SNIA DNSF | DATA, NETWORKING, & STORAGE

The Evolution of Congestion Management in Fibre Channel

Live Webinar

August 27, 2024

10:00 am PT / 1:00 pm ET

Today’s Presenters

Erik Smith
Distinguished Engineer
Dell Technologies

Howard Johnson
Chair INCITS/Fibre Channel
Principal Engineer
Broadcom BSN
(Brocade)

Harsha Bharadwaj
Distinguished Engineer
DC Switching BU
Cisco

Dale Kaisner
Principal Architect
Broadcom ECD
(Emulex)

Scott Rowlands
Software Engineer Technical Staff
Dell Technologies

The SNIA Community

200
Corporations, universities, startups, and individuals

2,500
Active contributing members

50,000
Worldwide IT end users and professionals

SNIA DNSF | DATA, NETWORKING & STORAGE

What We Do

Drive the awareness and adoption of a broad set of technologies, including:

- ✓ Storage Protocols (Block, File, Object)
- ✓ Traditional and software-defined storage
- ✓ Disaggregated, virtualized and hyperconverged
- ✓ AI, including storage and networking considerations
- ✓ Edge implementation opportunities and factors
- ✓ Storage and networking security
- ✓ Acceleration and offloads
- ✓ Programming frameworks
- ✓ Sustainability

How We Do It

By delivering:

- **Expert webinars and podcasts**
- **White papers**
- **Articles in trade journals**
- **Blogs**
- **Social Media**
- **Presentations at industry events**

www.snia.org/dnsf

SNIA Legal Notice

- The material contained in this presentation is copyrighted by SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Today’s Agenda

- Fabric Notifications overview
  - Fibre Channel Architecture
  - Classic congestion scenario
- Implementations and considerations
  - Fabric
  - Host (HBA)
  - Storage

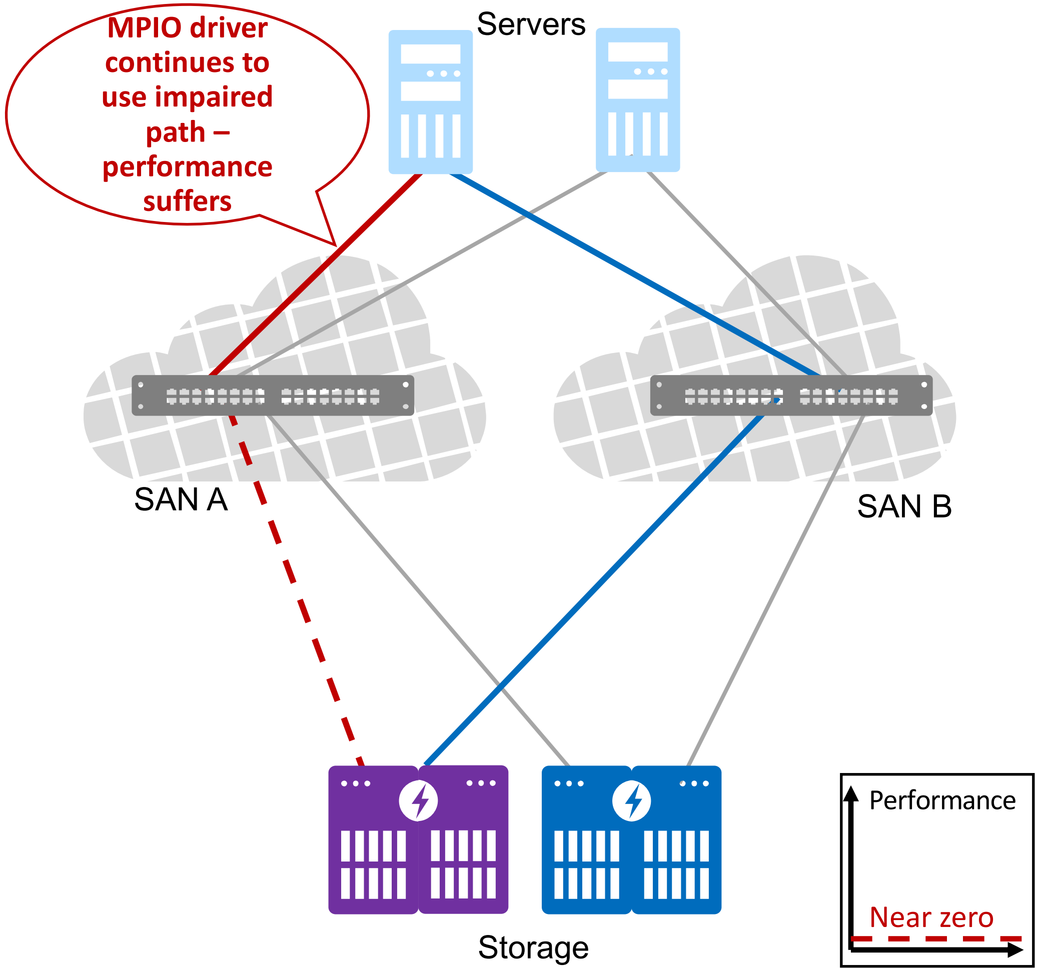
Fabric Notifications Overview

Howard Johnson
Broadcom BSN
(Brocade)

The Problem

- Persistent, intermittent errors
  - Significant role in customer escalations
  - Difficult for traditional solutions to resolve
  - Required manual intervention increases mitigation costs
  - MPIO solutions struggle with resolution, which impacts the dual fabric paradigm
- Causes
  - Marginal cables, SFPs, connections, etc
  - Congestion due to lost credit, credit stall, or oversubscription
- Why now?
  - Fibre Channel solutions are mature and diversified
  - Identification and mitigation tools have evolved
  - Customers are demanding more automation

Flakey paths

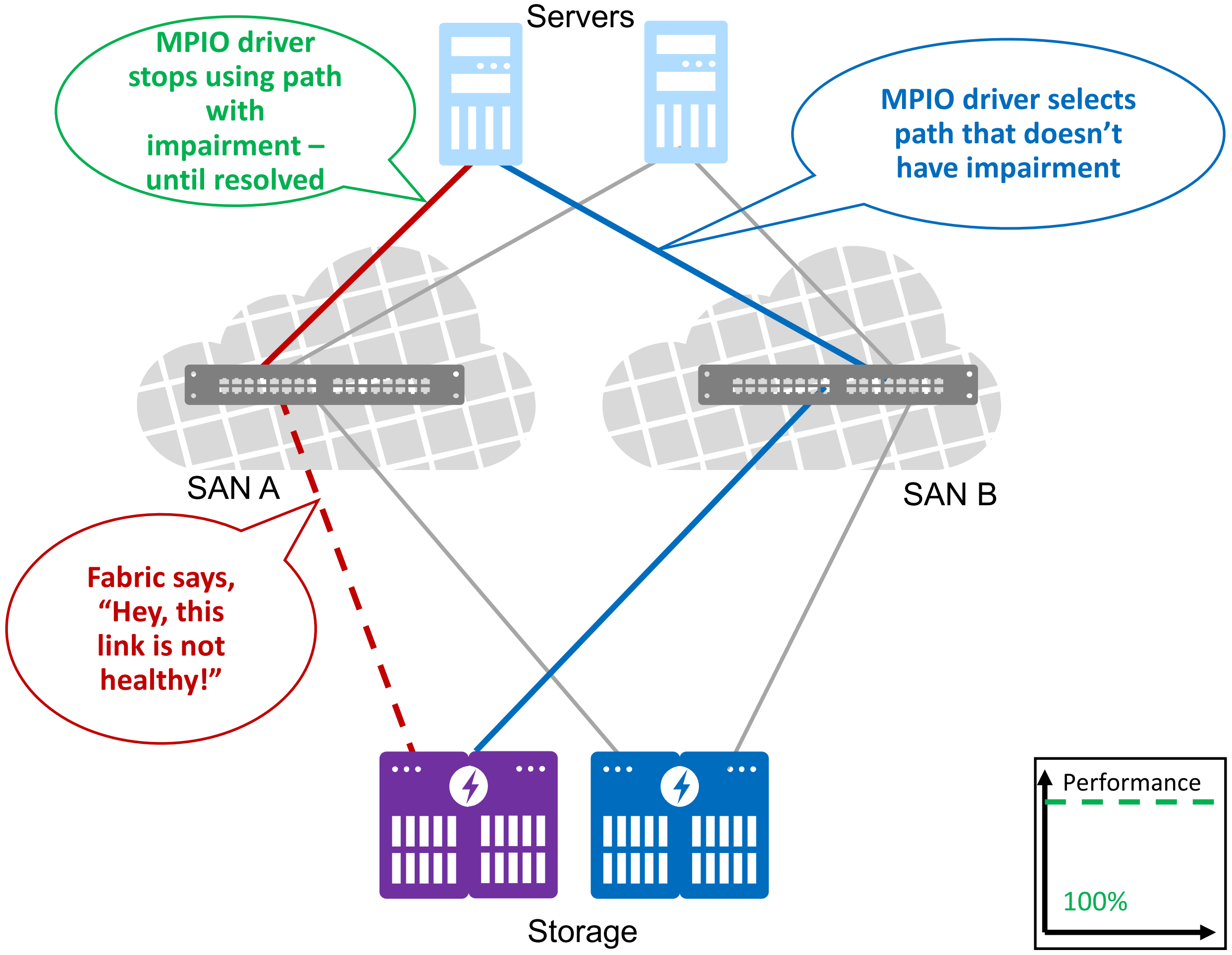


MPIO driver continues to use impaired path – performance suffers

The Solution

- Fabric Notifications
  - Notifications and signals
    - Generated by the fabric
    - Inform devices of impairments
- Notifications
  - Reporting: Events sent to registered devices
  - Diagnostics: Helps efficiently evaluate errors
  - Operation: Extended Link Services (ELS)
- Signals
  - Signaling: Report resource depletion to registered devices
  - Diagnostics: Transmitter indicates resource usage
  - Operation: Link level Primitive Signal

