
cisco Live!
ALL IN

The bridge to possible

Detecting, Alerting, Identifying and Preventing SAN Congestion

Congestion Best Practices

Paresh Gupta
Technical Leader, Technical Marketing Engineering, Cisco
@reach2paresh

BRKDCN-3241

Cisco Webex App

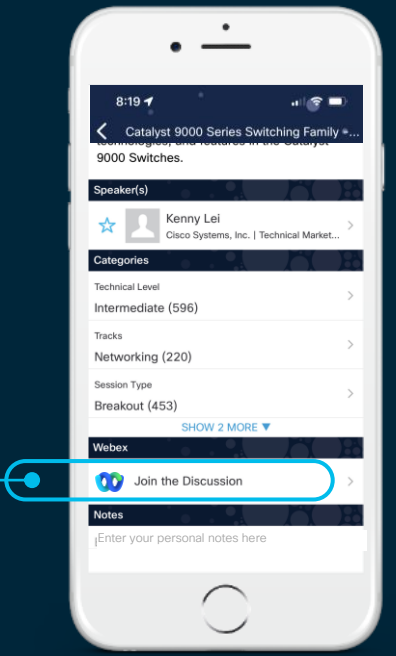
Questions?

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

How

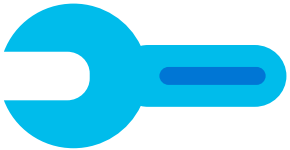
1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until June 17, 2022.



https://ciscolive.ciscoevents.com/ciscolivebot/#BRKDCN-3241

SAN Congestion & Slow Drain – Overview

Detection	Troubleshooting	Prevention
in 2.5µs	Detailed tools	in 1 ms
		
TxWait, slowport-monitor, Automatic alerting using port-monitor	NDFC Slow drain analysis * Long term trending and seasonality of congestion events	DIRL* FPIN Congestion Isolation No-credit-drop timeout

*Slow-drain Analysis is renamed to Congestion Analysis in NDFC 12.1.1e

* Cisco Innovation

SAN Congestion Management – Recommendations

Reactive

- NDFC/DCNM Slow-drain Analysis* – Maintains historic trends and seasonality of congestion metrics from switchports. Must be enabled manually.
- MDS collects TxWait, Datarate and other metrics by default
- Important for troubleshooting

Proactive

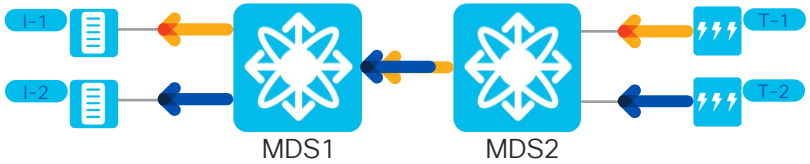
- MDS port-monitor (PMon) – Monitors congestion events at a low granularity. Generates alerts and takes automated actions when thresholds exceeds.
- A proper PMon policy must be configured manually.
- Important for proactive alerting and congestion prevention using DIRL.

Predictive

- SAN Analytics and SAN Insights – Collects I/O performance metrics to get visibility into application traffic patterns.
- Important for predicting the underlying root cause of congestion and predicting the ports where congestion is more likely than other ports in a fabric.
- Refer to BRKDCN-3645

*Slow-drain Analysis is renamed to Congestion Analysis in NDFC 12.1.1e

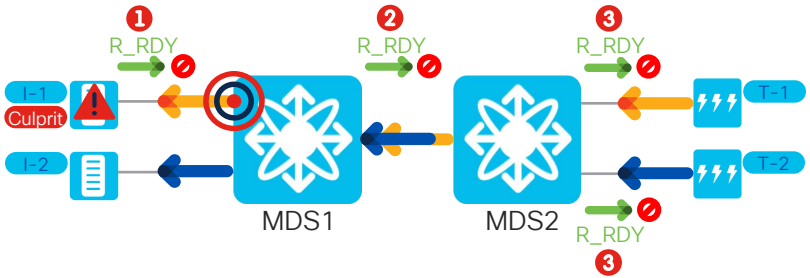
Common Causes of SAN Congestion



Common Causes of SAN Congestion

Slow Drain

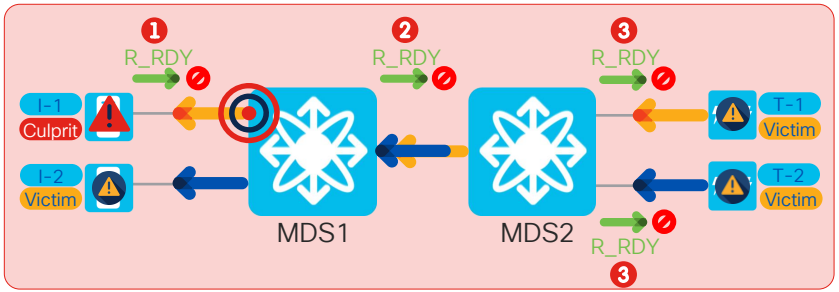
Tx B2B credit starvation



Common Causes of SAN Congestion

Slow Drain

Tx B2B credit starvation

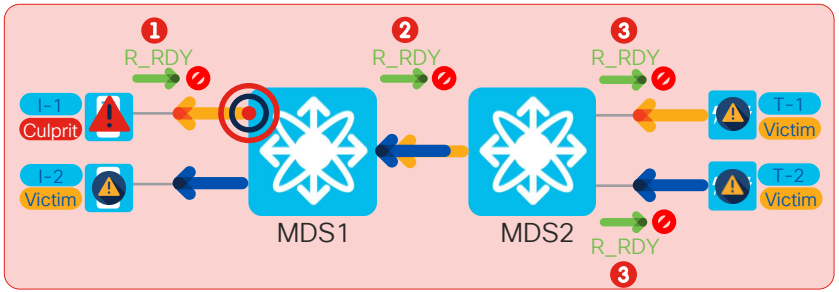


Frames are not dropped in FC fabric. Rather, they consume switch buffers causing a fabric-wide congestion

Common Causes of SAN Congestion

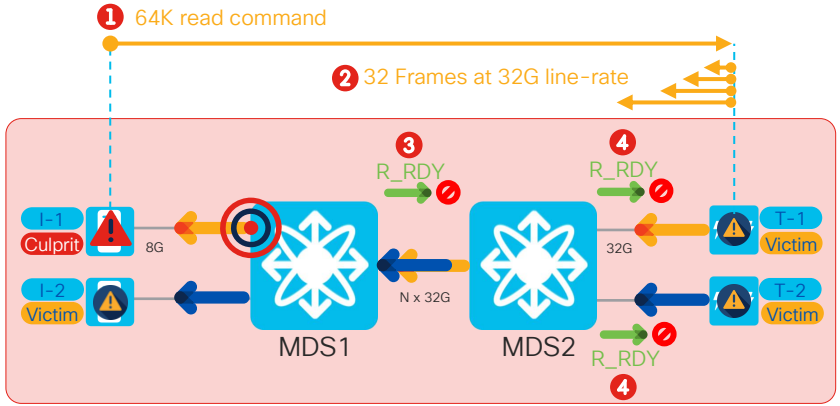
Slow Drain

Tx B2B credit starvation



Over-utilization

Receive data rate on ISL port is faster than the host port speed



Frames are not dropped in FC fabric. Rather, they consume switch buffers causing a fabric-wide congestion

