AMD

HA Overview

Balakrishnan Raman

Design Goals

- All connections setup before switchover should work reliably after planned and unplanned switchovers
- 0 downtime planned switchover, <2 sec downtime unplanned switchover
- Data packets should not be dropped due to flow replication delays
- Sync connection setup and teardown at datapath rate to support high CPS
- Sync only required packets to conserve PPS for data traffic

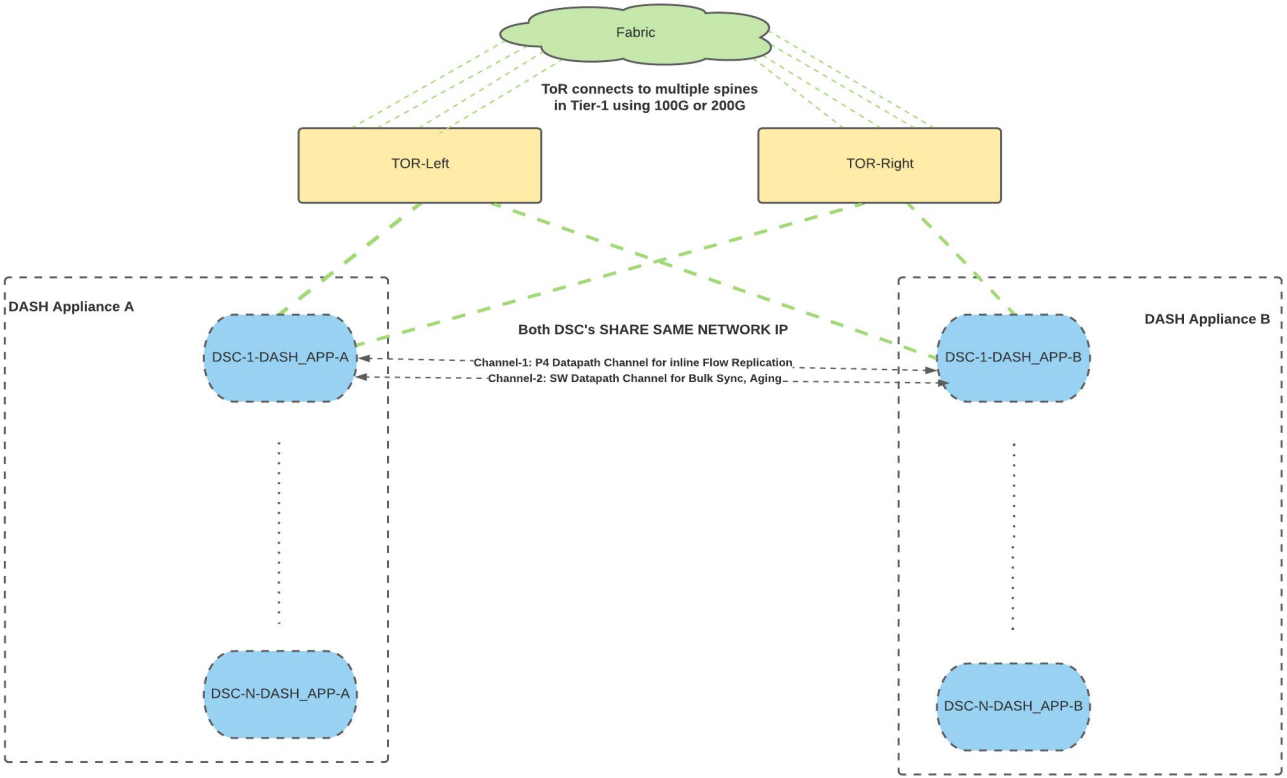
Flow Replication Options

- Async Flow Replication
  - Flows established but not synced with secondary inline
  - Flows are synced offline with secondary after forwarding packet to endpoint
  - Connections are not reliably synced so they can fail after switchover
- Inline Flow Replication with Secondary Forwarding
  - Inline flow sync using data packets and no ack for sync from the peer
  - Secondary forwards the data packet
  - Reliable sync at datapath rate to avoid connection failure after switchover and to sustain high CPS
  - Sync using data packets to avoid buffering or drops due to any sync delays
  - Connections can go out of sync if LAST-ACK/RST packets get dropped between HA pairs
- Inline Flow Replication with Primary Forwarding
  - Inline flow sync and ack for sync from secondary by forwarding data packets back to primary
  - Primary forwards the packet on receiving ack
  - In addition to advantages of the above option, connections don’t go out of sync on LAST-ACK/RST transaction due to sync acknowledgement

Inline flow replication with primary forwarding is the option we have implemented.

