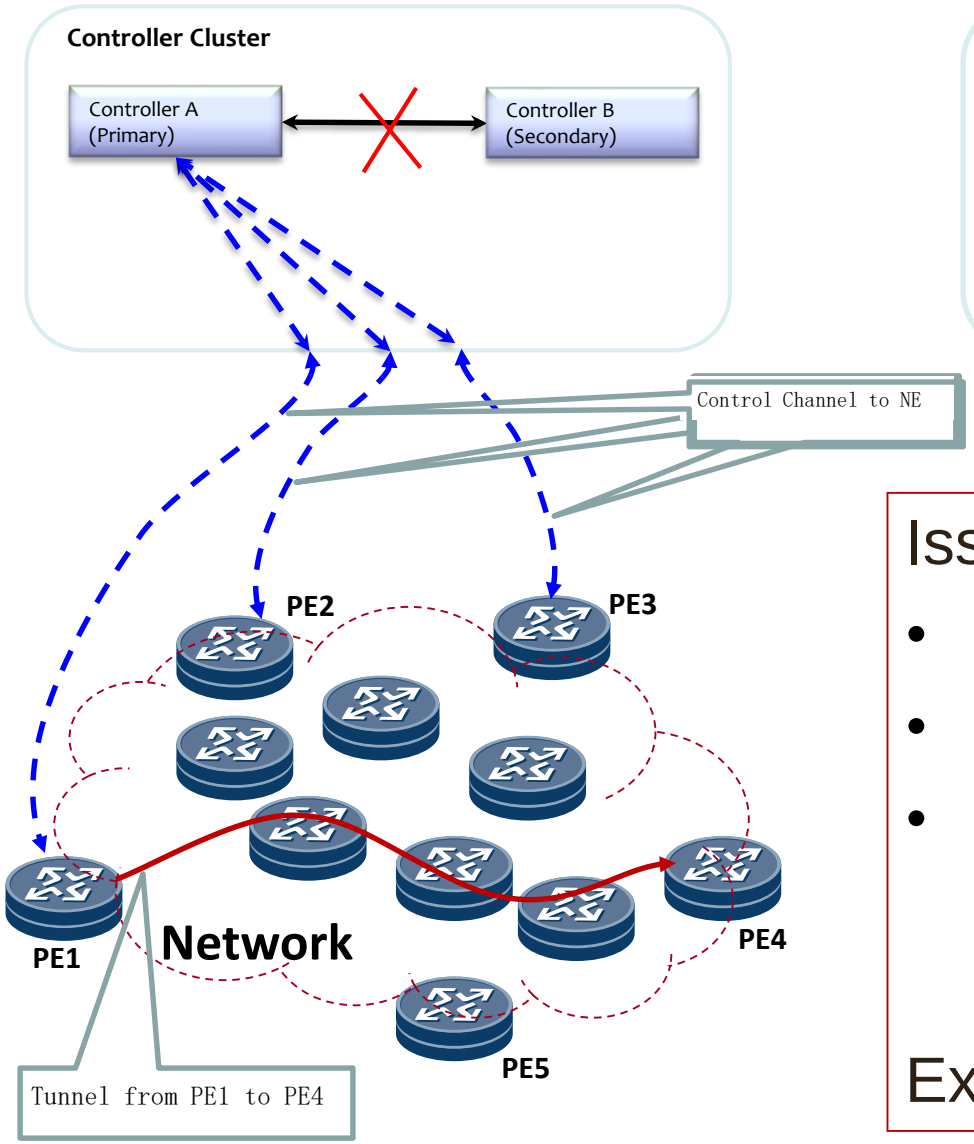
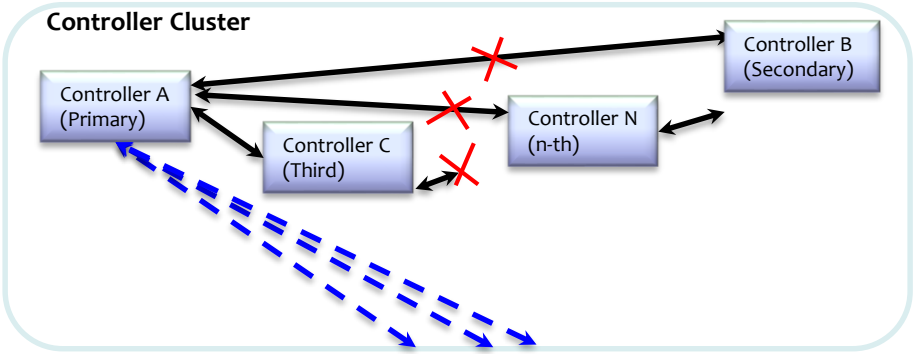
BGP for Network High Availability

