T11 FC-SW-6

# Out of Order Can Happen

Patty Driever

# Background

- We agreed in 13-134v0 that NPRDs (routing distributions) have priority over N_Ports joining or leaving the fabric (NPZDs)

- We agreed in 13-057v0 that a distribution tree would be used to control the distribution of VA_Port SW_ILS commands to the FDFs

  - It was asserted that such a distribution tree would prevent SW_ILSs from being received in a different order than the order in which they were sent
  - Sequence number descriptor fields were removed from NPRDs and AZADs, ***but*** left in NPZDs and managed on a per FCDF basis and rules for handling 'out-of-order' NPZDs were documented

# Background

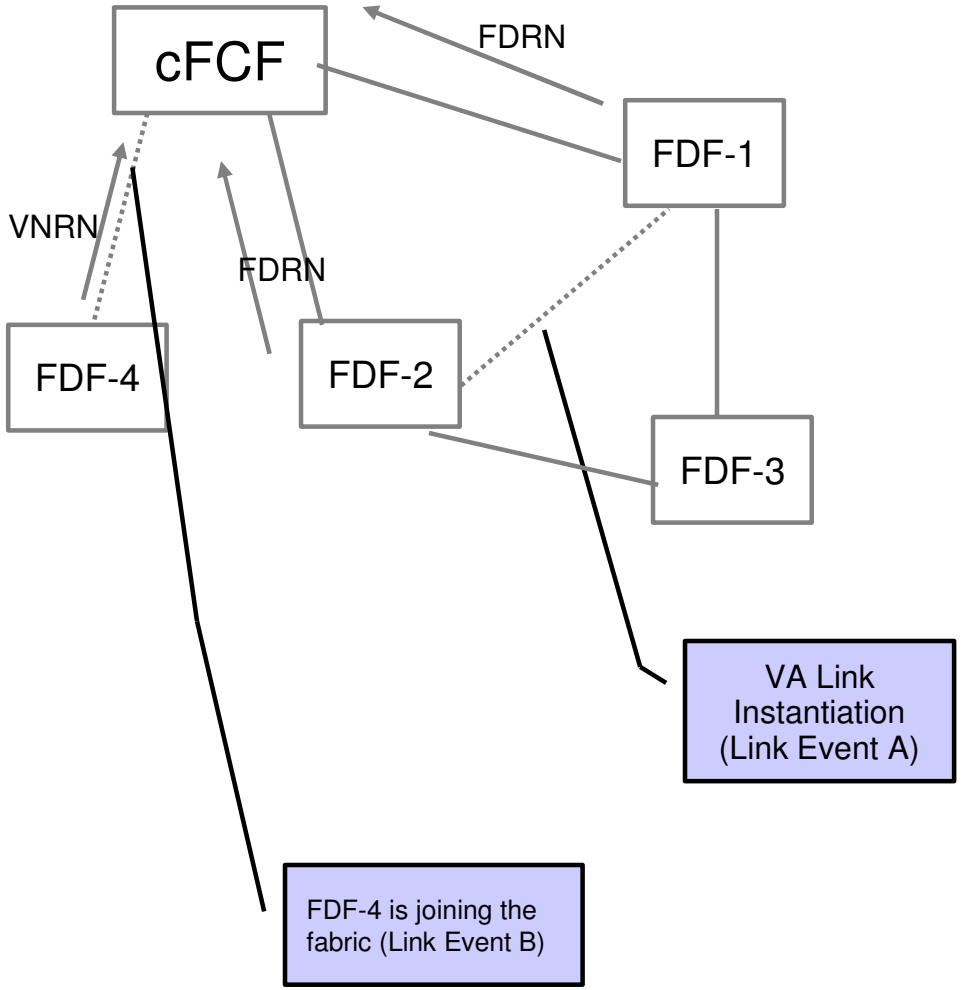
- In response to FC-BB-6 letter ballot comments IBM-H1 and Juniper-006, the following change was made to FC-BB-6:

  "FC-BB-E devices shall provide in order delivery of FCoE frames on at least a per Exchange basis within the Lossless Ethernet network."

