BROCADE

FC-SW-6
LINK LENGTH PROPOSALS

Howard L. Johnson

Dave Peterson

T11/13-133v1

Table of Contents

Link Length Proposal

- Summary of FC-BB-6 Proposed Text
- Review of Proposed Method
- Discuss Action

FC-BB-6

Proposed Text Changes

FC-BB-6 (Current Text)

4.4.4 QoS and bandwidth

- FC-BB_IP recommends that some form of preferential QoS be used for the FCIP traffic in the IP network to minimize latency and packet drops although no particular form of QoS is recommended. See RFC 3821.
- FC-BB_GFPT has no specific transport service requirements.
- FC-BB_PW recommends that Primitive Sequences are carried with low latency and no loss over the MPLS network. In addition to these properties, FC data traffic should be provided with assurance of some amount of bandwidth, however no specific recommendation is made in this standard. The Differentiated Services EF PHB (see RFC 3246) is an example of a mechanism that may be used for FC-BB_PW traffic management.
- FC-BB_E is intended to operate over an Ethernet network that does not discard frames in the presence of congestion. Such an Ethernet network is called Lossless Ethernet in this standard. Lossless Ethernet may be implemented through the use of some Ethernet extensions. A possible Ethernet extension to implement Lossless Ethernet is the PAUSE mechanism defined in IEEE 802.3-2008. Another possible Ethernet extension to implement Lossless Ethernet is the Priority-based Flow Control (PFC) mechanism defined in IEEE 802.1Qbb. When PFC is used to implement Lossless Ethernet, FCoE frames shall use a lossless priority (see IEEE 802.1Qbb).

FC-BB-6 (Proposed Text)

4.4.4 QoS and bandwidth

- FC-BB_IP recommends that some form of preferential QoS be used for the FCIP traffic in the IP network to minimize latency and packet drops although no particular form of QoS is recommended. See RFC 3821.
- FC-BB_GFPT has no specific transport service requirements.
- FC-BB_PW recommends that Primitive Sequences are carried with low latency and no loss over the MPLS network. In addition to these properties, FC data traffic should be provided with assurance of some amount of bandwidth, however no specific recommendation is made in this standard. The Differentiated Services EF PHB (see RFC 3246) is an example of a mechanism that may be used for FC-BB_PW traffic management.
- FC-BB_E is intended to operate over an Ethernet network that does not discard frames in the presence of congestion. Such an Ethernet network is called Lossless Ethernet in this standard. Lossless Ethernet may be implemented through the use of some Ethernet extensions. Suitable extensions include: the PAUSE mechanism defined in IEEE 802.3-2088, or the Priority-based Flow Control (PFC) mechanism defined in IEEE 802.1Qbb; where, FCoE frames shall use a lossless priority (see IEEE 802.1Qbb). The Precision Time Protocol (PTP) mechanism may be used to determine the link latency (see IEEE 1588- 2008 or IEEE 802.1AS).

FC-BB-6 (Current Text)

7.2 FC-BB_E overview

- This clause discusses aspects of the FC-BB_E mapping, including initialization and procedures for the mapping of Fibre Channel frames over Ethernet.
- Figure 28 shows how FC-BB_E maps the Fibre Channel levels and sublevels over IEEE 802.3 layers.
- *Figure 28*
- Figure 29 shows how the FC-BB_E mapping applies to FCoE Forwarders (FCF) and FCoE Nodes (ENodes).
- *Figure 29*
- FC-BB_E defines a direct mapping of Fibre Channel over Ethernet (FCoE). Although a generic Ethernet network may lose frames due to congestion, a proper implementation of appropriate Ethernet extensions (e.g., the PAUSE mechanism defined in IEEE 802.3-2008) allows a full duplex Ethernet link to provide a lossless behavior equivalent to the one provided by the buffer-to-buffer credit mechanism (see FC-FS-3). The protocol mapping defined by FC-BB_E is referred to as Fibre Channel over Ethernet (FCoE) and shall use an underlying Ethernet layer (i.e., composed only of full duplex links and providing a lossless behavior when carrying FCoE frames (see 4.4.4)). The Lossless Ethernet layer provides sequential delivery of FCoE frames.
- *Remainder of clause*

