T11 FC-SW-6

Out of Order Can Happen 13-227v2

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

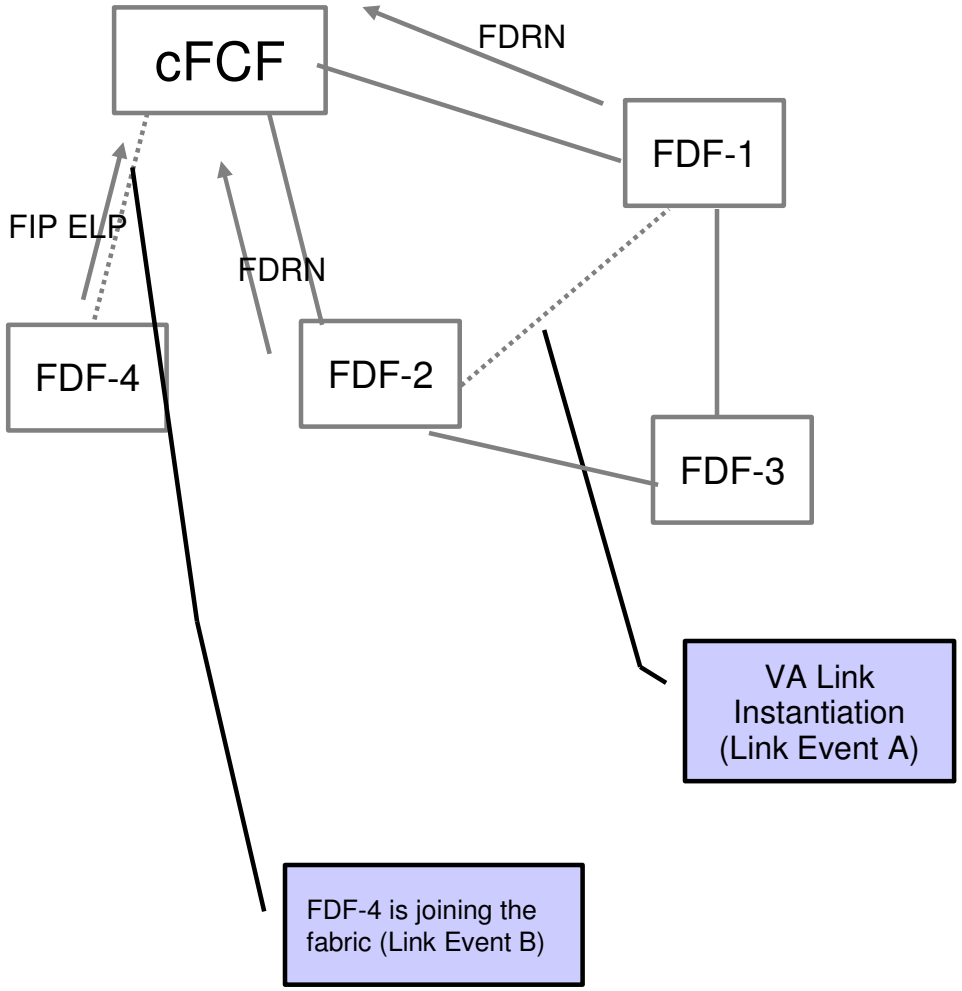
- In response to FC-BB-6 letter ballot comments IBM-H1 and Juniper-006, the following change was made to FC-BB-6:

  "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

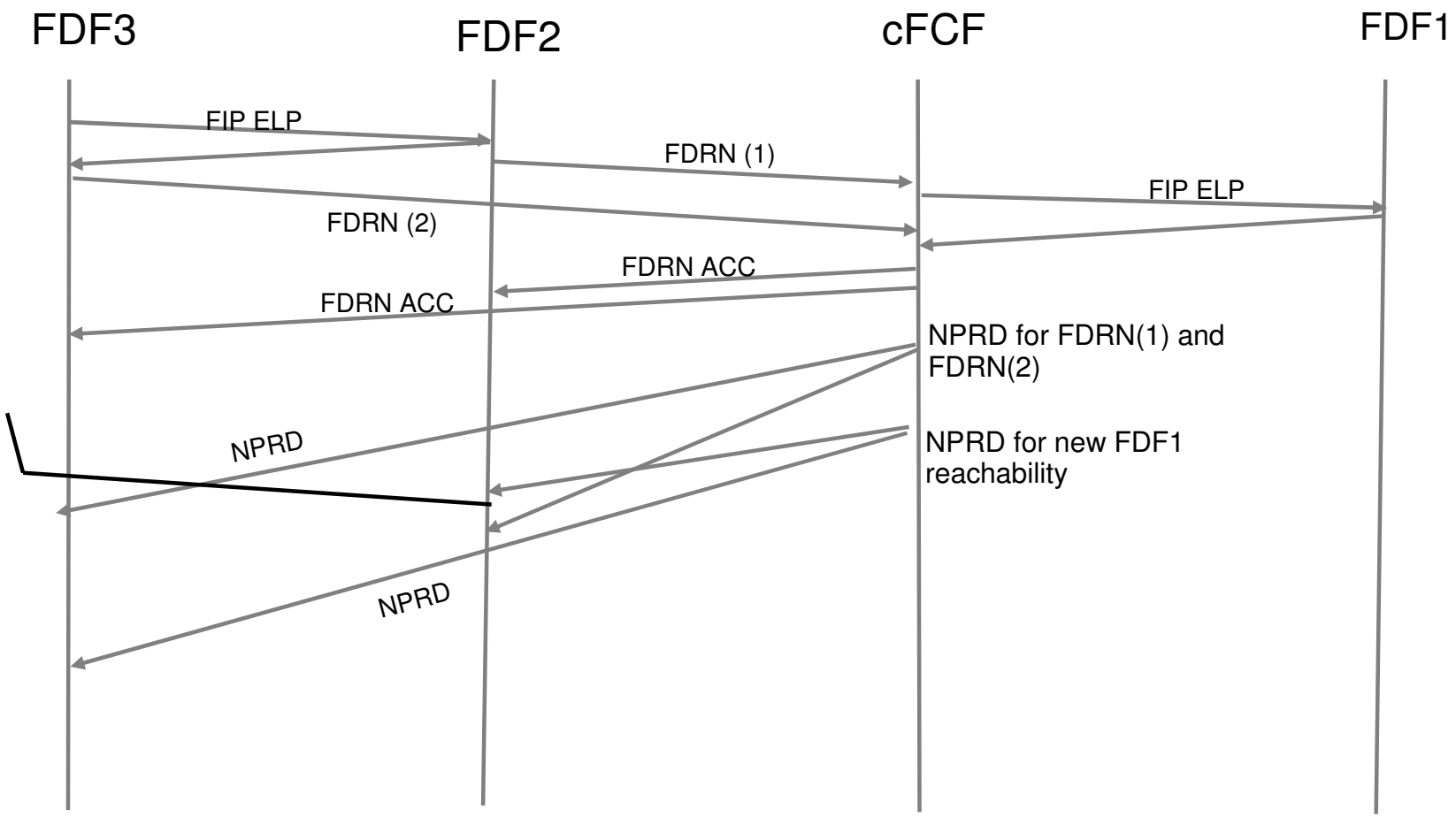
NPRDs sent in succession



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)

