White Paper

Fabric Failover Scenarios in the Cisco Unified Computing System

What you will Learn

Fabric failover is a feature that is unique to the Cisco Unified Computing System (UCS) and is available in End Host Ethernet switching mode for Ethernet traffic only. It enables system failover between virtual network interface cards (vNICs) in the fabric instead of implementing it on the host as it is typically done with a teaming driver. This feature is only available with the Cisco UCS M71KR-Q QLogic Converged Network Adapter and the Cisco UCS M81KR and Virtual Interface Card (VIC) 1280.

This white paper features an overview of the fabric failover feature and as it pertains to different operating systems running on blades and hypervisor-based technologies. Also included is a description of the FabricSync feature, available beginning with Cisco UCS Manager version 1.4, that enables dynamic MACs belonging to virtual machines (VMs) to be synced or replicated to the peer Fabric Interconnect for a fabric failover-enabled virtual network interface card (vNIC).

Cisco UCS Architecture for Fabric Failover

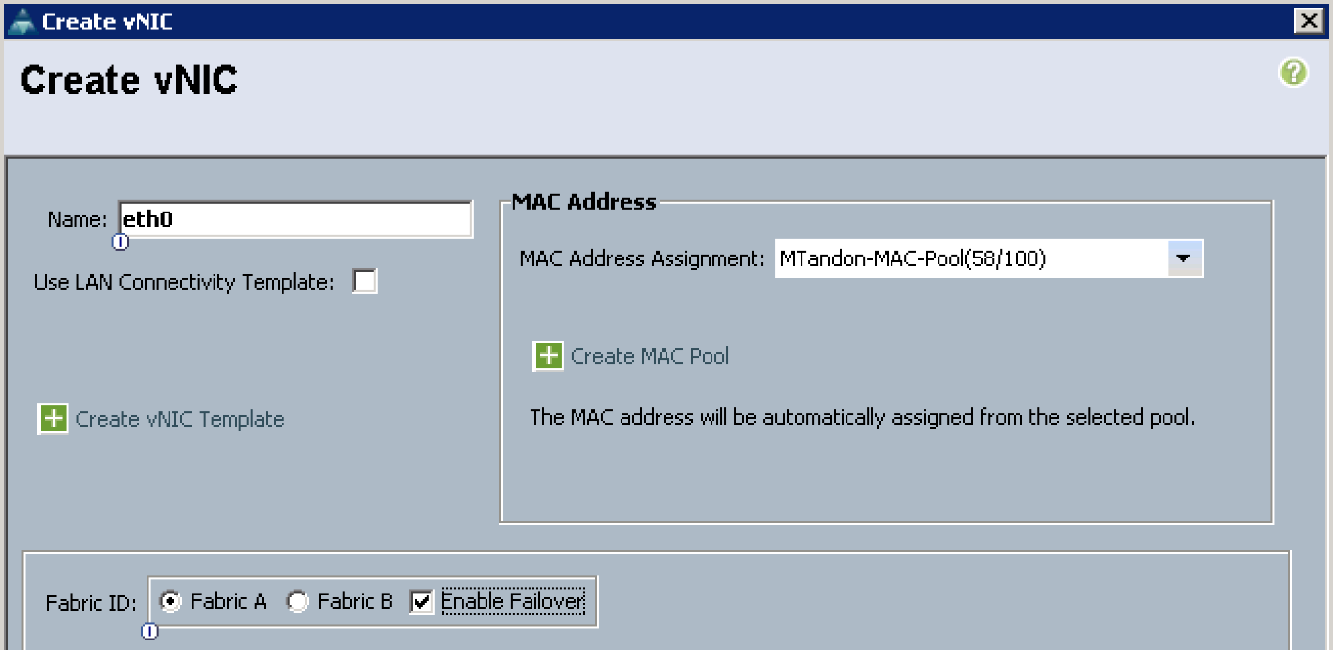
A fully redundant Cisco UCS is composed of two independent fabric planes (referred to here as Fabric A and Fabric B) with each plane comprised of a Fabric Interconnect (FI) connected to an I/O module in each blade chassis. It is important to note that the two fabric planes in the system are completely independent of each other from a data plane perspective. All network endpoints such as host adapters and management entities like the Cisco Integrated Management Controller (CIMC) are dual connected to both fabric planes and hence can work in an active-active manner.