FC-BB-6 (Proposed Text)

7.2 FC-BB_E overview

- This clause discusses aspects of the FC-BB_E mapping, including initialization and procedures for the mapping of Fibre Channel frames over Ethernet.
- Figure 28 shows how FC-BB_E maps the Fibre Channel levels and sublevels over IEEE 802.3 layers.
- *Figure 28*
- Figure 29 shows how the FC-BB_E mapping applies to FCoE Forwarders (FCF) and FCoE Nodes (ENodes).
- *Figure 29*
- FC-BB_E defines a direct mapping of Fibre Channel over Ethernet (FCoE). Although a generic Ethernet network may lose frames due to congestion, a proper implementation of appropriate Ethernet extension (~~e.g., the PAUSE mechanism defined in IEEE 802.3-2008~~ see 4.4.6) allows a full duplex Ethernet link to provide a lossless behavior equivalent to the one provided by the buffer-to-buffer credit mechanism (see FC-FS-3). The protocol mapping defined by FC-BB_E is referred to as Fibre Channel over Ethernet (FCoE) and shall use an underlying Ethernet layer (i.e., composed only of full duplex links and providing a lossless behavior when carrying FCoE frames (see 4.4.4). The Lossless Ethernet layer provides sequential delivery of FCoE frames.
- *Remainder of clause*

