Distribution of Device Discovery Information in NVMe Over RoCEv2 Storage Network Using BGP

draft-wang-idr-bgp-nof-nlri-03

Ruixue Wang(China Mobile)

Changwang Lin(New H3C Technologies)

Fengwei Qin (China Mobile)

Qi Zhang (China Mobile)

Wenxuan Wang (China Mobile)

Rui Zhuang (China Mobile)

Background and Motivation

- NVMe over RoCEv2 is currently a popular solution of storage network based on Ethernet for accelerating communication performance.

- A host accesses to an NVMe storage subsystem via Ethernet Fabric with RoCEv2 protocol, **but this approach has the following problems:**

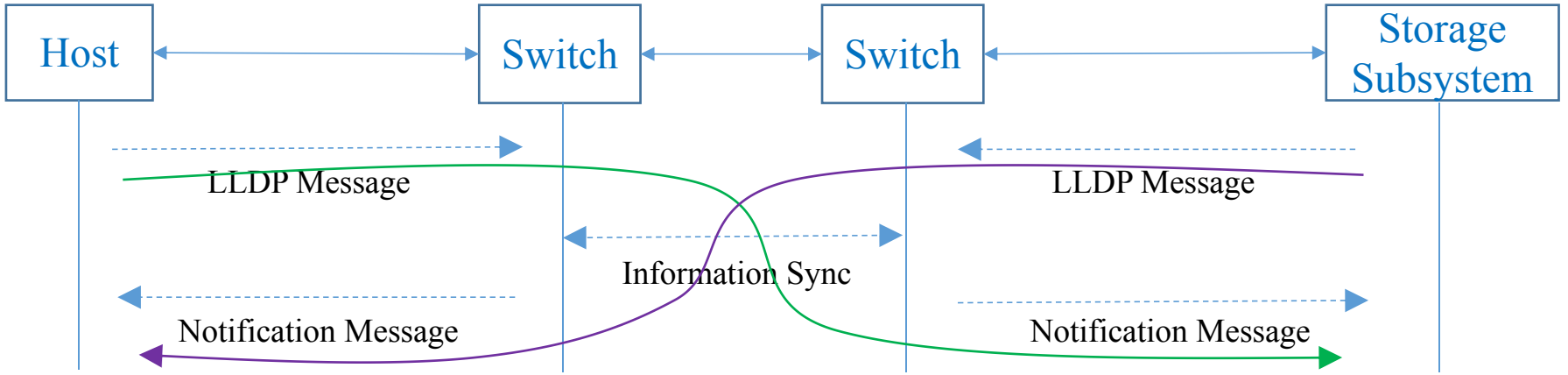
a. Require Manual intervention for the discovery of host and storage subsystem, which is not conducive to management and maintenance.

b. The reaction speed is slow when a device goes online or offline, making it hard to realize hot-plug and failover.

- To solve these problems, automatic discovery method should be deployed.

