November 2019

FASTER CLUSTER CONVERGENCE AND LIVELINESS DETECTION THROUGH HIERARCHAL BIDIRECTIONAL FORWARDING DETECTION

Tapraj Singh

Jeffrey Lo

FASTER CLUSTER CONVERGENCE AND LIVELINESS DETECTION THROUGH HIERARCHAL BIDIRECTIONAL FORWARDING DETECTION

AUTHORS:
Tapraj Singh
Jeffrey Lo

ABSTRACT

Techniques are described herein for hierarchal Bidirectional Forwarding Detection (BFD) to scale liveliness detection. Scalability of BFD sessions to Master and Standby nodes is a concern in current implementations. A fast detection mechanism such as BFD is required to support fast detection of liveliness of cluster nodes.

DETAILED DESCRIPTION

In clustering use cases for enterprise networks, a large number of nodes (e.g., greater than 4,000) form a cluster. The cluster is formed based on Internet Protocol (IP) as the underlay transport. In order to detect the liveliness and for faster convergence of the cluster, an IP Operations, Administration, and Management (OAM) mechanism can be used. However, convergence and liveliness detection through this IP mechanism is very slow. For fast detection of the liveliness of cluster nodes, it is a requirement to support a fast detection mechanism. Bidirectional Forwarding Detection (BFD) is one such mechanism for fast detection of liveliness. Because scalability of BFD sessions to Master and Standby nodes is a concern in current implementations, techniques are described herein which use hierarchal BFD for liveliness detection and faster convergence of cluster nodes.

Figure 1 below illustrates an example clustering architecture. The clustering network is a hierarchal network of devices. The Master is the main device (e.g., wireless controller or high capacity switch). There are also Standby switches/controllers.

