARP
  - Multicast Layer 2 and Layer 3 Forwarding

Troubleshooting

CoPP — Essentials

Goal

CoPP protects the SUP against the following classes of traffic

- Control Plane packets, such as Protocols Hellos and other Receives
- Data Plan transit packets, such as Glean, Exceptions, and Redirects
- Management Plane packets, such as SNMP, and SSH

Operation

- NX-OS device segregates different packets destined to the inband interface into different classes.
- Once these classes are identified, the NX-OS device polices or marks down packets, which ensure that the supervisor module is not overwhelmed.
- CoPP policer is attached to the interface “control-plane”

Implementation

CoPP Policing is implemented on each forwarding engine independently:

- the configured policer’s values apply on a per forwarding engine basis and the aggregate traffic prone to hit the CPU is the sum of the conformed/transmit traffic on all of the forwarding engines
- CoPP can be modified from the default VDC only.

Troubleshooting

CoPP — Tighten the grip on Received packets (OSPF example)

Problem

Flapping OSPF neighbors!!

- A faulty OSPF neighbor or an offending server is blasting the switch with Hello packets.
- Default CoPP is rate-limiting as designed, but that results on dropping legitimate neighbors packets as well.

```
N7K-1# show policy-map interface control-plane module 2 | egrep "service-
policy|critical|ospf|police cir 39600|malicious"
  service-policy  input: copp-system-policy
    class-map copp-system-class-critical (match-any)
      match access-grp name copp-system-acl-ospf
      match access-grp name copp-system-acl-ospf6
      police cir 39600 kbps , bc 250 ms

N7K-1# show class-map type control-plane copp-system-class-critical | egrep
class|ospf
    class-map type control-plane match-any copp-system-class-critical
      match access-grp name copp-system-acl-ospf
      match access-grp name copp-system-acl-ospf6

N7K-1# show ip access-lists copp-system-acl-ospf
IP access list copp-system-acl-ospf
        10 permit ospf any any
```

No “malicious” class to block malicious traffic

Troubleshooting

CoPP — Tighten the grip on Received packets (OSPF example) Cont.

Modify

copp-system-acl-ospf
to permit the neighbors only

```
N7K-1# show ip access-lists copp-system-acl-ospf
IP access list copp-system-acl-ospf
        10 permit ospf any any
        20 permit ip 40.9.0.0/16 224.0.0.5/32
        30 permit ip 40.9.0.0/16 224.0.0.6/32
```

Remove

Add neighbors

Create

copp-system- acl-malicious
access-list

```
N7K-1# show ip access-lists copp-system-acl-malicious
IP access list copp-system-acl-malicious
        10 permit ip any 224.0.0.0/24
```

Add

copp-system-class-malicious class, right before the last class default, with zero-rate policer to block all malicious traffic.

```
N7K-1# show policy-map interface control-plane module 2 | egrep
"service-policy|critical|ospf|police cir 39600|malicious|police cir 1 "
  service-policy  input: copp-system-policy
    class-map copp-system-class-critical (match-any)
      match access-grp name copp-system-acl-ospf
      match access-grp name copp-system-acl-ospf6
      police cir 39600 kbps , bc 250 ms
    class-map copp-system-class-malicious (match-any)
      match access-grp name copp-system-acl-malicious
      police cir 1 bps , bc 200 ms
```

Troubleshooting

CoPP — Tighten the grip on Received packets (OSPF example) Cont.

Verify

Check the CoPP policer for drops

- The new class-map shows high rate of dropped packets.
- Furthermore, the statistics results point to the module where the offending device is connected .

```
N7K-1# show policy-map interface control-plane module 2 class copp-system-class-
  malicious

control Plane
  service-policy  input: copp-system-policy
    class-map copp-system-class-malicious (match-any)
      match access-grp name copp-system-acl-malicious
      police cir 1 bps , bc 200 ms
      module 2 :
        conformed 0 bytes; action: drop
        violated 1799505072 bytes; action: drop

N7K-1# show policy-map interface control-plane module 1 class copp-system-class-
  malicious

control Plane
  service-policy  input: copp-system-policy

    class-map copp-system-class-malicious (match-any)
      match access-grp name copp-system-acl-malicious
      police cir 1 bps , bc 200 ms
      module 1 :
        conformed 0 bytes; action: drop
        violated 0 bytes; action: drop
```

Troubleshooting

Control Plan — Hardware Rate-limiters

Essentials

- Rate-limiters can prevent redirected packets for egress exceptions from overwhelming the supervisor module
- As with CoPP policers, modifying the default rates should be carefully planned before any configuration changes.

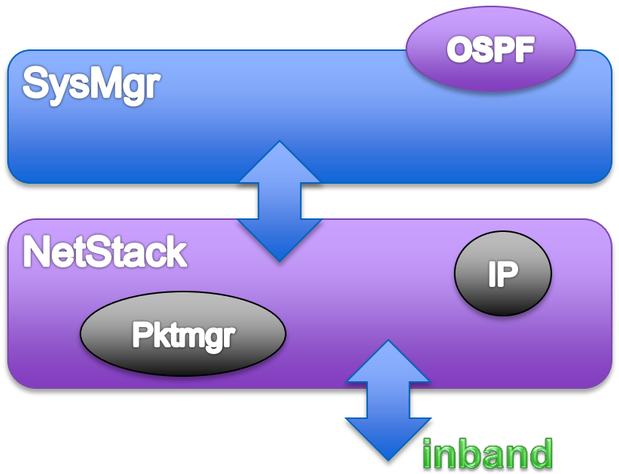
```
N7K-1# show hardware rate-limiter ?
  [snip]
  access-list-log  Packets copied to supervisor for access-list logging
  copy             Data and control packets copied to supervisor
  f1               Control packets from F1 modules to supervisor
  layer-2          Layer-2 control and Bridged packets
  layer-3          Layer-3 control and Routed packets
  module           Optionally specify a module number
  receive          Packets redirected to supervisor
  |                Pipe command output to filter

N7K-1# show hardware rate-limiter layer-2 mcast-snooping module 1
Units for Config: packets per second
Allowed, Dropped & Total: aggregated since last clear counters
Rate Limiter Class                          Parameters
-----------------------------------------------------------
layer-2 mcast-snooping                      Config    : 1500
                                            Allowed   : 302128
                                            Dropped   : 0
                                            Total     : 302128
```

Troubleshooting

Control Plan — Verifying Software Services health (OSPF example)

Sysmgr

The System Manager handles processes and monitors their health. It keeps the mapping of PIDs to UUIDs.




```
N7K-1-PeerA# show system internal sysmgr service name ospf
Service "__inst_001__ospf" ("ospf", 14):
        UUID = 0x41000119, PID = 3725, SAP = 320
        State: SRV_STATE_HANDSHAKED (entered at time Wed Mar 14 15:47:34 2012).
        Restart count: 1
        Time of last restart: Wed Mar 14 15:47:33 2012.
        The service never crashed since the last reboot.
        Tag = 1
        Plugin ID: 1

N7K-1-PeerA# show system internal sysmgr service all | egrep -i netstack|name
Name                        UUID      PID     SAP   state  Start count      Tag  Plugin ID
netstack                0x00000221     5588    246   s0009          1        N/A           0
```

NOTE

Remember this:

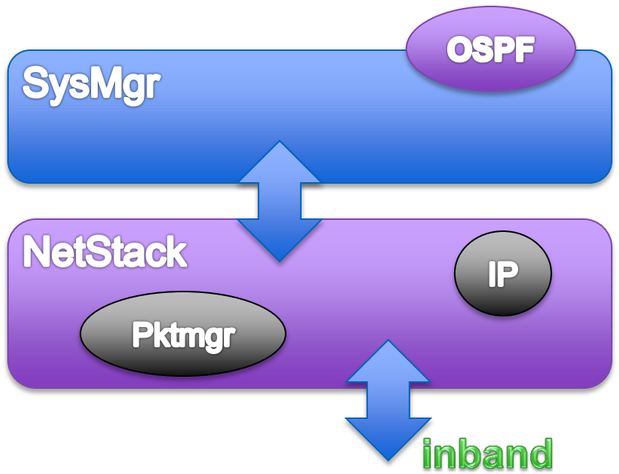
SAP = 320

Troubleshooting

Control Plan — Verifying Software Services health (OSPF example) Cont.

NetStack

Netstack is a full Network Stack designed with Modularity, High availability, and Virtualization implementation goals.




```
N7K-1-PeerA# show ip client ospf
Client: ospf-6467, uuid: 1090519321, pid: 3981, extended pid: 3981
 Protocol: 89, client-index: 19, routing VRF id: 65535
 Data MTS-SAP: 2339
 Data messages, send successful: 209867328, failed: 13263152

N7K-1-PeerA# show system internal pktmgr client 0x221
Client uuid: 545, 4 filters, pid 3841
 Filter 1: EthType 0x0800,
 Rx: 299923608, Drop: 0
 Filter 2: EthType 0x86dd,
 Rx: 1412579, Drop: 0
 [snip]
 Total Rx: 301346464, Drop: 0, Tx: 144295338, Drop: 0
 COS=0 Rx: 15993531, Tx: 87699456    COS=1 Rx: 1903980, Tx: 0
 COS=2 Rx: 0, Tx: 0    COS=3 Rx: 0, Tx: 0
 COS=4 Rx: 0, Tx: 0    COS=5 Rx: 3694169, Tx: 1
 COS=6 Rx: 56191519, Tx: 56595881    COS=7 Rx: 223563265, Tx: 0
```

Check for OSPF IP client failures

Check for L2 client packet drops

Troubleshooting

Control Plan — Verifying Software Services health (OSPF example) Cont.

MTS

- "Messages and Transactional Services". MTS offers SAPs (Service Access Points) to allow services to exchange messages
- MTS provides complete fault isolation by handling data structure communications.

```
N7K-1-PeerA# show system internal mts sup sap 320 stats
msg  tx: 3328
byte tx: 396657
msg  rx: 527
byte rx: 65045
opc sent to myself: 8927
max_q_size q_len limit (soft q limit): 1024
max_q_size q_bytes limit (soft q limit): 15%
max_q_size ever reached: 17
max_fast_q_size (hard q limit): 4096
rebind count: 0
Waiting for response: none
buf in transit: 0
bytes in transit: 0
```

Make sure the counters are incrementing (no memory leak)

```
N7k# show system internal mts buffers summary
node    sapno   recv_q  pers_q  npers_q log_q
sup     320     0       0       4592    0
```

npers high value indicates OSPF MTS buffer leak

Unicast L2 and L3 Forwarding, ARP

Control Plan — Golden rule

In case the issue you have encountered is urgent, complicated or you can't figure it out, collect **show tech-support** output asap!