- Fibre Channel does not guarantee in-order delivery across exchanges
  - Implementations are known that do not guarantee such in-order delivery
  - FC-LS-3 specifically declares that "The ordering relationship and deliverability of Sequences between two separate Exchanges is outside the scope of this standard"
  - In Link Aggregation implementations that hash on an exchange basis, different exchanges can flow on different physical links, potentially arriving out-of-order at the target with respect to each other

- In the presence of error conditions (e.g. links going up/down...leading to routing changes), out-of-order can also happen

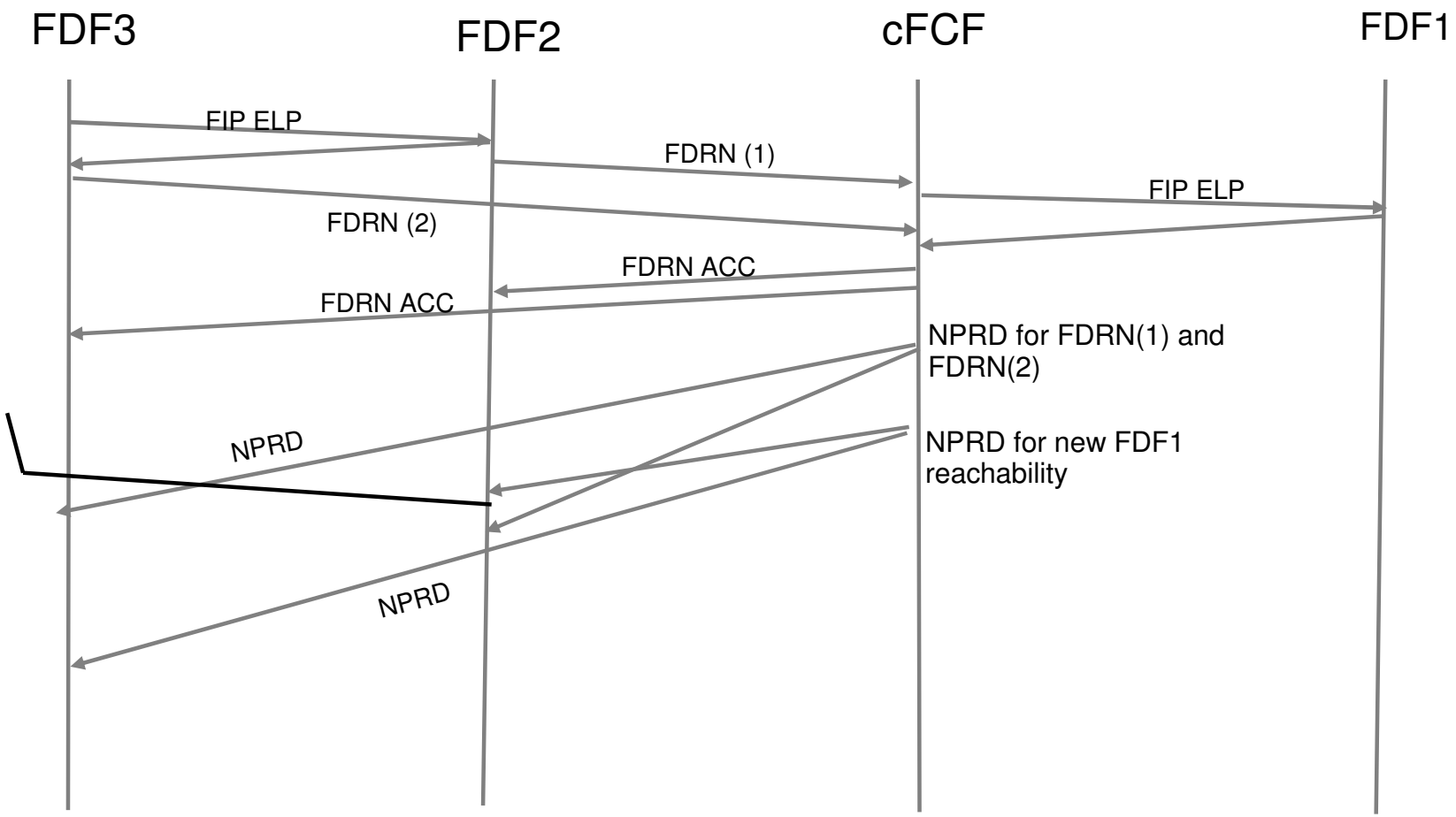
# NPRDs sent in succession




Multiple NPRD-generating events occur in close succession from the cFCF's perspective:

- The VA link between FDF-1 and FDF-2 is instantiated, reported via FDRNs from both affected FDFs to the cFCF
- The VA link between FDF-4 and the cFCF occurs via FIP ELP just after the previous FDRNs are received (i.e. FDF-4 joins the fabric)

