Understanding, Identifying and Resolving Congestion in a UCS Converged Infrastructure

De-blackboxifying congestion in the UCS

Edward Mazurek
Principal Engineer CX, @TheRealEdMaz

CISCO Live!

BRKCOM-2657

Cisco Webex App

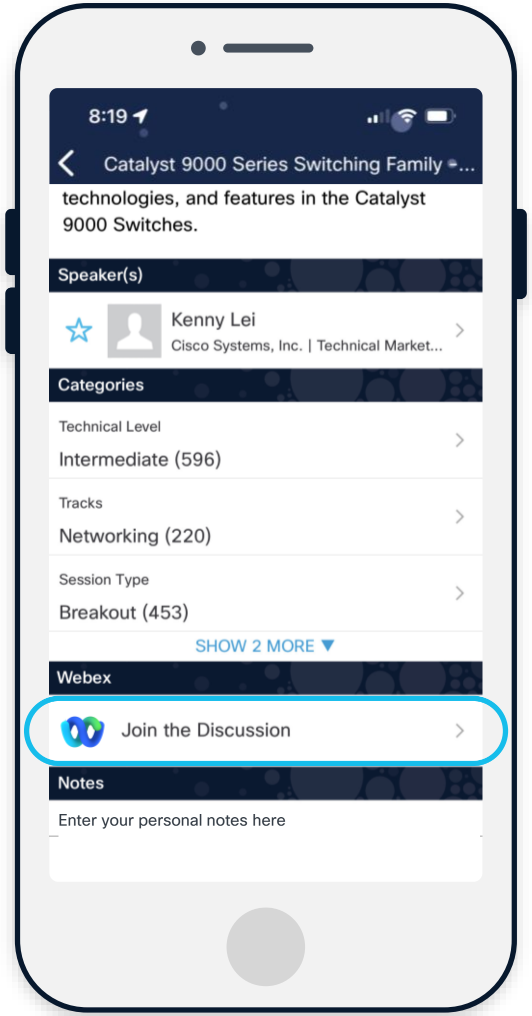
Questions?

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 13, 2025.



Agenda

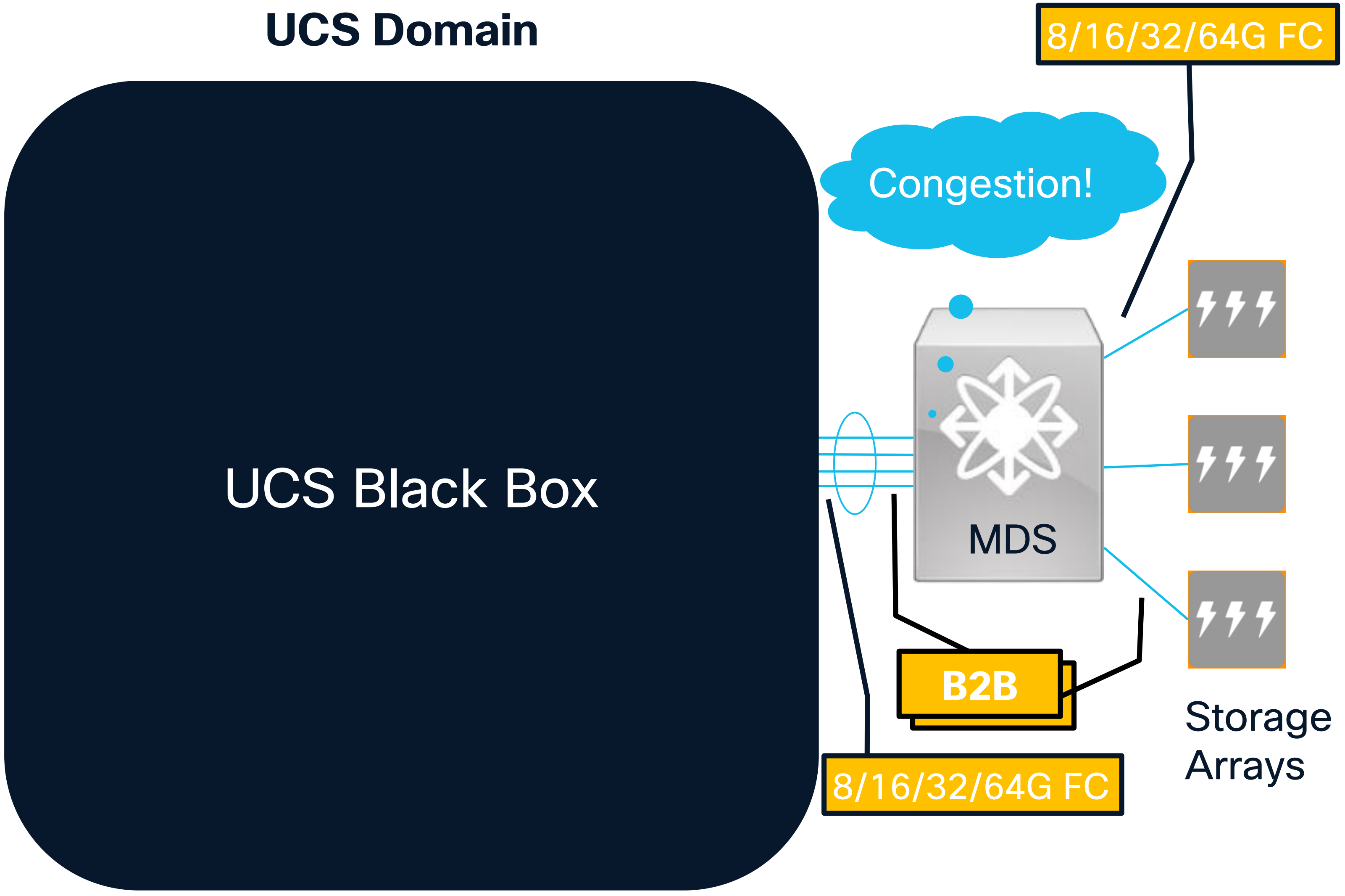
The UCS Congestion Problem

UCS Congestion Problem

Problem Types

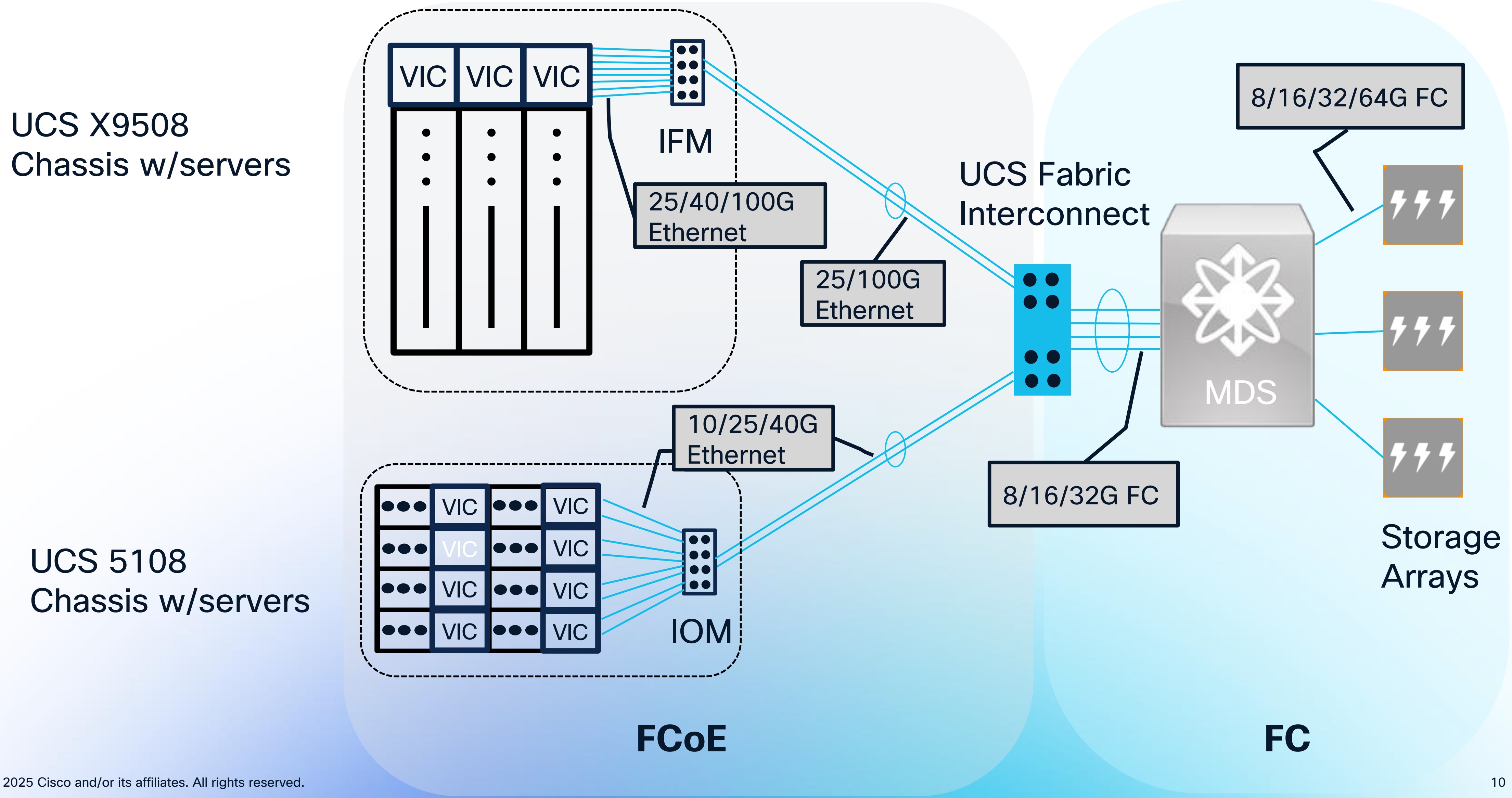
- SAN administrators report congestion on links to UCS
- Applications on servers report slow performance
- Storage arrays report slow performance/congestion or errors

How can the UCS "Black Box" be investigated to determine the reason for the congestion?



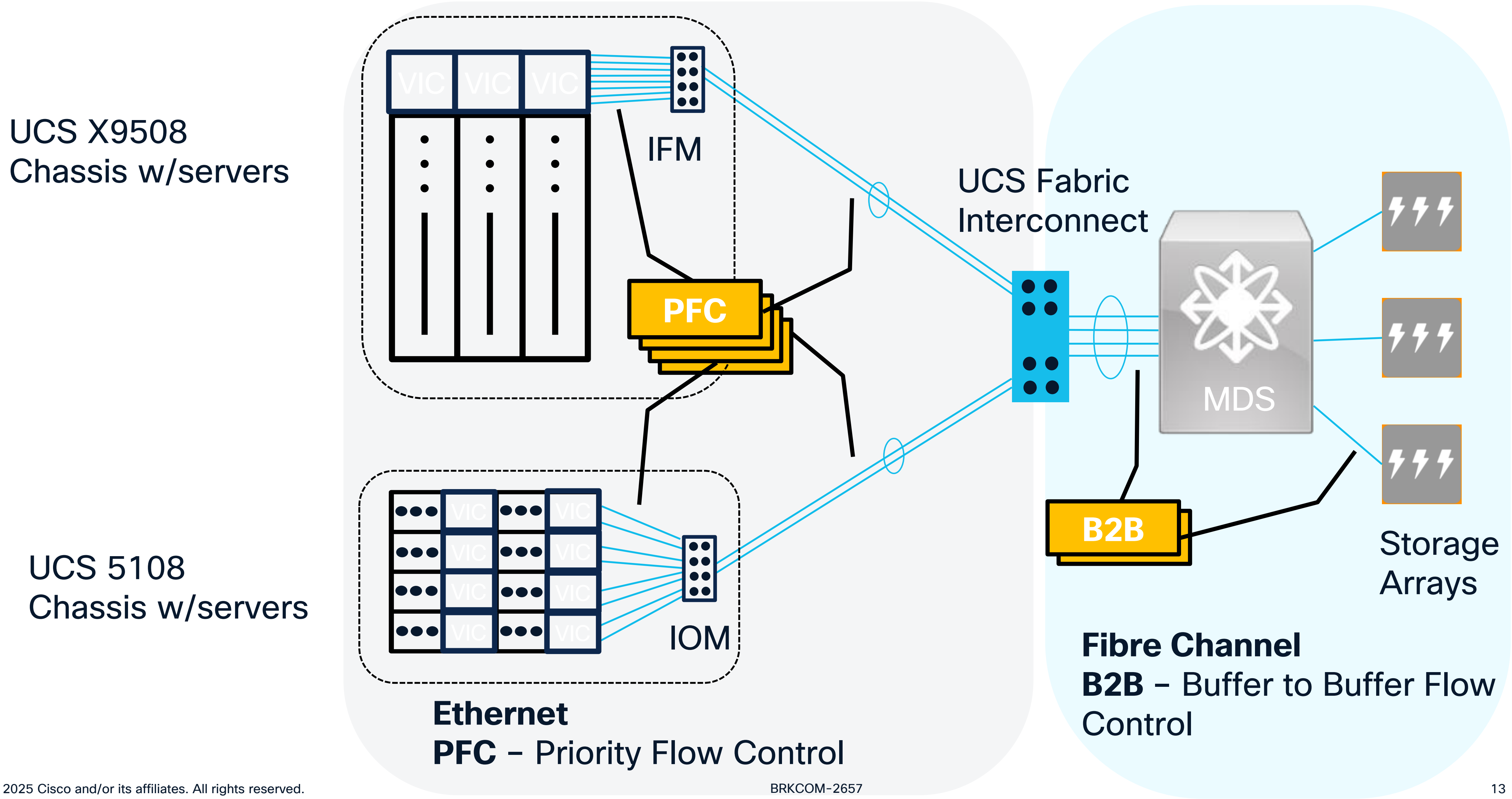
UCS Domain Logical Network Layout

UCS Domain Topology and Components



Flow Control Mechanisims

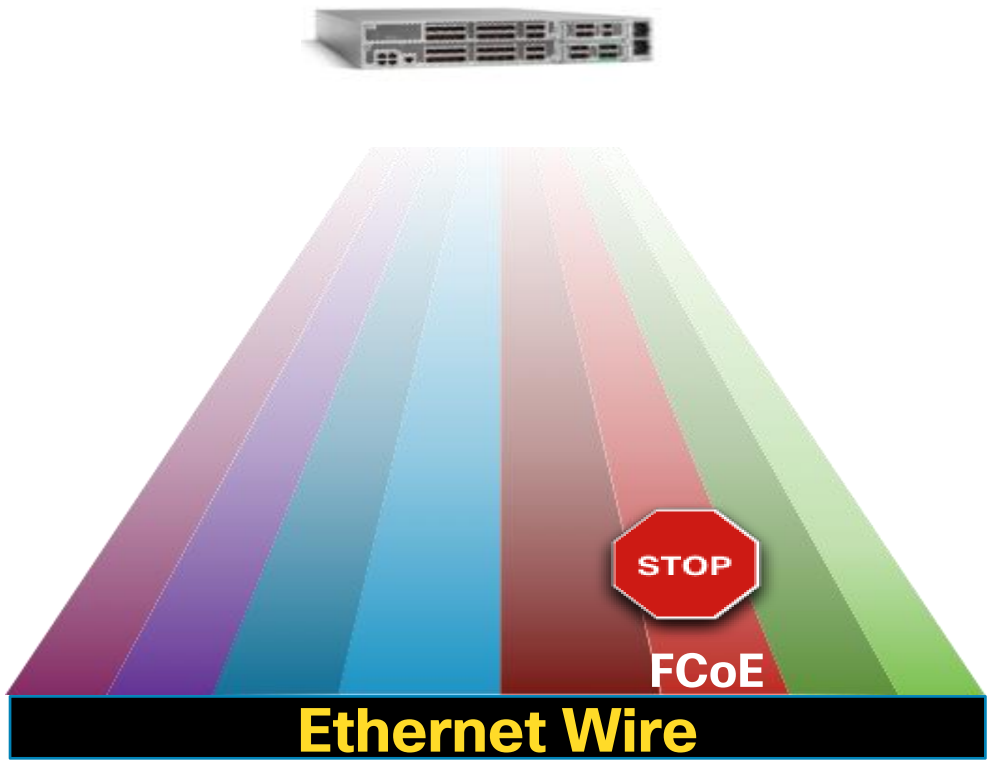
UCS Domain Flow Control Mechanisms



Ethernet Priority Flow Control - PFC

Just the facts

- PFC makes Ethernet into a lossless network protocol a.k.a “converged Ethernet’
- Priority Flow Control defined in IEEE 802.1Qbb
- Used by FCoE and RoCE
- Receiver can pause the FCoE/RoCE classes when congested
- Two types of Pauses
  - **Pause** (non-zero quanta) – Receiver must immediately stop sending **or continue not sending**
  - **UnPause** (zero quanta) – Receiver can immediately resume sending



Priority Flow Control - PFC

Just the Facts continued... PFC Quanta

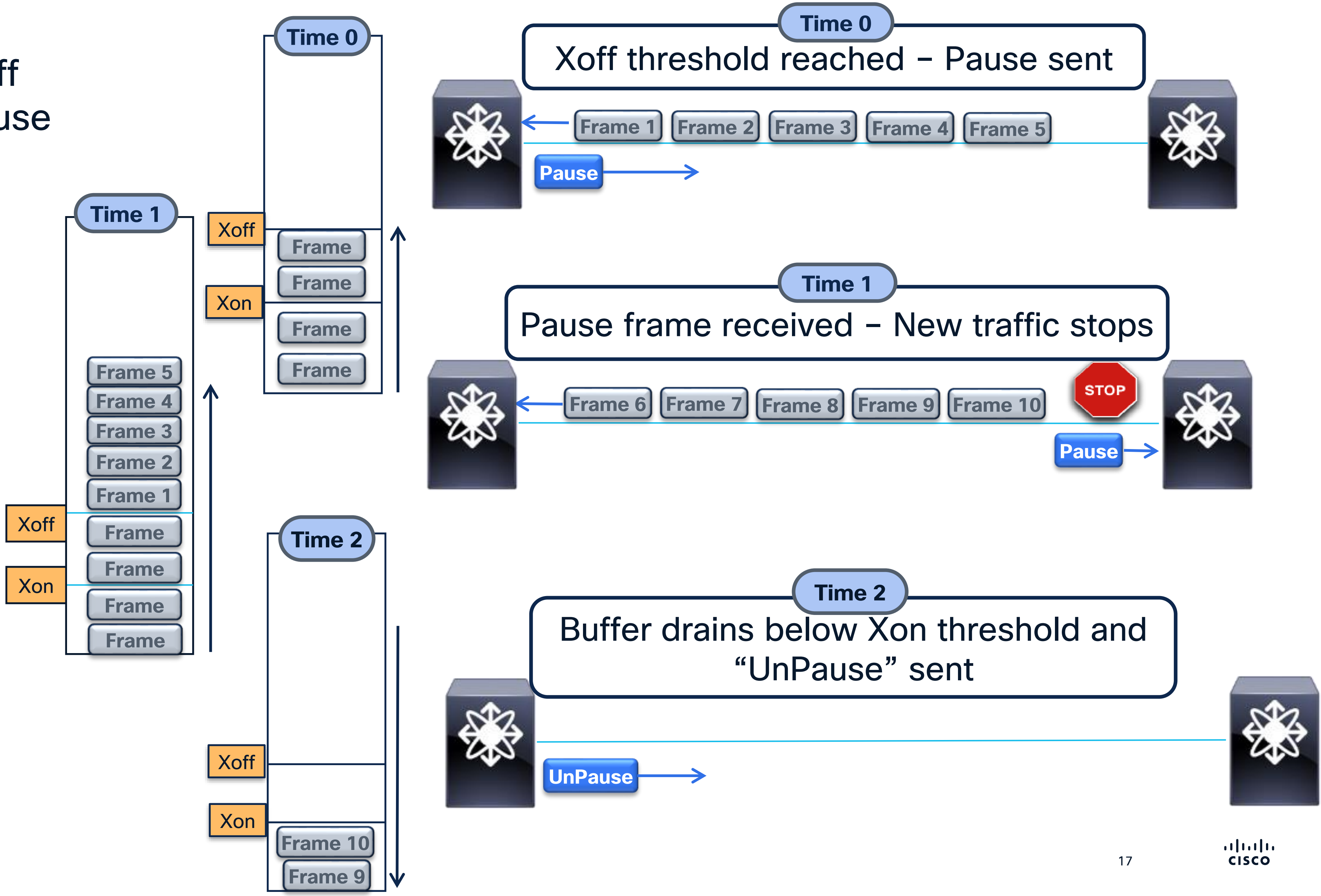
- Frames are transmitted unless Pause w/non-zero **quanta** has been received
- Amount of time class is paused is measured in **quanta**
- Each **quanta** is the amount of time taken to send 512 bits at link speed
- **Quanta** is contained in a two-byte field – Maximum value is 0xFFFF (65535)
- As speed increases the time represented by the **quanta** decreases
    - 0xFFFF at 10G 3.3ms
    - 0xFFFF at 100G 0.33ms
- Most Pauses do not ‘expire’ (pause traffic for the quanta duration)
  - Most Pauses are followed quickly by UnPauses
- The time the class is paused is the duration between Pause and UnPause
- Very important to understand which side of the link is reporting Tx/Rx Pauses!

