Link Recovery Times

Patty Driever

Roger Hathorn

Recovery from Link Down Conditions

- When a link goes down due to an error condition, SB-6 recovery architecture distinguishes between
  - Situations where the link remained down ‘briefly’ and
  - Situations where the link remained down for a longer period of time (SB_TOV) such that additional recovery actions are required

- SB_TOV is set to a value of (~1.5 seconds) based on the determination that if the link stayed down longer than this value the condition could have been caused by someone manually moving a cable from one port to another (i.e. the same two host and CU entities are no longer connected)
  - Recovery times exceeding SB_TOV can also have adverse impacts on other systems in a shared environment

Recovery From ‘Brief’ Link Down Conditions (on the fabric egress port)

- Class 3 frames on route to the target get dropped
  - FC-SB channels operating in ‘command mode’ detect conditions where a command was sent on this link but had not yet been acknowledged by the control unit
    - In this case such operations are terminated with appropriate status to the host operating system
    - Operations that proceeded beyond this point but that hadn’t completed will be detected by the host operating system missing interrupt handler (e.g. 15-45s)
  - FC-SB channels operating in ‘transport mode’ time all operations and use REC to determine the state of each operation
- A detected timeout (PTOV = 2s) or an RSCN received from fabric initiates “Test Initialization” process.
- Special commands (RNID, TIN) will be sent in Class 2 by the channel to determine if the link recovered quickly or not or if logical connections between the host and CU were lost for any other reason related to the link down
  - If the link did come back and logical connections were maintained, no further recovery actions are required
  - If one of these is not true, →

Recovery From Links Down Exceeding SB_TOV

- Since it is deemed manual intervention could have occurred (either accidental or malicious) ,
  - Logical connections are removed
  - The host OS is notified of the path removal and does ‘resetting event’ recovery for every device previously connected on this path
    - Determines that the entity it is now connected to is the same one it was connected to on this path before the link incident
    - Re-establishes path connections to the device and sets path back into appropriate multi-path groupings
    - Device-type-specific initialization is performed (i.e. for DASD vs tape vs other device types)
  - At the control unit, removal of logical paths releases device reserves as well as holds on ‘tracks’ (formatted sections of the device), tossing the active operations and enabling other operations on other paths needing such access to proceed

What Might Change Due to Increased Link Recovery Times?

- If SB_TOV were increased (i.e. time link is down before logical paths removed), OS or protocol layer could do a 'shortened' form of 'path validation' to ensure connection is to same device as before link went down
- However without removing logical paths, device reserves and track holds would remain in place for an elongated time, impacting operations of other connected systems.
- With today's relatively quick link recovery times, many errors types are 'lumped together' into a single recovery action that resets the entire adapter resulting in dropping light on the link
  - Would likely need to tackle additional complexity of designing more tailored recovery actions
- Current recovery paradigm is built on 50+ years of experience and implementation
  - Concern exists regarding amount of uncertainty and instability that would be introduced
- Changing this timeout value would also likely have a cascading effect on other timeouts, requiring new careful tuning

Other events that cause link down recovery

- Port reset from error
- Port reset after firmware updates
- In these cases, preservation of logical connections is desired and controlled by the port performing the recovery. (I know the cause of the link down, so I can rule out human intervention.)
- This is enabled by use of a logical path timeout process using LP_TOV
- LP_TOV in current implementations is 4s.
- If the aforementioned Test Initialization process does not complete successfully within LP_TOV, logical connections are removed.

Backups

FC-SB-6 Link Failure

- 10.2.3 FC-SB-6 Timeout Value
  - The FC-SB-6 timeout value (SB_TOV) is the maximum time allowed for a channel or a control unit to remain in the FC-FS-4 link failure state before an FC-SB-6 link failure is detected. Whenever SB_TOV is exceeded, an FC-SB-6 link failure shall be recognized. The value of SB_TOV shall be 1.5 seconds.
- 10.3 FC-SB-6 Link Failure
  - An FC-SB-6 link failure shall be recognized by a channel or control unit when an FC-FS-4 link failure persists for a period in excess of SB_TOV (see FC-FS-4 and 10.2.3).
- An FC-SB-6 Link Failure causes removal of logical paths which has an elongated recovery involving path re-establishment and path validation on every path to every device address (e.g. 256 devices per path, 2K paths on a port)
- It also causes Link Incident Records to be reported for service.

Logical Path Timeouts

- 10.4 Logical Path Timeout Error
  - A logical path timeout error shall be recognized by a control unit/channel when it has attempted to send a TIN IU to a channel/control unit after recognizing a logical path timeout condition for that channel and does not receive a TIR IU in response (see 6.4.7 and 10.2.4).
- Whenever LP_TOV is exceeded, a logical path timeout condition shall be recognized.
- The value of LP_TOV is model dependent within the range of 4-10 seconds. (implemented as 4, but looks like some leeway.)
- A logical path timeout error causes removal of logical paths which has an elongated recovery involving path re-establishment and path validation on every path to every device address (e.g. 256 devices per path, 2K paths on a port)

End of Presentation

Questions and Comments

please send to:

Patty Driever (pgd@us.ibm.com)
Roger Hathorn (rhathorn@us.ibm.com)