Preventing Reorder and Duplication Using Topology Sequence Numbers

Michael Takefman

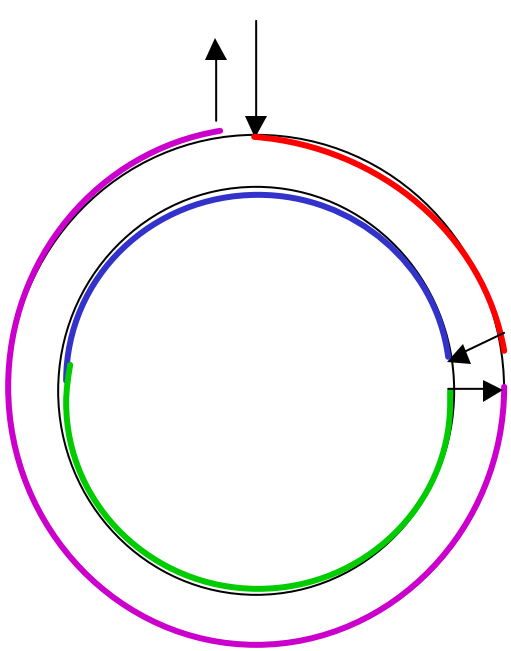
Wrapping Scenarios

- Wrapping systems do not suffer from reorder under normal failure operation, but there are special cases.
- 1) The packet is launched and then a nodal failure at the source removes it from the ring
- 2) A failure occurs at some point on the ring and the packet is wrapped. A subsequent failure occurs between the wrap point and the source wrapping the packet back on the original ring

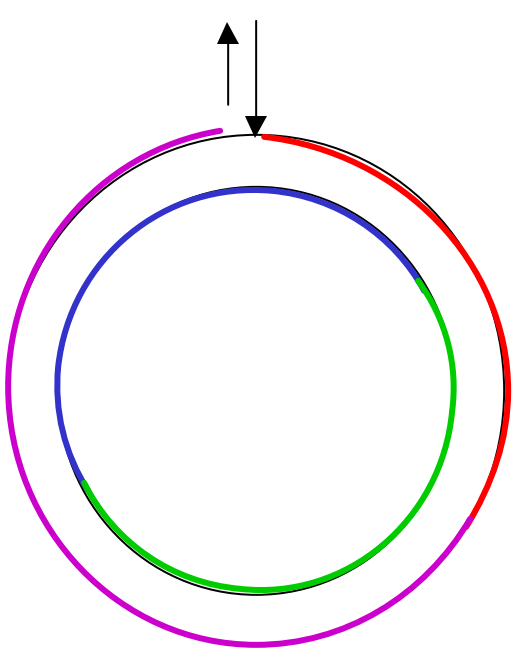
Wrapping Scenarios

- 3) A failure occurs at some point on the ring and the packet is wrapped. The packet travels around the ring and is wrapped back and then a subsequent failure occurs before the packet reaches the source.
- 4) When failures are cascaded close together, it is possible to reorder packets due to the packets that were "trapped" on the wrong ring (and expected to be deleted by TTL reaching zero) get wrapped back on the original ring.

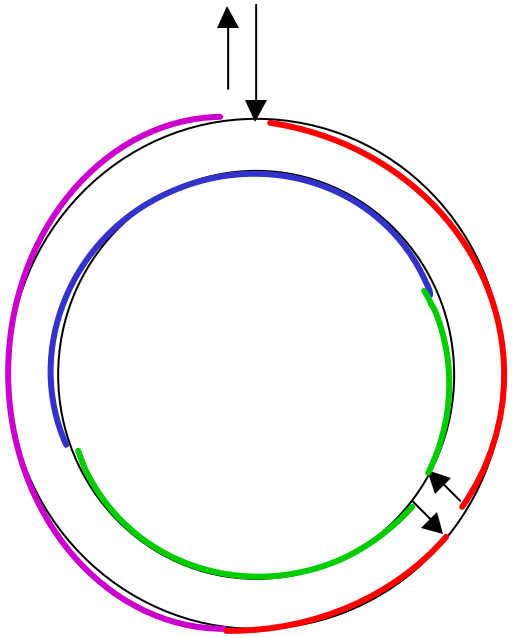
Wrapping Scenarios



Wrapped



Healed



Wrapped &
Misordered

Wrapping Scenarios

- Scenarios 1-3 are solved by adding 1 bit to the header to indicate the current WrapState
- WS = 0 when the packet is launched and the packet can only be wrapped onto the secondary ring when WS = 0
- Once the packet has transited the source node WS is set to 1 and the packet can only be re-wrapped onto the primary ring when WS == 1
- Once rewrapped the packet cannot be wrapped again.