Related show tech(s)

```
N7K-1-VDC2# show tech-support sysmgr
N7K-1-VDC2# show tech-support netstack detail
N7K-1-VDC2# show tech-support pktmgr
N7K-1-VDC2# show tech-support <service>
```

Agenda

- Before You Get Started
  - Traditional Versus NX-OS Troubleshooting Approach
  - Nexus 7000 Built-in Troubleshooting Tools
  - Architecture Overview
- Troubleshooting
  - CPU
  - Control Plane
  - Memory Utilization
  - ACL
  - vPC
  - Unicast Layer 2 and Layer 3 Forwarding and ARP
  - Multicast Layer 2 and Layer 3 Forwarding

Troubleshooting

CoPP — Essentials

Protection

- A single process creates the shared-memory segment and only it can write to it.
- API clients can only READ from the segment.
- If clients inadvertently write to the segment (due to a software bug), the client process will crash, but other processes will be protected.

Memory Locks

- The NX-OS shared-memory locking philosophy is to keep the system robust by having as few locks as possible.
- When a process fully exports a shared-memory to an API client, the client‘s code acquires the write lock since it is the only process that has write permission to the shared-memory segment

Device Drivers

- Device drivers are typically implemented from within the kernel space in Linux
- NX-OS provides a —user space driverll infrastructure to implement device drivers as real-time priority scheduled tasks
- A Crash of such a driver task is isolated within the address space of the driver task, instead of a Kernel crash.

Troubleshooting

Memory Utilization — System Memory

Facts

NX-OS built-in memory monitoring

- From 4.2(4) , the default memory alert thresholds are 85% Minor 90% Severe 95% Critical
- System memory issues affect all vDCs

```
N7K-1# show logging logfile | grep -b 5 -i memory | grep "Mar 22"
2011 Mar 22 15:40:13 N7K-1 %BGP-5-MEMALERT:  bgp-1 [3439]  BGP memory status
changed from OK to Minor Alert
2011 Mar 22 15:40:13 N7K-1 %PLATFORM-2-MEMORY_ALERT: Memory Status Alert : MINOR.
Usage 85% of Available Memory

N7K-1# show system internal memory-status
MemStatus: Minor Alert

N7K-1# show system internal memory-alerts-log
 MINOR ALERT INFO
Tue Mar 22 15:40:13 PDT 2011
***** /proc/memory_events *****
Alert MINOR Reached at 1300833613.000287556
***** /proc/meminfo *****
MemTotal:      4115232 kB
MemFree:        318452 kB
Buffers:         81524 kB
Cached:        1726848 kB
[snip]

N7k-3(config)# system memory-thresholds minor 85 severe 90 critical 95
```

System memory alert threshold can be modified as required

Troubleshooting

Memory Utilization — System Memory General Health Check

```
N7k-3-VDC3# show system internal kernel meminfo
MemTotal:       4115232 kB
MemFree:         263684 kB
Buffers:          82400 kB
Cached:         1817788 kB
ShmFS:          1533324 kB
Allowed:        1028808 Pages
Free:             65921 Pages
Available:       164026 Pages
SwapCached:           0 kB
Active:         2080320 kB
Inactive:       1433752 kB
HighTotal:      3338960 kB
HighFree:          4092 kB
LowTotal:        776272 kB
LowFree:         259592 kB
SwapTotal:            0 kB
SwapFree:             0 kB
Dirty:                0 kB
Writeback:            0 kB
AnonPages:      1613748 kB
Mapped:          456088 kB
Slab:            142884 kB
```

Examine the counters for memory leak indicators

Glossary

- MemTotal - Total amount of memory in the system (4GB in N7K Sup1)
- Cached - Memory used by page cache (tmp fs mounts and data cached from bootflash)
- Available - Amount of free memory in pages (takes into account the space that could be made available in page cache and free lists
- Mapped - Memory mapped into page tables (data being used by non-kernel processes)
- Slab - Rough indication of kernel memory consumption

Troubleshooting

Memory Utilization — System Memory General Health Check

```
N7K-1-VDC2# show system resources
Load average:   1 minute: 0.11   5 minutes: 0.09   15 minutes: 0.14
Processes    :   1241 total, 2 running
CPU states   :   2.0% user,   3.4% kernel,   94.6% idle
Memory usage:   4115232K total,   3606556K used,    508676K free

N7K-1-VDC2# show processes memory | egrep "PID|--|ospf|bgp"
PID     MemAlloc  MemLimit    MemUsed     StackBase/Ptr       Process
-----  --------  ----------  ----------  ------------------  ----------------
 3981  43761664  446487641   247361536   bff070c0/bff06b80  ospf
 3982   9428992  446266867   230895616   bff070c0/bff06b80  ospfv3
 3986  18247680  2411763200  271065088   bfe7a850/bfe7a760  bgp

N7K-1-VDC2# show system internal processes memory | egrep "PID|ospf|bgp"
  PID TTY      STAT     TIME MAJFLT  TRS   RSS    VSZ %MEM COMMAND
 3981 ?        Ssl  11:52:06      0  690 64840 176028  1.5 /isan/bin/routing-sw/ospf -t 6467
 4392 ?        Ssl  02:15:41      0  690 63136 157424  1.5 /isan/bin/routing-sw/ospf -t 6467
 4396 ?        Ssl  00:35:01      0 1460 40856 180744  0.9 /isan/bin/routing-sw/bgp -t 1204
 3986 ?        Ssl  00:37:57      0 1460 39944 199176  0.9 /isan/bin/routing-sw/bgp -t 1203
 3982 ?        Ssl  01:16:17      0  728 22448 159948  0.5 /isan/bin/routing-sw/ospfv3 -t 8893
 4393 ?        Ssl  01:14:42      0  728 21436 141808  0.5 /isan/bin/routing-sw/ospfv3 -t 8893
 3431 ?        Ssl  01:09:00      0  728 15356 173136  0.3 /isan/bin/routing-sw/ospfv3 -t 1
 3430 ?        Ssl  01:08:23      0  690 15144 142376  0.3 /isan/bin/routing-sw/ospf -t 1
 4811 ?        Ssl  01:08:52      0  690 14832 123944  0.3 /isan/bin/routing-sw/ospf -t 1
 3436 ?        Ssl  01:07:37      0  690 14416 141872  0.3 /isan/bin/routing-sw/ospf -t 6467
```

Glossary

- MemAlloc – Data Segment Size
- MemLimit – Max memory process can use set by susmgr
- MemUsed – Virtual Memory
- TRS – Test Resident Set
- RSS – Resident Set Size (physical memory used
- VZS – Virtual Set Size (RSS + swap)

Troubleshooting

Memory Utilization — System Memory Per Process Utilization

```
N7K-1-VDC2# show system internal kernel memory uuid 0x11B
MEMORY TYPE                              TOTAL    RSS    PSS SHARED PRIVATE
bgp                                TEXT   1464   1224   1224   1204      20
bgp                                DATA     24     16     16      0      16
Anonymous                          HEAP   8328   8308   8308      0    8308
ld-2.8.so                          TEXT    104    100    100    100       0
ld-2.8.so                       RO_DATA      4      4      4      0       4
ld-2.8.so                          DATA      4      4      4      0       4
libc-2.8.o                         TEXT   1252    440    440    440       0
[snip]

N7K-1# show system internal pktmgr internal mem-stats detail | grep -b 13 -a 3 TCP_MEM_client_t
Private Mem stats for UUID : Transmission Control Protocol (TCP)(271) Max types:
 21
--------------------------------------------------------------------------------
TYPE NAME                                          ALLOCS                  BYTES
                                                 CURR     MAX        CURR        MAX
   2 TCP_MEM_inpcb                                 18      66        3240      11880
   3 TCP_MEM_socket                                18      66       11160      40920
   4 TCP_MEM_getsockaddr                            0       1           0         40
   5 TCP_MEM_tcp_msg_t                             17      17       14892      14892
   6 TCP_MEM_tseg_qent                              0       1           0         28
   7 TCP_MEM_tcpcb                                  3      51         732      12444
   9 TCP_MEM_sockaddr_in_dcos                       0       1           0         24
  10 TCP_MEM_syncache                               0      33           0       4620
  11 TCP_MEM_syncache_head                          1       1       12296      12296
  12 TCP_MEM_client_t                            4153    4154    71099360   71116480
--------------------------------------------------------------------------------
Total bytes: 71141680 (69474k)
--------------------------------------------------------------------------------
```

BGP UUID. Examine the counters for memory leak

Growing daily.
Symptoms indicate memory leak in TCP_MEM_client

Troubleshooting

Memory Utilization — System Memory , Estimated Utilization

```
N7k-3-VDC3# show routing ip multicast memory estimate groups 200 sources-per-group
16 oifs-per-entry 16
Shared memory estimates:
  Current max       8 MB;     204 groups
                                16 sources-per-group
                                16 oifs-per-entry
  In-use            4 MB;       1 groups
                                 1 sources-per-group (average)
                                 0 oifs-per-entry (average
  Configured max    8 MB;     204 groups
                                16 sources-per-group
                                16 oifs-per-entry
  Estimate          8 MB;     200 groups
                                16 sources-per-group
                                16 oifs-per-entry

N7k-3-VDC3# show routing ip unicast memory estimate routes 180000 next-hops 4
Shared memory estimates:
  Current max       8 MB;    6868 routes with 16 nhs
           in-use   1 MB;     143 routes with  2 nhs (average)
  Configured max    8 MB;    6868 routes with 16 nhs
  Estimate         69 MB; 180000 routes with  4 nhs
```

Useful cli to predict mrib shared memory utilization based on number of multicast groups, sources and output interfaces (oifs)

Useful cli to predict urib shared memory utilization based on number of unicast prefixes and next-hops

Troubleshooting

Memory Utilization — System Memory , Estimated Utilization

```
2010 Jun 12 15:05:13 N7K-1-VDC2%MRIB-3-MALLOC_FAILED:  mrib [6971]  sm_malloc()
failed for mrib_notify_buffer
2010 Jun 12 15:05:23 N7K-1-VDC2 %MRIB-4-SYSLOG_SL_MSG_WARNING:MRIB-3-MALLOC_FAILED:
message repeated 3835 times in last 60 sec

N7K-1-VDC2# show resource

      Resource          Min       Max       Used      Unused    Avail
      --------          ---       ---       ----      ------    -----
      vlan              16        4094      603       0         3491
      monitor-session   0         2         0         0         1
      monitor-session-erspan-dst 0         23        0         0
      vrf               16        200       2         14        198
      port-channel      0         768       2         0         759
      u4route-mem       8         8         1         7         7
      u6route-mem       4         4         1         3         3
      m4route-mem       8         8         8         0         0
      m6route-mem       5         5         1         4         4

N7K-1(config-vdc)# limit-resource m4route-mem minimum 24 maximum 24

N7K-1-VDC2# show resource | egrep "Resource|---|m4route-mem"
      Resource          Min       Max       Used      Unused    Avail
      --------          ---       ---       ----      ------    -----
      m4route-mem       24        24        4         20        20
```