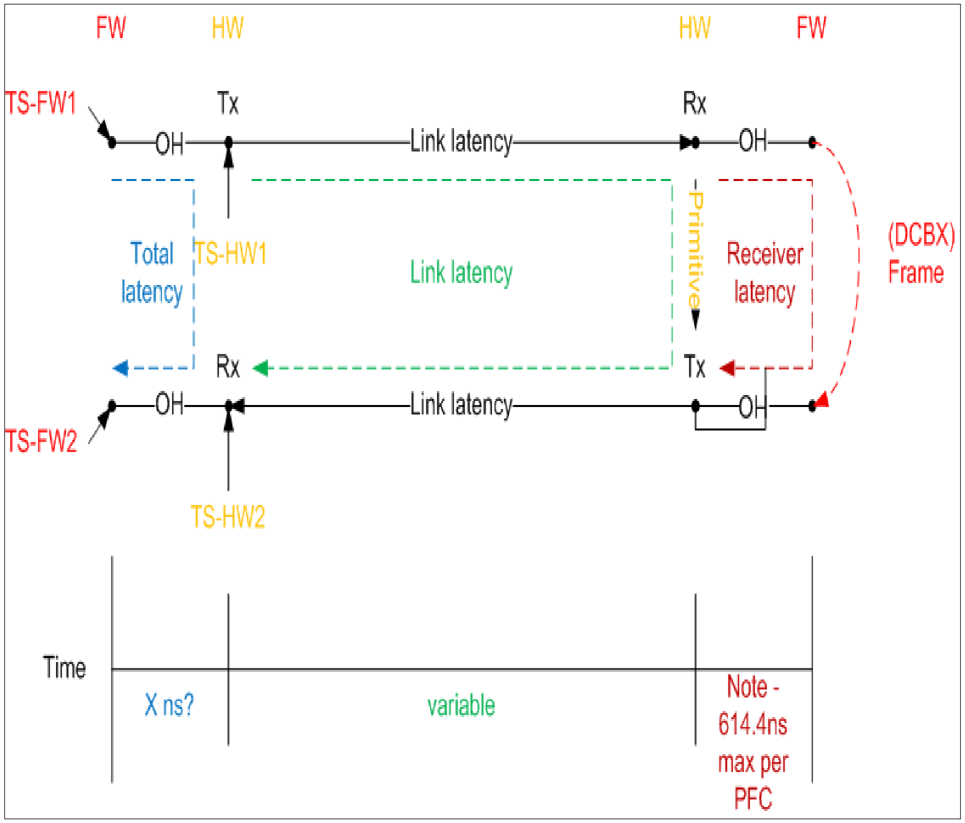
FC-SW-6

Proposed Method

Proposal

Link Length Determination Objectives

- Use low level mechanism
  - Point to point measurement
  - Consistency for realistic time
- Determine link length
  - MARK primitive
- Use results to establish buffer memory resources
  - BB Credits for FC
  - Receiver memory for FCoE
  - Pause/resume thresholds
- Provide notification for unsupportable link lengths



FC-SW Link Length Determination

Skew and Roundtrip Delay using MARK primitive

- FC-AL-2 MARK primitive
  - MARK primitive translates into KD character K28.5, D31.2
  - Character is 0xBC5F0000
- Measurements
  - Link round trip measurement (0xBC5FFF01)
    - MARK sent with timer active
    - MARK receiver simply returns the MARK immediately
    - The timer stops when transmitting side receives the returned MARK
    - The round trip measurement is recorded
  - Link skew (0xBC5F0000)
    - MARK sent out of multiple parallel links simultaneously
    - The receiver side clocks in the differences between the links
    - Executed on both sides to eliminate one-way skews that the round trip skew measurement could not detect

FC-BB Link Length Determination

Roundtrip Delay using MARK primitive

- Non-standard Ethernet Ordered Set
  - MARK primitive (0xBC5F0000)
    - MARK primitive translates into KD char is K28.5, D31.2
- Measurements
  - Link round trip measurement (0xBC5F0000)
    - MARK sent with timer active
    - MARK receiver simply returns the MARK immediately
    - The timer stops when transmitting side receives the returned MARK
    - The round trip measurement is recorded

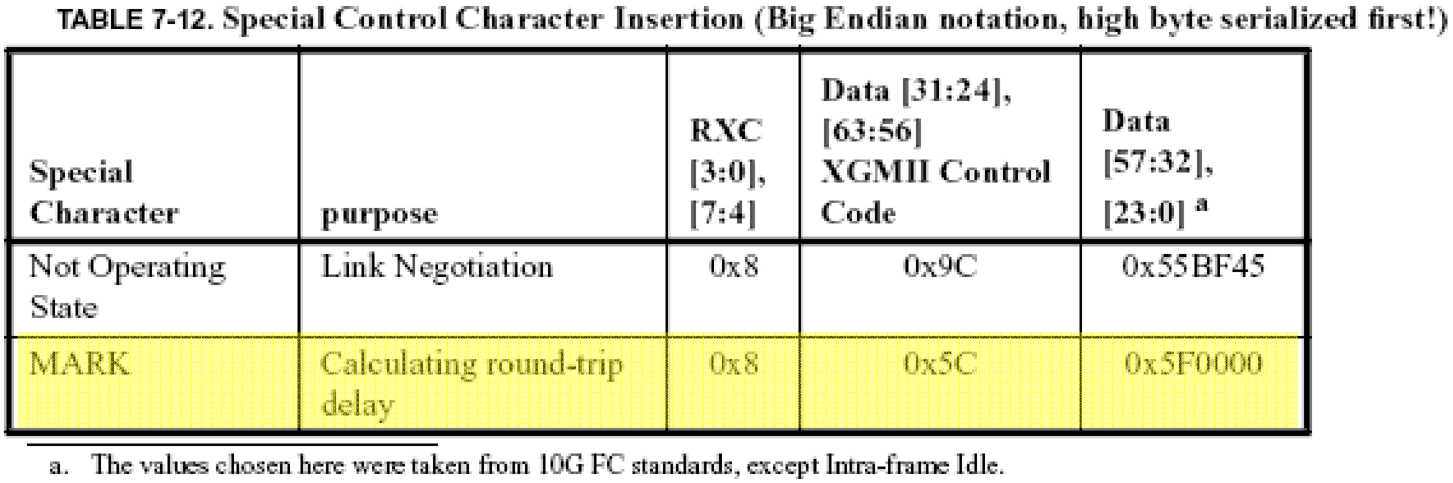


TABLE 7-12. Special Control Character Insertion (Big Endian notation, high byte serialized first!)