Common Causes of SAN Congestion

Slow Drain

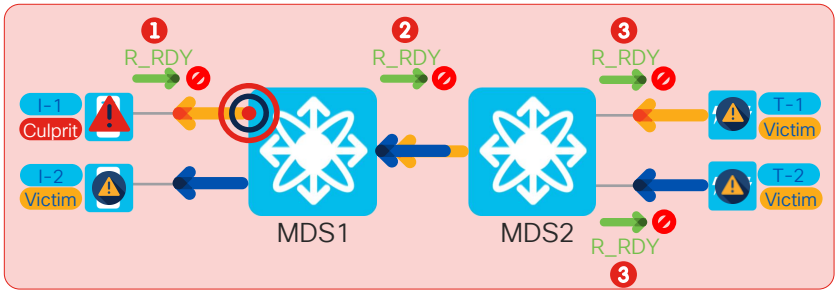
Tx B2B credit starvation

I-1 is busy

I-1 slows down its ingress traffic rate by slowing down sending of R_RDY to MDS1

Tx Utilization % (tx-datarate)

Tx Congestion % (txwait)



Over-utilization

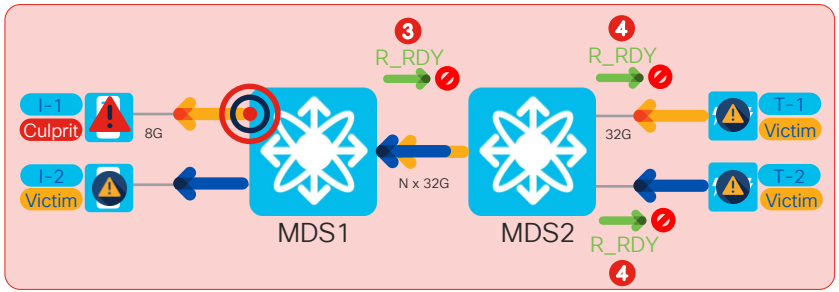
Receive data rate on ISL port is faster than the host port speed

I-1 is receiving at full capacity

MDS1 slows down traffic rate to I-1 by slowing down sending of R_RDY to MDS2

Tx Utilization % (tx-datarate)

Tx Congestion % (txwait)



Frames are not dropped in FC fabric. Rather, they consume switch buffers causing a fabric-wide congestion

Understanding TxWait

- TxWait is an ASIC counter that increments by 1 when a port is unable to transmit a queued frame for 2.5 microseconds due to Tx B2B credit unavailability

```
mds9710-1# show interface fc1/1 counters | inclue ignore-case wait
 26009409536 2.5us TxWait due to lack of transmit credits
```

- Convert TxWait to seconds by (TxWait * 2.5) / 1000000
- In the above output, **26009409536** * 2.5/1000000 = 65,023 seconds
  - MDS was not able to transmit for 65,023 seconds since the counter was last cleared
- MDS enriches the raw TxWait counter:
  - For storing on switch OBFL (On-board Failure Logging (Buffer)) for troubleshooting
  - For automated alerting and actions by port-monitor (PMon)
  - Export via SNMP or NX-API to remote systems like NDFC/DCNM slow drain analysis

TxWait history on MDS

- TxWait delta value is logged periodically(20 seconds) into OBFL, if delta value >=100ms.
- Displays TxWait time in 2.5µs ticks as well as in seconds.
- Congestion value is displayed in percentage over period of 20 seconds.
- Timestamp of event occurrence also recorded.

```
MDS9706-C# sh logging onboard txwait
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  |            |                    |
-------------------------------------------------------------------------------------
|   fc1/15  | None         |    86510    |    0     |       1%   | Thu Feb 10 15:11:42 2022 |
|   fc1/15  | None         |    46459    |    0     |       0%   | Thu Feb 10 15:11:22 2022 |
|   fc1/15  | None         |  1129160    |    2     |      14%   | Sat Oct 16 00:09:52 2021 |
|   fc1/15  | None         |   658894    |    1     |       8%   | Tue Oct 12 02:18:50 2021 |
```

Tx-datarate – Port Traffic

- Pmon on MDS measures datarate in percent utilization. For example,
  - Tx-datarate: tx utilization > 80% (*) continuously for 10 seconds (*)
  - Tx-datarate-burst: 5 (*) times in 10 seconds (*) tx utilization > 90% (*) continuously for 1 second

Fibre Channel speeds	Baud rate (GBd)	Bit Rate (Gbps)	Data rate (MB/s)
1GFC	1.0625	1.0625	100
2GFC	2.125	2.125	200
4GFC	4.25	4.25	400
8GFC	8.5	8.5	800
16GFC	14.025	14.025	1600
32GFC	28.025	28.025	3200
64GFC	28.900	57.8	6400

- Important: Max datarate of FC interfaces is lower than the used notation
  - What's the % utilization of 25Gbps a on 32GFC port? (Wrong: 25/32 = 78%. Correct: 25/28 = 89%)
  - Use correct max bit-rate when DIYing link-utilization calculation or using 3rd party monitoring apps

For all practical purposes, due to longer polling intervals in production environments, treat any occurrence of high utilization the same as over-utilization, which may cause congestion

Tx-datarate history on MDS

- High-utilization events are stored on the switch

```
MDS9706-C# show logging onboard datarate
 ------------------------------------------------------------------------------------
| Interface |  Speed  |     Alarm-types                |    Rate   |         Timestamp        |
 ------------------------------------------------------------------------------------
|  fc1/13   |    4G   |  TX_DATARATE_BURST_FALLING     |    0@0%   | Fri Apr 29 16:41:06 2022 |
|  fc1/13   |    4G   |  TX_DATARATE_FALLING           |    63%    | Fri Apr 29 16:40:56 2022 |
|  fc1/13   |    4G   |  TX_DATARATE_RISING            |    98%    | Fri Apr 29 16:34:03 2022 |
|  fc1/13   |    4G   |  TX_DATARATE_BURST_RISING      |    6@98%  | Fri Apr 29 16:34:00 2022 |
|  fc1/13   |    4G   |  TX_DATARATE_BURST_FALLING     |    0@0%   | Fri Apr 29 16:33:04 2022 |
|  fc1/13   |    4G   |  TX_DATARATE_FALLING           |    54%    | Fri Apr 29 16:32:53 2022 |
|  fc1/13   |    4G   |  TX_DATARATE_RISING            |    98%    | Fri Apr 29 16:25:41 2022 |
```

TX_DATARATE_RISING it started at 10 seconds prior to when it was recorded 16:25:31and ended 10 seconds prior to when the TX_DATARATE_FALLING was recorded 16:32:43. There was high utilization for 7 min 12 seconds.