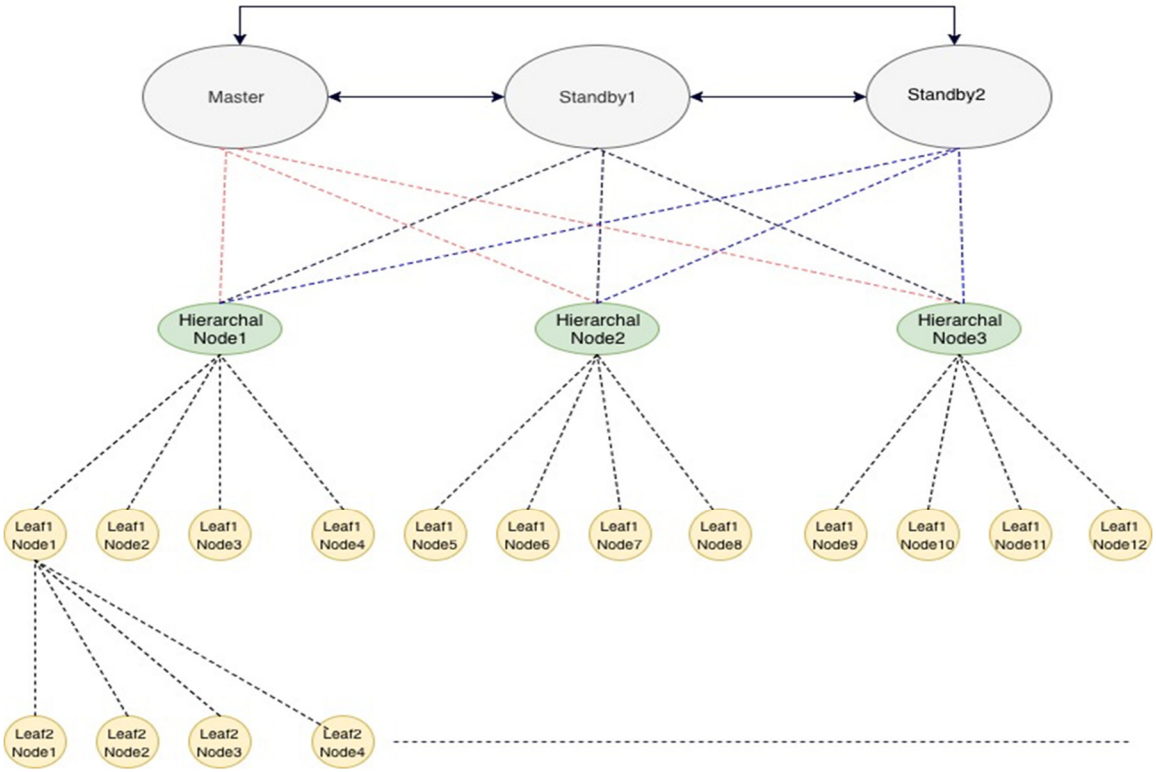


Figure 1

Figure 2 below illustrates an example BFD packet format. The diag field of BFD may be used to identify a BFD packet as a hierarchal BFD packet. Because diag field values 0-7 are already used, diag field value 8 may identify that the packet is a hierarchal BFD packet.

Vers	Diag	Stat	P	F	C	A	D	R	Detect Multi	Length
My Discriminator										
Your Discriminator										
Desired Min Tx Interval										
Required Min Rx Interval										
Required Min Echo RX interval										

Daig	Description
0	No Diagnostic
1	Control Detection Time Expired
2	Echo Function Failed
3	Neighbor Signaled Session Down
4	Forwarding Pane Reset
5	Path Down
6	Concatenated Path Down

Daig	Description
7	Administratively Down
8-31	Reserved for Future Use

Figure 2

Figure 3 below illustrates an example clustering architecture that implements hierarchical BFD. Hierarchy level 0 is used for hierarchal nodes 1-3. Hierarchal nodes 1-3 have BFD sessions to the Active/Master node and Standby nodes. There is also a second level of hierarchy represented as leaf1 nodes 1-12. Hierarchy level 1 nodes have BFD sessions to hierarchal nodes 1-3. Similarly, leaf1 node1 has a BFD session to leaf2 nodes 1-4 at hierarchy level 2.

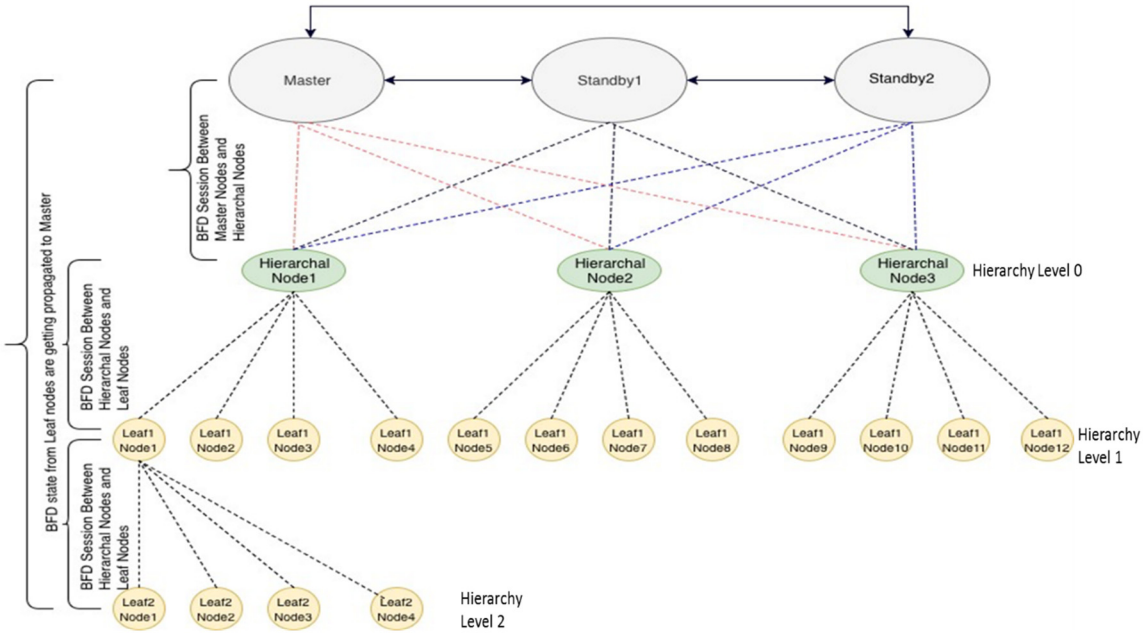
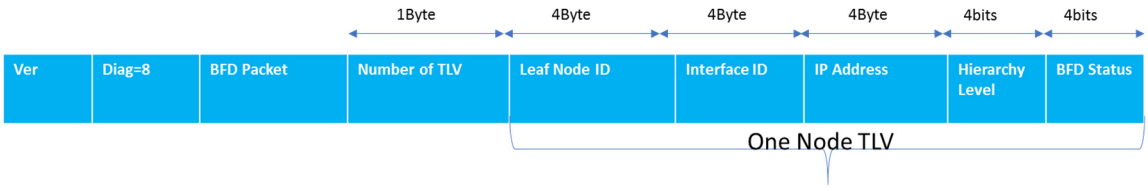


