
cisco
You make possible

Advanced Storage Area Network Design

Edward Mazurek
Technical Leader Data Center Storage Area Networking
emazurek@cisco.com
@TheRealEdMaz

BRKSAN-2883

cisco Live!
Barcelona | January 27–31, 2020

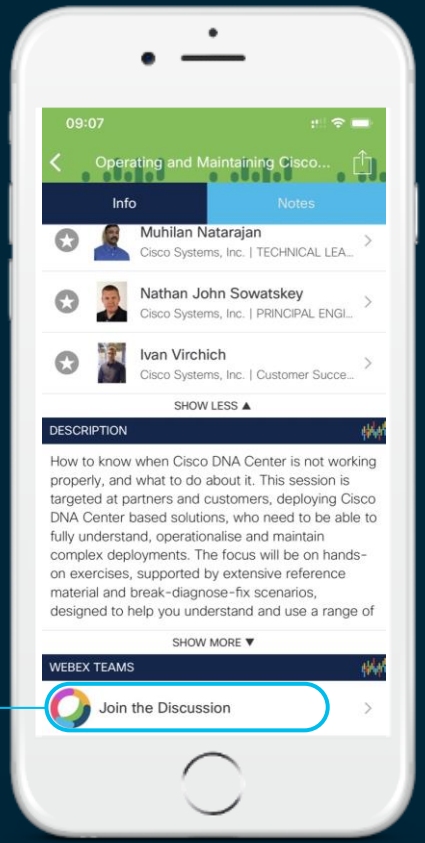
Cisco Webex Teams

Questions?

Use Cisco Webex Teams to chat with the speaker after the session

How

1. Find this session in the Cisco Events Mobile App
2. Click “Join the Discussion”
3. Install Webex Teams or go directly to the team space
4. Enter messages/questions in the team space



Agenda

- Welcome and Introduction
- Design Principles – Best Practices
- Design Principles for Slow Drain and Congestion Isolation
- Design Principles for SAN Analytics
- Q&A

Introduction

Introduction

- Assumptions:
  - Most SANs are reliable and have few problems
- Move from:
  - Occasional problems that sometimes cause outages
  - Performance difficult to ascertain
- Move to:
  - Infrequent problems and almost no outages
  - Transparent and easily obtained SCSI/NVMe performance information
- How can your SAN be more reliable, robust and less prone to errors?
- How can your SAN communicate actual performance?

Introduction

- Cisco FC/FCoE SAN switches provide a host of advanced features that can make your SANs more
- Robust
- Scalable
- Fault tolerant
- High performance
- Easy to Manage
- Easy to investigate / troubleshoot
- New features are being added every release!

Design Principles
Best Practices

Design Principles

- VSANs
- Zoning, Smart Zoning and Autozone
- N-Port Virtualization
- Trunking and Port-channeling
- MDS Internal CRC handling
- Device-alias
- SAN Security
- Misc: FCDomain, FEC, BB_Sc_N, CFS, Clock Settings, Uniform Timestamps

Zoning

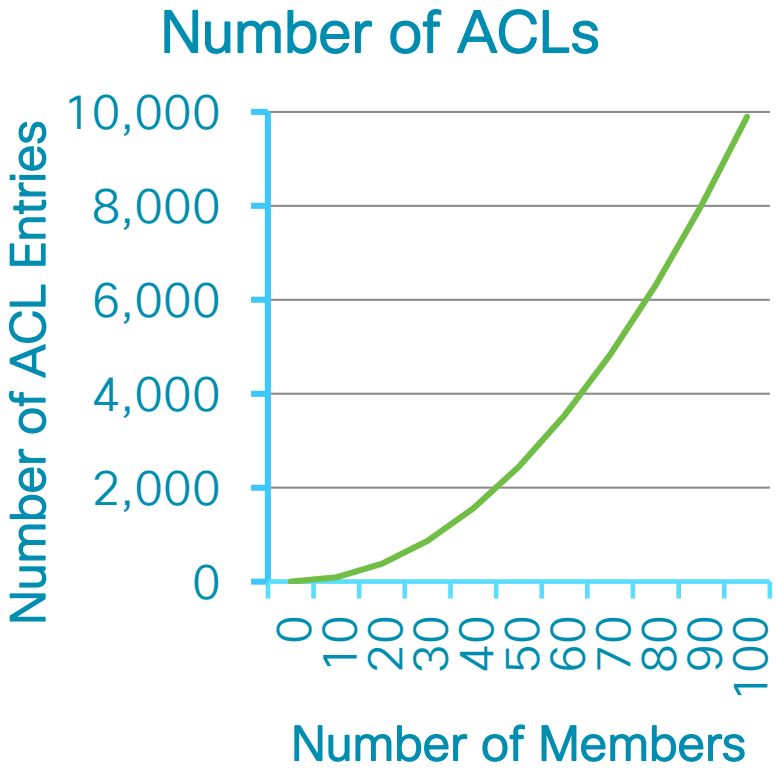
- Non-zoned devices are members of the default zone
- A physical fabric can have a maximum of 16,000 zones (9700-only network)
- Attributes can include pWWN, FC alias, FCID, FWWN, Switch Interface fc x/y, Symbolic node name, Device alias
- Recommended: Device-alias and/or PWWN

```
zone name AS01_NetApp vsan 42
  member pwwn 20:03:00:25:b5:0a:00:06
  member pwwn 50:0a:09:84:9d:53:43:54
```

```
device-alias name AS01
   pwwn 20:03:00:25:b5:0a:00:06
device-alias name NTAP
   member pwwn 50:0a:09:84:9d:53:43:54
zone name AS01_NetApp vsan 42
   member device-alias AS01
   member device-alias NTAP
```

The Trouble with sizable Zoning
All Zone Members are Created Equal

- Standard zoning model just has “members”
- Any member can talk to any other member
- Recommendation: 1-1 zoning
- Each pair consumes two ACL entries in TCAM
- Result: n*(n-1) entries per zone



Smart Zoning

Operation	1:1 Zoning			Today –Many - Many			Smart Zoning		
	Zones	Cmds	ACLs	Zones	Cmds	ACLs	Zones	Cmds	ACLs
Create zones(s)	32	96	64	1	13	132	1	13	64
Add an initiator	+4	+12	+8		+1	+24		+1	+8
Add a target	+8	+24	+16		+1	+24		+1	+16

8 x I
4 x T

- Feature added in NX-OS 5.2(6)
- Allows storage admins to create larger zones while still keeping premise of single initiator & single target
- Dramatic reduction SAN administrative time for zoning
- Utility to convert existing zone or zoneset to Smart Zoning

Autozone

- Automates zoning in single switch fabrics
- Automatically zones initiators with targets
- As devices come online they are added to the zoneset in VSAN 1 and the zoneset activated
- All initiators are zoned to all targets
- Runs every 5 minutes to check for new devices
- Introduced in NX-OS 8.3(1)
- 8.4(1) - Added the --enable , --enableautosave and --disableautosave options.
- Cisco MDS 9132T, 9148T, and 9396T fabric switches only.

Zoning Best Practices

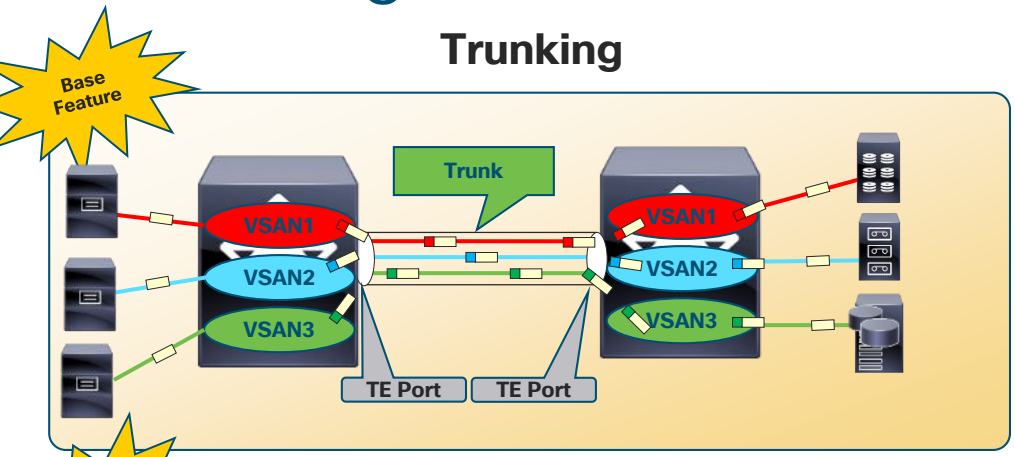
- **no zone default-zone permit**
  - All devices must be explicitly zoned
- **zone mode enhanced**
  - Acquires lock on all switches while zoning changes are underway
  - Enables full zoneset distribution
- **zoneset distribute full**
  - If not using enhanced mode this is a must to ensure common full zone database!
- **zone smart-zoning enable**
  - Allows for more efficient, easier zoning
- **zone confirm-commit**
  - Causes zoning changes to be displayed during zone commit
- **zoneset overwrite-control** – New in NX-OS 6.2(13)
  - Prevents a different zoneset than the currently activated zoneset from being inadvertently activated

Zoning Best Practices - continued

- Use smart zoning or single-initiator single-target zones
- Use device-alias
  - Names are much better than pWWNs
  - Do not contribute to the size of the zoning database
  - Device-alias enhanced mode allows device-alias in zoneset
- Do not use fWWN, sWWN or interface zoning for NPV switches
  - All devices on that NPV link will be zoned together!
- Check ACLTCAM usage after zoning changes
  - show system internal acl tcam-usage
  - show system internal acltcam-soc tcam-usage

Trunking & Port Channels

Trunking

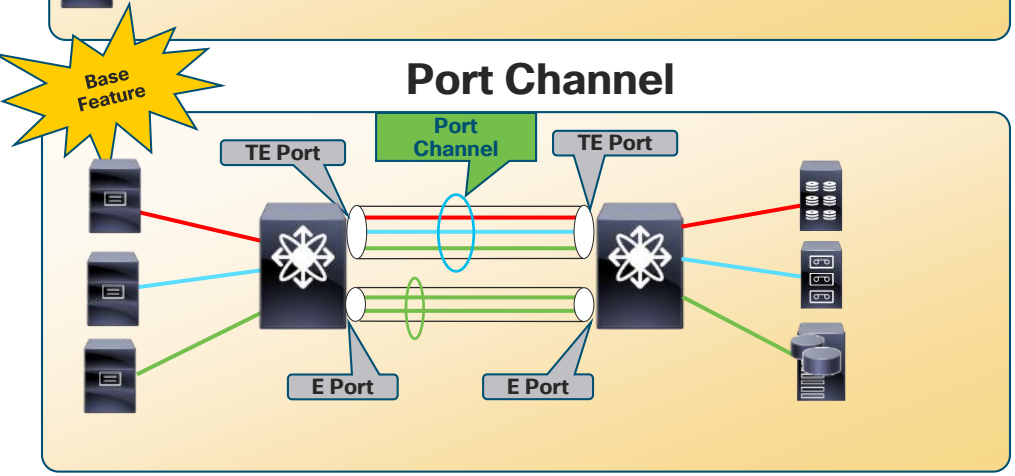


Single-link ISL or Port Channel ISL can be configured to become EISL – (TE_Port)

Traffic engineering with pruning VSANs on/off the trunk

Efficient use of ISL bandwidth

Port Channel



Up to 16 links can be combined into a Port Channel increasing the aggregate bandwidth by distributing traffic granularly among all functional links in the channel

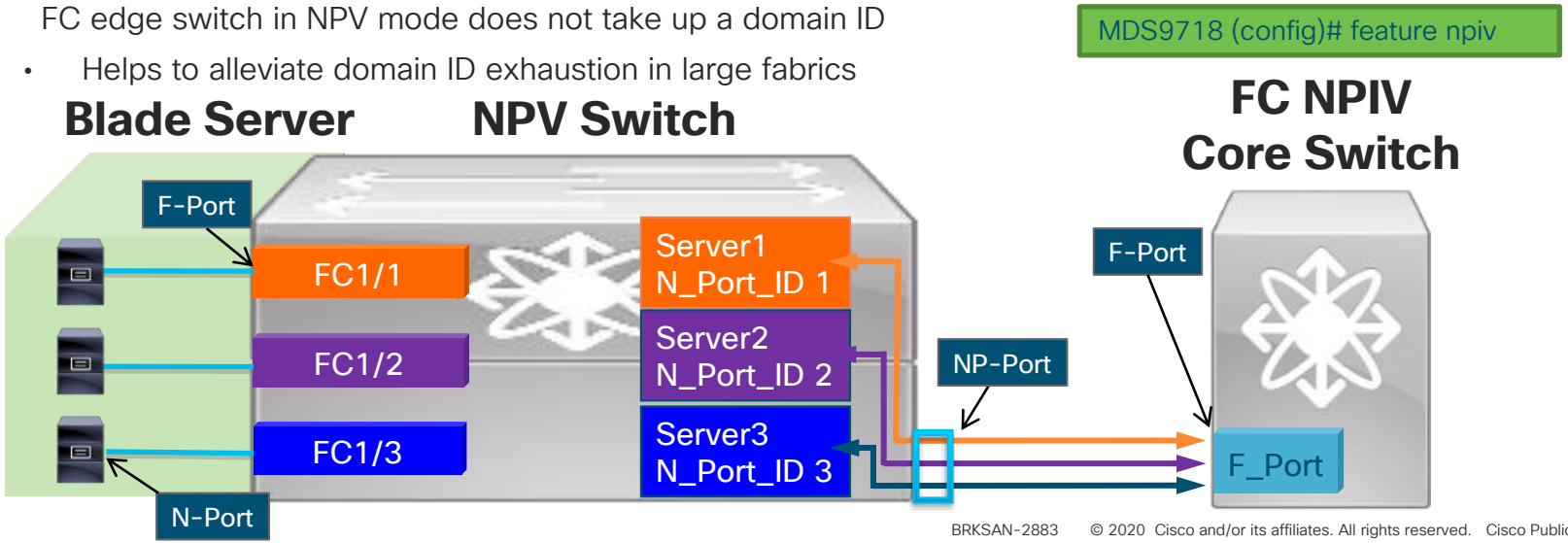
Load balances across multiple links and maintains optimum bandwidth utilization. Load balancing is based on the source ID, destination ID, and exchange ID

If one link fails, traffic previously carried on this link is switched to the remaining links. To the upper protocol, the link is still there, although the bandwidth is diminished. The routing tables are not affected by link failure