Message indicates that there was lack of shared memory for multicast rib and default setting adjustment was required

Minimum and maximum shared memory allocation must be equal

Switchover, vDC reload or system reload is required to get new shared memory allocation into effect

Agenda

- Before You Get Started
  - Traditional Versus NX-OS Troubleshooting Approach
  - Nexus 7000 Built-in Troubleshooting Tools
  - Architecture Overview
- Troubleshooting
  - CPU
  - Control-Plane – CoPP
  - Memory Utilization
  - ACL
  - vPC
  - Unicast Layer 2 and Layer 3 Forwarding and ARP
  - Multicast Layer 2 and Layer 3 Forwarding

Troubleshooting

ACL — Operation

Characteristic

- Atomic/hitless update of existing applied ACL while modified
  - temporary label swap (no use of default-result)
  - two acl copies in tcam, if there is no enough space, process fails
- ACL tcam banks chaining supported
- L4OPs/LOUs only used for expansion beyond 5 lines, configurable
- 10 L4op per acl limit

```
N7K-1-VDC3# show system internal access-list globals module 1
  Atomic Update : ENABLED
  Default ACL   : PERMIT
  Bank Chaining : DISABLED
  LOU Threshold Value : 5
N7K-1(config)# hardware access-list resource ?
  pooling  Enable ACL programming across TCAM banks

N7K-1(config)# hardware access-list update ?
  atomic          Enable atomic update of access-list in hardware
  default-result  Default access-list result during non-atomic hardware update

N7K-1(config)# hardware access-list lou resource threshold 10

NOTE: Operation in progress, please check the status using
'show hardware access-list lou resource threshold' command
```

TCAM chaining (2x32K TCAMs, 2 banks each)

Disable atomic update if there is not enough space in TCAM

Hidden cli available in 5.1.X code

Troubleshooting

ACL — ingress ACL Hardware configuration

Characteristic

- Ingress ACLs are programmed only to required I/O modules (localization support)
- Egress access-lists are programmed to all I/O modules as they are executed on ingress
- ACL statistics in software must be enabled via configuration

```
N7K-1-VDC3# show run interface ethernet 1/1 | i access
  ip access-group tcp_flags in
  ip access-group test_punt out

N7K-1-VDC3# show hardware access-list interface ethernet 1/1 input config
module 1

          Policy id:  1, Type: QoS, Protocol: IPv4   Name: *
          Policy id:  3, Type: RACL, Protocol: IPv4   Name: tcp_flags

permit tcp 0.0.0.0/0.0.0.0  0.0.0.0/0.0.0.0  syn   log
permit tcp 0.0.0.0/0.0.0.0  0.0.0.0/0.0.0.0  ack   log
permit ip 0.0.0.0/0.0.0.0  0.0.0.0/0.0.0.0
deny ip 0.0.0.0/0.0.0.0  0.0.0.0/0.0.0.0  *

N7K-1-VDC3# show hardware access-list interface ethernet 1/1 input config
module 2
no policy found
```

ACL TCAM on I/O Module 2 does not contain access-lists configured on I/O Module 1 interface

Troubleshooting

Egress ACL Hardware Configuration

Characteristic

- Specific applications (dhcp, bfd) may install their own ACLs which must merge with user configured racl,vacl,pacl
- Some combination of ACL based applications may not be supported
- Data collection: show tech-support aclmgr detail

Both I/O Module 1 and 2 have egress acl configured

```
N7K-1-VDC3# show hardware access-list interface ethernet 1/1 output config
module 1

          Policy id:  2, Type: QoS, Protocol: IPv4   Name: *
          Policy id:  5, Type: RACL, Protocol: IPv4   Name: test_punt

permit udp 172.222.222.64/255.255.255.255  172.31.31.250/255.255.255.255  log
permit icmp 9.9.9.9/255.255.255.255  172.31.31.250/255.255.255.255  log
permit icmp 9.9.9.9/255.255.255.255  14.14.14.14/255.255.255.255  log
permit ip 0.0.0.0/0.0.0.0  0.0.0.0/0.0.0.0
deny ip 0.0.0.0/0.0.0.0  0.0.0.0/0.0.0.0  *

N7K-1-VDC3# show hardware access-list interface ethernet 1/1 output config
module 2

          Policy id:  4, Type: RACL, Protocol: IPv4   Name: test_punt

permit udp 172.222.222.64/255.255.255.255  172.31.31.250/255.255.255.255  log
permit icmp 9.9.9.9/255.255.255.255  172.31.31.250/255.255.255.255  log
permit icmp 9.9.9.9/255.255.255.255  14.14.14.14/255.255.255.255  log
permit ip 0.0.0.0/0.0.0.0  0.0.0.0/0.0.0.0
deny ip 0.0.0.0/0.0.0.0  0.0.0.0/0.0.0.0  *
```

Troubleshooting

ACL — Feature ACLs Merge

```
N7K-1-VDC2# show hardware access-list vlan 33 input statistics module 1
Tcam 1 resource usage:
----------------------
Label_b = 0x3
 Bank 0
 ------
   IPv4 Class
     Policies: DHCP Snooping() BFD()   [Merged]
     Entries:
       [Index] Entry [Stats]
       ---------------------
[0014] redirect(0x43024) udp 0.0.0.0/0 0.0.0.0/0 eq 3785 ttl eq 254    [185050]
[0015] redirect(0x43024) udp 0.0.0.0/0 0.0.0.0/0 eq 3784 ttl eq 255    [5783]
[0016] redirect(0x800) udp 0.0.0.0/0 255.255.255.255/32 eq 68   [0]
[0017] redirect(0x800) udp 0.0.0.0/0 255.255.255.255/32 eq 67   [0]
[0018] redirect(0x800) udp 0.0.0.0/0 eq 68 255.255.255.255/32   [0]
[0019] redirect(0x800) udp 0.0.0.0/0 eq 67 255.255.255.255/32   [0]
[0020] permit ip 0.0.0.0/0 0.0.0.0/0   [240021]

N7K-1-VDC2-CS1# show hardware access-list vl 33 input l4ops module 1
Tcam 1 resource usage:
----------------------
Lou usage:
   Lou     sw_id    l4op_bit    ref_count           Operation
-------------------------------------------------------------------------
   2(A)        0         0          1       IPTTL EQ(255)
   2(B)        1         1          1       IPTTL EQ(254)

TCP flags usage: none
```

Number of packets matching access-list entry (ACE)

BFD acl

DHCP relay agent acl

CPU Inband is not part of BD for IP packets and therefore DHCP has to be caught by ACL to be directed to rp for processing via special ltl index

10 l4op_bits maximum
N7K3-VDC4(config-if)# ip access-group l4optest in
ERROR: l4op bits exhausted in label

Troubleshooting

ACL — Feature ACLs and RACL Merge

```
N7K-1-PeerA# show hardware access-list vlan 33 input statistics module 1
Tcam 1 resource usage:
----------------------
Label_b = 0x8
 Bank 0
 ------
   IPv4 Class
     Policies: RACL(test_lou) DHCP Snooping() BFD()   [Merged]
     Entries:
       [Index] Entry [Stats]
       ---------------------
[0013] permit tcp 1.1.1.0/24 2.2.2.0/24 fragment  [0]
[0014] permit tcp 1.1.1.0/24 2.2.2.0/24 eq 179  [0]
[0015] permit tcp 1.1.1.0/24 eq 179 2.2.2.0/24  [0]
[0016] deny-routed udp 0.0.0.0/0 0.0.0.0/0 range 2000 2300   [0]
[0017] deny-routed tcp 10.0.0.0/8 20.0.0.0/24 range 1500 1900   [0]
[0054] redirect(0x43035) udp 0.0.0.0/0 0.0.0.0/0 eq 3785 ttl eq 254   [152]
[0055] redirect(0x43035) udp 0.0.0.0/0 0.0.0.0/0 eq 3784 ttl eq 255   [3]
[0056] redirect(0x800) udp 0.0.0.0/0 255.255.255.255/32 eq 68  [0]
[0057] redirect(0x800) udp 0.0.0.0/0 255.255.255.255/32 eq 67  [0]
[0058] redirect(0x800) udp 0.0.0.0/0 eq 68 255.255.255.255/32  [0]
[0059] redirect(0x800) udp 0.0.0.0/0 eq 67 255.255.255.255/32  [0]
[0060] permit ip 0.0.0.0/0 0.0.0.0/0  [124]
```

Merging Verification

Use the following commands to verify the ACLs merger on both directions:

- show hardware access-list vlan 33 input merge module 1
- show hardware access-list vlan 33 output merge module 1

Troubleshooting

ACL — Per I/O Module Summary, VDC Wide ACL Summary

```
N7K-1-PeerA# show hardware access-list resource utilization module 1
         ACL Hardware Resource Utilization (Module 1)
         ------------------------------------------
                              Used    Free    Percent
                                              Utilization
-----------------------------------------------------
Tcam 0, Bank 0                5       16379   0.03
Tcam 0, Bank 1                3       16381   0.01
Tcam 1, Bank 0                55      16329   0.33
Tcam 1, Bank 1                151     16233   0.92
LOU                           5       99      4.80
Both LOU Operands             3
Single LOU Operands           2
--snip –
```

```
N7K-1-PeerA# show access-lists summary | egrep -a 5 tcp|lou
IPV4 ACL tcp_zoom
        Total ACEs Configured: 5
        Configured on interfaces:
                port-channel111 - ingress (Router ACL)
        Active on interfaces:
                port-channel111 - ingress (Router ACL)
--
IPV4 ACL test_lou
        Total ACEs Configured: 5
        Configured on interfaces:
                Vlan33 - ingress (Router ACL)
        Active on interfaces:
                Vlan33 - ingress (Router ACL)
```

Usage

Cumulative usage of I/O Module 1 ACL TCAM hardware resources by all type of programmed access-lists

ACL

Golden rule

In case the issue you have encountered is urgent, complicated or you can’t figure it out, collect **show tech-support** output asap!

