
cisco Live!
TURN
IT
UP
#CiscoLive

The bridge to possible

Dynamic Ingress Rate Limiting

A *Real* Solution to SAN Congestion

Edward Mazurek Technical Leader CX

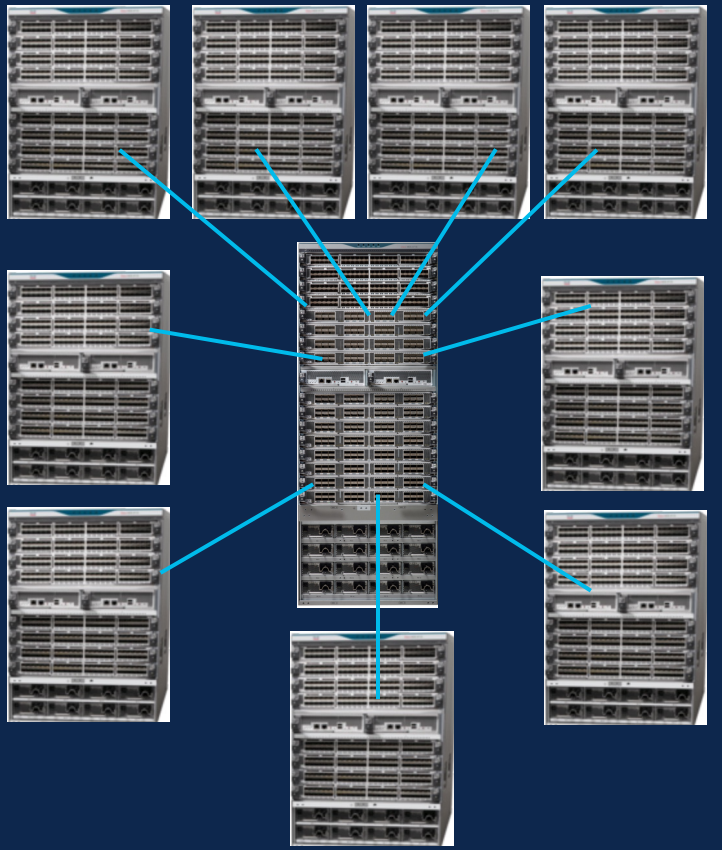
BRKDCN-3002

cisco Live!

Agenda

- Introduction
- Background
  - SAN Congestion
  - Current Mitigation Mechanisms
  - SCSI/NVMe IO Flows
- Function
- Configuration
- Operation
- Conclusion – Benefits

Introduction



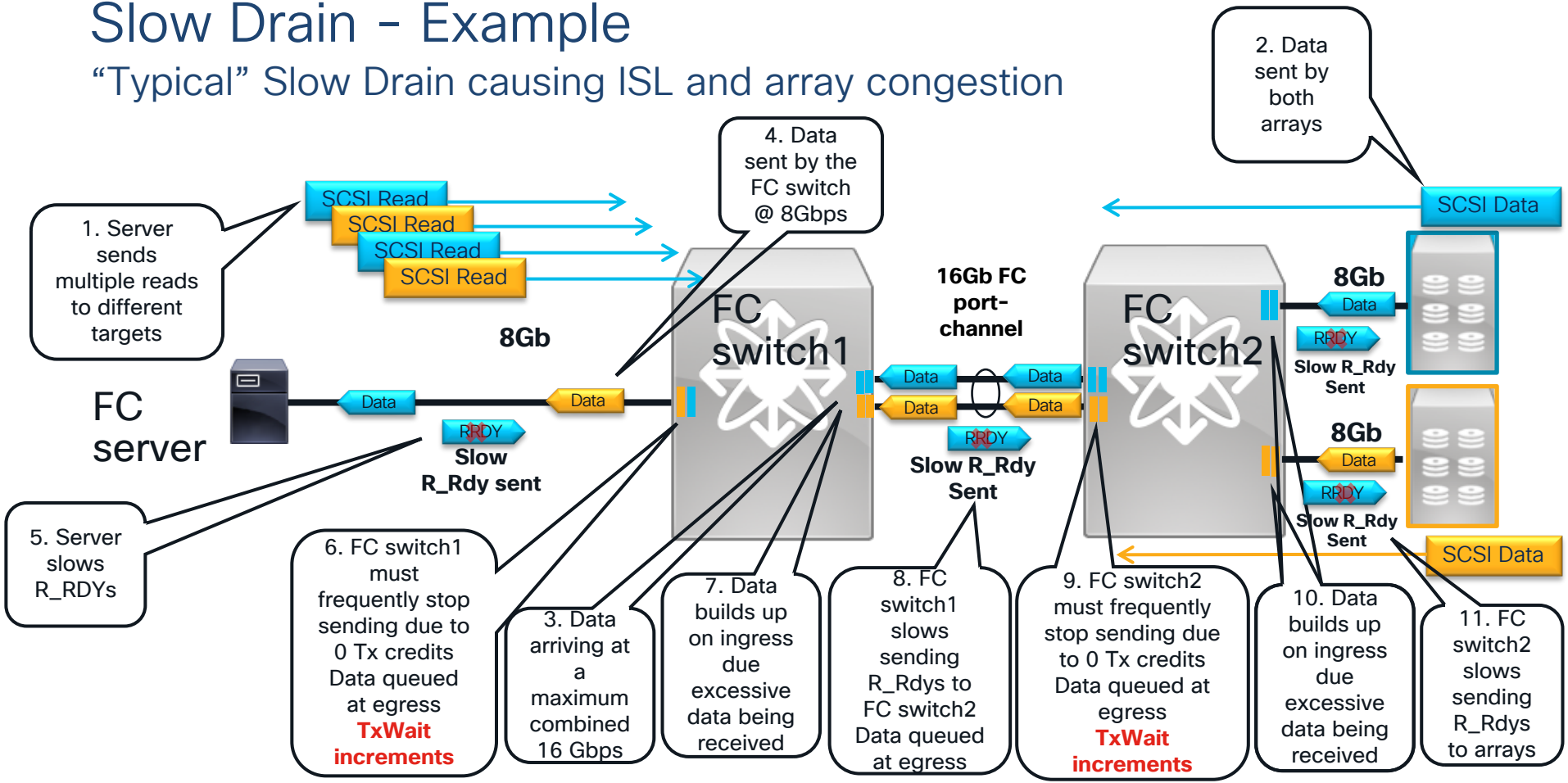
Background – SAN Congestion

Background – SAN Congestion

- “Classic” Slow Drain
  - Still very popular… after all it’s a “classic”!
  - End devices withhold B2B credits/R-RDYs slowing rate of data transmission causing congestion
  - Congestion works its way back across ISLs to data source
  - Detect using TxWait on F ports
  - To resolve investigate end device for internal bottleneck (ex. PCIe bus or HBA)
- Over Utilization
  - End devices requesting more data than can be transmitted to them
  - B2B credits/R-RDYs are not slowed
  - Data transmission is at/near 100%
  - Seeing increasing amounts of this
  - Much more difficult to detect
  - Detected using Port Monitor tx-datarate/tx-datarate-burst counters
  - To resolve increase HBA capacity in end device by increasing HBA speed or quantity