Figure 3

Figure 4 below illustrates an example hierarchical BFD packet format. As shown, TLV fields have been added. The number of TLV field identifies a number of nodes corresponding to this hierarchal BFD packet. The leaf node ID field is four bytes and is used to identify the relevant leaf node. The interface ID field is also four bytes and is used to identify the interface on which BFD is operating. The IP address field is used to detect failure of a given service. The hierarchal level field identifies whether the BFD is operating for the first level of hierarchy or the second level of hierarchy. The BFD status field indicates the status of the hierarchal session.



Status	Description
0	Init State
1	Hierarchal BFD Session UP
2	Hierarchal BFD Session Down
3	Terminated, bz Configuration Removed
4-above	Reserved for Future Use

Figure 4

Figure 5 below illustrates the BFD session between the Master/Active, Standby, and hierarchal nodes 1-3. BFD sessions may be established after the IP underlay. The Master/Active/Leader may send a BFD packet with the diag field value set to 8 and the number of TLV field set to 0. The BFD packet from hierarchal nodes 1-3 may include the diag field value set to 8. The number of TLV field depends on the number of BFD sessions below the hierarchal node. In this example there is no active BFD session, and therefore the hierarchal node sets the number of TLV field value to 0. After the BFD session is established, periodic BFD packets are exchanged.