Wrapping Scenarios

- The fourth scenario is solved by having all packets with the wrong RI deleted from a ringlet following the healing of a protection event.

Passthru Failures

- In Passthru the TTL will not be decremented correctly
  - 1st Killer scenario is for bi-directional flood doing a double delivery due to the last node in the flood space going into passthru
  - 2nd Killer scenario is all but one node goes into passthru, Packet circulates N times
  - 3rd Killer scenario is really wonky, but is solved by forcing stations coming out of passthru to wait a period of time before they begin to receive traffic
  - If passthru is illegal, then all passthru failures become Node Down failures.

A Key Definition

- For a ring using bi-directional flooding the ***flooding scope*** of a node is a tuple which indicates the distance that a packet should travel on each ring under normal conditions
- For an N node ring, there are N flooding scopes $FS_n(I_n,J_n)$; where $I_n+J_n = N-1$ for $n = 1 .. N$
- Whenever a protection event occurs, the FS for all but one of the nodes change. When the ring is healed it is likely that the FS return to the pre-fault values.

A Key Observation

- For bi-directionally flooded rings, the MAC replicates the packet for both attachments.
- Due to internal buffering within the MAC, it is critical that the FS of the packet is set when the Client gives the MAC the packet
- Setting the FS just prior to attach could have the same packet get two different flooding scopes that overlap

Topology Race Conditions

- This is the real killer scenario
- The database within the MAC may be inconsistent with packets traveling around the ring
  - This can cause a double delivery and reorder for bi-directional flood
- Each time a topology event occurs and the Flooding Scope changes, there will be packets in flight from different flooding scopes

Topology Race Conditions

- For rings with 2 TBs and bi-directional flooding the amount of storage in those nodes can cause a very large number of packets to be stored with different FSs if failures and restorations are coming rapidly
- 50 ms switch times cannot be guaranteed if one waits for a flush operation to clear the MAC and ring of packets from previous FSs

Flushing and Steering

- An assumption exists that the source node sends out flush messages, and once they are received the steer operation can safely occur
  - How does this message get handled at the failure point?
    - Wrap it ☺ ?
  - Can the ring go quiescent in 50 ms?
  - Is there are race condition with packet that “just” made it past the fault.

Flushing and Steering

- Consider the packets in flight towards a station
  - When a break occurs, some packets are downstream from the break and some are upstream
    - The upstream packets disappear into the failure but
    - The downstream packets are in a race with new packets that are steered from the source
  - Therefore, the stations to either side of the fault need to initiate the flush, and all other stations that are going to steer have to wait for the flush packets to pass them before sending.
    - This is faster than the source nodes sending out a flush message that has to traverse the entire ring!
    - It is the flush packet on the inner ringlet that triggers transmission on the outer ringlet (and vice-versa)

Flushing and Resilience

- What happens if a flush packet is lost?
  - Nodes waiting for the flush packet will not restart transmission
  - Flush packet must be retransmitted periodically
    - What period?
    - When does one stop sending the packet?
    - Packet has to be differentiable with regard to any future (or past) faults

Topology Sequence Numbers

- A simple and elegant method of insuring no misorder and duplication and allowing steering to occur in minimal time without the need to flush
- Each node maintains a local sequence number for its current Flooding Scope
- Anytime the scope changes (maximum distance or DA endpoint changes) the TSN is incremented
- The TSN and FS is advertised by the node when its flooding behavior is changing, the node should increment the sequence number
- All data packets carry the TSN of the source node and are flooded in accordance with the corresponding FS

Topology Sequence Numbers

- Destination nodes receive TSN and FS from all nodes.
- The Destination determines which ring it should receive bi-directional packets on for the given TSN
- A database associating SA and TSN is stored for each ringlet attachment
- Incoming packets are checked against the destination node's current TSN database and are deleted if the TSN does not match

Robustness Proof

- Packets arrive at the MAC SAP in order and are associated with the current TSN
- Bi-directional flooding with destination selection guarantees that only ½ of the packets that might appear at a MAC will be received thus preventing duplication and that those received are also strictly in order
- When a Topology / Steering change occurs, the TSN is advanced. The packets within that TSN are in order and are “newer” than packets from a previous TSN
- Eventually the destination receives instructions to move to the new TSN for that source. Once received, packets from the old TSN are discarded, and packets from the new TSN are received in order

Robustness Proof

- As long as the destination always uses the newest TSN that it is informed of, it will never reorder a packet as it will never accept a packet that is possibly stale