draft-chen-idr-ctr-availability-00

Huaimo Chen(Futurewei)
Yanhe Fan (Casa)
Aijun Wang (China Telecom)
Lei Liu (Fujitsu)
Xufeng Liu (Volta Networks)

Introduction



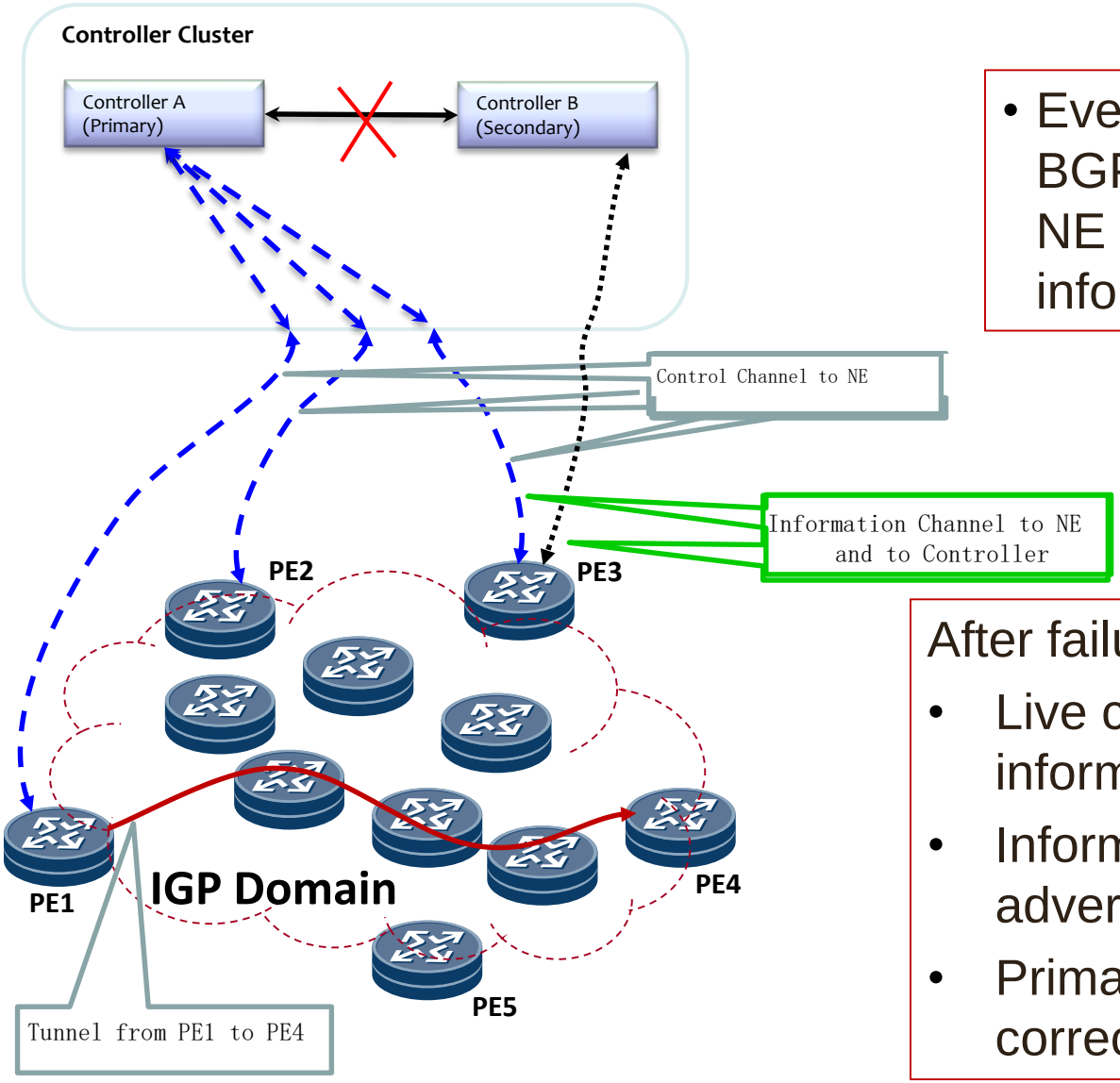


Issues by failures in cluster:

- Split into separated groups
- Out of Synchronization
- Multiple primary groups selected to control network

Extend BGP to resolve these

Overview of Mechanism



- Every controller has BGP session to same NE (e.g., PE3) as information channel

After failures in cluster

- Live controller has information channel to NE
- Information on controller is advertised via the channel
- Primary group is selected correctly to control network

Information on Controller

Normally, A (Primary) advertises the information about the controllers connected to it:

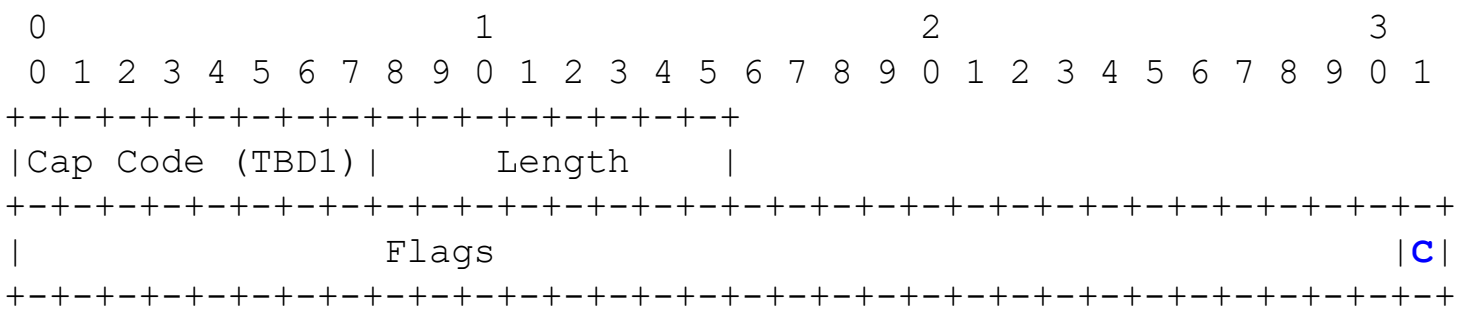
C = 1, A's current Position = 1, A's OldPosition = 1, A's Priority, NoControllers = n, A's ID, B's ID, …

After failures in cluster, for each separated group

- Intent primary, secondary controller, and so on are elected
- Intent primary controller advertises information about its group
- Every group has information about others. Primary group is selected.
- In case of tie, group with the highest old position controller (e.g., the old primary controller) wins in one policy

Extensions to BGP: Capability

New Controller HA Support Capability Triple is defined in Capabilities Optional Parameter in OPEN message