PFC Flow Control Thresholds

PFC Buffer Xon and Xoff thresholds and PFC Pause

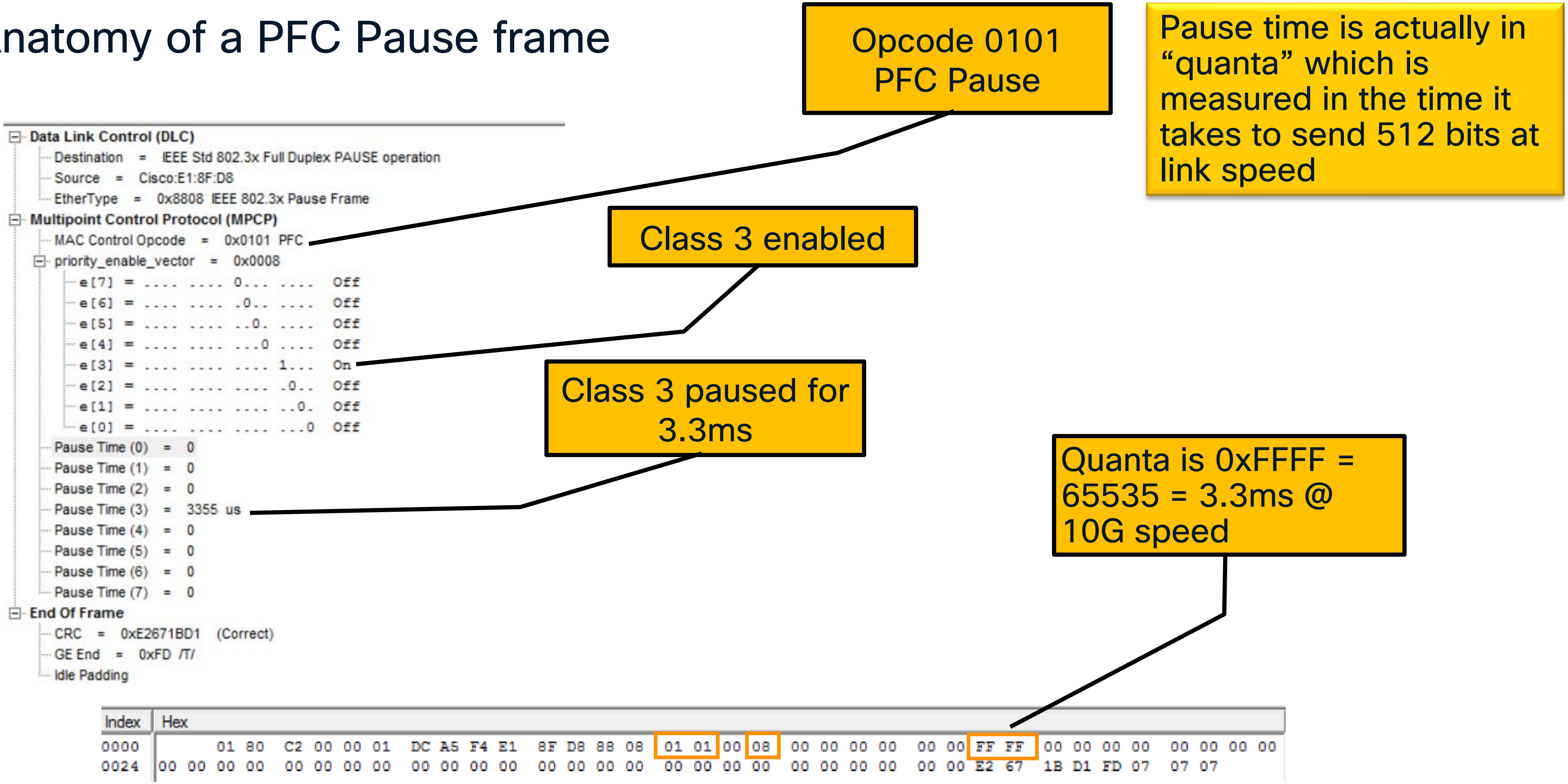
Pauses are sent based on Xoff and Xon thresholds being reached

Pauses are not necessarily sent because pauses have been received from other interfaces



Priority Flow Control

Anatomy of a PFC Pause frame



Priority Flow Control

Anatomy of a PFC UnPause frame

```
Data Link Control (DLC)
    Destination  =  IEEE Std 802.3x Full Duplex PAUSE operation
    Source  =  Cisco:E1:8F:D8
    EtherType  =  0x8808  IEEE 802.3x Pause Frame
Multipoint Control Protocol (MPCP)
    MAC Control Opcode  =  0x0101  PFC
    priority_enable_vector  =  0x0008
        e[7] = .... .... 0... ....  Off
        e[6] = .... .... .0.. ....  Off
        e[5] = .... .... ..0. ....  Off
        e[4] = .... .... ...0 ....  Off
        e[3] = .... .... .... 1...  On
        e[2] = .... .... .... .0..  Off
        e[1] = .... .... .... ..0.  Off
        e[0] = .... .... .... ...0  Off
    Pause Time (0)  =  0
    Pause Time (1)  =  0
    Pause Time (2)  =  0
    Pause Time (3)  =  0
    Pause Time (4)  =  0
    Pause Time (5)  =  0
    Pause Time (6)  =  0
    Pause Time (7)  =  0
End Of Frame
    CRC  =  0x83E44ABE   (Correct)
    GE End  =  0xFD  /T/
    Idle Padding
```

Opcode 0101
PFC Pause

Class 3 enabled

Class 3 "unpaused" since quanta is zero

When the priority enable bit is on for a class and the quanta is 0 the pause frame cancels a previous pause and allows for immediate resumption of transmission by the receiver

Quanta is 0x0000 = 0ms

```
Index  Hex
0000         01 80  C2 00 00 01  DC A5 F4 E1  8F D8 88 08  01 01 00 08  00 00 00 00  00 00 00 00  00 00 00 00  00 00 00 00
0024   00 00 00 00  00 00 00 00  00 00 00 00  00 00 00 00  00 00 00 00  00 00 00 00  00 00 83 E4  4A BE FD 07  07 07
```

Priority Flow Control

Pause and UnPause timing Example

PFC Pause

Resume(UnPause) 57us later

Icon	Bookmark	us.ns	Delt		e A	Side B	Protoc	Summary
Beg								
10 BM	Pause	64.334		GE Port(1,2,2)		Ether Frame	MPCP	PFC; Pause(3) = 3355;
10 BM	Resume 57us	121.259	56.925	GE Port(1,2,2)		Ether Frame	MPCP	PFC;
10 BM	Pause	176.974	55.715	GE Port(1,2,2)		Ether Frame	MPCP	PFC; Pause(3) = 3355;
10 BM	Resume 57us	233.944	56.970	GE Port(1,2,2)		Ether Frame	MPCP	PFC;
10 BM	Pause	289.664	55.720	GE Port(1,2,2)		Ether Frame	MPCP	PFC; Pause(3) = 3355;
10 BM	Resume 57us	346.689	57.025	GE Port(1,2,2)		Ether Frame	MPCP	PFC;
10 BM	Pause	402.399	55.710	GE Port(1,2,2)		Ether Frame	MPCP	PFC; Pause(3) = 3355;
10 BM	Resume 57us	459.329	56.930	GE Port(1,2,2)		Ether Frame	MPCP	PFC;
10 BM	Pause	515.044	55.715	GE Port(1,2,2)		Ether Frame	MPCP	PFC; Pause(3) = 3355;
10 BM	Resume 57us	571.914	56.870	GE Port(1,2,2)		Ether Frame	MPCP	PFC;
10 BM	Pause	627.634	55.720	GE Port(1,2,2)		Ether Frame	MPCP	PFC; Pause(3) = 3355;
10 BM	Resume 57us	684.564	56.930	GE Port(1,2,2)		Ether Frame	MPCP	PFC;
10 BM	Pause	740.279	55.715	GE Port(1,2,2)		Ether Frame	MPCP	PFC; Pause(3) = 3355;
10 BM	Resume 57us	797.244	56.965	GE Port(1,2,2)		Ether Frame	MPCP	PFC;
10 BM	Pause	852.964	55.720	GE Port(1,2,2)		Ether Frame	MPCP	PFC; Pause(3) = 3355;
10 BM	Resume 57us	909.919	56.955	GE Port(1,2,2)		Ether Frame	MPCP	PFC;
10 FR		965.639	55.720	GE Port(1,2,2)		Ether Frame	MPCP	PFC; Pause(3) = 3355;
10 FR		1022.540	56.901	GE Port(1,2,2)		Ether Frame	MPCP	PFC;

Although each pause specifies 3.3ms the actual time paused is usually much less than that

Total pause time for 8 PFC Pauses ~456us

Priority Flow Control

Pause and UnPause timing Example – Severe congestion

Event View

PFC Pause

Icon	Bookmark	mm:ss.ms_us_ns_ps		A	Side B	Summary	...	Bytes
Beg								
10 BM	1st Pause	29:44.655_825_773_9		Port(1,1,6)	MPCP	PFC; Pause Time (3) = 3355 us;		65
10 BM	1st UnPause	29:44.655_973_616_4	147.8425	Port(1,1,6)	MPCP	PFC;		65
10 BM	Start of Pause block	29:44.656_094_546_4	120.9301	Port(1,1,6)	MPCP	PFC; Pause Time (3) = 3355 us;		
10 BM	Pause continued	29:44.659_449_200_8	3354.6545	Port(1,1,6)	MPCP	PFC; Pause Time (3) = 3355 us;		
10 BM	Pause continued	29:44.662_803_851_4	3354.6506	Port(1,1,6)	MPCP	PFC; Pause Time (3) = 3355 us;		
10 BM	Pause continued	29:44.666_158_503_9	3354.6526	Port(1,1,6)	MPCP	PFC; Pause Time (3) = 3355 us;		
10 BM	Pause continued	29:44.669_513_156_4	3354.6526	Port(1,1,6)	MPCP	PFC; Pause Time (3) = 3355 us;		
10 BM	Pause continued	29:44.672_867_803_9	3354.6476	Port(1,1,6)	MPCP	PFC; Pause Time (3) = 3355 us;		
10 BM	Pause continued	29:44.676_222_456_4	3354.6526	Port(1,1,6)	MPCP	PFC; Pause Time (3) = 3355 us;		
10 BM	Pause continued	29:44.679_577_103_9	3354.6476	Port(1,1,6)	MPCP	PFC; Pause Time (3) = 3355 us;		
10 BM	Pause continued	29:44.682_931_756_4	3354.6525	Port(1,1,6)	MPCP	PFC; Pause Time (3) = 3355 us;		
10 BM	UnPause	29:44.685_662_188_9	2730.4325	Port(1,1,6)	MPCP	PFC;		
10 FR		29:44.685_777_856_4	115.6676	Port(1,1,6)	MPCP	PFC; Pause Time (3) = 3355 us;		65
10 FR		29:44.685_938_258_9	160.4025	Port(1,1,6)	MPCP	PFC;		65
10 FR		29:44.686_073_278_9	135.0200	Port(1,1,6)	MPCP	PFC; Pause Time (3) = 3355 us;		65
10 FR		29:44.689_427_932_5	3354.6537	Port(1,1,6)	MPCP	PFC; Pause Time (3) = 3355 us;		65

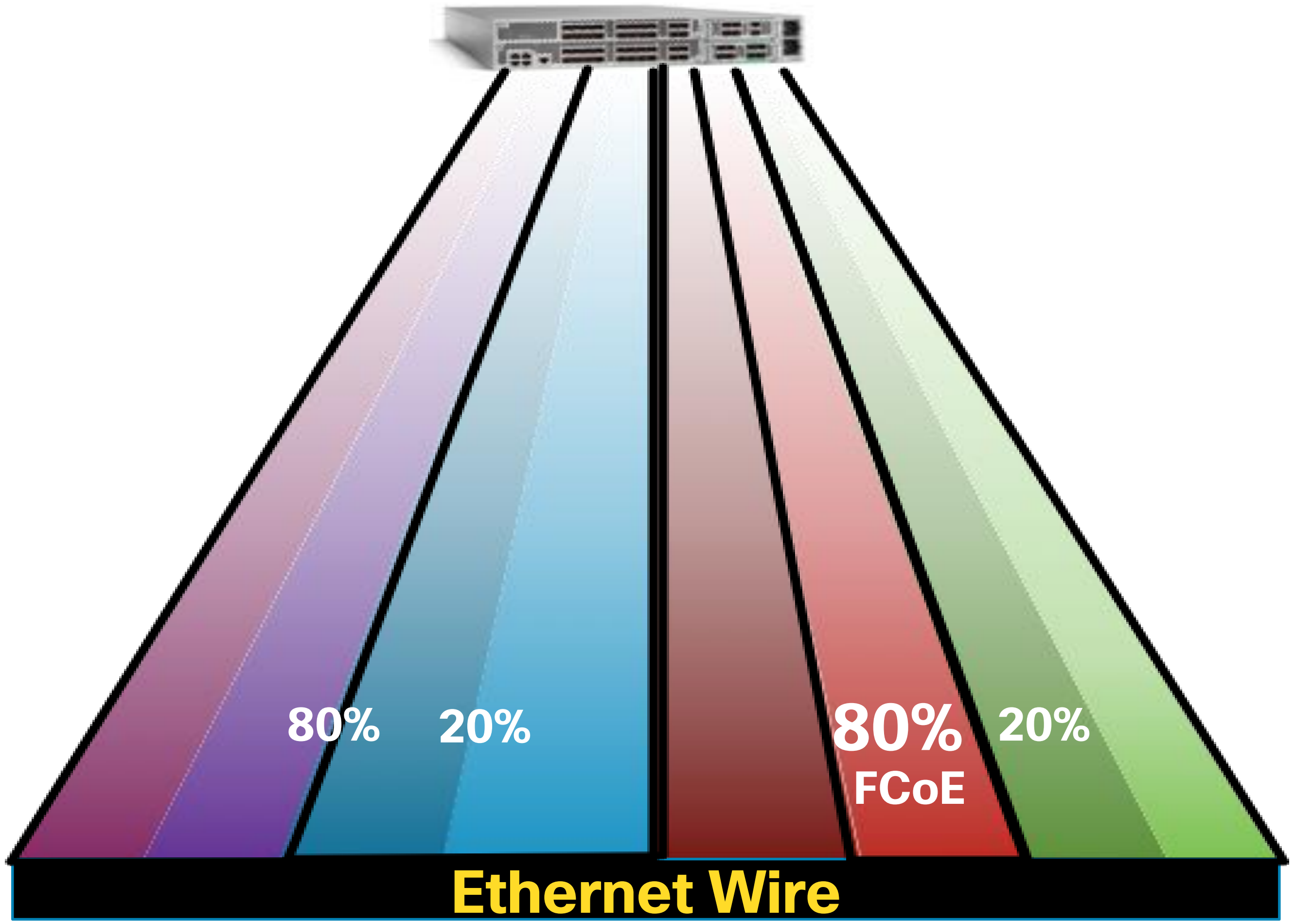
This example shows Pauses that are continued multiple times

UnPause 19ms later!

Enhanced Transmission Selection - ETS

Just the facts

- Guarantees each class a minimum percentage of the link
- Enhanced Transmission Selection defined in IEEE 802.1Qbb
- When there is available link bandwidth classes can transmit at higher rates
- Defined in egress queuing policy-map
- Congestion may appear only in the presence of other traffic!



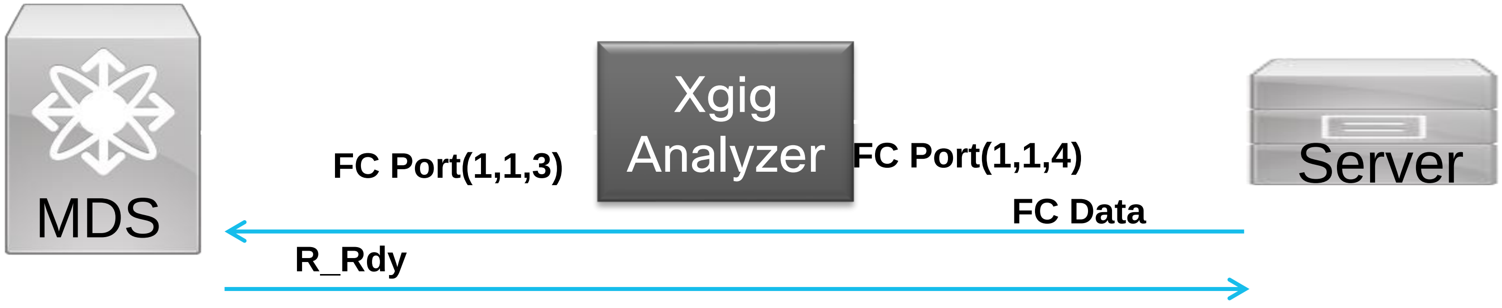
Fibre Channel Buffer to Buffer Flow Control

Just the Facts

- Fibre Channel is a 'lossless' network protocol by design
- Sender does not send a frame unless the receiver has a buffer
- Fibre Channel utilizes Buffer-to-Buffer(B2B) Credit based flow control
- Each side of link informs adjacent side of the number of buffers/credits
- Each frame sent requires a B2B credit to be returned
- B2B credits are also called 'R_RDYs'
- **Frame receivers can slow rate of ingress traffic by 'withholding' credits**
- **If a sender runs out of credits it must stop sending until it receives one**

Fibre Channel Flow Control – Example

Normal flow



ports

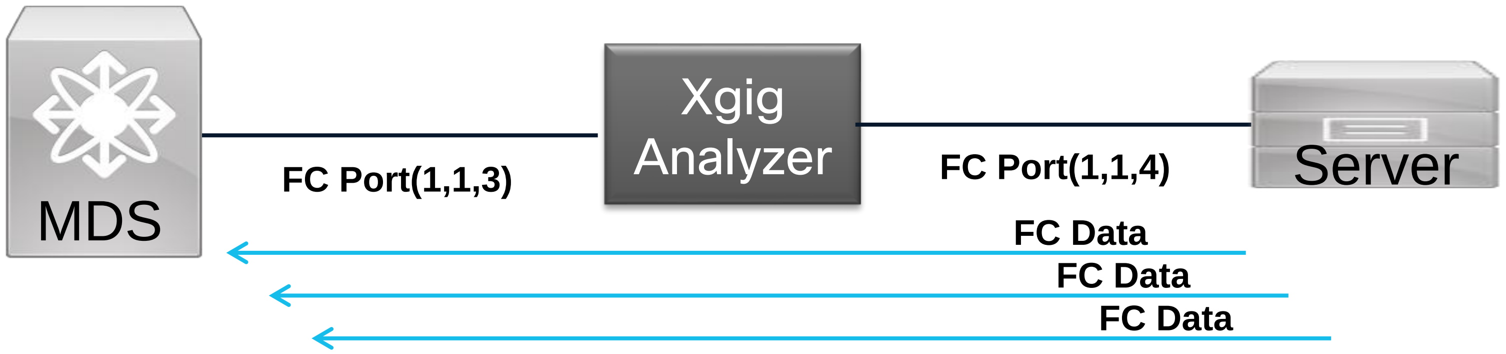
Icon	Bool	mm:ss.ms_us_ns (Ri	Delta Time	Port	Side A	Side B	Summary	Bytes	Source	Destinat	LUN	OX_Id
4 FR		143:34.083_879_824	34.720	FC Port(1,1,4)		FC4Cmd	Write(10); LUN = 0x0020; LBA = 0x55C64C08; FCP_DL = 0x00002000;	68	780087	280085	0020	4DCD
4 OS		143:34.083_880_516	0.692	FC Port(1,1,3)	1 - R-Rdy			4				
4 FR		143:34.083_882_668	2.152	FC Port(1,1,4)		FC4Cmd	Write(10); LUN = 0x0020; LBA = 0x56C20020; FCP_DL = 0x00002000;	68	780087	280085	0020	4FC2
4 OS		143:34.083_883_396	0.728	FC Port(1,1,3)	1 - R-Rdy			4				
4 FR		143:34.083_900_616	17.220	FC Port(1,1,4)		FC4SData	FC4SData; SCSI FCP; Offset = 0x00000000; Len = 0x0800;	164/2084	780084	280085	0027	41DB
4 FR		143:34.083_905_660	5.044	FC Port(1,1,4)		FC4SData	FC4SData; SCSI FCP; Offset = 0x00000800; Len = 0x0800;	164/2084	780084	280085	0027	41DB
4 [illegible]		[illegible]	4.922	FC Port(1,1,4)		1 - R-Rdy		4				
4 [illegible]		[illegible]	0.334	FC Port(1,1,3)	FC4SData		FC4SData; SCSI FCP; Offset = 0x00007800; Len = 0x0800;	164/2084	280085	780085	004D	4535
4 OS		143:34.083_915_838	4.922	FC Port(1,1,3)	1 - R-Rdy			4				
4 OS		143:34.083_915_866	0.028	FC Port(1,1,3)	1 - R-Rdy			4				
4 OS		143:34.083_922_712	6.846	FC Port(1,1,4)		1 - R-Rdy		4				
4 FR		143:34.083_923_044	0.332	FC Port(1,1,3)	FC4SData		FC4SData; SCSI FCP; Offset = 0x00008000; Len = 0x0800;	164/2084	280085	780085	004D	4535
4 OS		143:34.083_967_712	44.668	FC Port(1,1,4)		1 - R-Rdy		4				
4 FR		143:34.083_968_040	0.328	FC Port(1,1,3)	FC4SData		FC4SData; SCSI FCP; Offset = 0x00008800; Len = 0x0800;	164/2084	280085	780085	004D	4535
4 OS		143:34.083_996_876	28.836	FC Port(1,1,4)		1 - R-Rdy		4				
4 FR		143:34.083_997_216	0.340	FC Port(1,1,3)	FC4SData		FC4SData; SCSI FCP; Offset = 0x00009000; Len = 0x0800;	164/2084	280085	780085	004D	4535
4 FR		143:34.084_013_284	16.068	FC Port(1,1,4)		FC4Cmd	Read(10); LUN = 0x0004; LBA = 0x011975A8; FCP_DL = 0x00002000;	68	780096	280085	0004	486F
4 OS		143:34.084_013_996	0.712	FC Port(1,1,3)	1 - R-Rdy			4				
4 OS		143:34.084_067_740	53.744	FC Port(1,1,4)		1 - R-Rdy		4				

