HPE Virtual Connect traffic flow with HPE Synergy

Contents

Introduction ................................................................ .................................................. 2
HPE Synergy composable fabric ................................................................ ........ 2
Virtual Connect technology and HPE Synergy. ..................................................... 4
  Network types supported by HPE VC technology .............................................. 4
  Network management by HPE OneView .......................................................... 5
HPE VC features ................................................................ ....................................... 7
  Uplink sets ........................................................................................................... 7
  Edge-safe operation ............................................................................................ 7
  FlexQbg support ................................................................................................. 9
  Loop protection ................................................................................................. 9
HPE VC configurations for fault tolerance and load balancing ...................... 10
  Active-Standby configurations...................................................................... 10
  Classic Active-Active configurations............................................................... 14
  MLAG traffic flow ............................................................................................. 16
Advanced VC network features .......................................................................... 21
  IGMP Snooping ............................................................................................... 22
  Pause flood protection ...................................................................................... 22
  Storm control ................................................................................................ 22
Resources ............................................................................................................... 23
Introduction

HPE Synergy, unlike other platforms, is architected from the ground up as a truly composable infrastructure. Through a single user interface—HPE OneView—all compute, storage, and fabric resources become composable, and can be configured and optimized for any application.

HPE Virtual Connect (VC) technology provides a wire-once, edge-safe, and change-ready environment that makes it easy for administrators to manage dynamic network environments. Introduced by HP (now Hewlett Packard Enterprise) in 2007, HPE VC technology has become dominant for server-to-server and server-to-network connections in HPE BladeSystem environments.

The HPE Synergy composable fabric enhances the HPE VC experience by adding fabric disaggregation, frictionless scaling, and flexible bandwidth pools—all managed by HPE OneView, a template-based software-defined infrastructure (SDI) with a rich set of interconnect features. This document describes control and data traffic flow within the composable fabric of an HPE Synergy system. It explains how the HPE Synergy architecture uses HPE VC technology to deliver simple, composable bandwidth resources. This document assumes basic familiarity with HPE Synergy and HPE VC concepts, and specifically discusses how HPE VC technology is implemented in HPE Synergy.

HPE Synergy composable fabric

The HPE Synergy frame is the basic unit of hardware infrastructure upon which management of all system resources—compute, storage, and fabric—are based. The HPE Synergy Composer is the heart of HPE Synergy management. Powered by HPE OneView, HPE Synergy Composer manages all components within the HPE Synergy infrastructure, including fabric resources, the HPE Frame Link Module, and interconnect modules (ICMs). As shown in Figure 1, each HPE Synergy frame accommodates multiple management, link, and interconnect modules for complete redundancy. Multiple frames can be linked to form a ring, and into larger groups or domains through a dedicated 10 Gb management network that can be connected to a customer’s management LAN.

![Figure 1. Management and interconnect components of the HPE Synergy 12000 frame enclosure](image)

The HPE Synergy composable fabric includes three fabric ICM types: HPE VC ICMs, switch ICMs, and pass-through ICMs. HPE VC ICMs are managed through the HPE Synergy Composer. Switch ICMs can be managed through a command-line interface (CLI) and monitored through the HPE Intelligent Management Center (IMC).

Figure 2 shows six HPE Synergy interconnect bays designed for full redundancy of three air-gapped fabrics. Each HPE Synergy compute module can accommodate between 3 to 10 mezzanine (M) cards, with each mezzanine card connecting to Fabric 1, 2, or 3.
HPE Synergy fabric uses a master/satellite architecture (Figure 3) to consolidate data center network. For high availability and full fabric redundancy, there are two Master HPE VC ICMs per composable fabric, with the second Master HPE VC ICM installed in a frame other than the first Master ICM. The actual physical view of the three frames shows a Master HPE VC ICM in frame 1, Master and Satellite ICMs installed in frame 2, and Satellite ICMs in frame 3. The HPE VC ICM acting as Master contains intelligent networking capabilities that extend connectivity to satellite frames. Therefore, a Master ICM is similar (in a hierarchical function) to a top-of-rack (ToR) switch.
Uplinks from the Master HPE VC ICM can connect to Ethernet LAN and/or FC/FCoE networks. The downlink ports are connected to the compute modules through the HPE Synergy midplane. If 10 Gb ICMs are used, the Master HPE VC ICM can connect to up to 60 compute modules or five frames. With 20 Gb ICMs, the Master HPE VC ICM can connect up to 36 compute modules or three frames. Using pass-through modules would require connecting to a ToR using 1-1 ports while having a Master HPE VC ICM hides those extra connections and reduces the number of ports at the aggregation layer. With a Master HPE VC ICM, legacy ToR roles remain, but the complexity has been reduced since any additional satellite frames are connected to the ICM instead of a rackmount ToR switch.