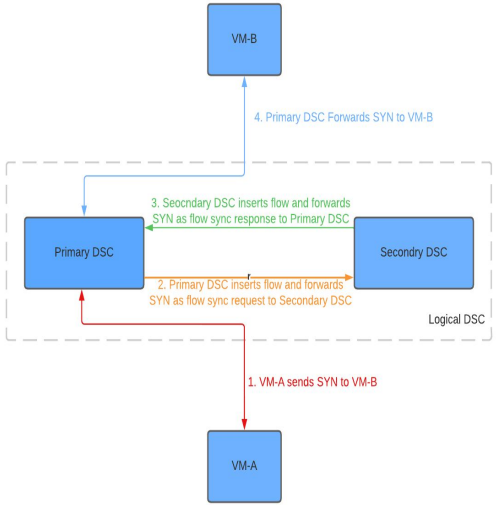
Network Deployment Topology

DSC = DPU



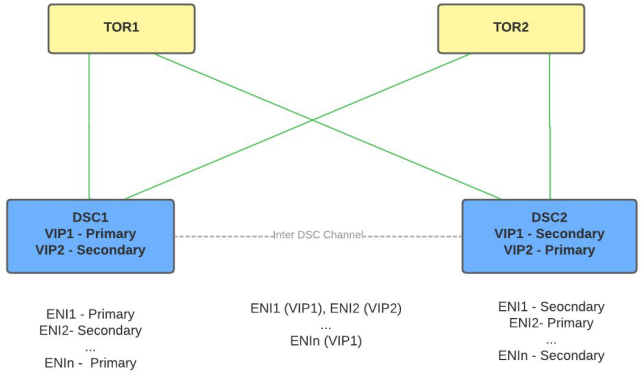
Inline Flow Replication with Primary Forwarding

- Primary and Secondary DSCs modeled as a single Logical DSC
- Flow is setup/cleaned in the logical DSC before forwarding packets to endpoint
- Co-ordinated flow setup/cleanup in Primary and Secondary by exchanging packets (syn/syn-ack/ack, fin/fin-ack/ack, rst for TCP, one or more packets for udp) with metadata in the P4 datapath
- Connection is setup and synced to Secondary using Primary's policy evaluation and rewrites
- Co-ordinated idle aging in Primary and Secondary by SW datapath for flow cleanup
- Bulk sync by SW datapath to sync on new DSC pairing with existing DSC
- Packet drops between DSCs perceived as network drops and triggers retransmission from endpoint
- Periodic Heartbeat exchanges between DSCs to detect peer reachability



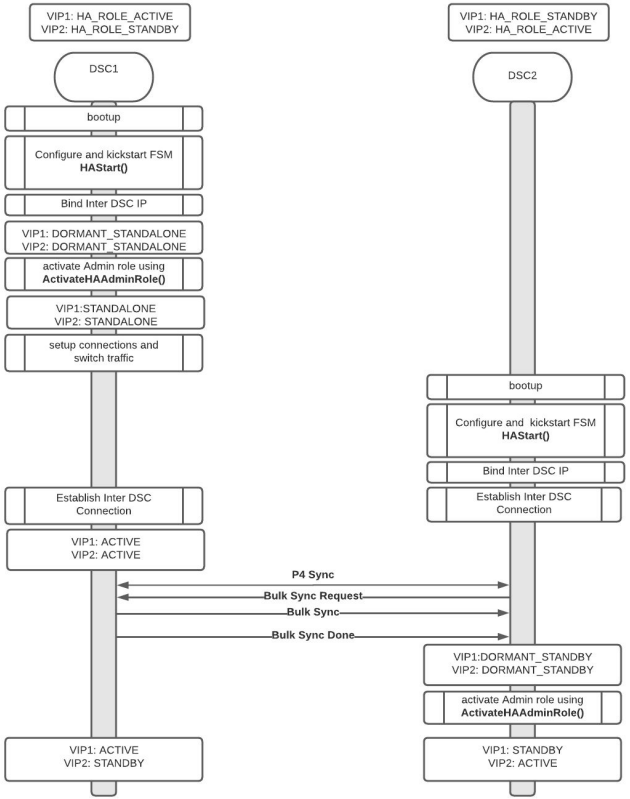
ENI Based Active-Active

- Each DSC pair has two Virtual IPs (or PA IPs)
  - DSC1 in DASH Appliance A has VIP1 as primary AND VIP2 as secondary
  - DSC2 in DASH Appliance B has VIP2 as primary AND VIP1 as secondary
- DSC uses BGP to attract traffic such that packets are sent only to DSC where the VIP is primary
- Cloud controller associates ENIs to VIP and hence which DSC an ENI is primary or secondary
- Allows both DSC to be actively forwarding for respective primary ENIs
- In the failure case surviving DSC owns and forwards traffic for both VIP1 and VIP2