Delta time ~0.7us

FC data and R_Rdy

Fibre Channel Flow Control – Example cont’

Delayed/No R_RDYs



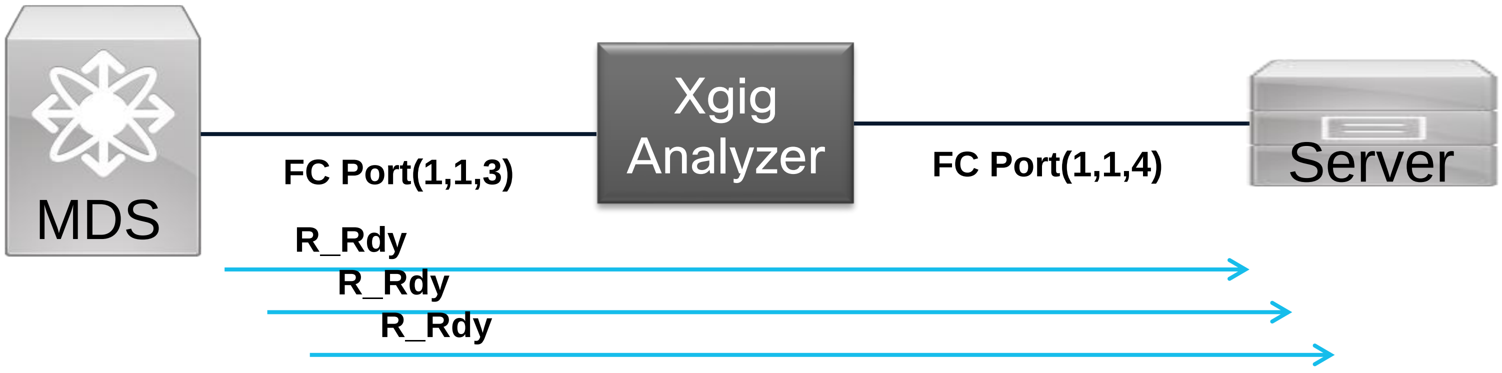
Icon	Bool	mm:ss.ms_us_ns (R	Delta Time	Port	Side A	Side B	Summary	Bytes	Source	Destinat	LUN	OX_Id
4 FR		143:49.879_845_500	5.024	FC Port(1,1,4)		FC4SData	FC4SData; SCSI FCP; Offset = 0x00000800; Len = 0x0800;	164/2084	780087	280085	0004	4F20
4 FR		143:49.879_850_460	4.960	FC Port(1,1,4)		FC4SData	FC4SData; SCSI FCP; Offset = 0x00001000; Len = 0x0800;	164/2084	780087	280085	0004	4F20
4 FR		143:49.879_855_420	4.960	FC Port(1,1,4)		FC4SData	FC4SData; SCSI FCP; Offset = 0x00001800; Len = 0x0800;	164/2084	780087	280085	0004	4F20
4 FR		[illegible]	[illegible]	FC Port(1,1,4)		FC4SData	FC4SData; SCSI FCP; Offset = 0x00000000; Len = 0x0800;	164/2084	780086	280085	0040	4AA2
4 FR		[illegible]	[illegible]	FC Port(1,1,4)		FC4SData	FC4SData; SCSI FCP; Offset = 0x00000800; Len = 0x0800;	164/2084	780086	280085	0040	4AA2
4 FR		[illegible]	[illegible]	FC Port(1,1,4)		FC4SData	FC4SData; SCSI FCP; Offset = 0x00001000; Len = 0x0800;	164/2084	780086	280085	0040	4AA2
4 FR		143:49.879_927_228	5.004	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4AA2
4 FR		143:49.879_932_452	5.224	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4AA2
4 FR		143:49.879_937_496	5.044	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4AA2
4 FR		143:49.879_942_456	4.960	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4AA2
4 FR		143:49.879_947_416	4.960	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4AA2
4 FR		143:49.879_952_396	4.980	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4AA2
4 FR		143:49.879_957_356	4.960	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4AA2
4 FR		143:49.879_962_336	4.980	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4AA2
4 FR		143:49.879_967_548	5.212	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4AA2
4 FR		143:49.879_972_592	5.044	FC Port(1,1,4)		FC4SData	FC4SData; SCSI FCP; Offset = 0x00006000; Len = 0x0800;	164/2084	780086	280085	0040	4AA2

Only data – no R_Rdys

FC4SData; SCSI FCP; Offset = 0x00000800; Len = 0x0800;
FC4SData; SCSI FCP; Offset = 0x00001000; Len = 0x0800;
FC4SData; SCSI FCP; Offset = 0x00001800; Len = 0x0800;
FC4SData; SCSI FCP; Offset = 0x00000000; Len = 0x0800;

Fibre Channel Flow Control – Example cont’

R-RDY recovery



Icon	Bool	mm:ss.ms_us_ns (R	Delta Time	Port	Side A	Side B	Summary	Bytes	Source	Destinat	LUN
4 os		143:49.883_147_828	4.764	FC Port(1,1,3)	1 - R-Rdy			4			
4 os		143:49.883_149_372	1.544	FC Port(1,1,3)	1 - R-Rdy			4			
4 os		143:49.883_150_924	1.552	FC Port(1,1,3)	1 - R-Rdy			4			
4 os		143:49.883_152_524	1.600	FC Port(1,1,3)	1 - R-Rdy			4			
4 FR		143:49.883_157_588	5.064	FC Port(1,1,3)	FC4SData		FC4SData; SCSI FCP; Offset = 0x00000000; Len = 0x0800;	164/2084	280085	780084	0053
4 os		143:49.883_157_960	0.372	FC Port(1,1,4)		1 - R-Rdy		4			
4 os		143:49.883_162_510	4.550	FC Port(1,1,3)	1 - R-Rdy			4			
4 os		143:49.883_162_538	0.028	FC Port(1,1,3)	1 - R-Rdy			4			
4 os		143:49.883_162_566	0.028	FC Port(1,1,3)	1 - R-Rdy			4			
4 FR		143:49.883_162_594	0.028	FC Port(1,1,3)	FC4SData		FC4SData; SCSI FCP; Offset = 0x00000800; Len = 0x0800;	164/2084	280085	780084	0053
4 os		143:49.883_162_980	0.386	FC Port(1,1,4)		1 - R-Rdy		4			
4 os		143:49.883_167_518	4.538	FC Port(1,1,3)	1 - R-Rdy			4			
4 os		143:49.883_167_546	0.028	FC Port(1,1,3)	1 - R-Rdy			4			
4 os		143:49.883_167_574	0.028	FC Port(1,1,3)	1 - R-Rdy			4			

R_Rdys start arriving

More R_Rdys

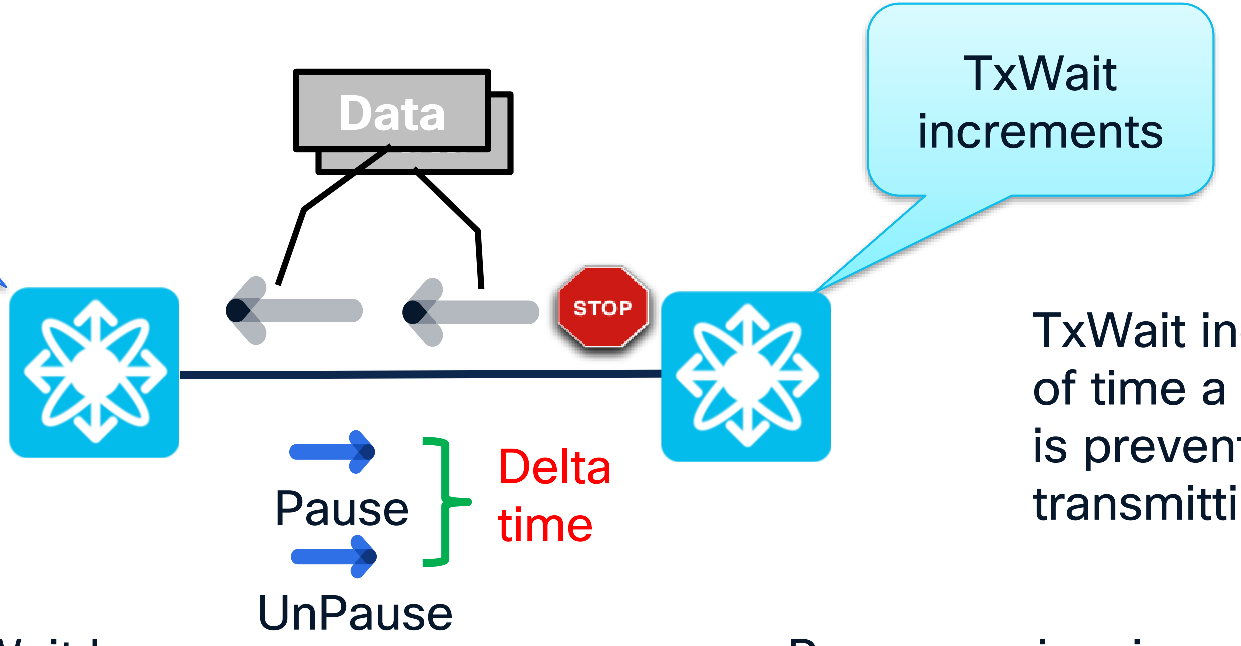
More R_Rdys

TxWait and RxWait

TxWait and RxWait

- **TxWait and RxWait exist in both Fibre Channel and PFC Ethernet**
- **Both are measures of time in 2.5μs increments (2.5 microseconds)**
- **Increment separately by class**
- **TxWait increments when frame sender cannot transmit**
  - In FC frame sender has zero Tx credits
  - In Ethernet frame sender has received PFC Pause
- **RxWait increments when frame receiver has prevented reception of frames**
  - In FC frame receiver has withheld all credits (zero Rx credits)
  - In Ethernet frame receiver has transmitted PFC Pause
- Not all Ethernet ASICs support TxWait and RxWait

TxWait and RxWait in Ethernet PFC



RxWait indicates the amount of time a frame receiver is preventing frames from being sent to it.

TxWait indicates the amount of time a frame transmitter is prevented from transmitting frames.

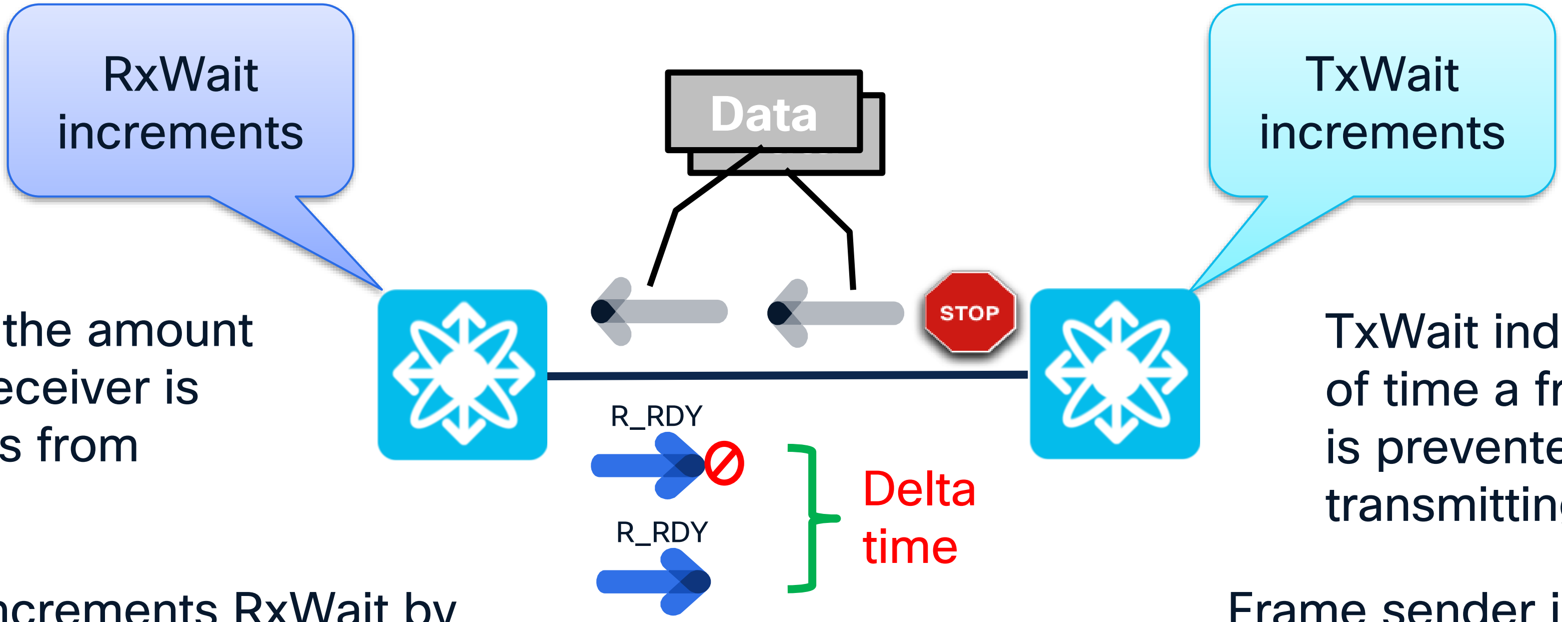
Pause sender increments RxWait by PFC delta time

Pause receiver increments TxWait by PFC delta time

RxWait increments when frame receiver has transmitted PFC Pause

TxWait increments when frame sender receives PFC Pause

TxWait and RxWait in Fibre Channel



RxWait indicates the amount of time a frame receiver is preventing frames from being sent to it.

TxWait indicates the amount of time a frame transmitter is prevented from transmitting frames.

Frame receiver increments RxWait by delta time between 0 Rx credits remaining and the next R_RDY

Frame sender increments TxWait by delta time between 0 Tx credits remaining and the next R_RDY

RxWait increments when frame receiver has withheld all credits

TxWait increments when frame sender runs out of credits

Congestion Types

Slow Drain and Over-Utilization

Two main types or causes of congestion

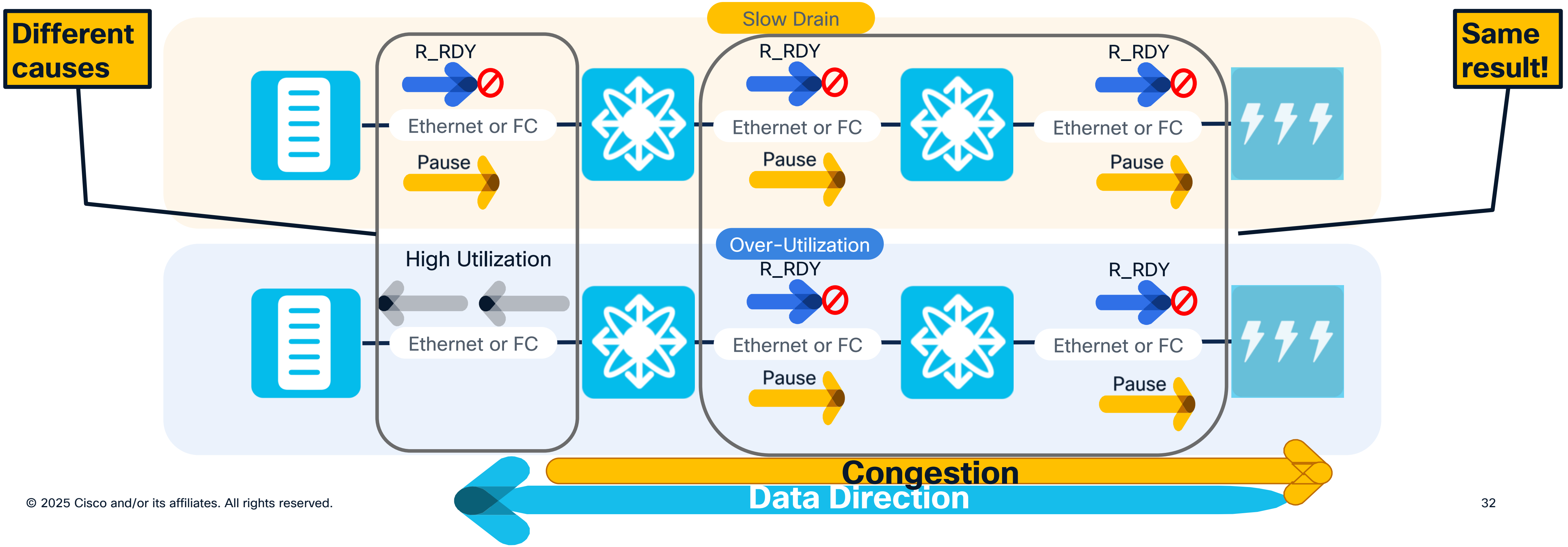
1. **Slow Drain**

- Frame receiver utilizes flow control to reduce the rate of incoming frames

2. **Over-utilization**

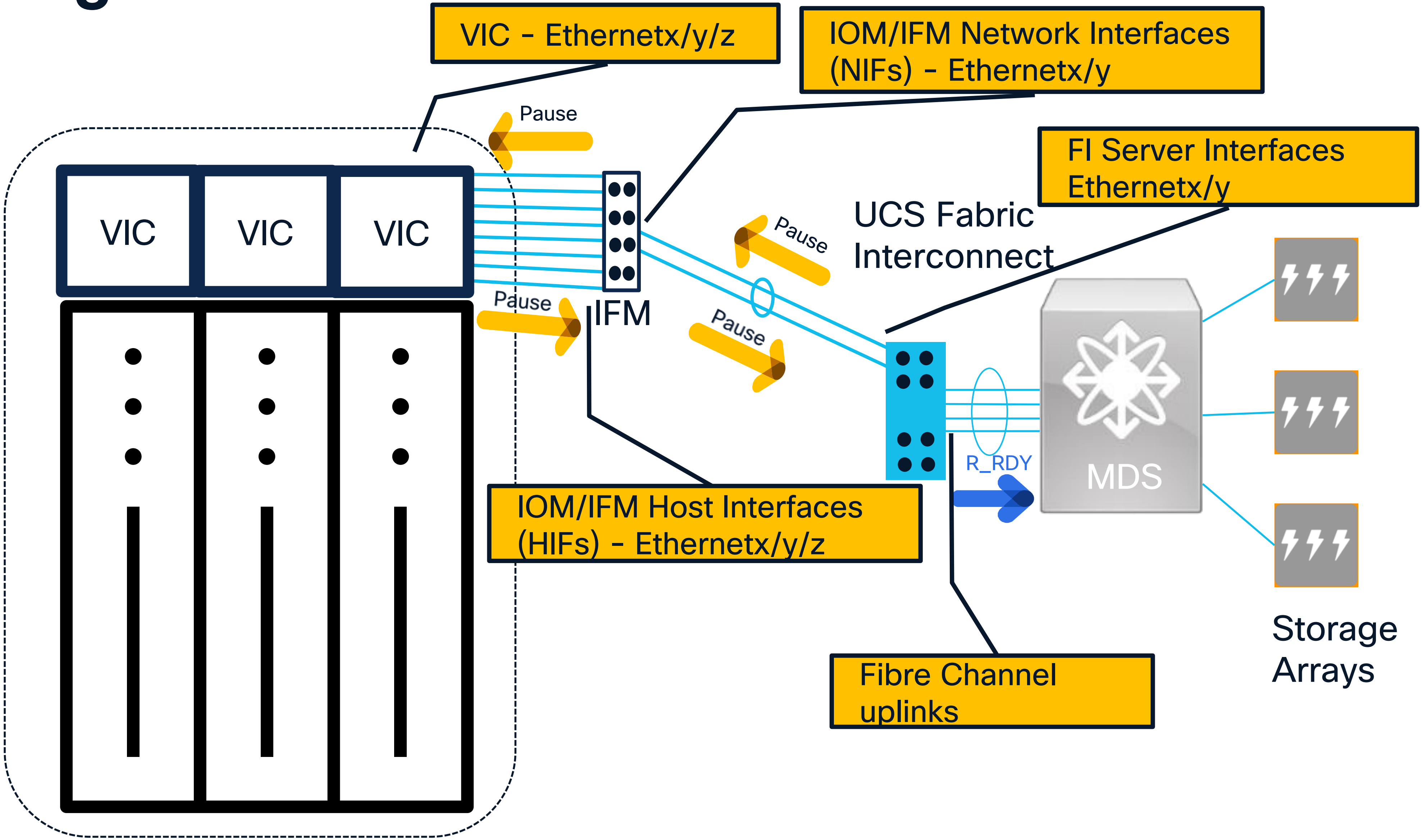
- Tx utilization at or near 100% with more data arriving than can be transmitted at link rate.
- Frame receiver is NOT reducing the rate of incoming frames via flow control