Slow Drain – Example

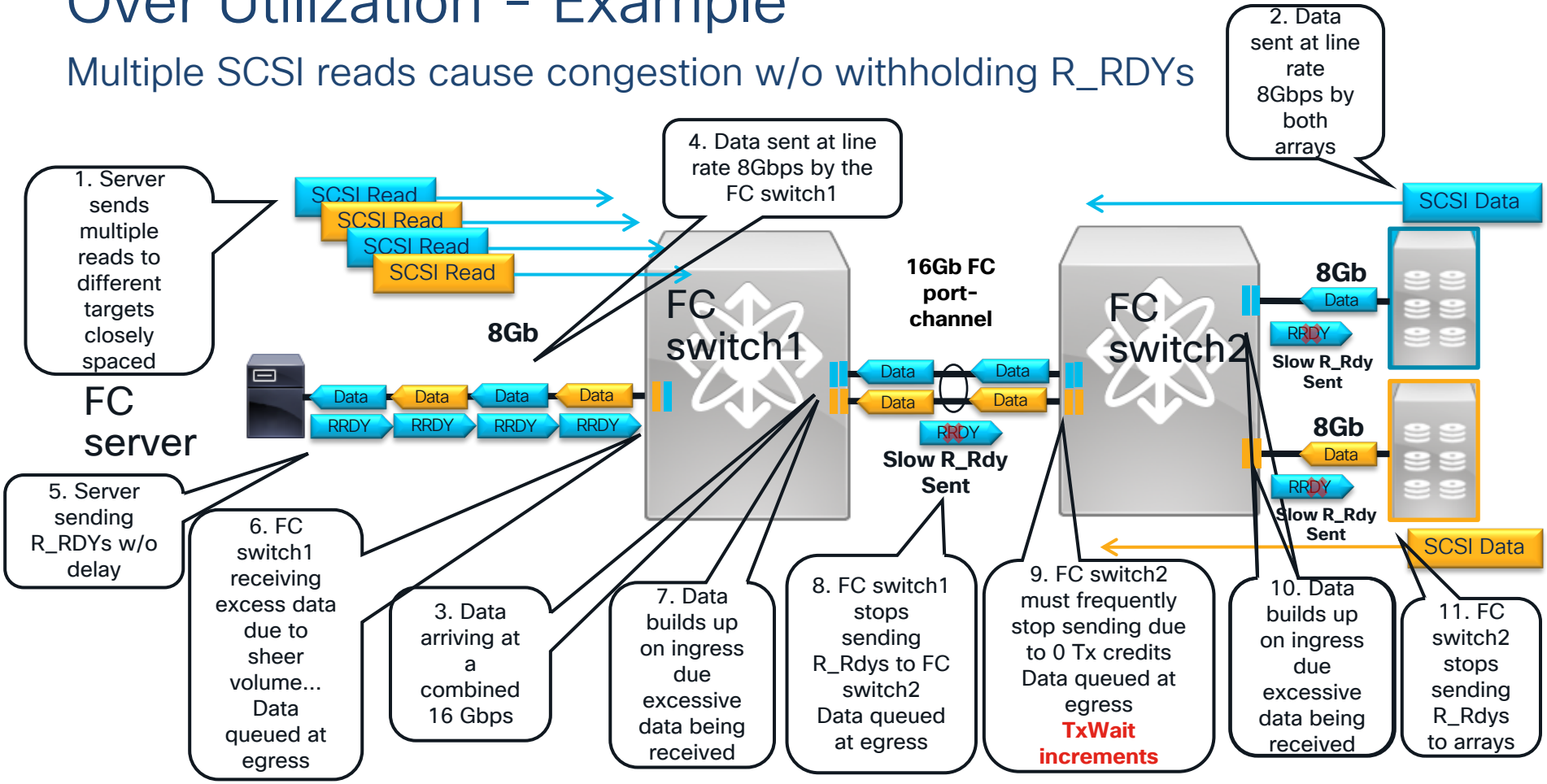
"Typical" Slow Drain causing ISL and array congestion



Both arrays and all devices utilizing ISLs are affected!

Over Utilization – Example

Multiple SCSI reads cause congestion w/o withholding R_RDYs



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

Background – Current Mitigation Mechanisms

Current Mitigation Mechanisms

- No-credit-drop – Drop all frames destined to a port that is at 0 Tx credits for xxx ms
  - Effective at mitigating the ‘Slow Drain’ case at lower values
  - Can be very disruptive to end device
  - Only applicable to the ‘slow drain’ case
- Port Monitor portgard errdisable|flap
  - Port-monitor takes action to shutdown(error-disable) a port when a counter hits a threshold
  - Extremely disruptive

Current Mitigation Mechanisms

- Congestion-isolation
  - Reduces scope of congestion but devices related by zoning still affected
  - Only works on multi-switch fabrics (across ISLs)
  - “All or nothing” approach
  - Automatic de-isolation mechanism now available in NX-OS 8.5(1)
  - Only applicable to the ‘slow drain’ case
- Fabric Performance Impact Notification(FPIN)
  - Available in NX-OS 8.5(1)
  - Only applies to brand new devices/HBAs
  - New standard, unknown benefits and results

