Resilient Network Interconnect:
Split Brian Avoidance

Version 0

Stephen Haddock
Extreme Networks
December 9, 2010

Introduction

- Link Aggregation, as a Resilient Network Network Interface (RNNI) control protocol, is no more susceptible to split-brain issues than any other protocol that could be used!

- A “split-brain” situation arises when:
  1. In normal operation, two (or more) devices depend upon a control path to coordinate their operation such that they function as a single virtual entity with a single identity; and
  2. Upon failure of the common control path, the two (or more) devices operate independently but
     a) Each assumes the full functionality of the single virtual entity; and/or
     b) Each continues to use the identity of the single virtual entity.
- Split-brain issues are avoided if the solution is designed so that conditions 2a and 2b do not occur.
  - There are two general approaches to achieving this.

Approach A: Easy Split-Brain Avoidance

- Prevent condition 2b by:
  - Assuring that all devices, or all but one pre-determined device, always switch to a unique identity (different from the identity of the single virtual device) upon failure of the control path.
- Prevent condition 2a by either:
  i. Assuring one and only one device assumes the full functionality of the single virtual device upon failure of the control path; or
  ii. Assuring that each device deterministically assumes a subset of the functionality that does not overlap or conflict with the subset assumed by another device.
- Link Aggregation, using the standard protocol without any changes running across the NNI, achieves this.
  - Details in subsequent slides.
- Characterized as “easy” because this approach does not require distinguishing whether a node failure or a link failure resulted in the loss of the control path.

Approach B: Hard Split-Brain Avoidance

- Prevent condition 2b by:
  - Assuring that one and only one device continues to operate with the identity of the single virtual device upon failure of the control path.
  - Note that with hard split-brain avoidance there is always one device continuing to operate with the identity of the single virtual device, whereas with easy split-brain avoidance there may or may not be a device that continues to operate with the identity of the single virtual device.
- Prevention of condition 2a:
  - The options for prevention of condition 2a are the same for both easy and hard split-brain avoidance. This is because once the identity issue is resolved, there are many possible ways to resolve the division of functionality.
- Characterized as “hard” because this approach requires distinguishing whether a node failure or a link failure resulted in the loss of the control path.

Easy vs. Hard

- Essential behavioral difference between easy and hard:
  - Whether there is guaranteed to be one device continuing to use the identity of the single virtual entity following a failure of the control path.
- The presentation given on the Nov. 23, 2010 conference call described 5 situations where this might be desirable.

  http://www.ieee802.org/1/files/public/docs2010/new-haddock-resilient-network-interconnect-addressing-1110-v1.pdf
- If we decide these are not important:
  - Then easy split-brain avoidance for Link Aggregation is sufficient, and can be achieved without modification to the standard protocol running across the NNI.
  - Other potential NNI protocols either don't need a single addressable virtual entity at the portal at all, or also use easy split-brain avoidance.
- If we decide these are important:
  - Then any NNI protocol (whether based on Link Aggregation or not) must support a single addressable virtual entity at each portal and must assure that this address continues to be used by one and only device following the failure of the control path within the portal.
  - Which means any NNI protocol needs to deal with hard split-brain avoidance.

For review: slide from previous presentation

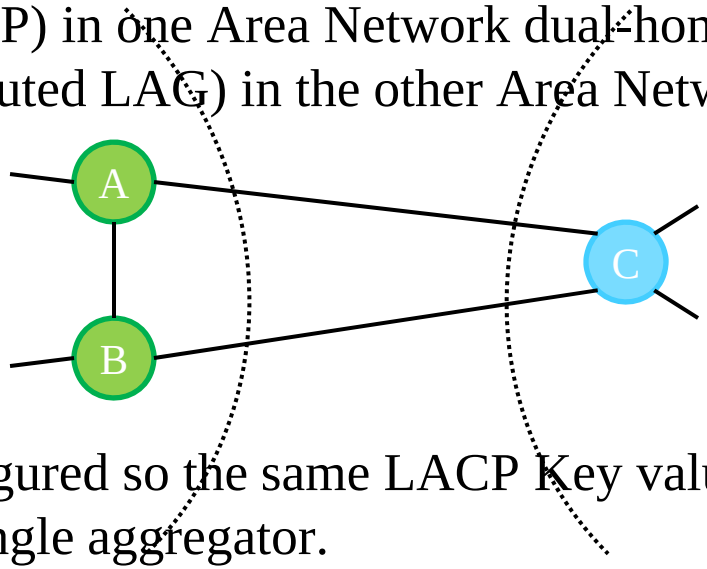
Having a single logical control element at the RNNI

- Even if LACP is not chosen as the RNNI protocol, there are other situations where there is value in having a single addressable control element at the RNNI.
- These situations are independent of:
  - The Data Plane Model
    - i.e. the need is common to Norm's, Steve's, and Zehavit's models.
  - The Control Plane Protocol used in the Buffer/Interconnect Network
- These situations include:
  1. Protection Switching in the Area Network
  2. Backbone Area Network with an S-tagged RNNI
  3. Backbone Area Network with an I-tagged RNNI
  4. E-Line services and point-to-point OVCs
  5. Service OAM (CFM and Y.1731)

Easy Split-Brain Avoidance with Distributed Link Aggregation

Standard LACP: Dual-homing

- Start with the simplest case: an NNI with a single node (running standard LACP) in one Area Network dual-homed to a pair of nodes (using Distributed LAG) in the other Area Network.