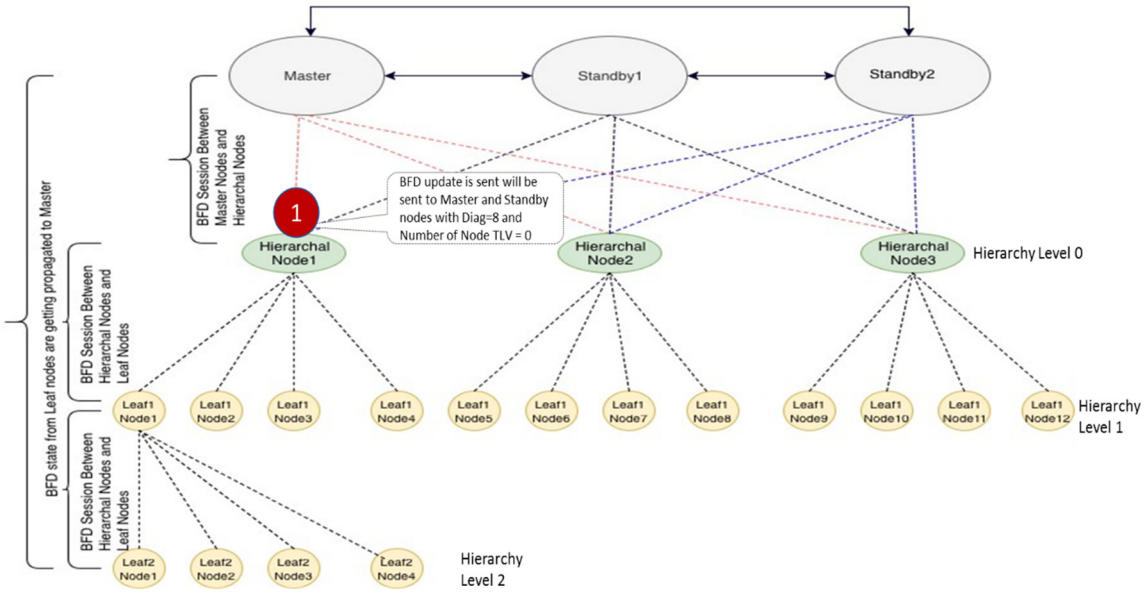


Figure 5

As illustrated in Figures 6-8 below, BFD sessions between hierarchal nodes 1-3 and leaf1 nodes 1-12 may be established using BFD packets. Leaf1 nodes 1-4 have BFD sessions to hierarchal node 1, leaf1 nodes 5-8 have BFD sessions to hierarchal node 2, and leaf1 nodes 9-12 have BFD sessions to hierarchal node 3. The hierarchal nodes may send BFD packets to the leaf nodes with diag field values set to 8 and the number of TLV field set to 0. Similarly, for a BFD packet from a first leaf1 node, a second leaf1 node may send BFD packets with the diag field value set to 8 and the number of TLV field set to 0.

Figure 6 below illustrates a BFD session for hierarchal node 1 and leaf1 node 1.

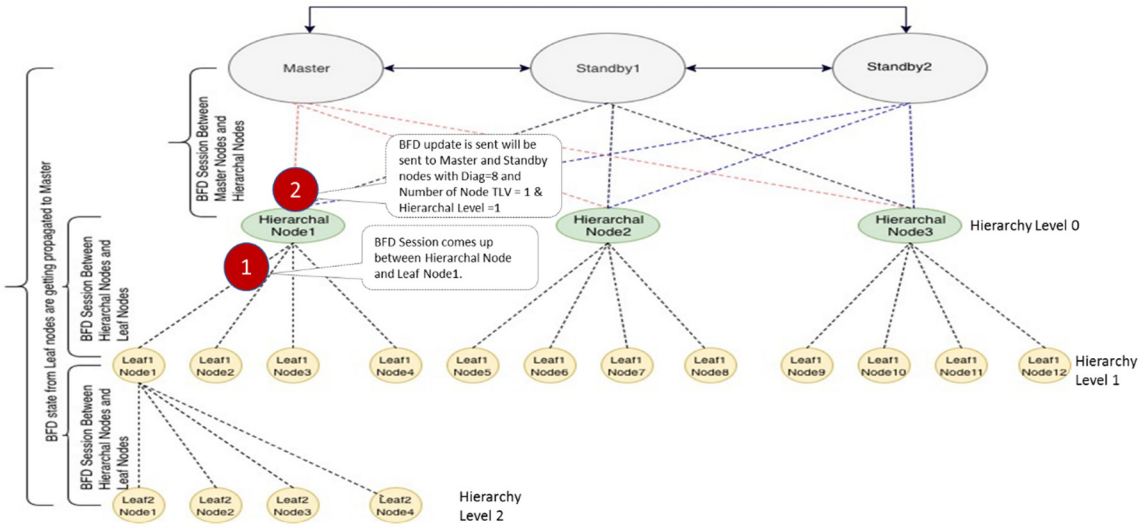


Figure 6

Figure 7 below illustrates a BFD session for hierarchal node 1 and leaf1 node 2.