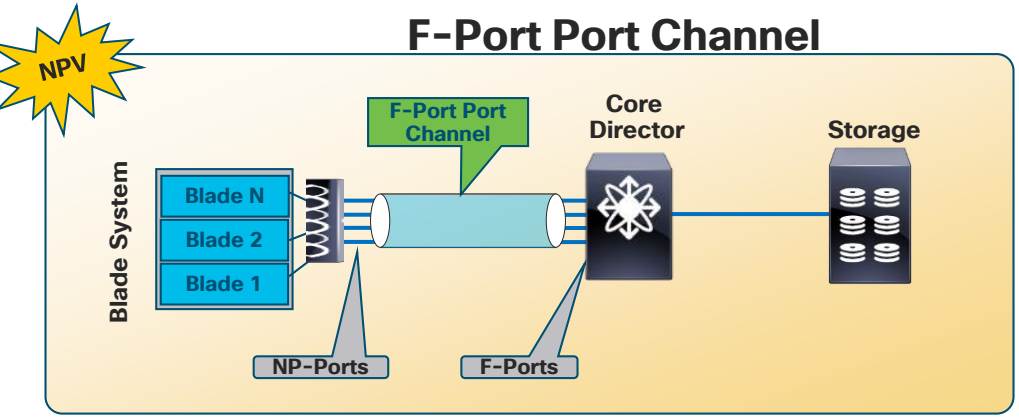
N-Port Virtualization

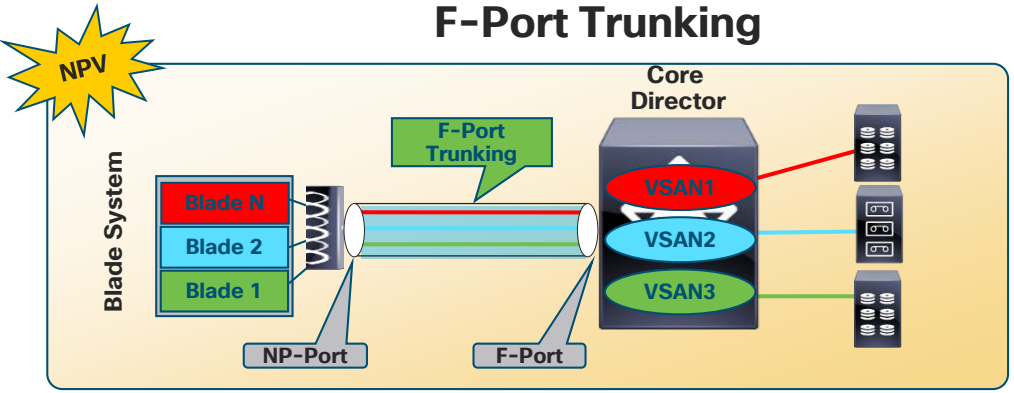
Scaling Fabrics with Stability

- N-Port Virtualizer (NPV) utilizes NPIV functionality to allow a “switch” to act like a server/HBA performing multiple fabric logins through a single physical link
- Physical servers connect to the NPV switch and login to the upstream NPIV core switch
- No local switching is done in NPV mode switch
- FC edge switch in NPV mode does not take up a domain ID
  - Helps to alleviate domain ID exhaustion in large fabrics



F-Port Port Channel and F-Port Trunking Enhanced Blade Switch Resiliency





F-Port Port Channel w/ NPV

- Bundle multiple ports in to 1 logical link
  - Any port, any module
- High-Availability (HA)
  - Blade Servers are transparent if a cable, port, or line cards fails
- Traffic Management
  - Higher aggregate bandwidth
  - Hardware-based load balancing

F-Port Trunking w/ NPV

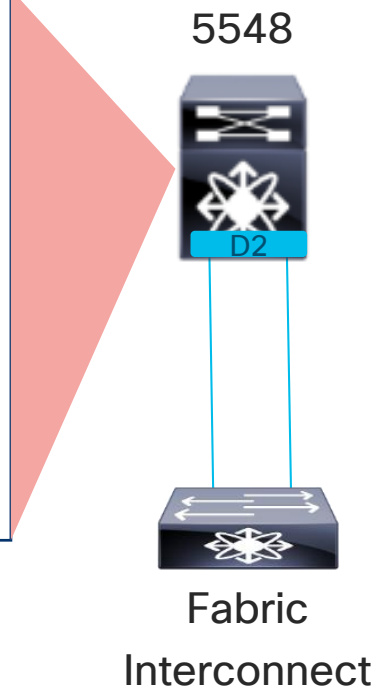
- Partition F-Port to carry traffic for multiple VSANs
- Extend VSAN benefits to Blade Servers
  - Separate management domains
  - Separate fault isolation domains
  - Differentiated services: QoS, Security

FLOGI – Before Port Channel

```
phx2-5548-3# show flogi database
--------------------------------------------------------------------------------
INTERFACE        VSAN    FCID           PORT NAME               NODE NAME
--------------------------------------------------------------------------------
fc2/9            12      0x020000  20:41:00:0d:ec:fd:9e:00 20:0c:00:0d:ec:fd:9e:01
fc2/9            12      0x020001  20:02:00:25:b5:0b:00:02 20:02:00:25:b5:00:00:02
fc2/9            12      0x020002  20:02:00:25:b5:0b:00:04 20:02:00:25:b5:00:00:04
fc2/9            12      0x020003  20:02:00:25:b5:0b:00:01 20:02:00:25:b5:00:00:01
fc2/10           12      0x020020  20:42:00:0d:ec:fd:9e:00 20:0c:00:0d:ec:fd:9e:01
fc2/10           12      0x020021  20:02:00:25:b5:0b:00:03 20:02:00:25:b5:00:00:03
fc2/10           12      0x020022  20:02:00:25:b5:0b:00:00 20:02:00:25:b5:00:00:00

Total number of flogi = 7

phx2-5548-3#
```



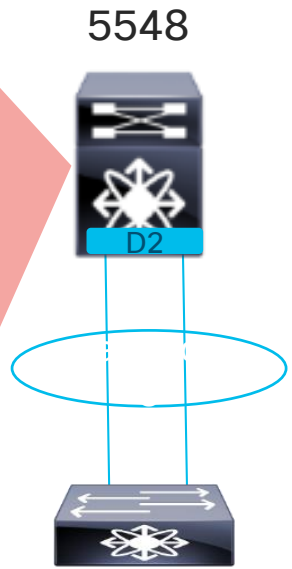
FLOGI- After port channel

```
N5548-3# show flogi database
--------------------------------------------------------------------------------
INTERFACE        VSAN    FCID           PORT NAME               NODE NAME
--------------------------------------------------------------------------------
San-po3          12      0x020040   24:0c:00:0d:ec:fd:9e:00 20:0c:00:0d:ec:fd:9e:01
San-po3          12      0x020001   20:02:00:25:b5:0b:00:02 20:02:00:25:b5:00:00:02
San-po3          12      0x020002   20:02:00:25:b5:0b:00:04 20:02:00:25:b5:00:00:04
San-po3          12      0x020003   20:02:00:25:b5:0b:00:01 20:02:00:25:b5:00:00:01
San-po3          12      0x020021   20:02:00:25:b5:0b:00:03 20:02:00:25:b5:00:00:03
San-po3          12      0x020022   20:02:00:25:b5:0b:00:00 20:02:00:25:b5:00:00:00

Total number of flogi = 6

phx2-5548-3#
```

All devices associated with port-channel now



Port Channel design considerations

All types of switches

- Name port channels the same on both sides (for clarity)
- Common port allocation in both fabrics
- ISL speeds should be >= edge device speeds
- Maximum 16 members per port channel allowed
- Multiple port channels to same adjacent switch should be equal BW
  - BW determines FSPF cost
- Member of VSAN 1 + trunk other VSANs

Port-channel design considerations

All types of switches

- Use channel mode active
  - This ensures misconfigurations and cabling are checked
  - Required for analytics
  - Default changed to “channel mode active” in NXOS 8.4(1)!

- Check TCAM usage on NPIV core switch
  - show system internal acl tcam-usage
  - show system internal acltcam-soc tcam-usage

Port Channel design considerations

Director class

- Distribute members across multiple line cards
- When possible use same port on each LC (for clarity):
  - Ex. fc1/5, fc2/5, fc3/5, fc4/5, etc.
- If multiple members per linecard distribute across Fwd Engines and port-groups
  - show port-resources module x

Port Channel design considerations

Fabric switches – MDS 9250i, 9148S, 9396S, 9132T, 9148T and 9396T

- **Ensure enough credits for distance**
  - Can “rob” buffers from other ports in port-group that are “out-of-service”
- **Split PC members across *different FWD* engines to distribute ACLTCAM**
  - For F port-channels to NPV switches (like UCS FIs)
    - Each device’s zoning ACLTCAM programming **will be repeated on each member**
  - For E port-channels(or just ISLs) using IVR
    - Each host/target session that gets translated will take up ACLTCAM on each member
  - Ex. On a 9148S a 6 member port-channel could be split across the 3 fwd engines as follows: fc1/1, fc1/2, fc1/17, fc1/18, fc1/33 and fc1/34

Port Channel design considerations

Fabric switches – 9250i, 9148S, 9396S, 9132T, 9148T and 9396T

If ACLTCAM usage is high…

- Split large F port-channels into two separate port-channels each with half members
- Consider MDS 9396S, 9132T, 9148T and 9396T for larger scale deployments
  - S = Sixteen (16G switches)
  - T = Thirtytwo (32G switches)
  - **These fabric switches contain director class ASICs with much higher limits**

Port Channel design considerations

Fabric switches – 9250i, 9148S

- Check TCAM usage after major zoning operations

9148S has 3 FWD engines

98 entries in use due to zoning

Zoning region is the most likely to be exceeded

```
MDS9148S-1# show system internal acltcam-soc tcam-usage

TCAM Entries:
=============
                          Region1    Region2     Region3      Region4    Region5    Region6
Mod Fwd    Dir            TOP SYS    SECURITY    ZONING       BOTTOM     FCC DIS    FCC ENA
    Eng                   Use/Total  Use/Total   Use/Total    Use/Total  Use/Total  Use/Total
--- ---   ------ ---------- --------- ------------ --------- --------- ---------
1   1     INPUT      19/407      1/407     98/2852 *     4/407      0/0       0/0
1   1     OUTPUT      0/25       0/25        0/140        0/25       0/12      1/25
1   2     INPUT      19/407      1/407       0/2852 *     4/407      0/0       0/0
1   2     OUTPUT      0/25       0/25        0/140        0/25       0/12      1/25
1   3     INPUT      19/407      1/407       0/2852 *     4/407      0/0       0/0
1   3     OUTPUT      0/25       0/25        0/140        0/25       0/12      1/25
-------------------------------------------------------
```

ACLTCAM alert system messages

- New in NX-OS 8.3(1) and 8.4(1)
- Two types of alert system messages:
  - Region - When TCAM usage in a fwd-engine and region cross 80%
  - Total – When total TCAM usage crosses 60%

%ACLTCAM-SLOT1-4-REGION_RISING_THRESHOLD: ACL (region) (input | output) region usage (num of in use entries of total entries) exceeded 80% on forwarding engine (num)

%ACLTCAM-SLOT1-4-TOTAL_RISING_THRESHOLD: ACL total (input | output) usage (num of in use entries of total entries) exceeded 60% on forwarding engine (num)

NX-OS 8.3(1) includes all switches except MDS 9148S and 9250i.

NX-OS 8.4(1) includes MDS 9148S and 9250i!

F port-channel design considerations

Ports are allocated to fwd-engines according the following table:

Switch Type	Fwd Engines	Port Range(s)	Fwd-Eng Number	Zoning Region Entries
MDS 9148	3	fc1/25-36 & fc1/45-48	1	2852
		fc1/5-12 & fc1/37-44	2	2852
		1-4 & 13-24	3	2852
MDS 9250i	4	fc1/5-12 & eth1/1-8	1	2852
		fc1/1-4 & fc1/13-20 & fc1/37-40	2	2852
		fc1/21-36	3	2852
		ips1/1-2	4	2852

F port-channel design considerations

Continued…

Switch Type	Fwd Engines	Port Range(s)	Fwd-Eng Number	Zoning Region Entries
MDS 9148S	3	fc1/1-16	1	2852
		fc1/17-32	2	2852
		fc1/33-48	3	2852
MDS 9396S	12	fc1/1-8	0	49136
		fc1/9-16	1	49136
		…etc…	2-10	49136
		fc1/89-96	3	49136

F port-channel design considerations

Continued…

Switch Type	Fwd Engines	Port Range(s)	Fwd-Eng Number	Zoning Region Entries
MDS 9132T	2	fc1/1-16	0	49136
		fc1/17-32	1	49136
MDS 9148T	3	fc1/1-16	0	49136
		fc1/17-32	1	49136
		fc1/33-48	2	49136
MDS 9396T	6	fc1/1-16	0	49136
		fc1/17-32	1	49136
		…etc…	2-4	49136
		fc1/81-96	5	49136

For others see the Zoning Configuration Guide – **“Zoning Best Practice”**

MDS Internal CRC handling

Overview

- When MDS receives a good frame on a port it forwards it to the egress
- In rare cases frames can get corrupted internally due to bad hardware
  - These are then dropped
  - Sometimes difficult to detect
  - 5 possible stages where frames can get corrupted
- Frames that are *received* corrupted and cannot be FEC corrected are dropped at the ingress port –
  - These are standard CRC error frames and not included in this topic
- Several features handle this condition

1. Hardware fabric crc threshold
2. Port-monitor

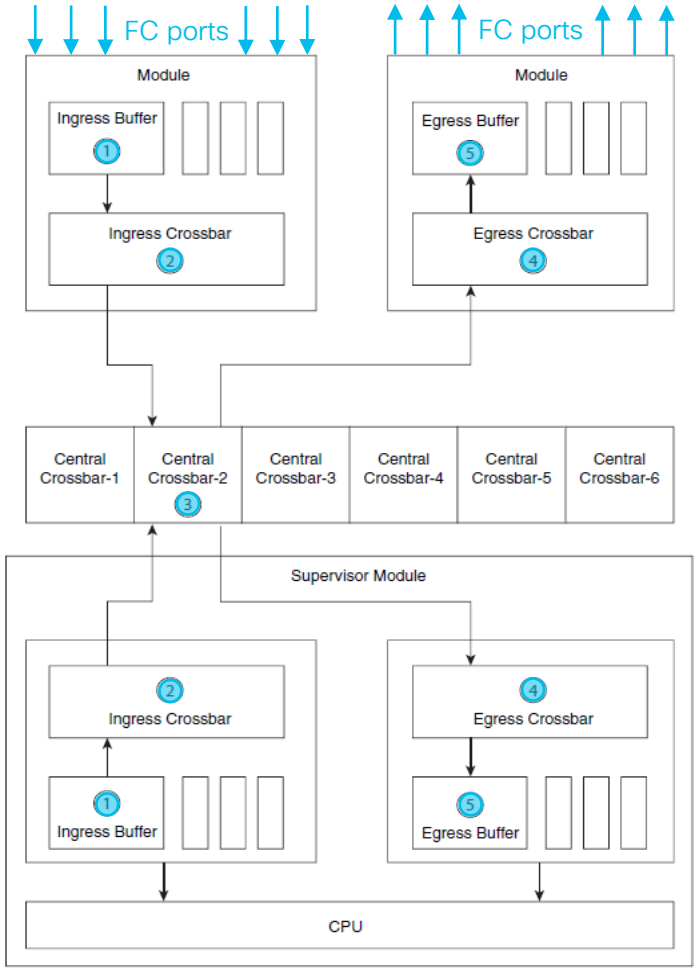
Internal CRC handling

Stages of Internal CRC Detection and Isolation

The five possible stages at which internal CRC errors may occur in a switch:

1. Ingress buffer of a module
2. Ingress crossbar of a module
3. Crossbar of a fabric module
4. Egress crossbar of a module
5. Egress buffer of a module

- *Normally* it is the lowest "stage" that detects the errors is the bad HW component



Internal CRC handling

hardware fabric crc threshold command

- Detects and powers down the module causing the internal CRC errors
- Supported on MDS 9700s only – all supervisors, modules and XBARs
- Enabled via the following configuration command:
  - hardware fabric crc threshold 1-100
- When detected failing module is powered down
- Threshold is per 24 hour period
- New in NX-OS 6.2(13)