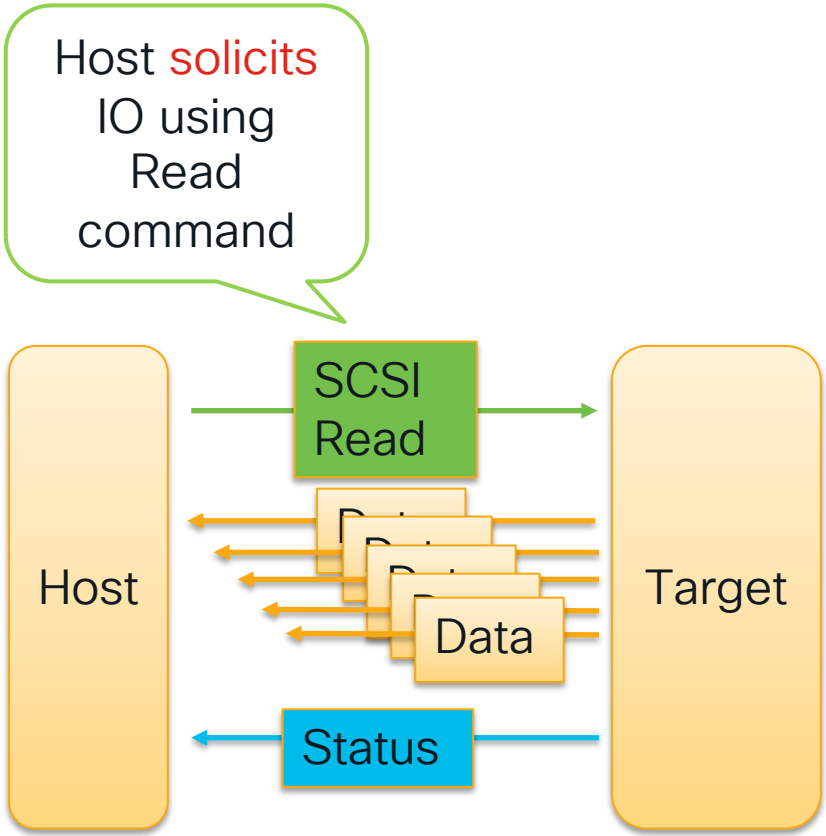
Note: FPIN has beta status in 8.5(1). This beta status will change to regular production status in an upcoming release.

Background – SCSI/NVMe IO

Background – SCSI/NVMe IO

- All data transfer in SCSI and NVMe is “solicited”
- Data is only transferred when the end device (host or target) specifically requests it
  - One minor exception SCSI Write ‘First Burst’

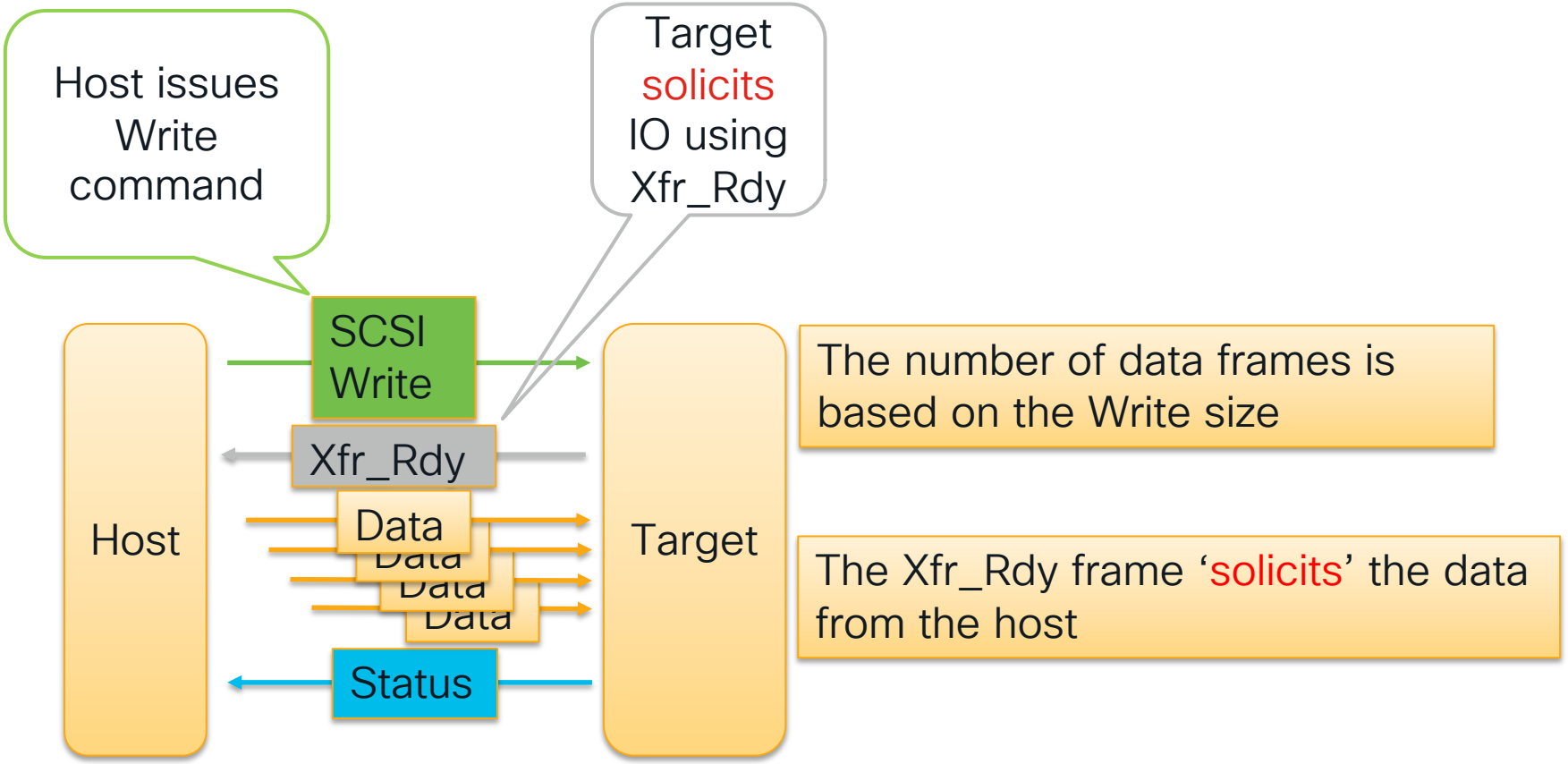
SCSI/NVMe Read IO Flow



The number of data frames is based on the Read size

The Read command ‘solicits’ the data from the target

SCSI/NVMe Write IO Flow



The number of data frames is based on the Write size

The Xfr_Rdy frame 'solicits' the data from the host

Dynamic Ingress Rate Limiting - Function

Dynamic Ingress Rate Limiting – Theory

3 premises, 1 Idea and 1 Conclusion

Premise 1-Slow Drain

Devices exhibit Slow Drain symptoms(TxWait) to reduce the rate of traffic they are receiving. Reducing the amount of data they are receiving will stop them from exhibiting slow drain symptoms (ex. TxWait)

Premise 2-Over Utilization

Reducing amount of solicited data will prevent the “backup” of data queued in the SAN

Premise 3-Ingress Rate Limit

Reducing the rate of ingress IO solicitation requests will make a proportional reduction in egress traffic rate

Dynamic Ingress Rate Limiting – Theory

3 premises, 1 idea and 1 Conclusion

Idea

Reduce ingress rate to reduce egress rate!