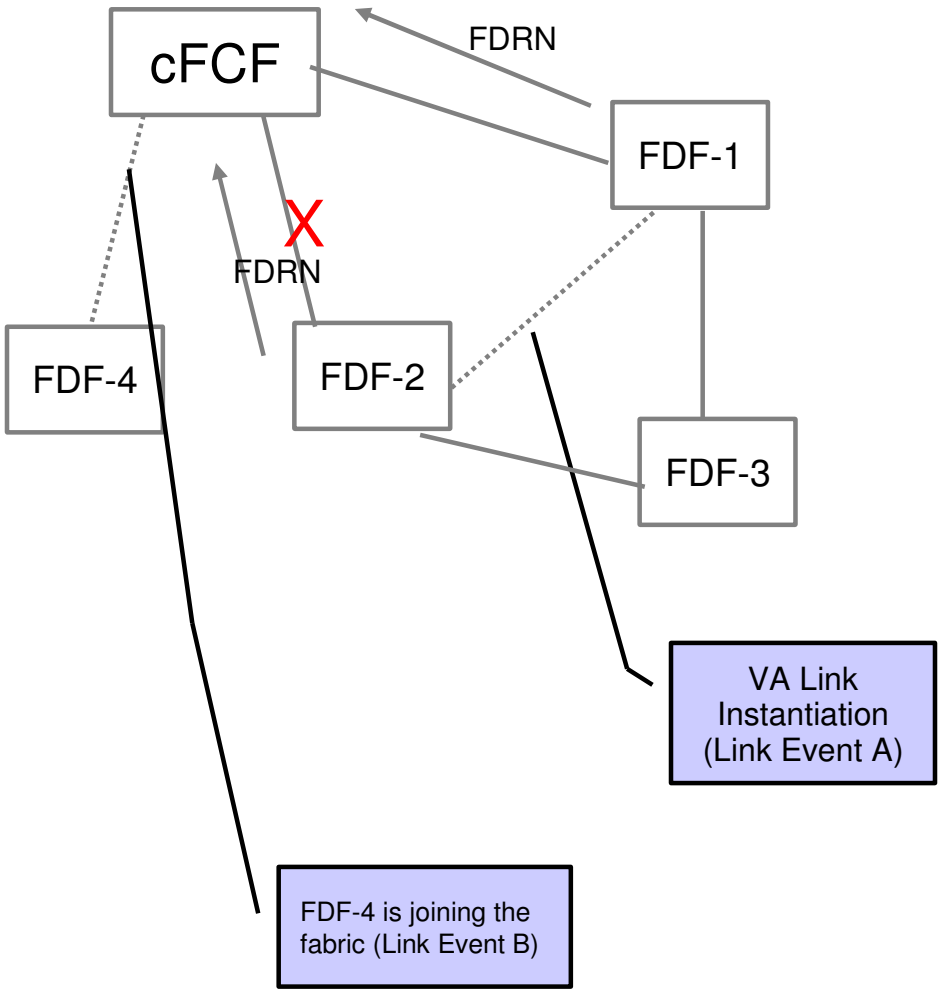
# NPRDs Arriving Out-of-order in the Absence of an Error




Older NPRD arrives last, wiping out new FDF1 reachability information.

FDF2 does not have reachability information of FDF1.

# NPRDs Arriving Out-of-order in the Presence of an Error




The distribution tree indicates that the primary (least cost) path to FDF-3 goes through FDF-2

In between sending the two required NPRDs to each FDF in the fabric, the link between FDF-2 and the cFCF goes down

If the Link Event A NPRD made it across the wire before the link failed, since the Link Event B NPRD takes a different path, it's possible that conditions in the fabric are such that the second arrives before the first

- Easier to conceive of such timings in larger cascaded fabrics

Bottom line:

- When two different paths are used to send NPRDs in succession, the NPRDs can arrive out-of-order
- Rules must exist to handle this potential out-of-order case

# Assertions

- Blanket serialization mandating that all NPRDs wait for the previous NPRDs to complete unnecessarily slows down system processing of link up/down-related events that affect fabric routing tables
    - The idea of complete serialization of NPZDs was rejected, leaving sequence numbers in place for NPZDs, along with rules for handling such ‘out-of-order’ events
- Sequence number descriptor could be placed back in the NPRD command descriptor, and a set of rules for detection and handling such out-of-order events could be put in place
    - Each NPRD sends a new complete routing distribution tree, so if multiple NPRDs are sent the LAST one sent must be processed LAST

# Relevant Prior Spec Changes

13-057 added at the end of section 17.9.2:

"The distribution of NPRD Requests shall take precedence over the distribution of AZAD and NPZD Requests."