**Virtual Connect technology and HPE Synergy**

**HPE Virtual Connect (VC) technology** provides a wire-once, edge-safe, and change-ready environment that makes it easy for administrators to manage dynamic network environments. HPE VC technology virtualizes compute module connectivity by adding a hardware abstraction layer that removes the direct coupling between servers and LAN or SAN network infrastructure.

HPE VC technology uses standard Ethernet networking to carry virtual LAN (VLAN) tagged and untagged traffic through the composable fabric. VLAN tagged networks can be aggregated into a trunk and assigned to a specific set of uplink or downlink ports forming a broadcast domain.

**Network types supported by HPE VC technology**

HPE Synergy and HPE OneView support the following types of networks:

- Ethernet, including tagged, untagged, or tunnel networks
- Fibre Channel, including fabric-attach (SAN) Fibre Channel (FC) connections and direct-attach (Flat SAN) Fibre Channel connections
- Fibre Channel over Ethernet (FCoE) for storage networks where storage traffic is carried over a dedicated Ethernet VLAN

**Ethernet untagged network**

An untagged network is a dedicated network with a dedicated set of uplink ports used to pass traffic without VLAN tags. Any tagged packets are dropped, and forwarding is by MAC address. Administrators may want to configure an untagged network for iSCSI storage traffic or set up networks without configuring VLANs. When an uplink or a server s-channel is a member of a dedicated network, it cannot be a member of any other network.

**Ethernet tunnel network**

A tunnel network is a dedicated network with a dedicated set of uplink ports used to pass either a group of VLANs or untagged traffic. The presence of a VLAN tag is ignored when forwarding traffic to a dedicated set of uplinks and server connections. A tunnel network is implemented internally as a QinQ network. All traffic exiting an untagged or tunnel network will have the internal S-VLAN tag removed. This S-VLAN comes from internal dedicated range of VLANs. A minimum of 60 VLANs are reserved from the default range of 4034 to 4094. The S-VLAN range can be extended up to a maximum of 128 VLANs, and all learning and forwarding are based on this internal S-VLAN.

Similar to an untagged network, when an uplink or a server s-channel is a member of a tunnel network, it cannot be used by any other network.

**Ethernet network**

Ethernet networks are regular 802.1Q virtual LANs (VLANs), allowing multiple Ethernet networks to use the same physical uplinks. All Ethernet network frames either explicitly contain a C-VLAN tag or, if configured to accept untagged frames, the untagged frames are mapped to a C-VLAN. All learning and forwarding behavior are based on the C-VLAN tag.

Multiple Ethernet networks can be assigned to the same uplink or uplinks. Traffic entering uplinks must contain configured C-VLAN tags before being accepted. Traffic containing C-VLAN tags that are not configured for a given uplink is dropped. Untagged traffic is accepted if and only if one of the Ethernet networks is configured as native. If no network is designated as native, untagged traffic is dropped. VLAN 1 is a preferred default VLAN for untagged traffic but can be changed to any VLAN.

**Note**

On uplinks, all traffic egresses the HPE OneView Synergy domain with the appropriate C-VLAN tags. Native network traffic will ingress/egress untagged.
With HPE OneView, there are two possibilities for configuring Ethernet networks on the downlink:

• **Single network assignment (untagged traffic only):** Only one Ethernet network is assigned to a connection. All frames that egress to the connection will be untagged. Any frame that ingresses from the s-channel must be untagged as well. Any C-VLAN tagged frame that ingresses from the connection will be dropped, even if the C-VID matches the VID of the Ethernet network.

• **Network Set assignment (tagged and untagged traffic):** A Network Set is a collection of Ethernet networks that can be assigned to an s-channel port. Only Ethernet networks can be grouped into a Network Set. FCoE, untagged, or tunnel networks cannot be included. An Ethernet network can be a member of different Network Sets. For each Network Set, only one Ethernet network can be designated as untagged. The behavior of this network is similar to the single network assignment described previously. Note that unlike uplink native network behavior, downlink untagged networks always transmit untagged frames only.

**FCoE network**

Fibre Channel over Ethernet (FCoE) is a technology that encapsulates Fibre Channel frames over an Ethernet network. FCoE Priority Flow Control (PFC), Enhanced Transmission Selection (ETS), and Data Center Bridging Exchange (DCBX) are supported for lossless Ethernet. Two types of Ethernet traffic are involved: FCoE Initialization Protocol (FIP) (Ethertype = 0x8914) and FCoE (Ethertype = 0x8906).