Special Character	purpose	RXC [3:0], [7:4]	Data [31:24], [63:56] XGMII Control Code	Data [57:32], [23:0] [a]
Not Operating State	Link Negotiation	0x8	0x9C	0x55BF45
MARK	Calculating round-trip delay	0x8	0x5C	0x5F0000

[a] The values chosen here were taken from 10G FC standards, except Intra-frame Idle.

Next Steps

Proposal

- Interest from group

- Move forward with detailed proposal

BROCADE

This slide intentionally left blank

Thank You

Reference

Slides from February 2012 Meeting

Summary

Improvements for handling long links

- Concern
  - Handling links longer than supported by configured buffer capacity
- An Opportunity to improve the state of the art
  - Provide link length/latency determination and validation
    - Executed at link initialization
    - Exchange of buffer depths at or prior to Fabric Login
    - Provide feedback for incorrectly configured resources
  - Technique
    - Use low level primitive for accurate point to point measurements
- Agreement
  - Recommend use of PTP
- Standard Reference
  - Precision Time Protocol IEEE 1588

Existing Text

FC-BB-6 4.4.4 QOS and Bandwidth

- *"FC-BB_E is intended to operate over an Ethernet network that does not discard frames in the presence of congestion. Such an Ethernet network is called Lossless Ethernet in this standard. Lossless Ethernet may be implemented through the use of some Ethernet extensions. A possible Ethernet extension to implement Lossless Ethernet is the PAUSE mechanism defined in IEEE 802.3-2008. Another possible Ethernet extension to implement Lossless Ethernet is the Priority-based Flow Control (PFC) mechanism defined in IEEE 802.1Qbb. When PFC is used to implement Lossless Ethernet, FCoE frames shall use a lossless priority (see IEEE 802.1Qbb)."*

Proposed Text

FC-BB-6 4.4.4 QOS and Bandwidth

- *Text*
  - *"… as well as the latency determination mechanism defined in the Precision Time Protocol IEEE 1588 …"*
- *Placement*
  - *"FC-BB_E is intended to operate over an Ethernet network that does not discard frames in the presence of congestion. Such an Ethernet network is called Lossless Ethernet in this standard. Lossless Ethernet may be implemented through the use of some Ethernet extensions. A possible Ethernet extension to implement Lossless Ethernet is the PAUSE mechanism defined in IEEE 802.3-2008* as well as the latency determination mechanism defined in the Precision Time Protocol IEEE 1588*. Another possible Ethernet extension to implement Lossless Ethernet is the Priority-based Flow Control (PFC) mechanism defined in IEEE 802.1Qbb. When PFC is used to implement Lossless Ethernet, FCoE frames shall use a lossless priority (see IEEE 802.1Qbb)."*

Existing Text

FC-BB-6 7.2 FC-BB_E Overview

- *"FC-BB_E defines a direct mapping of Fibre Channel over Ethernet (FCoE). Although a generic Ethernet network may lose frames due to congestion, a proper implementation of appropriate Ethernet extensions (e.g., the PAUSE mechanism defined in IEEE 802.3-2008) allows a full duplex Ethernet link to provide a lossless behavior equivalent to the one provided by the buffer-to-buffer credit mechanism (see FC-FS-3). The protocol mapping defined by FC-BB_E is referred to as Fibre Channel over Ethernet (FCoE) and shall use an underlying Ethernet layer (i.e., composed only of full duplex links and providing a lossless behavior when carrying FCoE frames (see 4.4.4)). The Lossless Ethernet layer provides sequential delivery of FCoE frames."*