Related show tech(s)

```
N7K-1-VDC2# show tech-support forwarding L3 multicast
N7K-1-VDC2# show tech-support ip pim
N7K-1-VDC2# show tech-support ip multicast
N7K-1-VDC2# show tech-support igmp brief
N7K-1-VDC2# show tech-support ip igmp snooping
N7K-1-VDC2# show tech-support ip mfwd
N7K-1-VDC2# show tech-support forwarding L2 multicast
```

Agenda

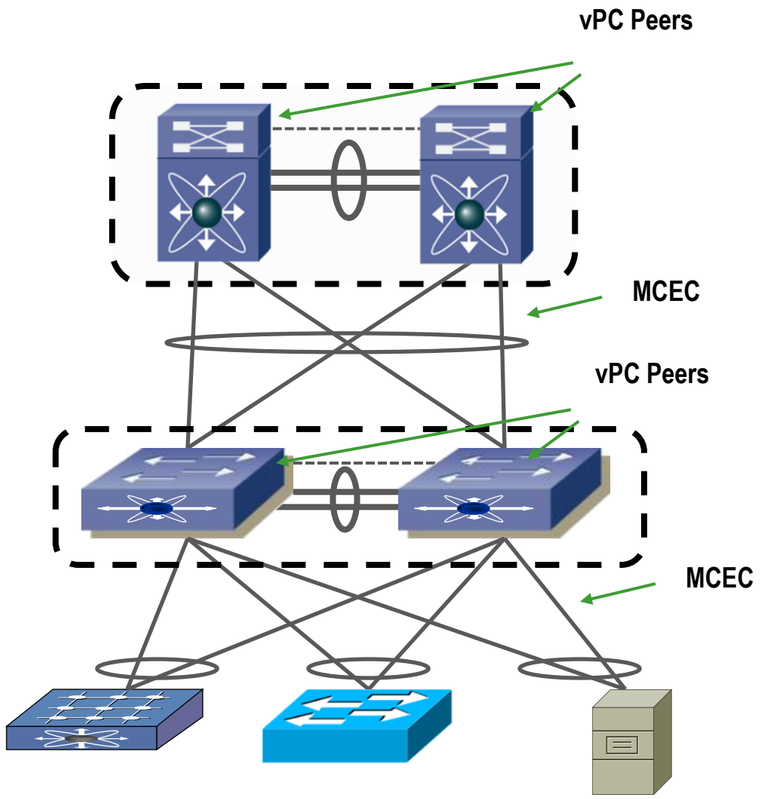
- Before You Get Started
  - Traditional Versus NX-OS Troubleshooting Approach
  - Nexus 7000 Built-in Troubleshooting Tools
  - Nexus 7000 Module and Forwarding Engine Architecture Overview
- Troubleshooting
  - CPU
  - Control Plane
  - Memory Utilization
  - ACL
  - vPC
  - Unicast Layer 2 and Layer 3 Forwarding and ARP
  - Multicast Layer 2 and Layer 3 Forwarding

Troubleshooting

vPC — Essentials

Characteristic

- Dual control-plane
- Eliminates STP blocking ports
- FHRP active/active mode
- Loop-avoidance logic (drop packet received on vPC peer link (PL) and destined to another vPC port-channel, VSL bit set on ingress and checked on egress)
- Cisco Fabric Services (CFS) protocol is used to synchronize configuration and state machines between vpc peers (igmp, pim etc)



Unsupported

- L3 adjacencies between vpc peers and 3rd device behind vpc port-channel connected to L2 switch
- non-default pim/ospf/hsrp timers
- PIM-DM, SSM. PIM bi-dir
- pim spt-threshold infinity

In case your network has any of the above configured, eliminate it before spending any time troubleshooting your network issue.

Troubleshooting

vPC — Essentials

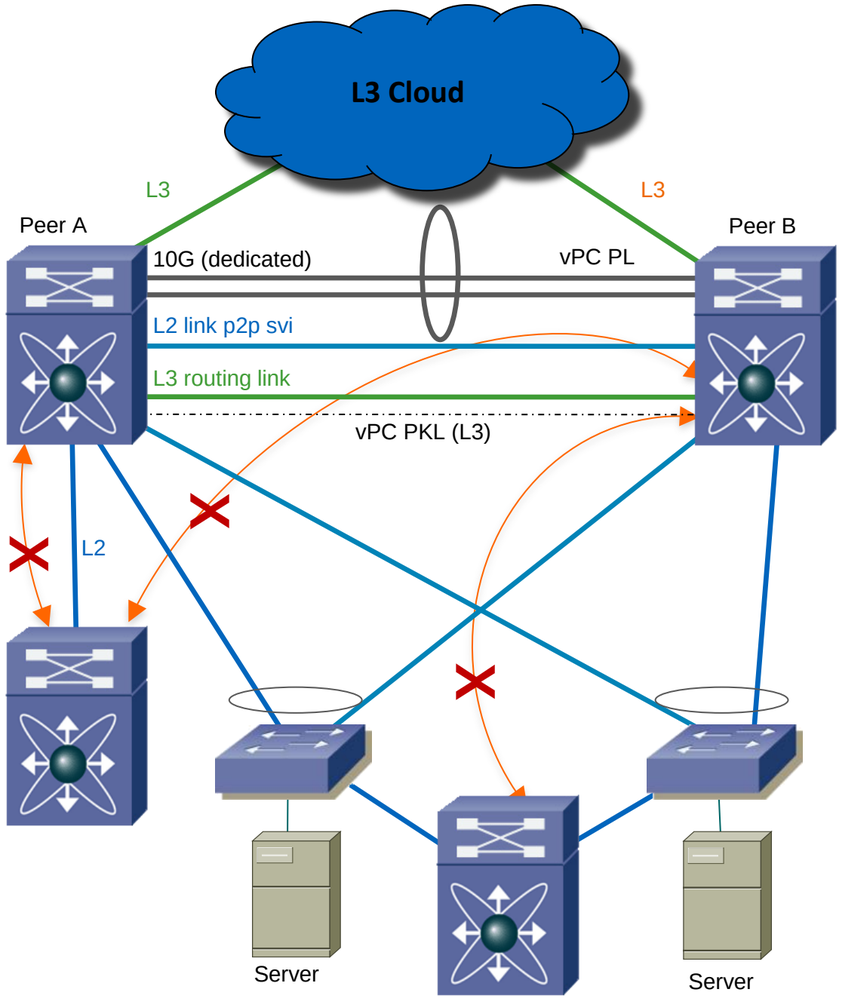
Operation

Generic vPC recommendations

- PL 10G ports (only) in dedicated mode
- Dedicated L3 vPC peer keel-alive (PKL) link
- peer-gateway to accommodate non RFC compliant hosts connected to L2 switch
- peer-gateway exclude <vlan-list> in case vPC PL resides on F1 I/O module
- peer-switch for faster stp convergence (both peers appear to be roots for rest of L2 topology)

Routing vPC recommendations

- Dedicated L3 link between vPC peers or
- Dedicated L2 link between vPC peers with p2p svi interfaces or
- Dedicated vlan carried on vPC PL and not extended to vPC connected L2 switch with p2p svi interfaces
- ip pim pre-build-spt for faster multicast failover



Troubleshooting

vPC — General Health Check

```
PeerB# show vpc brief
Legend:
                (*) - local vPC is down, forwarding via vPC peer-link
vPC domain id                     : 64
Peer status                       : peer adjacency formed ok
vPC keep-alive status             : peer is alive
Configuration consistency status: success
Type-2 consistency status         : failed
Type-2 consistency reason         : SVI type-2 configuration incompatible
vPC role                          : primary
Number of vPCs configured         : 2
Peer Gateway                      : Disabled
Peer gateway excluded VLANs       : -
Dual-active excluded VLANs        : -
vPC Peer-link status
---------------------------------------------------------------------
id   Port   Status Active vlans
--   ----   ------ --------------------------------------------------
1    Po664  up     1,19,31-35,2000,4092-4093

vPC status
----------------------------------------------------------------------------
id   Port   Status Consistency Reason                     Active vlans
--   ----   ------ ----------- ------                     ------------
667  Po667  up     success     success                    1,19,31-35,
                                                         2000,4092
4093 Po4093 up     success     success                    4093
```

Detect

Type-2 inconsistency indicates that one vPC peer has SVI configured and in up/up state and the other does not have it.

Troubleshooting

vPC — General Health Check

```
PeerA# show vpc consistency-parameters global
    Legend:
        Type 1 : vPC will be suspended in case of mismatch

Name                        Type  Local Value            Peer Value
-------------               ----  ---------------------- -----------------------
STP Mode                    1     Rapid-PVST             Rapid-PVST
STP Disabled                1     None                   None
STP MST Region Name         1     ""                     ""
STP MST Region Revision     1     0                      0
STP MST Region Instance to  1
 VLAN Mapping
STP Loopguard               1     Disabled               Disabled
STP Bridge Assurance        1     Enabled                Enabled
STP Port Type, Edge         1     Normal, Disabled,      Normal, Disabled,
BPDUFilter, Edge BPDUGuard        Disabled               Disabled
STP MST Simulate PVST       1     Enabled                Enabled
VTP domain                  2
VTP version                 2     2                      2
VTP mode                    2     Server                 Server
VTP password                2
VTP pruning status          2     Disabled               Disabled
Interface-vlan admin up     2     19,31-35,2000,4092-409 19,31-35,4092-4093
                                  3
Interface-vlan routing      2     1,19,31-35,2000,4092-4 1,19,31-35,4092-4093
capability                        093
Allowed VLANs               -     1,19,31-35,2000,4092-4 1,19,31-35,2000,4092-4
Local suspended VLANs       -     -                      -
```

Detect

- Note: Both vPC peers will be in active (primary) state if both PL and PKL fail and stay active if only PL is recovered. In case only PL fails, secondary vPC peer suspends all of its vPCs.
- Note that interface vlan2000 is missing!

Troubleshooting

vPC — General Health Check

```
PeerA# show vpc brief
Legend:
                (*) - local vPC is down, forwarding via vPC peer-link
vPC domain id                     : 64
Peer status                       : peer adjacency formed ok
vPC keep-alive status             : peer is alive
Configuration consistency status: failed
Configuration consistency reason: vPC type-1 configuration incompatible -
  STP Mode inconsistent
Type-2 consistency status         : success
vPC role                          : secondary
Number of vPCs configured         : 2
Peer Gateway                      : Disabled
Peer gateway excluded VLANs       : -
Dual-active excluded VLANs        : -

vPC Peer-link status
---------------------------------------------------------------------
id   Port   Status Active vlans
--   ----   ------ --------------------------------------------------
1    Po664  up     -

vPC status
----------------------------------------------------------------------------
id   Port   Status Consistency Reason                     Active vlans
--   ----   ------ ----------- ------                     ------------
667  Po667  up     failed      Global compat check failed -
4093 Po4093 up     failed      Global compat check failed -
```

Detect

- STP incompatibility was introduced and vpc was suspended
- Vlan2000 SVI issue was fixed

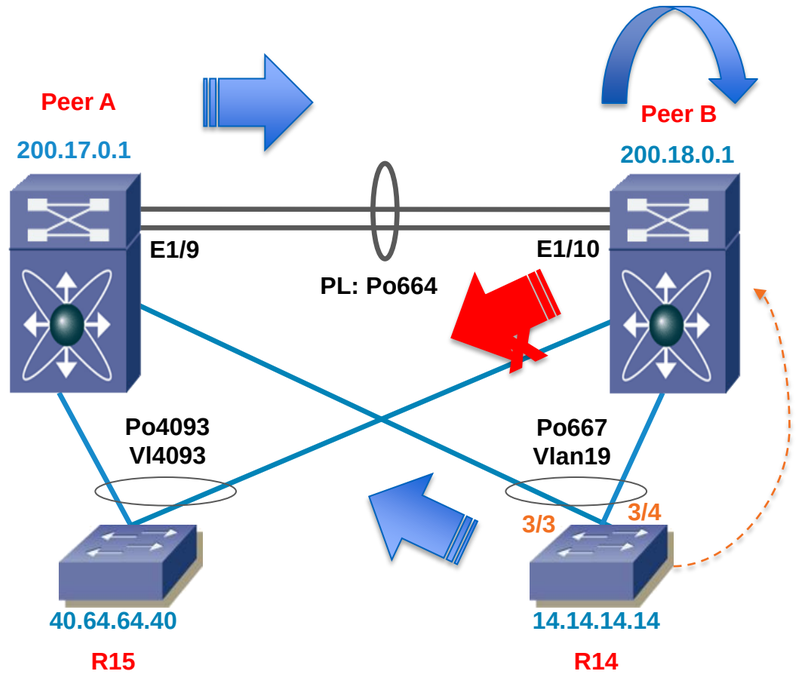
Troubleshooting

vPC — Why Routing Doesn't Work Without Peer-Gateway?