FIP traffic entering uplinks may or may not contain VLAN tags. Untagged FIP traffic is accepted through the default VLAN and egresses the HPE OneView stacking domain without a tag. Tagged FIP traffic will be forwarded based on the C-VLAN tag and egresses the HPE OneView stacking domain with the C-VLAN tag.

FCoE traffic entering uplinks must contain configured C-VLAN tags before being accepted. Traffic containing C-VLAN tags that are not configured for a given uplink is dropped. FCoE traffic egresses the HPE OneView stacking domain with the appropriate C-VLAN tags.

**Note**

Only one FCoE network can be assigned to an s-channel. HPE Synergy supports up to 32 FCoE networks per uplink set. An FCoE network may not be assigned to an uplink set spanning more than one interconnect.

**Network management by HPE OneView**

HPE OneView manages HPE Synergy networking resources as, in ascending order, interconnects, logical interconnects, and logical interconnect groups. The interconnect is the basic level of network resource and represented by an ICM. A logical interconnect (LI) is a single administrative entity consisting of a set of one or two master interconnects and up to four satellite interconnects per master. A logical interconnect group (LIG) consists of multiple LIs and represents the top level of network management.

All ports of an interconnect, LI, or LIG are assigned a role. Port roles are an important concept with HPE Virtual Connect since they help dictate the traffic forwarding behavior for a port. Each port may be assigned one of the following roles:

• **Uplink**—A network-facing port with the following attributes:
  – Must be a member of an uplink set (and potentially a member of an LAG or Active-Standby set)
  – Eligible for forwarding traffic to/from the external network

• **Downlink**—A server-facing port with the following attributes:
  – Uses active loop detection and prevention mechanisms
  – Implements HPE VC Flex Technology (when connected to multiple FlexNICs and/or FlexHBAs) that allows bandwidth tuning for optimized connectivity with different applications. Bandwidth on Physical Function NIC ports can be set from 200 Mbps to 20 Gbps (in increments of 200 Mbps per NIC\(^1\))

---

\(^1\) The combined bandwidth of NICs cannot exceed port bandwidth (that is, 20 Gbps).
• **Cluster**—Peer ICM interconnect port (where one ICM is designated the master for protocols) with the following attributes:
  – Provides peer-to-peer connection between ICMs, enabling state coordination for features such as MLAG
  – Enables Ethernet traffic to flow between ICMs for direct ICM-to-ICM communication although FCoE traffic is not permitted to cross the cluster links between ICMs
  – Used to exchange control plane protocol data enabling state coordination for features such as MLAG

• **Service**—Port that connects to the HPE Image Streamer, with the following attributes:
  – Restricted to connecting with the HPE Image Streamer and will only be enabled if 1) it is directly connected to the image service and 2) the image service has been enabled in HPE OneView.
  – Must enable forwarding based on attached port.
  – Permits traffic only to/from the HPE Image Streamer module.
  – Neighbor is verified using LLDP.

• **Unused**—Default port state, that is, role, configuration: Port does not forward traffic but is able to collect neighbor LLDP information (for example, LLDP enabled with port in the blocked state).

Uplink sets define the potential members of a link aggregation group (LAG) or multi-module LAG (MLAG). Only one port/LAG is active at any given time—all others are in standby. Uplink sets map data center networks to physical uplink ports. If no uplink sets are defined, the logical interconnect cannot connect to data center networks, and the servers attached to the downlinks of the logical interconnect cannot connect to data center networks. Uplink sets exhibit edge-safe behavior, which requires that traffic must never be forwarded between two different uplinks or uplink sets to prevent formation of a network loop.

Figure 4 shows a typical LI configuration with two ICMs. The Q1 ports handling traffic of VLAN 100 thru 199 of both ICMs are members of a common uplink set to form either an LAG or Active-Standby pair.

![Figure 4. LI configuration](image-url)
HPE VC features

HPE VC provides a simple and safe network abstraction used for interconnecting servers with the data center storage and data fabrics. HPE VC hardware abstraction allows you to configure and connect physical and virtual servers. HPE OneView allows you to change, move, or redeploy any server within a single HPE VC domain. The following sections describe key HPE VC features that provide safe network abstraction.

Uplink sets

HPE VC uplinks are associated with the networks to provide network connectivity. Assigning more than one HPE VC uplink to the same set of networks provides network redundancy, load balancing, or both for the servers assigned to that network (within an enclosure). An uplink set without uplink ports can provide server-to-server communication. Server NIC ports assigned to the same network can communicate directly with one another without exiting the LI. As a rule, traffic is never forwarded between two uplink ports within the same uplink set or between two uplink sets to prevent formation of a network loop.