Active/Active NIC teaming, based on the IEEE 802.3ad Link Aggregation Control Protocol (LACP) or static mode from server interfaces, is not supported in Cisco UCS and the FIs are not Virtual PortChannel (vPC) peers.

Active/Passive teaming—adapter fault tolerance (AFT), switch fault tolerance (SFT), and adaptive load balancing (ALB)—is possible using teaming software on the host. The fabric failover feature in UCS provides active/passive functionality without any teaming software requirement and configuration at the host level.

The fabric failover functionality is enabled by selecting “Fabric Failover” while creating the vNIC in UCS Manager, (Figure 1). On creating the vNIC, depending on whether Fabric A or B is primary, the vNIC is referred to as A-B or B-A.

Figure 1. Creating a Fabric Failover vNIC in Cisco UCS Manager



Create vNIC

Create vNIC

Name: eth0

Use LAN Connectivity Template: □

MAC Address

MAC Address Assignment: MTandon-MAC-Pool(58/100)

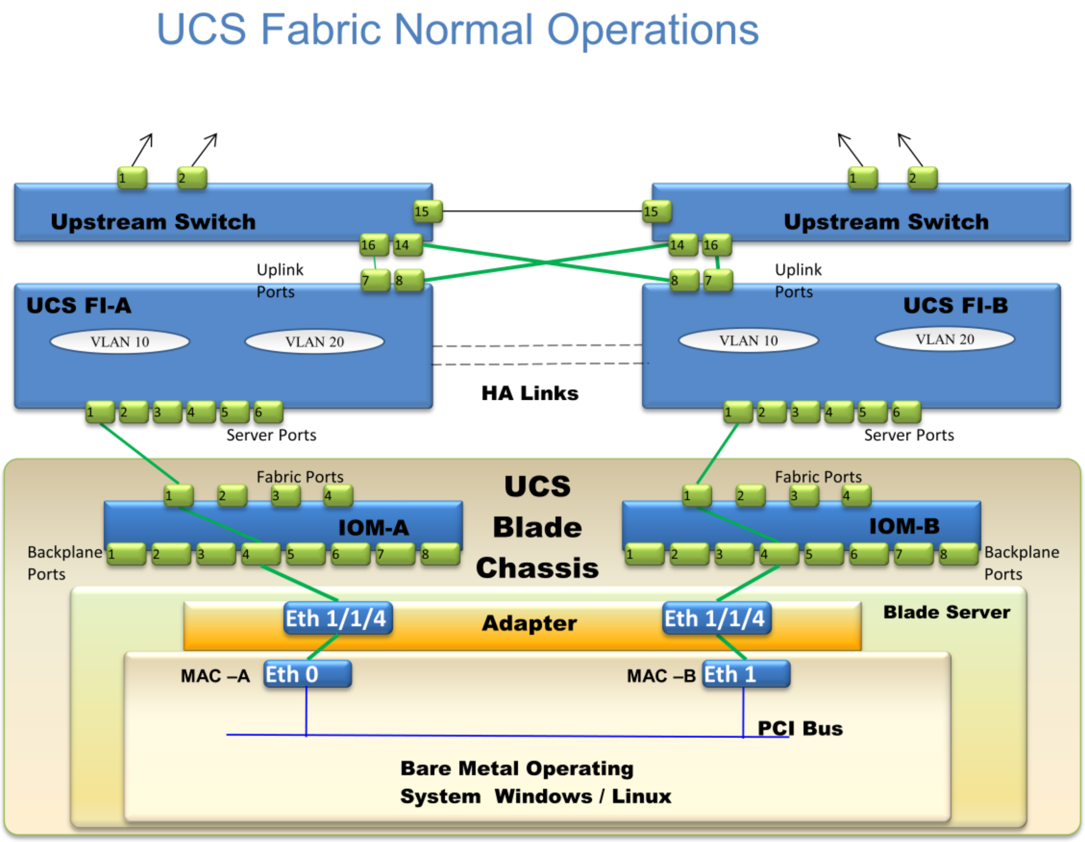
The MAC address will be automatically assigned from the selected pool.