Proposed Text

FC-BB-6 7.2 FC-BB_E Overview

- Text
  - *"… and the latency determination mechanism defined in the Precision Time Protocol IEEE 1588 …"*
- Placement
  - *"FC-BB_E defines a direct mapping of Fibre Channel over Ethernet (FCoE). Although a generic Ethernet network may lose frames due to congestion, a proper implementation of appropriate Ethernet extensions (e.g., the PAUSE mechanism defined in IEEE 802.3-2008 and the* latency determination mechanism defined in the Precision Time Protocol IEEE 1588*) allows a full duplex Ethernet link to provide a lossless behavior equivalent to the one provided by the buffer-to-buffer credit mechanism (see FC-FS-3). The protocol mapping defined by FC-BB_E is referred to as Fibre Channel over Ethernet (FCoE) and shall use an underlying Ethernet layer (i.e., composed only of full duplex links and providing a lossless behavior when carrying FCoE frames (see 4.4.4)). The Lossless Ethernet layer provides sequential delivery of FCoE frames."*

Informative Annex

Link Latency Determination

- Describe concern
- Define objectives and methods
- Provide sample solution using PTP IEEE 1588

Reference

Slides from December 2011 Meeting

Link Length

Administration

- Fibre Channel
  - Optimal configuration is multidimensional
    - Link speed, average payload size, and link length
  - Administrative configuration is imperfect
    - Errors result in under utilized links or wasted buffer resources
- Ethernet
  - Optimal configuration is multidimensional
    - Link speed and link length
    - Pause buffering capacity, I/O Consolidation, Convergence
  - Administrative configuration is imperfect
    - Errors may result in wasted buffer resources or potentially lost frames on a congested link
- User Feedback
  - Clamoring for this “little stuff” to be automatic

GRAHAM CHAPMAN JOHN CLEESE TERRY GILLIAM
ERIC IDLE TERRY JONES MICHAEL PALIN
MONTY PYTHON'S
And Now For Something
COMPLETELY DIFFERENT
DVD
The BEST of Monty Python's Flying Circus