Internal CRC handling

hardware fabric crc threshold command

- Sample messages when XBAR 5 was detected on MDS 9718 causing internal CRC errors
  - Note ": Fab_slot-23" is XBAR 5 MDS 9718 (23 – 18 = 5)

%XBAR-2-XBAR_MONITOR_INTERNAL_CRC_ERR: Fab_slot-23 detects CRC error at ingress stage2, putting it in failure state
%MODULE-2-XBAR_DIAG_FAIL: Xbar 5 (Serial number: sn) reported failure 23/1-23/0 due to XBM - CRC error detected on Module in device DEV_XBAR_COMPLEX (device error 0x0)
%PLATFORM-2-XBAR_DETECT: Xbar 5 detected (Serial number sn)
%PLATFORM-5-XBAR_PWRUP: Xbar 5 powered up (Serial number sn)
%PLATFORM-5-MOD_STATUS: Fabric-Module 5 current-status is MOD_STATUS_POWERED_UP
%MODULE-5-XBAR_OK: Xbar 5 is online (Serial number: sn)
%PLATFORM-5-MOD_STATUS: Fabric-Module 5 current-status is MOD_STATUS_ONLINE/OK
... 2 retries
%MODULE-2-XBAR_FAIL: Initialization of xbar 5 (Serial number: sn) failed
%PLATFORM-5-XBAR_PWRDN: Xbar 5 powered down (Serial number sn)
%PLATFORM-5-MOD_STATUS: Fabric-Module 5 current-status is MOD_STATUS_CONFIGPOWERED_DOWN

Internal CRC handling

Port-monitor

- Port-monitor can alert on internal CRC errors to and from the XBAR
- Add the following two counters into one port-monitor policy
  1. **`counter err-pkt-to-xbar poll-interval 300 delta rising-threshold 5 event 3 falling-threshold 0 event 3`**
  2. **`counter err-pkt-from-xbar poll-interval 300 delta rising-threshold 5 event 3 falling-threshold 0 event 3`**
- One policy will cover all ports

Sample messages

%PMON-SLOT2-3-RISING_THRESHOLD_REACHED: ASIC Error Pkt to xbar has reached the rising threshold (port=fc2/41 [0x10a8000], value=5) .
%PMON-SLOT1-3-RISING_THRESHOLD_REACHED: ASIC Error Pkt from xbar has reached the rising threshold (port=fc1/37 [0x1024000], value=5) .
%PMON-SLOT2-3-RISING_THRESHOLD_REACHED: ASIC Error Pkt to xbar has reached the rising threshold (port=fc2/41 [0x10a8000], value=8) .
%PMON-SLOT2-3-RISING_THRESHOLD_REACHED: ASIC Error Pkt to xbar has reached the rising threshold (port=fc2/41 [0x10a8000], value=5) .

Internal CRC handling

New messages

- NX-OS 8.3(2) introduced new messages to indicate a module receiving CRC error frames:
- No configuration necessary
- The modules that support this functionality are:
  - Cisco MDS 9700 48-Port 32-Gbps Fibre Channel Switching Module
  - Cisco MDS 9700 Fabric Module 3
  - Cisco MDS 9700 Supervisor 4

Sample messages

```
%SM15_USD-SLOT15-2-SM15_CRC_ERR: SM15 1 received packet(s) with CRC error on downlink(s) 0
%SM15_USD-SLOT12-2-SM15_CRC_ERR: SM15 1 received packet(s) with CRC error on downlink(s) 0
%SM15_USD-SLOT13-2-SM15_CRC_ERR: SM15 1 received packet(s) with CRC error on downlink(s) 0
%SM15_USD-SLOT16-2-SM15_CRC_ERR: SM15 1 received packet(s) with CRC error on downlink(s) 0
```

Device-alias

- device-alias(DA) is a way of naming PWWNs
- DAs are distributed on a fabric basis via CFS
- device-alias database is independent of VSANs
  - If a device is moved from one VSAN to another no DA changes are needed
- device-alias can run in two modes:
  - **Basic** – device-alias names used but PWWNs are substituted in config
  - **Enhanced** – device-alias names exist in configuration natively – Allows rename without zoneset re-activations
- device-alias are used in zoning, IVR zoning and port-security
- **copy running-config startup-config *fabric*** after making changes!

SAN Security

Secure management access

- Role-based access control – CLI, SNMP, Web

Secure management protocols

- SSH, SFTP, and SNMPv3
- Disable *feature telnet*

Secure switch control protocols

- TrustSec
- FC-SP (DH-CHAP)

AAA - RADIUS, TACACS+ and LDAP

- User, switch and iSCSI host authentication

Fabric Binding

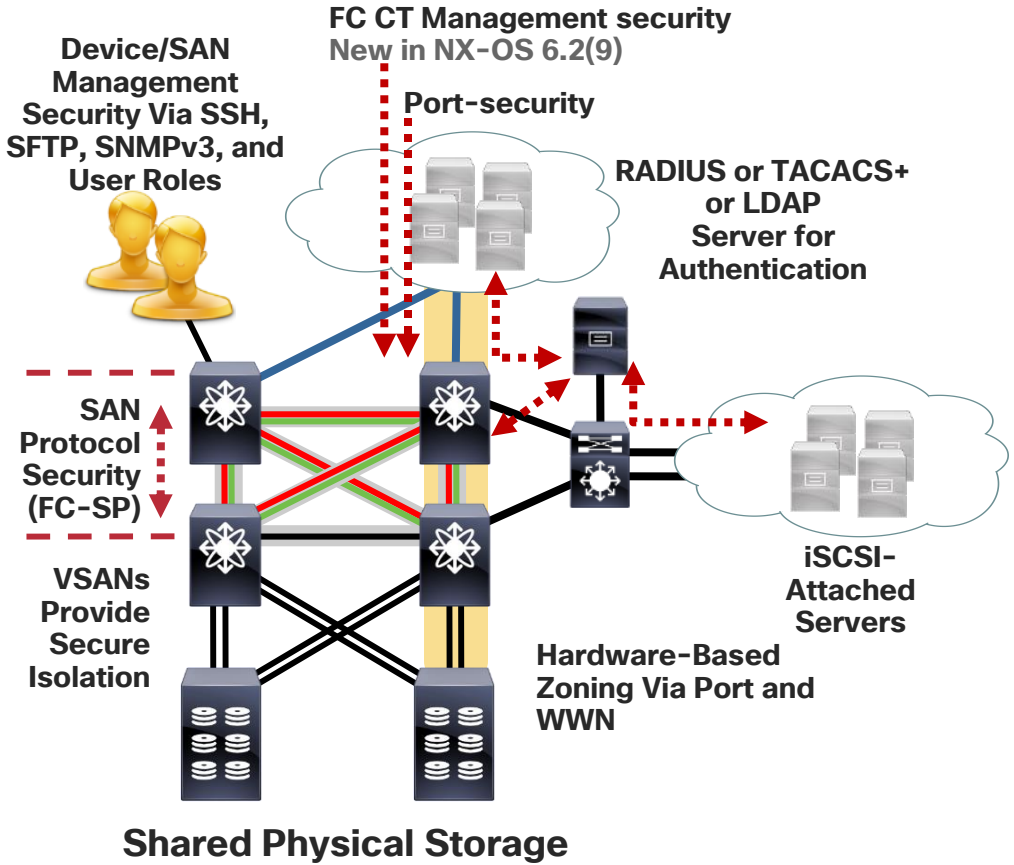
- Prevent unauthorized switches from joining fabric

Port-security

- Ensure only approved devices login to fabric

FC CT Management Security

- Ensure only approved devices send FC CT cmds



FC Management Security

- FC servers can send management requests into switches via FC
- By default the MDS does not prohibit this
- Two possibilities
  1. Zone Generic Services

     zone service via inband management

     Defaults to allowing queries *and* zoning changes
  2. Fibre Channel Common Transport (FC-CT) Management Security

     Prevents unauthorized FC-CT queries into a network

     Defaults to allow inband access to management information

FC Management Security

Zone Generic Services

- Defaults to read-write to allow local devices to make zoning changes
- To disable configure:
  - no zone gs read-write vsan x
- Or
  - **no system default zone gs read-write**

Switch default for all VSANs

FC Management Security

Fibre Channel Common Transport (FC-CT) Management Security

- Attached FC devices can send in Common Transport commands to query the Name Server
- Ensure only approved devices send FC CT cmds
- Introduced in NX-OS 6.2(9)
- To enable
  - fc-management enable
- If devices are to be permitted to query the name server then they must be added into the database:
  - fc-management database vsan <vsan>
  - pwwn <pwwn> feature <feature or all> operation <both or read>

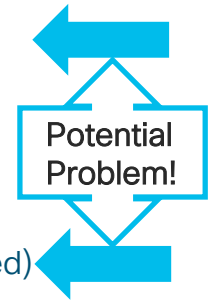
Static fcdomains

- Ensure static domain IDs are configured
- Configure: fcdomain domain 23 static vsan 237

MDS9710# show fcdomain vsan 237

VSAN 237
Local switch run time information:
State: Stable
Local switch WWN: 20:ed:54:7f:ee:ea:72:81
Running fabric name: 20:ed:54:7f:ee:ea:72:81
Running priority: 128
Current domain ID: 0x17(23)

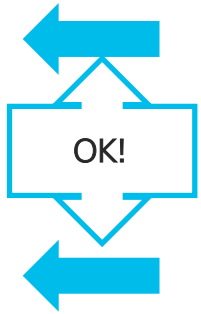
Local switch configuration information:
State: Enabled
...
Configured priority: 128
Configured domain ID: 0x00(0) (preferred)



MDS9710# show fcdomain vsan 237

VSAN 237
Local switch run time information:
State: Stable
Local switch WWN: 20:ed:54:7f:ee:ea:72:81
Running fabric name: 20:ed:54:7f:ee:ea:72:81
Running priority: 128
Current domain ID: 0x17(23)

Local switch configuration information:
State: Enabled
...
Configured priority: 128
Configured domain ID: 0x17(23) (static)



Static domain IDs prevent FCID reallocations during fabric reconfiguration

Forward Error Correction – FEC

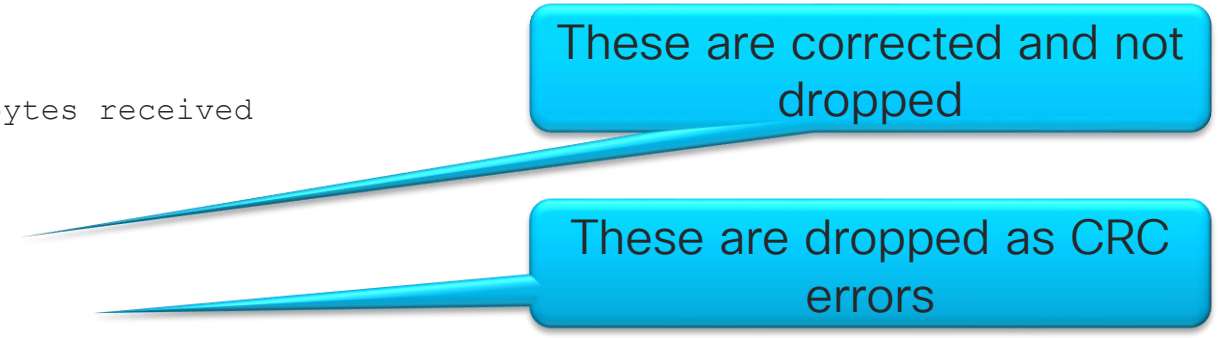
- Allows for the correction of some frame errors
- Almost zero latency penalty
- Can prevent SCSI timeouts/aborts
- Applies to 9700 FC, 9396S/T, 9132T, 9148T
- 16G – Applies fixed speed FC ISLs only
- 32G – On by default
- Configured via:
- switchport fec tts
- No reason not to use it!

```
9710-2# show interface fc1/8
fc1/8 is trunking
...
    Port mode is TE
    Port vsan is 1
    Speed is 16 Gbps
    Rate mode is dedicated
    Transmit B2B Credit is 500
    Receive B2B Credit is 500
    B2B State Change Number is 14
    Receive data field Size is 2112
    Beacon is turned off
    admin fec state is up
    oper fec state is up
    Trunk vsans (admin allowed and active) (1-2,20,237)
```

FEC is operational

Forward Error Correction - FEC

```
MDS9710-1# show interface fc9/1 counters details
fc9/1
      861181 frames, 75211536 bytes received
…
      0 fec corrected blocks
      0 fec uncorrected blocks
      Percentage Tx credits not available for last 1s/1m/1h/72h: 0%/0%/0%/0%
```



FEC corrected blocks can be used as a warning of some frame corruption

BB_Sc_N – Buffer-to-Buffer Credit Recovery

- A B2B credit can be lost in either of these scenarios:
  1. An error corrupts the start-of-frame (SoF) delimiter of a frame.
  2. An error corrupts an R_RDY primitive.

- For both cases a credit will be lost.
- Longstanding feature of ISLs
- Enabled by default on MDS for F/N ports starting in NX-OS 8.2(1)
- Enabled by default on MDS for F/NP ports starting in NX-OS 8.4(1)

- HBA must also support the feature

- Negotiates a value in FLOGI/ACC(FLOGI)

BB_Sc_N – Buffer-to-Buffer Credit Recovery

8.4(1)

```
MDS9706-2# show interface fc6/20
fc6/20 is up
    Hardware is Fibre Channel, SFP is short wave laser w/o OFC (SN)
...
    Transmit B2B Credit is 12
    Receive B2B Credit is 32
    B2B State Change: Admin(on), Oper(up), Negotiated Value(14)

MDS9706-2# show interface fc6/20 counters detailed
fc6/20
...
    0 BB_SCs credit resend actions, 0 BB_SCr Tx credit increment actions
```

BB_Sc_N is up

No credits recovered (yet)

CFS Distribution over IP

- Cisco Fabric Services(CFS) is used to distribute various features
  - callhome Callhome server
  - device-alias DDAS Daemon
  - fctimer Fibre Channel timer
  - ivr Inter-VSAN Routing
  - ntp Network Time Server
  - port-security Port Security Manager
  - radius Radius Daemon
  - role Role
  - syslogd System Logger Daemon
  - tacacs Tacacs Daemon
  - And several others…

CFS Distribution over IP

- CFS distributes the information via FC/FCoE ISLs (by default)
- Topologies with NPV switches do not get this distribution
- To distribute to NPV switches distribute via IP multicast
- Multicast addresses should be unique per fabric

```
show cfs status
Distribution : Enabled
Distribution over IP : Enabled - mode IPv4
IPv4 multicast address : 239.255.70.83
```

- Problems can arise when multicast address are non-unique per fabric
- Can lead to high CPU and CFS distribution problems if enabled
- Ensure “show cfs peers…” shows appropriate per-fabric peers only!