Under both conditions congestion spreads back to the frame senders



UCS Congestion Points

UCS Congestion Points



UCS Congestion Points

Four possible congestion points in a UCS

1. FC Uplinks
2. Network interfaces (NIFs/FI Server interfaces)
3. Host Interfaces (HIFs)
4. VICs

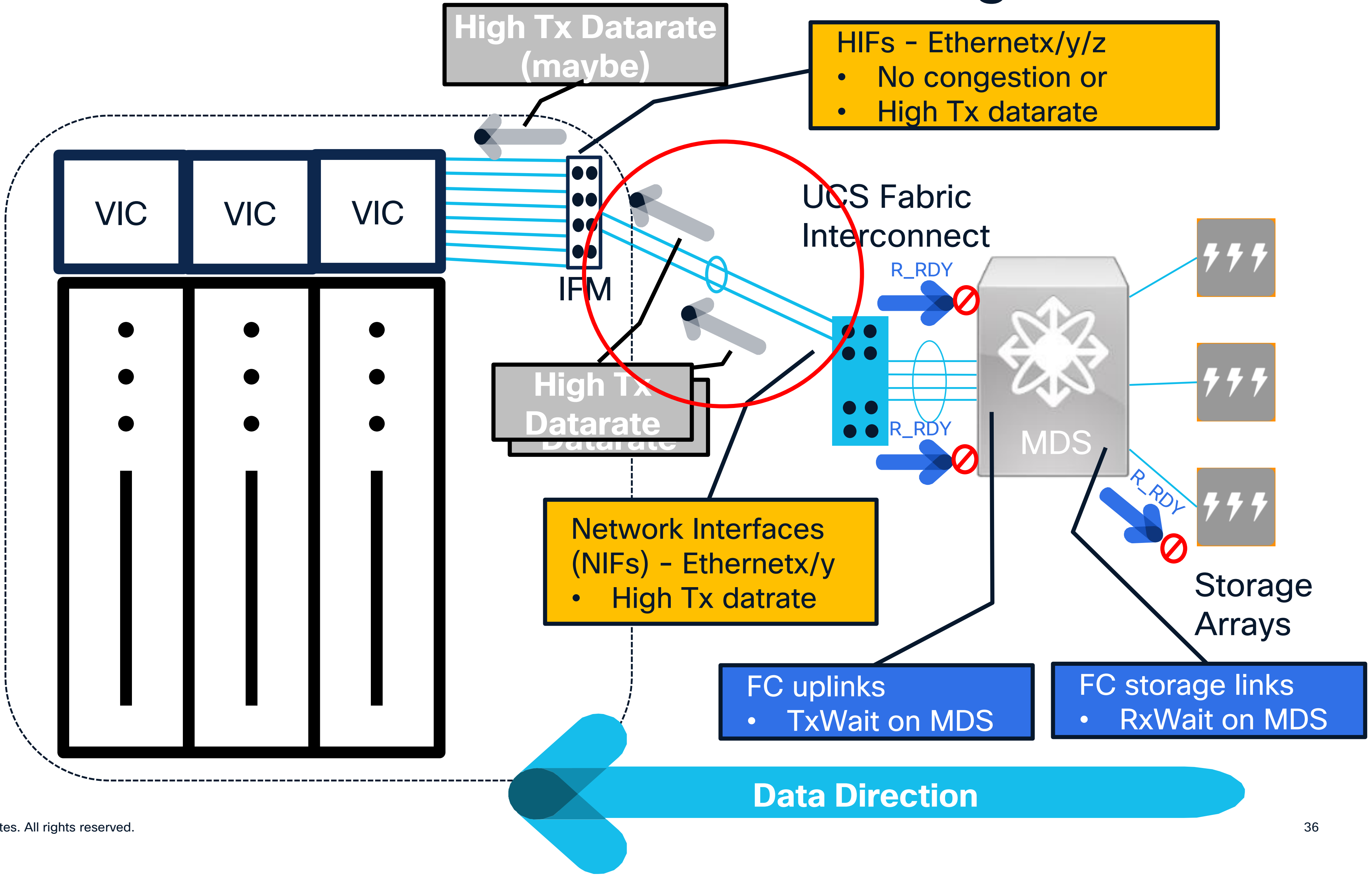
- Two main reasons

1. Over-utilization – More data being received for the link than it can handle

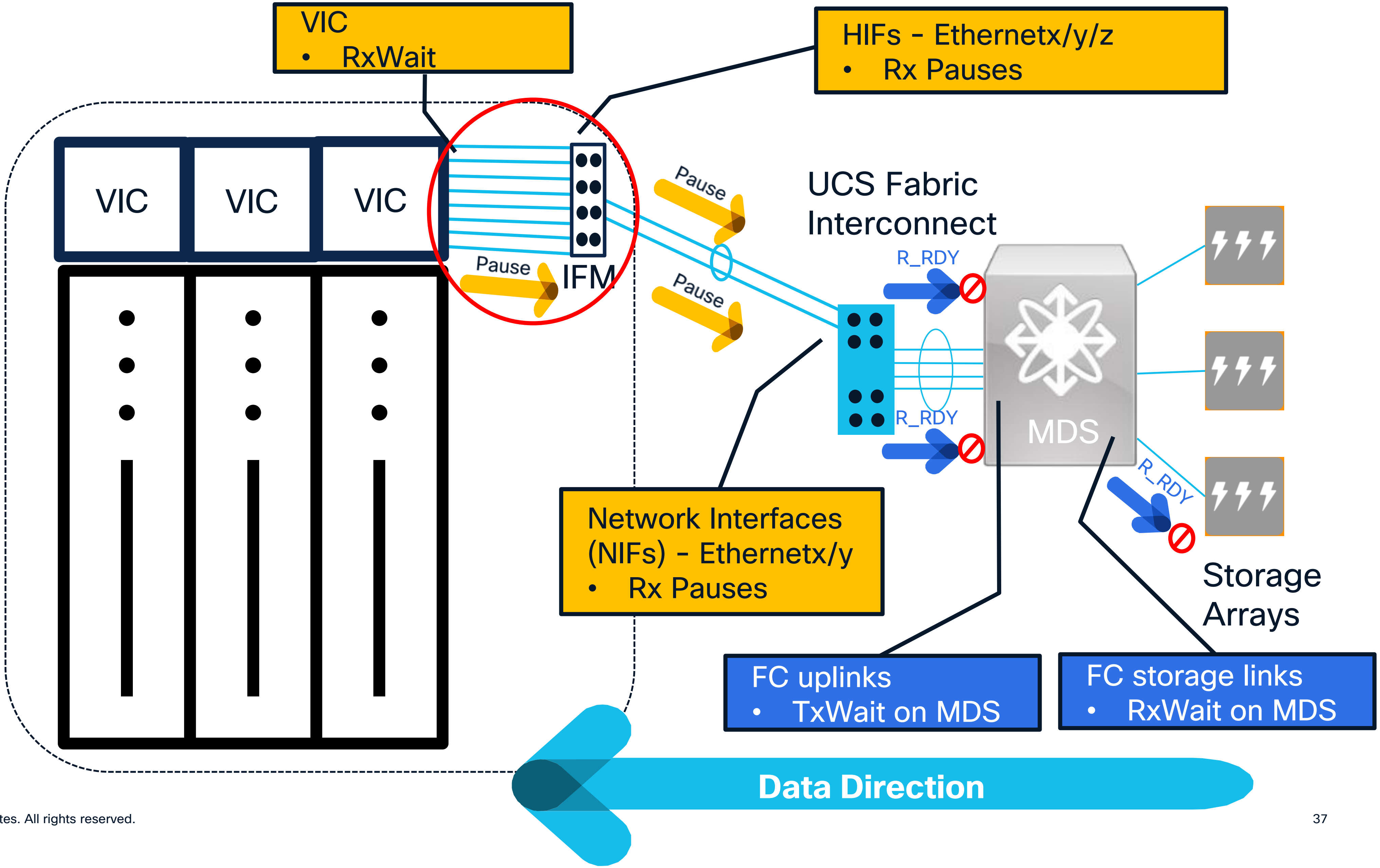
   1a – This includes speed mismatches

2. Slow Drain – Flow control (Either PFC or B2B) slowing down link.

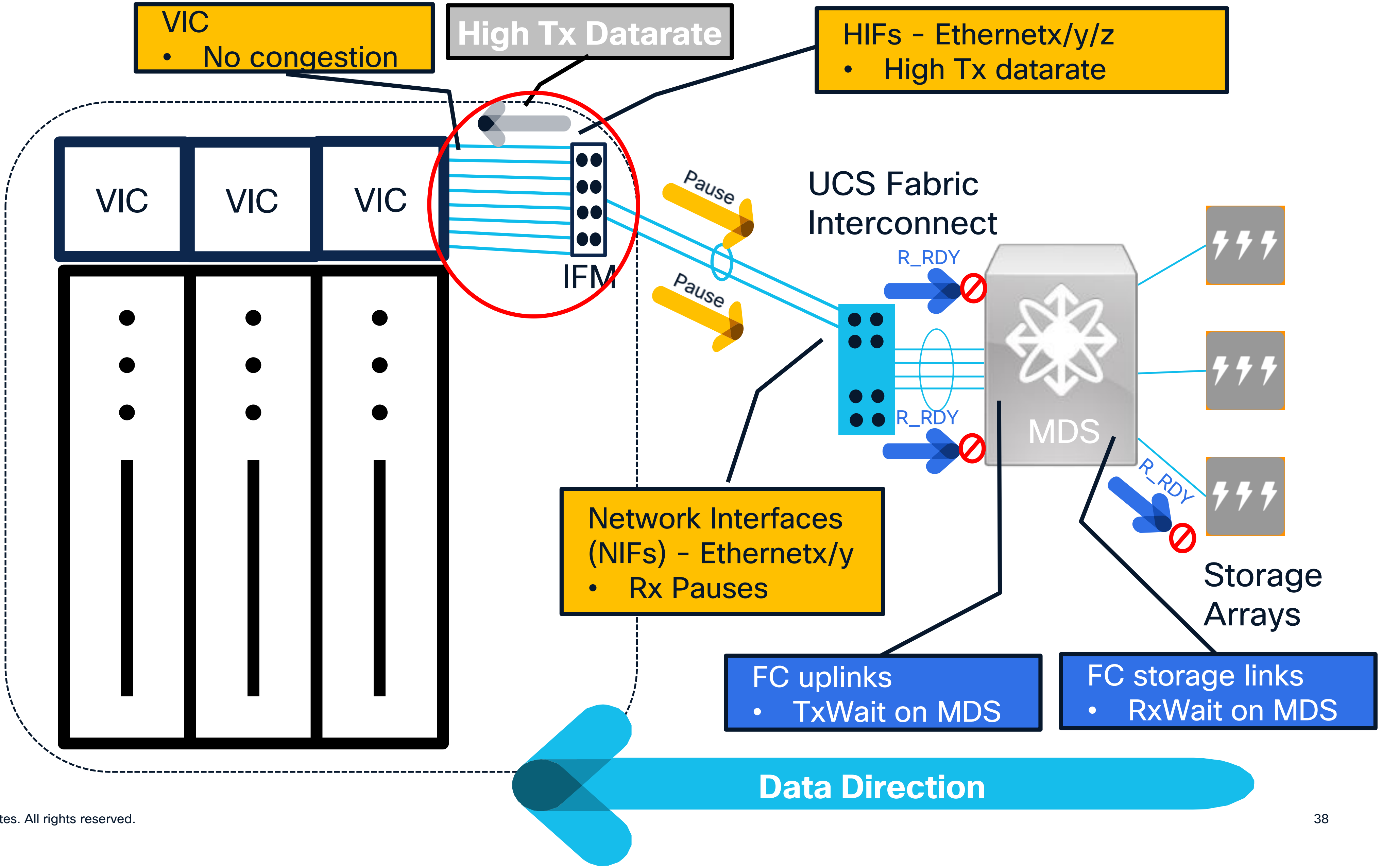
Scenario 1 – FI to NIF Over-Utilization Congestion



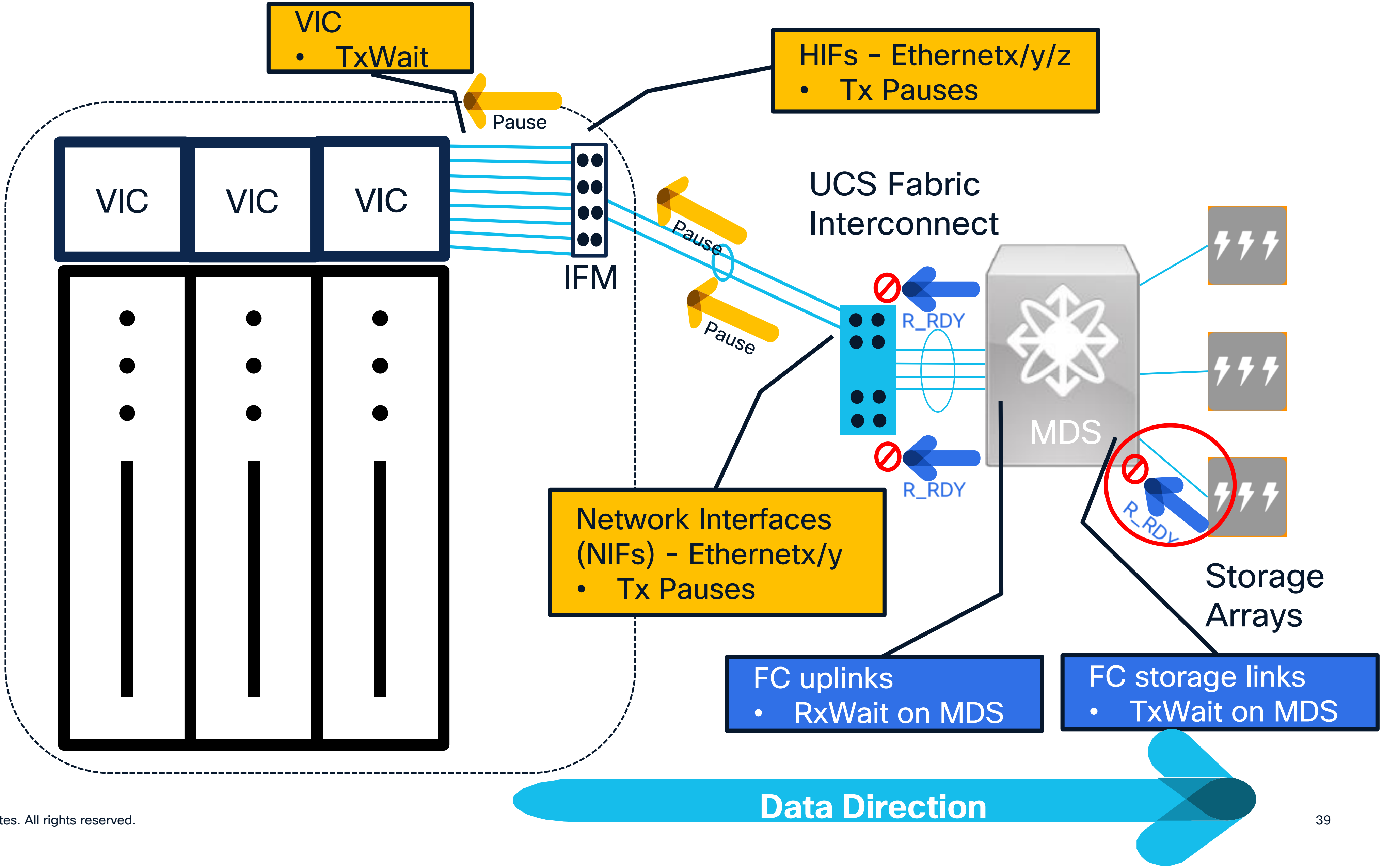
Scenario 2 – VIC to HIF Slow Drain Congestion



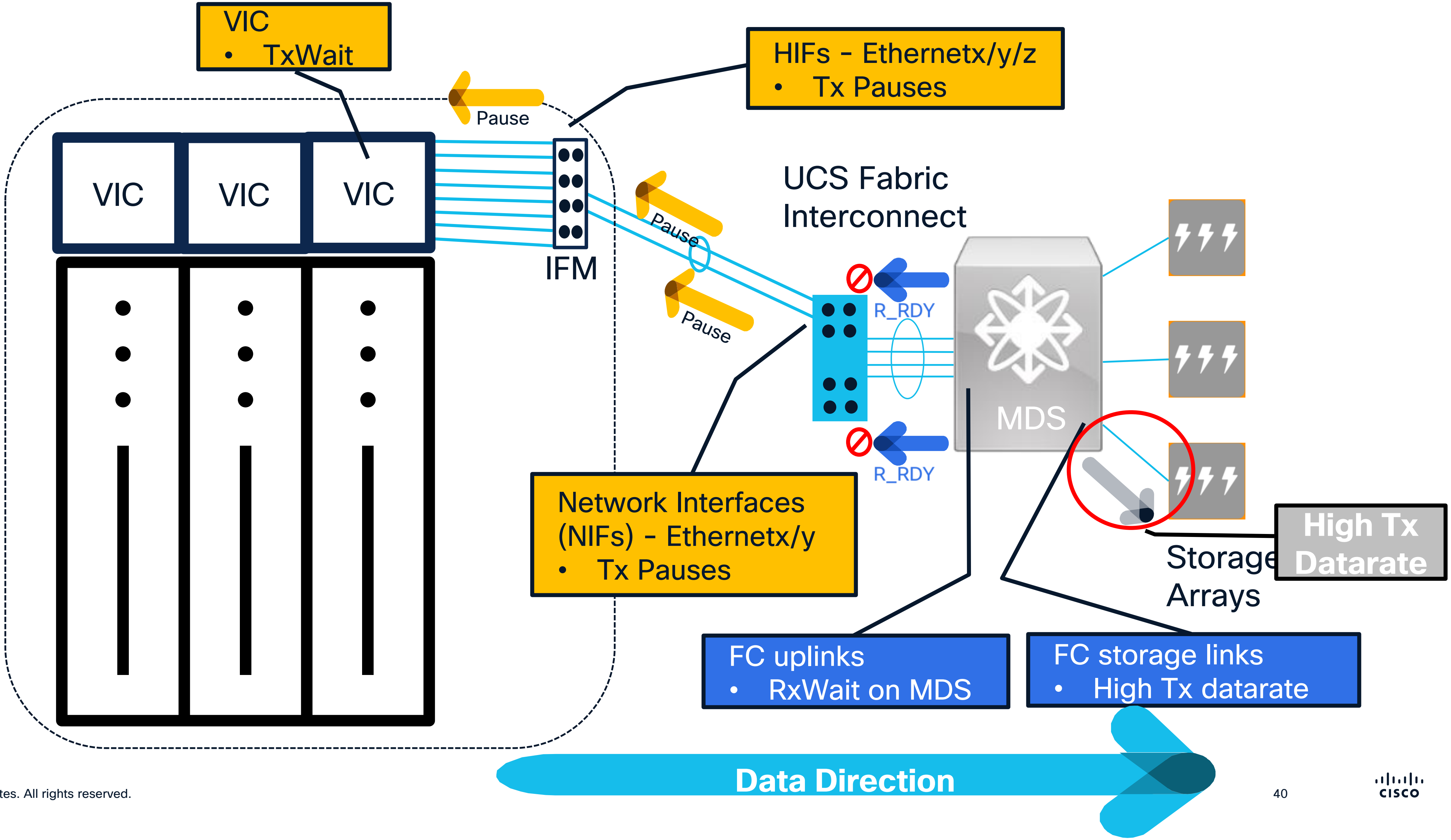
Scenario 3 – HIF to VIC Over-Utilization Congestion