Automated Alerting and Congestion Prevention

Port-monitor (PMon) on Cisco MDS

PMon monitors each switchport at a low granularity (as low as 1 second). When threshold exceed, PMon automatically takes actions like generating alerts, shutting down (errdisable) ports, flapping the port, isolating the port, or DIRL.

How to configure PMon?

1. Start by enabling PMon for sending alerts
2. Refine the thresholds over weeks/months. Solve the real culprits. Avoid too many alerts.
3. Finally, enable actions, such as congestion prevention using DIRL
4. Go to step 2

Sample PMon policies: https://www.cisco.com/c/en/us/support/docs/storage-networking/mds-9000-nx-os-software-release-62/200102-Sample-MDS-port-monitor-policy-for-alert.html

PMon Policy on MDS

```
#
port-monitor name fabricmon_edge_policy
  logical-type edge
 counter txwait poll-interval 1 delta rising-threshold 30 event 4 falling-threshold 10 event 4 alerts syslog rmon portguard DIRL
 counter tx-datarate poll-interval 10 delta rising-threshold 80 event 4 falling-threshold 70 event 4 alerts syslog rmon obfl portguard DIRL
 counter tx-datarate-burst poll-interval 10 delta rising-threshold 5 event 4 falling-threshold 1 event 4 alerts syslog rmon obfl datarate 90


# Show port-monitor
Policy Name  : fabricmon_edge_policy
Admin status : Not Active
Oper status  : Not Active
Port type    : All Edge Ports
```

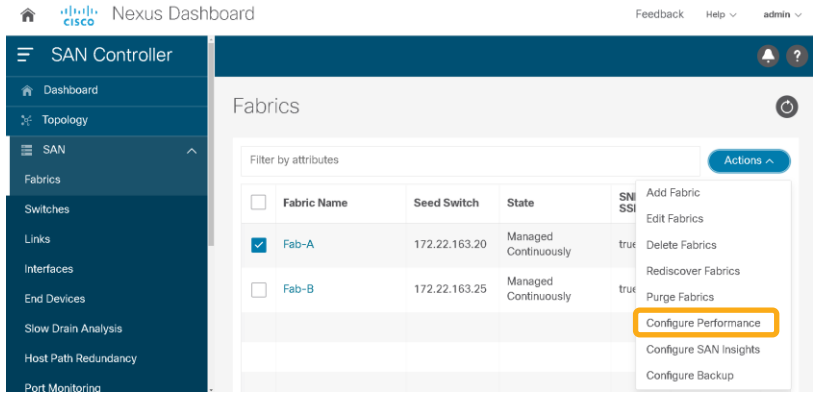
Counter	Threshold Type	Interval (Secs)	Warning		Thresholds		Rising/Falling actions			Congestion-signal	
			Threshold	Alerts	Rising	Falling	Event	Alerts	PortGuard	Warning	Alarm
Link Loss	Delta	30	none	n/a	5	1	4	syslog,rmon	FPIN	n/a	n/a
Sync Loss	Delta	30	none	n/a	5	1	4	syslog,rmon	FPIN	n/a	n/a
Signal Loss	Delta	30	none	n/a	5	1	4	syslog,rmon	FPIN	n/a	n/a
Invalid Words	Delta	30	none	n/a	1	0	4	syslog,rmon	FPIN	n/a	n/a
Invalid CRC's	Delta	30	none	n/a	5	1	4	syslog,rmon	FPIN	n/a	n/a
State Change	Delta	60	none	n/a	5	0	4	syslog,rmon	none	n/a	n/a
TX Discards	Delta	60	none	n/a	200	10	4	syslog,rmon	none	n/a	n/a
LR RX	Delta	60	none	n/a	5	1	4	syslog,rmon	none	n/a	n/a
LR TX	Delta	60	none	n/a	5	1	4	syslog,rmon	none	n/a	n/a
Timeout Discards	Delta	60	none	n/a	200	10	4	syslog,rmon	none	n/a	n/a
Credit Loss Reco	Delta	1	none	n/a	1	0	4	syslog,rmon	none	n/a	n/a
TX Credit Not Available	Delta	1	none	n/a	10%	0%	4	syslog,rmon	none	n/a	n/a
RX Datarate	Delta	10	none	n/a	80%	70%	4	syslog,rmon,obfl	none	n/a	n/a
TX Datarate	Delta	10	none	n/a	80%	70%	4	syslog,rmon,obfl	DIRL	n/a	n/a
TX-Slowport-Oper-Delay	Absolute	1	none	n/a	50ms	0ms	4	syslog,rmon	none	n/a	n/a
TXWait	Delta	1	none	n/a	30%	10%	4	syslog,rmon	DIRL	40%	60%
RX Datarate Burst	Delta	10	none	n/a	5@90%	1@90%	4	syslog,rmon,obfl	none	n/a	n/a
Input Errors	Delta	60	none	n/a	5	1	4	syslog,rmon	none	n/a	n/a

```
 On falling threshold portguard actions FPIN, DIRL, Cong-Isolate-Recover will initiate auto recovery of ports.
```

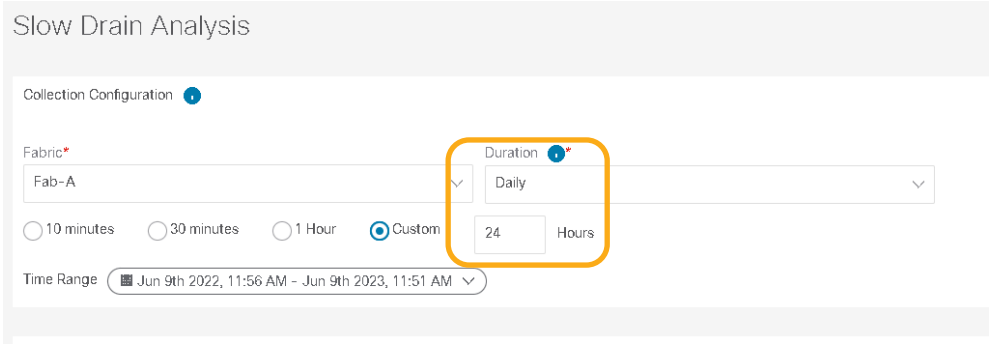
NDFC Congestion/Slow-drain Analysis

Best Practice – Run in always-on mode.

- Slow-drain analysis is not enabled by default



- After adding a new fabric:
  - Enable performance monitoring
  - Schedule to run slow drain analysis daily for 24 hours.