Problem

Ping fails from R14 (14.14.14.14) to R15 (40.64.64.40)

- To stabilize OSPF in VLAN19, peer-gateway was disabled.
- A packet capture on R15 confirmed that the ICMP request packets is NOT being received.



Analysis

- R14 next-hop is PeerB, as OSPF route depicts.
- R14 EC hashes to Ten3/3 connected to PeerA
- PeerA forwards the L2 packet over PL in vlan19 towards PeerB
- PeerB sets the VSL bit on PL ingress, and routes the packets into VLAN4093
- On the egress module, PeerB performs a VSL bit check and drops the packet.

Troubleshooting

vPC — Why Routing Doesn't Work Without Peer-Gateway? (Cont.)

Find

Po4093 members

Map

Ports to ASIC

Verify

that the error counter is incrementing continuously as the Ping continues

Extra — ELAM capture?

```
PeerB# show system internal pixm info interface Po4093
--snip--
Member rbh rbh_cnt
Eth3/34    0x000000f0    0x04
Eth3/32    0x0000000f    0x04
```

```
module-3# show hardware internal dev-port-map | egrep "32|34|FP"
FP port|PHYS |SECUR |MAC_0 |RWR_0 |L2LKP |L3LKP |QUEUE |SWICHF
  32      3      7      2      1      0      0      0      0
  34      4      8      2      1      0      0      0      0
```

```
PeerB#show hardware internal statistics module 3 device mac errors port 32 |
egrep -b 9 aric
|-----------------------------------------------------------------------|
| Device:R2D2                          Role:MAC                    Mod: 3    |
| Last cleared @ Mon Mar 28 21:46:42 2011
| Device Statistics Category :: ERROR
|-----------------------------------------------------------------------|
Instance:2
ID    Name                                                Value              Ports
--    ----                                                -----              -----
4422 mstat_rdrop                                          0000000000001791   32 -
28688 aric_no_port_select_error                           0000000000001037   25-36
```

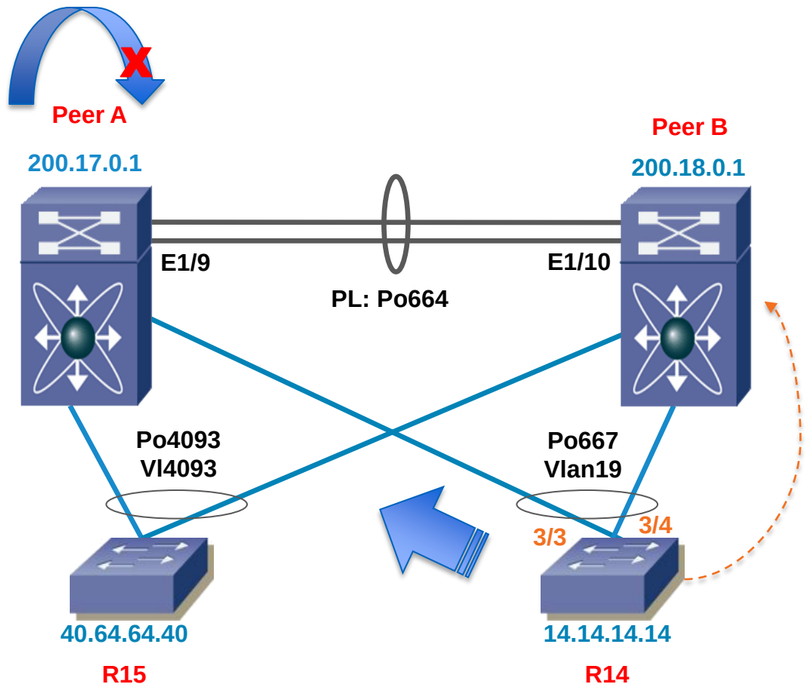
Troubleshooting

vPC — Why Routing Doesn't Work With Peer-Gateway?

Problem

OSPF is unable to come up!!

- To remedy the previous failure scenario, peer-gateway was enabled.
- OSPF neighbors went down and the switch logged the incident indicating "Too many retransmissions".



Analysis

- OSPF multicast packets are OK
- R14 send the unicast Hello packet to PeerB.
- R14 EC hashes to Ten3/3 connected to PeerA
- PeerA punts the packet to the CPU since TTL=1 and G-Bit is set
- PeerB SUP drops the unicast Hello packet

Troubleshooting

vPC — Why Routing Doesn't Work With Peer-Gateway? (Cont.)

CLI

shows OSPF status

```
PeerA# show ip ospf neighbor vlan19 | grep -a 2 Neighbor
 Neighbor ID    Pri State            Up Time  Address         Interface
 14.14.14.14      1 EXSTART/DROTHER  00:01:47 192.251.19.14   Vlan19
 200.18.0.1       1 FULL/BDR         1d05h    192.251.19.22   Vlan19
```

Syslog

report OSPF failure

```
Mar 29 16:53:55.691: %OSPF-5-ADJCHG: Process 6467, Nbr 200.18.0.1 on Vlan19
from EXCHANGE to DOWN, Neighbor Down: Too many retransmissions
```

Ethanalyzer

can be used to verify that PeerB is receiving the Hello packet and punting it.

```
PeerA#ethanalyzer local interface inband decode-internal capture-filter "proto
89 and host 192.251.19.14 and host 192.251.19.22" limit-captured-frames 1
detail > bootflash:ospf_neighbor.txt

Use "write" to create a pcap file which can later be analyzed by GUI wireshark
```

Troubleshooting

vPC — Golden rule

In case the issue you have encountered is urgent, complicated or you can’t figure it out, collect **show tech-support** output asap!

Related show tech(s)

```
N7K-1-VDC2# show tech-support vpc
N7K-1-VDC2# show tech-support stp
N7K-1-VDC2# show tech-support vtp
N7K-1-VDC2# show tech-support pixm
N&K-1-VDC2# show tech-support forwarding l2 unicast
```

Agenda

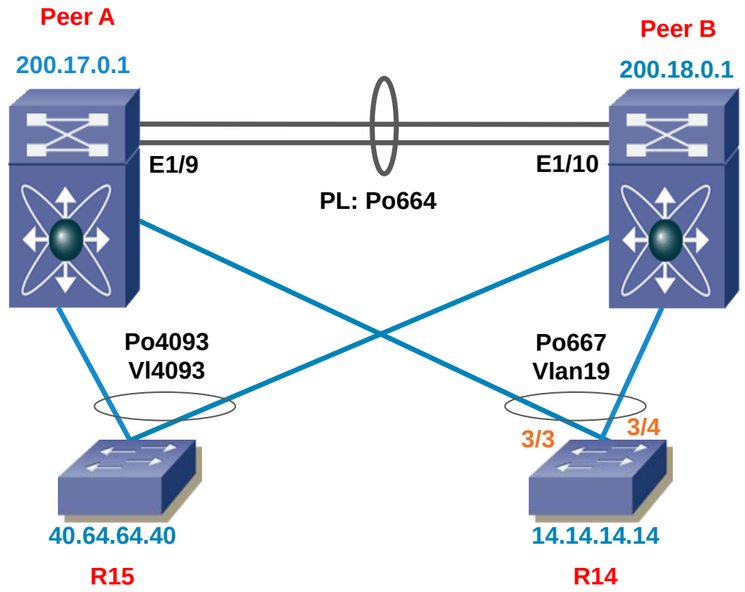
- Before You Get Started
  - Traditional Versus NX-OS Troubleshooting Approach
  - Nexus 7000 Built-in Troubleshooting Tools
  - Architecture Overview
- Troubleshooting
  - CPU
  - Control Plane
  - Memory Utilization
  - ACL
  - vPC
  - Unicast Layer 2 and Layer 3 Forwarding and ARP
  - Multicast Layer 2 and Layer 3 Forwarding

Unicast L2 and L3 Forwarding, ARP

L2 — Essentials

Tasks

- Track the MAC address!!
- Check the ASICs errors counters
- Verify Spanning-Tree status
- Adjacency
- ARP and Glean Throttling
- Forwarding Engine Error counters



Key Points

- A sound knowledge of the software architecture and hardware programming is essential to troubleshoot the internal data path
- Start by confirming that the way the feature is configured is supported and follows the recommendations
- Verify the hardware programming before jumping to capture the packets

Unicast L2 and L3 Forwarding, ARP

L2 — Identify your Physical/Logical port

```
N7K-1-PeerA# show system internal pixm info interface port-channel 664 vdc 2

PC_TYPE     PORT     LTL       RES_ID        LTL_FLAG       CB_FLAG       MEMB_CNT
------------------------------------------------------------------------------
--
Normal      Po664      0x0a40     0x16000297    0x00000000     0x00000002    1

Member    rbh           rbh_cnt
Eth1/9    0x000000ff    0x08
VLAN| BD| CBL |BD-St & CBL Direction:
-------------------------------------
[snip …]
32 | 0x3be | FORWARDING | INCLUDE_IF_IN_BD | BOTH
33 | 0x3bf | FORWARDING | INCLUDE_IF_IN_BD | BOTH

-- snip --
```

PIXM (Port Index Manager) component manages various hardware indexes tables used by the system forwarding architecture. The PIXM server runs on the SUP and interacts with PIXM clients on the different linecards.

Know

- LTL (Local Target Logic) index is assigned to each physical/logical port in the system, and used by the Forwarding Decision Engines to forward frames.
- BD (Broadcast Domain) index is assigned to each VLAN, and used by Forwarding Decision Engines to flood frame to all ports in the VLAN
- If the LTL index in ELAM RBUS result is different than the port LTL, then the frame is not sent out that port.