Edge-safe operation

To qualify as being edge-safe, an interconnect must exhibit all of the following properties:

- Behave as a leaf node (that is, edge or access port) to the external network
- Have port assigned roles with well-defined forwarding semantics
- Not forward traffic across uplinks

As a leaf node to the external network, edge safe refers to an Ethernet interconnect that does not participate in Spanning Tree Protocol or other external loop prevention protocols on uplink ports. More generally, this means an HPE VC interconnect or fabric does not expose its internal topology and protocols to the external network if doing so could cause an outage or destabilize the external network. If two or more interconnects are clustered, their collective forwarding behavior must be indistinguishable from that of a single leaf switch.

To ensure loop-free forwarding behavior, ports are assigned roles with well-defined behavior and state machines. Rigorous discovery protocols are used to confirm that a port's neighbor is consistent with its role before the port is permitted to pass traffic. The port role defines the permitted uses, enabled protocols, and required edge-safe behavior of the port. An HPE VC module is designed to never create or cause a network loop.

HPE VC technology offers edge-safe operation in single- or multi-module configurations (Figure 5). The module does not participate in the Spanning Tree, and the forwarding/bridging of traffic across network-facing ports is not allowed. All traffic through the port is blocked unless it is configured and assigned a role. With a dual module configuration, the two ICMs operate as a single logical interconnect that behaves as a single administrative entity that ensures edge-safe operation. Stacking links can forward traffic between ICMs and path redundancy is supported without introducing loops.

You cannot use a network to bridge multiple HPE VC uplink ports for connectivity between two external devices or to connect two external networks. HPE VC ICMs are not transit devices and cannot be configured as such.
Figure 5. Edge-safe operation with HPE VC technology
FlexQbg support
HPE Flex10Qbg technology leverages the Edge Virtual Bridging (EVB, or the IEEE802.1Qbg) standard to create s-channels and physical functions (PFs) in the compute module. Extensions of the Link Layer Data Protocol (LLDP) such as Data Center Bridging Exchange (DCBX) and Channel Discovery and Configuration Protocol (CDCP) use Type, Length, and Value (TLV) elements to allocate s-channels (Figure 6).

HPE FlexQbg adds the following functions to the 802.1Qbg standard:

- Control link status of channel
- Channel description: type, traffic class, BW (max. and min.), channel ID

![Diagram](image)

Figure 6. FlexQbg operation

Loop protection
Depending on the role of the HPE VC Ethernet port, HPE VC can use several loop avoidance mechanisms. As described previously, an HPE VC Ethernet port can be an uplink, a downlink, or a clustered link. HPE VC ICMs will combine multiple uplink ports into a single aggregation (LAG or MLAG); therefore, it ensures that only a single physical or logical port is actively forwarding traffic while the remaining uplink ports remain on standby. By ensuring there is only one network path from HPE VC ICMs to the external Ethernet fabric, HPE VC ICMs avoid network loops between HPE VC ICM uplink ports and the external switch ports.

If a user mistakenly bridges two NIC ports, a network loop may occur, but the HPE VC ICM network loop prevention (NLP) feature detects bridged NIC ports and disables the offending ports. The loop prevention feature is activated when a downlink port receives a designated multicast frame. These may be Per-VLAN Spanning Tree + Bridge Protocol Data Unit (PVST+ BPDU) packets from a server NIC or internally injected probe frame sent to a well-defined multicast address.

Quality of service
The quality of service (QoS) feature enables administrators to configure traffic queues for different priority network traffic, categorize and prioritize ingress traffic, and adjust priority settings on egress traffic. Administrators can use these settings to ensure that important traffic receives the highest priority handling while less important traffic is handled at a lower priority. Network traffic is categorized and then classified. After being classified, traffic is given priorities and scheduled for transmission. For end-to-end QoS, all hops along the way must be configured with similar QoS policies of classification and traffic management.
QoS configuration is defined in the logical interconnect group and applied to the logical interconnect. QoS statistics are collected by the interconnects.

Traffic prioritization happens because of two things in an end-to-end QoS policy. At the interconnect, the packets are egressed based on the associated queue bandwidth. Egress dot1p remarking helps achieve priority at the next hops in the network. If the queue egress traffic is remarked to a dot1p value and that dot1p value is mapped to a queue in the next hops with higher bandwidth, then these packets in the end-to-end network are treated with higher priority.