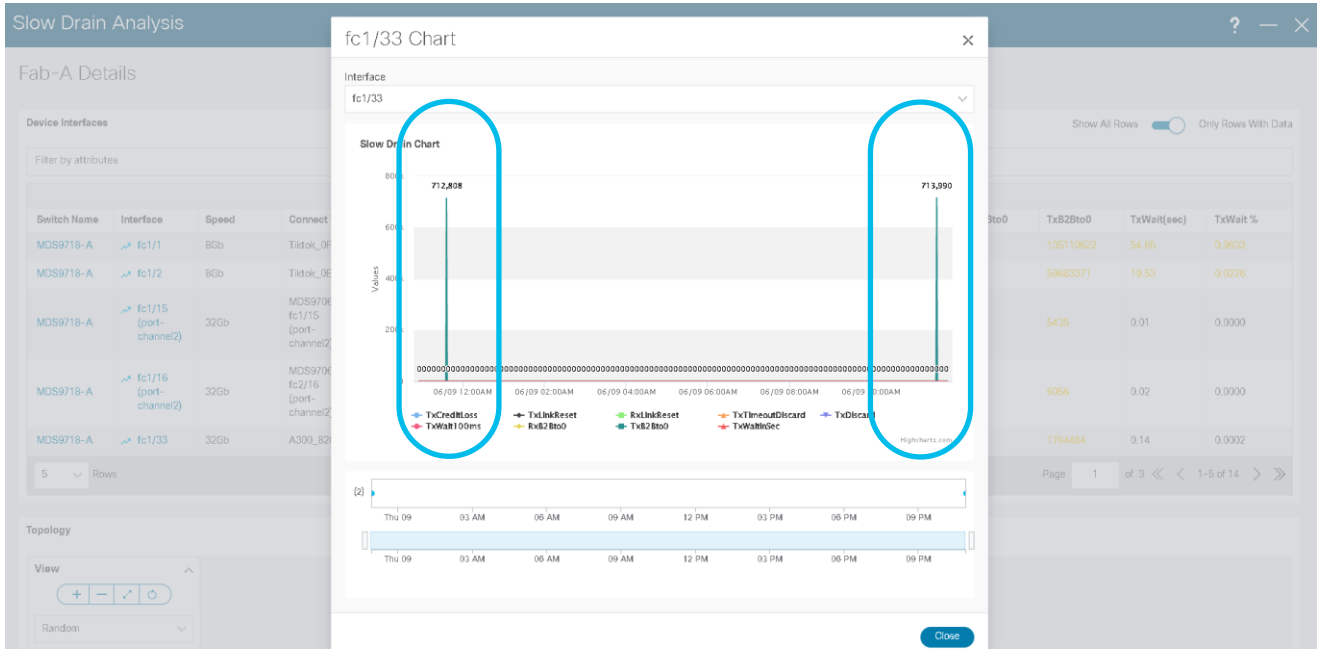
*Slow-drain Analysis is renamed to Congestion Analysis in NDFC 12.1.1e

- DCNM/NDFC slow-drain analysis has minimal/negligible effect on the switches

NDFC Congestion/Slow-drain Analysis

Always-on, historical view with trending and seasonality

- fc1/33 is congested in Tx direction
- TxWait increases but not all the time. Only two spikes in last 12 hours.
- Next Steps –
  - Correlate with host and app. Does it correlate with a cron job on the host?
  - Look at SAN Insights metrics to find the root cause.