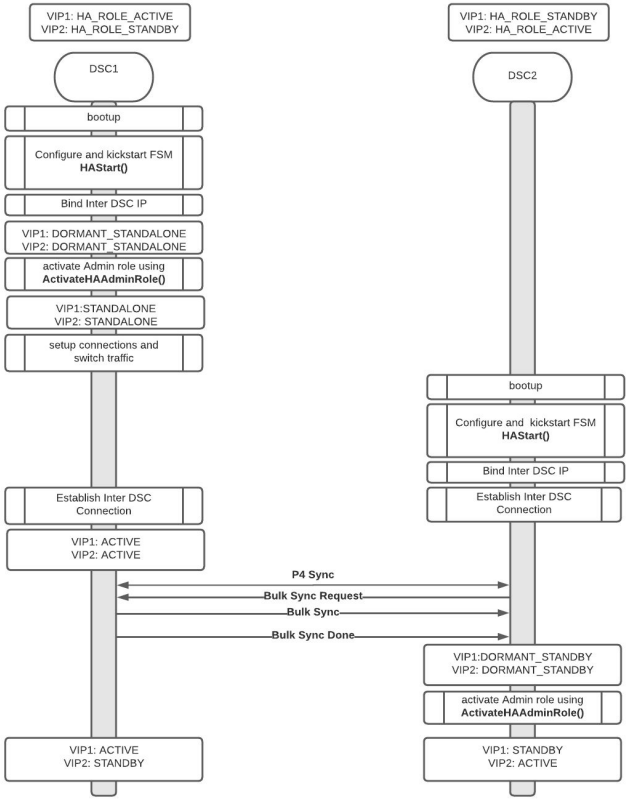
Node Pairing

- Controller kick starts HA FSM (HAStart()) after DSC bootup and configuration
- If no peer DSC exists, both DSC1 VIPs take over as Dormant Standalone
- In Dormant role, VIP is not advertised but ready for ha pairing and state syncing
- Controller triggers FSM (ActivateHAAdminRole()) to advertise VIPs and to take over admin roles
- If no peer DSC, both DSC1 VIPs become Standalone
- After DSC2 bootup and HA FSM start, DSC1 takes over as Active for both VIPs
- DSC1 bulk syncs existing connection state to DSC2 using SW datapath channel
- New connections are simultaneously synced using P4 channel
- DSC2 becomes Dormant Standby after bulk sync, where states are continuously synced but VIPs are not advertised
- On DSC2 FSM trigger, one VIP becomes Standby; Other VIP preempts role and becomes Active



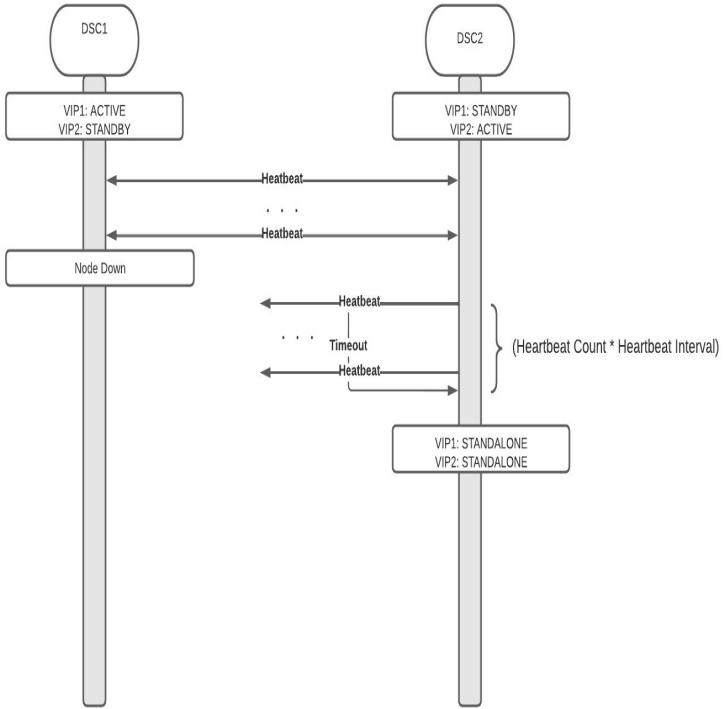
P4 Sync and Bulk Sync on Node Pairing

- Sync required for the following on node pairing:
  - Already setup connections
  - Teardown (aging or endpoint termination) of existing connections
  - New connections getting setup during pairing
  - Flow re-simulation due to config change
- Track existing connections vs connections newly getting setup during pairing
- SW datapath does bulk sync of all existing connections and aging
- P4 datapath does inline sync of connection setup and termination



Unplanned Switchover

- Software driven periodic heartbeat exchanges to detect peer reachability
- Configurable heartbeat timeout and count
- Heartbeat sent every heartbeat timeout interval
- Consecutive heartbeat misses upto heartbeat count result in role takeover as standalone



Planned Switchover

- To gracefully bring down a node for maintenance without disruption to connections
- Both VIPs become Active in DSC2 upon planned switchover trigger (StartHASwitchover()) in DSC1
- DSC1 withdraws VIP routes and after timeout1 (route convergence timeout) notifies DSC2 to takeover as Standalone
- DSC2 waits for timeout2 (inter DSC channel flush timeout) and takes over as Standalone