Link Length

Automated Buffer Resource Determination

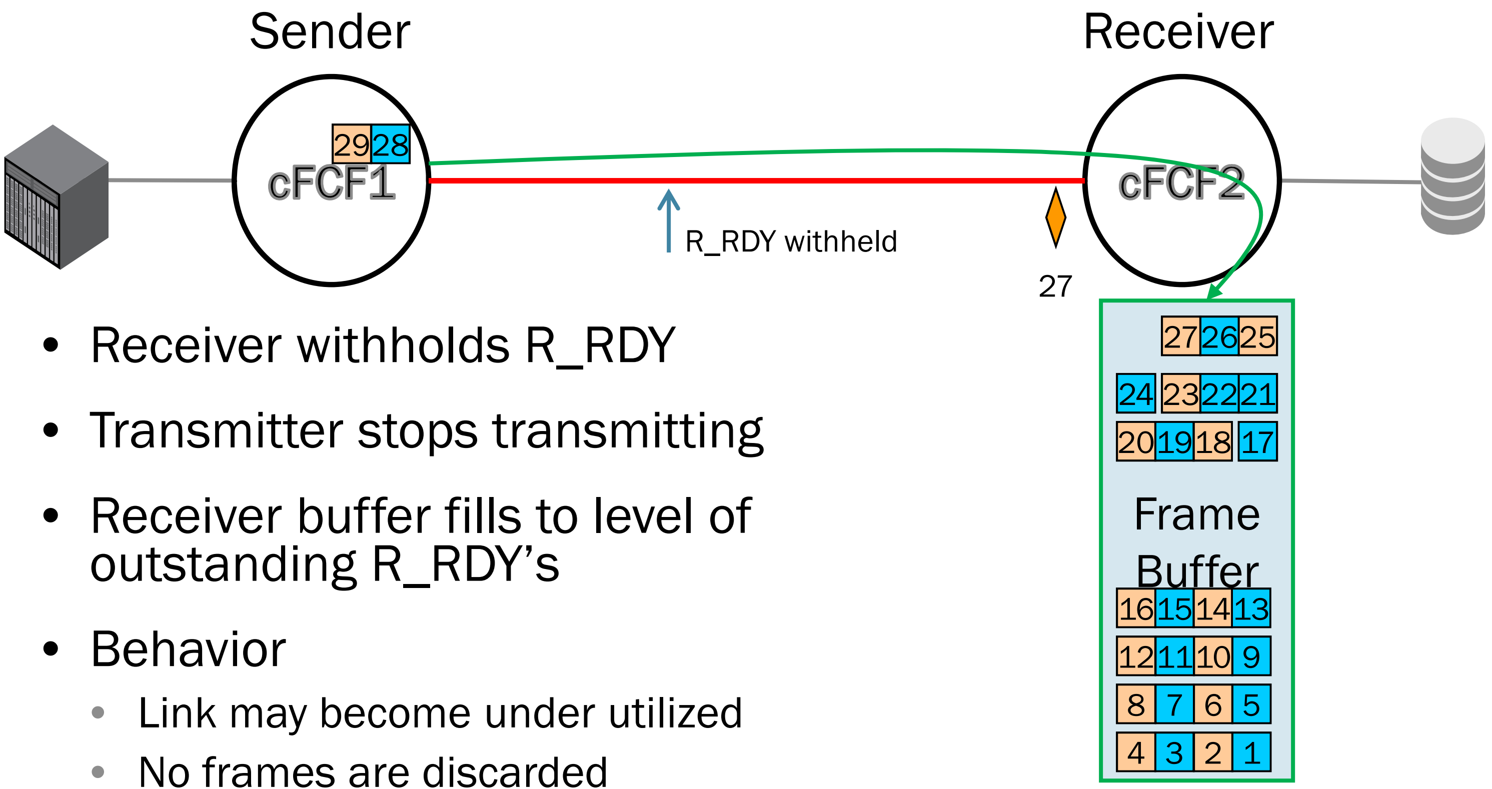
Mechanisms

Well known, effective processes

- Proactive method (e.g. credit based)
  - Exchange of buffering capacity at discovery
  - Transmitter responsible for managing flow
- Reactive method (e.g. pause based)
  - Receiver buffer is configured to accommodate bandwidth delay
  - Receiver responsible for managing flow
- Concern is around a particular error scenario
  - Link longer than configured buffering resources
  - Unable to forward frames due to congestion
  - Errors occurring intermittently – just a little bit mis-configured
    - Receiver overrun is a possible behavior of a properly functioning but incorrectly configured device