Conclusion

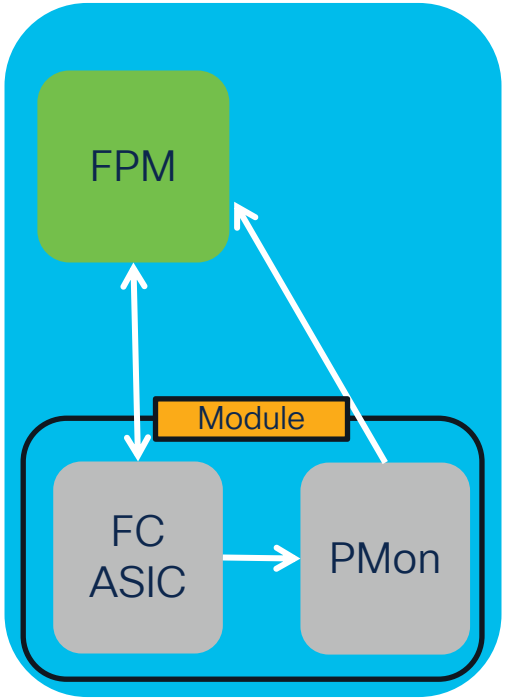
Once the egress rate has been reduced both “Slow Drain” and “Over Utilization” will be eliminated!

Ingress Rate Limiting

- Ingress rate limiting is a feature of all current MDS FC ASICs
- Works by slowing Buffer-to-Buffer credits to end device
- Hardware registers are programmed to set the ingress rate
- This can prevent the end device from transmitting any frames to the switchport for a time
- When ingress rate limit is applied SCSI Reads (host side) or Xfr_Rdys (target side) will be limited (as well as other frames)
- Hardware limits exist

Dynamic Ingress Rate Limiting – Overview

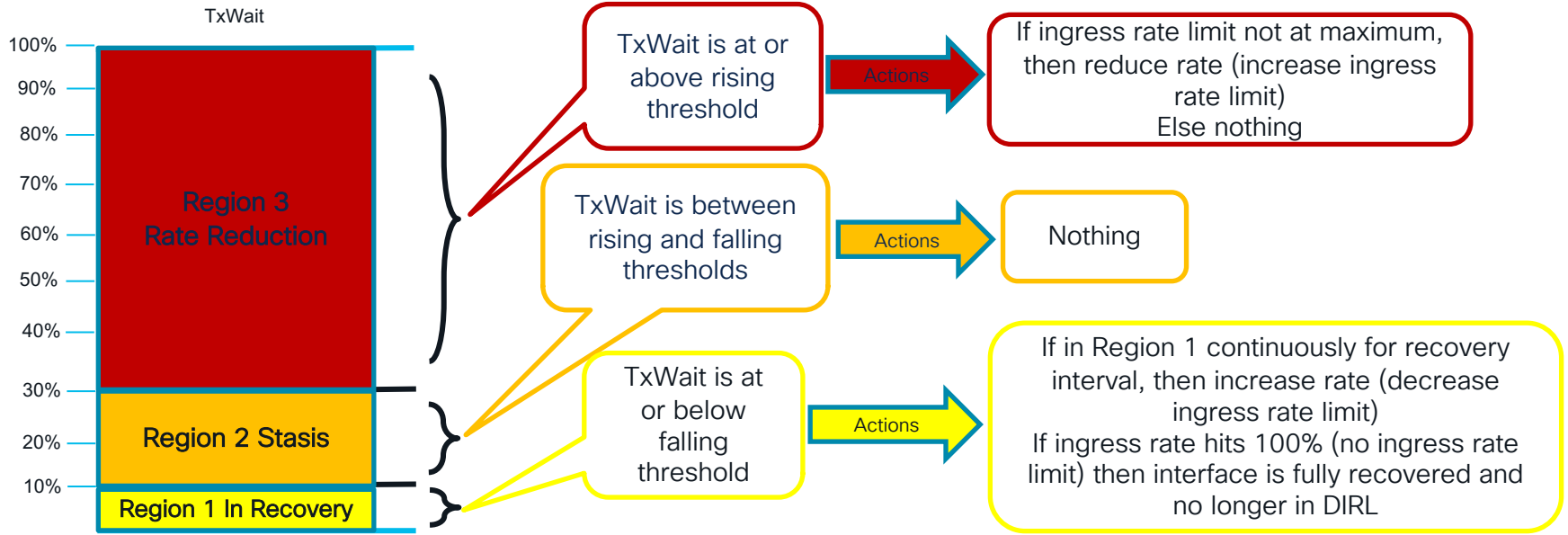
- DIRL utilizes the following components
  - **Port Monitor(Pmon)**
    - Detects slow drain/overutilization using the following 3 counters:
      1. Txwait
      2. Tx-datarate
      3. Tx-datarate-burst(new)
    - 'portguard dirl' is specified on the Pmon counter to activate
    - Maintains 'recovery interval timer'
  - **Fabric Performance Monitor – FPM –** New in NX-OS 8.5(1)
    - Handles the indications from Pmon to do rate reductions and increases
    - Signals port FC ASIC to reduce/increase ingress rate limit
  - **FC ASIC(FCMAC)**
    - Maintains counters for TxWait and Tx-datarate for Pmon
    - Programs ingress rate limit into port