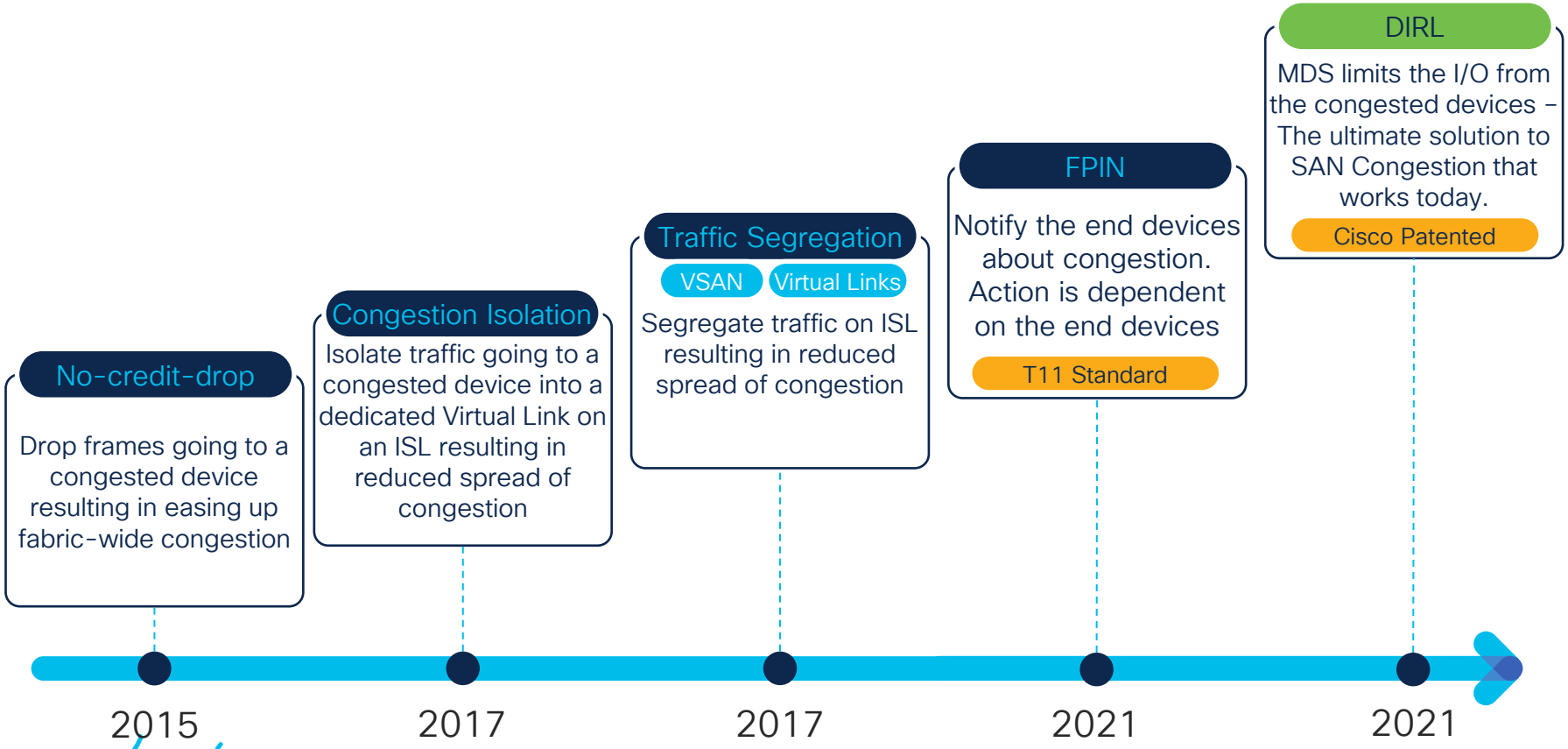
SAN Congestion

including Slow Drain and Over-Utilization

We talked about
Detection and
Alerting

Now, let's talk about
Prevention

SAN Congestion Innovation on Cisco MDS



Not an exhaustive list

Common Causes of SAN Congestion

Slow Drain

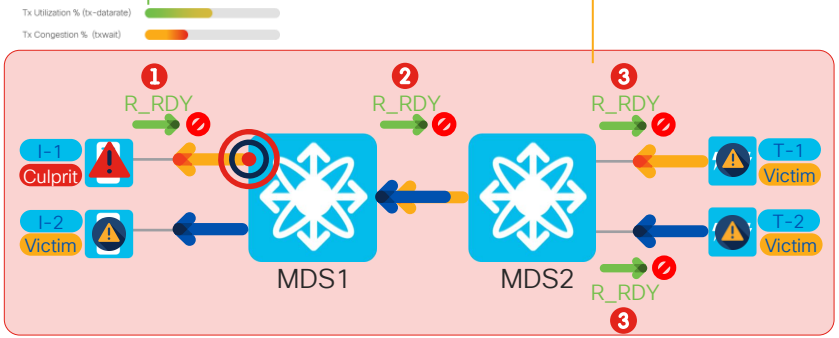
Over-Utilization

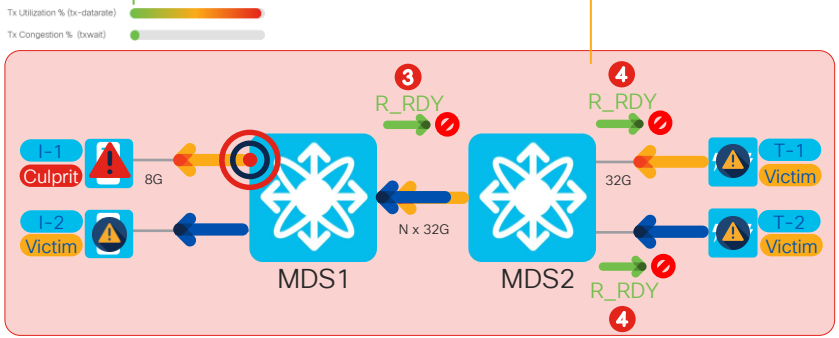
Different detection

On the switch port connected to the culprit device

Similar effects

On fabric and victim devices





Common Causes of SAN Congestion

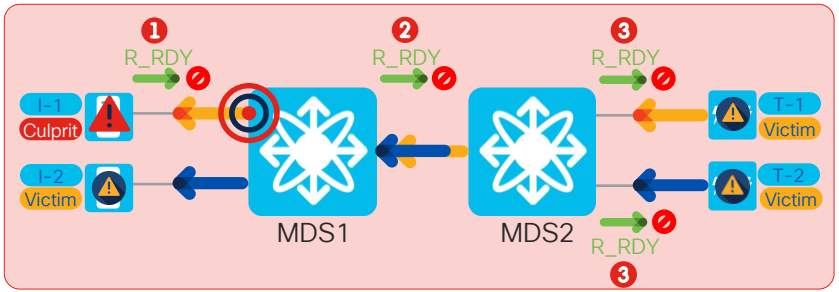
Slow Drain

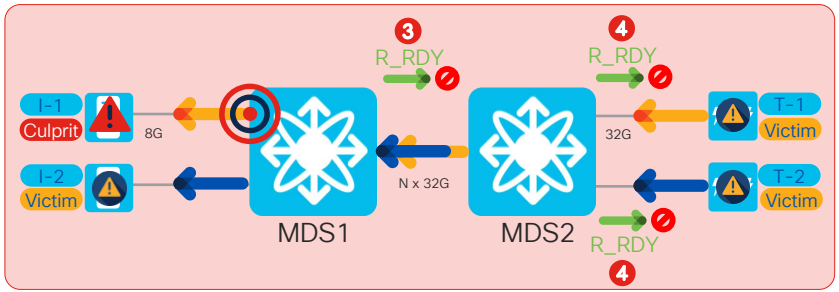
Over-Utilization

Different detection
On the switch port connected to the culprit device

Similar effects
On fabric and victim devices

Same Root Cause
The culprit device is receiving more than it can ingest





Different Detection, Similar Effect, Same Root Cause

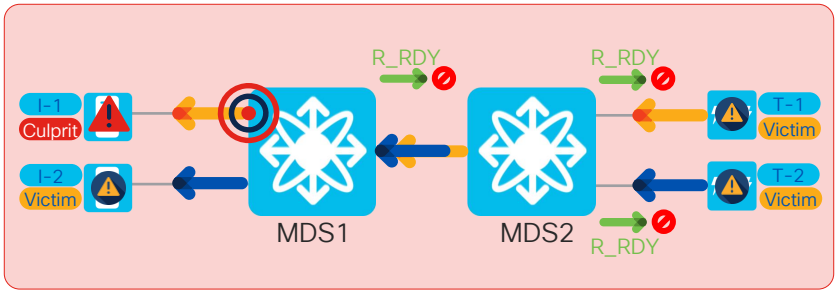
Slow Drain

Over-Utilization

Different detection
On the switch port connected to the culprit device

Similar effects
On fabric and victim devices

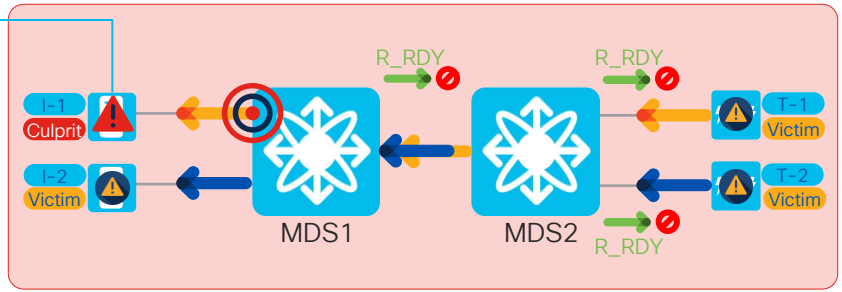
Same Root Cause
The culprit device is receiving more than it can ingest



The Root Cause of SAN Congestion

The Root Cause

I-1 is receiving more than it can ingest



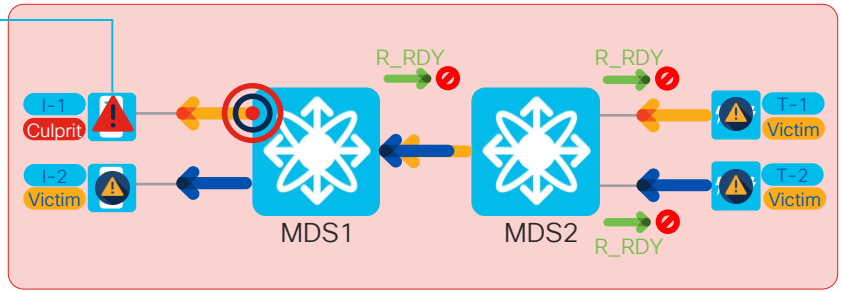
The Root Cause of SAN Congestion

The Root Cause

I-1 is receiving more than it can ingest

Why is I-1 receiving more than it can ingest?

...because I-1 is asking for it.