Proactive Behavior during Error Scenario

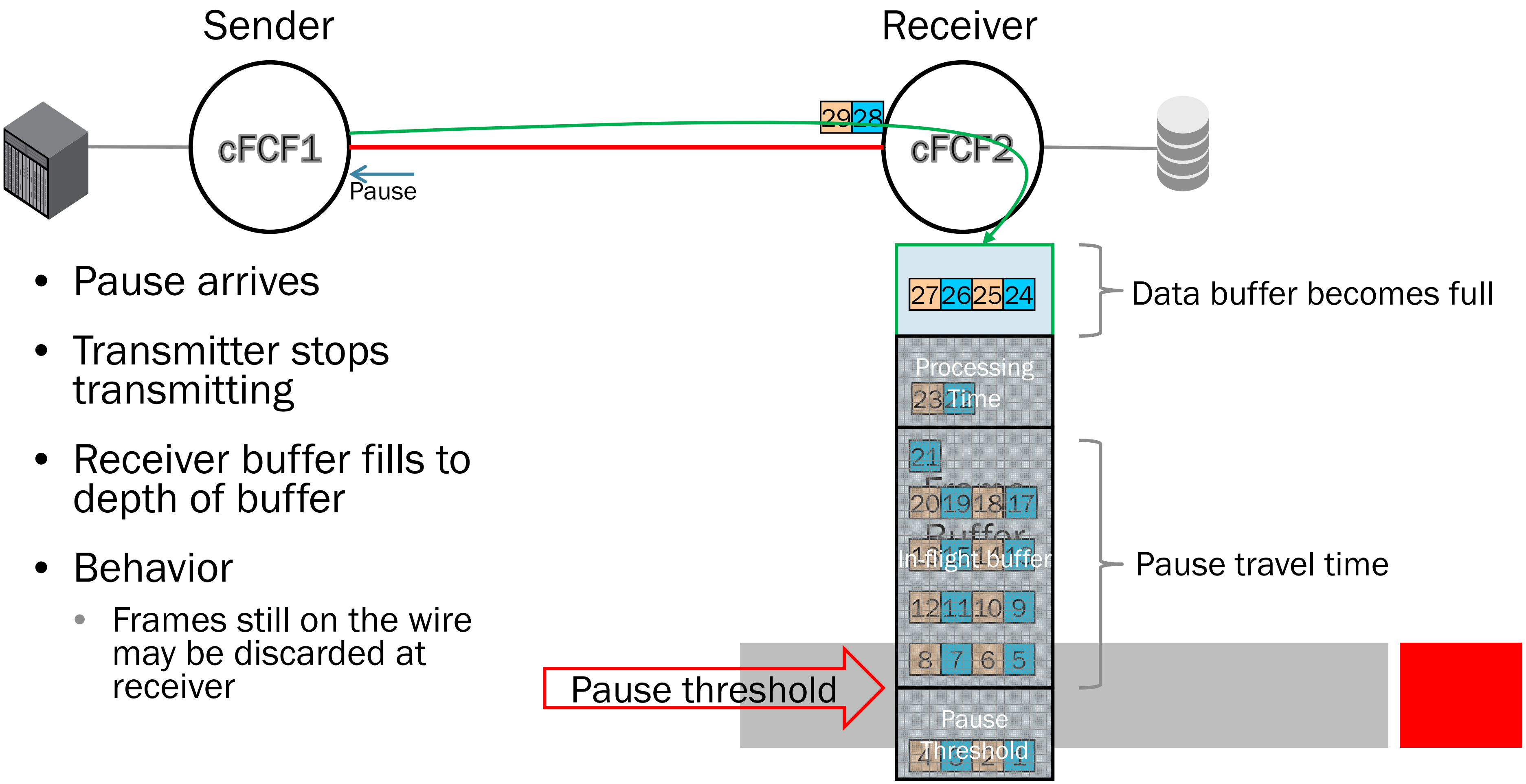
Link is longer than configured credit capacity



- Receiver withholds R_RDY
- Transmitter stops transmitting
- Receiver buffer fills to level of outstanding R_RDY's
- Behavior
  - Link may become under utilized
  - No frames are discarded
  - Consistent complaint among traditional FC users

Reactive Behavior during Error Scenario

Link is longer than configured receive buffer depth



Data buffer becomes full

Pause travel time

Pause threshold

- Pause arrives
- Transmitter stops transmitting
- Receiver buffer fills to depth of buffer
- Behavior
  - Frames still on the wire may be discarded at receiver

Error Scenario Summary

The link is longer than the configured buffer resources support

- Proactive flow control (credits)
  - Transmitter throttles flow
  - Link capacity may be under-utilized
  - Device is not directly aware of performance penalty
  - Must be administratively corrected
- Reactive flow control (pause)
  - Receiver throttles flow
  - Overflow frames may be discarded
  - Device level recovery employed
  - Must be administratively corrected