Port Monitor Thresholds, Regions & DIRL Actions

Port Monitor Regions

Txwait with rising-threshold 30% falling-threshold 10%



Dynamic Ingress Rate Limiting – Overview

- **Step 1** – Port Monitor detects rising-threshold(Region 3) of counter configured with 'portguard dirl' on F port
- **Step 2** – Port Monitor notifies FPM to reduce the ingress rate by 50% from the **current 1 second ingress rate**
- **Step 3** – If Port Monitor continues to detect rising-threshold(Region 3) then it continues to notify FPM to reduce the ingress rate by 50%
- **Step 4** – If hardware limits are reached then issue syslog and leave port as is
- **Step 5** – When Port Monitor counter is between rising and falling threshold(Region 2) then no rate reductions or rate increases are done. Port Monitor starts checking if the counter is at or below the falling threshold(Region 1) continually for the entire recovery interval
- **Step 6** – Port Monitor counter is at or below the falling threshold(Region 1) then "Slow Drain" or "Over Utilization" is no longer detected, slowly increase the ingress rate to recover device (approx. 25% increase)

DIRL States

1. None – Port not currently in DIRL
2. **DIRL Rate reduction**
   - In Pmon Region 3
   - Notify FPM to reduce rate (increase ingress rate limit)
   - Short lived state
3. **DIRL Rate reduction Max**
   - In Pmon Region 3
   - Ingress Rate Limiting has reach max - No further rate limiting will be done
   - Could be long lived state

3. **DIRL Stasis**
   - In Pmon Regions 1 & 2(oscillating) or just Region 2
   - Start recovery interval timer
   - Take no actions
   - Could be long lived state
4. **DIRL In Recovery**
   - In Pmon Region 1 for entire recovery interval
   - Notify FPM to increase ingress rate (decrease ingress rate limit)
   - Long lived state since recovery interval is between 30 seconds and 5 minutes

Configuration

DIRL Configuration

2 (or 3) easy Steps

1. Configure feature fpm
2. Configure Port Monitor
3. Optional
   - Configure DIRL reduction percentage
   - Configure DIRL recovery percentage
   - Configure Port Monitor recovery interval

DIRL Configuration

Step 1 – Configure feature fpm

Fabric Performance Monitor needs to be enabled:

```
MDS(config)# feature fpm?
  fpm  Enable/Disable Fabric Performance Monitor
```

DIRL Configuration

Step 2 - Configure Port-monitor

Port-monitor needs to be configured with at least one of the following:

1. Txwait – This counter is for ‘slow drain’
2. Tx-datarate - This counter is for ‘Over Utilization’
3. Tx-datarate-burst - This counter is for ‘Over Utilization’ - New in NX-OS 8.5(1)

DIRL Configuration – Pmon counters

1. **TxWait – Measures the amount of time at 0 Tx credits**

Configured as a percentage of the polling interval

Example: poll-interval 1 (second) and rising/falling-threshold 30%/10%

This will trigger a rising-threshold alert when there is 300ms or more of TxWait in a 1 second interval

This will trigger a falling-threshold alert when there is 100ms or less of TxWait in a 1 second interval

DIRL Configuration – Pmon counters

2. Tx-Datarate – Measures high utilization leading to congestion

Configured as a percentage of the operational link speed

Example: poll-interval 10 (seconds) and rising/falling-threshold 80%/70%

This will trigger a rising-threshold alert when the Tx link utilization is 80% or more in a 10 second interval

This will trigger a falling-threshold alert when the Tx Utilization is 70% or less in a 10 second interval

Tx-datarate is an average over the poll-interval!

DIRL Configuration – Pmon counters

3. **Tx-Datarate-Burst – Measures high utilization in 1 second bursts leading to congestion**

Configured as the number of 1 second bursts at a Tx utilization percentage

Example: poll-interval 10 (seconds) and rising/falling-threshold 6/2 with datarate 90%

This will trigger a rising-threshold alert when the Tx link utilization is 90% or more for 6 one second intervals in a 10 second interval

This will trigger a falling-threshold alert when there are 2 or less 1 second intervals where Tx Utilization is 90% or more in a 10 second interval

New Pmon counter in NX-OS 8.5(1)

DIRL Configuration – PMon

The following is a very basic, minimum Pmon policy that will configure DIRL for both ‘Slow Drain’ and ‘Over Utilization’

```
MDS(config)# port-monitor name edge-dirl
```

Create policy named edge-dirl

```
MDS(config-port-monitor)# logical-type edge
```

Apply to edge ports only

```
MDS(config-port-monitor)# no monitor counter all
```

Shortcut to turn off monitoring for all counters not needed

DIRL Configuration – PMon

```
MDS(config)# monitor counter txwait
MDS(config)# counter txwait poll-interval 1 delta rising-threshold 30 event 4 falling-threshold 10 portguard dirl
```

Monitor and Configure Txwait w/DIRL for 'Slow Drain'

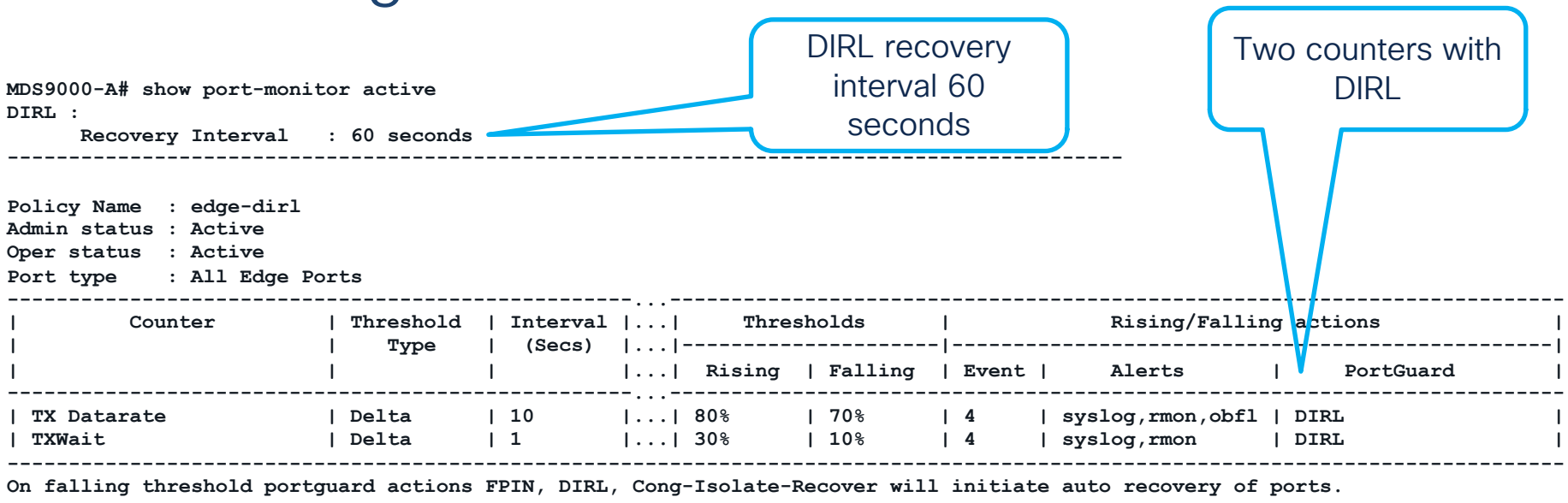
```
MDS(config)# monitor counter tx-datarate
MDS(config)# counter tx-datarate poll-interval 10 delta rising-threshold 80 event 4 falling-threshold 70 portguard dirl
```