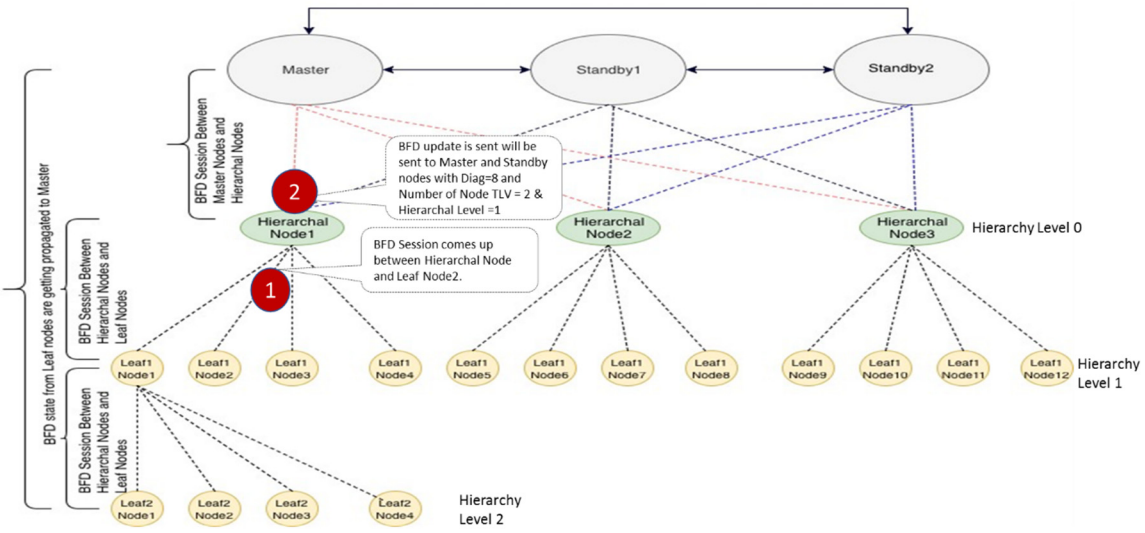


Figure 7

Figure 8 below illustrates a BFD session for hierarchal node 1 and leaf1 nodes 3-4.

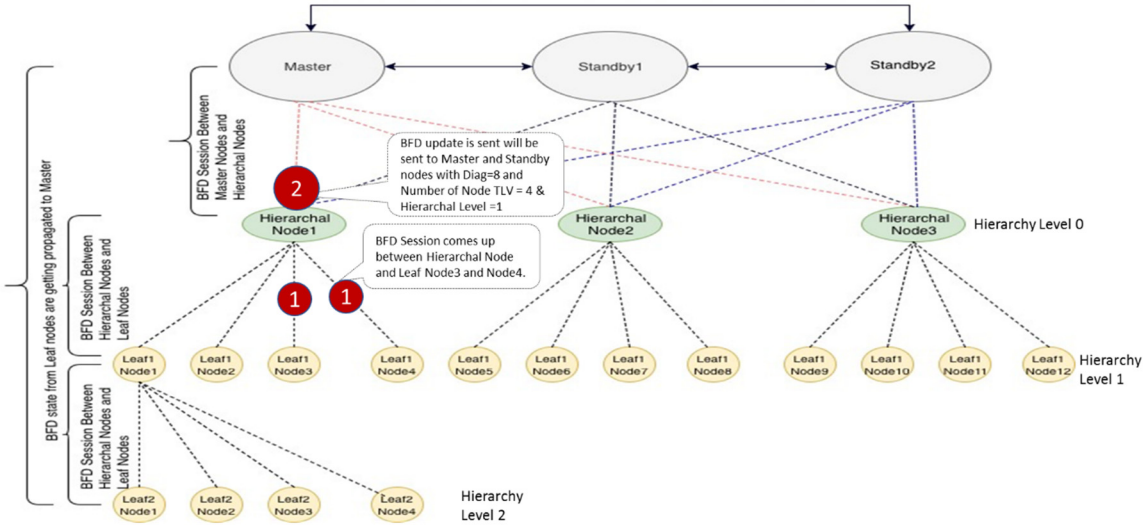


Figure 8

Figure 9 below illustrates an example BFD session periodic update. For a BFD packet from a hierarchal node to a Leader/Master and Standby node, the hierarchal node may send a periodic BFD session update to the Leader/Master and Standby node with followings attributes. The BFD packet may have a diag field value of 8, number of TLV set to the number of BFD sessions to a leaf1 node, a node ID set to the leaf1 node, an interface ID and an IP address both set to a value of 0 for the node level BFD, a hierarchal level set to 1, and the BFD status set according to the status of the BFD session.