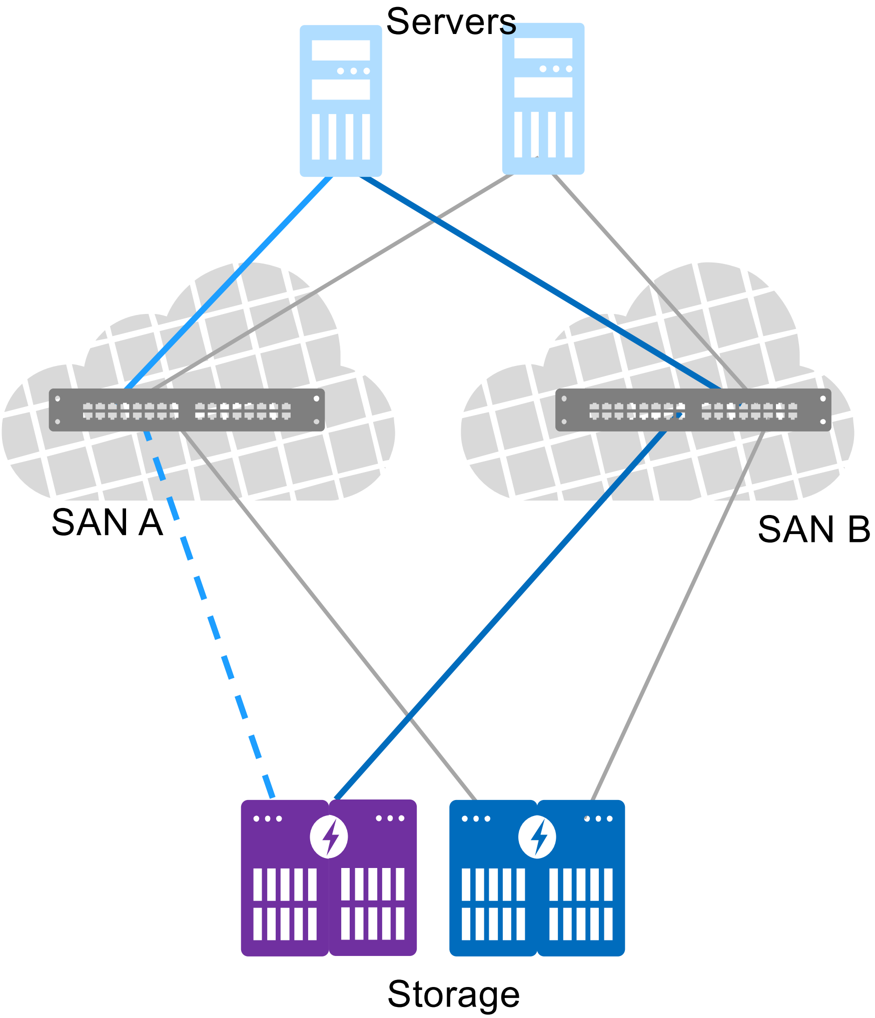
Fabric Notifications


Servers
MPIO driver stops using path with impairment – until resolved
MPIO driver selects path that doesn't have impairment
SAN A
SAN B
Fabric says, "Hey, this link is not healthy!"
Storage
Performance
100%

Fibre Channel Standards

- Standards History
  - Began in December 2018
  - Fully specified in April 2022
  - Standards complete in June 2024

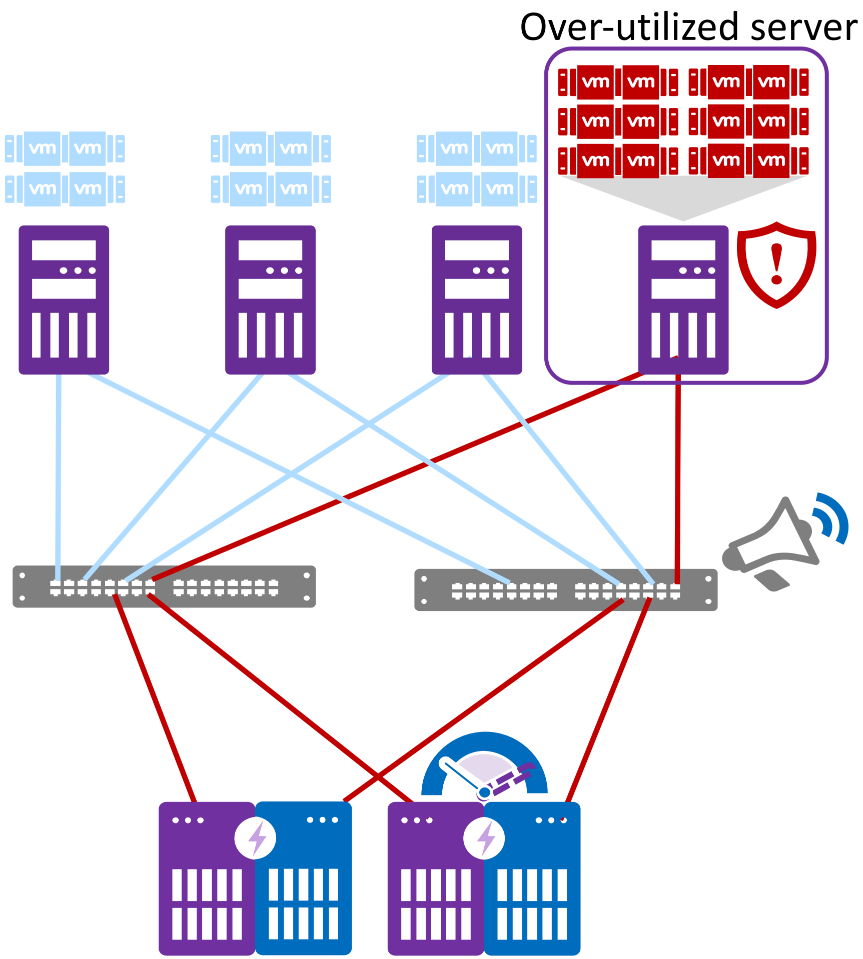
- Draft standards
  - FC-FS-6: Congestion Signals (r0.3)
    - ANSI Standard
  - FC-LS-5: Notifications (r5.01)
    - INCITS final draft
  - FC-SW-8: Fabric detection and generation (r1.01)
    - INCITS final draft



Fabric Notifications

- Software-based FPIN
  - Extended Link Services commands
  - Fabric Performance Impact Notification (FPIN)
- Hardware-based Congestion Signal primitives
  - Defined as Primitive Signal characters
  - Warning and Alarm Signals

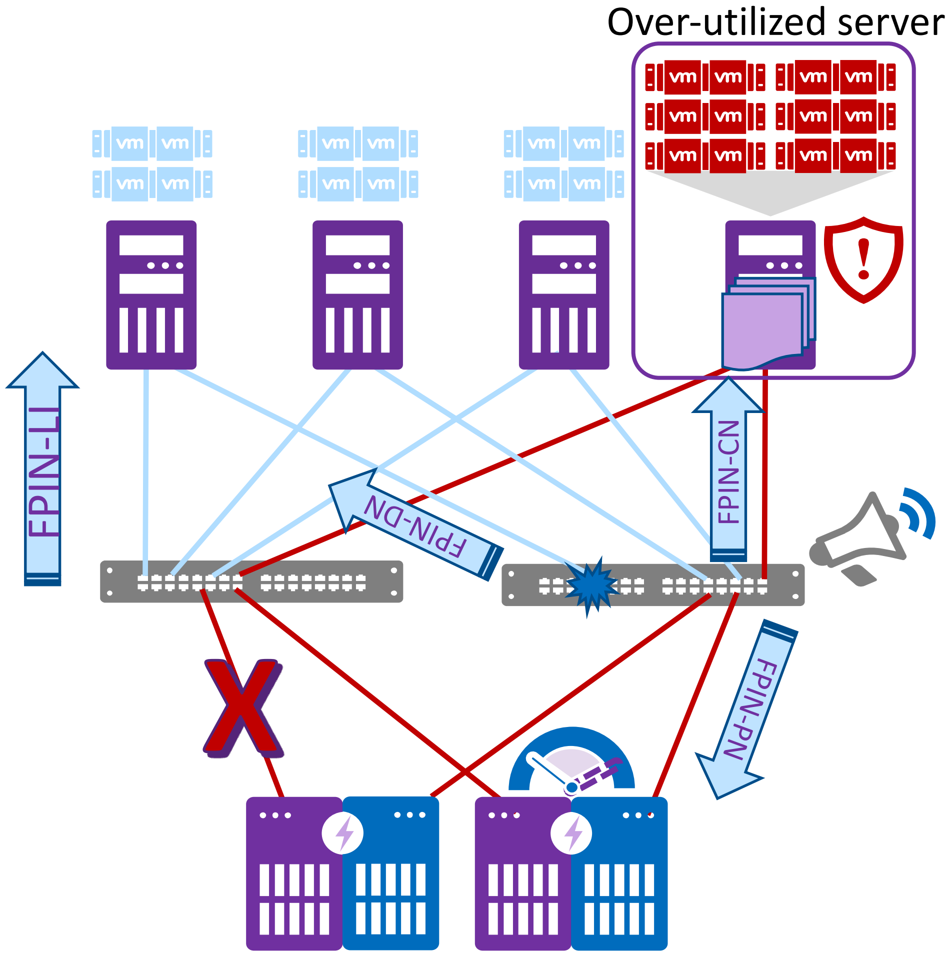
Fabric Notifications



Fabric Notifications

- Link Integrity Notifications
  - Link Integrity notifications are received by MPIO drivers, which update the path selection to avoid the impaired path
  - The Link Integrity notifications allow the MPIO driver to take the appropriate action for errors (e.g., CRC, ITW)
- Congestion and Peer Congestion Notifications
  - Congestion notifications are the software equivalent of the Congestion Signal and are sent to congesting end devices
  - Peer congestion notifications are sent to registered and “in-zone” peers of end devices that are experiencing congestion
- SCSI Command Delivery Notifications
  - Delivery notifications are sent when a fabric discards a SCSI command or status frame to notify the initiator of the failure