Unicast L2 and L3 Forwarding, ARP

L2 — Mac Addresses, Software Entry

```
N7K-1-PeerA# show mac address-table vlan 32
Legend:
        * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
        age - seconds since last seen,+ - primary entry using vPC Peer-Link
   VLAN     MAC Address      Type      age     Secure NTFY Ports/SWID.SSID.LID
---------+-----------------+--------+---------+------+----+------------------
G 32       0000.0c07.ac20    static       -       F    F  sup-eth1(R)
G 32       0011.3232.3232    static       -       F    F  sup-eth1(R)
* 32       0022.3232.3232    static       -       F    F  vPC Peer-Link
* 32       0000.98b9.4868    dynamic   60         F    F  Po667
```

```
N7K-1-PeerA# show mac address-table vlan 32 | egrep "G|Vlan|--"
Legend:
        * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
---------+-----------------+--------+---------+------+----+------------------
G 32       0000.0c07.ac20    static       -       F    F  sup-eth1(R)
G 32       0011.3232.3232    static       -       F    F  sup-eth1(R)
G 32       0022.3232.3232    static       -       F    F  vPC Peer-Link(R)
* 32       0000.98b9.4868    dynamic   420        F    F  Po667
* 32       0013.5f1f.46c0    dynamic   480        F    F  Po667
```

Observe

- The G flag is set for all local SVI mac addresses, which is expected since vPC HSRP “acts” in active/active mode.
- With peer-gateway disabled, the mac addresses of the remote SVI interfaces point to the peer-link.
- With peer-gateway enabled, the G flag is set for the remote SVI mac addresses and all routing is done locally.

- NOT seeing the G bit set when peer-gateway is enabled point to a switch programming issue.

Unicast L2 and L3 Forwarding, ARP

L2 — Mac Addresses, Hardware Entry

```
module-1# show hardware mac address-table vlan 32 vdc 2

FE  | Valid| PI|  BD  |      MAC       |  Index| Stat| SW  | Modi| Age| Tmr| GM| Sec| TR| NT| RM| RMA| Cap| Fld|Always
    |      |   |      |                |       |  ic |     | fied|Byte| Sel|   | ure| AP| FY|   |    |TURE|    | Learn
----+------+---+------+----------------+-------+-----+-----+-----+----+----+---+----+---+---+---+----+----+----+------
 0     1      1   958    0000.98aa.8ac9  0x00a42   0    0x003   0    215   1    0    0   0   0   0    0    0    0     0
 0     1      1   958    0022.3232.3232  0x00a40   1    0x000   0    42    1    1    0   0   0   0    0    0    0     0
 0     1      1   958    0000.0c07.ac20  0x00400   1    0x000   0    56    1    1    0   0   0   0    0    0    0     0
 0     1      0   958    0100.0cff.fffe  0x07ffc   1    0x001   0    169   0    0    0   0   0   0    0    1    0     0
```

0x00a42 is Po667 LTL index

0x00400 is Inband LTL index

L2lu – L2 forwarding engine (Lookup Unit)

This output shows only non zero error counters

There are no errors which would indicate L2 forwarding issue

```
module-1# show hardware internal statistics device l2lu errors
|-------------------------------------------------------------------------|
| Device:Eureka                    Role:L2                      Mod: 1    |
| Last cleared @ Fri Feb 25 21:30:09 2011
| Device Statistics Category :: ERROR
|-------------------------------------------------------------------------|
Instance:0
ID    Name                                                Value              Ports
--    ----                                                -----              -----
185   Non-flood packets sent with drop-index              0000000000000039   1-32 I1
```

Unicast L2 and L3 Forwarding, ARP

MAC ASICs Statistics

```
N7K-1-PeerA# slot 1 show hardware internal statistics device mac pktflow port 2
| grep -v ^$
|-----------------------------------------------------------------------|
| Device:R2D2                    Role:MAC                      Mod: 1     |
| Last cleared @ Wed Apr 13 08:32:20 2011
|-----------------------------------------------------------------------|
Instance:0
ID    Name                                          Value              Ports
--    ----                                          -----              -----
4096 mstat_rx_pkts                                  0000000000656926   2,4,6,8 -
4128 mstat_rx_pkts_65_127                           0000000000436692   2,4,6,8 -

--snip -

|-----------------------------------------------------------------------|
| Device:Naxos                   Role:MAC SECURITY             Mod: 1     |
| Last cleared @ Wed Apr 13 08:32:20 2011
|-----------------------------------------------------------------------|
Instance:0
ID    Name                                          Value              Ports
--    ----                                          -----              -----
11    sys_egress_octets                             0000054061058144   2 -
12    sys_egress_unicast_frames                     0000000000574369   2 -
--snip–
33    phy_ingress_octets                            0000001111728397   2 -
34    phy_ingress_unicast_frames                    0000000000370327   2 -
```

Inspect

- mstat – mac level counters
- sys – fabric-side counters
- phy – network-side counters
- ingress – ingress from n7k perspective
- egress – egress from n7k perspective

- This output gives good clues on related issues If the counters are NOT incrementing (or rapidly incrementing) .

Unicast L2 and L3 Forwarding, ARP

L2 — Spanning-Tree

```
N7K-1-PeerA#show spanning-tree vlan 32 | grep -v "^$"
VLAN0032
  Spanning tree enabled protocol rstp
  Root ID     Priority    32
              Address     0023.04ee.be40
              This bridge is the root
              Hello Time  2   sec  Max Age 20 sec  Forward Delay 15 sec
  Bridge ID   Priority    32     (priority 0 sys-id-ext 32)
              Address     0023.04ee.be40
              Hello Time  2   sec  Max Age 20 sec  Forward Delay 15 sec
Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
-
Po664            Desg FWD 2         128.4759 (vPC peer-link) Network P2p
```

```
N7K-1-PeerA# show spanning-tree internal event-history tree 32 interface
port-channel 664 | grep -v "^$"
VDC02 VLAN0032 <port-channel664>
0) Transition at 145271 usecs after Sat Apr  2 16:04:38 2011 State: BLK
Role: Desg  Age:  0 Inc: no  [STP_PORT_STATE_CHANGE]
```

Detect

- When Peer-switch is configured, both peers should report to be the root with the same ID.
- Spanning-Tree event-history is a good place to look for related events when troubleshooting packet loss or flooding.

Troubleshooting

L3 — Essentials

Theory of Operation

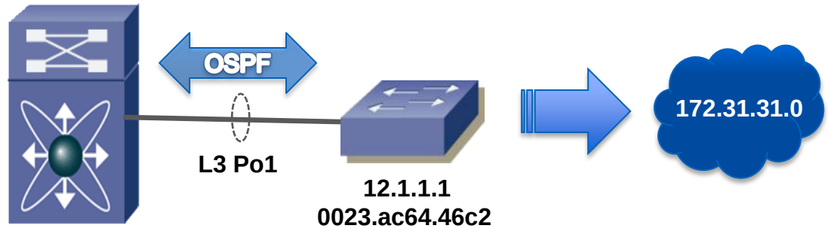
Network

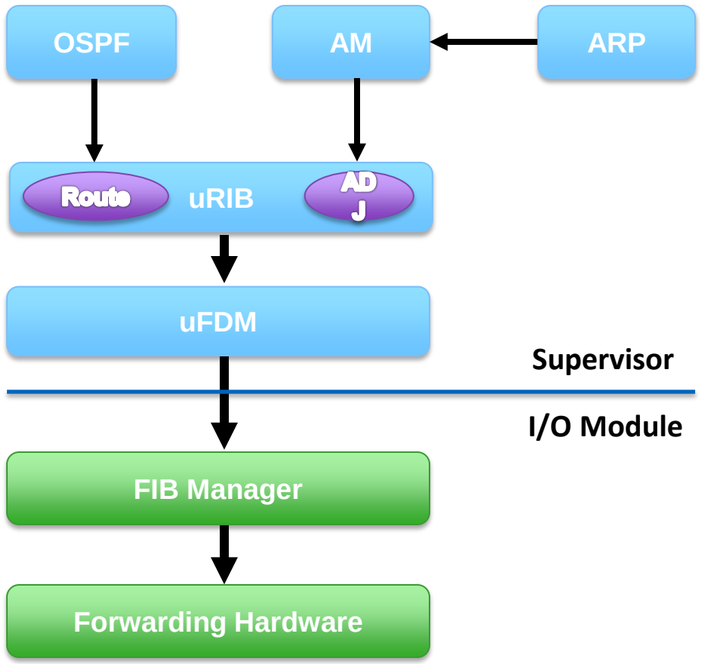
- OSPF is running between the Nexus switch and upstream router via a layer 3 port-channel
- Remote-site routes are advertised back to the Nexus

Software/Hardware Programming

- OSPF communicates with uRIBto build the routing table
- AM builds the next-hop adjacency entry
- uFDM distributes the information to the linecards
- IP FIB (running on the linecards) programs the ASIC components with the forwarding and adjacency information.

Remember: Software forwarding by the SUP is only used for control and exception packets





Unicast L2 and L3 Forwarding, ARP

L3 — Software Entry

Next-Hop

Check the routing table

```
N7K-1-VDC3# show ip route 172.31.31.0 | grep -b 2 via

172.31.31.0/24, ubest/mbest: 2/0
    *via 12.1.1.1, Po1, [110/80], 00:12:09, ospf-6467, intra
```

ARP/MAC

Check the ARP Table

```
N7K-1-VDC3# show ip arp | egrep "12.1.1.1|12.111.111.1"
12.1.1.1         00:01:36  0023.ac64.46c2  port-channel1
```

Adjacency

Check the ARP Table

```
N7K-1-VDC3# show ip adjacency 12.1.1.1 | grep -b 3 12.1
IP Adjacency Table for VRF default
Total number of entries: 1
Address         MAC Address     Pref Source   Interface
12.1.1.1        0023.ac64.46c2  50   arp      port-channel1
```

uRIB

Check the uRIBtable for next-hop info

```
N7K-1-VDC3# show forwarding ip route 172.31.31.0/24 module 1

IPv4 routes for table default/base

------------------+------------------+---------------------
Prefix            | Next-hop         | Interface
------------------+------------------+---------------------
172.31.31.0/24      12.1.1.1           port-channel1
```

Unicast L2 and L3 Forwarding, ARP

L3 — Hardware Entry

IP FIB

Check the FIB for ADJ entry

```
N7K-1-VDC3# show system internal forwarding ip route 172.31.31.0/24 detail module 1
RPF Flags legend:
          S - Directly attached route (S_Star)
          V - RPF valid
          M - SMAC IP check enabled
          G - SGT valid
          E - RPF External table valid
        172.31.31.0/24      ,  port-channel1
        Dev: 1 , Idx: 0xf1f6   , RPF Flags: V      , DGT: 0 , VPN: 33
        RPF_Intf_5:   port-channel1   (0x4018   )
        AdjIdx: 0x43032, LIFB: 0    , LIF: port-channel1   (0x4018   ), DI: 0xa46
        DMAC: 0023.ac64.46c2 SMAC: 0023.ac64.46c3
```