The Solution

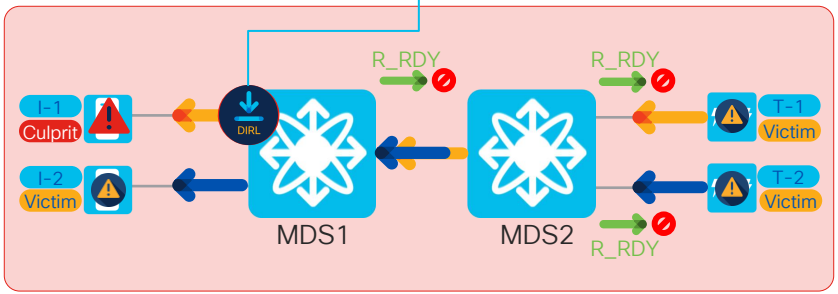
Cisco Dynamic Ingress Rate Limiting

DIRL

I-1 is asking for more than it can ingest

DIRL limits I-1's asking rate to reduce its receiving rate

DIRL dynamically changes I-1's asking rate to adapt to its traffic profile



The Solution

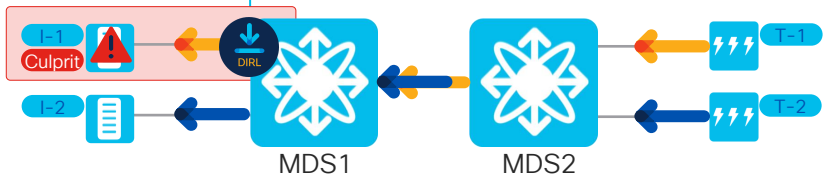
Cisco Dynamic Ingress Rate Limiting

DIRL

I-1 is asking for more than it can ingest

DIRL limits I-1's asking rate to reduce its receiving rate

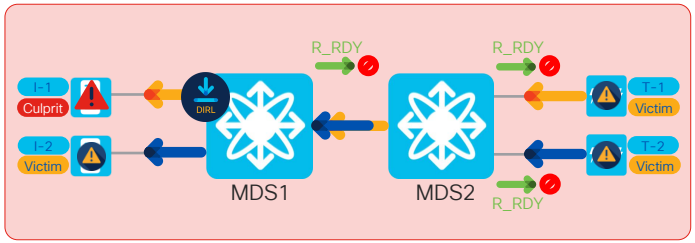
DIRL dynamically changes I-1's asking rate to adapt to its traffic profile



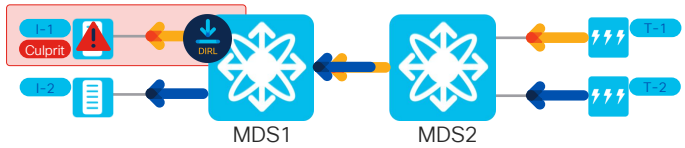
DIRL prevents SAN Congestion due to slow-drain and over-utilization.

Cisco Dynamic Ingress Rate Limiting

Without Cisco DIRL



With Cisco DIRL



Cisco Dynamic Ingress Rate Limiting

End-device independent
Upgrading of end-devices is not needed

Adaptive
DIRL dynamically adjusts as per the traffic profile of the host

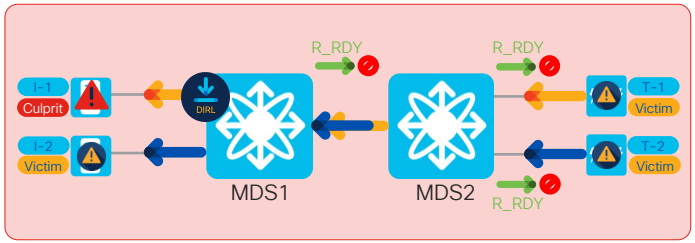
No side effects
Rate limits congested hosts only. Other non-congested hosts and storage ports are not impacted

Easy adoption
DIRL is available on MDS switches after a software-only upgrade.

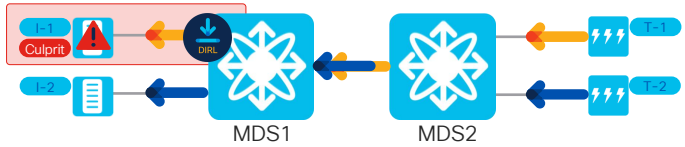
Affordable
No additional license needed

Topology independent
DIRL works in edge-core, edge-core-edge, or collapsed core (single switch fabric) topologies

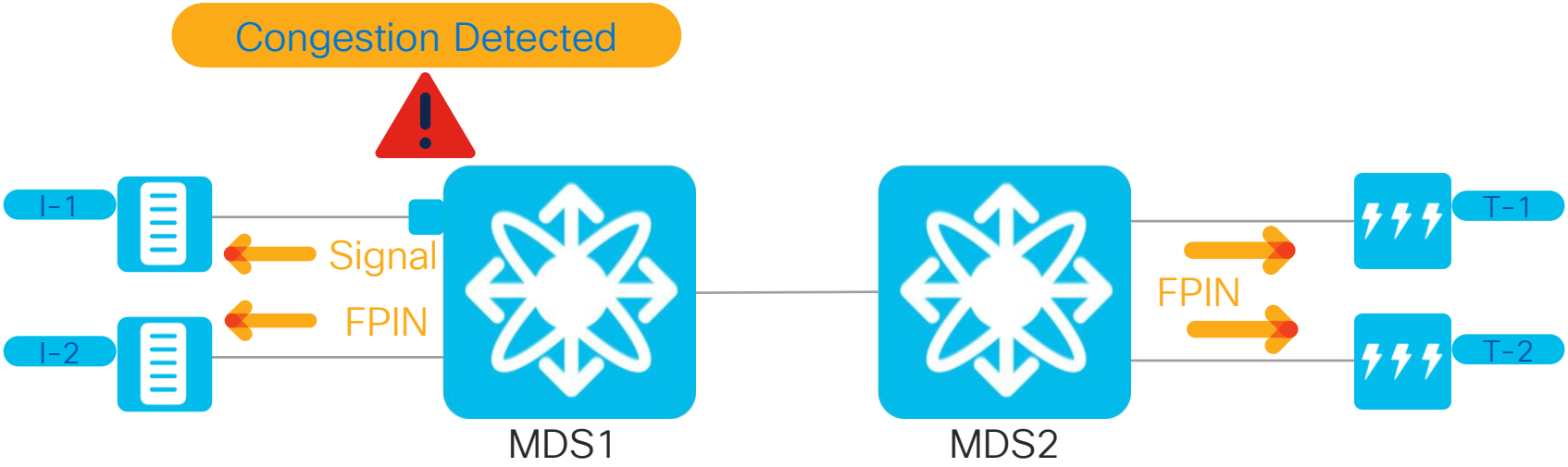
Without Cisco DIRL



With Cisco DIRL



Notifications and Signals in Fibre Channel

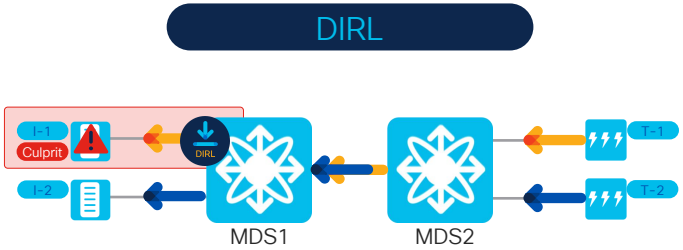


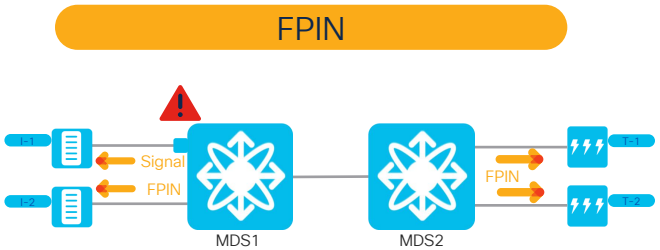
Available on Cisco MDS in NX-OS 8.5(1) onwards