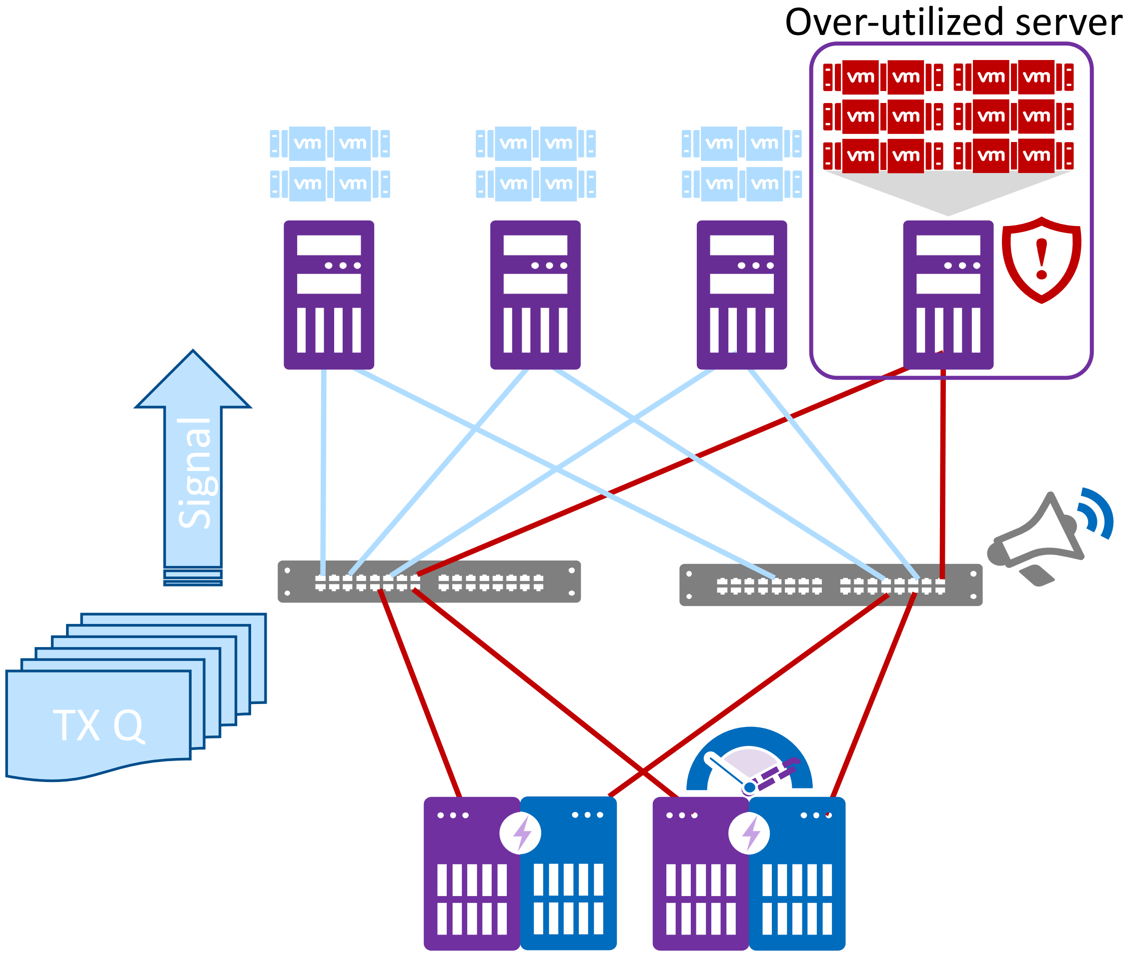
Fabric Performance Impact Notifications (FPIN)



Fabric Notifications

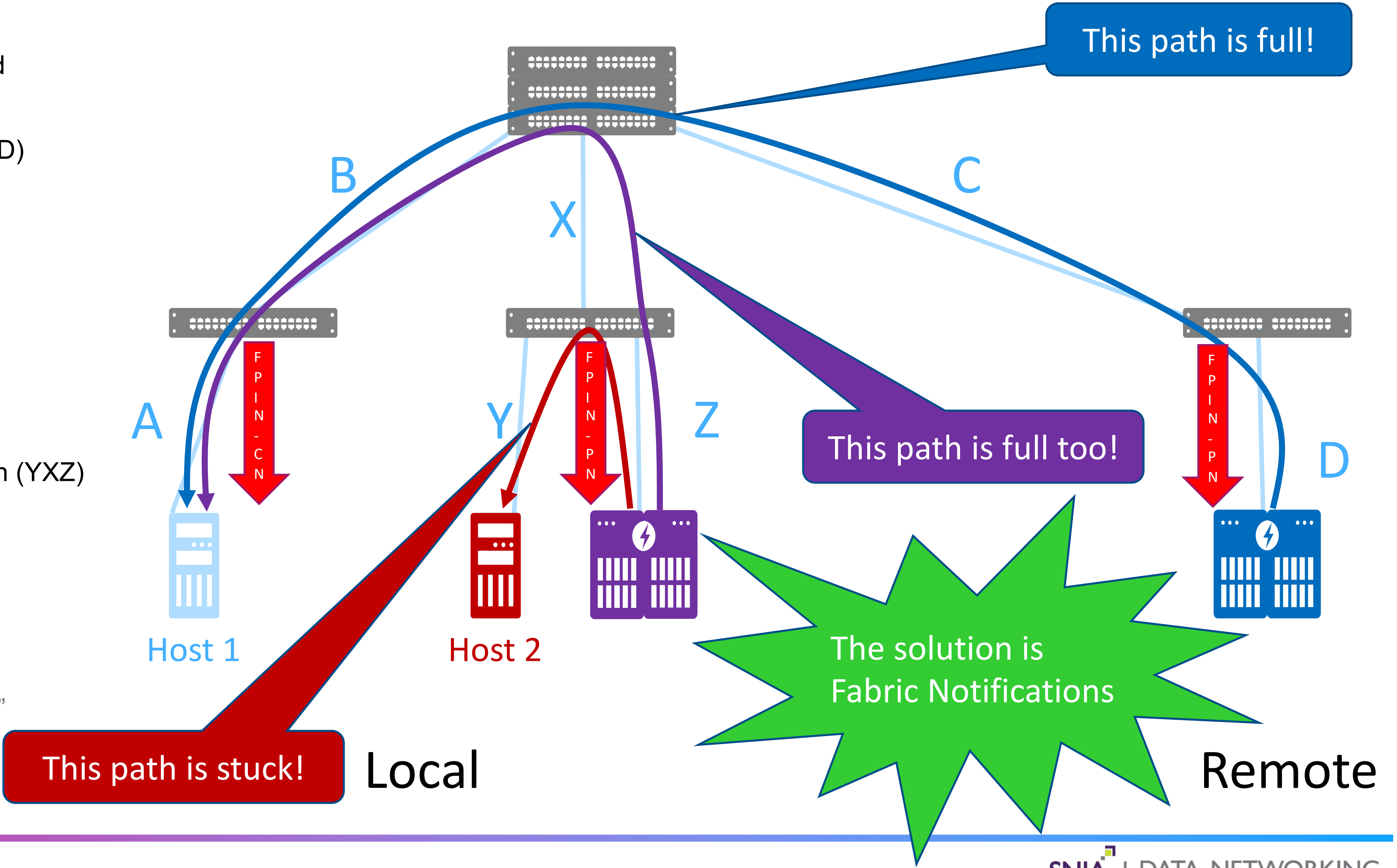
- Congestion Signals
  - Immediate feedback mechanism
  - Indicates transmission resources are consumed
- Link level communication
  - Transmitter to receiver

Congestion Signals



Congestion scenario

- All links running at the same speed
- Host 1 starts remote backup (ABCD)
  - Happy host
  - Links are full
  - No problems
- Host 1 starts local backup (AXZ)
  - Happy host
  - Links are full
  - No "noticeable" problems
- Host 2 starts production application (YXZ)
  - Unhappy host
  - Link barely running
  - There are problems
- Conclusion
  - "The Purple storage is broken!"
  - "Call the Storage Admin!"
  - "And the Storage vendor!"