Monitor and Configure tx-datarate w/DIRL for 'Over Utilization'

```
MDS(config)# port-monitor activate edge-dirl
```

Activate new policy

DIRL Configuration – PMon




```
MDS9000-A# show port-monitor active
DIRL :
      Recovery Interval    : 60 seconds
-------------------------------------------------------------------------------------------

Policy Name  : edge-dirl
Admin status : Active
Oper status  : Active
Port type    : All Edge Ports
-------------------------------------------------...-------------------------------------------------------------------
|         Counter          | Threshold  | Interval |...|     Thresholds      |             Rising/Falling actions             |
|                          |    Type    |  (Secs)  |...|---------------------|------------------------------------------------|
|                          |            |          |...|  Rising  | Falling  | Event |     Alerts       |     PortGuard        |
-------------------------------------------------...-------------------------------------------------------------------
| TX Datarate              | Delta      | 10       |...| 80%      | 70%      | 4     | syslog,rmon,obfl | DIRL                 |
| TXWait                   | Delta      | 1        |...| 30%      | 10%      | 4     | syslog,rmon      | DIRL                 |
-------------------------------------------------------------------------------------------------------------------------
On falling threshold portguard actions FPIN, DIRL, Cong-Isolate-Recover will initiate auto recovery of ports.
```

- Note above output modified to fit

DIRL Configuration – Optional

1. Configure recovery interval

```
MDS(config)# port-monitor dirl recovery-interval 120
```

DIRL recovery interval configured in Pmon
Recovery-interval defaults to 60 seconds

2. Configure reduction/recovery rates

```
MDS(config)# fpm dirl reduction 40 recovery 10
```

Reduction and recovery rates are configured in FPM
Reduction Rate defaults to 50%
Recovery Rate defaults to 25%

3. Include Target Devices

```
MDS# show fpm dirl exclude
All target device connected interfaces are excluded from DIRL

MDS(config)# fpm dirl exclude list
MDS(config-dirl-excl)# no member fc4-feature target

MDS# show fpm dirl exclude
No interface excluded from dynamic ingress rate limit.
```

Targets are excluded by default

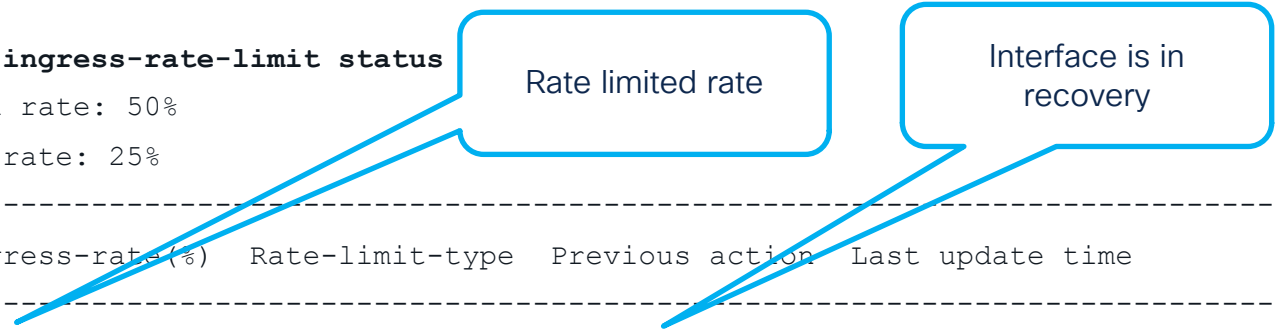
No devices excluded from DIRL now

Operation

DIRL Commands - Status

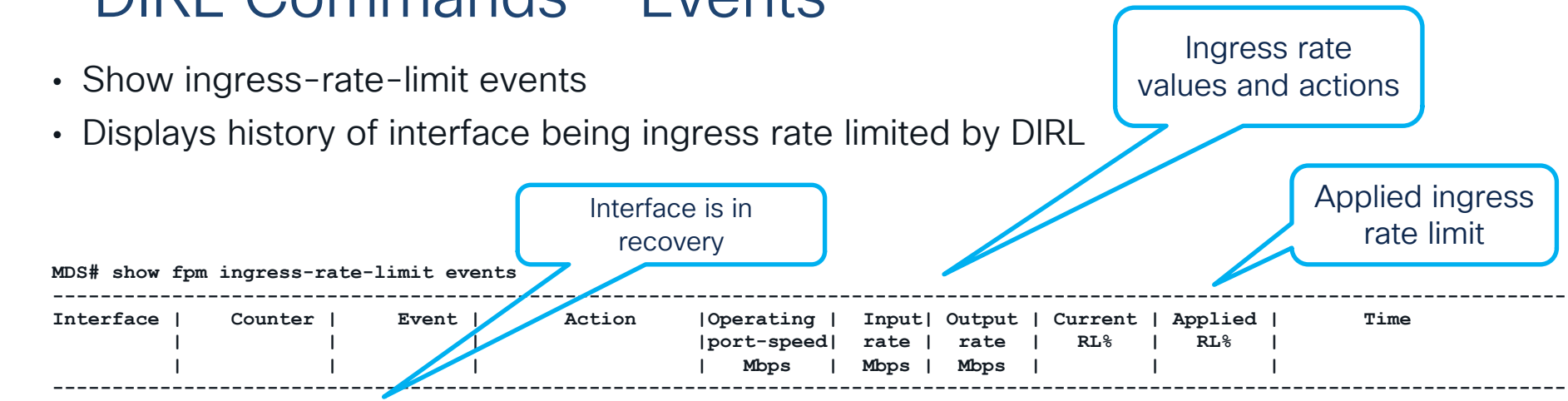
- Show ingress-rate-limit status
- Displays list of devices currently being ingress rate limited by DIRL

```
MDS# show fpm ingress-rate-limit status
DIRL reduction rate: 50%
DIRL recovery rate: 25%
-------------------------------------------------------------------------------
Interface  Ingress-rate(%)  Rate-limit-type  Previous action  Last update time
-------------------------------------------------------------------------------
fc9/17     0.2378           dynamic          rate-recovery    Thu Feb 11 13:43:24 2021
```