Fabric ID: ● Fabric A ○ Fabric B ☑ Enable Failover

Fabric Failover on Windows or Linux OS Running on Bare Metal Blades in UCS

The IP traffic path for a fully redundant Cisco UCS system with an adapter is shown in Figure 2.

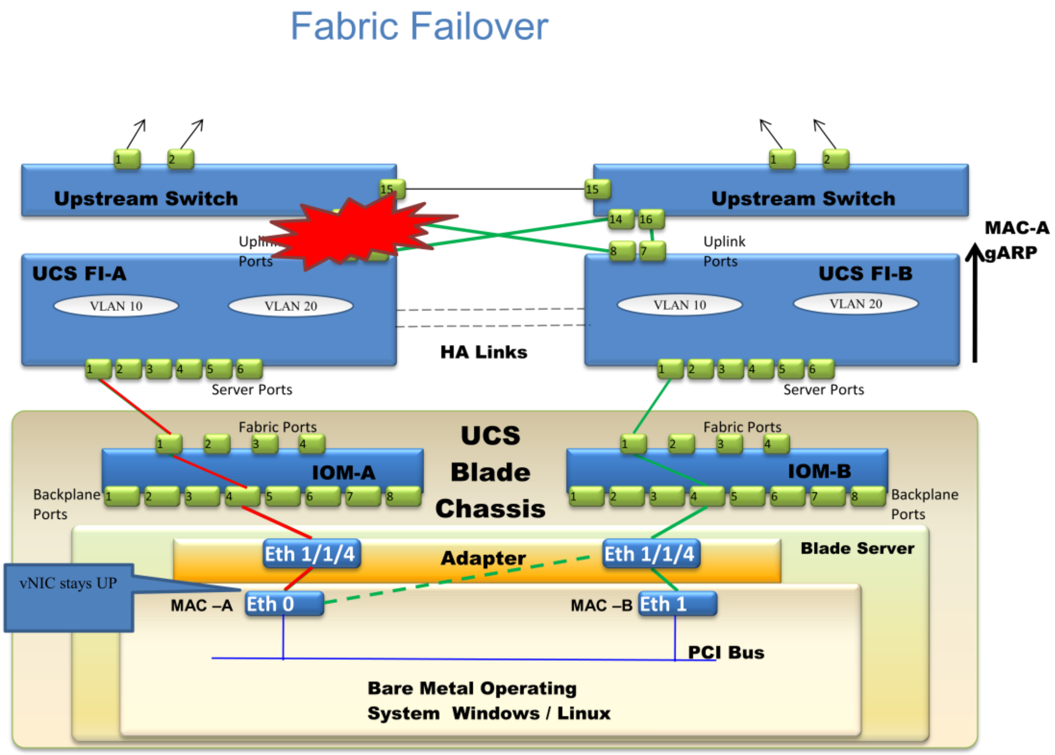
Figure 2. Data Flow Within the Unified Computing System



The two port adapters connect to Fabric A and B respectively and work in active/active fashion. Apart from the two Data Center Bridging (DCB) ports (Eth1/1/4), there are 2 10 Gbps Ethernet ports exposed to the OS. The OS sees the interfaces as eth0/eth1 and Local Area Connection 0/1 in Linux and Windows respectively and not the DCB ports.

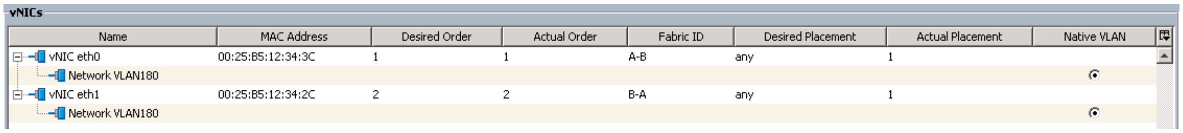
As shown in Figure 3, when Fabric A fails, for example, (e.g., failure of link between the IOM-A and the FI or all uplink failures on the FI), the adapter sends a path enable message to the IOM-B (FI-B). As part of the failover, the FI-B sends gratuitous Address Resolution Protocol (ARP) messages upstream for MAC-A so that the external network knows that the new path is through FI-B.