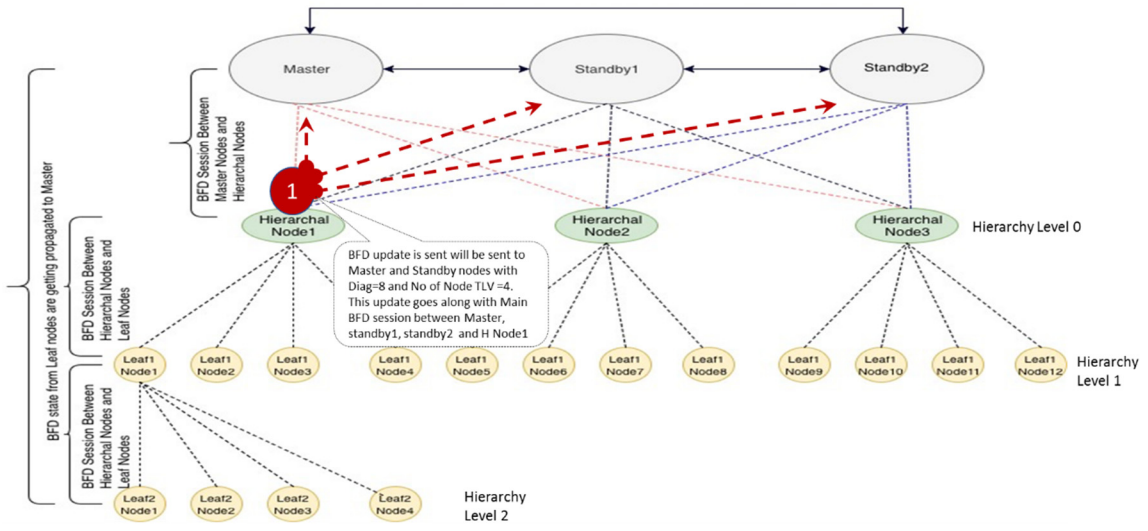


Figure 9

Figure 10 below illustrates an example BFD session between leaf1 node 1 and leaf2 node1. BFD sessions between leaf1 node 1 and leaf2 nodes 1-4 are established using BFD packet fields. A BFD packet from one of leaf2 nodes 1-4 to leaf1 node 1 has a diag field value set to 8 and a number of TLV field set to 0. A BFD packet from leaf1 node 1 to leaf2 node 1 has a diag field value set to 8 and a number of TLV field set to 0. A BFD packet from leaf1 node 1 to hierarchal node 1 has a diag field value set to 8, the number of TLV field set to the number of BFD sessions to leaf1 node1 from the leaf2 node, the node ID for the leaf2 node, an interface ID and IP address set to a value of 0 for the node level BFD, the hierarchal level set to 1, and the BFD status set to the status of the BFD session.

As further illustrated, a BFD packet from hierarchal node 1 to the Leader/Master and Standby nodes has a diag field value set to 8, the number of TLV field set to the number of BFD sessions to leaf1 node 1 from leaf2 node 1, the node ID for leaf2 node 1, the interface ID and the IP address set to a value of 0 for node level BFD, the hierarchal level set to 2, and the BFD status set to the status of the BFD session.