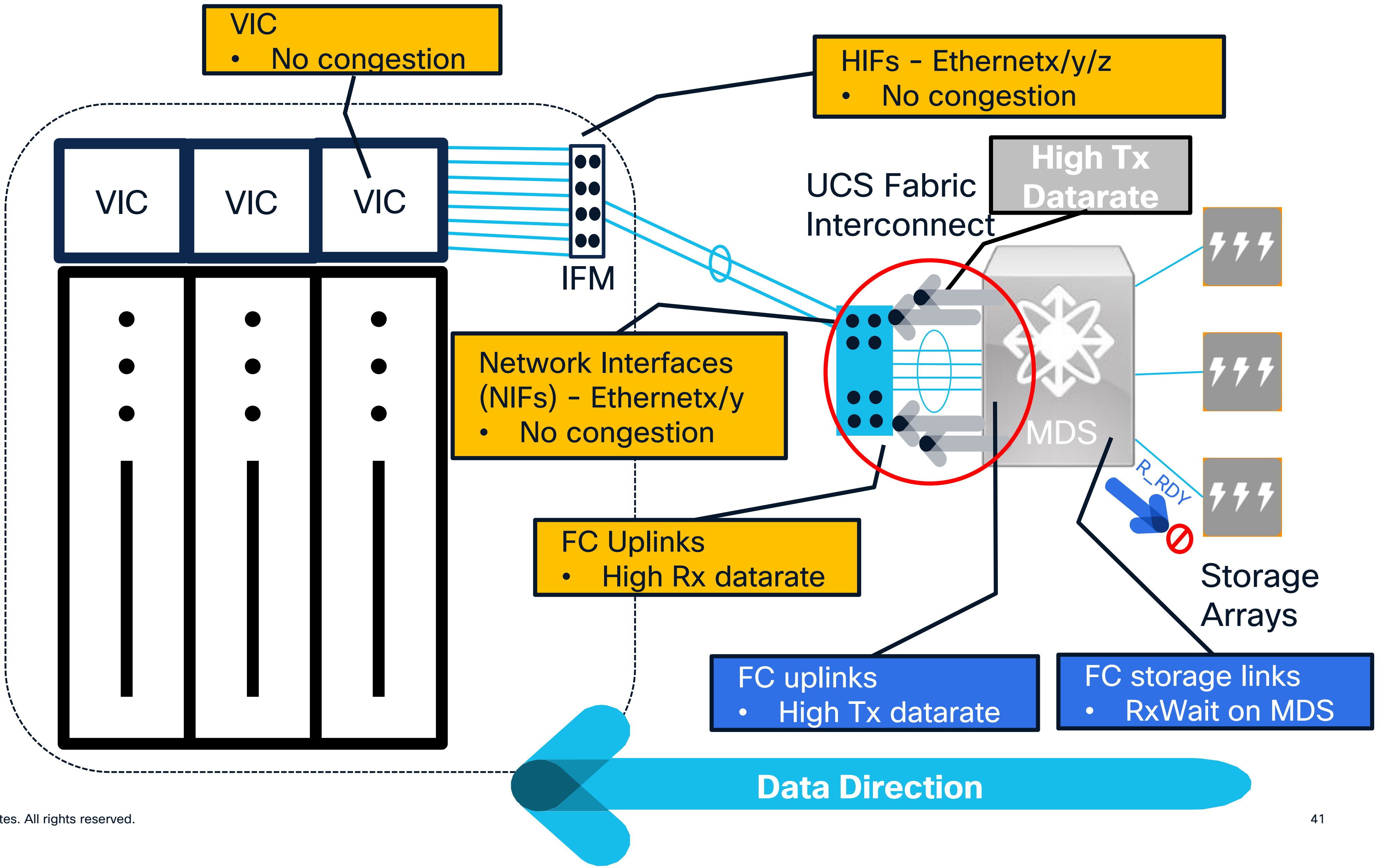
Scenario 4 – Array Slow Drain Congestion



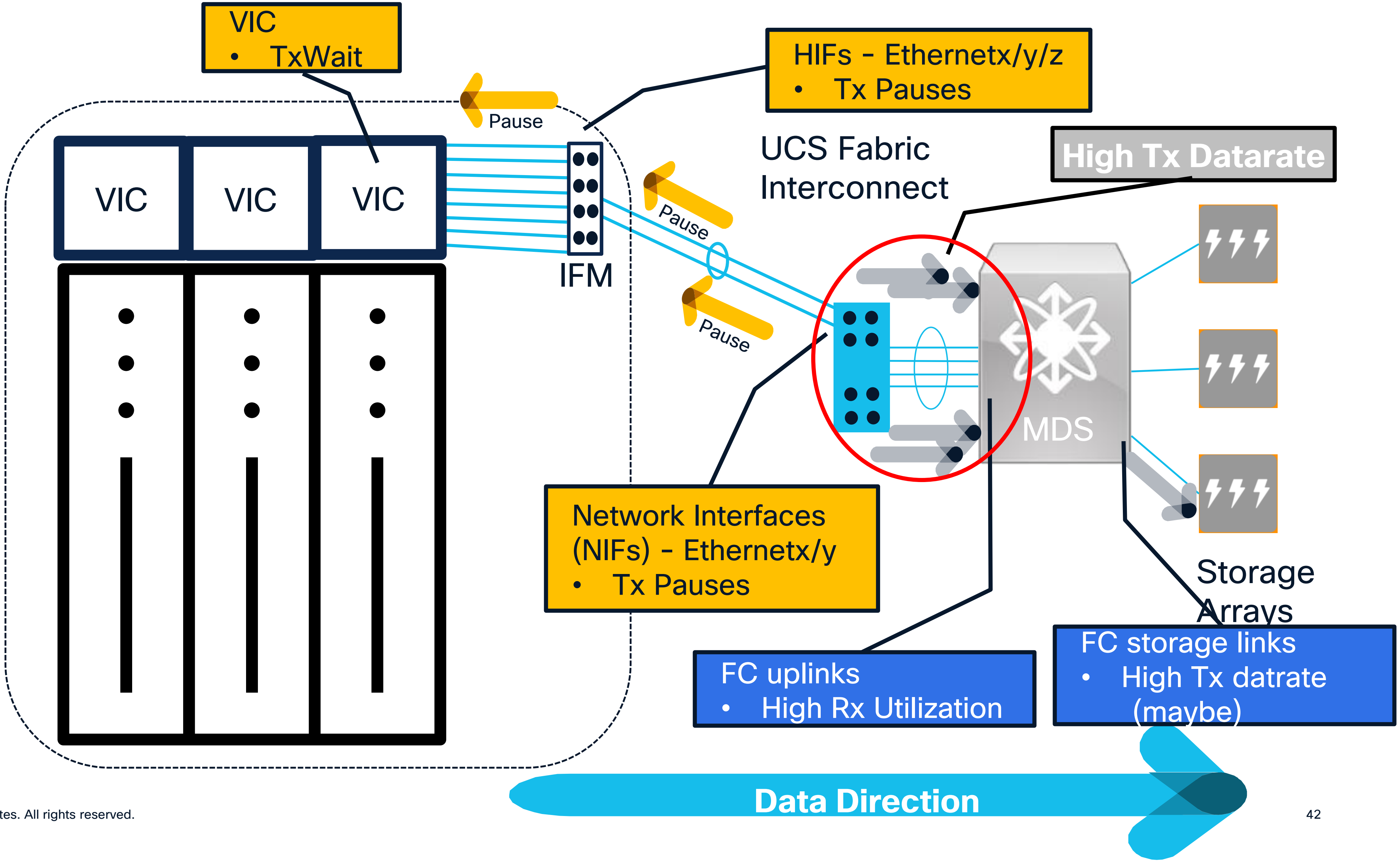
Scenario 5 – Array Tx Over-Utilization Congestion



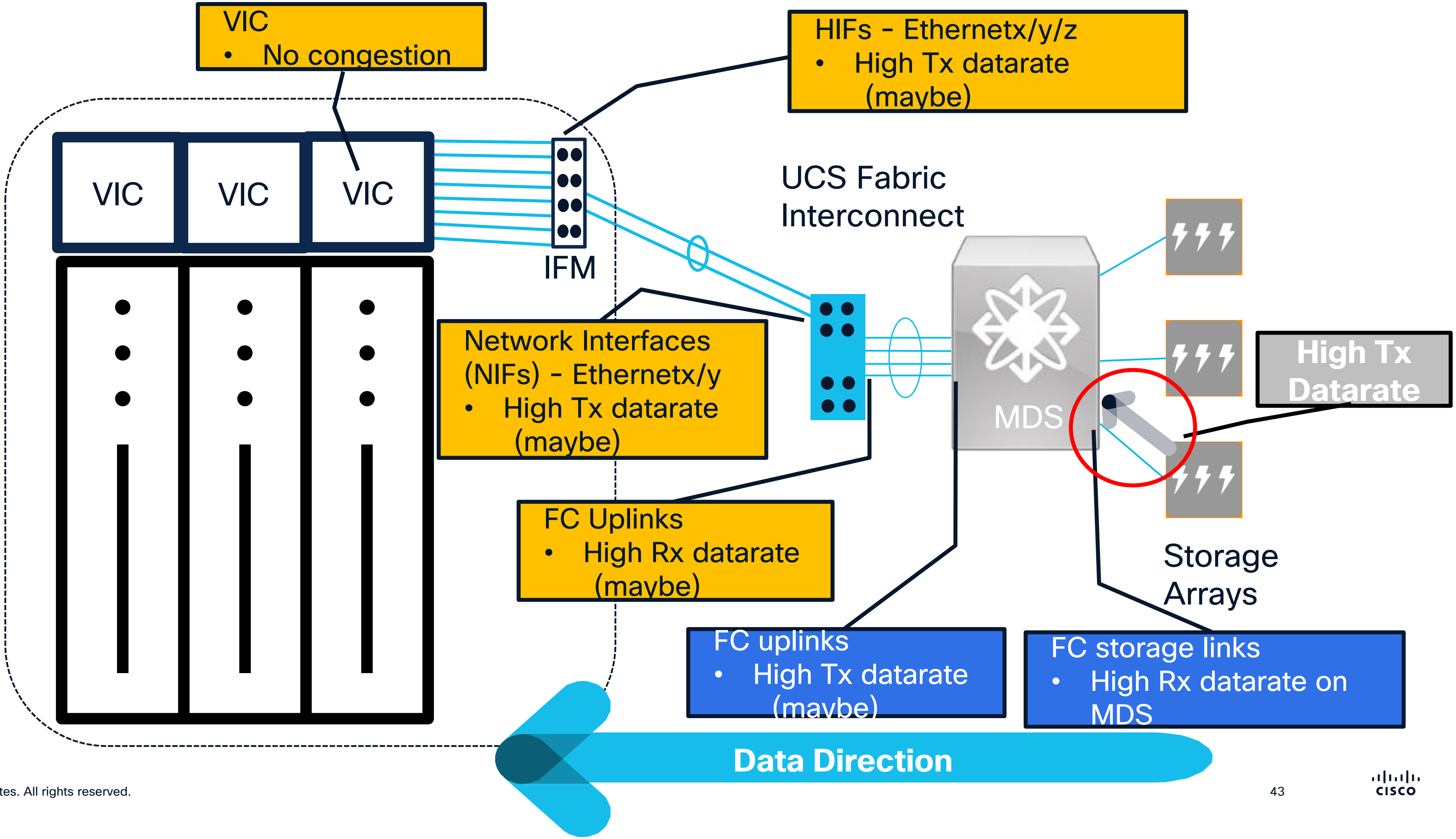
Scenario 6 - FC Uplink Over-Utilization Congestion



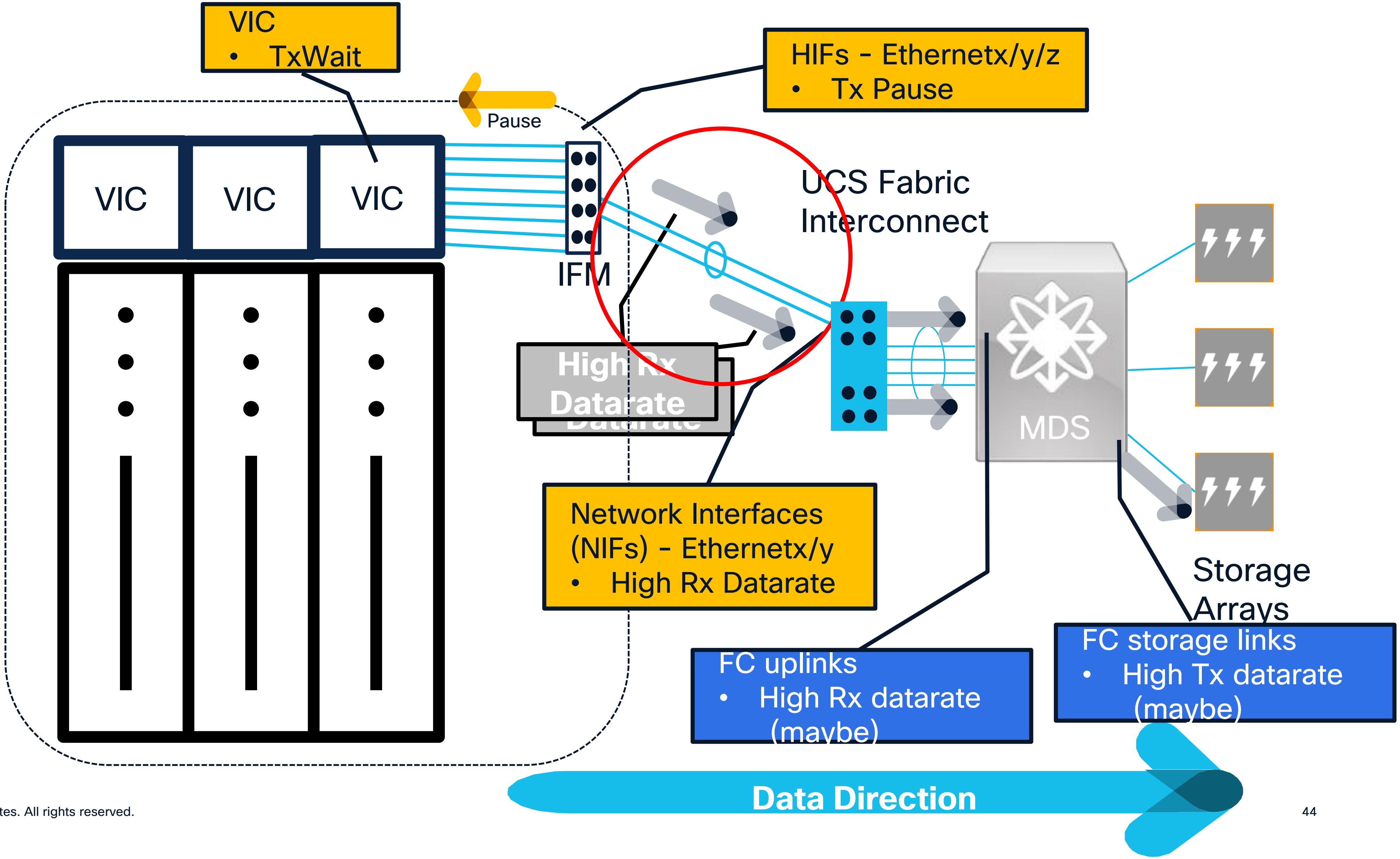
Scenario 7 - FC Uplink Over-Utilization Congestion



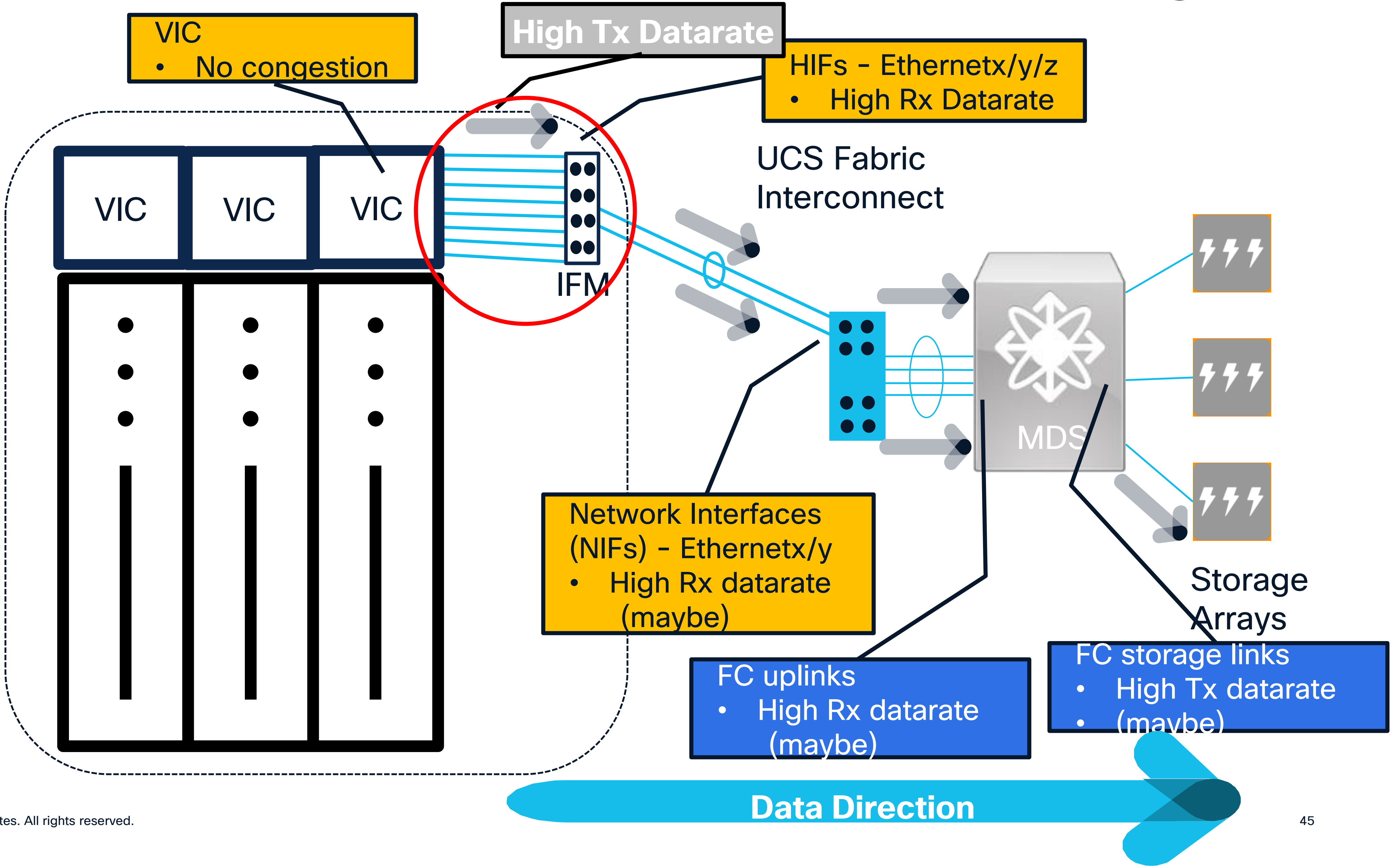
Scenario 8 - Array Rx Over-Utilization Congestion



Scenario 9 - NIF to FI Over-Utilization Congestion



Scenario 10 – VIC to HIF link Over-Utilization Congestion



Troubleshooting Workflows

Culprits and Victims

New terminology to describe devices causing problems and those affected

Culprits

Those devices **causing** congestion

Victims

Those devices **affected** by the congestion

Three types

1. **Direct** – Devices zoned with the culprit
2. **Indirect** – Devices zoned with the “direct: **or** “same-path” victims
3. **Same-path** – Devices *transmitting* over the congested network path

Understanding culprits and victims explains the scope of the congestion

See the following session for more details:
BRKDCN-2942 - SAN Congestion Management and SAN Analytics – Thursday June 12th 10:30 AM

“Normal” congestion

Due to different link speeds between FC and Ethernet there can be normal PFC pauses

Ethernet has speeds: 10G, 25G, 40G, 100G etc.

Fibre Channel has speeds: 8G, 16G, 32G and 64G

As an FCoE exchange traverses different link speeds bottlenecks may be encountered

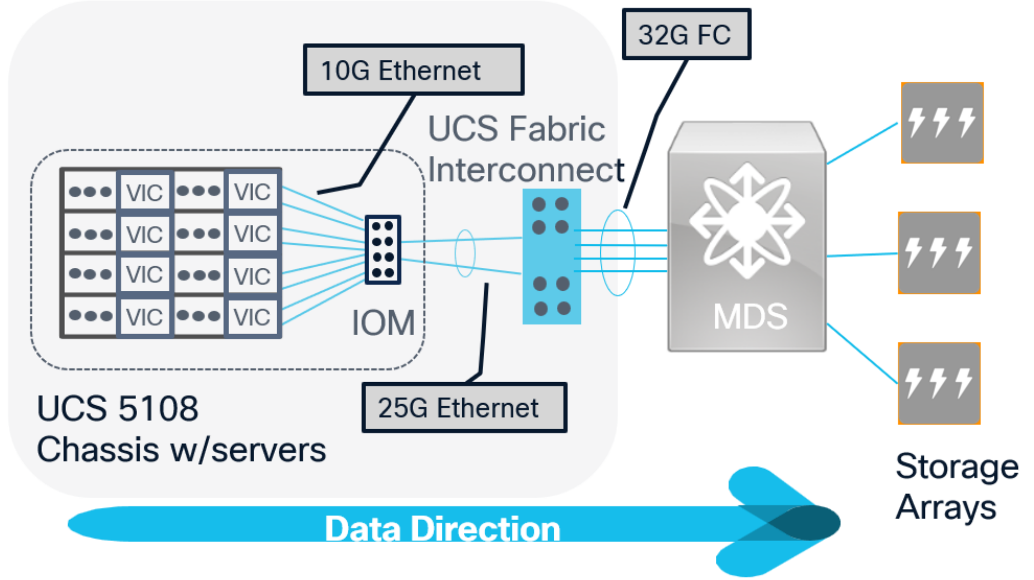
FCoE exchanges are mapped to individual links to ensure in-order delivery

This will result in PFC Pauses being sent to equalize transmission rates

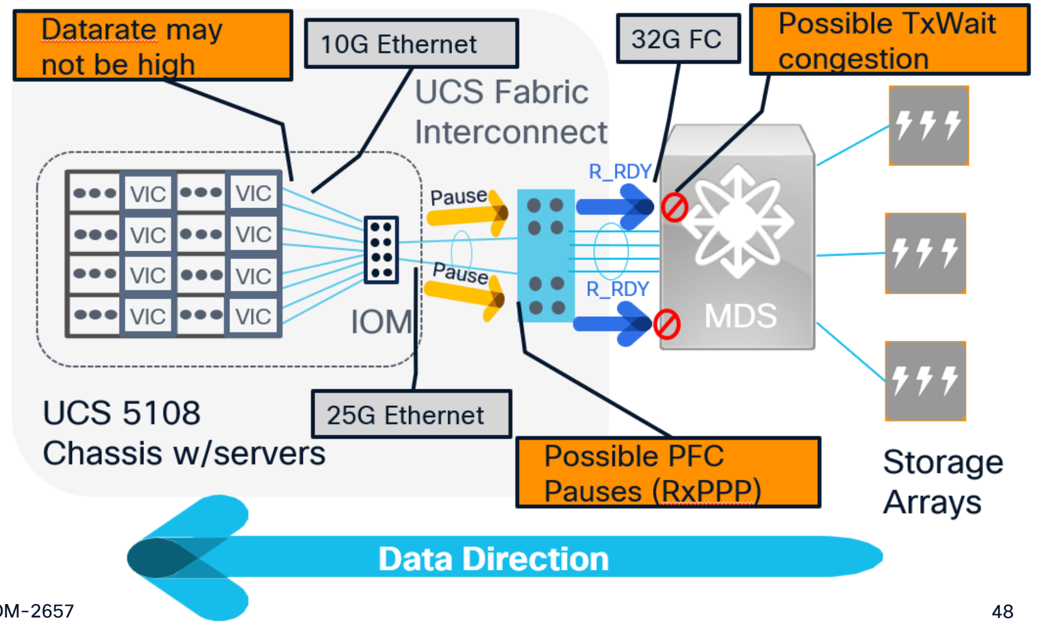
These can be minimized but cannot be completely avoided!