Three predefined QoS configuration types are provided:

- **Pass-through**: Incoming non-FCoE packets are not classified or altered. There are no traffic classes, maps, or rules. This is the default behavior.
- **Custom with FCoE**: This enables QoS and allows customized configuration including the FCoE class. It includes two system classes (FCoE Lossless and Best Effort) and up to six custom classes for non-FCoE Ethernet traffic. Administrators can configure the traffic class parameters and the traffic classification.
- **Custom without FCoE**: This enables QoS and allows customized configuration, which does not include an FCoE class. It includes one system class (Best Effort) and up to seven custom classes for non-FCoE Ethernet traffic. Administrators can configure the traffic class parameters and traffic classification.

**HPE VC configurations for fault tolerance and load balancing**

The following sections describe network configurations supported by HPE Synergy VC ICMs. These configurations provide various levels of fault tolerance and load distribution for Ethernet network topologies associated with uplink ports and compute module NIC ports. The configurations include:

- Active-standby
- Active-active
- Multi-module link aggregation (MLAG)

For all network configurations listed previously, HPE VC ICMs implement load-balancing algorithms to distribute Ethernet frames across the physical ports within the LAG or MLAG. The primary concern in distributing frames across multiple physical ports is frame ordering. An application or TCP endpoint receiving frames out of order could result in it assuming, incorrectly, that a packet was lost, thus impacting the application and TCP's perceived network congestion. Optimally, for any communication between two network devices and for specific application flows, the network infrastructure delivers the frames for that flow in order. This means that any load-balancing algorithm used by the LACP must load balance frames while maintaining frame ordering within a flow.

HPE Synergy VC ICMs support load distribution based on the source and destination MAC addresses for all packets that traverse the ports within link aggregation groups (LAGs) or MLAGs on the module. The value of the source and destination MAC address in the packet are used to calculate the rule for distributing the Ethernet traffic among the aggregated links and to select the port on which the traffic should flow.

Also discussed is the impact of (with or without) local port affinity (LPA), and how each HPE VC configuration affects server link speed and stacking link congestion.

**Active-Standby configurations**

Active-Standby is a legacy method of using HPE VC technology for a compute module-to-network interconnect. Figure 7 shows a basic network architecture using one LI in a redundant, active-standby configuration. Note that there is no link aggregation configured on the top-of-rack (ToR) switches. In Figure 7, the upstream network switch connects a network to a single port on each HPE VC ICM.

---

**Note**

In practice, Active-Standby can apply to an LAG on a single ICM with other ICM uplinks acting in standby.
In this configuration, unused uplinks limit bandwidth to the external network but do offer a failover mechanism. Traffic is not evenly distributed across ICMs and excessive loading on stacking links can occur.

![Traffic flow for HPE VC Active-Standby configuration](image)

Figure 7. Traffic flow for HPE VC Active-Standby configuration

Figure 8 shows traffic flow for an HPE VC Active-Standby configuration for an uplink set without MLAG to an aggregation switch and then to a Remote Device (such as another compute module, server, or another frame). When an uplink set is defined without enabling MLAG on the upstream (ToR) switch, then one port (or LAG on a single ICM) is selected to be active (ICM B and ToR-SW2 in this example) while the remainder (ICM A and ToR-SW1) are placed in standby mode. In this case, all traffic to the compute modules will pass through the single uplink port and there will be no load balancing of traffic. In this instance, both ToRs are configured with port P1 enabled. Since HPE Synergy ICMs only placed one port (Q1 on ICM B in Figure 8) into the forwarding state, this ensures that the HPE VC ICM will not create a loop within the attached Ethernet/VLAN network.

**Note**

There may be multiple LAGs for a given uplink set. At most, one LAG will be placed in the active (forwarding) state and all the others will be in the standby (blocking) state. If the active LAG should fail, then a standby LAG will be selected and made active.
Figure 9 shows traffic flow for an HPE VC Active-Standby configuration for link failure, without MLAG for a failover mode. When an uplink fails on HPE Synergy ICMs without MLAG, all traffic is redirected to the standby port (that is, one of the standby ports is selected to become active). As shown in Figure 9, a fault (indicated by the X on the link between ICM B and ToR-SW2) will cause the ICM to update its MAC forwarding table. Likewise, ToR-SW2 will detect the link failure and update its MAC forwarding table as shown. The fact that ToR-SW1 doesn’t list the server MAC address (MAC.X1 and MAC.X2) is not a problem since this results in packets destined to the missing MAC address to be flooded to all ports that are members of the same VLAN. The potential problem is if Core-SW3 retains invalid forwarding information resulting in any packets from the Remote Device to Server X being sent to ToR-SW2, where they will be dropped (referred to as a black hole situation).

A black hole is prevented or mitigated by one of the following:

- Compute Module X sends a packet, which is received by Core SW3 causing its MAC table to learn the updated information. Compute modules can generally expected to send traffic bi-directionally, forcing MAC table updates within the network.