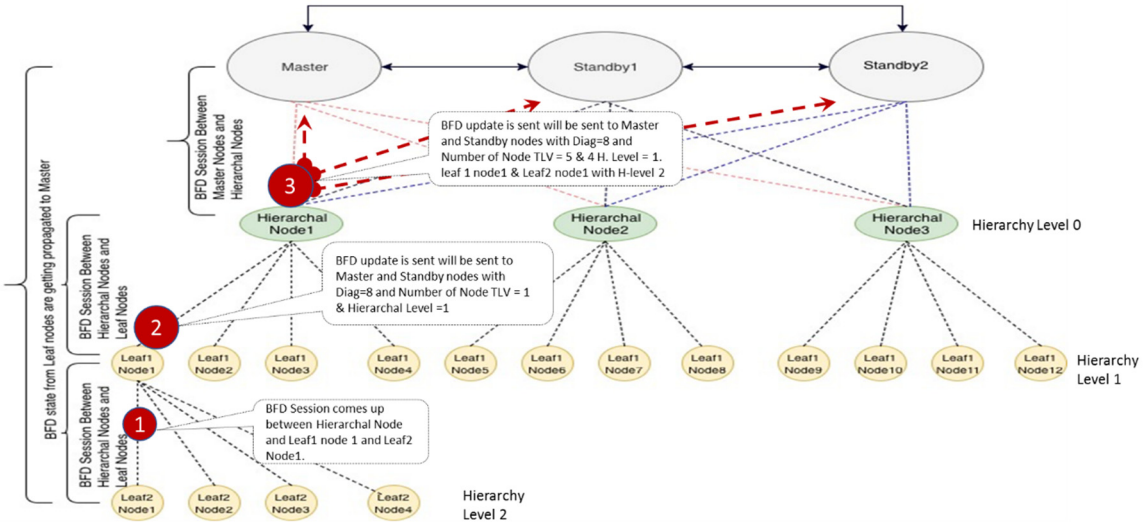


Figure 10

Figure 11 illustrates an example BFD session between leaf1 node 1 and leaf2 nodes 2-4.

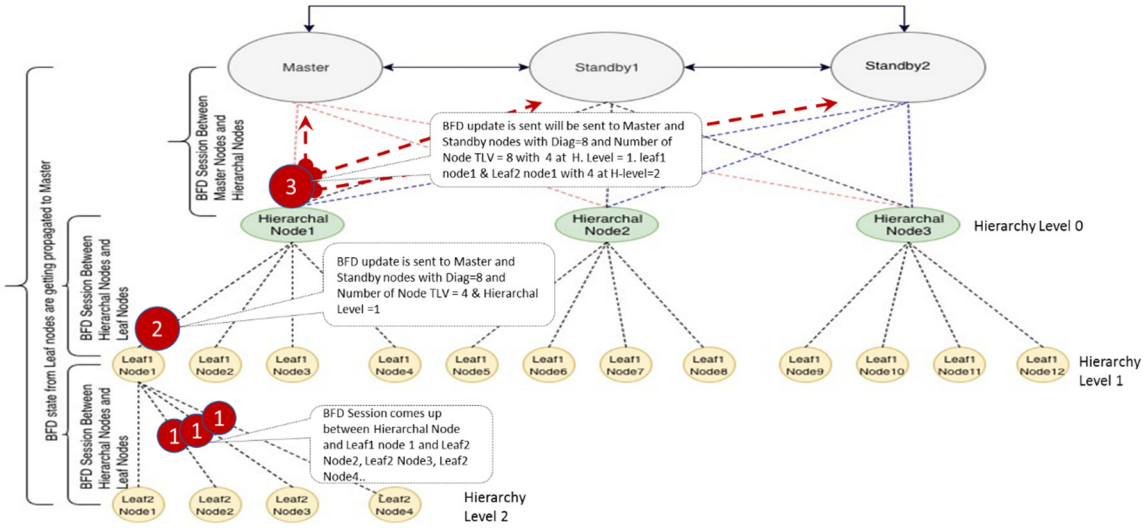


Figure 11

After the BFD sessions are established, the Leader and Standby nodes may quickly detect liveliness, failure, and convergence.

Figure 12 below illustrates example fault propagation between leaf1 node 1 and leaf2 node 1. Failure between leaf1 node 1 and leaf2 node 1 is detected because leaf2 node 1 is isolated and the BFD session goes down. As soon as the BFD session goes down between leaf1 node 1 and leaf2 node 1, leaf1 node 1 sends a BFD packet to hierarchal node 1 with the status indicated as down for this node. Leaf1 node 1 sends all other BFD sessions along with this session.