PFC Pauses are a normal occurrence and will not normally be zero

No bottlenecks



Bottlenecks!



UCS Congestion Troubleshooting Methodology

- SCSI/NVMe Read commands generate data towards the servers/VICs
- SCSI/NVMe Write commands generate data away from the servers/VICs

Follow Tx congestion to the source to find the culprit

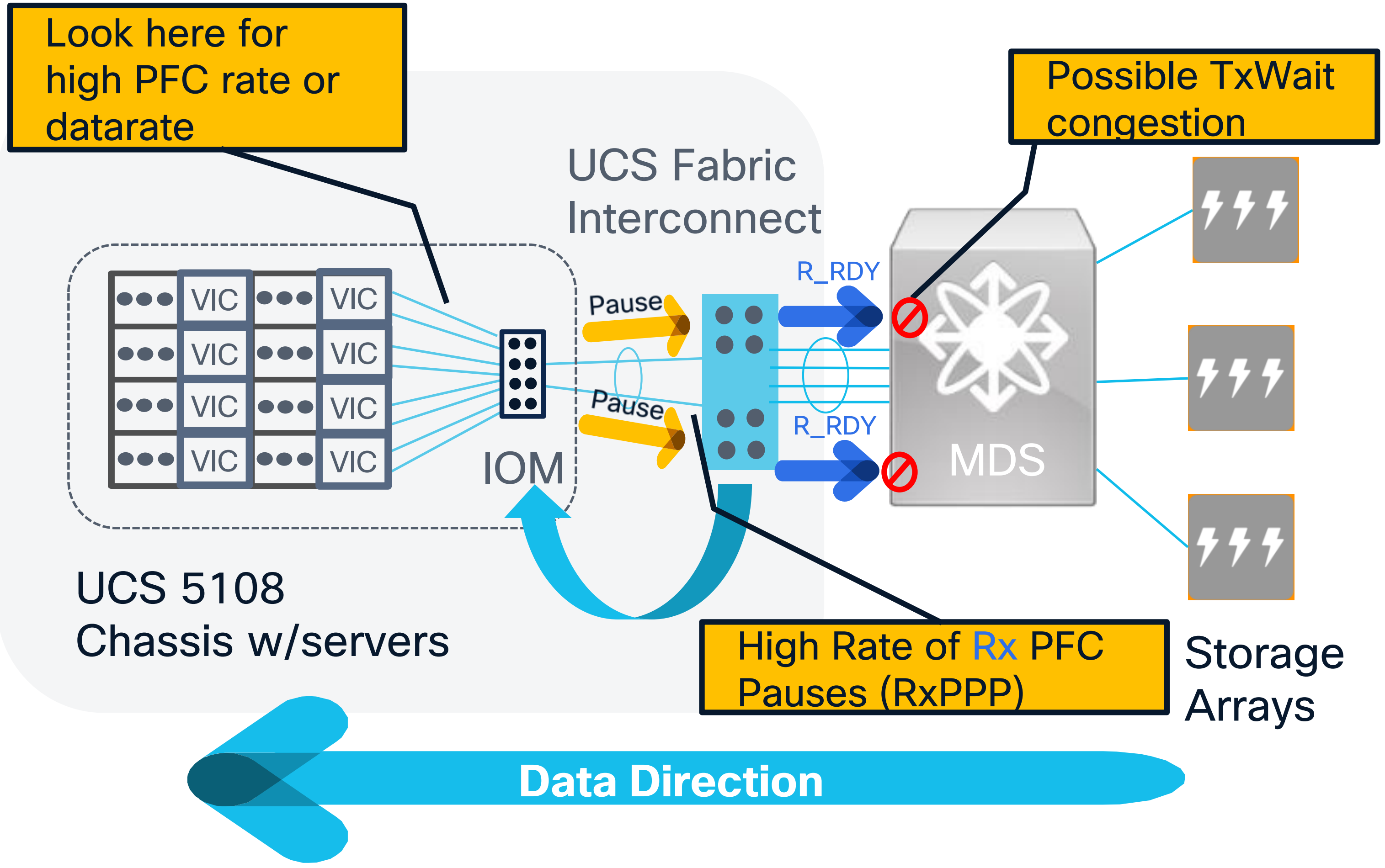
- For Tx congestion (e.g. high Rx PFC rate)
  - Check next hop where the PFCs Pauses are being received from

Follow Rx congestion to the source of the data to find the victim(s)

- For Rx congestion (e.g. high Tx PFC rate)
  - Check next hop where data is being received from

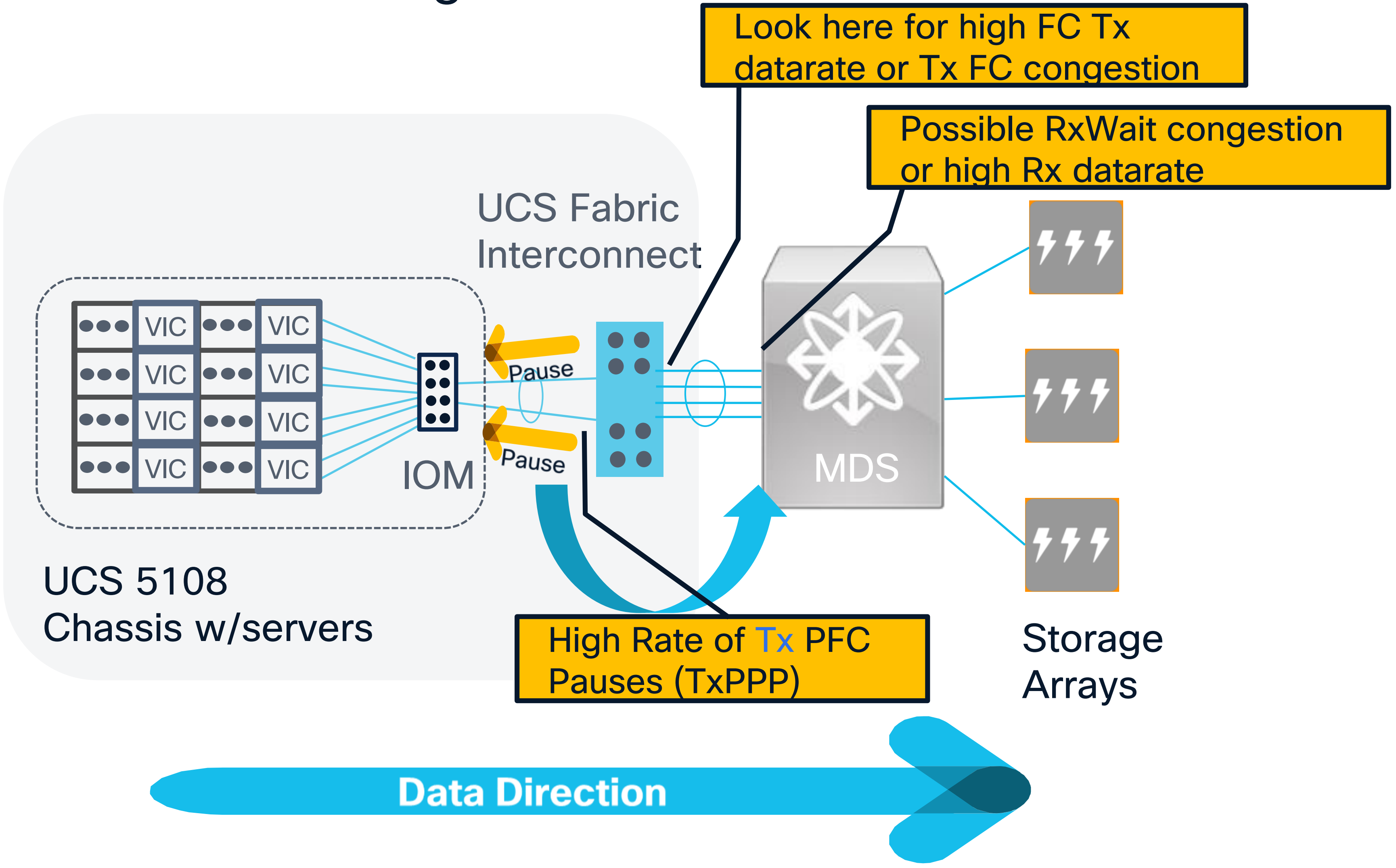
Tx Congestion

Check next hop where the PFCs Pauses are being received from



Rx Congestion

Check next hop where data is being received from



Troubleshooting Workflows

Where to start?

Recommendation is to start in the MDS fabric and look for congestion

- MDS has excellent time and date stamped congestion logs

4 possibilities

- Upstream MDS is reporting TxWait congestion on FC uplinks to FI
  - Scenarios 1, 2 and 3
- Upstream MDS is reporting RxWait congestion on FC uplinks to FI
  - Scenarios 4 and 5
- Upstream MDS is reporting high Tx or Rx utilization
  - Scenarios 6 and 7
- Upstream MDS is **not** reporting any congestion
  - Scenarios 8, 9 and 10

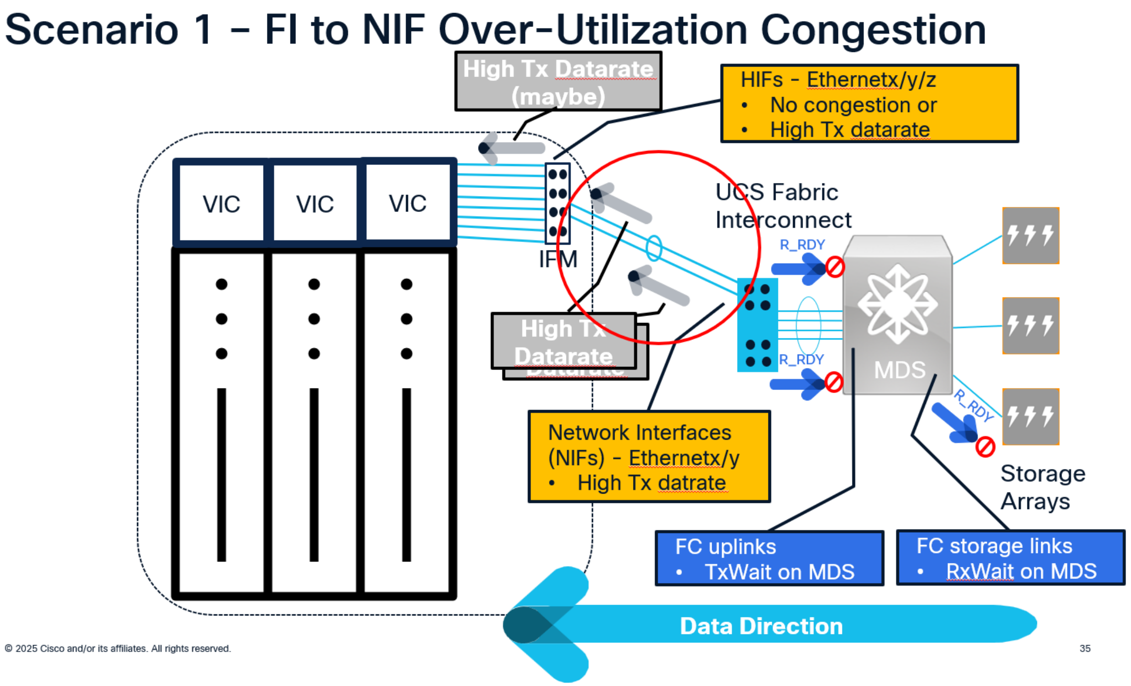
Troubleshooting Workflow 1

If upstream MDS is reporting TxWait congestion on FC uplinks to UCS FI

Congestion is within the UCS domain

Step 1

- Check Tx Over-utilization on Ethernetx/y links to IOM/IFM – Scenario 1
  - If found increase capacity on those links
- Check PFC Pause received(RxPPP) from IOM/IFM on Ethernetx/y links
  - If RxPPP found go to Step 2

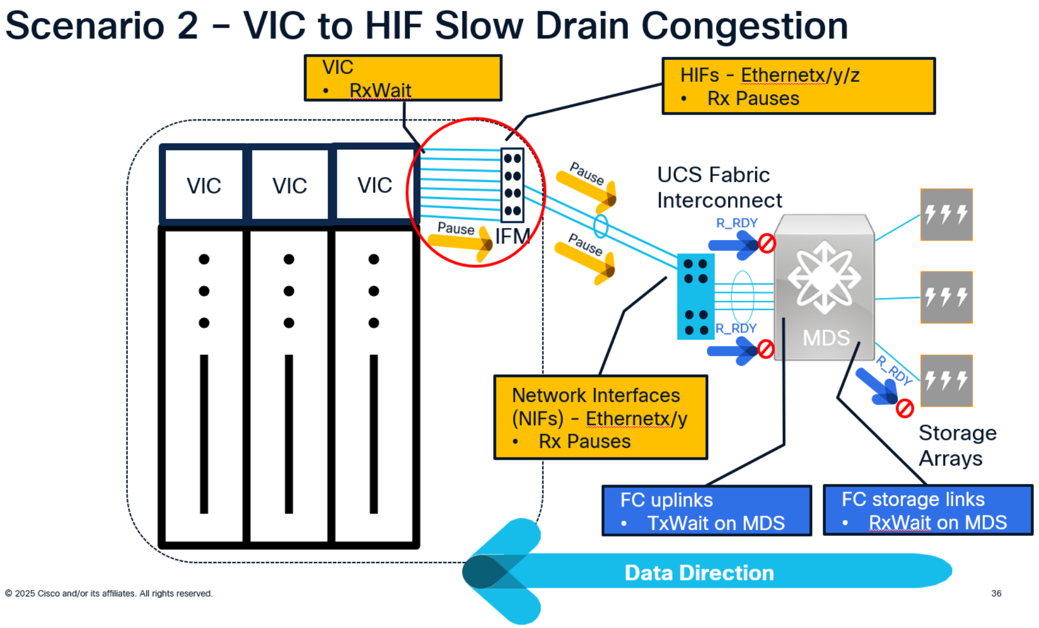


Troubleshooting Workflow 1 – continued

If upstream MDS is reporting TxWait congestion on FC uplinks to UCS FI

Step 2

- Check PFC Pause received (RxPPP) from VIC on Ethernetx/y/z links
  - If found check RxWait on VIC on Ethernetx/y/z link – Scenario 2
    - If significant RxWait found investigate individual server for internal bottlenecks
  - If no PFC Pause received (RxPPP) or no RxWait found go to Step 3

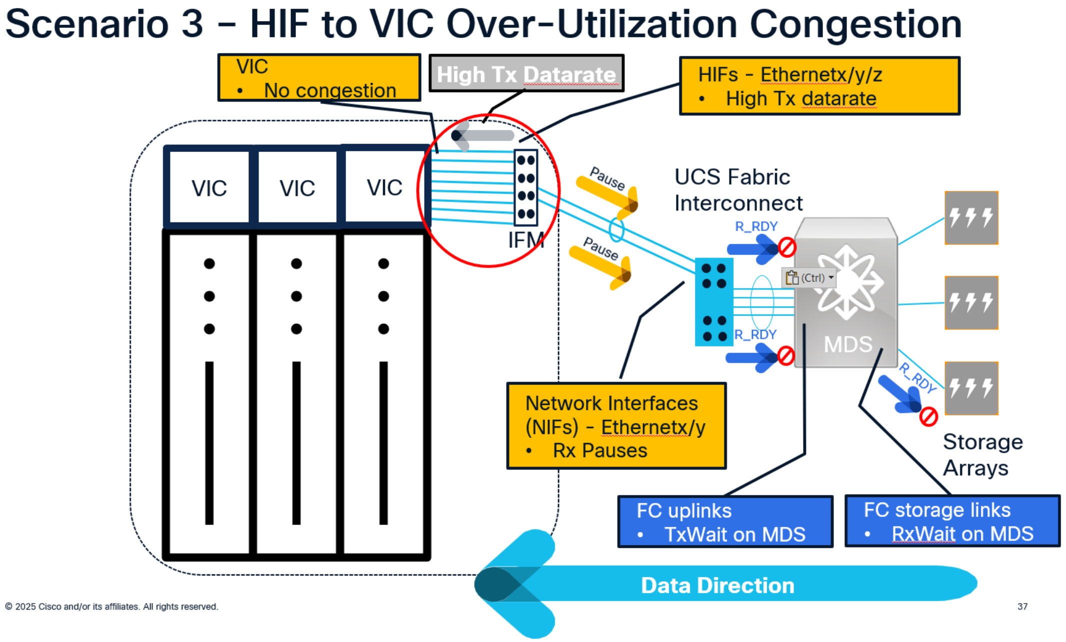


Troubleshooting Workflow 1 - continued

If upstream MDS is reporting TxWait congestion on FC uplinks to UCS FI

Step 3

- Check Tx Over-utilization on Ethernetx/y/z links - Scenario 3
  - If found increase capacity from IOM/IFM to VIC
    - Port expander
    - Updated IOM/IFM and VIC



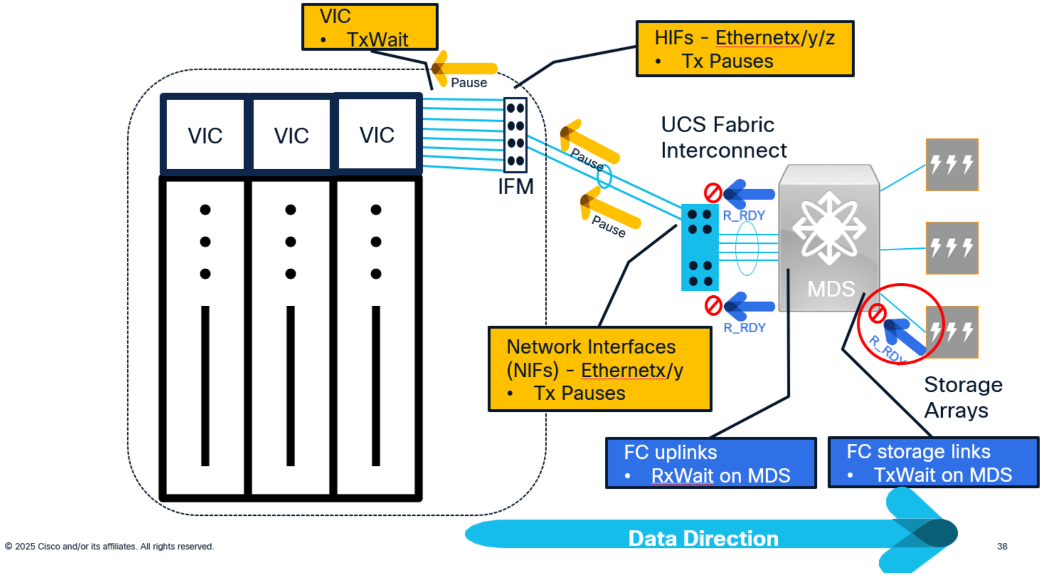
Troubleshooting Workflow 2

If upstream MDS is reporting RxWait congestion on FC uplinks to UCS FI

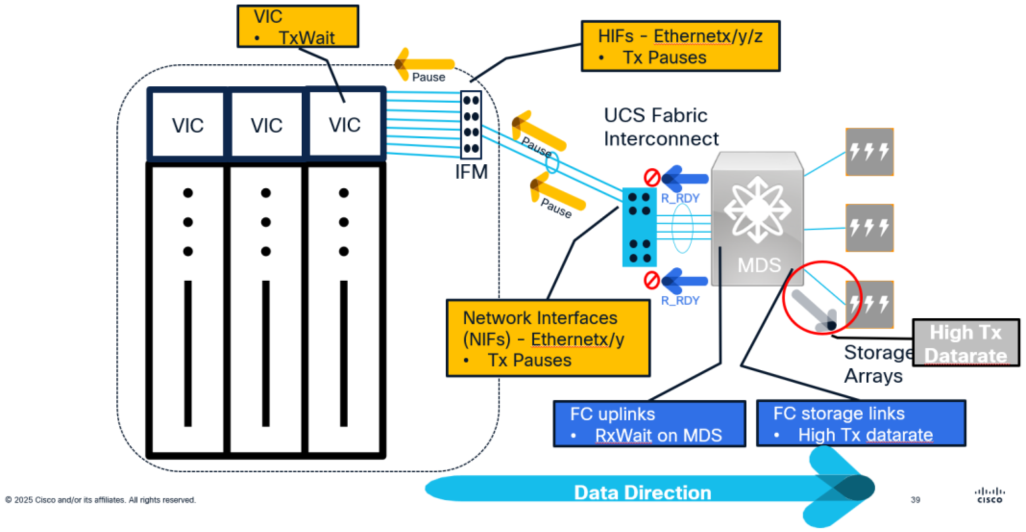
Congestion is within MDS SAN

- Indications in UCS include
  - PFC Pause transmitted(TxPPP) from FI on Ethernetx/y links
  - PFC Pause transmitted (TxPPP) from IOM/IFM on Ethernetx/y/z links
  - Scenarios 4 and 5
  - Troubleshoot FC congestion issue (See BRKDCN-2942)