- Core-SW3 MAC table entries for Compute Module X time-out and the switch remove the invalid entries. All Ethernet switches define a MAC table entry aging value where entries are removed if a packet is not received from the source MAC address within the specified time-out. Most vendors default to five minutes. This bounds any outage to an average of 2.5 minutes and worst case of 5 minutes (although the administrator could change the age interval).

- User manually flushes/clears the incorrect MAC table entries in SW3’s MAC table.
Figure 9. Traffic flow for HPE VC Active-Standby configuration for link failure, without MLAG (failover mode)
Classic Active-Active configurations

Active-Active is another method of connecting HPE VC to the data center. Active-Active configurations require two LLs and use all available uplinks across the ICMs, maximizing bandwidth to external networks and reducing the oversubscription rates for compute module-to-core network traffic. There are no stacking links, and compute modules are responsible for load balancing across ICMs cross-connected to the uplink infrastructure. Figure 10 shows a typical traffic flow for an Active-Active HPE VC connection. Note that Figure 10 is a simplified diagram. MAC.X1 traffic can use either MAC.Y1 or MAC.Y2 paths (as can MAC.X2 and reverse traffic), but there will not be any traffic over stacking links, only over the high-bandwidth links between tiers.

The issue with active-active as described and shown in Figure 10 is that link errors between the ICM and ToR would not be detected by the server, causing a black hole for traffic until the failed link is repaired. To address this potentially crippling fault, HPE has created the SmartLink feature. Supported with HPE OneView 4.0, SmartLink enables a compute module to discover a fault in a path and redirect traffic accordingly. As indicated by the X Fault in Figure 11, the uplink between Q1 of ICM B and P1 of ToR-SW2 has failed. Without SmartLink, Compute Module X would continue to send traffic out its P2 port, only to have ICM B drop this traffic since it has no available uplink on which to send it.

With SmartLink, ICM B Port D1 is disabled since no corresponding uplink is available, and the compute module, detecting this loss, fails over to the alternate (P1) port.
Figure 11. Traffic flow for link failure with SmartLink in Active-Active configuration.
**MLAG traffic flow**

As an alternative Active-Active configuration, HPE Synergy VC ICMs allow MLAG configurations that logically aggregate ports across multiple modules for increased bandwidth and higher availability. MLAG can be implemented for uplink and downlink traffic flow as described in the following sections.

**Uplinks with MLAG without cross-connecting ToRs to ICMs**

Figure 12 shows the traffic flow using uplink MLAG without cross-connecting the ToRs and ICMs. Green arrows indicate traffic flow from Compute Module X Port 1 to Compute Module Y Port 1 while orange arrows denote traffic flow from Compute Module X Port 2 to Compute Module Y Port 2. ICM A Port Q1 is aggregated with ICM B Port Q2 (for both LI 1 and LI 2) for increased bandwidth and availability. All uplinks across ICMs are enabled, allowing the ICMs to provide load balancing and failover functions.

![Diagram of MLAG traffic flow](image)

**Figure 12.** Traffic flow for MLAG (uplink only, without LPA) without ToR-to-ICM cross-connects

Prior to HPE OneView 4.0, local port affinity (LPA) was not available, in which case each ICM would load balance traffic across the ICM uplinks, sending half of the traffic from each ICM over the stacking links. This can result in congestion on the stacking links since there is greater northbound bandwidth available than with inter-ICM.

To address this potential source of congestion, HPE implemented the local port affinity feature, which will only send traffic over the stacking link if there are no available uplinks in the local ICM. Note that server-to-server traffic is contained within the LI and is not forwarded to ToRs.
As shown in Figure 13, with LPA enabled (HPE OneView 4.0 and HPE VC firmware version 1.2.0 and later), traffic from the same side does not use stacking links and a local uplink is always selected if available and in the active/forwarding state (Otherwise, packets are sent to the peer ICM). Note that you can still have multiple local uplinks and the ICM will load balance across the local ports. The compute module will load balance across the ICMs. OS teaming implementations will load balance across the compute module NICs, thereby load balancing their traffic across the ICMs. The ICMs in turn will load balance traffic on an ICM to the active local uplinks. Not using LPA would result in the undesirable effect of having both the ICMs and the compute module trying to load balance across the ICMs.

Figure 13. Traffic flow for MLAG (uplink only, with LPA) without ToR-to-ICM cross-connects

Figure 14 shows traffic flow for an uplink set with MLAG and LPA to aggregation switch without MLAG on s-channel. A compute module’s teaming driver will load balance across the two interfaces. Traffic from HPE Synergy compute modules uses redundant paths through the ICMs and into the aggregation switch (ToR-SW1 and -SW2) while traffic from a Remote Device (server, compute module, another frame, and others) will be load balanced across the two paths as determined by the core switch configuration.