Fibre Channel: The Once and Future King

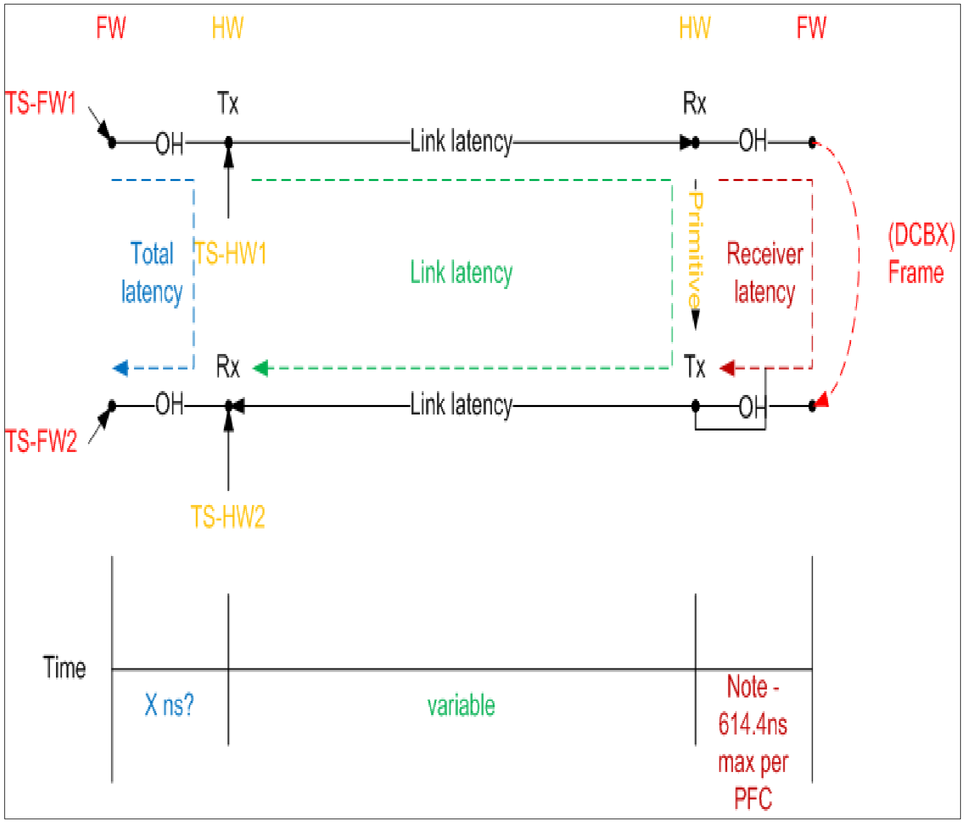
A chance to improve the state of the art

- Concern is with infrequent error scenarios
  - Difficult to discover and correct
  - Dependent on device level recovery
    - In some environments, frame discards are elevated to service calls
    - Under-utilization is even harder to figure out
- FCoE adoption
  - Differences in “corner case” behavior could hinder acceptance of FCoE
  - The invocation of device level recovery could create a negative perception of FCoE
- Opportunity
  - Improve the state of the art by adding link length determination

Proposal

Link Length Determination Objectives

- Use low level mechanism
  - Point to point measurement
  - Consistency for realistic time
- Determine link length
  - The specific method is TBD
- Use results to establish buffer memory resources
  - BB Credits for FC
  - Receiver memory for FCoE
  - Pause/resume thresholds
- Provide notification for unsupportable link lengths



Summary

Improvements for handling long links

- Concern
  - Handling links longer than supported by configured buffer capacity
- An Opportunity to improve the state of the art
  - Provide link length/latency determination and validation
    - Executed at link initialization
    - Exchange of buffer depths at or prior to Fabric Login
    - Provide feedback for incorrectly configured resources
  - Technique
    - Use low level primitive for accurate point to point measurements
- Consideration
  - What link lengths should be supported by FCoE?
- Standards Process
  - Can FC-BB request such changes?
  - Can FC-BB specify a particular solution?

BROCADE

This slide intentionally left blank

Thank You