Figure 3. Fabric Failover Feature Traffic Flow



The fabric properties for the two vNICs exposed to the OS shows them as A-B and B-A, Figure 4.

Figure 4. Properties for Fabric Failover vNICs

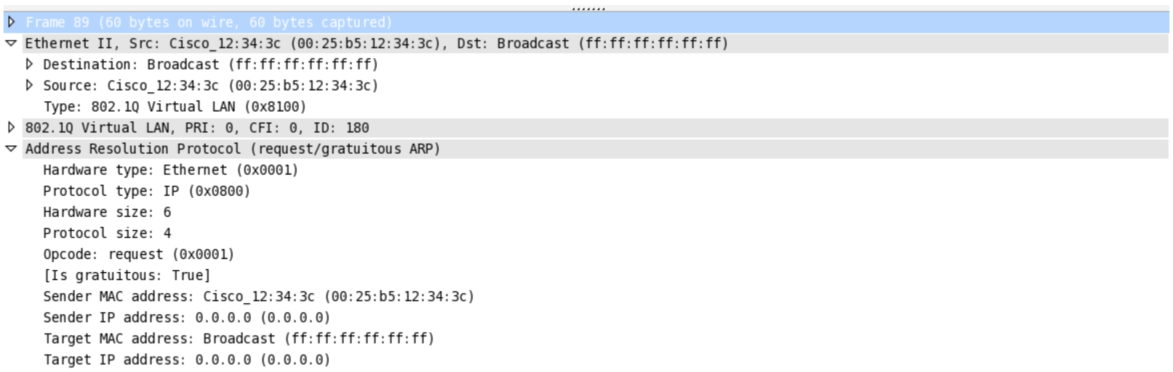


vNICs

Name	MAC Address	Desired Order	Actual Order	Fabric ID	Desired Placement	Actual Placement	Native VLAN
vNIC eth0	00:25:B5:12:34:3C	1	1	A-B	any	1	
Network VLAN180							◉
vNIC eth1	00:25:B5:12:34:2C	2	2	B-A	any	1	
Network VLAN180							◉

The sniffer trace of the gARP packet as seen on the upstream network (originated by FI-B) is shown in Figure 5.

Figure 5. Sniffer Trace for gARP as seen on Upstream Switch Upon Failure of One Fabric



```
▷ Frame 89 (60 bytes on wire, 60 bytes captured)
▽ Ethernet II, Src: Cisco_12:34:3c (00:25:b5:12:34:3c), Dst: Broadcast (ff:ff:ff:ff:ff:ff)
  ▷ Destination: Broadcast (ff:ff:ff:ff:ff:ff)
  ▷ Source: Cisco_12:34:3c (00:25:b5:12:34:3c)
    Type: 802.1Q Virtual LAN (0x8100)
▷ 802.1Q Virtual LAN, PRI: 0, CFI: 0, ID: 180
▽ Address Resolution Protocol (request/gratuitous ARP)
    Hardware type: Ethernet (0x0001)
    Protocol type: IP (0x0800)
    Hardware size: 6
    Protocol size: 4
    Opcode: request (0x0001)
    [Is gratuitous: True]
    Sender MAC address: Cisco_12:34:3c (00:25:b5:12:34:3c)
    Sender IP address: 0.0.0.0 (0.0.0.0)
    Target MAC address: Broadcast (ff:ff:ff:ff:ff:ff)
    Target IP address: 0.0.0.0 (0.0.0.0)
```

It is important to note that the OS does not see Eth0 as down and traffic is transparently redirected via Fabric B. When Fabric A comes back online, the traffic from Eth0 reverts back to its original path.

If the gratuitous ARPs are not sent, the old path will remain upstream untll the time the server sends out packets so that upstream switches learn of the new MAC through the new path (i.e. FI-B).
This is known as the “silent server” issue (i.e. failover does not occur unless the server is chatty and will be referenced later in this paper).