Configuration summary

- **ToR**: A pair configured with MLAG enabled, with SW1.P1 and SW2.P1 defined as an MLAG, and VLAN 100 enabled on both MLAGs.
- **HPE Synergy ICM**: Clustered, with a single uplink set with uplinks ICM A Port Q1 and ICM B Port Q1 defined, and VLAN 100 added to the uplink set.
- **Compute module**: Connections on Compute Module X Port P1 and Port P2. In the compute module’s OS, the teaming driver is enabled with both ports active and assigned VLAN 100. A switch-independent teaming policy is used, and the compute module/VM MAC address should always have the same interface.
If LPA is not available, then each ICM will load balance traffic across the two ICM uplinks (sending half of the traffic from each ICM over the stacking links).

With LPA enabled, a local uplink is always selected if one is available and in the active/forwarding state. Otherwise, packets are sent to the peer ICM.

**Figure 14.** Traffic flow for an uplink set with MLAG to aggregation switch (without MLAG on s-channel)
Figure 15 illustrates the circumstance of a link failure using MLAG but without MLAG on s-channel. With MLAG, when an uplink fails, all traffic will be redirected to the remaining forwarding uplinks in the MLAG group. For example, in Figure 15, a failure of ICM B Port Q1 causes all traffic to be redirected to ICM A Port Q1 over the stacking links. The MAC forwarding table references the logical port (the LAG or MLAG), so as long as at least one member port is in a forwarding state, traffic will continue to be forwarded.

In reality, whenever a port fails there is a small window of time\(^2\) when packets may be lost due to protocol convergence (time required to update state) or because some packets were accepted/queued at the physical port itself (the large egress queues are associated with the logical MLAG port).

![MAC table](image)

\[\text{MAC table}\]

<table>
<thead>
<tr>
<th>VLAN</th>
<th>D MAC</th>
<th>Port</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>MACX1</td>
<td>LAG (P1,P2)</td>
</tr>
<tr>
<td>100</td>
<td>MACX2</td>
<td>LAG (P1,P2)</td>
</tr>
<tr>
<td>100</td>
<td>MACRD</td>
<td>P3</td>
</tr>
</tbody>
</table>

\[\text{Shared MAC table}\]

<table>
<thead>
<tr>
<th>VLAN</th>
<th>D MAC</th>
<th>Port</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>MACX1</td>
<td>ICMAD1</td>
</tr>
<tr>
<td>100</td>
<td>MACX2</td>
<td>ICMED1</td>
</tr>
<tr>
<td>100</td>
<td>MACRD</td>
<td>LAG1 (CMAX,ICBMX)</td>
</tr>
</tbody>
</table>

\[\text{Failover with an MLAG/LAG will occur within} 350 \text{ ms if ports are in the same ICM or within one second if they are on different ICMs.}\]

\[\text{Figure 15. Link failure using MLAG (without MLAG on s-channel)}\]
Uplink MLAG with cross-connecting ToRs to ICMs

Figure 16 shows traffic flow using uplink MLAG with cross connects between the ToRs and ICMs. All uplinks across ICMs are enabled allowing the ICMs to provide load balancing and failover functions. The cross-connecting of links between the switching tiers help to evenly distribute traffic across spine/core switches. If LPA is not used, traffic will also be hashed across stacking links (LPA was not available prior to HPE OneView 4.0). Without MLAG on downlinks, this configuration can result in traffic at the destination ICM having to cross the local stacking link, possibly resulting in congestion.

Adding MLAG over s-channels

HPE OneView 4.0 and HPE VC firmware version 1.2.0 and later add support for MLAG over s-channels (Figure 17).

The benefits of using MLAG on compute module downlinks include the following:

- Provides seamless failover between adapter ports
- Enhances server traffic load balancing
- Allows server admins to use switch-assisted NIC teaming policies
Benefits of LPA
HPE Synergy ICMs restrict stacking to two links and sending traffic over stacking links can result in congestion. Not using LPA for uplink, MLAG configurations will enable all uplinks across ICMs. In this circumstance, load balancing will be implemented and half the traffic will be sent across stacking links. Not using LPA can also result in:

- MLAG traffic hashed across ICMs contributing to stacking link congestion
- Traffic not evenly distributed across spine or core switches
- Intercompute module LI traffic transiting stacking links (assume 50% of local traffic)
- Support for only switch-independent teaming (traffic load balancing causes inter-ICM traffic)

With LPA on uplinks, only a small percentage of traffic will be sent over stacking links. With HPE OneView 4.0, LPA is enabled by default for uplink MLAG configurations.