DIRL Commands - Events

- Show ingress-rate-limit events
- Displays history of interface being ingress rate limited by DIRL




```
MDS# show fpm ingress-rate-limit events
```

Interface	Counter	Event	Action	Operating port-speed Mbps	Input rate Mbps	Output rate Mbps	Current RL%	Applied RL%	Time
fc9/17	txwait	recovery	rate-recovery	4000	0.00	0.00	0.1218	0.1522	Thu Feb 11 13:41:24 2021
fc9/17	txwait	recovery	rate-recovery	4000	0.00	0.00	0.0974	0.1218	Thu Feb 11 13:40:24 2021
fc9/17	txwait	rising	rate-reduction-max	4000	4.23	0.04	0.0991	0.0974	Thu Feb 11 13:38:49 2021
fc9/17	txwait	rising	rate-reduction	4000	8.42	0.04	0.2916	0.0991	Thu Feb 11 13:38:48 2021
fc9/17	txwait	rising	rate-reduction	4000	24.78	0.04	0.6758	0.2916	Thu Feb 11 13:38:47 2021
fc9/17	txwait	rising	rate-reduction	4000	57.44	0.04	1.4037	0.6758	Thu Feb 11 13:38:46 2021
fc9/17	txwait	rising	rate-reduction	4000	119.31	0.04	2.8882	1.4037	Thu Feb 11 13:38:45 2021
fc9/17	txwait	rising	rate-reduction	4000	245.50	0.04	5.8746	2.8882	Thu Feb 11 13:38:44 2021
fc9/17	txwait	rising	rate-reduction	4000	499.34	0.04	11.9811	5.8746	Thu Feb 11 13:38:43 2021
fc9/17	txwait	rising	rate-reduction	4000	1018.39	0.04	24.3411	11.9811	Thu Feb 11 13:38:42 2021
fc9/17	txwait	rising	rate-reduction	4000	2068.99	0.04	49.2637	24.3411	Thu Feb 11 13:38:41 2021
fc9/17	txwait	rising	rate-reduction	4000	4187.41	0.04	100.0000	49.2637	Thu Feb 11 13:38:40 2021

DIRL syslog messages

When a device hits a rising-threshold with DIRL configured

%PMON-SLOT9-2-PMON_PORTGUARD_ACTION: PortGuard Config (DIRL) rate-limit action sent to FPM for Port fc9/33(1420000) Counter <counter_name>

Note: The above may happen more than once

When a device is recovering

%PMON-SLOT9-2-PMON_PORTGUARD_ACTION: PortGuard Config (DIRL) recovery action sent to FPM for Port fc9/2(1401000) Counter <counter_name>