Scenario 4 – Array Slow Drain Congestion



Scenario 5 – Array Tx Over-Utilization Congestion



Troubleshooting Workflow 3

If UCS FI or upstream MDS is reporting high datarate on FC uplinks

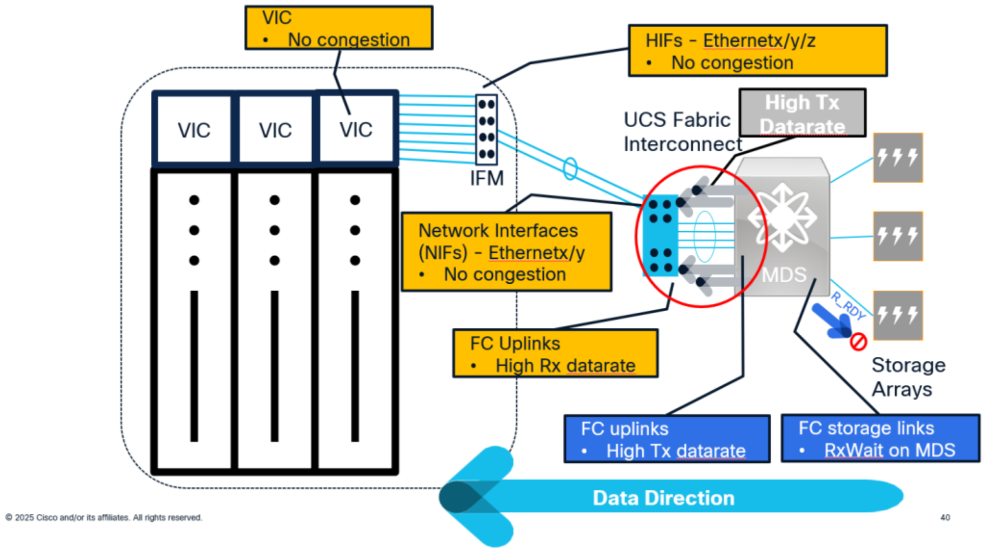
Congestion is due to insufficient FC bandwidth

Scenarios 6 & 7

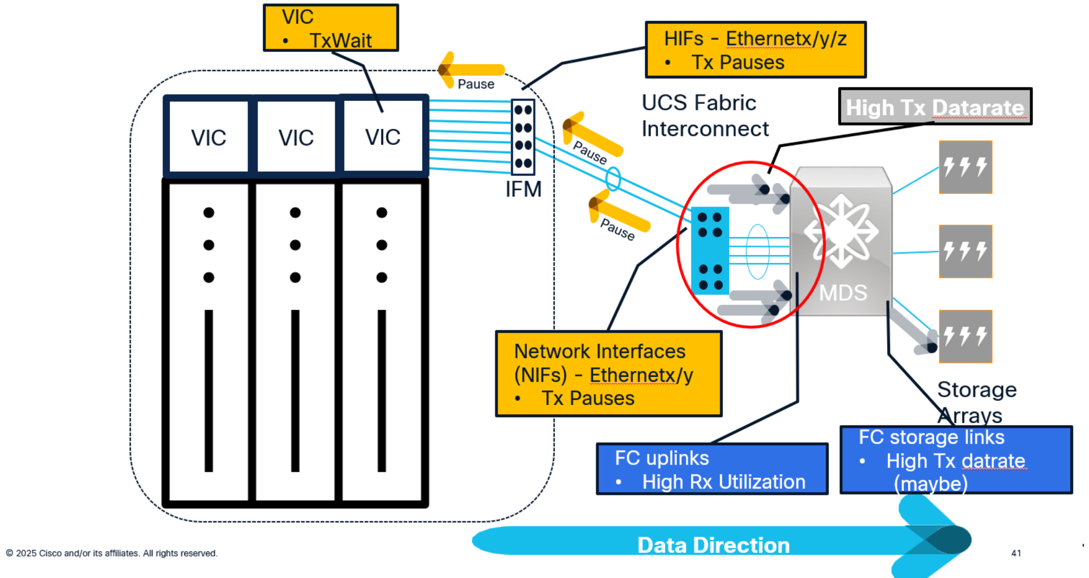
Increase FC link bandwidth via:

- Higher speed links
- More links into port-channels
- Consider splitting port-channel into 2-4 port-channels

Scenario 6 – FC Uplink Over-Utilization Congestion



Scenario 7 – FC Uplink Over-Utilization Congestion



Troubleshooting Workflow 4

If upstream MDS is not reporting any congestion on FC uplinks to UCS FI

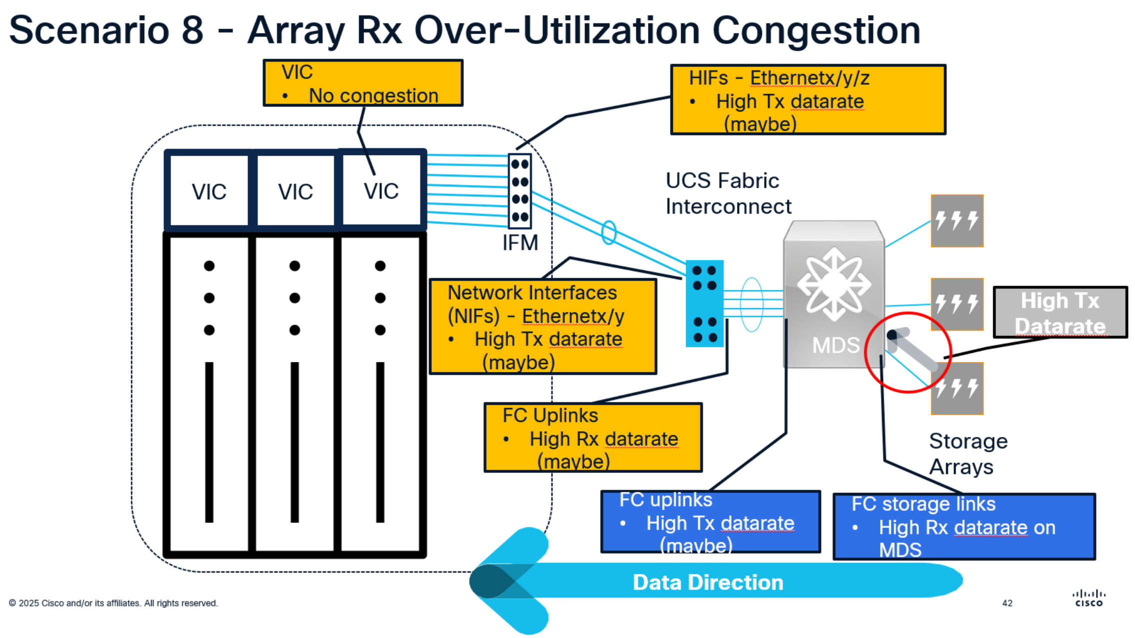
Congestion could be within the UCS domain
Congestion could be within MDS SAN

Check links for high Tx/Rx datarate

Scenarios 8, 9 and 10

Step 1

- Check FC links for high utilization
- Scenario 5
- If found increase capacity on links



Troubleshooting Workflow 4 - continued

If upstream MDS is not reporting any congestion on FC uplinks to UCS FI

Step 2

Check high Rx utilization on IOM/IFM to FI on Ethernetx/y links (Rx from FI's perspective)

Scenario 9

- If found there also should be PFC Pause transmitted (TxPPP) from IOM/IFM to VICs on Ethernetx/y/z links
- If found increase bandwidth on IOM/IFM to FI links

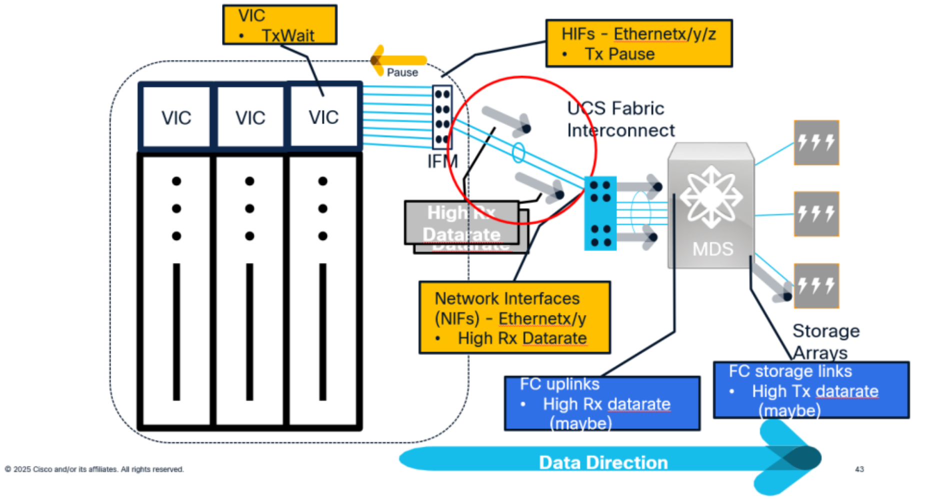
Step 3

Check high Rx utilization on VIC to IOM/IFM Ethernetx/y/z links (Rx from IOM/IFM's perspective)

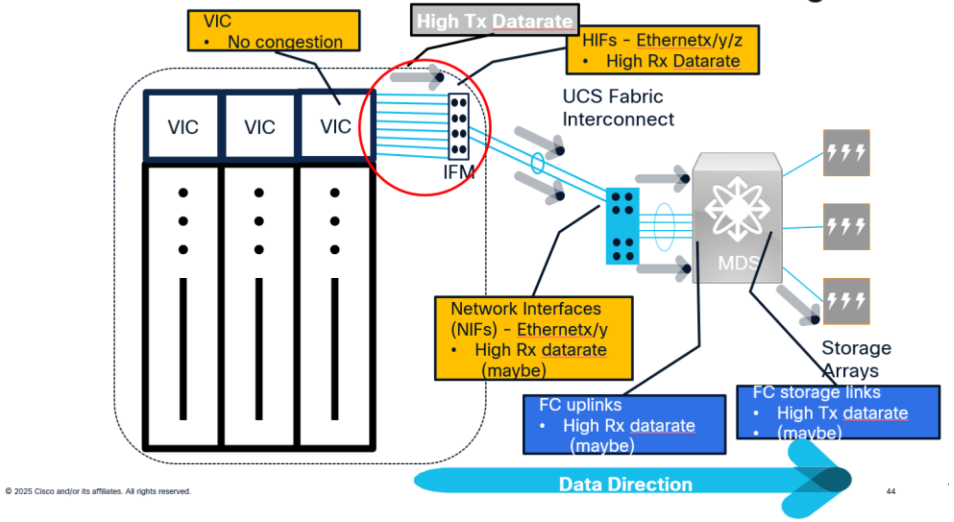
Scenario 10

- If found increase bandwidth on VIC to IOM/IFM links

Scenario 9 - NIF to FI Over-Utilization Congestion



Scenario 10 - VIC to HIF link Over-Utilization Congestion



UCS Congestion Commands

show interface fc

show interface contains only basic FC info.

```
fc1/1 is trunking
    Hardware is Fibre Channel, SFP is short wave laser w/o OFC (SN)
    Port WWN is 20:01:00:3a:9c:f2:83:d8
    Admin port mode is NP, trunk mode is on
    snmp link state traps are enabled
    Port mode is TNP
    Port vsan is 1
    Operating Speed is 32 Gbps
    Admin Speed is 32 Gbps
    Transmit B2B Credit is 32
    Receive B2B Credit is 64
    Receive data field Size is 2112
    Beacon is turned off
    fec state is enabled by default
    Belongs to san-port-channel11
    Trunk vsans (admin allowed and active) (1,3511)
    Trunk vsans (up)                       (1,3511)
    Trunk vsans (isolated)                 ()
    Trunk vsans (initializing)             ()
    5 minutes input rate 1745371760 bits/sec,218171470 bytes/sec, 110101 frames/sec
    5 minutes output rate 177147816 bits/sec,22143477 bytes/sec, 14531 frames/sec
      1280970913463 frames input,2428652799045228 bytes
        0 discards,0 errors
        0 invalid CRC/FCS,0 unknown class
        0 too long,0 too short
      473576843374 frames output,821659884226636 bytes
        0 discards,0 errors
      2 input OLS,35  LRR,2 NOS,0 loop inits
      9 output OLS,4 LRR, 2 NOS, 0 loop inits
    Receive B2B Credit performance buffers is 0
      32 transmit B2B credit remaining
      0 low priority transmit B2B credit remaining
    Last clearing of "show interface" counters  :never
```

Tx/Rx Credits agreed to

Tx Credits remaining (real time)

show interface priority-flow-control

show interface priority-flow-control displays the total PFC pauses from the FI's perspective

Ethernetx/y links are links to IOM/IFMs

```
show interface priority-flow-control

slot  1
=======

============================================================
Port                 Mode Oper(VL bmap)  RxPPP      TxPPP
============================================================

...
Ethernet1/17         On   On   (8)       49230      0
Ethernet1/18         On   On   (8)       48042      0
Ethernet1/19         On   On   (8)       51126      0
Ethernet1/20         On   On   (8)       53434      0
Ethernet1/21         On   On   (8)       7991400    0
Ethernet1/22         On   On   (8)       8060310    0
Ethernet1/23         On   On   (8)       7851333    0
Ethernet1/24         On   On   (8)       7220608    0
```

- **Mode** – PFC mode – Should be On or Auto
- **Oper** – Should be On
- **VL Bmap** – Indicates classes using PFC
  - 8 is class 3
  - Might be two classes if RoCE
- **RxPPP** – Count of Received Per Priority Pause(PFC) by FI from IOM/IFM
- **TxPPP** – Count of Transmitted Per Priority Pause(PFC) by FI from IOM/IFM
- Can be cleared using '**clear qos statistics**'

Count is total of both Pause and UnPause!

If PFC is functioning properly these will not normally be zero!

show interface priority-flow-control - continued

Ethernetx/y/z links are links from the IOM/IFMs to server VICs

```
…continued
=============================================================
Port                Mode Oper(VL bmap)  RxPPP       TxPPP
=============================================================
Ethernet1/1/1       Auto On  (8)       38          0
Ethernet1/1/2       Auto Off           0           0
Ethernet1/1/3       Auto On  (8)       40          0
Ethernet1/1/4       Auto Off           0           0
…
Ethernet4/1/30      Auto Off           0           0
Ethernet4/1/31      Auto On  (8)       6642        0
Ethernet4/1/32      Auto Off           0           0
Ethernet4/1/33      Auto Off           0           0
```

Ethernet1/x/y are links to chassis 1
Ethernet4/x/y are links to chassis 4
...etc...

- Mode – PFC mode – Should be on or Auto
- Oper – Should be On
- VL Bmap – Indicates classes using PFC
  - 8 is class 3
  - Might be two classes if RoCE
- RxPPP – Count of Received Per Priority Pause(PFC) by HIF on IOM/IFM
- TxPPP – Count of Transmitted Per Priority Pause(PFC) by HIF on IOM/IFM
- Can be cleared using ‘clear qos statistics’

show fex detail

- **show fex detail shows which Ethernetx/y links go to which chassis**
- **Ethernetx/y/ links are links from the server VICs to IOM/IFMs**

```
show fex detail
FEX: 1 Description: FEX0001   state: Online
  FEX version: 9.3(5)I43(5a) [Switch version: 9.3(5)I43(5a)]
...
  Fabric interface state:
    Po1025 - Interface Up. State: Active
    Eth1/17 - Interface Up. State: Active
    Eth1/18 - Interface Up. State: Active
    Eth1/19 - Interface Up. State: Active
    Eth1/20 - Interface Up. State: Active
```

These 4 Ethernet interfaces connect to chassis 1 (FEX/IOM/IFM 1) via a 4 member port-channel1025

Datarate and PFC rate monitoring

- 4.3(4a) introduced:
  - show logging onboard fc-datarate
  - show logging onboard datarate
  - show logging onboard pfc
  - Applies to 4G and 5G Fis (6454, 64108, 6536, etc.)
- 4.3(6a) introduced:
  - show logging onboard datarate fex <x>
  - show logging onboard pfc fex <xx>
  - Applies to 2408 IOMs and 9108 IFMs
- Thresholds
  - Datarate
    - Datarate rising-threshold logged when 10 second average >= 80%
    - Datarate falling-threshold logged when 10 second average <= 70%
  - PFC
    - PFC Thresholds(PFC per second): 10G - 300, 25G - 750, 40G - 1200, 100G - 3000
    - PFC Rising threshold: 30% (e.g. 90 PFC/sec @ 10G)
    - PFC Falling threshold: 10% (e.g. 30 PFC/sec @ 10G)

FC uplink datarate monitoring

4.3(4a)

show logging onboard fc-datarate

- **Shows high Tx and Rx datarate on the upstream FC links to MDS**
- **Average over 10 seconds**

```
(nx-os)# show logging onboard fc-datarate
----------------------------
    Module:  1
----------------------------
    Switch OBFL Log:                                    Enabled
Notes:
  - Sampling period is 10 seconds
    Date and Time     : Interface  :   Speed   : Direction : KB per Sec : Link Utilization : Rising/Falling Thresholds
=======================================================================================================================
[11/01/2023 23:23:29] :   fc1/7        16G          RX         28976            90%              80%/70%
[11/01/2023 23:23:29] :   fc1/7        16G          TX         28957            90%              80%/70%
[11/01/2023 23:23:49] :   fc1/7        16G          RX             0             0%              80%/70%
[11/01/2023 23:23:49] :   fc1/7        16G          TX             0             0%              80%/70%
[11/01/2023 23:24:59] :   fc1/7        16G          RX         30873            96%              90%/70%
[11/01/2023 23:24:59] :   fc1/7        16G          TX         30859            96%              90%/70%
```

Ethernet datarate monitoring on FI Server Ports

4.3(4a)

show logging onboard datarate

- **Shows high Tx and Rx datarate on ethernetx/y links from FI to the IOM/IFM**
- **Average over 10 seconds**

```
(nx-os)# show logging onboard datarate
…

Notes:

Switch OBFL Log:  Enabled

Notes:
- Datarate Sampling Interval: 10 seconds
- Datarate Rising threshold: 80%
- Datarate Falling threshold: 70%

===========================================================================================================
===============
Date and Time              : Interface        : Speed   : Direction : Counter          : Gb per Sec  : Link Utilization :
Rising/Falling Thresholds
===========================================================================================================
===============
11/16/2023 15:07:22.939773 : Ethernet1/15     : 10Gb/s  : Rx        : data_rate_falling : 0.07       :  0%             : 80%/70%
11/16/2023 15:15:53.388784 : Ethernet1/15     : 10Gb/s  : Rx        : data_rate_rising  : 9.05       : 90%             : 80%/70%
11/16/2023 15:16:03.390267 : Ethernet1/15     : 10Gb/s  : Tx        : data_rate_rising  : 8.01       : 80%             : 80%/70%
11/16/2023 15:18:33.457884 : Ethernet1/15     : 10Gb/s  : Rx        : data_rate_falling : 5.57       : 55%             : 80%/70%
11/16/2023 15:18:43.458456 : Ethernet1/15     : 10Gb/s  : Tx        : data_rate_falling : 4.77       : 47%             : 80%/70%
11/16/2023 15:07:12.938718 : Ethernet1/16     : 10Gb/s  : Rx        : data_rate_falling : 0.87       :  8%             : 80%/70%
```

Ethernet datarate monitoring on IOM/IFM HIF Ports

4.3(6a)

show logging onboard datarate fex 3

- **Shows high Tx and Rx datarate on Ethernetx/y/z links on links from the IOM/IFM to VIC**
- **Average over 10 seconds**
- Applies to 24xx IOMs and 91xx IFMs

```
(nx-os)# show logging onboard datarate fex 3

slot  1
=======

Notes:

Switch OBFL Log:  Enabled

Notes:
- Datarate Sampling Interval: 10 seconds
- Datarate Rising threshold: 80%
- Datarate Falling threshold: 70%

==========================================================================================================================================================
Date and Time              : Interface          : Speed   : Direction : Counter             : Gb per Sec   : Link Utilization : Rising/Falling Thresholds
==========================================================================================================================================================
10/23/2024 13:30:53.764663 : Ethernet3/1/1      : 10Gb/s  : Rx        : data_rate_rising    : 8.79         : 87%              : 80%/70%
10/23/2024 13:31:03.829259 : Ethernet3/1/1      : 10Gb/s  : Tx        : data_rate_falling   : 1.97         : 19%              : 80%/70%
```