Clock Settings and Unified Timestamp

8.4(1)

- Important that switch clocks be set correctly and are consistent
- Especially important for diagnosis in multi-switch fabrics
- Use NTP!
- Sample configuration:
  - ntp server x.x.x.x
    clock timezone PST -8 0
    clock summer-time PDT 2 Sunday March 02:00 1 Sunday November 02:00 60
- Unified Timestamp – system timestamp format rfc5424

8.4(1)

2019-05-24T12:21:57Z MDS9710 %PORT-5-IF_UP: %$VSAN 237%$ Interface fc1/13 is up in mode F

"Z" means Zulu - UTC

(multi)pathtrace

Traces a path to a destination FCID or domain

Gives basic interface stats along the way both Ingress and Egress

- Speed
- Tx/Rx Bytes/sec
- Tx/Rx B2B credit
- Error
- Discards
- CRC
- TxWait(1s/1m/1h/72h)
- FibDrops
- ZoneDrops

(multi)pathtrace


```
mds9706-2# pathtrace domain 232 vsan 237 multipath
 I - Ingress
 E -  Egress
 M - Member Port-channel
 * - Fport
--------------------------------------------------------------------------------
PATH 1   MDS9706-2           MDS9710-1           MDS9396S-1
Domain   236                 10                  232
--------------------------------------------------------------------------------
Hop 1   MDS9706-2(port-channel6)(E)-------(I)(port-channel6)MDS9710-1
--------------------------------------------------------------------------------
Interface            Spd(G)  Tx(B/s)        Rx(B/s)        TxB2B    RxB2B   Errors  Discards  CRC
TxWait(1s/1m/1h/72h)         FibDrops           ZoneDrops
--------------------------------------------------------------------------------
(E)port-channel6      64.0    640            128            -        -       0       0         0
0%/0%/0%/0%                   -                 -

   (M)fc6/1           32.0    640            84             500      500     0       0         0
0%/0%/0%/0%                   0                 0

   (M)fc6/48          32.0    0              44             500      500     0       0         0
0%/0%/0%/0%                   0                 0

(I)port-channel6      64.0    132            640            -        -       0       0         0
0%/0%/0%/0%                   -                 -

   (M)fc9/1           32.0    88             640            500      500     0       0         0
0%/0%/0%/0%                   0                 0

   (M)fc9/48          32.0    352            0              500      500     0       0         0
```

Design Principles for Slow Drain and Congestion Isolation

SAN Congestion

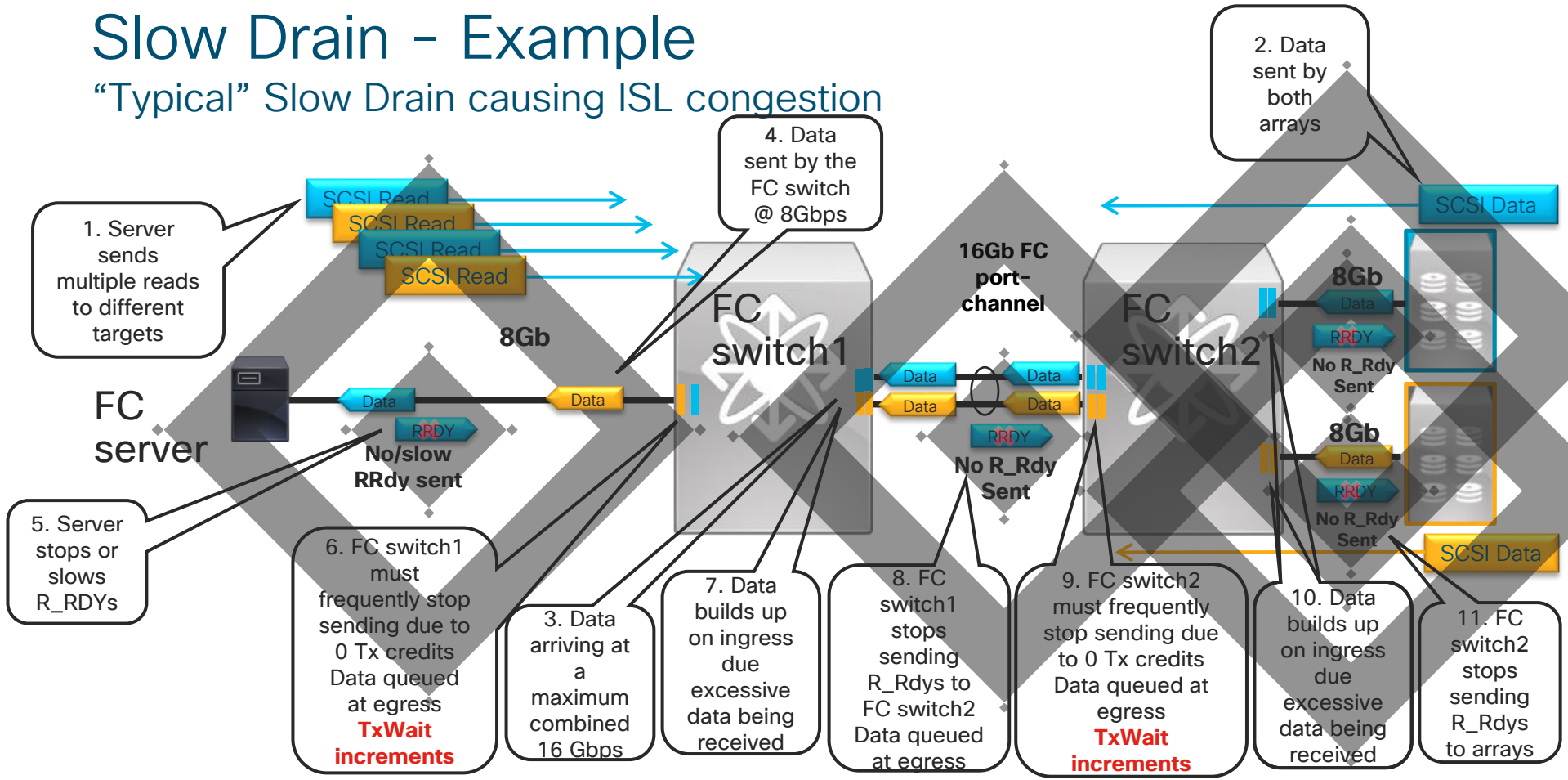
What is SAN congestion?

- SAN congestion is when some part of the SAN has frames that cannot be immediately transmitted
- Caused by two main reasons

1. “Traditional” slow drain
   - Devices purposely withholding buffer to buffer credits
   - Well known cause of poor performance
   - Easy to spot – TxWait is great!
2. Overutilization / Oversubscription
   - Devices requesting more data than they can receive at their link rate
   - Less known and definitely less understood
   - More difficult and tricky to spot

Slow Drain – Example

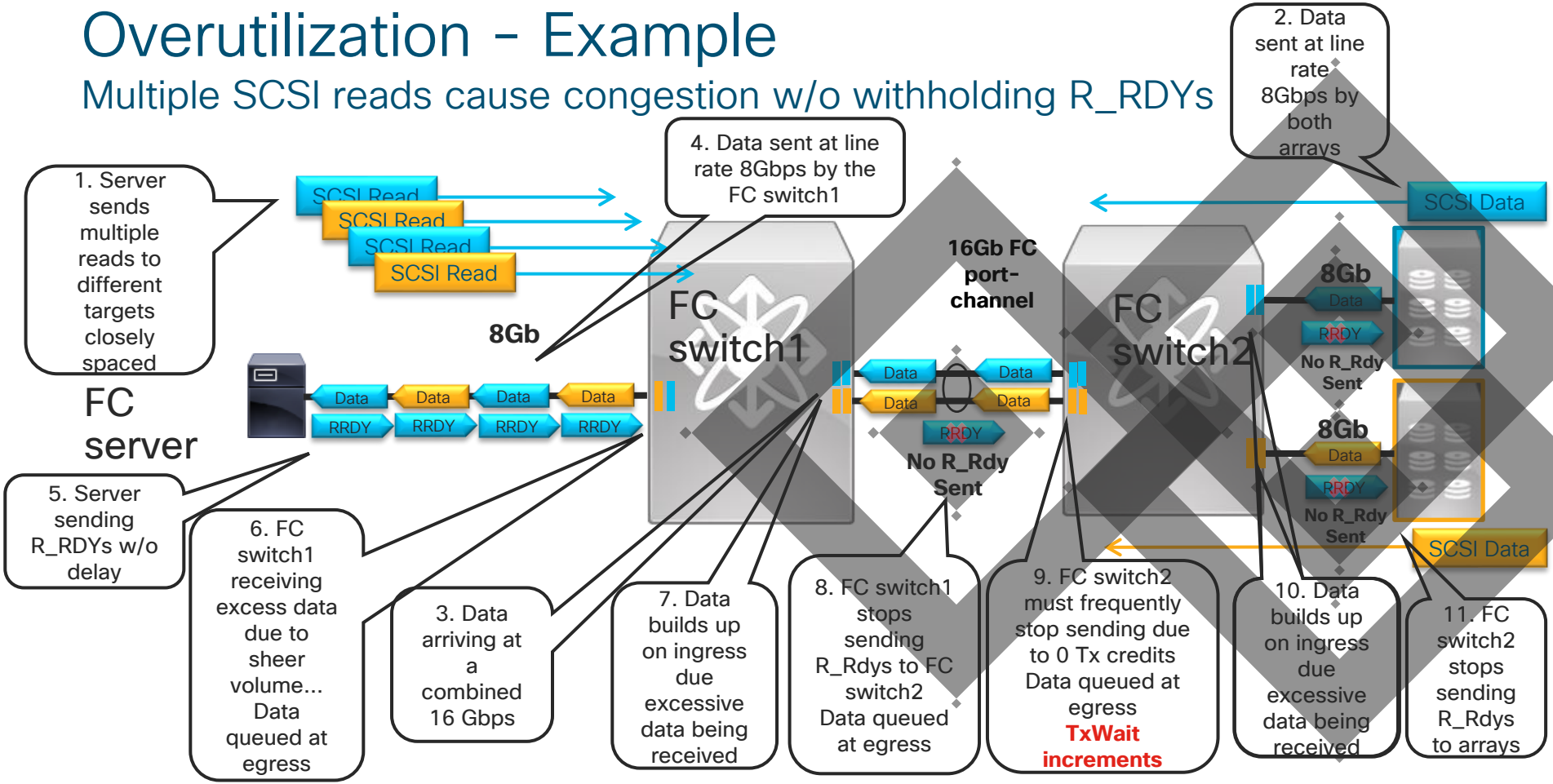
"Typical" Slow Drain causing ISL congestion



Both arrays and all devices utilizing ISLs are affected!

Overutilization – Example

Multiple SCSI reads cause congestion w/o withholding R_RDYs

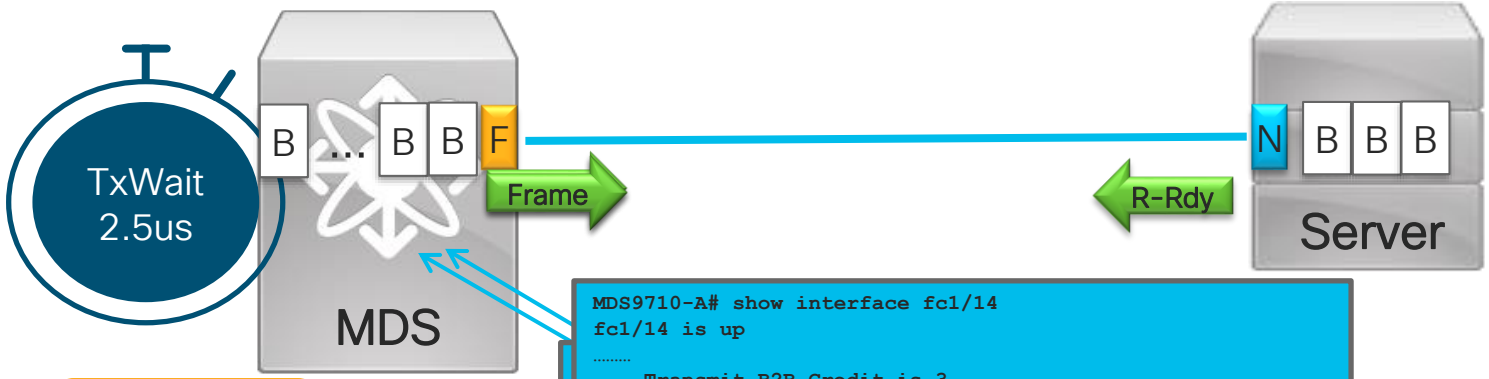


Not strictly “slow drain” but the effects are exactly the same!

SAN Congestion – Slow Drain

TxWait

- Time @ zero Tx credits is measured by TxWait
- Every 2.5us a port is at @ zero Tx credits TxWait increments by 1
- TxWait * 2.5 / 1,000,000 = Seconds
- TxWait is the foundation for ALL slow drain troubleshooting!



SAN Congestion – Slow Drain

show logging onboard txwait

- **Every 20 seconds TxWait is checked on every port**
- **Entry is recorded if there is 100ms or more TxWait in that 20 second interval**

```
---------------------------------
 Module: 1 txwait
---------------------------------
 Notes:
      - Sampling period is 20 seconds
      - Only txwait delta >= 100 m  are logged

-------------------------------------------------------------------------------
| Interface | Delta TxWait Time          | Congestion | Timestamp                |
|           | 2.5us ticks | seconds      |            |                          |
-------------------------------------------------------------------------------
|   fc1/14  |  7275593    |   18         |      90%   | Thu Nov 21 11:44:22 2019 |
|   fc1/13  |    89451    |    0         |       1%   | Thu Nov 21 11:44:22 2019 |
|   fc1/4   |   307260    |    0         |       3%   | Thu Nov 21 11:44:22 2019 |
|   fc1/3   |  3919949    |    9         |      48%   | Thu Nov 21 11:44:22 2019 |
|   fc1/38  |  2460156    |    6         |      30%   | Thu Nov 21 11:44:02 2019 |
|   fc1/37  |    80394    |    0         |       1%   | Thu Nov 21 11:44:02 2019 |
|   fc1/36  |    62623    |    0         |       0%   | Thu Nov 21 11:44:02 2019 |
|   fc1/35  |    85958    |    0         |       1%   | Thu Nov 21 11:44:02 2019 |
|   fc1/14  |  6067089    |   15         |      75%   | Thu Nov 21 11:44:02 2019 |
```

TxWait ticks in 20s

Converted to seconds

Congestion percentage is calculated

Recorded every 20 seconds

SAN Congestion Alerting and Prevention

Alerting – Overutilization - Port-monitor – tx-datarate counter

- **tx-datarate counter is used for detecting “overutilization”**
- Configure it as follows:

  counter tx-datarate poll-interval 10 delta rising-threshold 80 event 4 falling-threshold 79 event 4
- New in NX-OS 8.2(1) this is added into logging onboard

  show logging onboard datarate
- This is critical to being able to identify “over utilization”!
- When a port is running at 80+% for 10 seconds a “rising-threshold” alert
- When a port is running at 79-% for 10 seconds a “falling-threshold alert
- Time between alerts is when port is running at high utilization

SAN Congestion Alerting and Prevention

Alerting – Overutilization - Port-monitor – tx-datarate counter

```
MDS9710-1# show logging onboard datarate
--------------------------------
 Module: 1 datarate
--------------------------------
      - DATARATE INFORMATION FROM FCMAC
 ------------------------------------------------------------------------------------
| Interface |  Speed  |     Alarm-types      |   Rate  | Timestamp                  |
 ------------------------------------------------------------------------------------
|  fc1/13   |    4G   |  TX_DATARATE_FALLING |     1%  | Thu Nov  2 09:16:11 2017 |
|  fc1/13   |    4G   |  TX_DATARATE_RISING  |    83%  | Thu Nov  2 08:56:10 2017 |
|  fc1/13   |    4G   |  TX_DATARATE_FALLING |    73%  | Thu Nov  2 11:19:46 2017 |
|  fc1/13   |    4G   |  TX_DATARATE_RISING  |    83%  | Wed Nov  1 08:49:04 2017 |
```

Port was running at 80+% for over 26 hours!