Implementations and Considerations

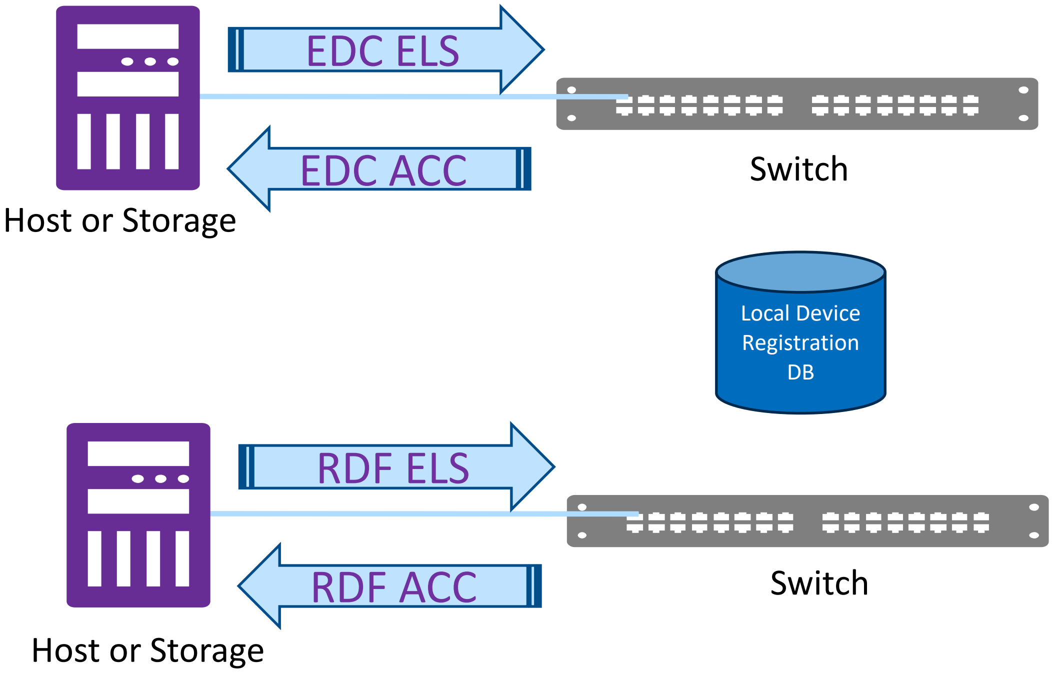
Fabric

Harsha Bharadwaj

Cisco

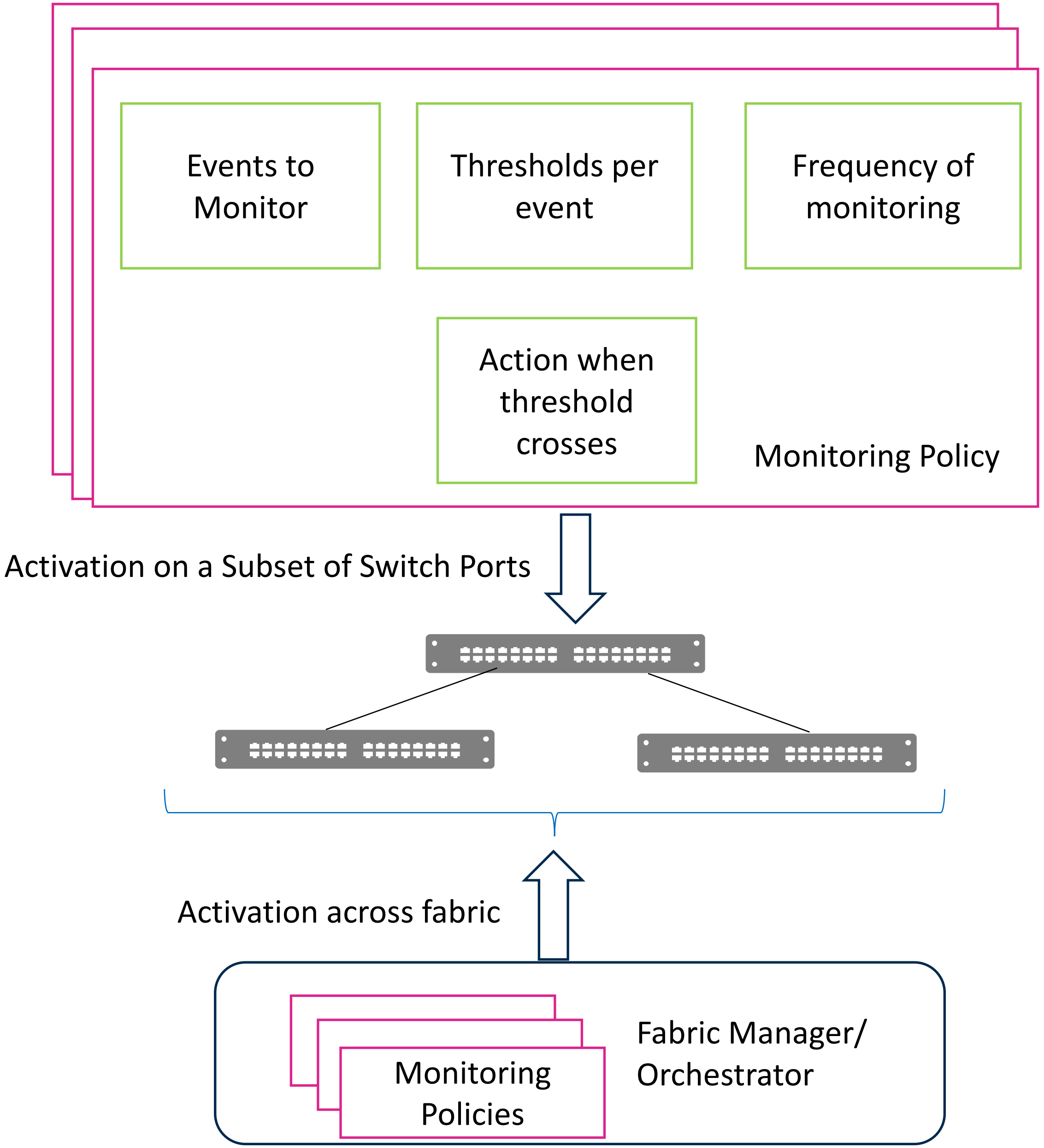
End Devices Register with Switch

- For Congestion Signals using EDC ELS
  - Host/Storage sends Exchange Diagnostic Capabilities (EDC) indicating its capabilities
    - Rx/Tx, Severity Levels (Warning/Alarm), Frequency
  - Switch returns EDC Accept with its capabilities
    - Rx/Tx, Severity Levels (Warning/Alarm), Frequency
  - ‘Least capable’ values become operational
- For FPIN using RDF ELS
  - Host/Storage sends Register Diagnostic Function (RDF) indicating types of FPIN it is interested
    - Congestion, Peer-Congestion, Link Integrity, Delivery Failure
  - Switch returns RDF Accept for supported FPIN types
- Registered Devices stored in a Database inside switch
- Host/Storage implementations choose to register for either FPIN or Congestion Signals or both
- Switch Rejects EDC/RDF if feature not supported/enabled



Switch Monitoring Policies

- FC Switch Monitoring function
  - Policy Configuration
    - Events (Eg: Congestion, Link Integrity)
    - Thresholds (Eg: Warning, Alarm)
    - Frequency (Eg: 5sec)
    - Actions (Eg: FPIN, Congestion Signals)
  - Activation (Enforcement)
    - Per-Switch (set of ports)
    - Across fabric
- Default policy may be good for most situations
  - Modify only for special use-cases
  - Consult switch vendor documentation for guidance

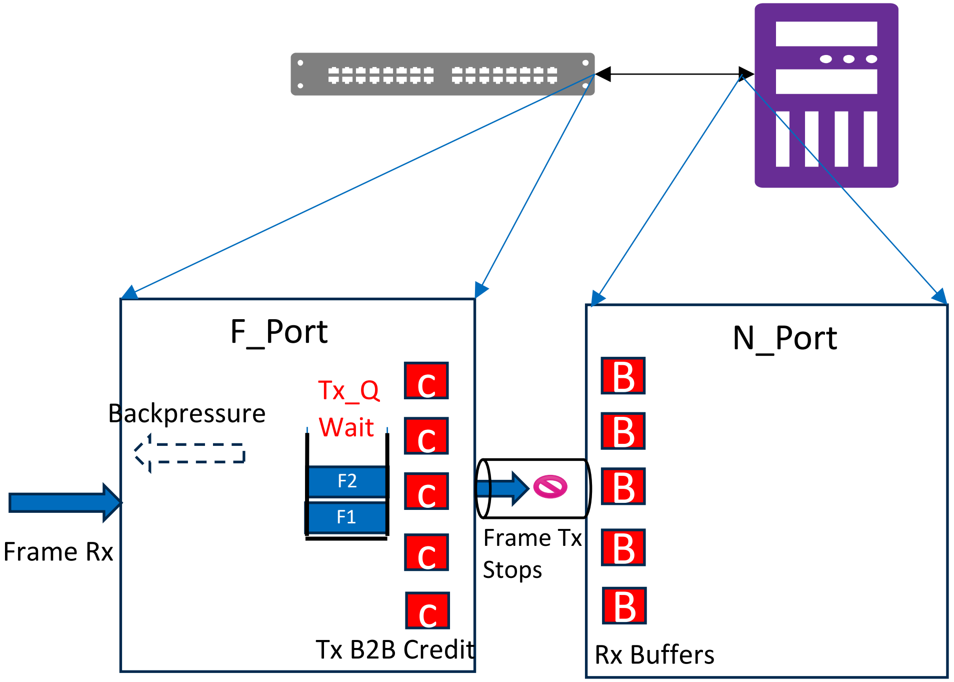


Congestion

- Switch ports experience two main types of congestion
  - Credit Stall
  - Oversubscription
- Root cause: Device behaviors that make them “not so good fabric citizens” in a no-drop FC fabric
  - Congestion originates at switch F_Port
  - Persistent F_port congestion causes congestion spreading to E_Ports, creating victim devices
- Vendor Specific switch centric congestion mitigation solutions existed
  - Eg: Congestion Isolation to quarantine VLs
- FPIN/Signals are standard notifications that put congestion mitigation responsibility on Devices
  - FPIN/Signal differ in scope and information they carry:
    - FPIN: Scope → Congestion causing device and zoned peers; Info → Type, Interval, Detecting & Attached Port WWPN, Severity etc
    - Signal: Scope → Congestion causing device; Info → Only indicates congestion detected and its severity
  - Devices may take mitigation actions in response to notifications - before congestion spreading
    - Actions typically involves some kind IO throttling
  - Relieved Congestion also notified to devices by
    - FPIN with Event-Type=Congestion Cleared (or)
    - Absence of Congestion Signals

Detecting Credit Stall

- B2B Crediting on FC links
- FC port Rx advertises ‘N’ MTU buffers to its peer port Tx as ‘N’ credits
  - No of Buffers at port initializes a Tx B2B Credit counter at peer during login
  - Every frame Tx to peer: Tx B2B Credit --
  - Every credit return (R_RDY) Rx from peer: Tx B2B Credit ++
- If peer device does not return credits in time, Tx B2B Credit eventually becomes 0, stopping all Tx
- Packets held back inside switch buffers, congestion spreading
- Switches detect credit stall per-port based on:
  - Tx B2B Credit == 0
  - Frames Tx stops for a contiguous time interval > threshold (Eg: 100ms)
- Switch Monitoring policy configured for Credit Stall on all F_ports
  - Action: Generate FPIN/Signals or Both

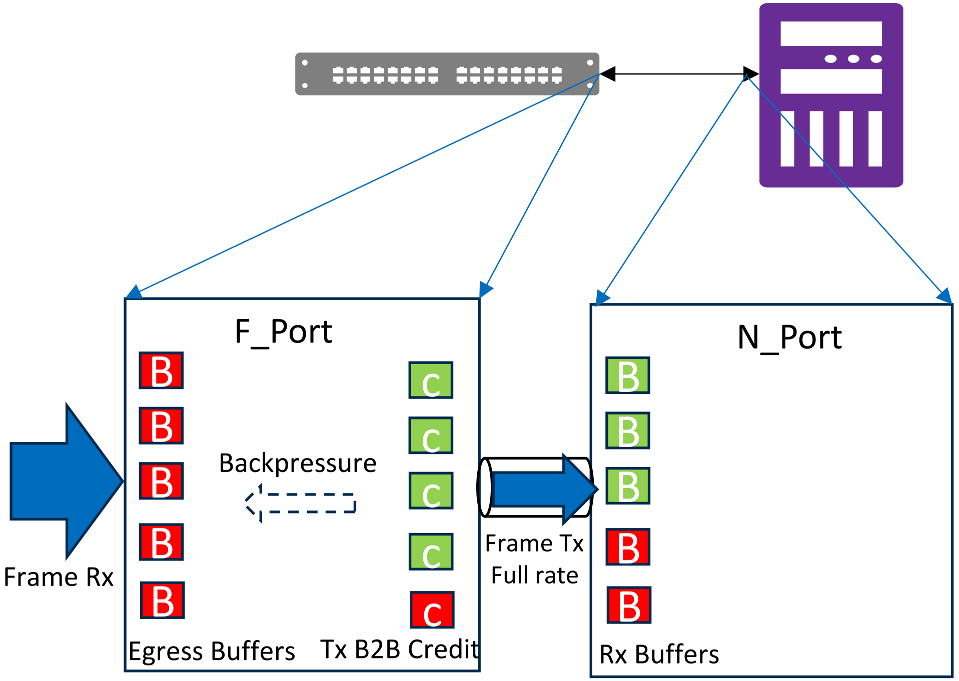


Credit Stall Condition on a F_port:

(Tx B2B credit == 0) && (Tx_Q Waittime > Threshold)

Detecting Oversubscription

- If a Port receives more traffic than the port bandwidth, residual traffic held back in switches due to no-drop nature of FC switches
  - Speed mismatched devices communicating
    - Eg: 32G Host ← 64G Target
  - 1:N traffic patterns
    - Eg: 3 x 32G Host → 32G Target
  - Host doing high IOPS/Throughput READ I/O but the response rate exceeds the link speed
    - Eg: 1MB READ request responses @ ~ 3.5K IOPS can saturate a 32G N_Port
- Different root cause than credit stall, but same side effect of congestion spreading
- Most common type of congestion today in FC fabrics
- Switches detect it per-port based on:
  - Tx data rate on a port very high (Eg: >80%)
  - Switch buffer buildup (or) higher packet switching latency
- Switch Monitoring policy configured for Oversubscription on all F_ports
  - Action: Generate FPIN/Signals or Both

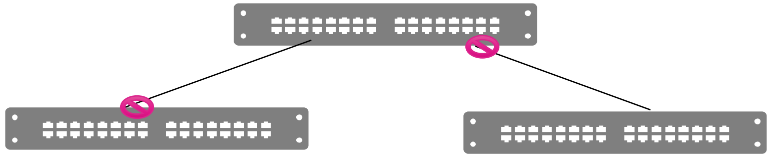


Oversubscription Condition on a F_port:

(Tx Rate > BW_Threshold) &&
(Switch buffer occupancy > Buffer_Threshold)

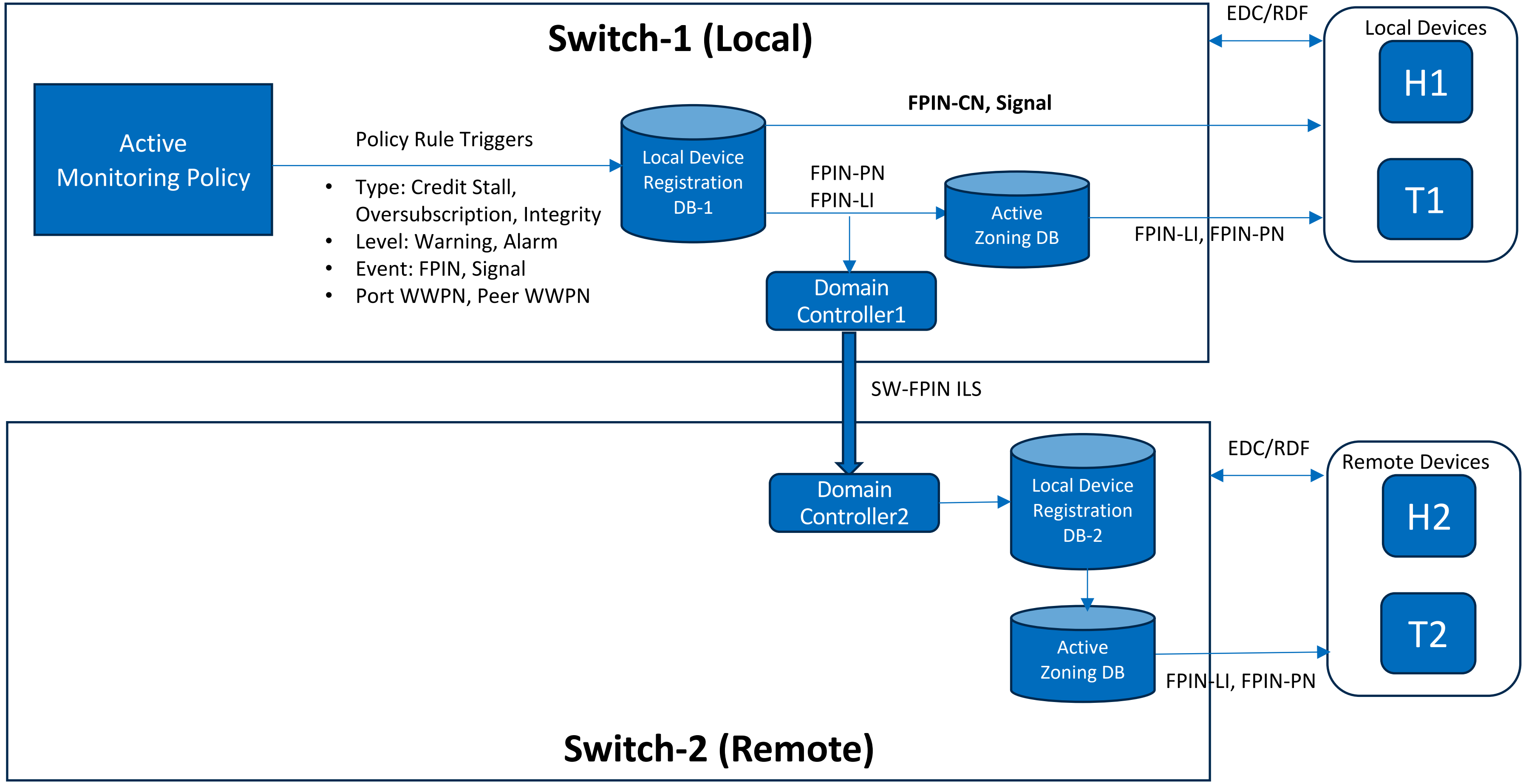
Detecting Link Integrity

- Faulty links can impact I/O performance
  - Bad SFPs, Cables, Hardware can cause packet drops, runt packets
  - I/O Aborts and Retries
  - Higher I/O latency
- Detection: Switch port HW error counter crossing the policy threshold
  - Link Loss
  - Sync Loss
  - Signal Loss
  - Invalid Words
  - Invalid CRC
  - Uncorrectable FEC
- Mitigation action may involve path modifications to bypass faulty links
- Switch Monitoring policy configured for Link Integrity on all (E/F) Switch Ports
  - Action: Generate FPIN



Link Integrity Condition F/E Port:
(Any Link Integrity Counter > Threshold)

Fabric Notification Delivery and Distribution



- Type: Credit Stall, Oversubscription, Integrity
- Level: Warning, Alarm
- Event: FPIN, Signal
- Port WWPN, Peer WWPN

Host (HBA)

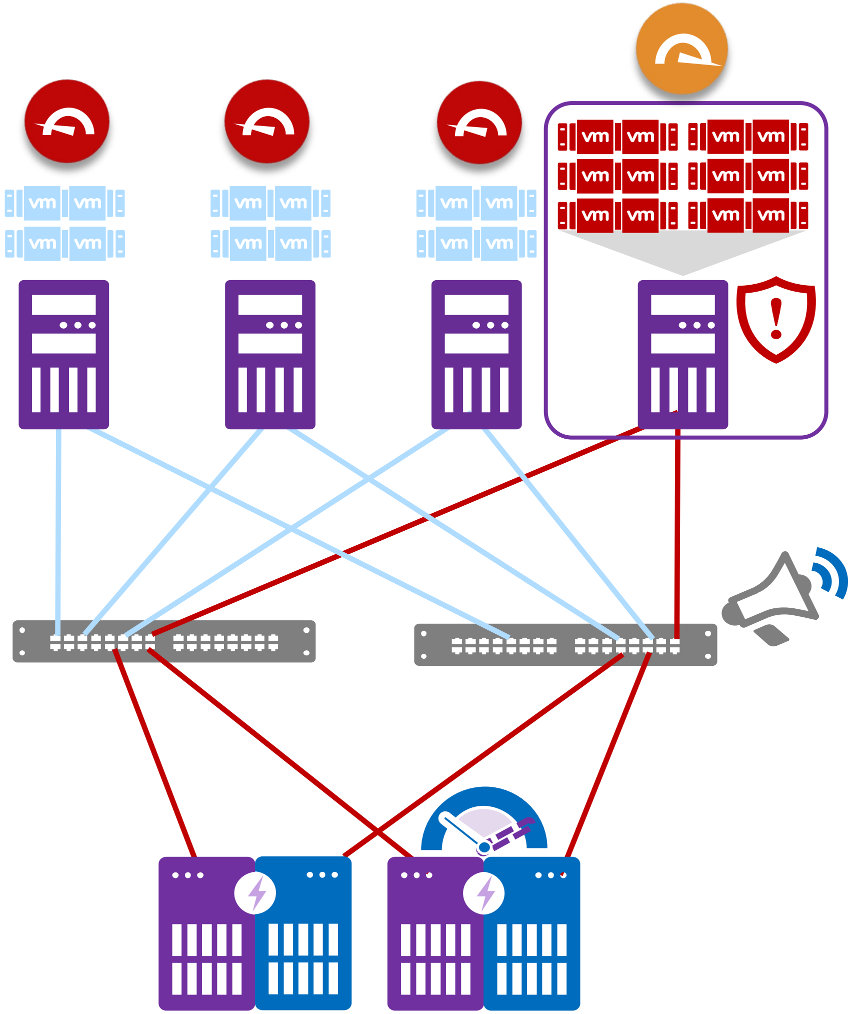
Dale Kaisner
Broadcom ECD
(Emulex)

HBA – Reaction to FPIN Notifications and Signals

- Direct action
  - Example: Active Congestion Management
- Forwarding to OS layer for action
  - Example: Link Integrity Alerts to MPIO Driver
- Logging & alerts
  - Example: Peer Notification Events

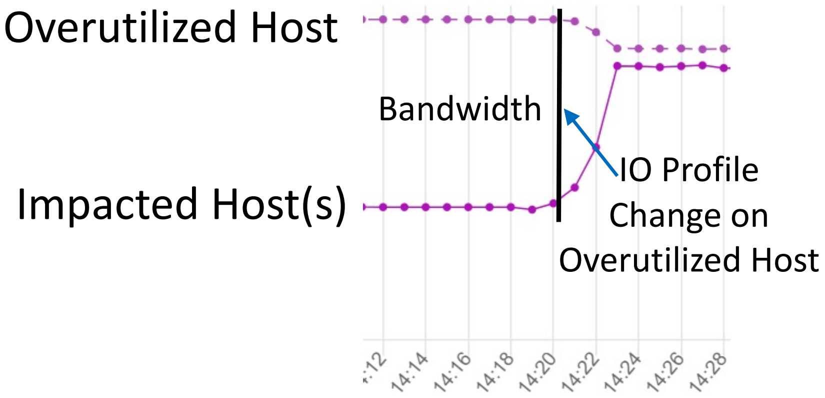
HBA – Congestion Example

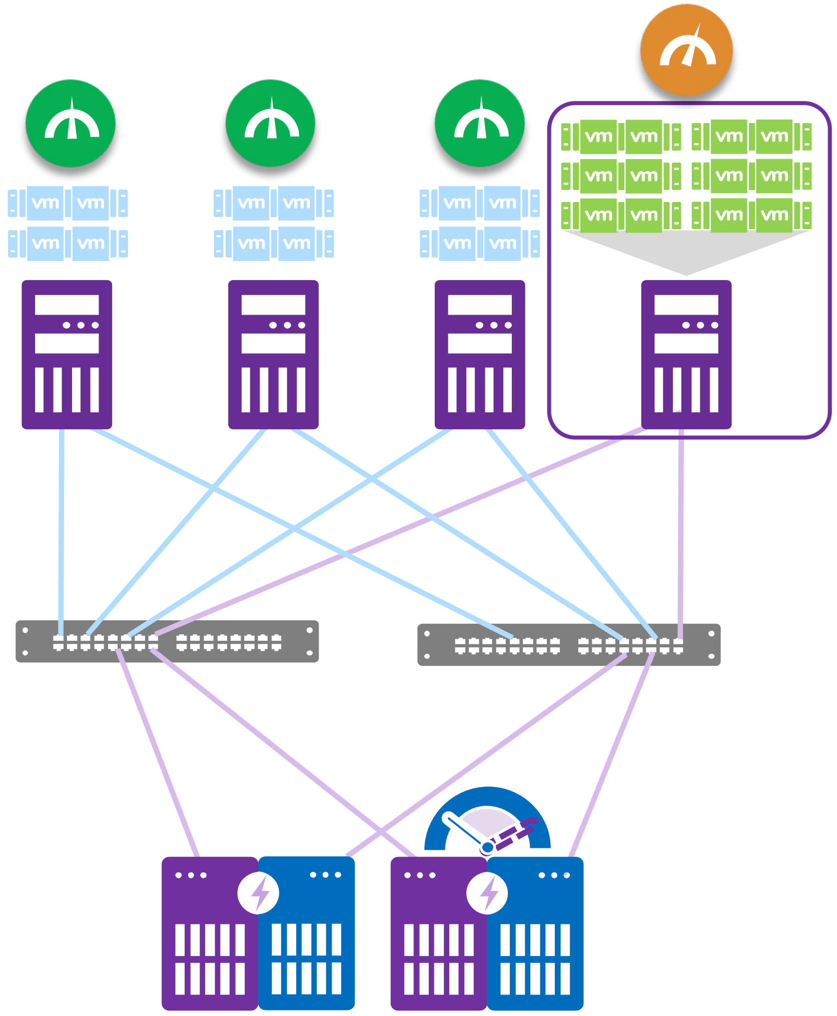
- Overutilized server creates congestion impact that affects other hosts
- FPIN-CN or signal from switch identifies offending host