Advanced VC network features
Without the proper safeguards, HPE Synergy VC ICMs would have the same issues with fault tolerance, load balancing, failover, congestion, and loop avoidance as external Ethernet networks. HPE VC ICMs employ many of the industry-standard methods to manage these issues in conventional networks and add the following features:

- Internet Group Multicast Protocol (IGMP) Snooping
- Pause flood protection
- Storm control

These features are configurable through the HPE OneView GUI as shown in Figure 18 and described in the following sections.
IGMP Snooping
IGMP Snooping allows HPE VC ICMs to analyze the IGMP IP multicast membership activities and configure the hardware Layer 2 switching behavior of multicast traffic to optimize network resource use. IGMP Snooping can be enabled on a logical interconnect (LI) for all Ethernet tagged VLANs or on the individually specified VLANs for IGMP v1, v2, and v3. IGMP Snooping is not supported on tunnel or FCoE VLANs. When IGMP Snooping is enabled on an LI, the interconnect will learn the hosts’ intention to listen to a specific multicast address and any packet received from the specified multicast address will be forwarded to the hosts listening for that address. By default, the interconnect will flood IGMP traffic until it receives join requests from specific hosts. Once those requests are received, traffic will be forwarded only to the subscribers of the IGMP stream. The IGMP idle time-out interval for removing an inactive port from the multicast group can range from 130 to 1225 seconds as set by the administrator. HPE OneView purges VLANs from the IGMP Snooping VLAN list if they are inactive for 260 seconds.

Pause flood protection
Ethernet switch interfaces use pause-frame-based flow control mechanisms to control data flow. When a pause frame is received on a flow-control-enabled interface, the transmit operation is stopped for the pause duration specified in the pause frame. All other frames destined for this interface are queued up. If another pause frame is received before the previous pause timer expires, the pause timer is refreshed to the new pause duration value. If a steady stream of pause frames is received for extended periods of time, the transmit queue for that interface continues to grow until all queuing resources are exhausted. This pause flood condition severely impacts the switch operation on other interfaces. All protocol operations on the switch are impacted because of the inability to transmit protocol frames. Both port pause and priority-based pause frames can cause the same resource exhaustion condition.

HPE VC technology provides the ability to analyze server downlink ports for pause flood conditions and take protective action by disabling the port. The polling interval is 30 seconds. The SNMP agent supports trap generation when a pause flood condition is detected or cleared.

Storm control
A network traffic storm occurs when packets flood a physical interface, creating excessive traffic and degrading network performance. In this situation, broadcast, multicast, and unknown destination frames (UDFs) are flooded on all ports in the same VLAN. These storms can increase device CPU utilization up to 100%, which affects the ability to process control plane traffic and can result in network performance degradation or outages.

The storm control feature allows the network administrator to suppress excessive inbound multicast, broadcast, and destination lookup failure (DLF) packets when a specified threshold is reached. When traffic exceeds the threshold, the offending traffic will be blocked until the start of the next polling interval. The threshold protects against broadcast storms causing disruption to Layer 2 networks being propagated further. Physical interfaces are not brought down so that other traffic types continue to flow.

Storm control is supported only on physical uplink and downlink ports, not on stacking links, and requires HPE OneView 4.0 or later.
Resources
HPE Synergy
hpe.com/info/synergy

HPE Synergy Management Infrastructure: Managing Composable Infrastructure
hpe.com/h20195/v2/Getdocument.aspx?docname=4AA6-3754ENW

HPE Synergy: The first platform architected for composability to bridge Traditional and Cloud Native apps
hpe.com/h20195/v2/Getdocument.aspx?docname=4AA6-3257ENW

HPE Synergy Composer, QuickSpecs
hpe.com/h20195/v2/Getdocument.aspx?docname=c04815139

Unlock the unlimited possibilities of HPE Synergy
hpe.com/h20195/v2/Getdocument.aspx?docname=4AA6-2753ENW

HPE Bridges Traditional and New IT with Composable Infrastructure, Moor Insights & Strategy white paper
hpe.com/h20195/v2/Getdocument.aspx?docname=4AA6-4729ENW

HPE OneView 4.0 User Guide
support.hpe.com/hpsc/doc/public/display?docId=emr_na-a00037746en_us&docLocale=en_US

Introduction to HPE OneView concepts for HPE Virtual Connect and HPE BladeSystem customers
hpe.com/h20195/v2/Getdocument.aspx?docname=4AA5-1386ENW

Learn more at
hpe.com/info/synergy