Note: Doesn't indicate overutilization backpressure by itself. Use in conjunction with other slow drain indications

SAN Congestion Alerting and Prevention

Alerting – Port-monitor – AllPorts Example no portguard

```
port-monitor name AllPorts
logical-type all
  counter link-loss poll-interval 60 delta rising-threshold 5 event 2 falling-threshold 0 event 2
  counter invalid-crc poll-interval 60 delta rising-threshold 5 event 3 falling-threshold 0 event 3
  counter tx-discards poll-interval 60 delta rising-threshold 50 event 3 falling-threshold 10 event 3
  counter lr-rx poll-interval 60 delta rising-threshold 5 event 2 falling-threshold 1 event 2
  counter lr-tx poll-interval 60 delta rising-threshold 5 event 2 falling-threshold 1 event 2
  counter timeout-discards poll-interval 60 delta rising-threshold 50 event 3 falling-threshold 10 event 3
  counter credit-loss-reco poll-interval 60 delta rising-threshold 1 event 2 falling-threshold 0 event 2
  counter tx-credit-not-available poll-interval 1 delta rising-threshold 10 event 4 falling-threshold 0 event 4
  counter tx-datarate poll-interval 10 delta rising-threshold 80 event 4 falling-threshold 79 event 4
  counter err-pkt-from-port poll-interval 300 delta rising-threshold 5 event 3 falling-threshold 0 event 3
  counter err-pkt-to-xbar poll-interval 300 delta rising-threshold 5 event 3 falling-threshold 0 event 3
  counter err-pkt-from-xbar poll-interval 300 delta rising-threshold 5 event 3 falling-threshold 0 event 3
  counter tx-slowport-oper-delay poll-interval 1 absolute rising-threshold 80 event 4 falling-threshold 0 event 4
  counter txwait poll-interval 1 delta rising-threshold 10 event 4 falling-threshold 0 event 4
  monitor counter err-pkt-from-port
  monitor counter err-pkt-to-xbar
  monitor counter err-pkt-from-xbar
  no monitor counter sync-loss
  no monitor counter signal-loss
  no monitor counter invalid-words
  no monitor counter state-change
  no monitor counter rx-datarate
port-monitor activate AllPorts
```

Event 2 – Critical
Event 3 – Error
Event 4 – Warning

Policy applies to all ports

Hardware internal CRC errors are monitored

These counters are not monitored

SAN Congestion Alerting and Prevention

Alerting - Port-monitor – Sample output

```
MDS9710-1# show port-monitor active
Policy Name  : AllPorts
Admin status : Active
Oper status  : Active
Port type    : All Ports
```

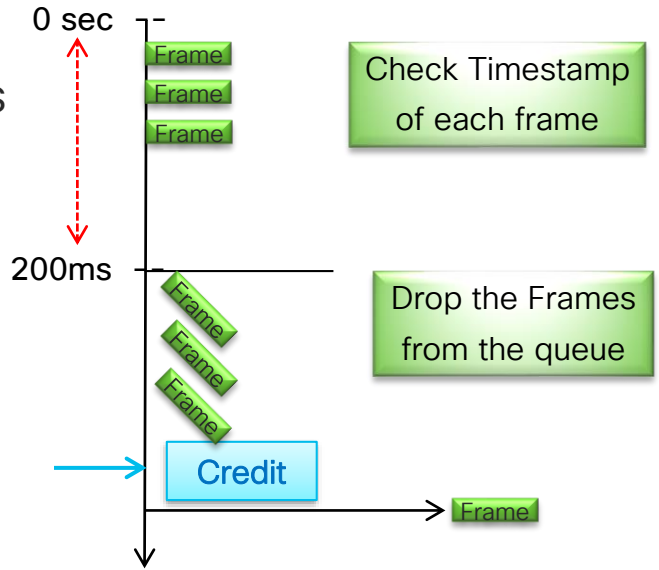
Counter	Threshold	Interval	Rising Threshold	event	Falling Threshold	event	Warning Threshold	PMON Portguard
Link Loss	Delta	60	5	2	0	2	Not enabled	Not enabled
Invalid CRC's	Delta	60	5	3	0	3	Not enabled	Not enabled
TX Discards	Delta	60	50	3	10	3	Not enabled	Not enabled
LR RX	Delta	60	5	2	1	2	Not enabled	Not enabled
LR TX	Delta	60	5	2	1	2	Not enabled	Not enabled
Timeout Discards	Delta	60	50	3	10	3	Not enabled	Not enabled
Credit Loss Reco	Delta	60	1	2	0	2	Not enabled	Not enabled
TX Credit Not Available	Delta	1	10%	4	0%	4	Not enabled	Not enabled
TX Datarate	Delta	10	80%	4	79%	4	Not enabled	Not enabled
ASIC Error Pkt from Port	Delta	300	5	3	0	3	Not enabled	Not enabled
ASIC Error Pkt to xbar	Delta	300	5	3	0	3	Not enabled	Not enabled
ASIC Error Pkt from xbar	Delta	300	5	3	0	3	Not enabled	Not enabled
TX-Slowport-Oper-Delay	Absolute	1	80ms	4	0ms	4	Not enabled	Not enabled
TXWait	Delta	1	10%	4	0%	4	Not enabled	Not enabled

Slow Drain Alerting and Prevention

Prevention – FC - Adjust Congestion Drop Threshold Lower

- Lowering congestion drop timeout value from 500ms to 200ms
- Frees up ingress buffer space quicker
- Can be set differently on F and E ports
- Congestion timeout for mode F should be smaller than(or equal to) mode E.
- Global command for switch
- Recommended for F ports
- Do not go below 200ms!

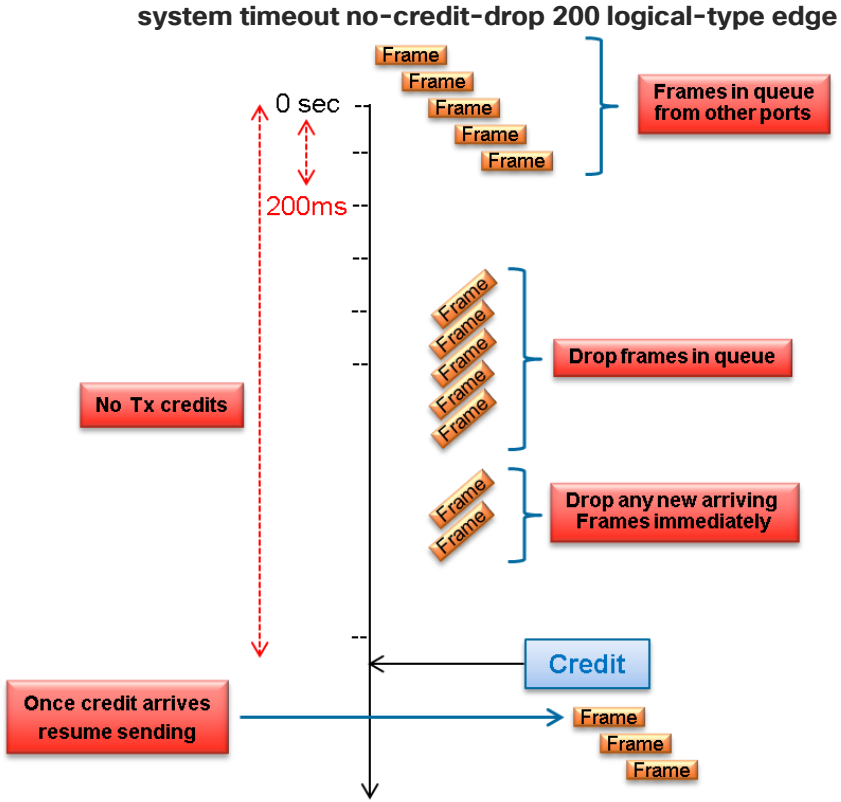
system timeout congestion-drop 200 logical-type edge



Slow Drain Alerting and Prevention

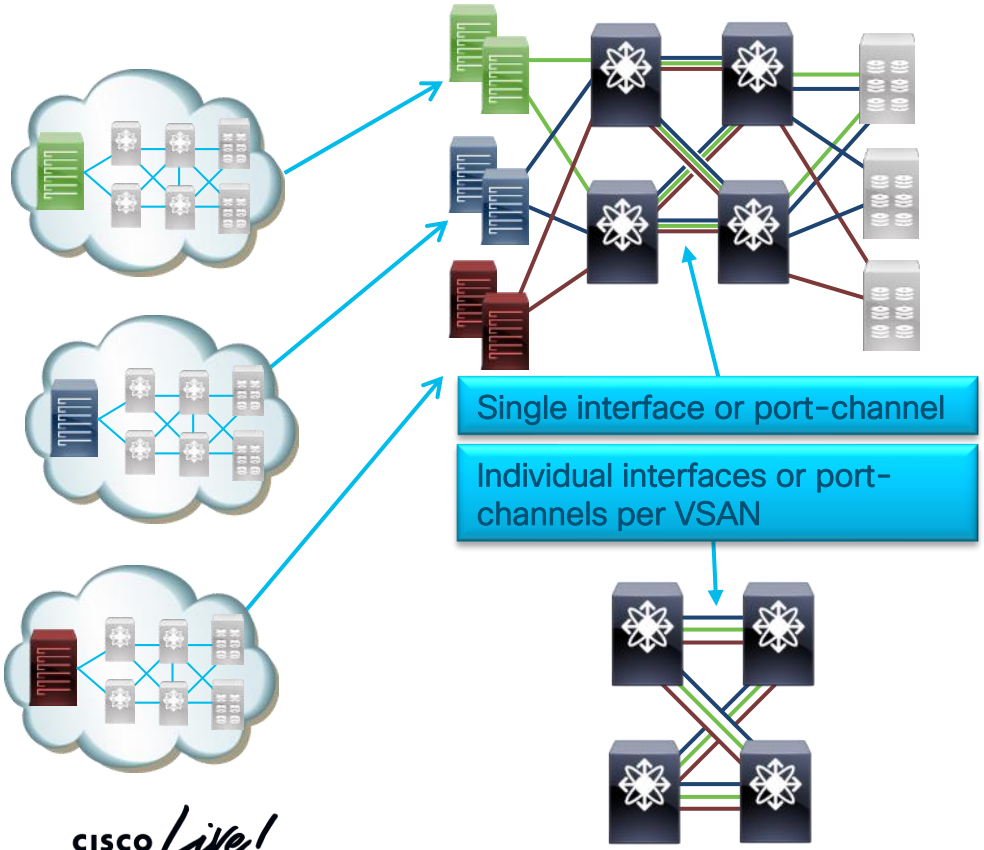
Prevention – FC - Setting the No Credit Drop Threshold

- No-credit-drop causes frames to be dropped immediately if the destination port is at 0 Tx credits for the time specified
- Should be used in conjunction with lowering congestion-drop threshold
- Recommended for F ports
- Can drastically improve ISL performance under slow drain conditions
- xxx_FORCE_TIMEOUT_ON/OFF counter
- By default no-credit-drop is not enabled



Slow Drain Alerting and Prevention

Prevention – SAN congestion - VSANs and ISLs



- ISLs trunking multiple VSANs will all experience congestion if one VSAN is experiencing it
- Separate physical ISLs for each VSAN provides better isolation
- With separate per-VSAN ISLs congestion on one VSAN will not affect the other VSANs

Slow Drain Alerting and Prevention

Congestion-Isolation of slow devices

- NX-OS 8.1(1) added the ER_RDY and Congestion-Isolation feature
- ISLs now have a new FC flow control mechanism – ER_RDY
- This internally partitions the physical link into 4 virtual links:
  - VL0 – Control traffic – 15 B2B credits
  - VL1 – High-priority traffic – 15 B2B credits
  - VL2 – Slow traffic – 40 B2B credits
  - VL3 – Normal traffic – 430 B2B credits
- ER_RDYs are now sent/managed separately by VL
- Initially all end device traffic is sent over the “Normal” VL - VL3
- Once a slow device is detected, it is put into the slow virtual link – VL2
- Note: B2B credit numbers are default for a 500 B2B credit ISL

Slow Drain Alerting and Prevention

Congestion-Isolation of slow devices

3 steps to enable

1. **ISLs must be put into ER_RDY Mode**
   - MDS-9710(config)# system fc flow-control er_rdy
   - ISLs flapped to change mode
   - Port-channel members can be flapped one at a time to be non-disruptive
2. **Feature congestion-isolation configured**
   - MDS9710(config)# feature congestion-isolation

Slow Drain Alerting and Prevention

Congestion-Isolation of slow devices

3. Port-monitor counters with portguard cong-isolate – 4 available
    1. credit-loss-reco
    2. tx-credit-not-available
    3. tx-slowport-oper-delay
    4. Txwait

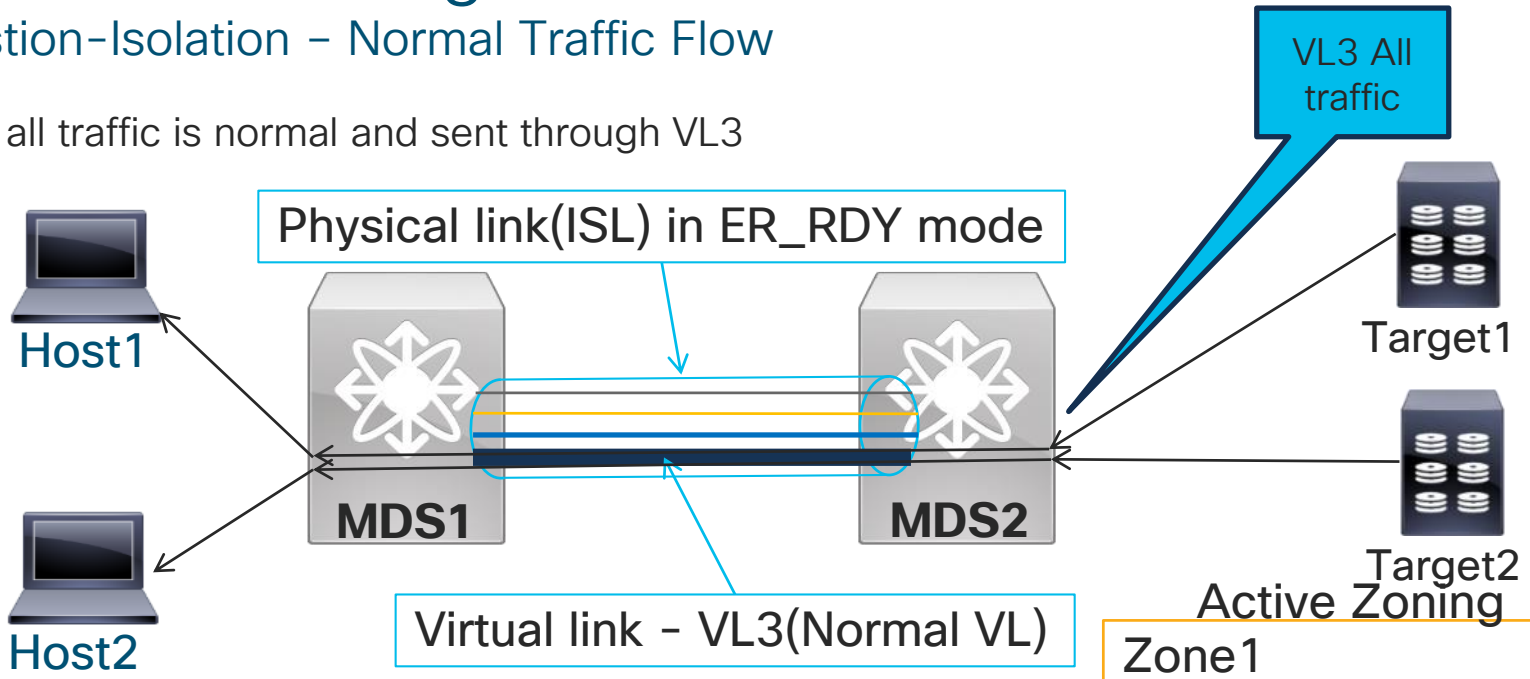
    Example:

- MDS9710(config-port-monitor)# counter txwait poll-interval 1 delta rising-threshold 40 event 4 falling-threshold 0 event 4 portguard cong-isolate
- The above will congestion isolate a device if it has 400ms of TxWait in a 1 second interval

Slow Drain Alerting and Prevention

Congestion-Isolation – Normal Traffic Flow

Initially all traffic is normal and sent through VL3



Active Zoning

Zone1

Host1 <-> Target1

Zone2

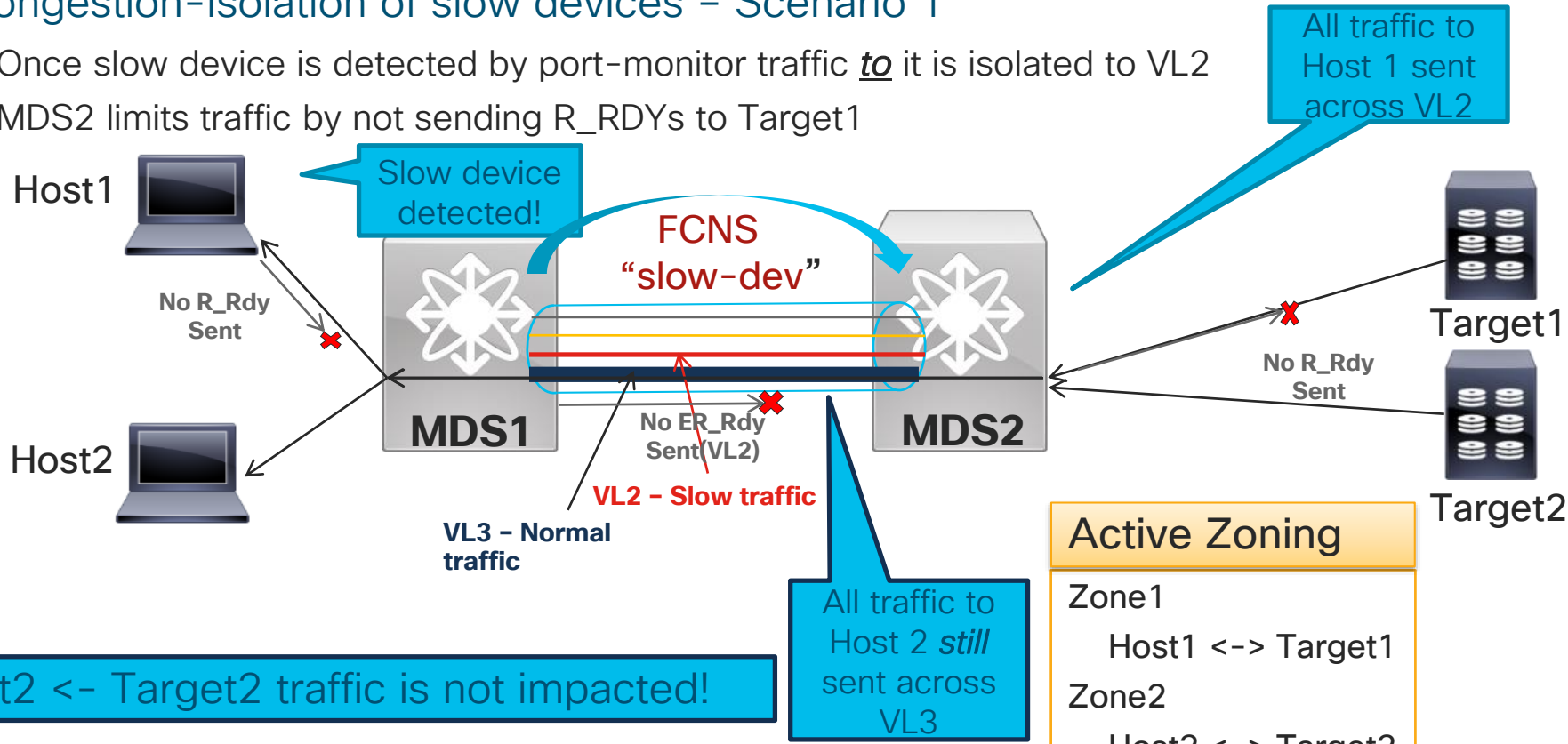
Host2 <-> Target2

Slow Drain Alerting and Prevention

Congestion-Isolation of slow devices – Scenario 1

Once slow device is detected by port-monitor traffic ***to*** it is isolated to VL2

MDS2 limits traffic by not sending R_RDYs to Target1

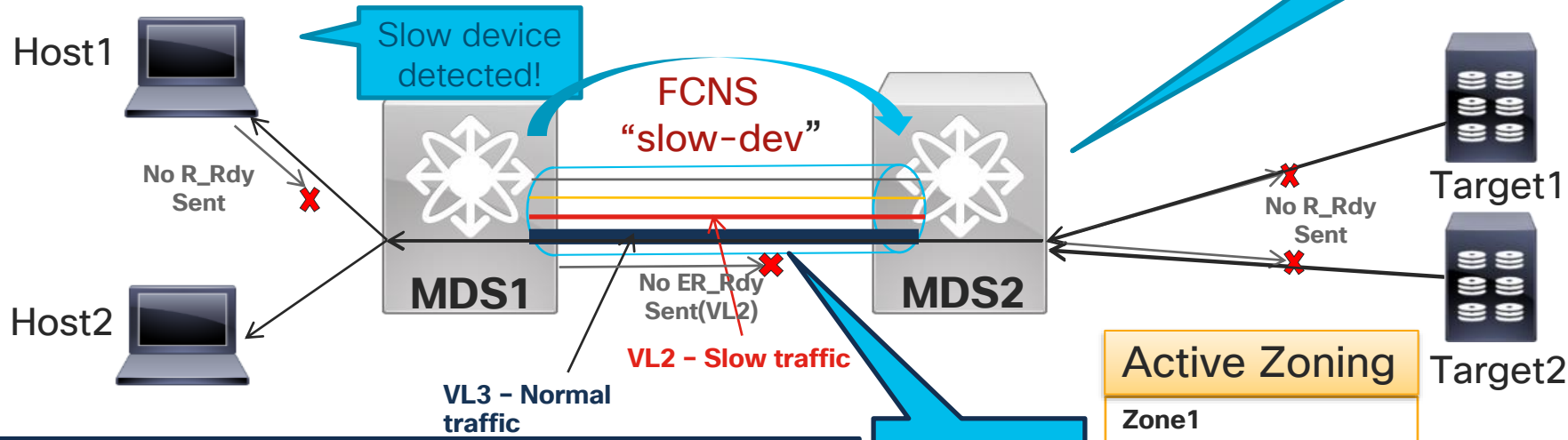


Slow Drain Alerting and Prevention

Congestion-Isolation of slow devices - Scenario 2

All traffic to the slow device is isolated to VL2 impacting Host 2 as well

Zoning overlaps cause traffic from common targets to be affected



Slow Drain Alerting and Prevention

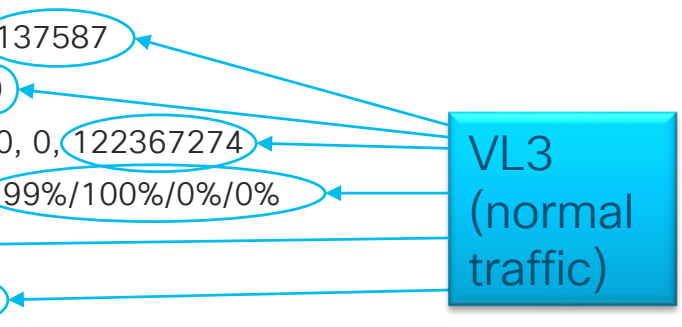
Congestion-Isolation – ER_RDY TxWait, Transitions to Zero

8.4(1)

TxWait and Transitions to Zero are now listed per Virtual Link(VL)

VL0 – Ctrl
VL1 – High
VL2 – Slow
VL3 – Data (Normal)

```
MDS9710-1# show interface fc9/48 counters
fc9/48
    5 minutes input rate 1440 bits/sec, 180 bytes/sec, 5 frames/sec
    5 minutes output rate 8711104 bits/sec, 1088888 bytes/sec, 451 frames/sec
...
    Transmit B2B credit transitions to zero for VL 0-3: 0, 0, 0, 137587
    Receive B2B credit transitions to zero for VL 0-3: 0, 0, 0, 0
    2.5us TxWait due to lack of transmit credits for VL 0-3: 0, 0, 0, 122367274
    Percentage Tx credits not available for last 1s/1m/1h/72h: 99%/100%/0%/0%
    Transmit B2B credit remaining for VL 0-3: 15, 15, 40, 0
    Receive B2B credit remaining for VL 0-3: 15, 15, 40, 430
```



Slow Drain Alerting and Prevention

Congestion-Isolation – ER_RDY TxWait, Transitions to Zero

8.4(1)

Logging Onboard TxWait are now listed per Virtual Link(VL)

```
MDS9710-1# show logging onboard txwait module 9

---------------------------------
 Module: 9 txwait count
---------------------------------
…
 Notes:
     - Sampling period is 20 seconds
     - Only txwait delta >= 100 ms are logged

-------------------------------------------------------------------------------------------
| Interface | Virtual Link | Delta TxWait Time     | Congestion | Timestamp               |
|           |              | 2.5us ticks | seconds |            |                         |
-------------------------------------------------------------------------------------------
|   fc9/48  | VL2(Slow)    |  8000000    |   20    |     100%   | Fri Apr 26 17:05:00 2019 |
|   fc9/48  | VL2(Slow)    |  3763318    |    9    |      47%   | Fri Apr 26 17:04:40 2019 |
|   fc9/48  | VL3(Data)    |  5433347    |   13    |      67%   | Fri Apr 26 17:01:19 2019 |
|   fc9/48  | VL3(Data)    |  8000000    |   20    |     100%   | Fri Apr 26 17:00:59 2019 |
```

VL0 – Ctrl
VL1 – High
VL2 - Slow
VL3 – Data (Normal)

Slow Drain Alerting and Prevention

Congestion-Isolation – Caveats, Disclaimers, Provisos, Fine Print, etc…

- Only works for “traditional” slow drain
- Only works across ISLs so has no effect in single switch fabrics
- FCoE is currently not supported
- If slow port is NPV connection all devices on port will be isolated
- No automatic de-isolation currently
- Several congestion-isolation commands available
  - Manually configure devices as slow
  - Display devices that are congestion-isolated
- See the Configuration Guide for more details

Slow Drain Alerting and Prevention

MDS 9700 FCIP TxWait and RxWait

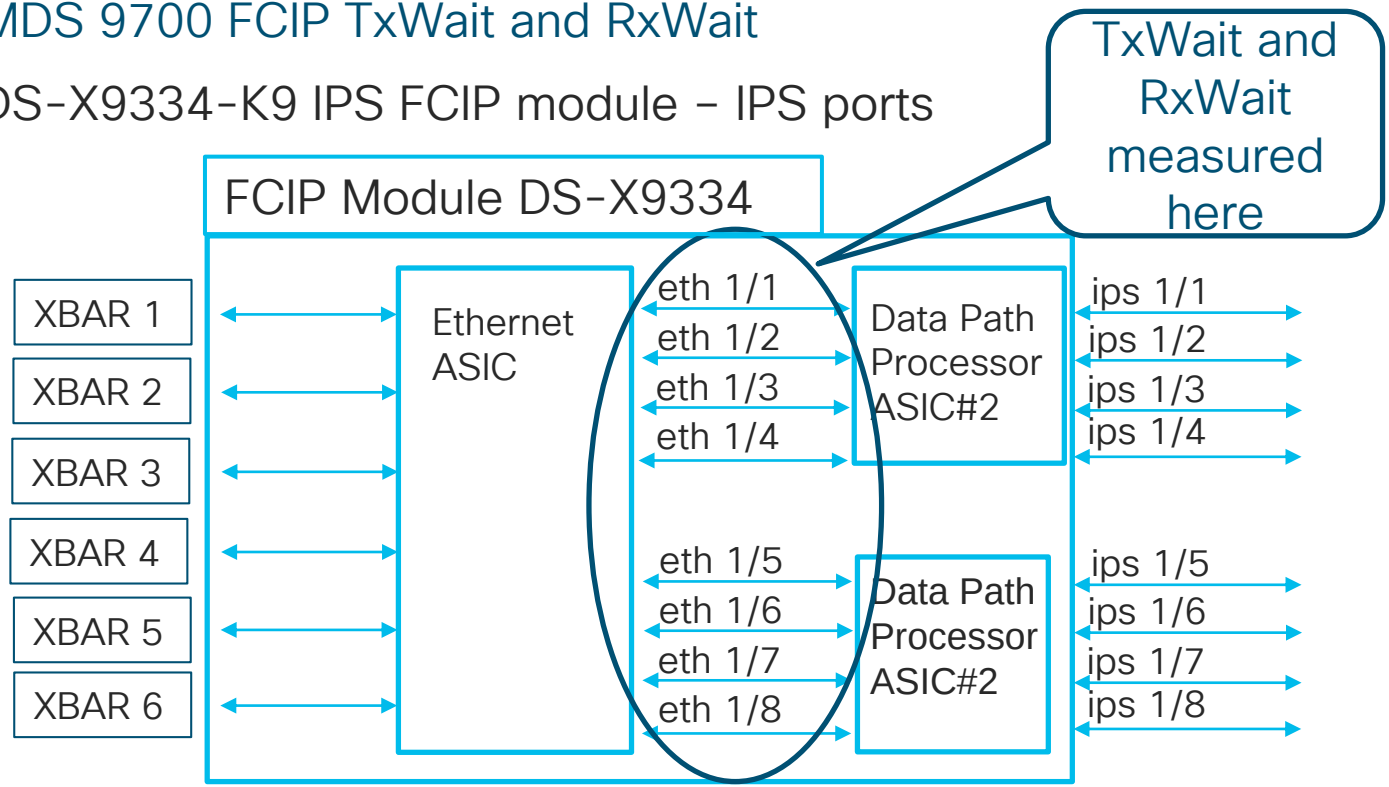
- DS-X9334-K9 IPS FCIP module has capability to measure TxWait and RxWait on an internal ethernet ASIC in the data path
- Occurs via Priority Flow Control(PFC) Pause
- Internal ethernet ASIC maps 1:1 to IPS ports
  - eth8/1 -> ips8/1
  - eth8/2 -> ips8/2
  - etc...
- show logging onboard txwait | rxwait
  - txwait will include internal ethernet ports as well as FC ports
- slot x show hardware internal txwait-history | rxwait-history

Slow Drain Alerting and Prevention

MDS 9700 FCIP TxWait and RxWait

8.4(1)