HBA – Congestion Example

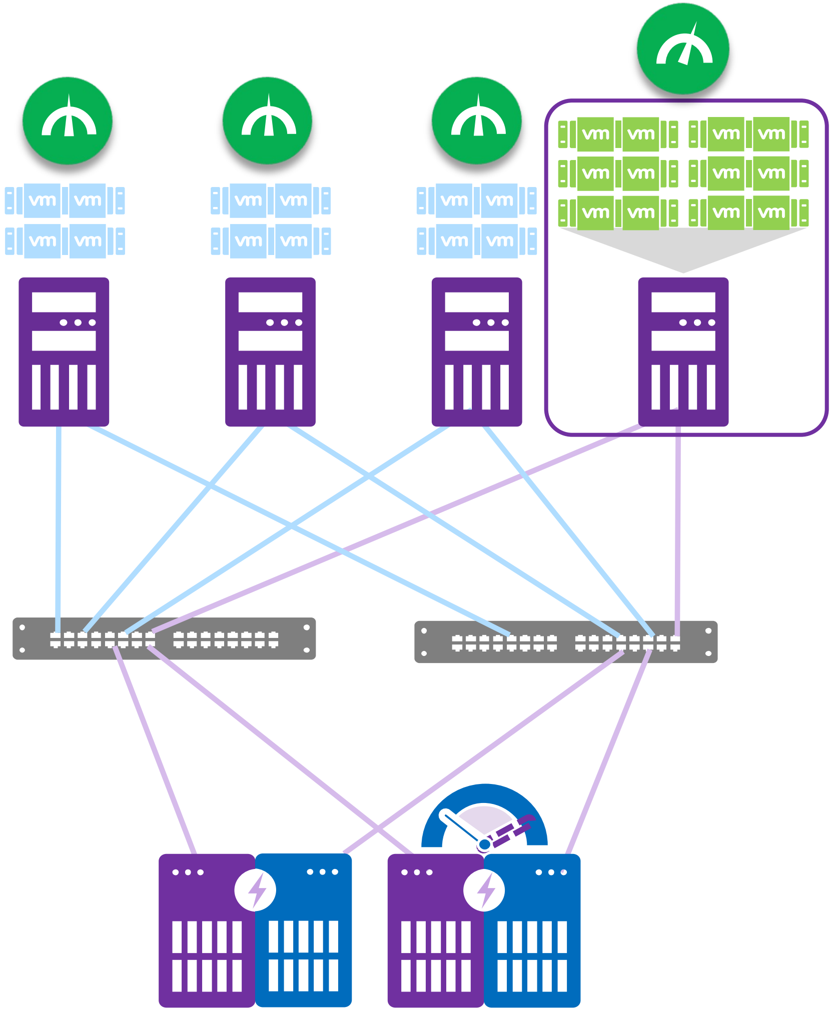
- Overutilized server creates congestion impact that affects other hosts
- FPIN-CN or signal from switch identifies overutilized host
- Overutilized host HBA automatically changes IO profile to alleviate congestion
- Other hosts return to expected performance levels





HBA – Congestion Example

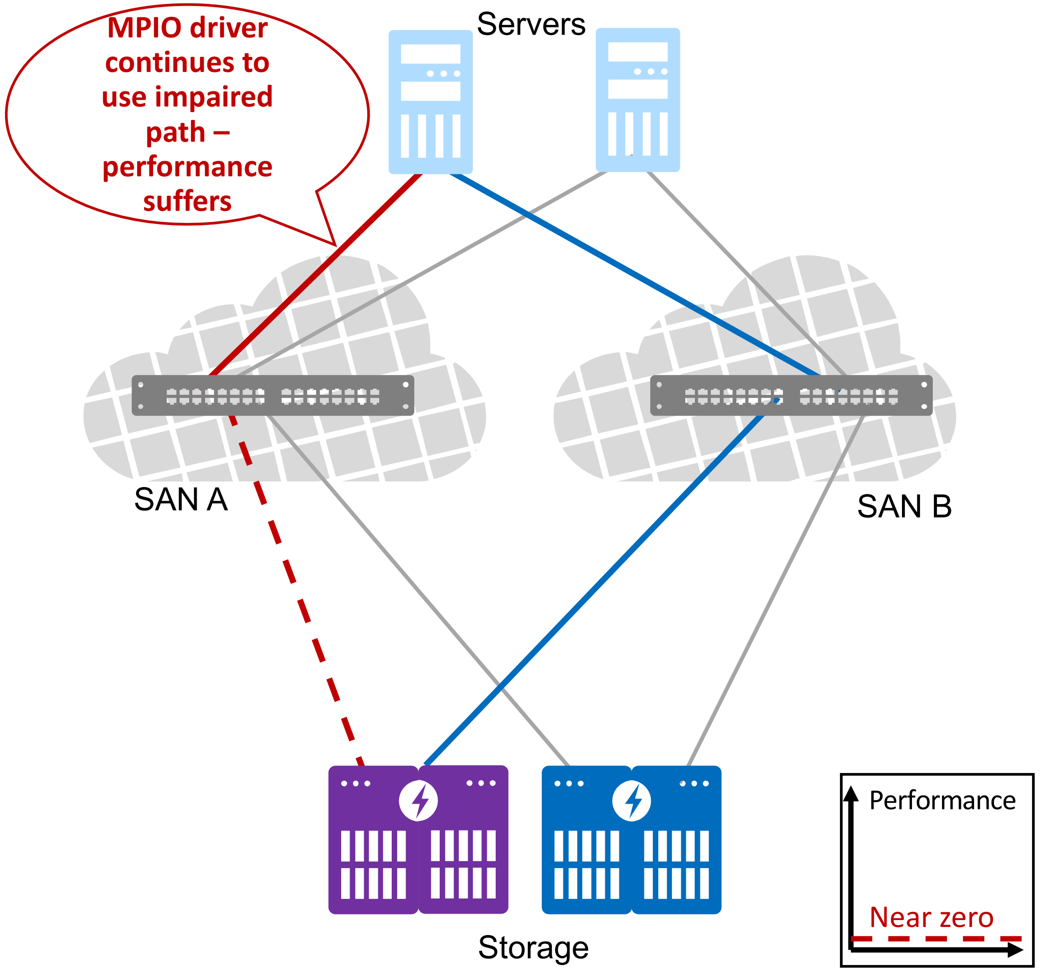
- Overutilized server creates congestion impact that affects other hosts
- FPIN-CN or signal from switch identifies overutilized host
- Overutilized host HBA automatically changes IO profile to alleviate congestion
- Other hosts return to expected performance levels
- Once the congestion clears the HBA IO profile returns to full performance



HBA – Link Integrity Example

- A flakey path is impacting IO on a multipath link
- Switch identifies the impaired link with an FPIN-LI

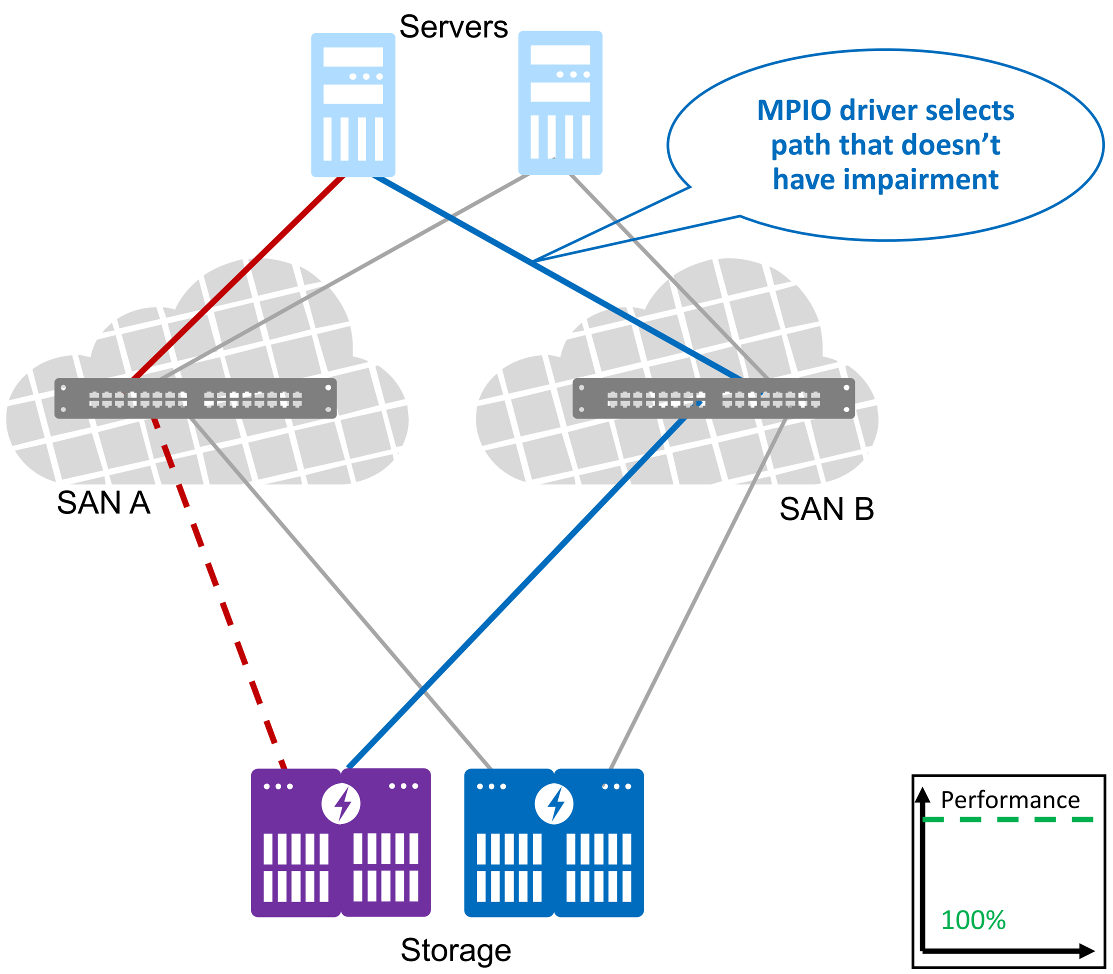
Flakey paths


MPIO driver continues to use impaired path – performance suffers
Servers
SAN A
SAN B
Storage
Performance
Near zero