Ethernet PFC monitoring on FI Server Ports

4.3(4a)

show logging onboard pfc

- **Shows high PFC rate on ethernetx/y links from FI to the IOM/IFM**
- **Note speed dependent thresholds**

```
(nx-os)# show logging onboard pfc

slot  1
=======


Switch OBFL Log:  Enabled

Notes:
- PFC Sampling Interval: 10 seconds
- PFC Thresholds(PFC per second): 10G - 300, 25G - 750, 40G - 1200, 100G - 3000
- PFC Rising threshold: 30%
- PFC Falling threshold: 10%
============================================================================================================================================
Date and Time              : Interface         : Speed   : Direction : Counter          : PFC per Sec : Percentage : Rising/Falling Thresholds
============================================================================================================================================
11/16/2023 16:04:18.500036 : Ethernet1/15      : 10Gb/s  : Rx        : pfc_rate_rising  : 290         : 96%        : 30%/10%
11/16/2023 16:06:18.579532 : Ethernet1/15      : 10Gb/s  : Rx        : pfc_rate_falling : 20          :  6%        : 30%/10%
```

Ethernet PFC monitoring on IOM/IFM HIF Ports

4.3(6a)

show logging onboard datarate pfc fex <x>

- **Shows high PFC rate on ethernetx/y/z links are links from IOM/IFM to server VIC**
- Applies to 24xx IOMs and 91xx IFMs

```
(nx-os)# show logging onboard pfc fex 3

slot  1
=======


Switch OBFL Log:  Enabled

Notes:
- PFC Sampling Interval: 10 seconds
- PFC Thresholds(PFC per second): 10G - 300, 25G - 750, 40G - 1200, 100G - 3000
- PFC Rising threshold: 30%
- PFC Falling threshold: 10%

=====================================================================================================
=====
Date and Time              : Interface         : Speed   : Direction : Counter          : PFC per Sec : Percentage : Rising/Falling
Thresholds
=====================================================================================================
=====
10/23/2024 13:26:13.675098 : Ethernet3/1/1     : 10Gb/s  : Rx        : pfc_rate_rising  : 200         : 66%        : 30%/10%
10/23/2024 13:26:53.676343 : Ethernet3/1/1     : 10Gb/s  : Rx        : pfc_rate_falling : 20          :  6%        : 30%/10%
```

VIC Txwait and Rxwait

VICs measure time paused in each direction

- **Shows TxWait and RxWait from VIC's perspective**

```
adapter (macd):3# waitstats 0

PFC priorities active: 3

1 MIN DELTA TOTAL DESCRIPTION
0 1 RxWait Pri#3 2.5us Count
0 9462815 TxWait Pri#3 2.5us Count
Percentage RxWait for Pri#3 for last 5s/1m/1h/72h: 0%/0%/0%/0%
Percentage TxWait for Pri#3 for last 5s/1m/1h/72h: 0%/0%/0%/0%

adapter (macd):6# waitstatshistory
236 events from last boot, showing last 236, maximum 1024
Notes:
- Sampling period is 20 seconds
- Only wait delta >= 100 ms are logged
--------------------------------------------------------------------------------
| Interface | Priority | Delta Wait Time          | Congestion | Timestamp
|           |          | 2.5us ticks | seconds   |            |
--------------------------------------------------------------------------------
| port 2    | TxPri#3  |      122636 |         0 |         1% | 250506-22:51:58
| port 0    | TxPri#3  |      119860 |         0 |         1% | 250506-22:51:58
| port 2    | TxPri#3  |       87142 |         0 |         1% | 250506-22:51:38
| port 0    | TxPri#3  |       77988 |         0 |         0% | 250506-22:51:38
| port 2    | TxPri#3  |       77442 |         0 |         0% | 250506-22:51:18
```

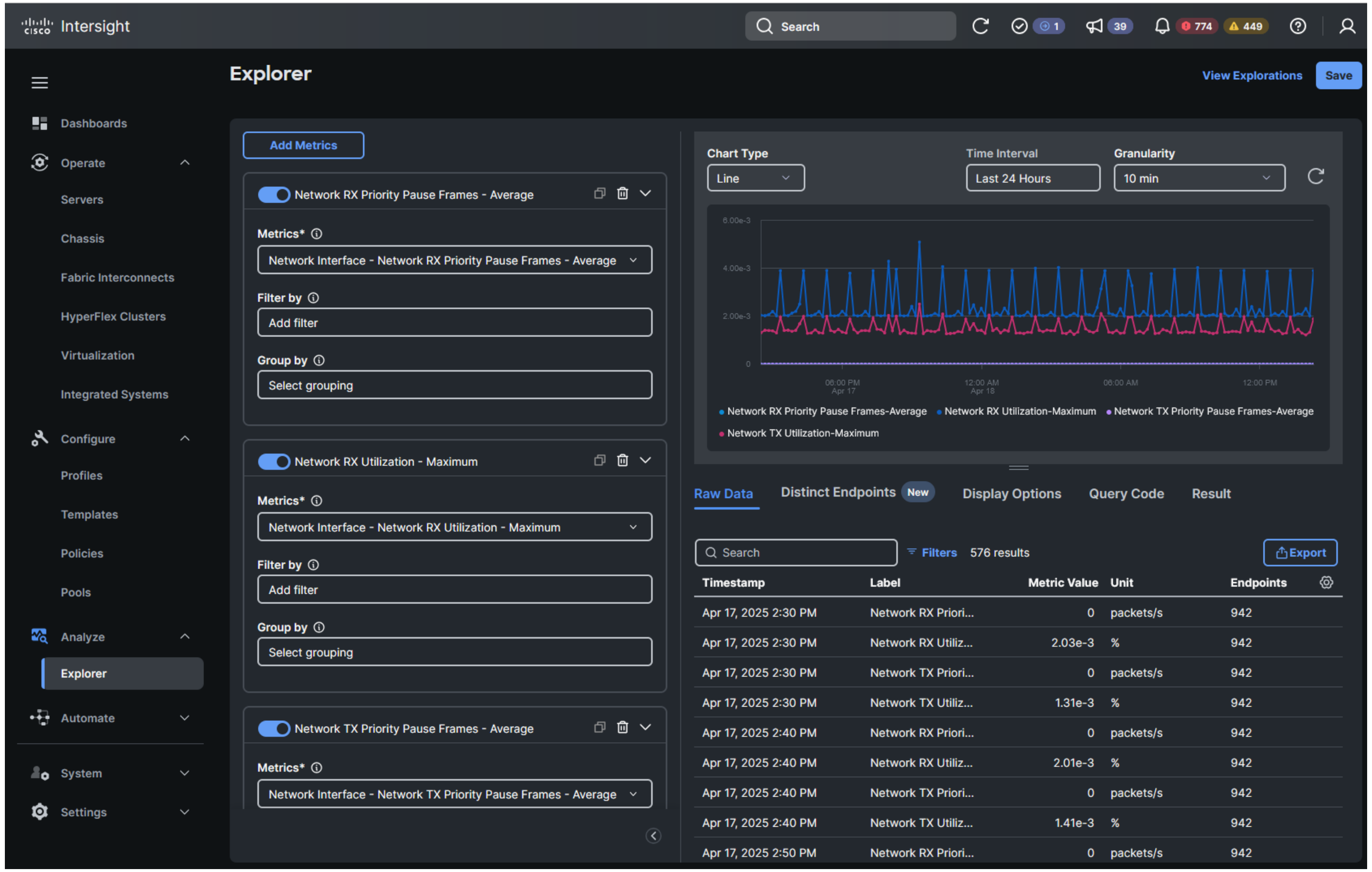
4.3(4a)

VIC logs with Tx/RxWait are included in
- UCS Chassis bundle
- IMM CIMC Bundle

Steps to issue VIC commands

1. ssh to FI
2. connect adapter 3/3/1
3. attach-mcp
4. vnic
5. attach-macd

Intersight Metrics Explorer



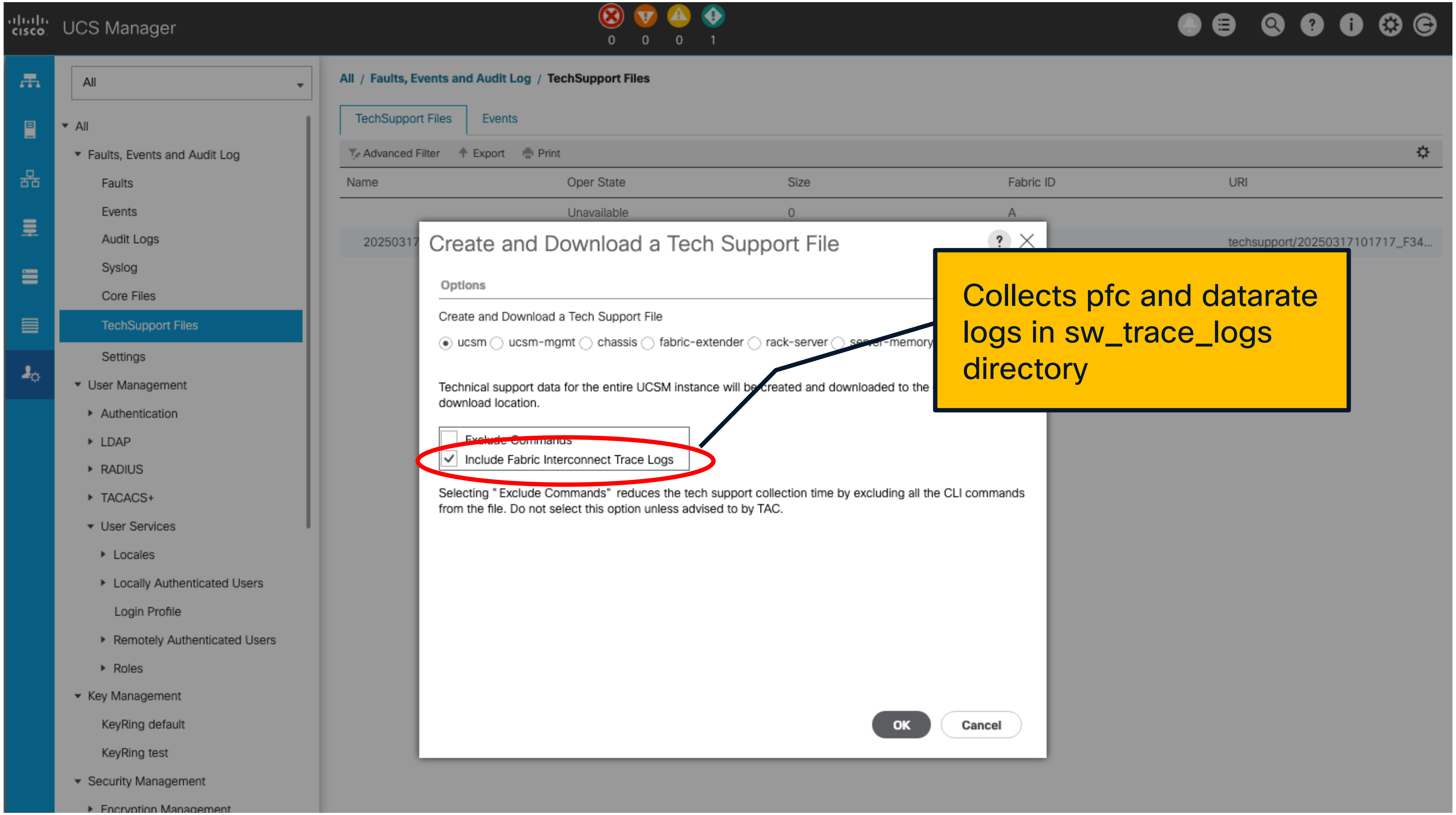
Intersight allows the graphing of various metrics.

Several metrics are there for congestion:

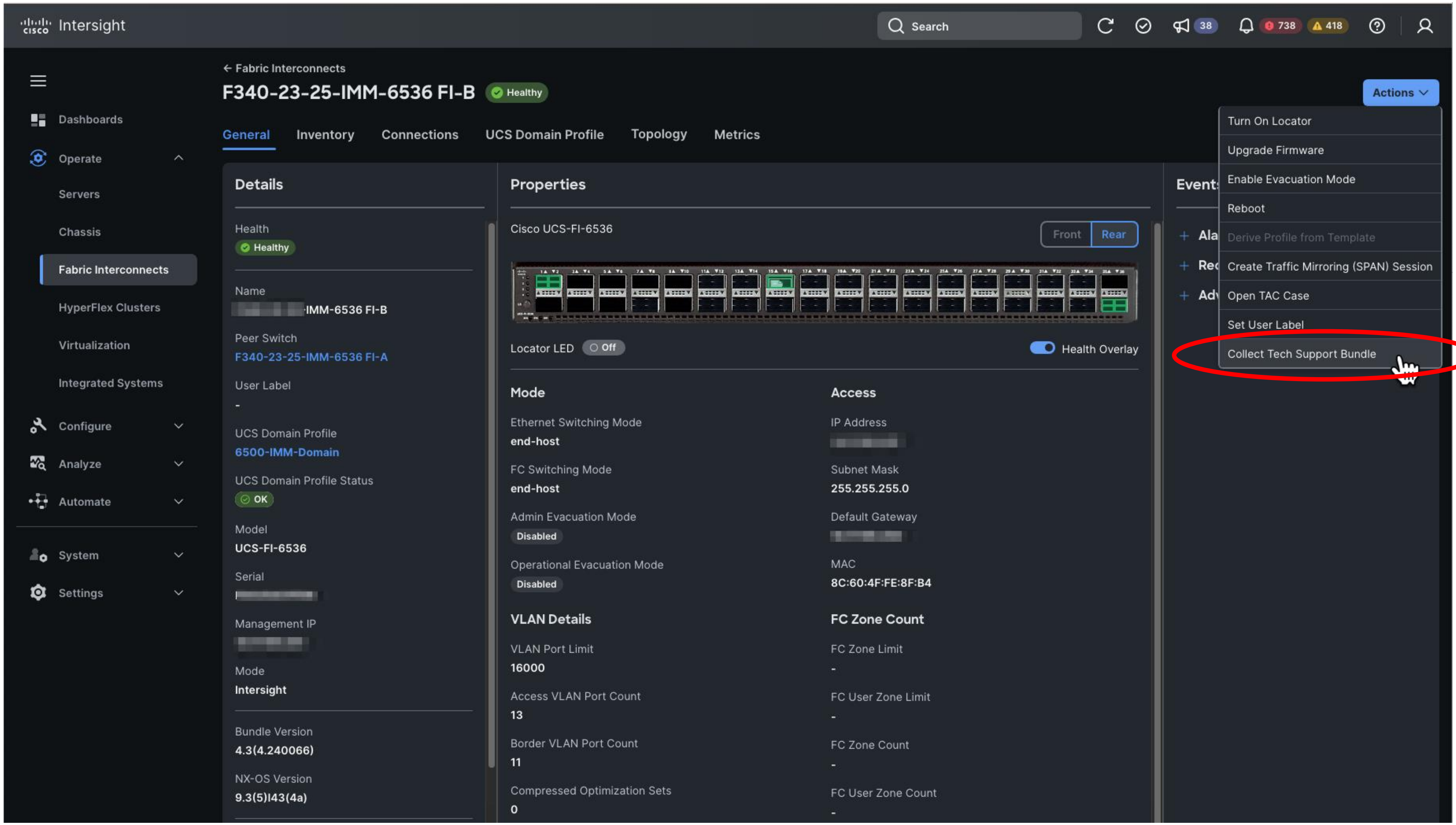
- Network RX Priority Pause Frames
- Network TX Priority Pause Frames
- Network RX Utilization
- Network TX Utilization

UCS Techsupport Commands

UCSM techsupport Include trace logs



IMM Techsupport



UCSM Techsupport

- The UCSM Techsupport bundle contains:
- Techsupports for both Fis A & B
  - sw_techsupportinfo - contains:
    - show interface priority-flow-control
    - show fex detail
    - show logging onboard fc-datarate
  - sw_trace_logs directory – contains:
    - datamon.log – show logging onboard datarate – Most recent entries
    - datamon_0.log – show logging onboard datarate – Older entries
    - pfcmon.log – show logging onboard pfc – Most current entries
    - pfcmon_0.log – show logging onboard pfc – Older entries
    - datamon_fexXXX.log – show logging onboard datarate fex X – Most recent entries
    - datamon_fexXXX_0.log – show logging onboard datarate fex X – Older entries
    - pfcmon_fexXXX.log – show logging onboard pfc fex X – Most recent entries
    - pfcmon_fexXXX_0.log – show logging onboard pfc fex X – Older entries

Conclusion

Conclusion

- Troubleshooting congestion within the UCS domain should no longer be a mystery.
- Ensure the UCS domain is at 4.3(6a) or later for the latest enhancements
- Follow troubleshooting workflows to determine congestion type and location
- If problem is determined to be slow drain at a server, then the server and applications must be investigated
- If problem is over-utilization, then the bottleneck connection should have its bandwidth increased
- More features are coming soon to make the data more accessible!

Agenda

Now available...

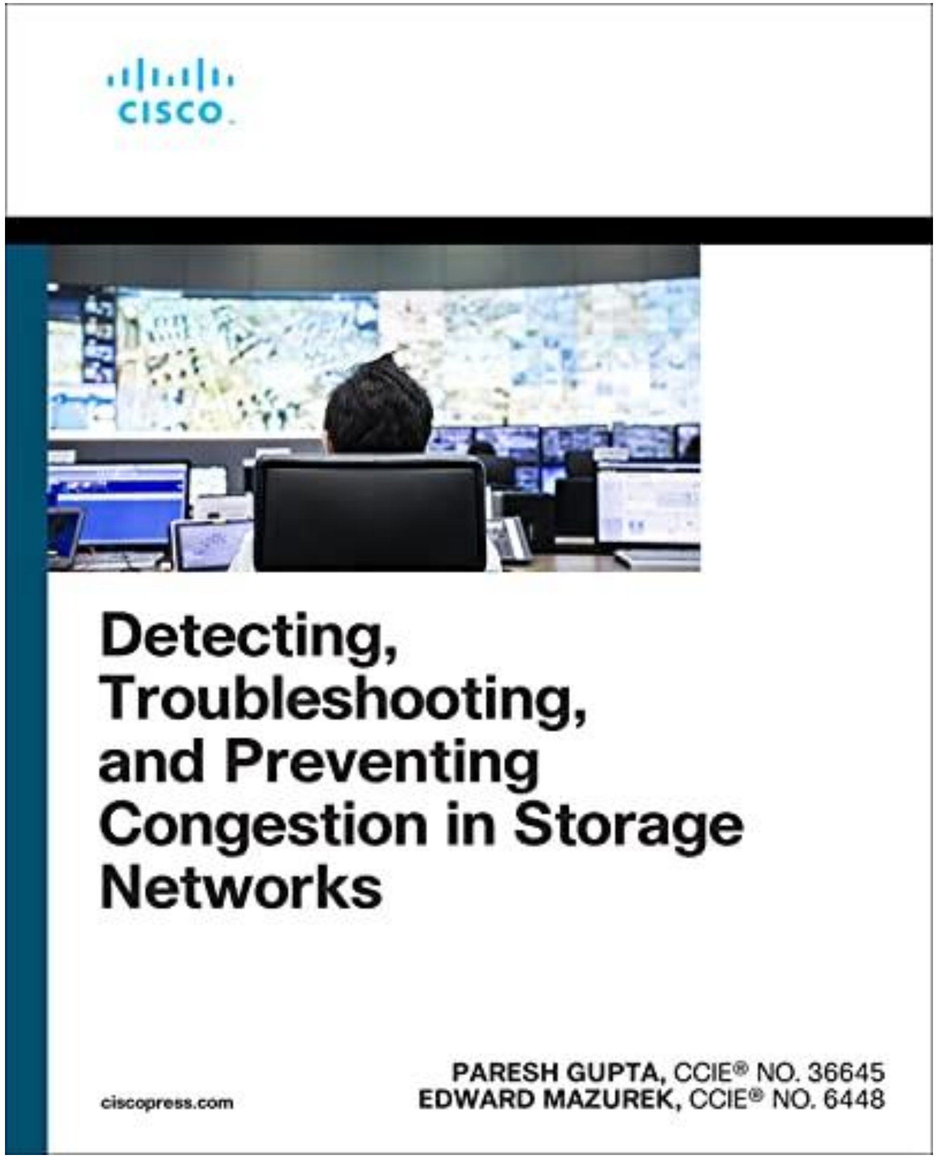


Table of Contents

Additional Relevant Sessions

Storage Area Networking

Reference

- Cisco Live 2025 San-Diego

  BRKDCN-2941 Advanced Storage Area Network Design – Tuesday June 10th 4:00 – 5:30 PM

  BRKDCN-2942 - SAN Congestion Management and SAN Analytics – Thursday June 12th 10:30 AM

- On Demand Library

  BRKDCN-2648 - What's New in NX-OS 8.x and 9.x for MDS and Nexus 9000 for SAN

  BRKDCN-3002 - Dynamic Ingress Rate Limiting as a Real Solution to SAN Congestion

  BRKDCN-3645 - SAN Insights-Real-time and Always-On NVMe Visibility at Scale

  BRKDCN-3641 - Manage, Operate, and Optimize High-Performance Storage Area Networks

Complete Your Session Evaluations

Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to win 1 of 5 full conference passes to Cisco Live 2026.

Earn 100 points per survey completed and compete on the Cisco Live Challenge leaderboard.

Level up and earn exclusive prizes!

Complete your surveys in the Cisco Live mobile app.

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

Contact me at: emazurek@cisco.com

Thank you

CISCO Live!

cisco