And before the last paragraph of 17.9.3:

"The Distribution of AZAD Requests shall take precedence over the distribution of NPZD Requests."

What does 'take precedence' mean?

Applies to when such commands are issued (distributed), but does not necessarily require strict serialization regarding the order in which they must be ***received*** and ***processed***

13-057 provided text for rules for handling out-of-order NPZD requests

13-057 removed the Sequence Number Descriptor field and its associated description from tables 240 and 244 (for NPRDs and AZADs)

# Proposed Changes

Proposed Text Changes:

1) Leave Sequence Number Descriptor fields as currently documented in current draft version of FC-SW-6 (13-047v0)
   - Includes Sequence Number Descriptors in NPRDs and AZADs

2) Modify the text added to section 17.9.2 by 13-057 as described on following chart

# Proposed 17.9.2 Modified Text

- The Primary Controlling Switch maintains two sequence numbers for each FCDF in the FCDF Set. One sequence number is incremented by one and included in the NPZD sequence number descriptor each time an NPZD Request is sent, and the second is incremented by one and included in the NPRD or AZAD sequence number descriptor each time an NPRD or AZAD request is sent, respectively.
- Upon receipt of an NPZD Request, an FCDF compares the sequence number in the received sequence number descriptor to that of the last processed NPZD Request, or to 00000000 00000001h if no NPZD has previously been processed. If the received sequence number is lower, except in the case where a sequence number wrap condition has been detected, the NPZD request shall be discarded and a VA_RJT shall be sent with Reason Code of 'Logical Error' and Reason Code Explanation of 'Out of Order'. If the received sequence number is higher or a wrap condition has been detected, then the NPZD is processed.
- An FCDF considers an N_Port_ID to be allocated when it has successfully received the N_Port_ID in an Allocation Entry of the current or previous NPZD Request. If an NPZD Request contains a peering entry with a Principal N_Port_ID that has not been allocated, that entire peering entry shall be ignored. If an NPZD Request contains a peering entry with a Principal N_Port_ID that is currently allocated, but that peering entry contains Peer N_Port_ID(s) that have not been allocated, then those Peer N_Port_ID(s) shall be ignored.
- Whenever an NPZD Request is retransmitted for any reason (e.g., timeout) the Zoning ACLs for the affected N_Port_IDs shall be recomputed and a new NPZD Request including a new sequence number and the newly computed peering entries shall be sent.
- If a Primary Controlling Switch receives a VA_RJT with a Reason Code of 'Logical Error' and Reason Code Explanation of 'Out of Order' in response to an NPZD Request, the Primary Controlling Switch shall retransmit the NPZD Request.

# Proposed 17.9.2 Modified Text

- Upon receipt of an NPRD Request or an AZAD Request, an FCDF compares the sequence number in the received sequence number descriptor to that of the sequence number in the last processed NPRD or AZAD Request, or to 00000000 00000001h if no NPRD or AZAD has previously been processed. If the received sequence number is lower, except in the case where a sequence number wrap condition has been detected, the NPRD or AZAD request shall be discarded. If the received sequence number is higher or a wrap condition has been detected, then the NPRD is processed.
- Whenever an NPRD Request is retransmitted for any reason (e.g., timeout) the Routing Distribution Tree for the affected FCDFs shall be recomputed and a new NPRD Request including a new sequence number and the newly computed Reachability Descriptors shall be sent.
- Whenever an AZAD Request is retransmitted for any reason (e.g., timeout) the Zoning ACLs for the new Zone Set shall be recomputed and a new AZAD Request including a new sequence number and the new Zone enforcement rules shall be sent..

# Concern

- Just as NPRDs can be received out-of-order with respect to each other under such error circumstances, so can NPRDs and NPZDs be received out-of-order with respect to each other (reference 13-134v0)
  - The content of NPRDs and NPZDs are distinct EXCEPT that the NPRD contains an N_Port_ID Reachability Descriptor with a list of allocated N_Port_IDs
  - If the FCDF is supposed to validate this information each time it is received and update it's allocation tables based on the information contained therein (as has been stated in the past), then if an NPRD is received out-of-order with respect to an NPZD, the potential exists that such updates could overwrite the latest information.
    - So is the value of this N_Port_ID reachability information limited to the case where it's the first NPRD that an FDF receives as part of joining the fabric?

# Thank you