DS-X9334-K9 IPS FCIP module – IPS ports

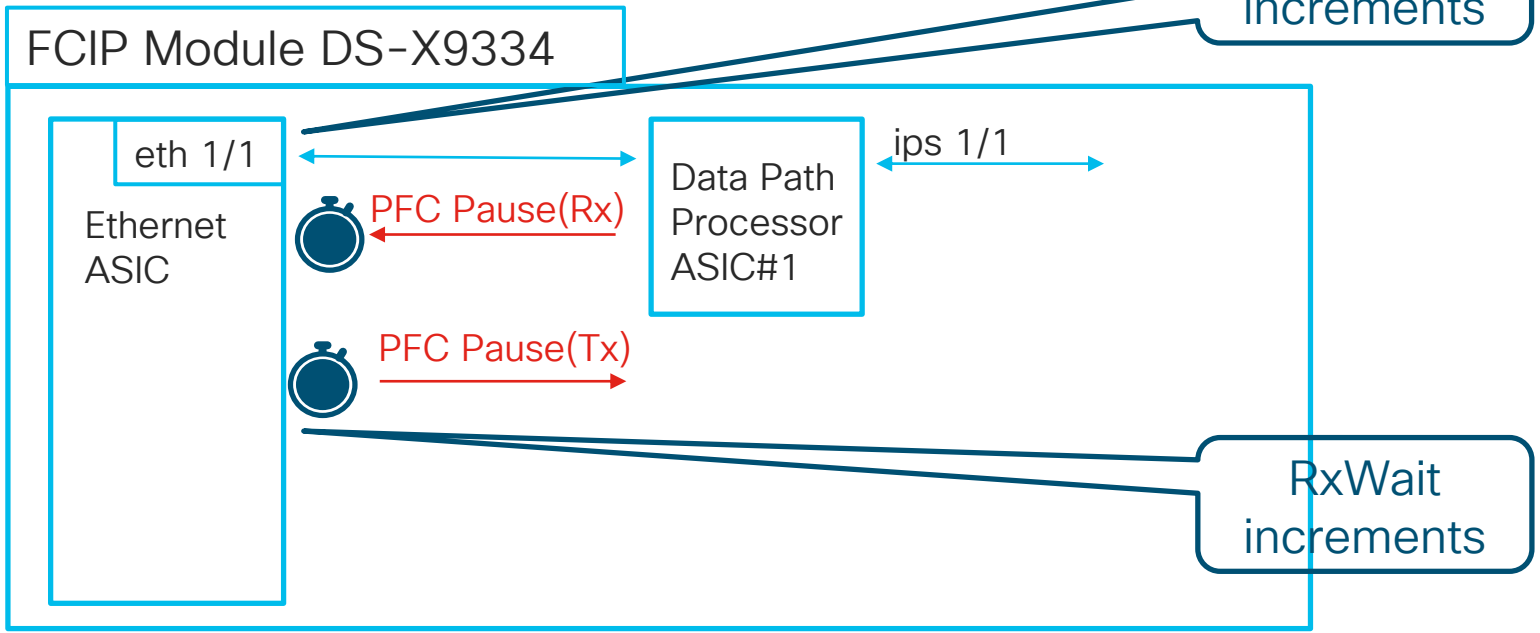


Slow Drain Alerting and Prevention

MDS 9700 FCIP TxWait and RxWait

8.4(1)

TxWait increments when Ethernet ASIC receives PFC Pause



RxWait increments when Ethernet ASIC transmits PFC Pause

Slow Drain Alerting and Prevention

MDS 9700 FCIP TxWait


```
MDS9710-1# show logging onboard txwait module 8
--------------------------------
 Module: 8 txwait
--------------------------------
 Notes:
     - Sampling period is 20 seconds
     - Only txwait delta >= 100 ms are logged

-------------------------------------------------------------------------------
| Interface     | Delta TxWait Time     | Congestion | Timestamp                |
|               | 2.5us ticks | seconds |            |                          |
-------------------------------------------------------------------------------
| Eth8/5(VL3)   |  7489866    |   18    |      93%   | Thu Jun  6 10:49:59 2019 |
| Eth8/1(VL3)   |  5678888    |   14    |      70%   | Thu Jun  6 10:49:39 2019 |
| Eth8/5(VL3)   |  2814708    |    7    |      35%   | Thu Jun  6 10:49:39 2019 |
| Eth8/1(VL3)   |  2669196    |    6    |      33%   | Thu Jun  6 10:49:19 2019 |
```

- TxWait indicates congestion going to the IPS ports
- RxWait indicates congestion going to local FC ports from the IPS ports
- Congestion could be due to high utilization or TCP retransmits on FCIP interfaces

Slow Drain Alerting and Prevention

Summary - Proactive

- Configure a lower congestion-drop on F ports
  - system timeout congestion-drop 200 logical-type edge
  - System timeout fcoe congestion-drop 200 mode edge
  - Don't go below 200ms!
- Configure no-credit-drop on F ports
  - system timeout no-credit-drop 100 logical-type edge
  - 200ms – safe, 100ms – aggressive, 50ms – Very aggressive
- Configure pause-drop on F ports
  - system timeout fcoe pause-drop 100 mode edge
  - 200ms – safe, 100ms – aggressive

Slow Drain Alerting and Prevention

Summary - Proactive

- Configure port-monitor policy(s)
  - Use samples included in port-monitor section
  - Make sure to include tx-datarate at 80%/79% polling-interval 10 seconds
- Configure switchport logical-type core on interfaces to NPV switches
  - MDS9710(config-if)# switchport logical-type core
- Consider separate physical ISL topology for older/slower VSANs
- Consider implementing congestion-isolation feature

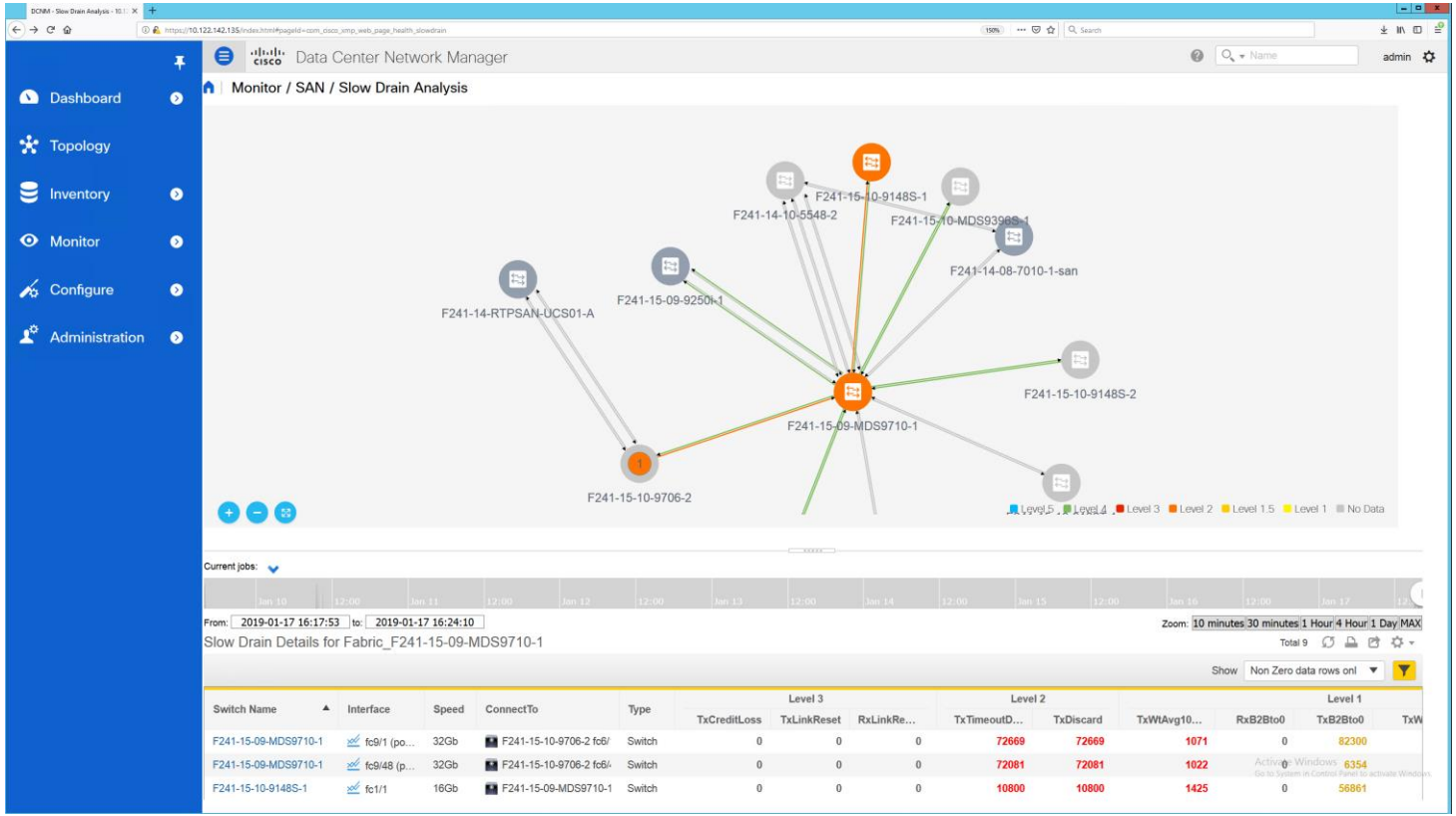
Overutilization Prevention

Summary - Reactive

- This is not an MDS switch issue
- First step is identification – use port-monitor tx-datarate counter
- Increase speed of HBA
- Increase number of HBAs
- Some arrays have the ability to rate limit data traffic to specific servers

Slow Drain Alerting and Prevention

DCNM 11.1(1) Slow drain Analysis enhancements – Topology graph



Design Principles for SAN Analytics

32G FC Analytics – Switch Native IO visibility

8.4(1)

DS-X9648-1536K9 4/8/16/32G Fibre Channel module for MDS 9700

MDS 9132T 32 Port 4/8/16/32G Fibre Channel Switch

MDS 9148T 48 Port 4/8/16/32G Fibre Channel Switch

8.4(1)

MDS 9396T 96 Port 4/8/16/32G Fibre Channel Switch

8.4(1)

Pervasive
No appliance
No probes
Always on

High Performance
Onboard analytics engine

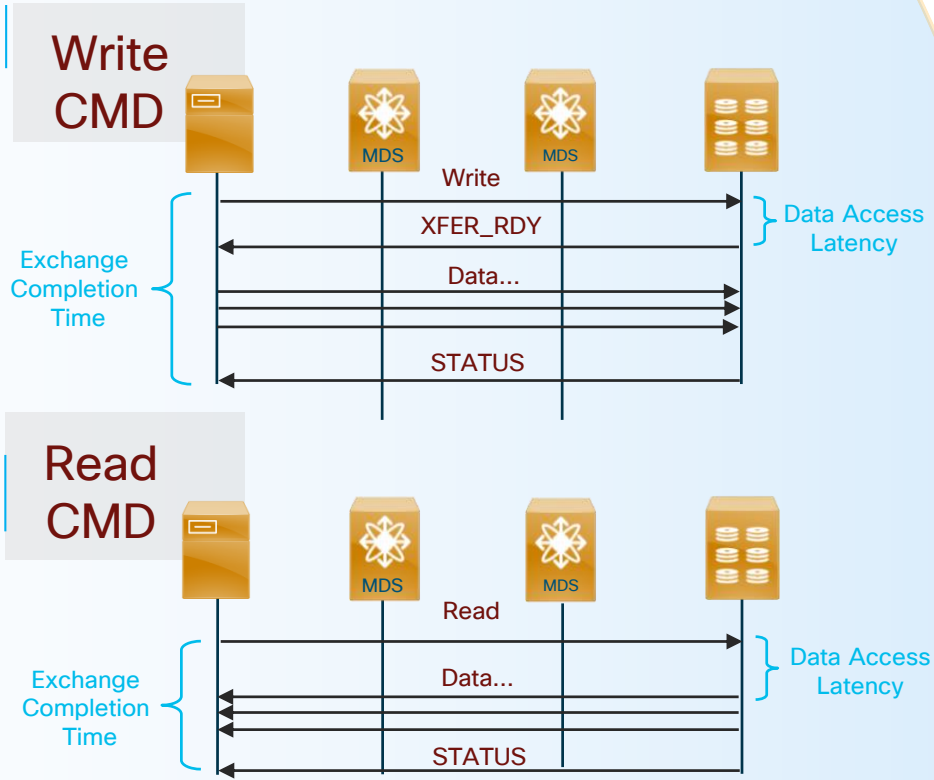
End to End Visibility
for trouble shooting

SCSI level flow data

Scale with MDS 9700
Director Platform

Analytics functionality available in NX-OS 8.2(1) – 9148T, 9396T added in 8.4(1)

SCSI Performance Exchange Metrics

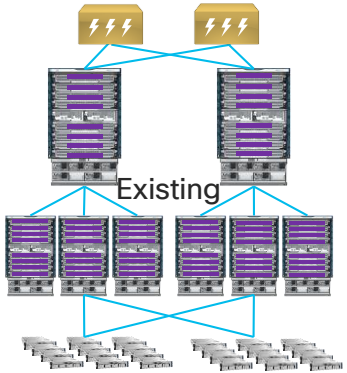


- Exchange Completion time
- Data Access Latency
- Outstanding IO
- IOPS & other flow level counters
- Per flow timeout drop frames
- Failed exchanges, IO retransmissions
- SCSI error conditions (Aborts, Rejects, etc)
- IO block size & other detailed flow level stats

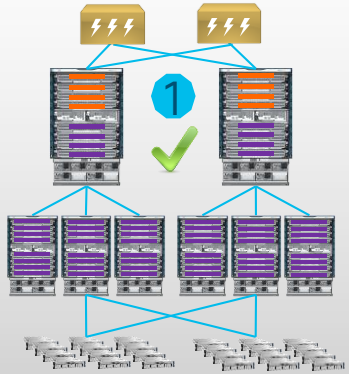
To be measured at SID-DID-LUN level

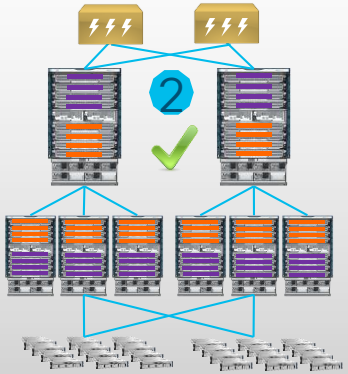
Best Practice for 32G FC module on MDS 9700

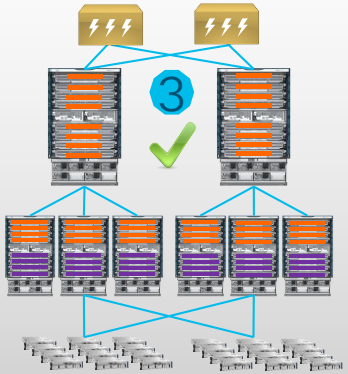
SAN upgrade with 16G FC on MDS 9700 – Seamless adoption of 32G FC



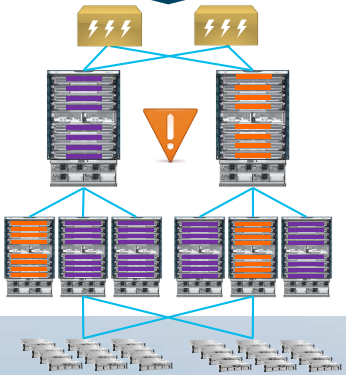
Cisco Recommended







Possible



- Native switch-integrated fabric-wide analytics
- 32G module should “see” all frames at least once!
- Investment protection of 16G FC module on MDS 9700
- Seamless & non-disruptive insertion of 32G FC module
- High speed ISL ➔ Increase performance with fewer links

32G FC module 16G FC module

SAN Analytics
What's new?