Verify

the ADJ entry counters and make sure its incrementing correctly

```
N7K-1-VDC3# show system internal forwarding adjacency entry 0x43032 detail module 1
Device: 1    Index: 0x43032    DMAC: 0023.ac64.46c2 SMAC: 0023.ac64.46c3
             LIF: 0x4018 (port-channel1) DI: 0xa46    ccc: 4    L2_FWD: NO   RDT: YES
             packets: 356523bytes: 534784500zone enforce: 0
```

Unicast L2 and L3 Forwarding, ARP

L3 — ARP, Glean Throttling

```
N7K-3-PeerB# show ip arp vlan 32 | grep -v ^$
Flags: * - Adjacencies learnt on non-active FHRP router
       + - Adjacencies synced via CFSoE
       # - Adjacencies Throttled for Glean
       D - Static Adjacencies attached to down interface
IP ARP Table
Total number of entries: 4
Address         Age       MAC Address     Interface
172.32.32.11    00:07:01  0011.3232.3232  Vlan32
172.32.32.14    00:06:35  0013.5f1f.46c0  Vlan32
172.32.32.150   00:01:26  INCOMPLETE      Vlan32        #
172.32.32.151   00:01:26  INCOMPLETE      Vlan32        #
```

```
N7K-3-PeerB# show ip adjacency 172.32.32.150 detail | b default | grep -v ^$
IP Adjacency Table for VRF default
Total number of entries: 1
Address :             172.32.32.150
MacAddr :             0000.0000.0000
Preference :          255
Source :              arp
Interface :           Vlan32
Physical Interface :  Vlan32
Packet Count :        62027
Byte Count :          5954592
Best :                Yes
Throttled :           Yes
```

Detect

- All packets destined to Incomplete entries are exceptions and get punted to supervisor for software forwarding.
- To protect the CPU, adjacency throttling kicks in and drops excess glean traffic

```
hardware ip glean throttle
hardware ip glean throttle maximum 1000
hardware ip glean throttle timeout 300
hardware ip glean throttle syslog 500
```

Unicast L2 and L3 Forwarding, ARP

L3 — Forwarding Engine Error Statistics

```
N7K-1-PeerA# show hardware internal statistics module 1 device L3lu errors port 2
Hardware statistics on module 01:
|-------------------------------------------------------------------------|
| Device:Lamira                    Role:L3                        Mod: 1    |
| Last cleared @ Fri Feb 25 21:30:09 2011
| Device Statistics Category :: ERROR
|-------------------------------------------------------------------------|
Instance:0
ID    Name                                              Value              Ports
--    ----                                              -----              -----
75    RP IPv4 L3 filtering Pkt drop                     0000000000000002   1-32 I1
76    RP IPv6 L3 filtering Pkt drop                     0000000000000001   1-32 I1
93    CL1 Same IF check Fail Pkt count                  0000000038480964   1-32 I1
188   PL OFE Global aggr drop pkt ctr                   0000000018316176   1-32 I1
189   PL OFE Global aggr drop byte ctr                  0000027451514923   1-32 I1
198   PL OFE Total police drop pkt ctr                  0000000018316176   1-32 I1
199   PL OFE Total police drop byte ctr                 0000027451514923   1-32 I1
207   PL OFE TTL expire pkt ctr                         0000000000037961   1-32 I1
259   L3 Fib Miss Pkt ctr                               0000000588018615   1-32 I1
260   L3 IPv4 Option Pkt ctr                            0000000000000357   1-32 I1
261   L3 IPv6 Option Pkt ctr                            0000000000046652   1-32 I1
262   L3 Non-Rpf Drop Pkt ctr                           0000000000240773   1-32 I1
305   NF L3 ACL deny pkt ctr                            0000006154091492   1-32 I1
449   Exception cause: ICMP UNREACH (Unicast)           0000000000538866   1-32 I1
454   Exception cause: L3 BRIDGE DROP (Unicast)         0000000733007752   1-32 I1
455   Exception cause: DROP (Unicast)                   0000000000000003   1-32 I1
461   Exception cause: OPTIONS (Multicast)              0000000000047009   1-32 I1
463   Exception cause: TWO MCAST RPF (Multicast)        0000000000000016   1-32 I1
464   Exception cause: L3 BRIDGE DROP (Multicast)       0000000001080488   1-32 I1
```

CoPP dropped packets

No route traffic drops

Acl dropped packets, when acl-log is configured packets hits also access-list-log rate-limiter

Packets received across vpc PL from mcast vpc forwarder

Unicast L2 and L3 Forwarding, ARP

L3 — Golden rule

In case the issue you have encountered is urgent, complicated or you can’t figure it out, collect **show tech-support** output asap!

Related show tech(s)

```
N7K-1-VDC2# show tech-support forwarding L3 unicast
N7K-1-VDC2# show tech-support netstack
N7K-1-VDC2# show tech-support arp
N7K-1-VDC2# show tech-support L2fm
```

Agenda

- Before You Get Started
  - Traditional Versus NX-OS Troubleshooting Approach
  - Nexus 7000 Built-in Troubleshooting Tools
  - Nexus 7000 Module and Forwarding Engine Architecture Overview
- Troubleshooting
  - CPU
  - Control-Plane – CoPP
  - Memory Utilization
  - ACL
  - vPC
  - Unicast Layer 2 and Layer 3 Forwarding and ARP
  - Multicast Layer 2 and Layer 3 Forwarding

Multicast L2 and L3 Forwarding

Multicast Replication

L2 Replication

- Copy of original packet for each output fabric-channel or switchport
- Performed by xbar (stage number 2) and port asics
- Driven by Multicast indexes (MI) at fabric level and LTL indexes at port level
- Multicast Distribution or MD copy is created by ingress replication asic

L3 Egress Replication

- Copy of original packet for each layer 3 interface (OIF)
- Performed by replication asic aka ‘rewrite’ or ‘RWR_0’
- Multicast Expansion Table (MET) in replication engines contains OIFs
- Nexus 7000 system supports egress Layer 3 replication

Asymmetric METs

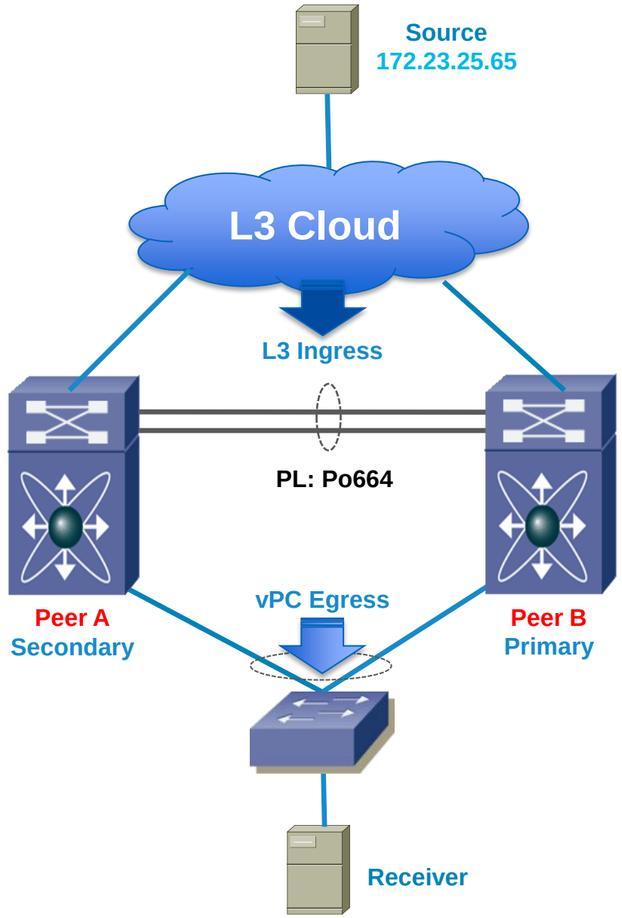
- Conserves replication asic and forwarding engine bandwidth (forwarding engine must provide lookup result for each individual packet copy)
- If OIF is SVI which L2 Vlan spans across multiple I/O modules, each I/O module creates copy of original packet even no receivers are present

Multicast L2 and L3 Forwarding

L2/L3 — Platform Independent, vPC Specific Check

Characteristic

- VPC supports PIM-SM only
- VPC uses CFS to sync IGMP state
- For sources in VPC domain, both VPC peers are forwarders
- Duplicates avoided via VPC loop-avoidance logic
- For sources in Layer 3 cloud, unicast best metric determines active forwarder (VPC operational primary in case of tie)
- CFS used to negotiate active forwarder role on per-source basis

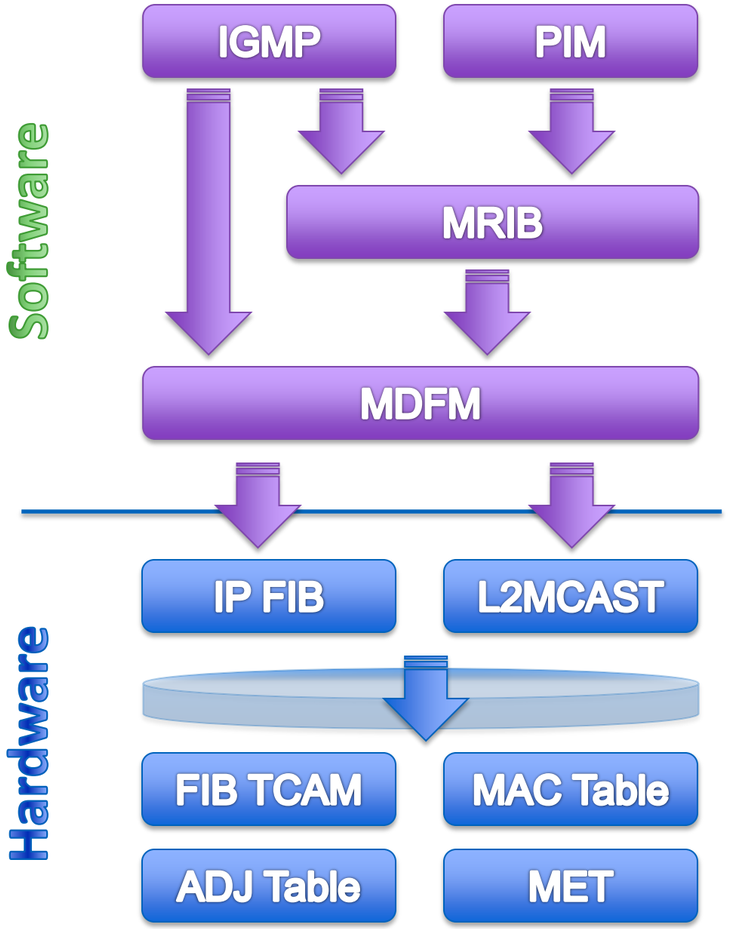


Multicast L2 and L3 Forwarding

L2/L3 — Platform Independent, vPC Specific Check

Facts

- IGMP and PIM processes learn routing information from neighbors and hosts respectively
- mRIB calculates multicast routing/RP/RPF/OIL information and populates the mroute and IGMP tables
- mFDM pushes the forwarding information down to the linecards
- IP FIB (on the linecards) programs the FIB TCAM, L2 multicast adjacency, and MET Tables