It is recommended that fabric failover be enabled for vNICs in an environment where OSs such as Linux and Windows are run on bare metal (i.e., non VM environments). Fabric failover alleviates the need for teaming software and its configuration. As fabric failover works at a layer below the OS, it stays consistent across bare metal OS platforms.

VMware ESX vSwitch Hypervisor, Distributed Virtual Switch, and Cisco Nexus 1000v Switch on UCS

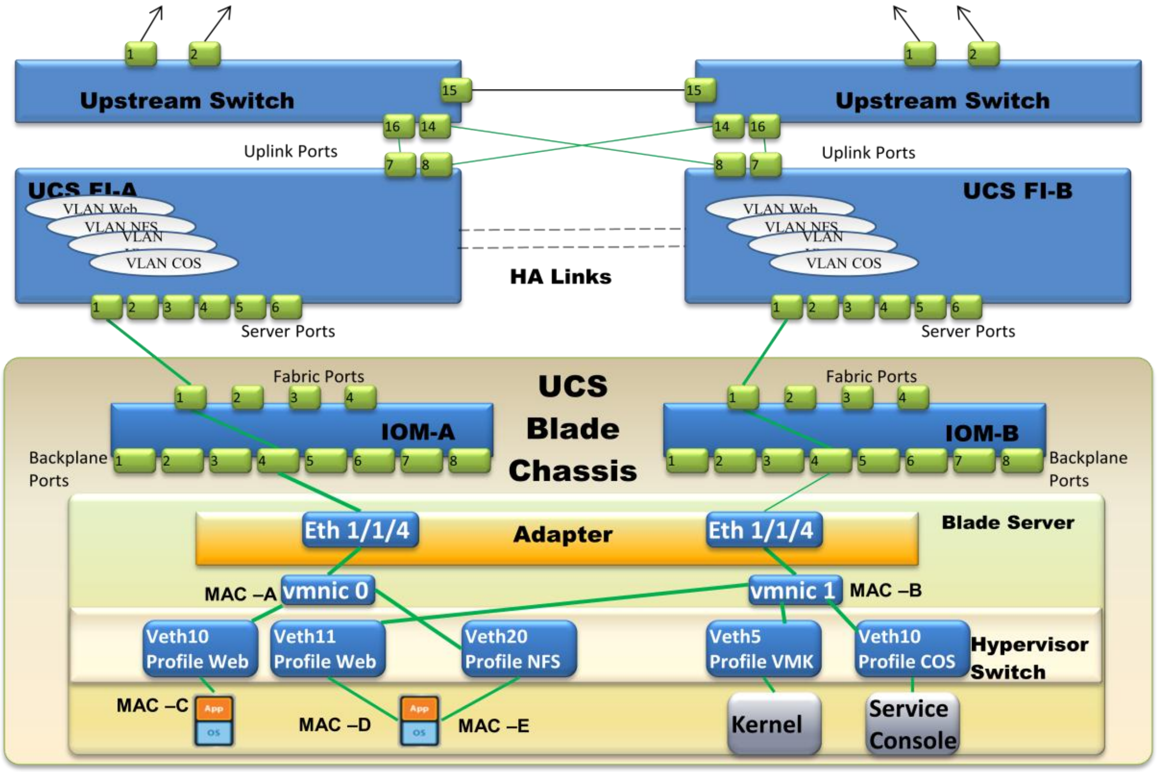
With the ESX or any hypervisor-based OS, there is a virtual switch (vSwitch/DVS/Nexus 1000v) that is connected to the peripheral component interconnect (PCI) exposed adapters. The VMs connect to the virtual switch instead of the adapters, which changes the way failover should work.

As shown in Figure 5, the VMs connect to the vSwitch/DVS/Nexus 1000v and the PCI adapters exposed are bound to the virtual switch. Note that the two adapters can be bound to the same virtual switch in case of a ESX. LACP/802.3ad is not required for active/active use of links in ESX with a virtual switch. The virtual switch does load sharing using route based on the originating virtual port ID or vPC-HM/mac-pinning to utilize both uplinks for active/active use of links. Each VM is pinned to one adapter port and in the event of a link/ adapter port failure traffic fails over to the other port.