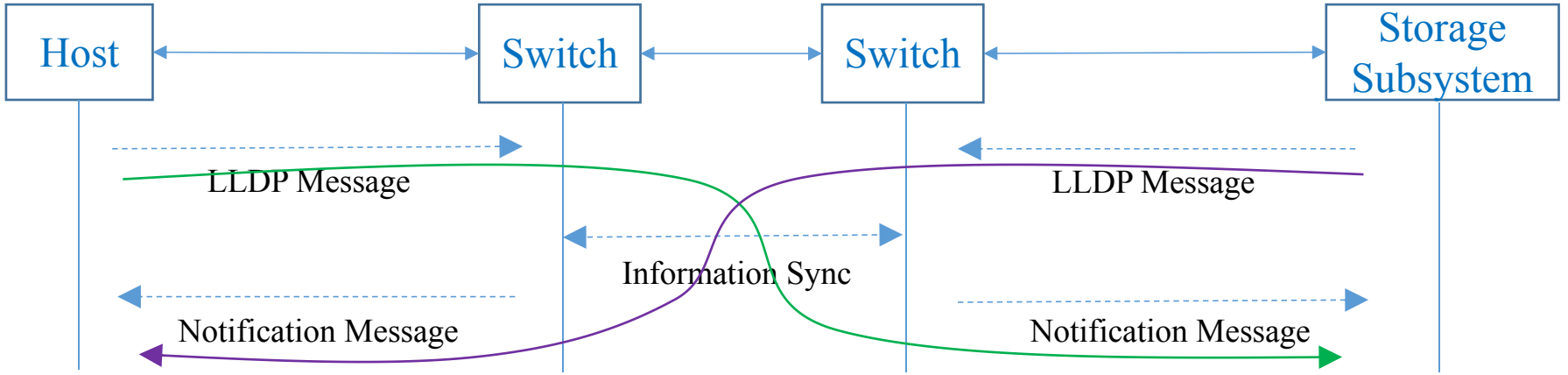
Background and Motivation

So, how to do the automatic discovery ?



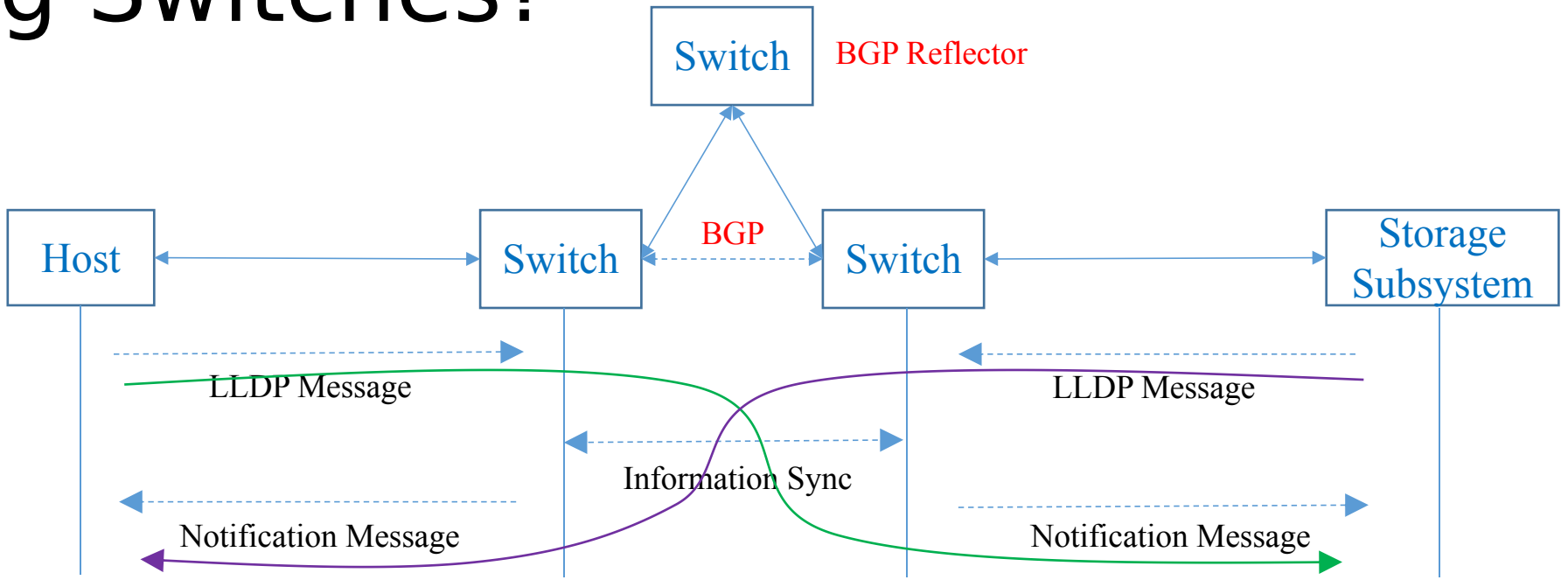
- When a host or storage subsystem is directly connected to a switch, it periodically reports its information to the switch using device discovery protocol like LLDP. The discovery information include its own information and subscription of interested storage subsystem.
- Then, the discovery information is distributed to others access switches by the access switch in the fabric.
- Finally, other devices (host or storage subsystem) get the discovery information via Notification Message from the switches which they directly connect with.
- So the host and storage subsystem know that each other gets online or offline.