HBA – Link Integrity Example

- A flakey path is impacting IO on a multipath link
- Switch identifies the impaired link with an FPIN-LI
- HBA passes that information up to the Linux MPIO layer
- MPIO driver moves traffic off impaired path

Fabric Notifications



HBA – Reviewing Settings for Notification and Signals

- Confirm Fabric Configuration
- Confirm Operating System Configuration
- Confirm HBA Configuration

Storage

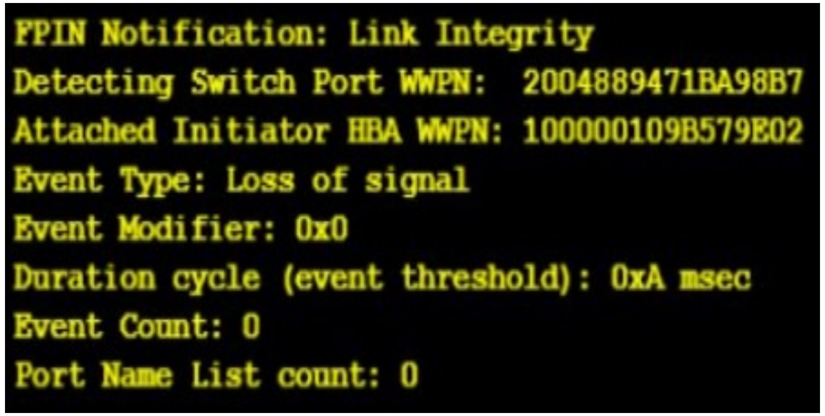
Scott Rowlands
Dell Technologies

Registration

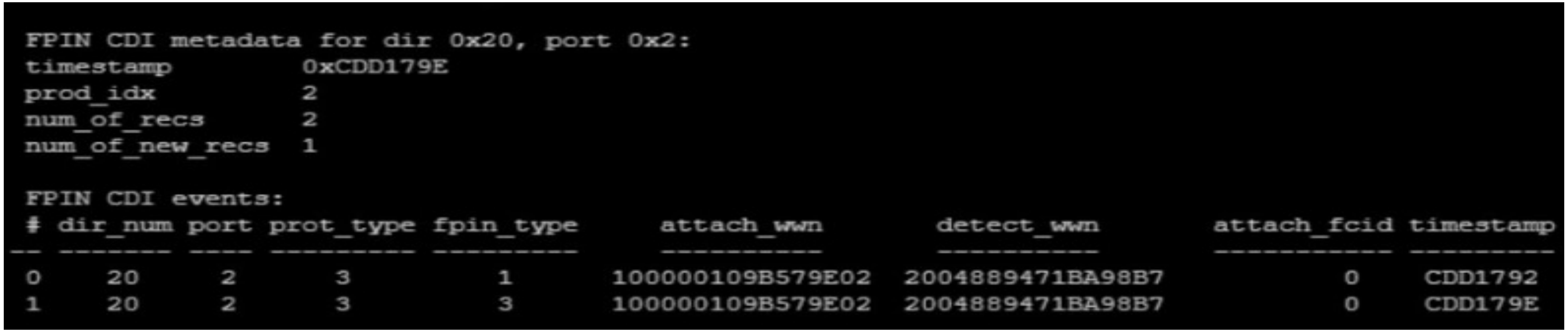
- Support can be added non-disruptively via SW update
- No user action is required to enable
- Each array port participates
- If the fabric supports FPINs, register for all notifications and signals
- User may choose to configure alerts or automatic remediation

Reaction to FPINs

- Catalog
  - Each array port keeps history of FPINs
  - Internal errors/logging for analysis/debug
  - Export to control plane



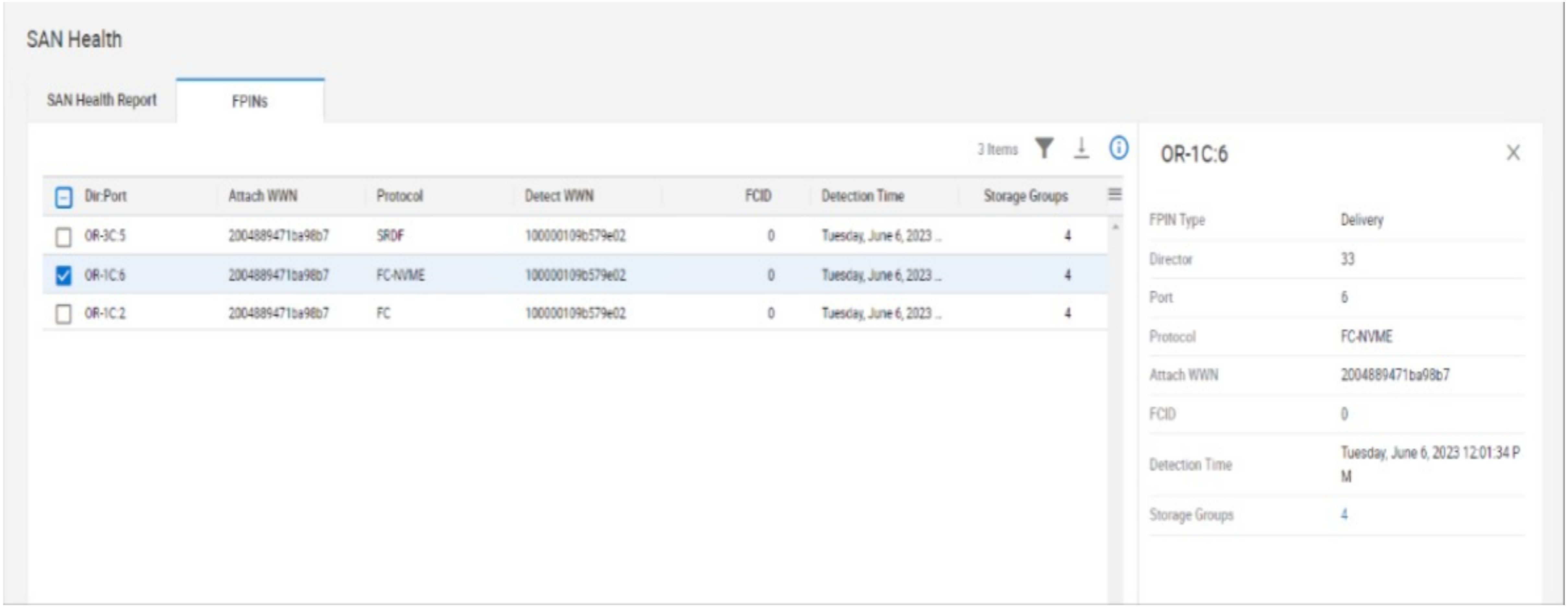
```
FPIN Notification: Link Integrity
Detecting Switch Port WWPN:  2004889471BA98B7
Attached Initiator HBA WWPN: 100000109B579E02
Event Type: Loss of signal
Event Modifier: 0x0
Duration cycle (event threshold): 0xA msec
Event Count: 0
Port Name List count: 0
```



```
FPIN CDI metadata for dir 0x20, port 0x2:
timestamp          0xCDD179E
prod_idx           2
num_of_recs        2
num_of_new_recs    1

FPIN CDI events:
# dir_num port prot_type fpin_type      attach_wwn         detect_wwn       attach_fcid timestamp
-- ------- ---- --------- ---------    -----------        -----------      ----------- ---------
0    20      2      3         1     100000109B579E02  2004889471BA98B7              0   CDD1792
1    20      2      3         3     100000109B579E02  2004889471BA98B7              0   CDD179E
```

Dashboard



SAN Health

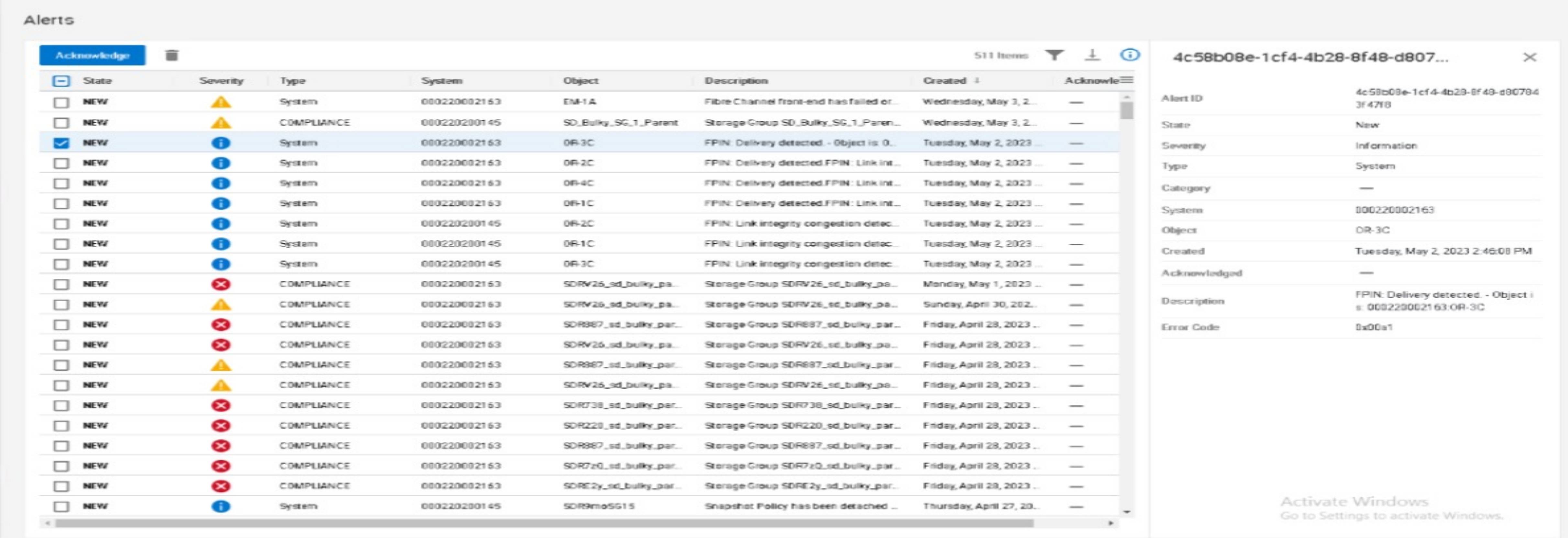
⊟	Dir:Port	Attach WWN	Protocol	Detect WWN	FCID	Detection Time	Storage Groups
□	OR-3C:5	2004889471ba98b7	SRDF	100000109b579e02	0	Tuesday, June 6, 2023 ...	4
☑	OR-1C:6	2004889471ba98b7	FC-NVME	100000109b579e02	0	Tuesday, June 6, 2023 ...	4
□	OR-1C:2	2004889471ba98b7	FC	100000109b579e02	0	Tuesday, June 6, 2023 ...	4