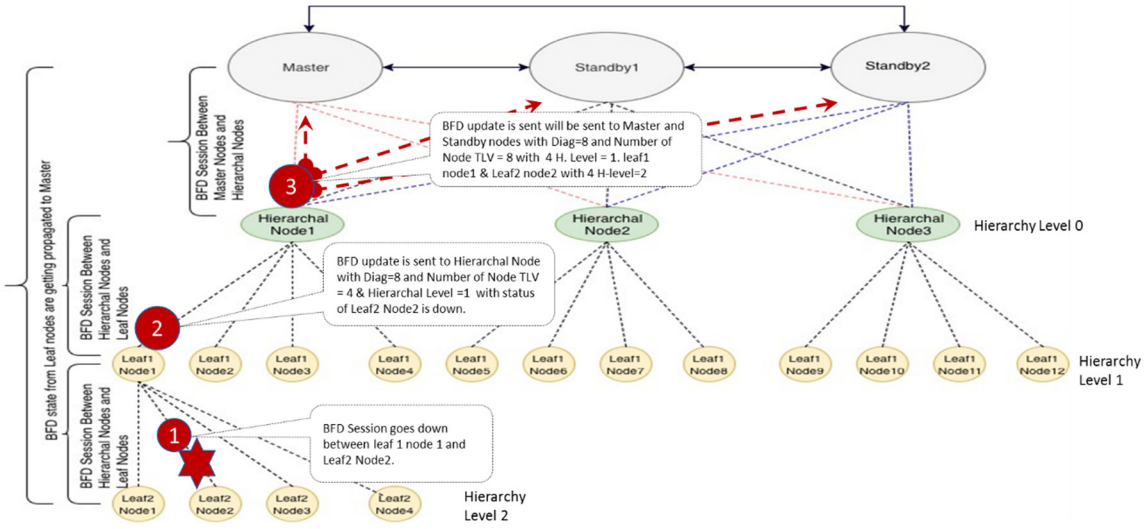


Figure 12

Figure 13 below illustrates example fault propagation from leaf1 node 1 to the Master node. Failure between leaf1 node 1 and hierarchal node 1 is detected because leaf1 node 1 is isolated and the BFD session goes down. As soon as the BFD session goes down between hierarchal node 1 and leaf1 node 1, hierarchal node 1 sends a BFD packet to the Master and Standby nodes with the status indicated as down for hierarchal node 1 and all leaf2 nodes below this hierarchy.

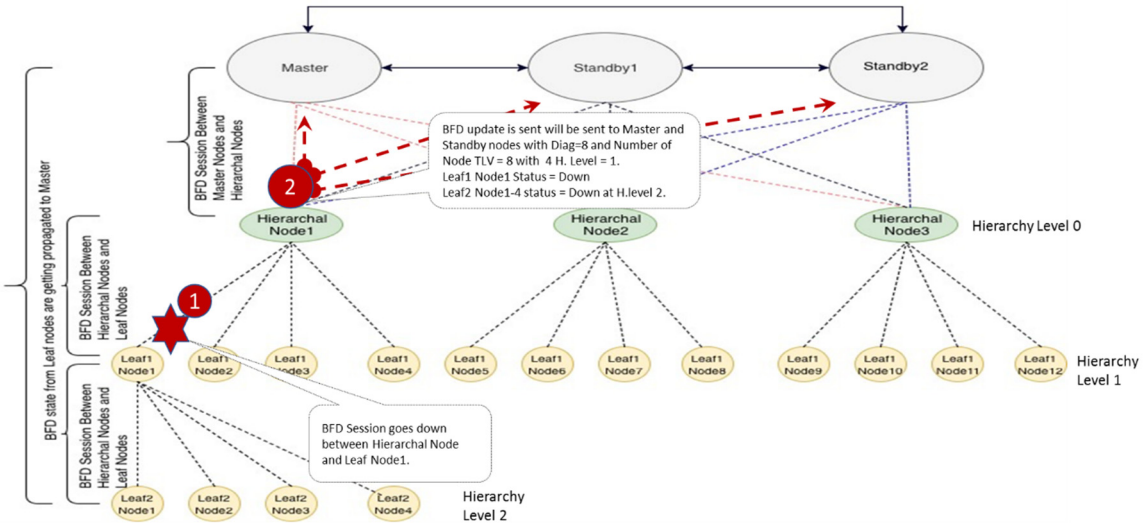


Figure 13

The detection of failure may occur quickly (e.g., 10ms) as most vendors support a 3.3ms period for BFD. There may be a small propagation delay from a leaf node to a Master node, but this mechanism is nonetheless very fast.

In summary, techniques are described herein for hierarchal BFD to scale liveliness detection. Scalability of BFD sessions to Master and Standby nodes is a concern in current implementations. A fast detection mechanism such as BFD is required to support fast detection of liveliness of cluster nodes.