Background and Motivation



- When the host knows that the storage subsystem gets online, the host will establish connection with the storage subsystem, and transmit data through NVMe over RoCEv2.
- When the storage subsystem failed, the host will quickly know the failed storage subsystem and disconnect from the storage subsystem, then the host can quickly switch over to the redundant service.
- In this solution, the LLDP messages and Notification messages between access switches and hosts or storage subsystems have been defined by ODCC-2020-05016 (the NVMe over RoCEv2 Network Control Optimization Technical Requirements and Test Specifications).
- However, the discovery information synchronization method between access switches has not yet defined .

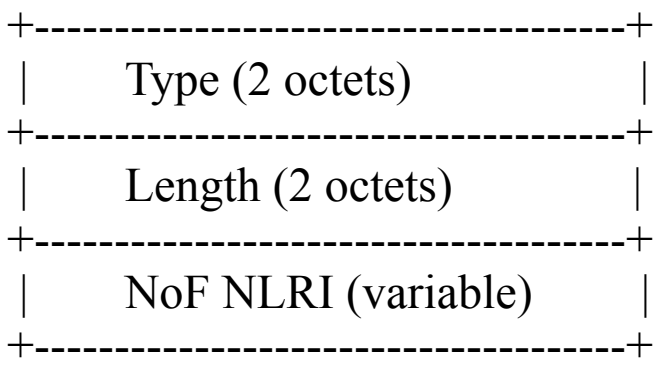
How to Transmit Discovery Information Among Switches?



- Proposes a new method of distributing device discovery information among switches by using BGP.
- All the access switches are BGP speakers, and the device discovery information is exchanged as BGP routes among them.
- The BGP Route Reflectors can be used to reduce the number of BGP connections.

So, in order to carry the device discovery information by BGP, a new BGP Network Layer Reachability Information (NLRI) is defined, named NoF NLRI

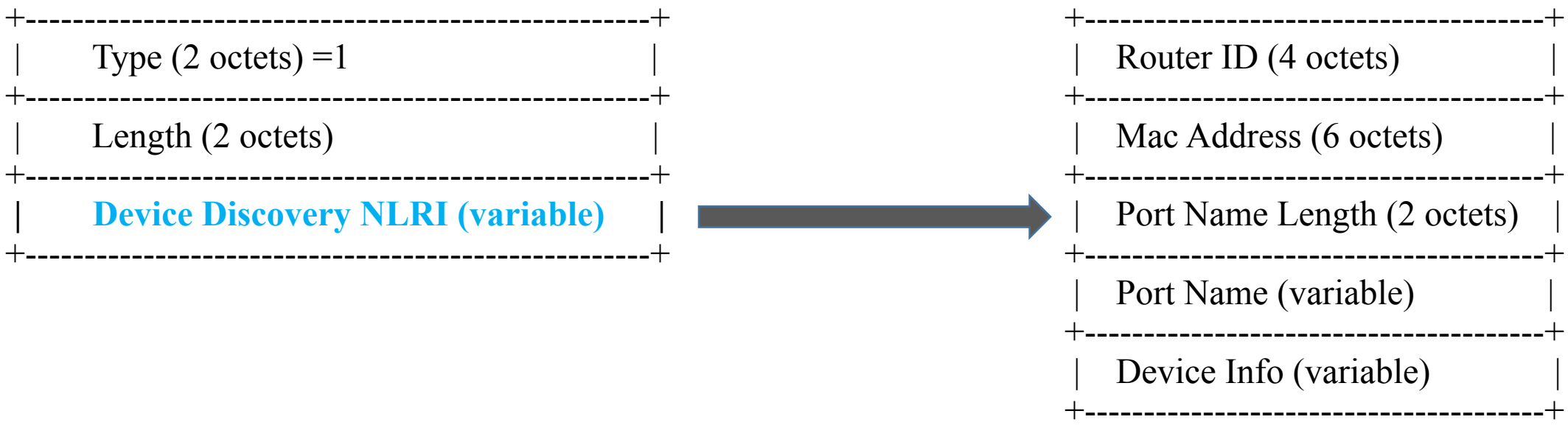
The Format of NoF NLRI



Two NoF NLRI type are defined, this is:

- Type = 1 : Device Discovery NLRI
  - Used to carry the discovery information of directly connected devices (host or storage subsystem).
- Type = 2 : Device Zone NLRI
  - Used to distribute the zone configuration of a device (host or storage subsystem).
  - In storage networks, hosts and storage subsystems are generally divided into several zones.
  - Only the devices in the same zone are allowed to discover and communicate with each other.

The Format of Device Discovery NLRI



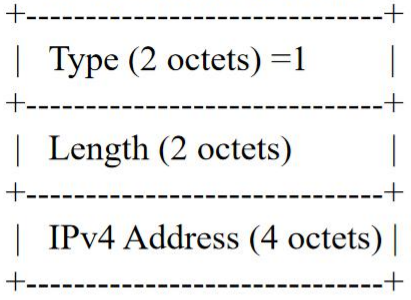
- **Router ID :** the Router ID of the access switch which originates the NLRI.
- **Mac Address :** the Mac Address of a connected device (host or storage subsystem).
- **Port Name :** the name of the connecting port, to distinguishing different ports which share the same Mac Address.
- **Device Info :** the specific information of the connected device and its connecting port, containing some TLVs.

The Key of Device Discovery NLRI is **Router ID**, **Mac Address** and **Port Name (Length)**, which are part of the prefix in the NLRI.

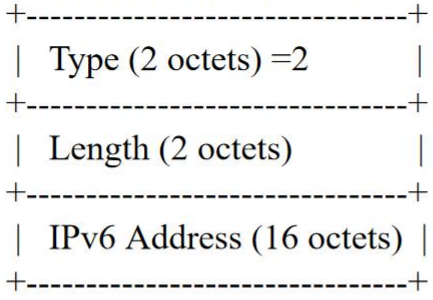
The Format of Device Discovery NLRI

The Format of Device Info TLVs:

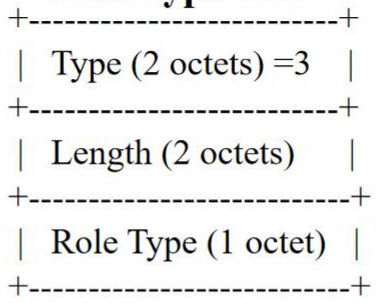
IPv4 Address TLV



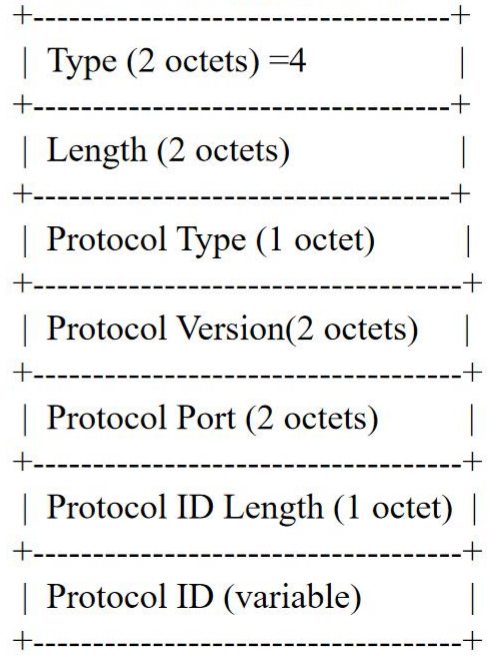
IPv6 Address TLV



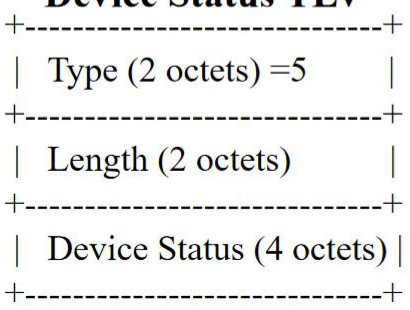
Role Type TLV



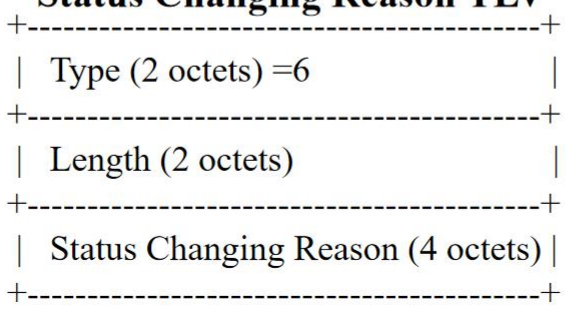
Service Protocol TLV



Device Status TLV

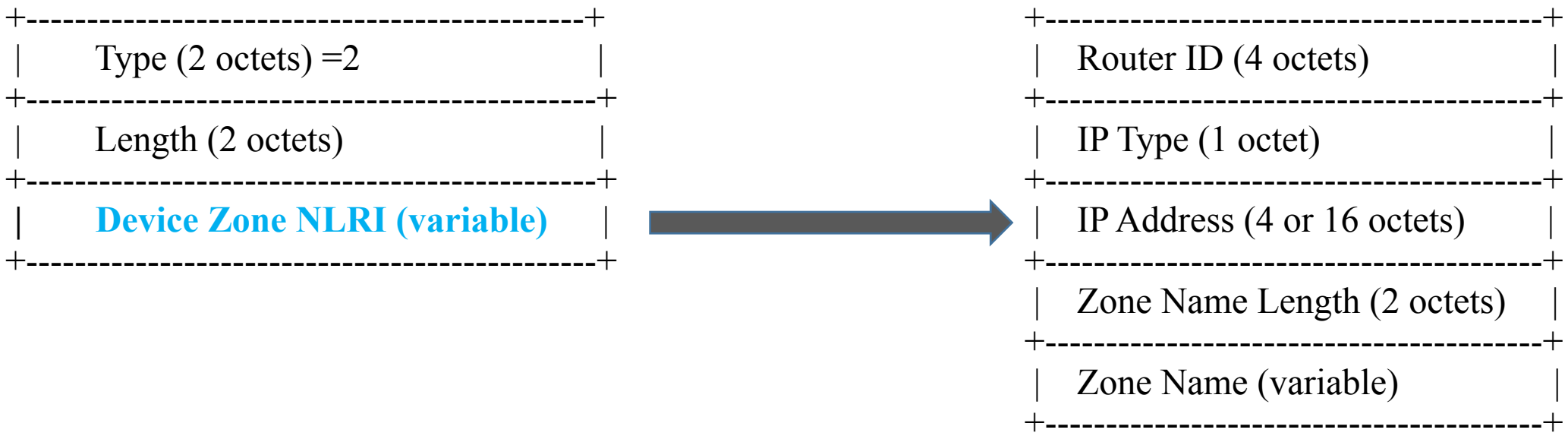


Status Changing Reason TLV



- **IPv4/6 Address :** the address of the connecting port.
- **Role Type :** the role of the device. Such as storage subsystem, host, and the device which can serve as both host and storage subsystem.
- **Protocol Type :** 0 (NVMe over RoCEv2)
- **Device Status :** 0 (offline),1 (online)
- **Status Changing Reason :** the reason of the device status changing.

The Format of Device Zone NLRI



- **Router ID :** the Router ID of the access switch which originates the NLRI.
- **IP Type :** indicating the type of IP Address, 0 (IPv4), 1 (IPv6).
- **IP Address:** the IPv4 or IPv6 Address of a connected device.
- **Zone Name:** the name of the zone which the connected device belongs to.

The Key of Device Zone NLRI is **Router ID**, **IP Address** and **Zone Name**, which are part of the prefix in the NLRI.

Next Step

- Request more review to improve the draft.
- Any questions or comments are Welcomed.