Multicast L2 and L3 Forwarding

L2/L3 — IGMP and PIM Verification

PeerA

```
N7K-1-PeerA# show ip igmp internal vpc
IGMP vPC operational state UP
IGMP vPC Operating Version: 2
IGMP vPC Domain ID: 64
IGMP vPC Peer-link Exclude feature enabled

N7K-1-PeerA# show ip pim internal vpc
PIM vPC operational state UP
VPC peer link is up on port-channel664
PIM vPC Operating Version: 2
PIM vPC Domain ID: 64

N7K-1-PeerA# show ip pim internal vpc rpf
Source: 172.23.25.65
  Pref/Metric: 110/63
  Source role: secondary
  Forwarding state: Win (forwarding)
```

Win state is per S,G

PeerB

```
N7K-3-PeerB# show ip igmp internal vpc
IGMP vPC operational state UP
IGMP vPC Operating Version: 2
IGMP vPC Domain ID: 64
IGMP vPC Peer-link Exclude feature enabled

N7K-3-PeerB# show ip pim internal vpc
PIM vPC operational state UP
VPC peer link is up on port-channel664
PIM vPC Operating Version: 2
PIM vPC Domain ID: 64

N7K-3-PeerB# show ip pim internal vpc
Source: 172.23.25.65
  Pref/Metric: 110/83
  Source role: primary
  Forwarding state: Lose (not forwarding)
```

Worse metric and therefore B is Loser

Multicast L2 and L3 Forwarding

L2 — IGMP and MRIB Verification

PeerA

```
N7K-1-PeerA# show ip igmp group 239.28.28.64
IGMP Connected Group Membership for VRF "default" - matching
  Group "239.28.28.64"
Type: S - Static, D - Dynamic, L - Local, T - SSM Translated
Group Address      Type Interface            Uptime    Expires   Last
Reporter
239.28.28.64       D    Vlan32               00:00:19  00:04:00
172.32.32.250
239.28.28.64       D    Vlan4093             00:01:19  00:03:49  40.9.3.12

N7K-1-PeerA# show ip igmp snooping groups 239.28.28.64 | grep -v ^$ | exc
*/*
Type: S - Static, D - Dynamic, R - Router port, F - Fabr
Vlan  Group Address      Ver  Type  Port list
32    239.28.28.64       v2   D     Po667
4093  239.28.28.64       v2   D     Po4093

N7K-1-PeerA# show ip mroute 239.28.28.64 flags
IP Multicast Routing Table for VRF "default"

(*, 239.28.28.64/32), uptime: 02:00:45, pim ip igmp
  Incoming interface: loopback88, RPF nbr: 64.67.88.93
  Outgoing interface list: (count: 2)
    Vlan32, uptime: 00:40:44, igmp
    Vlan4093, uptime: 00:41:44, igmp
```

Received IGMP join creates igmp snooping and (*,g) mroute entries Loopback88 is anycast-rp interface

PeerA is a vPC forwarder but PeerB has the same (*,g) entry

PeerB

```
N7K-3-PeerB# show ip igmp groups 239.28.28.64
IGMP Connected Group Membership for VRF "default" - matching
  Group "239.28.28.64"
Type: S - Static, D - Dynamic, L - Local, T - SSM Translated
Group Address      Type Interface            Uptime    Expires   Last Reporter
239.28.28.64       D    Vlan4093             00:01:22  00:03:47  40.9.3.12
239.28.28.64       D    Vlan32               00:00:22  00:03:57  172.32.32.250

N7K-3-PeerB# show ip igmp snooping groups 239.28.28.64 | grep -v ^$ | exc */*
Type: S - Static, D - Dynamic, R - Router port, F - Fabricpath core port
Vlan  Group Address      Ver  Type  Port list
32    239.28.28.64       v2   D     Po667
4093  239.28.28.64       v2   D     Po4093
```

Both peers have igmp state synchronized via CFS regardless to which of them igmp joined arrived to based on port-channel hashing

show ip igmp internal event-history event and show ip igmp asnooping internal event-history vpc outputs show events and CFS messaging between peers. Adjust event-history buffer via ip igmp event-history interface-events size <size> cli

Multicast L2 and L3 Forwarding

L2/L3 — MDFM and IP FIB Verification

```
N7K-1-PeerA# show forwarding distribution ip igmp snooping vlan
32 group 239.28.28.64

Vlan: 32, Group: 239.28.28.64, Source: 0.0.0.0
  Outgoing Interface List Index: 11
  Reference Count: 1
  Platform Index: 0x7ffa
  Number of Outgoing Interfaces: 2
    port-channel664
    port-channel667

N7K-1-PeerA# show forwarding multicast route group 239.28.28.64
source 172.23.25.65 module 1

  (172.23.25.65/32, 239.28.28.64/32), RPF Interface: port-
channel1, flags:
    Received Packets: 0 Bytes: 0
    Number of Outgoing Interfaces: 3
    Outgoing Interface List Index: 12
      Vlan32 Outgoing Packets:292302 Bytes:391684680
      Vlan4093 Outgoing Packets:146151 Bytes:195842340
      port-channel66 Outgoing Packets:0 Bytes:0
```

Verify the OIF entries for a specific group

Peer-link Po66 is not present on I/O Module 1

Multicast L2 and L3 Forwarding

L2/L3 — FIB TCAM, ADJ Table, and MET

```
N7K-1-PeerA# show system internal forwarding multicast route group 239.28.28.64 source
172.23.25.65 module 1 detail
(172.23.25.65/32, 239.28.28.64/32), Flags: *S
  Lamira: 1, HWIndex: 0x2202, VPN: 17
  RPF Interface: port-channel1, LIF: 0x40d9, PD oiflist index: 0x5
  ML3 Adj Idx: 0xa022, MD: 0x2007, MET0: 0x2008, MET1: 0x2008, MTU Idx: 0x1
  Metro Instance: 0
  Dev: 1 Index: 0xa038    Type: MDT      elif: 0xc0008
                          dest idx: 0x7ff0      recirc-dti: 0xe20000
  Dev: 1 Index: 0x60d9    Type: OIF      elif: 0x800d9     Vlan32
                          dest idx: 0x0         smac: 0011.3232.3232
  Metro Instance: 1
  Dev: 1 Index: 0xa038    Type: MDT      elif: 0xc0008
                          dest idx: 0x7ff0      recirc-dti: 0xe20000
  Metro Instance: 2
  Dev: 1 Index: 0xa038    Type: MDT      elif: 0xc0008
                          dest idx: 0x7ff0      recirc-dti: 0xe20000
  Metro Instance: 3
  Dev: 1 Index: 0xa038    Type: MDT      elif: 0xc0008
                          dest idx: 0x7ff0      recirc-dti: 0xe20000
  Dev: 1 Index: 0x60d9    Type: OIF      elif: 0x800d9     Vlan32
                          dest idx: 0x0         smac: 0011.3232.3232
  Dev: 1 Index: 0x6101    Type: OIF      elif: 0x80101     Vlan4093
                          dest idx: 0x0         smac: 0023.ac64.46c2
```

```
N7K-1-PeerA# show system internal forwarding adjacency entry 0x60d9 module 1 detail
Device: 1    Index: 0x60d9     DMAC: 0000.0000.0000 SMAC: 0011.3232.3232
             LIF: 0x800d9 (Vlan32) DI: 0x0      ccc: 4    L2_FWD: NO  RDT: NO
             packets: 12848bytes: 17216320zone enforce: 0
```

Module 1 is only egress module from multicast flow perspective

ML3 Adj Idx is same for all modules MET indexes do not need to be same for all modules

Empty MET tables in Metro 1 and 2 (no receivers, it saves replication and lookup resources

Index – OIF specific pointer to Adj table

DI – Dest index is zero as this information comes from L3 asic indicating L2 asic index will be used instead

Multicast L2 and L3 Forwarding

L2/L3 — Replication Engine Counters

```
N7K-1-PeerA# show hardware internal statistics module 1 device rewrite pktflow asic-all |
egrep Dev|Inst|Multicast

[snip]

| Device:Metropolis                Role:REWR                        Mod: 1     |

Instance:0
97  Multicast L3 MET replication pkt cnt          0000000000000500   2,4,6,8,10,12,14,16 I1
98   Multicast L3 PR replication pkt cnt          0000000000000500    2,4,6,8,10,12,14,16 I1

[snip]

96    Multicast L2 MET replication pkt cnt          0000000000000500    1,3,5,7,9,11,13,15 -
97    Multicast L3 MET replication pkt cnt          0000000000001000    1,3,5,7,9,11,13,15 -
98    Multicast L3 PR replication pkt cnt           0000000000001000    1,3,5,7,9,11,13,15 -
99    Multicast L2 PR replication pkt cnt           0000000000000500    1,3,5,7,9,11,13,15 -
```

L2/L3 MET – number of packets sent to replication
L2/L3 PR - number of copies created

Multicast L2 and L3 Forwarding

Golden rule

In case the issue you have encountered is urgent, complicated or you can't figure it out, collect **show tech-support** output asap!

Related show tech(s)

```
N7K-1-VDC2# show tech-support forwarding L3 multicast
N7K-1-VDC2# show tech-support ip pim
N7K-1-VDC2# show tech-support ip multicast
N7K-1-VDC2# show tech-support igmp brief
N7K-1-VDC2# show tech-support ip igmp snooping
N7K-1-VDC2# show tech-support ip mfwd
N7K-1-VDC2# show tech-support forwarding L2 multicast
```

Complete Your Online Session Evaluation

- Give us your feedback and you could win fabulous prizes. Winners announced daily.
- Receive 20 Cisco Daily Challenge points for each session evaluation you complete.
- Complete your session evaluation online now through either the mobile app or internet kiosk stations.

Maximize your Cisco Live experience with your free Cisco Live 365 account. Download session PDFs, view sessions on-demand and participate in live activities throughout the year. Click the Enter Cisco Live 365 button in your Cisco Live portal to log in.


CISCO
TM