Register Diagnostic Functions (RDF)
Exchange Diagnostic Capabilities (EDC)
Fabric Performance Impact Notifications (FPIN)
Congestion Signals (Primitives)

DIRL vs FPIN

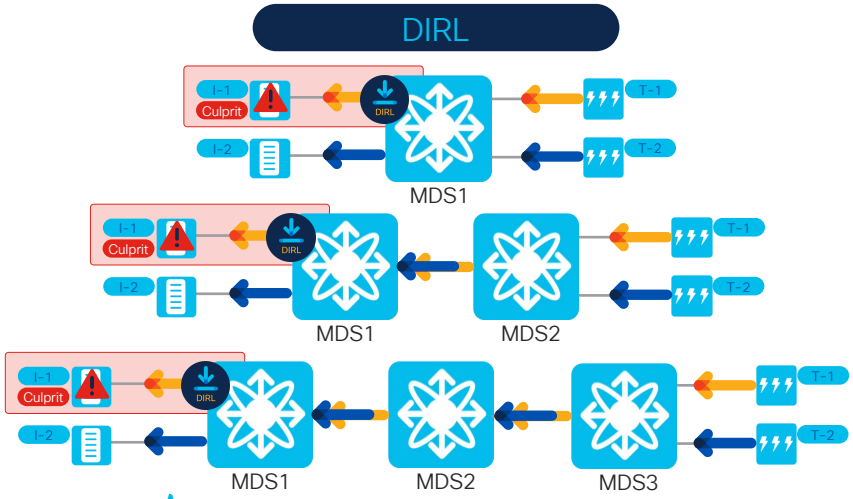
- DIRL helps today. FPIN readiness will take a few years.
  - DIRL is available on existing MDS switch after a software-only upgrade, without any dependency on end devices
  - Although FPIN is supported on MDS switches, action is dependent on the end devices
- DIRL is affordable
  - DIRL and FPIN work on existing MDS switches and don't need an additional license
  - Must upgrade end-devices to benefit from FPIN
- In the future, when you are ready for FPIN, DIRL will continue to be a complementary technology
  - What if a few devices don't react to FPIN and still cause congestion? DIRL within MDS switches will be the protection

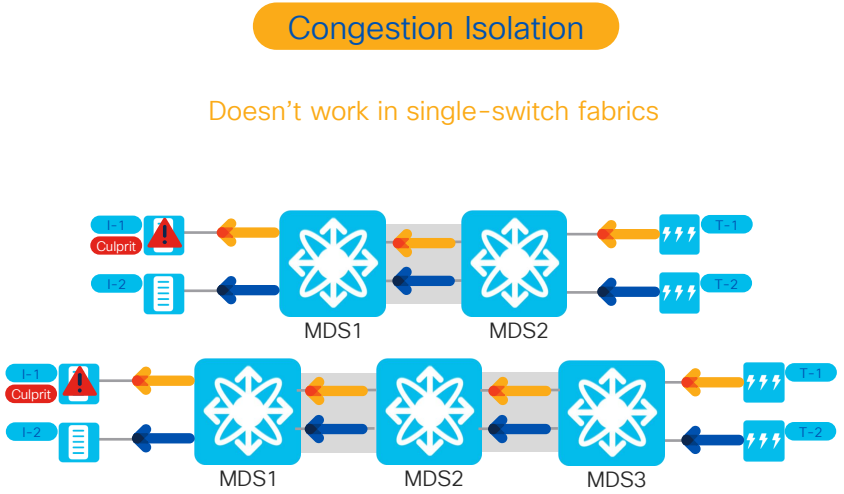




DIRL vs Congestion Isolation

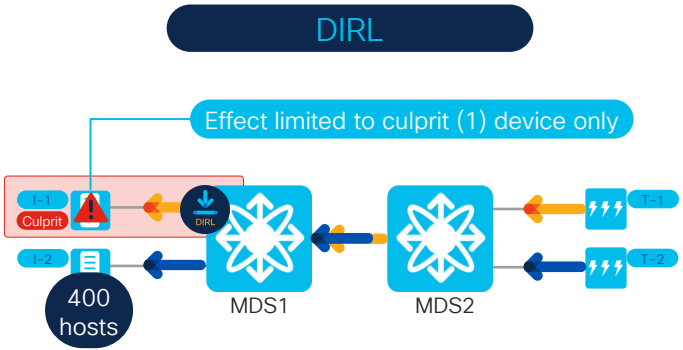
- DIRL is topology independent (shown on this slide)
  - DIRL works in edge-core, edge-core-edge, or collapsed core (single switch fabric) topologies
  - Congestion isolation works by isolating the traffic on ISLs thus it doesn't work in single-switch fabrics.
- DIRL confines congestion to the culprit host only (shown on next slide)
  - DIRL rate-limits the traffic from a congested host without any side effects on storage ports
  - Congestion Isolation confines the effect to the culprit hosts, storage ports that communicate with the culprit host, and all other hosts that communicate with the affected storage ports

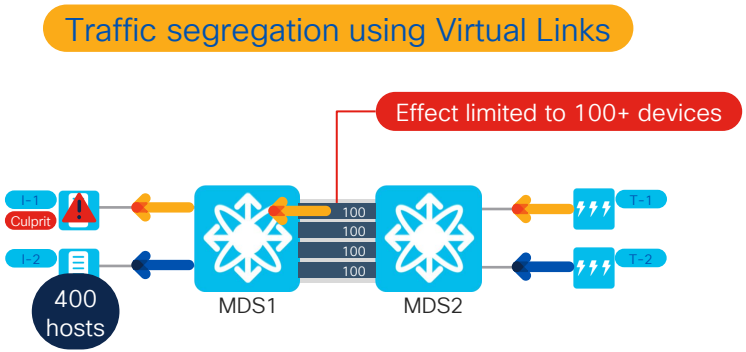




DIRL vs Traffic Segregation using Virtual Links

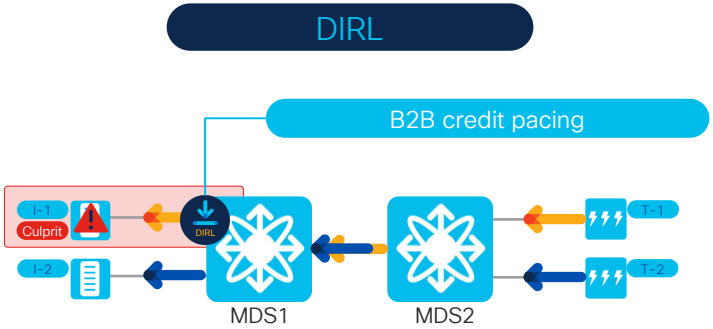
- DIRL is topology independent (shown on previous slide)
  - DIRL works in edge-core, edge-core-edge, or collapsed core (single switch fabric) topologies
  - Virtual links need ISLs, and hence they are of no use in single-switch fabrics
- DIRL confines congestion to the culprit host only (shown on this slide)
  - DIRL rate-limits the traffic from a congested host without any side effects on storage ports
  - Virtual links confines the effect to the culprit hosts, storage ports that communicate with the culprit host, and all other hosts that communicate with the affected storage ports