- NX-OS 8.3(2)
  - SAN Telemetry Streaming
  - Show analytics system-load
  - Updated query syntax – Added "asc" and "desc" sorting keywords

- NX-OS 8.4(1)
  - **Support for MDS 9148T and 9396T**
  - NVMe support
  - Showanalytics – Added --minmax, --errorsonly, --evaluate-npuload, --vsan-thput, --top, --outstanding-io, --alias, --limit options

SAN Analytics

Initial configuration - Licensing

8.4(1)

- NX-OS 8.3(1) and later license is SAN_ANALYTICS_PKG
  - 120 day grace period
- Switch commands
- To enable feature
  - feature analytics
- To enable a specific port
  - analytics type fc-scsi | fc-nvme | fc-all
- Note analytics is a licensed feature
  - Traditional license – 3 and 5 year term
  - Smart License – Subscription based

8.4(1)

SAN Analytics
Query Types

- 2 query types
  1. **Pull Query**
     - A one-time query used to extract the flow information that is stored in a database at the instant the query is executed. The output is in JSON format.
     - Overlay CLI—A predefined pull query that displays the flow metrics in a user-friendly tabular format.
  2. **Periodic-export(push) Query**
     - A recurring query installed to periodically extract the flow metrics that are stored in a database. The output is in JSON format.
     - Used in DCNM SAN Insights

SAN Analytics

Pull Query format

- show analytics query ‘query syntax’
- Query Syntax

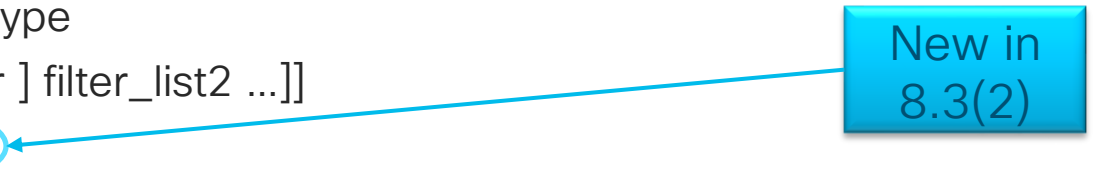
  select all | column1[,column2...]

  from analytics_type.view_type

  [where filter_list1 [[and | or] filter_list2 ...]]

  [sort column [asc | desc]

  [limit number]



SAN Analytics

Pull Query Examples

```
MDS9710-1# show analytics query 'select all from fc-scsi.scsi_initiator'
{ "values": {
        "1":    {
                "port": "fc9/16",
                "vsan": "1",
                "initiator_id": "0xc0100",
...
                "read_io_timeouts":     "129",
                "write_io_timeouts":    "0"
        },
        "2":    {
                "port": "fc9/32",
                "vsan": "1",
                "initiator_id": "0xc00e0",
...
                "read_io_timeouts":     "2",
                "write_io_timeouts":    "4"
        }
}}
```

60+ metrics total!

SAN Analytics

Error Metrics

- The following error metrics are available:

Error Metric	Description
read|write_io_aborts	ABTS sent for SCSI reads and writes
read|write_io_timeouts	SCSI reads and writes that did not complete in 2 seconds
read|write_io_failures	SCSI reads and writes with bad completion status

read_io_failures and write_io_failures include:

Error Metric	Description
read|write_io_scsi_reservation_conflict_count	0x18 – Reservation Conflict Status for SCSI reads and writes
read|write_io_scsi_queue_full_count	0x28 – A.K.A. Task Set Full Status for SCSI reads and writes
read|write_io_scsi_check_condition_count"	0x02 – Check Condition Status for SCSI reads and writes
read|write_io_scsi_busy_count	0x08 – Busy Status for SCSI reads and writes

Other completion status not tracked separately but included in failures

SAN Analytics

Pull Query Examples – Query with “where” key filter

```
MDS9710-1# show analytics query 'select all from fc-scsi.scsi_initiator where port=fc9/16'
{ "values": {
        "1":    {
                "port": "fc9/16",
                "vsan": "1",
                "initiator_id": "0xc0100",
...
                "write_io_initiation_time_max": "1277985",
                "read_io_inter_gap_time_min":   "26",
                "read_io_inter_gap_time_max":   "33834735",
                "write_io_inter_gap_time_min":  "26",
                "write_io_inter_gap_time_max":  "56612131",
                "read_io_aborts":       "0",
                "write_io_aborts":      "0",
                "read_io_failures":     "0",
                "write_io_failures":    "0",
                "read_io_timeouts":     "4",
                "write_io_timeouts":    "2"
```

SAN Analytics

showanalytics Overlay CLI – Syntax – New options in blue

MDS9710-1# showanalytics -help
ShowAnalytics --info <options> | --errors <options> | --errorsonly <options> | --minmax <options> | --evaluate-npuload <options> | --vsan-thput <options> | --top <options> | --outstanding-io <options> | --help

OPTIONS :

--info	Provide information about ITLs
--minmax	Provide Min/Max/Peak values of ITLs
--errors	Provides error metrics for all ITLs
--errorsonly	Provides error metrics for ITLs. Only display ITLs with non-zero errors
--evaluate-npuload	Provides per port NPU load
--vsan-thput	Provides per vsan scsi traffic rate for interface
--top	Provides top ITLs based on key. Default key is IOPS
--outstanding-io	Provides Outstanding io per ITL for an interface

SAN Analytics

showanalytics Overlay CLI – Syntax – New arguments in blue

8.4(1)

```
ARGUMENTS:
---------
      --initiator    <initiator_fcid>   Specifies initiator FCID in the format 0xDDAAPP
      --target       <target_fcid>      Specifies target FCID in the format 0xDDAAPP
      --lun          <lun_id>           Specifies LUN ID in the format XXXX-XXXX-XXXX-XXXX
      --interface    <interface>        Specifies Interface in format module/port
      --alias                           Prints device-alias for initiator and target.
      --limit        <itl_limit>        Maximum number of ITL records to display. Valid
                                        range 1-20000. Default = 20000
      --module       <mod1,mod2>        Specifies module list for --evaluate-npuload
                                        option example 1,2
      --key          <iops|thput|ect>   Defines the key value for the --top option
      --progress                        Provides progress for --top option. Should not be
                                        used on console
      --refresh                         Refreshes output of --outstanding-io
```

8.4(1)

SAN Analytics

showanalytics - Overlay CLI – Example - Info

```
MDS9710-1# showanalytics  --initiator-itl --info

 Interface fc9/32
+---------------------------------------+-----------+-----------------+----------------+
| VSAN|I|T|L                            | Avg IOPS  | Avg Thput (B/s) | Avg ECT (usec) |
+---------------------------------------+-----------+-----------------+----------------+
|                                       | Read|Write | Read|Write      | Read|Write     |
|                                       |           |                 |                |
| 1|0xc00e0|0x560020|0000-0000-0000-0000 | 0|0       | 0|0             | 10208|6032     |
+---------------------------------------+-----------+-----------------+----------------+

 Interface fc9/16
+---------------------------------------+-----------+--------------------+----------------+
| VSAN|I|T|L                            | Avg IOPS  | Avg Thput (B/s)    | Avg ECT (usec) |
+---------------------------------------+-----------+--------------------+----------------+
|                                       | Read|Write | Read|Write         | Read|Write     |
|                                       |           |                    |                |
| 1|0xc0100|0x560001|0000-0000-0000-0000 | 2311|1199 | 302940160|39305216 | 1936|2300      |
| 1|0xc0100|0x560020|0000-0000-0000-0000 | 2409|3661 | 78946304|119980032 | 219|5238       |
+---------------------------------------+-----------+--------------------+----------------+
```

SAN Analytics

showanalytics - Overlay CLI - Example – --errors and --errorsonly

8.4(1)

```
MDS9710-1# showanalytics --interface fc9/16 --initiator-itl --errors

 Interface fc9/16
+---------------------------------------+--------------------+-----------------+
| VSAN|I|T|L                            | Total SCSI Failures | Total FC Aborts |
+---------------------------------------+--------------------+-----------------+
|                                       | Read|Write          | Read|Write      |
|                                       |                    |                 |
| 1|0xc0100|0x560001|0000-0000-0000-0000 | 0|0               | 0|0             |
| 1|0xc0100|0x560020|0000-0000-0000-0000 | 0|0               | 0|2             |
+---------------------------------------+--------------------+-----------------+
```

8.4(1)

```
MDS9710-1# showanalytics --interface fc9/16 --initiator-itl --errorsonly

 Interface fc9/16
+---------------------------------------+--------------------+-----------------+
| VSAN|I|T|L                            | Total SCSI Failures | Total FC Aborts |
+---------------------------------------+--------------------+-----------------+
|                                       | Read|Write          | Read|Write      |
|                                       |                    |                 |
| 1|0xc0100|0x560020|0000-0000-0000-0000 | 0|0               | 0|2             |
+---------------------------------------+--------------------+-----------------+
```

SAN Analytics

showanalytics - Overlay CLI - Example - minmax


```
MDS9710-1# showanalytics --target-itl --minmax
2019-04-15 15:57:29.910279

 Interface fc9/32
+---------------------------------------+----------------------+--------------------------+----------------------+----------------------+
|        VSAN|Initiator|Target|LUN        |       Peak IOPS*     |     Peak Throughput*     |      Read ECT*       |      Write ECT*      |
+---------------------------------------+----------------------+--------------------------+----------------------+----------------------+
|                                       |    Read   |  Write   |     Read   |    Write    |    Min    |    Max     |    Min    |    Max     |
|                                       |           |          |            |             |           |            |           |            |
| 1|0x560000|0x0c0280|0000-0000-0000-0000 |     0     |   5045   |      0     | 157.7 MB/s  |    0      |    0       |  170.0 us|  250.5 ms |
| 1|0x560120|0x0c0280|0000-0000-0000-0000 |     0     |   4709   |      0     | 147.2 MB/s  |    0      |    0       |  170.0 us|  250.3 ms |
+---------------------------------------+----------------------+--------------------------+----------------------+----------------------+
*These values are calculated since the metrics were last cleared.
```

SAN Analytics

showanalytics - Overlay CLI – python source

“source copy-sys” command copies all scripts to bootflash:scripts

```
MDS9710-1# show file bootflash:/scripts/analytics.py
#!/usr/bin/env python

#########################################################################
# Copyright (c) 2017 by Cisco Systems, Inc.  #
#########################################################################

import sys
import argparse
import json
from prettytable import *

import cli
...
```

SAN Analytics

Monitoring ITLs and NPU Load

8.4(1)

- Each module/switch has a maximum number of ITLs supported
- If ITLs are exceeded it can lead to switch instability.
- The following error level messages are seen:
  - FTMGR_MOD_EXCESS_ITLS: Total monitored ITL count in module <num> exceeds module limit (<x> active ITLs). Analytics data may be incomplete.
  - FTMGR_SYS_EXCESS_ITLS: Total monitored ITL count in the system exceeds system limit (<x> active ITLs). Analytics data may be incomplete.
- Reduce the number of ports being monitored
- Check Configuration Limits Guide

9700

9700
9132T
9148T
9396T

SAN Analytics

Monitoring ITLs and NPU Load

8.4(1)

- Network Processing Unit(NPU) is also limited by IOPS
- If NPU capacity exceeded metrics will be incomplete/missing
- The following warning level messages are seen:
  - FTMGR_MOD_HIGH_NPU_LOAD: Module <num> is experiencing high NPU load.
  - FTMGR_SYS_HIGH_NPU_LOAD: Switch is experiencing high NPU load.
- Implement port-sampling
- Keep NPU loads below 90%

9700

9700
9132T
9148T
9396T

SAN Analytics

Monitoring ITLs and NPU Load


```
MDS9718# show analytics system-load
 n/a - not applicable
---------------------------------- Analytics System Load Info ----------------------------------
| Module | NPU Load (in %)  | ITLs    ITNs    Both    |        Hosts          |       Targets
|        | SCSI NVMe Total  | SCSI    NVMe    Total   | SCSI   NVMe   Total   | SCSI   NVMe   Total |
-------------------------------------------------------------------------------------------------
|   1    | 35   0    35     | 566     0       566     | 0      0      0       | 62     0      62    |
|   4    | 98   0    98     | 20770   0       20770   | 0      0      0       | 348    0      348   |
|   5    | 33   0    33     | 1756    0       1756    | 0      0      0       | 190    0      190   |
|   8    | 72   0    72     | 1119    0       1119    | 0      0      0       | 99     0      99    |
|   12   | 0    0    0      | 0       360     360     | 0      0      0       | 0      50     50    |
|   13   | 71   0    71     | 1032    0       1032    | 0      0      0       | 100    0      100   |
|   18   | 1    0    1      | 20036   0       20036   | 0      0      0       | 100    0      100   |
| Total  | n/a  n/a  n/a    | 45279   360     45639   | 0      0      0       | 899    50     949   |
-------------------------------------------------------------------------------------------------
```

SAN Analytics

Clearing metrics

- Metrics can be cleared to reset totals, min, max, etc.
- Clear all statistics in view fc-scsi.scsi_initiator
- clear analytics 'select all from fc-scsi.scsi_initiator’
- Clear all statistics in view fc-scsi.scsi_initiator only for port fc9/16
- clear analytics query 'select all from fc-scsi.scsi_initiator where port=fc9/16'

SAN Analytics
Documentation

- Cisco MDS 9000 Series NX-OS SAN Telemetry Streaming Configuration Guide

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/8_x/config/san_analytics/cisco-mds9000-san-analytics-telemetry-streaming-config-guide-8x/configuring-san-telemetry-streaming.html?dtid=osscdc000283

Summary

- New features are being introduced frequently
- Understand and implement the many features provided
- MDS release notes have information on new features

Additional Relevant Sessions

Storage Area Networking

- **Cisco Live Barcelona 2020**
  - PSODCN-2120 – Leverage High Fidelity Telemetry and Insights to Modernize SAN Infrastructure
    - Wednesday, January 29 | 12:30 PM - 01:00 PM
  - BRKDCN-3282 – NVMe over Fabrics (NVMe-oF) End-to-End Configuration with RoCEv2 and Fibre Channel
    - Friday, January 31 | 11:30 AM - 01:30 PM
- **Cisco Live San Diego 2019**
  - BRKDCN-2271 – DCNM San Insights – Next Generation Network Visability
  - BRKDCN-2291 – Cisco MDS and Nexus 9000: Continued SAN convergence solutions
  - BRKDCN-1008 – Managing Data Center Networks using Cisco Data Center Network Manager (DCNM)
  - BRKDCN-2494 – NVMe and NVMe over Fabrics deep dive
  - BRKDCN-2729 – The Networking Implications of NVMe over Fabrics (NVMe-oF)

Complete your online session survey

- Please complete your session survey after each session. Your feedback is very important.
- Complete a minimum of 4 session surveys and the Overall Conference survey (starting on Thursday) to receive your Cisco Live t-shirt.
- All surveys can be taken in the Cisco Events Mobile App or by logging in to the Content Catalog on ciscolive.com/emea.

Cisco Live sessions will be available for viewing on demand after the event at ciscolive.com.

cisco Live!

Continue your education

Demos in the Cisco campus

Walk-in labs

Meet the engineer 1:1 meetings

Related sessions

Thank you

cisco Live!

You make **possible**