OR-1C:6

FPIN Type	Delivery
Director	33
Port	6
Protocol	FC-NVME
Attach WWN	2004889471ba98b7
FCID	0
Detection Time	Tuesday, June 6, 2023 12:01:34 PM
Storage Groups	4

Alerts

- Optional alerts can expedite response



Alerts

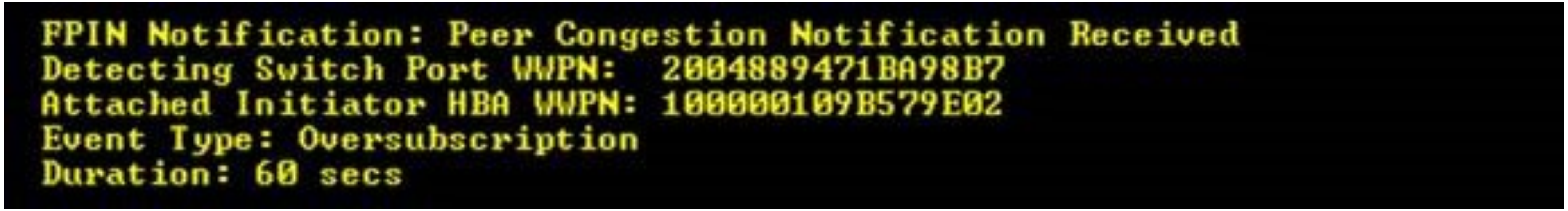
☐	State	Severity	Type	System	Object	Description	Created ↓	Acknowle
☐	NEW	⚠	System	000220002163	EM-1A	Fibre Channel front-end has failed or...	Wednesday, May 3, 2...	—
☐	NEW	⚠	COMPLIANCE	000220200145	SD_Bulky_SG_1_Parent	Storage Group SD_Bulky_SG_1_Paren...	Wednesday, May 3, 2...	—
☑	NEW	ⓘ	System	000220002163	0R-3C	FPIN: Delivery detected. - Object is: 0...	Tuesday, May 2, 2023 ...	—
☐	NEW	ⓘ	System	000220002163	0R-2C	FPIN: Delivery detected.FPIN : Link int...	Tuesday, May 2, 2023 ...	—
☐	NEW	ⓘ	System	000220002163	0R-4C	FPIN: Delivery detected.FPIN : Link int...	Tuesday, May 2, 2023 ...	—
☐	NEW	ⓘ	System	000220002163	0R-1C	FPIN: Delivery detected.FPIN : Link int...	Tuesday, May 2, 2023 ...	—
☐	NEW	ⓘ	System	000220200145	0R-2C	FPIN: Link integrity congestion detec...	Tuesday, May 2, 2023 ...	—
☐	NEW	ⓘ	System	000220200145	0R-1C	FPIN: Link integrity congestion detec...	Tuesday, May 2, 2023 ...	—
☐	NEW	ⓘ	System	000220200145	0R-3C	FPIN: Link integrity congestion detec...	Tuesday, May 2, 2023 ...	—
☐	NEW	✖	COMPLIANCE	000220002163	SDRV26_sd_bulky_pa...	Storage Group SDRV26_sd_bulky_pa...	Monday, May 1, 2023 ...	—
☐	NEW	⚠	COMPLIANCE	000220002163	SDRV26_sd_bulky_pa...	Storage Group SDRV26_sd_bulky_pa...	Sunday, April 30, 202...	—
☐	NEW	✖	COMPLIANCE	000220002163	SDR887_sd_bulky_par...	Storage Group SDR887_sd_bulky_par...	Friday, April 28, 2023 ...	—
☐	NEW	✖	COMPLIANCE	000220002163	SDRV26_sd_bulky_pa...	Storage Group SDRV26_sd_bulky_pa...	Friday, April 28, 2023 ...	—
☐	NEW	⚠	COMPLIANCE	000220002163	SDR887_sd_bulky_par...	Storage Group SDR887_sd_bulky_par...	Friday, April 28, 2023 ...	—
☐	NEW	⚠	COMPLIANCE	000220002163	SDRV26_sd_bulky_pa...	Storage Group SDRV26_sd_bulky_pa...	Friday, April 28, 2023 ...	—
☐	NEW	✖	COMPLIANCE	000220002163	SDR738_sd_bulky_par...	Storage Group SDR738_sd_bulky_par...	Friday, April 28, 2023 ...	—
☐	NEW	✖	COMPLIANCE	000220002163	SDR220_sd_bulky_par...	Storage Group SDR220_sd_bulky_par...	Friday, April 28, 2023 ...	—
☐	NEW	✖	COMPLIANCE	000220002163	SDR887_sd_bulky_par...	Storage Group SDR887_sd_bulky_par...	Friday, April 28, 2023 ...	—
☐	NEW	✖	COMPLIANCE	000220002163	SDR7z0_sd_bulky_par...	Storage Group SDR7z0_sd_bulky_par...	Friday, April 28, 2023 ...	—
☐	NEW	✖	COMPLIANCE	000220002163	SDRE2y_sd_bulky_par...	Storage Group SDRE2y_sd_bulky_par...	Friday, April 28, 2023 ...	—
☐	NEW	ⓘ	System	000220200145	SDR9moSG15	Snapshot Policy has been detached ...	Thursday, April 27, 20...	—

4c58b08e-1cf4-4b28-8f48-d807...

Alert ID	4c58b08e-1cf4-4b28-8f48-d807843f47f8
State	New
Severity	Information
Type	System
Category	—
System	000220002163
Object	OR-3C
Created	Tuesday, May 2, 2023 2:46:08 PM
Acknowledged	—
Description	FPIN: Delivery detected. - Object is: 000220002163:OR-3C
Error Code	0x00a1

Example Triage

- Customer escalates performance issue
- Array support personnel can quickly check for FPIN history system wide and for specific ports
- FPIN entries provide details about the type of event and suspect port
- CLI example



```
FPIN Notification: Peer Congestion Notification Received
Detecting Switch Port WWPN:  2004889471BA98B7
Attached Initiator HBA WWPN: 100000109B579E02
Event Type: Oversubscription
Duration: 60 secs
```

- Evidence is reported back to customer

Congestion Mitigation

- Peer (outbound)
  - Resolve Speed Mismatches
  - Throttling of reads (Initiator Specific)
    - ULP (IO) level – Helpful, but sequences still burst at line rate
    - Frame level – Better, but requires HW support
    - Application
      - Manual (support personnel typically involved)
      - Automatic – If selected, array will enable limits when congestion is detected and relieve them slowly when it clears
- Array (inbound)
  - ULP (IO) level throttling (via XFR_RDY)
  - Rebalance compute resources
  - Tune configuration
    - Add array ports to group
    - Isolate heavy duty applications

Challenges (Growing Pains)

- HW Support (switches/HBAs)
  - Often a mix of older and newer equipment
  - Updating to proper FW levels
  - Enablement
- Policy Configuration
  - Thresholds
  - Defaults changed quickly in early stages
  - Different switch models/FW-levels use different defaults
- Has Stabilized Over Time

Summary

- Fabric Notifications overview
  - Fibre Channel Architecture
- Congestion Use Case
  - Classic congestion issues
- Implementations and considerations
  - Fabric
  - Host (HBA)
  - Storage

Solutions

Fabrics and Storage

- Fabrics
  - Brocade
    - FOS 9.0.0, FOS 9.2.1
  - Cisco
    - NX-OS 9.2(1), NX-OS 9.4(2a)
  - Emulex
    - LPe3100x, LPe3200x, LPe3500x-M2
  - Marvell
    - QLE269x, QLE274x, QLE277x, QLE28xx
- Storage
  - Dell
    - PowerMax InfoScale 10.1
  - NetApp
    - OnTap 9.10
  - PureStorage
    - Oxygen

Multipath solutions

- Operating systems
  - IBM AIX
    - 7.2 TL5, 7.3 TL2
  - Redhat
    - RHEL 8.3 / EPEL 8, RHEL 9.0 / RHEL 8.7, RHEL 9.2 / RHEL 8.8, RHEL 9.3 / RHEL 8.9, RHEL9.4 / RHEL 8.10
  - SuSE
    - SLES15 SP4, SLES 15 SP5, SLES 15 SP6
  - Vmware
    - ESXi 8.0, ESXi 8.0U1, ESXi 8.0U2
- Multipath software
  - Dell
    - PowerPath 7.4
  - Veritas
    - InfoScale 8.0.2 DMP

References

- **Webinars**
  - "Introducing Fabric Notifications, From Awareness to Action" (FCIA BrightTalk presentation)
    - SNIA SDC 2021 EMEA virtual session (Part One and Part Two)
    - SNIA SDC 2021 virtual session (Presentation)
  - "Fabric Notifications – An Update from Awareness to Action"
    - SNIA SDC 2022 live session (Presentation)
  - "Fibre Channel Gen8 Update - 128GFC, Fabric Notifications, and Managing NVMe NQNs"
    - SNIA SDC 2024 live session (Presentation)
- **Industry**
  - IBM Power Community – AIX Support for Fabric Congestion Notification
  - PureStorage blog
  - Marvell SAN congestion mitigation Video
- **Articles**
  - The Autonomous SAN (FCIA Solutions guide)
  - Fabric Notifications Technical Brief (Brocade Whitepaper)
  - MPIO Load Balancing Recommendations (Brocade Whitepaper)
  - Cisco Fabric Notifications Blog
  - Dell Fabric Notifications Technical Brief
  - Emulex Fabric Notifications Product Brief
- **Videos**
  - Fabric Notifications Primer (Brocade video)
  - Fabric Notifications using RHEL 8.3 (Brocade video)
  - Fabric Notifications using IBM AIX 7.2 TL5 (Brocade video)

Q&A

After this Webinar

- Please rate this webinar and provide us with your feedback
- This webinar and a copy of the slides are available at the SNIA Educational Library https://www.snia.org/educational-library
- A Q&A from this webinar, including answers to questions we couldn't get to today, will be posted on our blog at https://sniansfblog.org/
- Follow us on X/Twitter @SNIANSF

Thank You