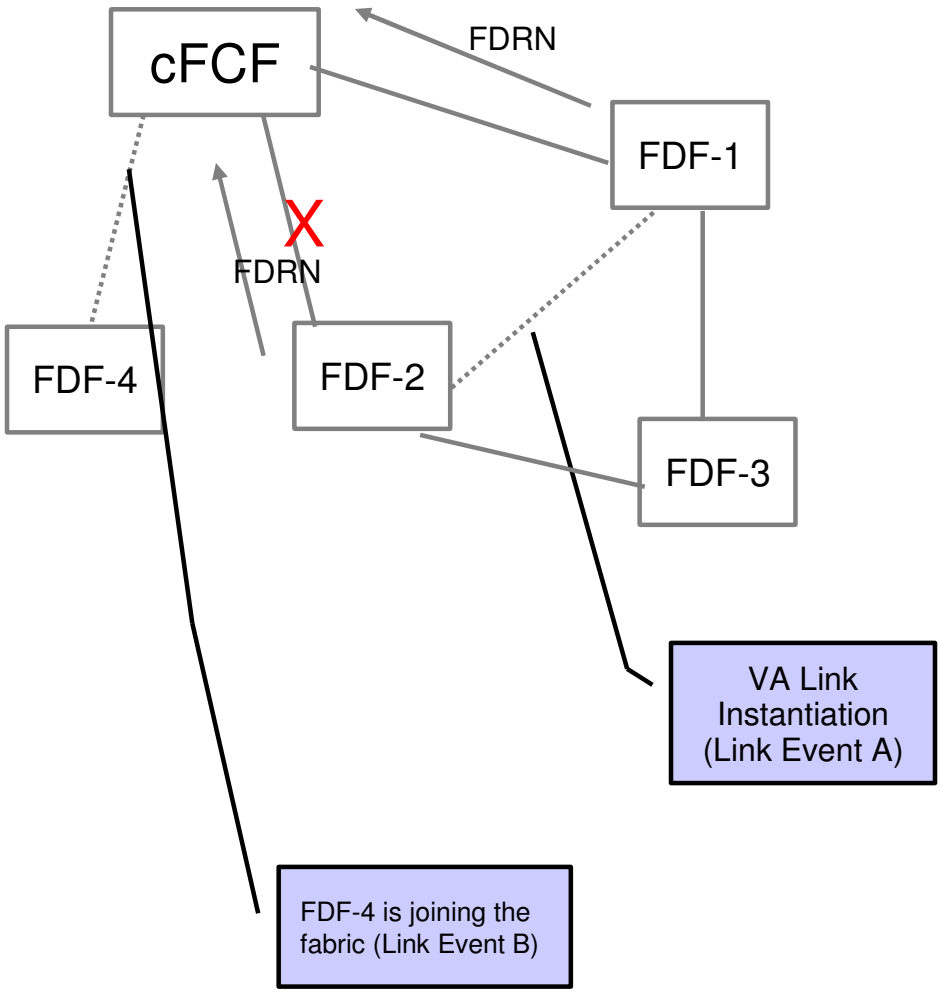
NPRDs Arriving Out-of-order in the Absence of an Error



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

Group Direction from Last Meeting

- NPRDs can be received out-of-order with respect to each other, and NPRDs, NPZDs and AZADs can be received out-of-order with respect to each other (reference 13-134v0)
  - The N_Port_ID Reachability Descriptor in NPRD and the Allocation Status Descriptor in NPZD both contain information that overlaps, specifically 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.
  - AZADs replace the entire zoning map, and similar 'out of order' issues exist with respect to the sequencing of AZADs with NPZDs and NPRDs

- Strict serialization of NPRDs with respect to NPZDs and AZADs and with other NPRDs was preferred over special rules for handling N_PortID information contained in NPRD routing distribution information

- Sequence number descriptors already exist in the NPRD and AZAD command descriptors (13-057 was amended for this in June FC-SW-6 meeting), but text does not exist to describe their use in all commands

Proposed 17.9.3 Modified Text

- The Primary Controlling Switch maintains a sequence number for each FCDF in the FCDF Set. The sequence number is incremented by one and included in the NPZD, NPRD, or AZAD sequence number descriptor each time an NPZD, NPRD, or AZAD Request is sent.
- Upon receipt of an NPZD, NPRD, or AZAD Request, an FCDF compares the sequence number in the received sequence number descriptor to that of the last processed NPZD, NPRD, or AZAD Request, or to 00000000 00000001h if none of these commands (NPZD, 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 received NPZD, NPRD, or AZAD 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 received NPZD, NPRD, or AZAD 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 17.9.3 Modified Text

- When a new Zone Set is activated in the Fabric, the Primary Controlling Switch shall recompute the Zoning ACLs for all N_Port_IDs allocated in the Virtual Domain and distribute them to the FCDFs of the Distributed Switch through AZAD Exchanges.
- If the Primary Controlling Switch has to send an AZAD request to an FCDF, any NPZD or NPRD requests outstanding to that FCDF shall first be completed. Any AZAD requests outstanding shall also be completed prior to initiating any subsequent NPZD or NPRD requests with that FCDF. (Delete: The Distribution of AZAD Requests shall take precedence over the distribution of NPZD Requests.)
- If the Primary Controlling Switch has to send an NPRD request to an FCDF, any NPZD or AZAD requests outstanding to that FCDF shall first be completed. Any NPRD requests outstanding shall also be completed prior to initiating any subsequent NPZD or AZAD requests with that FCDF.
- Upon receiving on a port a FLOGI Request or a NPIV FDISC Request from a Node, a Controlling Switch shall allocate to the newly reachable VN_Port an N_Port_ID from the Principal Domain_ID if it accepts the received FLOGI or NPIV FDISC Request.

Thank you