Behavior in UCS Manager Version 1.3 and Below

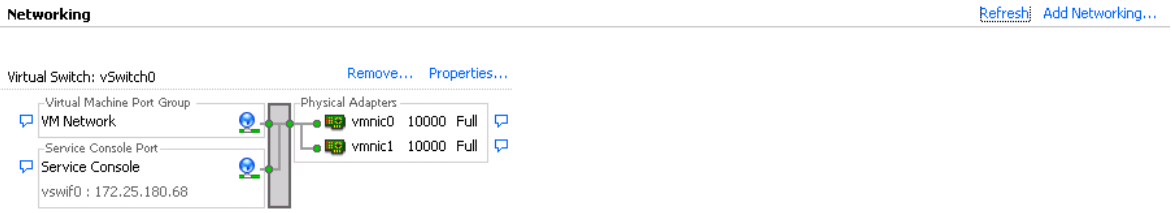
Figure 6 shows the data flow in a hypervisor switch environment.

Figure 6. Data flow when Hypervisor Switch is present



The two vNICs are assigned to the virtual switch (the vSwitch in this case) and not to the VMs themselves, as shown in Figure 7.

Figure 7. vmnic assignments to the vSwitch



Networking

Refresh Add Networking...

Virtual Switch: vSwitch0

Remove... Properties...

Virtual Machine Port Group

VM Network

Service Console Port

Service Console

vswif0 : 172.25.180.68

Physical Adapters

vmnic0 10000 Full

vmnic1 10000 Full

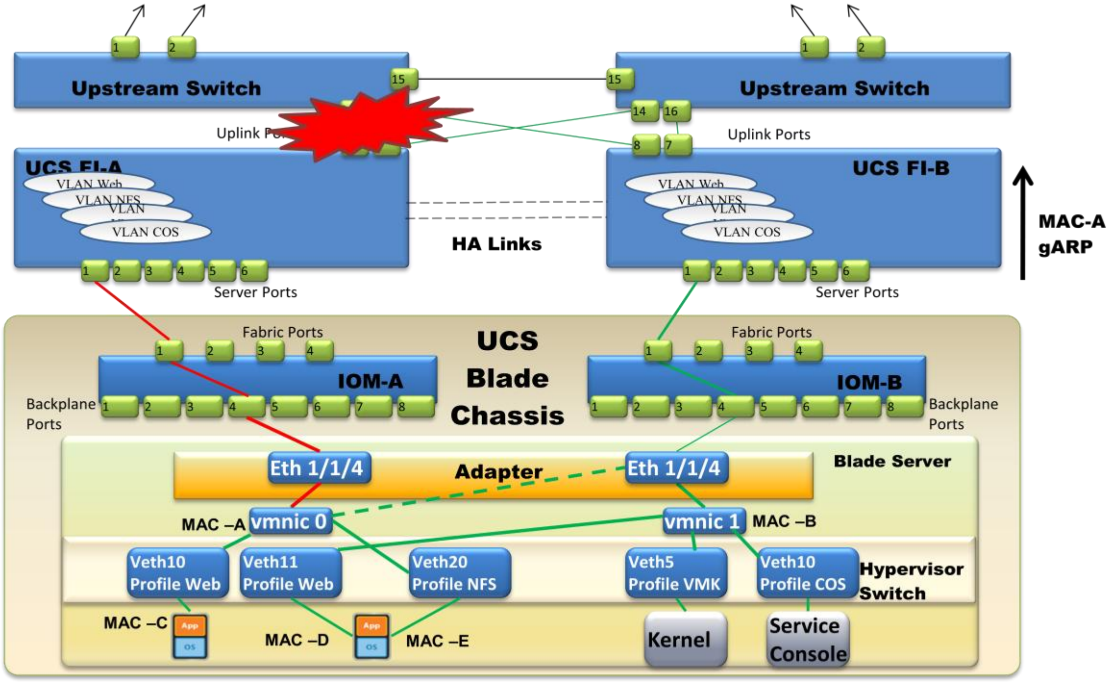
It is important to note that MAC-A and B, which are the MAC addresses on the vNICs, are not used for communication. Each VM has its own MAC and that is what is used by the VMs for traffic.