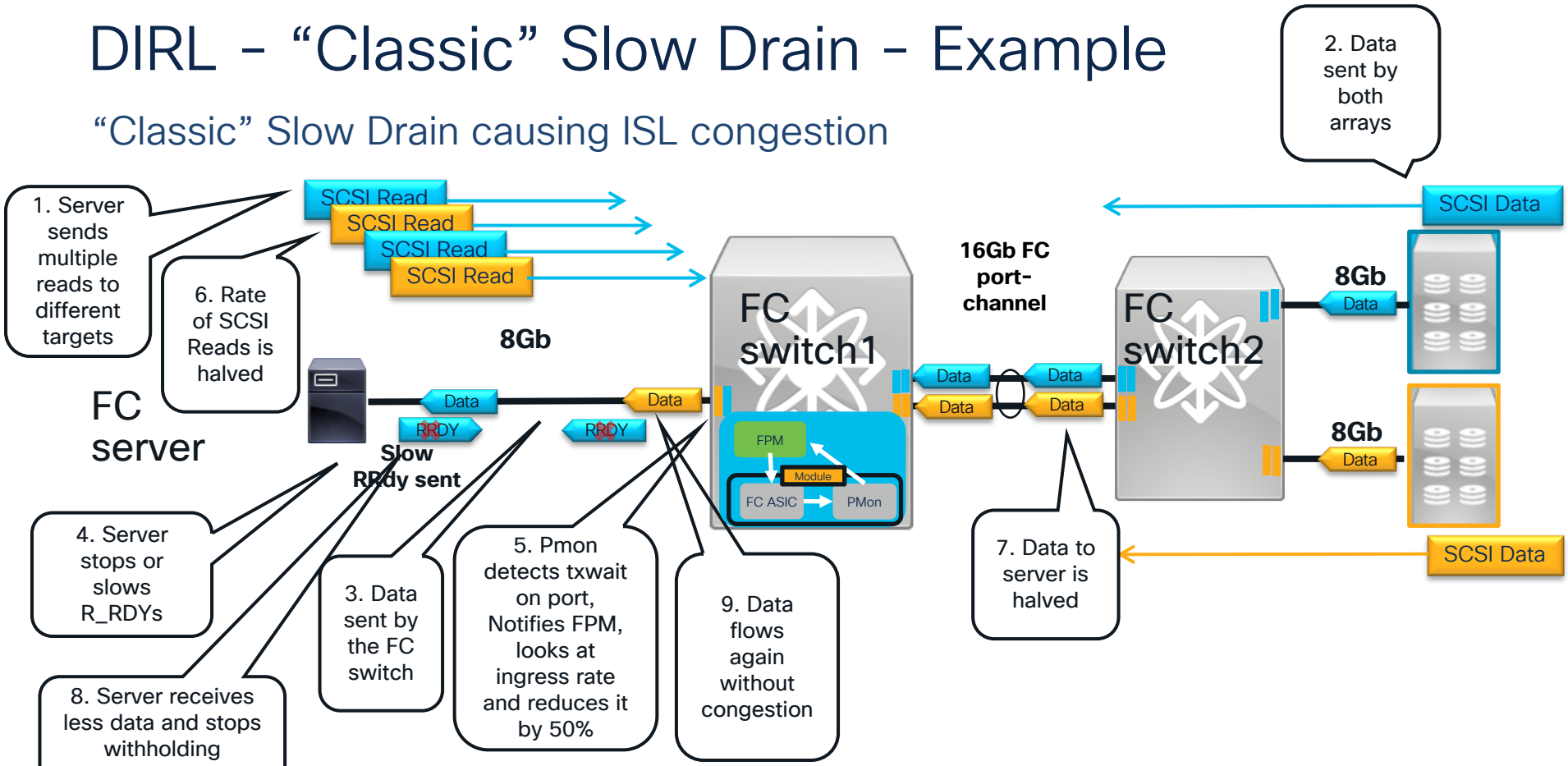
When a device is recovered

%FPM-5-DIRL_REC_COMPLETE_EVENT: fpm: Recovered interface <fc9/17> completely from dynamic ingress rate limit

%PMON-SLOT9-4-PMON_PORT_RECOVERED_FROM_CONGESTION: TXwait PG:DIRL port fc9/17[0x1410000] got recovered from rate-limit.

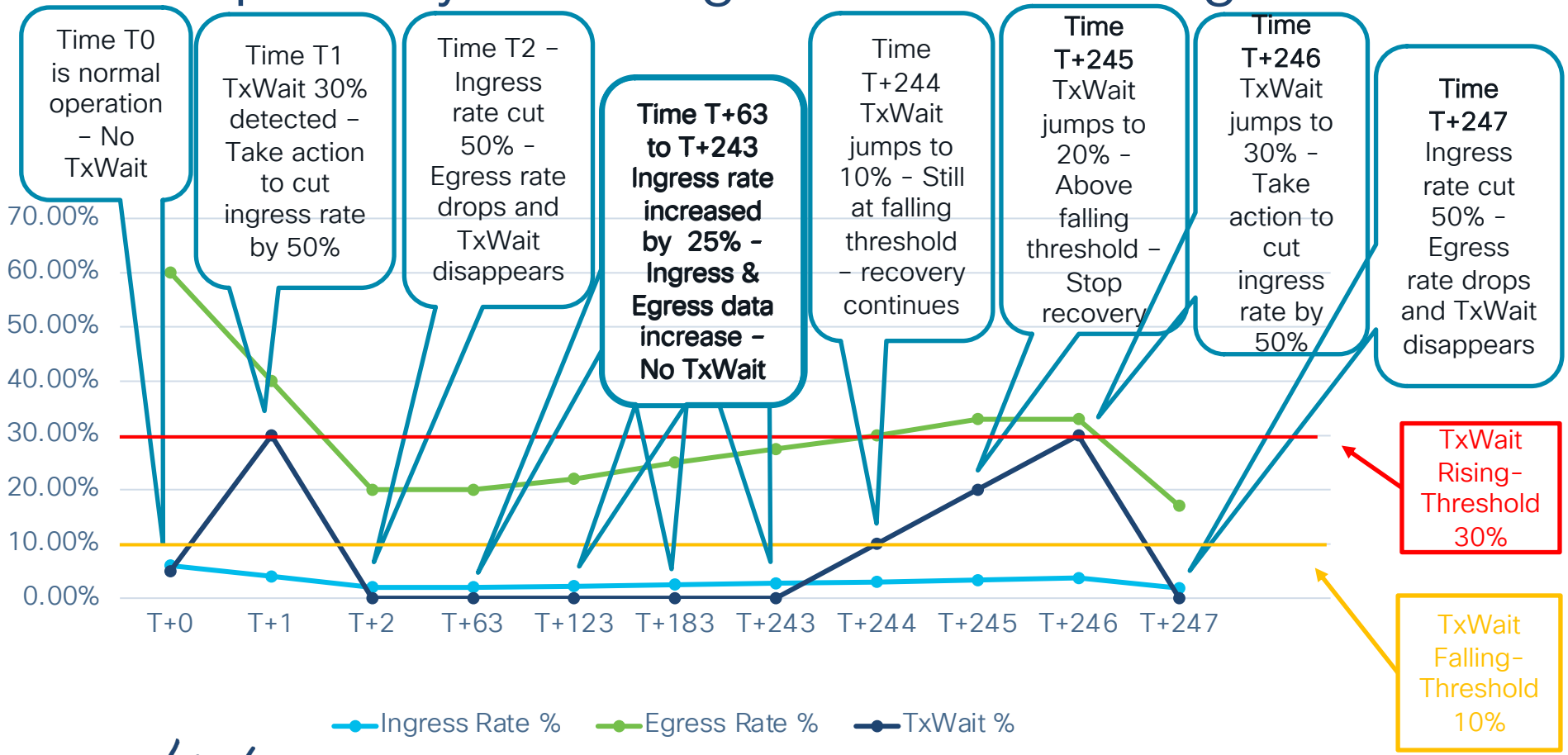
DIRL – "Classic" Slow Drain – Example

"Classic" Slow Drain causing ISL congestion



Congestion detected and mitigated without ancillary impact!

Example – Dynamic Ingress Rate Limiting



Conclusion – Benefits

DIRL – Benefits

- Limits effects of “Slow Draining”/”Over Utilization” device to itself only
- No packets are dropped from/to the device. Device will under perform but will not experience errors due to packet loss.
- Other related/ancillary devices(such as devices that are zoned with the device) are not affected
- Gradually, automatically returns the device to 100% ingress rate if it no longer exhibits evidence of problems. This could be very helpful to devices that are only slow during weekend backup periods.
- Does not require any inter switch communication. Both the determination of the “Slow Drain”/”Over Utilization” device and the rate limiting is on the same switch and switchport.
- Functions with single switch SANs.
- Can work on NPIV and NPV switches equally – NPV is for future release
- Works on all initiators and targets, all speeds and capabilities
- No modifications required from hosts and targets

Note: Not supported currently on MDS 9148S and 9250i

The bridge to possible

Thank you


cisco Live!
TURN
IT
UP
#CiscoLive