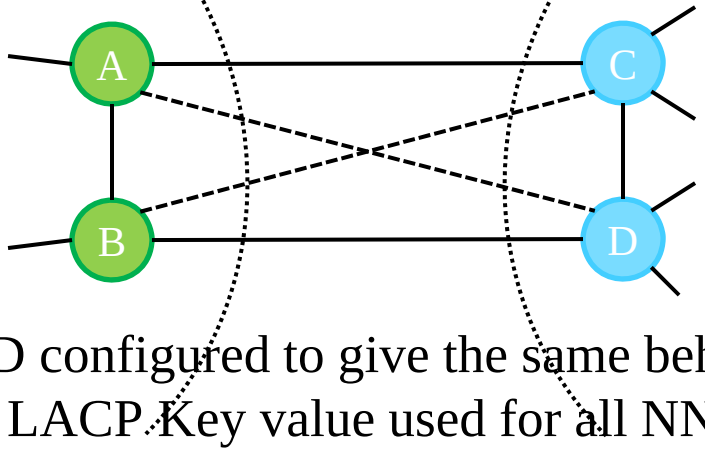
- Node C configured so the same LACP Key value is used for both NNI ports and a single aggregator.
  - This assures that either both ports come up as a LAG, or only one port comes up.
- Nodes A and B agree on a System ID and Key value to be used for the Distributed LAG when the link between A and B is operable.
  - It is not essential that this be the System ID of one of the nodes, but it is recommended in that it results in less disruption in case of a failure. A rule consistent with other 802.1 tie-breaking rules would be that each node propose a {System ID, Key} and select the one with the lowest numerical value.

Standard LACP operation (2)

- In normal operation:
  - Nodes A and B both use the mutually agreed System ID for LACP.
    - Assume it is the System ID for Node A.
  - Node C brings up both links in the same Link Aggregation Group.
- If the link between A and B fails:
  - Nodes A and B revert to using their own individual System ID for LACP.
  - Node C removes the link to node B from the Link Aggregation Group, leaving only the link to node A active.
    - Since Node C has no other aggregator with a Key value matching the link to Node B, that link is inactive.
    - If the mutually agreed System ID was not the individual System ID for either A or B, Node C would first remove both links from the Link Aggregation Group and then select one of them to re-connect to the aggregator as a new Link Aggregation Group.
- If Node B fails:
  - Node A continues to use the same System ID for LACP.
  - Node C removes the link to node B from the Link Aggregation Group, leaving only the link to node A active.
- If Node A fails:
  - Node B reverts to using its own individual System ID for LACP.
  - Node C first removes the link to node B from the Link Aggregation Group, then reconnects it to the aggregator as a new Link Aggregation Group.
    - Could optionally optimize this using Norm's "Graceful Name Change" proposal.

Extending to a dual-dual-homed case

- A Resilient NNI with a pair of nodes (using Distributed LAG) in each Area Network, connected with minimal connectivity (solid lines) or full mesh (solid + dotted lines).



- Nodes C and D configured to give the same behavior as a single node with the same LACP Key value used for all NNI ports and a single aggregator.
- Nodes A and B operate the same as in the dual-homing case.
- The response to a failure of Node A, or Node B, or the link between them is the same as in the dual-homing case.
  - The link(s) to either Node A or Node B, but not both, will continue to be operational.

Hard split-brain avoidance revisited

- There is no requirement that the LACP System ID and the NNI port address be the same.
- Since only the links to one of the Nodes A or B remain active, we could consider having both Nodes A and B continue to use the same NNI port address.
- This would resolve at least some of the “hard” split brain issues raised in the Nov 23 presentation.
  - Would resolve those related to using the NNI port address as a B-MAC address when Nodes A and B are Backbone Edge Bridges.
  - Need to investigate if there are control protocols that would then need to use this NNI port address, and how those would be affected.

Sensitivity to Node C configuration

- What happens if Node C (or Node C + D) mis-configured?
  - Specific mis-configuration of concern is if more than one aggregator uses the same key value as the NNI ports.
  - Then if link between A and B fails, they will revert to using distinct addresses for their system ID, but all NNI links remain active as two separate link aggregation groups.
- Impact is frame duplication or a potential forwarding loop.
  - If C continues to filter frames to a designated NNI link, then may get frame duplication in C's network.
  - If C treats both LAGs as distinct links and does not continue to select one for a given service, then may create a forwarding loop.
- Want to have some method for AB network to protect itself from mis-configuration faults at C.
  - If either frame duplication or looping occur, it is because there is a path between A and B in the AB network as well as through C. Nodes A and B can utilize this path to discover each other. Could attempt to discover each other through C as well. One path is direct but depends upon C; the other path may have longer forwarding latency but does not depend upon C or C's network.