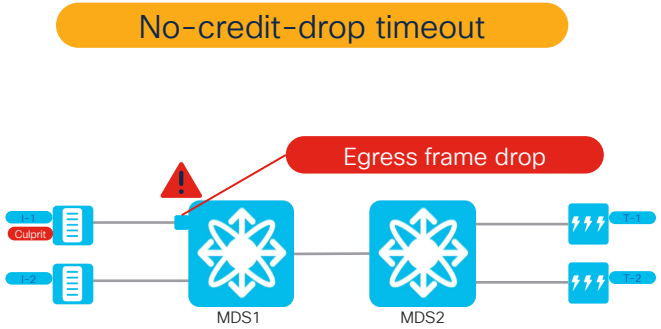




DIRL vs no-credit-drop timeout

- DIRL doesn't drop any frames
  - DIRL limits the traffic from a congested hosts by B2B credit pacing, without dropping any frames
  - No-credit-drop timeout drops frames going to a congested host
- DIRL is more granular
  - DIRL works using TxWait which detects congestion at a granularity of 2.5 microseconds
  - No-credit-drop uses a granularity of 1 ms



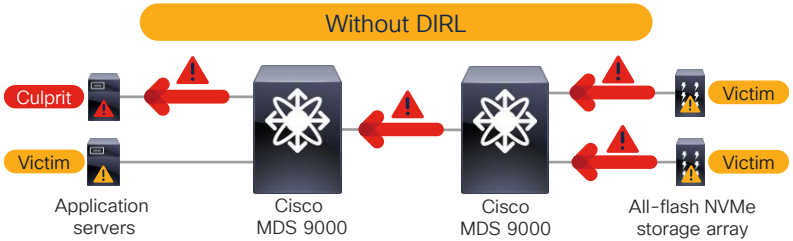


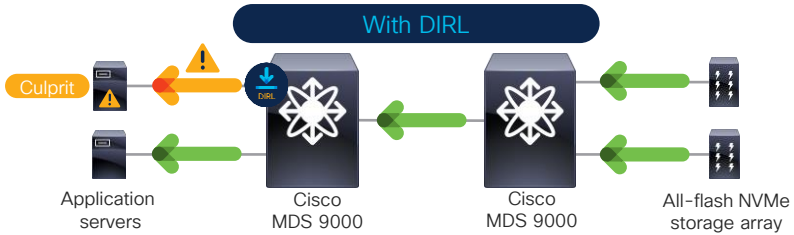
Prevent SAN Congestion and Slow Drain

Limit the culprit devices using

Cisco Dynamic Ingress Rate Limiting

to accelerate the performance of all-flash NVMe storage arrays





End-device independent
Upgrading of end-devices is not needed

Adaptive
DIRL dynamically adjusts as per the traffic profile of the host

No side effects
Rate limits congested hosts only. Other non-congested hosts and storage ports are not impacted

Easy adoption
DIRL is available on MDS switches after a software-only upgrade.

Affordable
No additional license needed

Topology independent
DIRL works in edge-core, edge-core-edge, or collapsed core (single switch fabric) topologies

SAN Congestion Management – Recommendations

Reactive

- Schedule NDFC/DCNM Slow-drain Analysis to run daily for 24 hours.
- Important for troubleshooting

Proactive

- Configure MDS port-monitor (PMon) for automated alerts and actions.
- Important for congestion prevention using DIRL.

Predictive

- Enable SAN Analytics and SAN Insights for getting visibility into application I/O traffic patterns.
- Important for finding the underlying root cause and predicting congestion

*Slow-drain Analysis is renamed to Congestion Analysis in NDFC 12.1.1e

Technical Session Surveys

- Attendees who fill out a minimum of four session surveys and the overall event survey will get Cisco Live branded socks!
- Attendees will also earn 100 points in the Cisco Live Game for every survey completed.
- These points help you get on the leaderboard and increase your chances of winning daily and grand prizes.

Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

cisco Live!

Cisco Learning and Certifications

From technology training and team development to Cisco certifications and learning plans, let us help you empower your business and career. www.cisco.com/go/certs

Pay for Learning with Cisco Learning Credits

(CLCs) are prepaid training vouchers redeemed directly with Cisco.

Learn

Cisco U.
IT learning hub that guides teams and learners toward their goals

Cisco Digital Learning
Subscription-based product, technology, and certification training

Cisco Modeling Labs
Network simulation platform for design, testing, and troubleshooting

Cisco Learning Network
Resource community portal for certifications and learning

Train

Cisco Training Bootcamps
Intensive team & individual automation and technology training programs

Cisco Learning Partner Program
Authorized training partners supporting Cisco technology and career certifications

Cisco Instructor-led and Virtual Instructor-led training
Accelerated curriculum of product, technology, and certification courses

Certify

Cisco Certifications and Specialist Certifications
Award-winning certification program empowers students and IT Professionals to advance their technical careers

Cisco Guided Study Groups
180-day certification prep program with learning and support

Cisco Continuing Education Program
Recertification training options for Cisco certified individuals

Here at the event? Visit us at **The Learning and Certifications lounge at the World of Solutions**

Related sessions

Session ID	Title	Time and Venue	Speaker
BRKDCN-3241	Detecting, Alerting, Identifying and Proactively Preventing SAN Congestion	Thursday, Jun 16, 8:00 AM - 8:45 AM PDT Level 2, Lagoon B	Paresh Gupta
BRKDCN-3645	SAN Insights - Real-time and always-on NVMe visibility at scale	Wednesday, Jun 15, 10:30 AM - 11:15 AM PDT Level 2, Lagoon H	Paresh Gupta
BRKDCN-3812	Dos and Don'ts of Deploying NVMe Over Fabrics	Tuesday, Jun 14, 2:30 PM - 3:15 PM PDT Level 2, Lagoon H	Kamal Bakshi
PSODCN-2355	Real-time NVMe and SCSI visibility using Cisco SAN Analytics	Wednesday, Jun 15, 2:00 PM - 2:30 PM PDT Level 3, South Seas H	Kiran Ranabhor
BRKDCN-2489	IP Fabric for Storage Networks Best Practice and Design	Wednesday, Jun 15, 4:00 PM - 4:45 PM PDT Level 3, South Seas D	Nemanja Kamenica
BRKDCN-1119	Introduction to NDFC: Simplifying Management of Your Data Center	Monday, Jun 13, 9:30 AM - 10:15 AM PDT Level 2, Lagoon G	Parth Patel

Thank you


cisco Live!
ALL IN