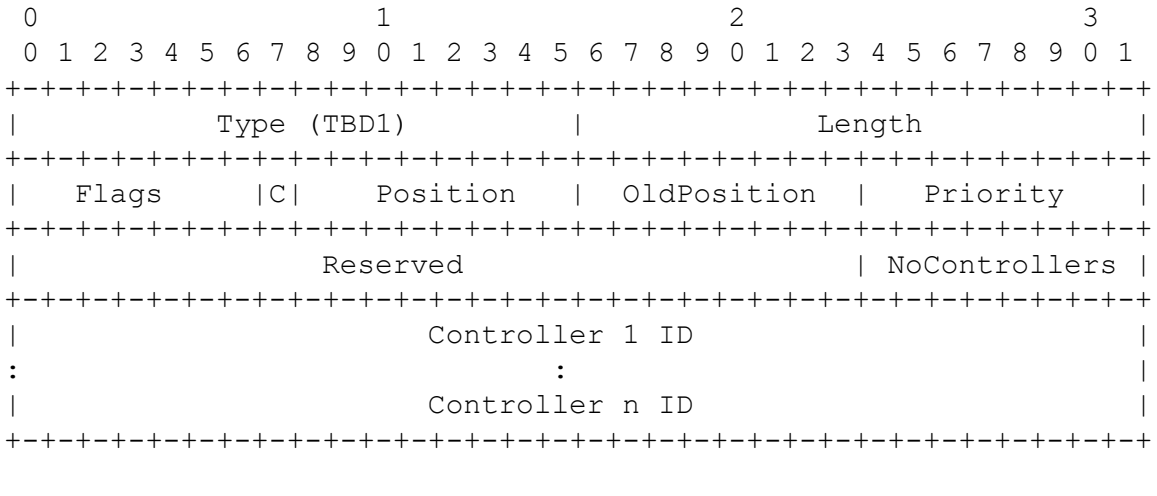
Controller HA Support Capability Triple

Flag (32 bits): One flag bit, **C-bit**, is defined.

- When it is set to **one**, it indicates that the BGP speaker supports the high availability of controller cluster as a **Controller**.
- When it is set to **zero**, it indicates that the BGP speaker supports the high availability of controller cluster as a **network element (NE)**.

Extensions to BGP: NLRI

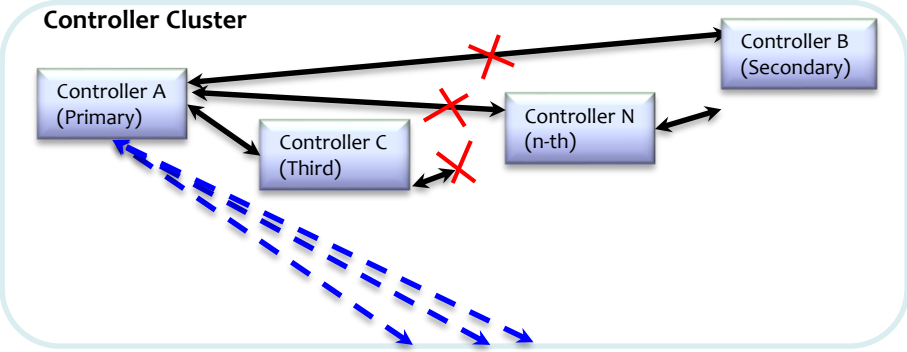
Under a new AFI and SAFI, a new NLRI (called Controllers NLRI) is defined for information about controllers



Controllers NLRI

- Flag (8 bits): One flag bit, **C-bit**, is defined. When set, it indicates that the position is the position of the **current active primary controller**.
- **Position** (8 bits): It indicates the current/intent position of the controller in the controller cluster or group. **1: primary (1st)** controller, **2: secondary** controller, …
- **OldPosition** (8 bits): It indicates the old position of the controller in the controller cluster before it is split.
- **Priority** (8 bits): It indicates the priority of the controller to be elected as a primary controller.
- **NoControllers** (8 bits): It indicates the number of controllers
- **Controller i ID** (32 bits): It represents the identifier (ID) of controller i at position i (i = 1, ..., n) in the cluster or group.

Recovery Procedure



- Cluster of n controllers: A, B, …, N with position 1, 2, …, n respectively
- Failures split cluster into:

Group 1: A, C;

Group 2: B, N

- Normally, A advertises information about controllers:

C=1, Position=1, OldPosition = 1, A's Priority, NoControllers=n, A's ID, B's ID, …, and N's ID.

- After failures, intent primary in each separated group advertises information about its controllers

A in group 1 advertises:

C=0, Position=1, OldPosition = 1, A's Priority, NoControllers=2, A's ID, C's ID.

B in group 2 advertises:

C=0, Position=1, OldPosition = 2, B's Priority, NoControllers=2, B's ID, N's ID.

Group 1 and 2 have the same number of controllers, which is 2. But OldPosition in group 1 is higher than that in group 2. Group 1 is elected as the primary group

Primary controller A in the primary group (i.e., group 1) advertises

C=1, Position=1, OldPosition = 1, A's Priority, NoControllers=2, A's ID, C's ID.

Next Step

Comments