As shown in Figure 8, if fabric failover is configured for the vNICs A and B, in the event of Fabric-A failure, for example, the vSwitch will not see anything change as the fabric failover feature prevents that from happening and re-routes traffic via the other fabric.

In UCS Manager version 1.3 and below, as part of the reroute the FI-B will send out gratuitous ARPs (MAC-A) as it is unaware of the VM MACs which are learned dynamically on the other FI (FI-A in this case). In this example, upstream learning of the new route to MAC-A does not achieve anything as the failover is required for the VMs (MAC-C).

The upstream network will not be able to communicate with MAC-C until the VMs transmit so that hardware MAC learning can happen on FI-B and the upstream switches, which result in the "silent server" issue discussed before (Figure 8). A chatty VM (where traffic originates from the VM) is required for failover to occur.

Figure 8. Fabric Failover Feature with vSwitch



To correct the above behavior (where failover occurs without the need of a "chatty" server) the recommended behavior is to let the vSwitch see the vNIC fail and take actions on the basis of that. This is possible by disabling fabric failover. When the ESX vSwitch sees the vNIC fail, it moves all vNICs to the other link and transmits gARPs on behalf of the VMs, as shown in Figure 9.

Figure 9. gARP for MAC C Sent by the vSwitch (Correct Behavior)

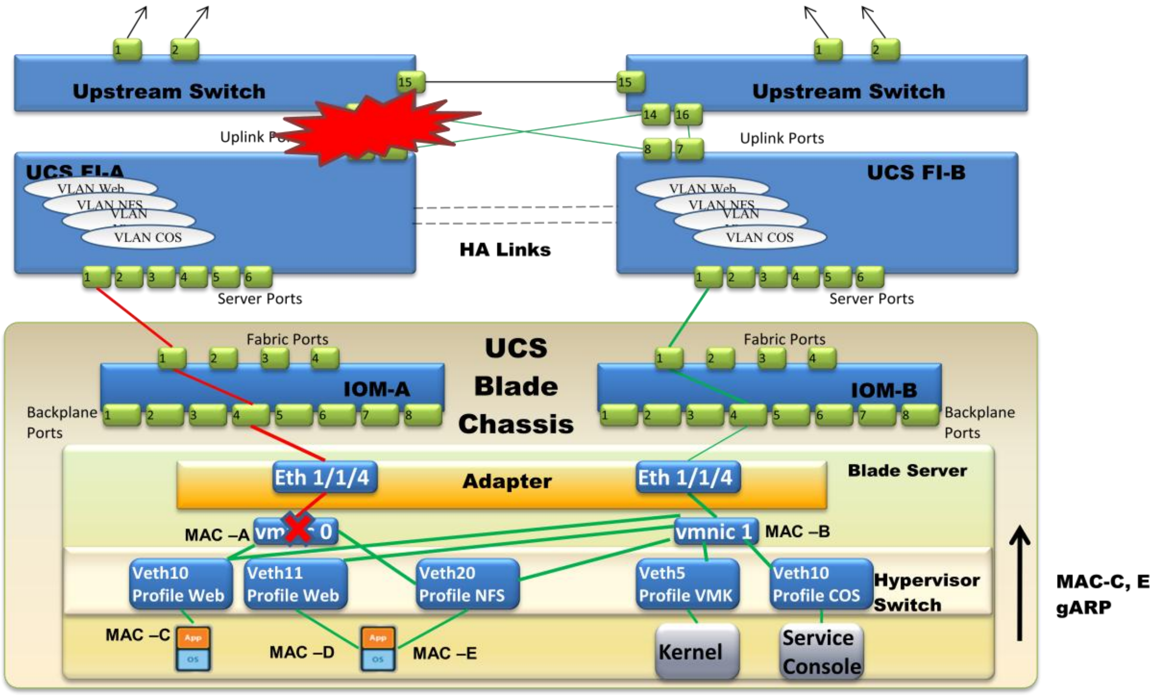
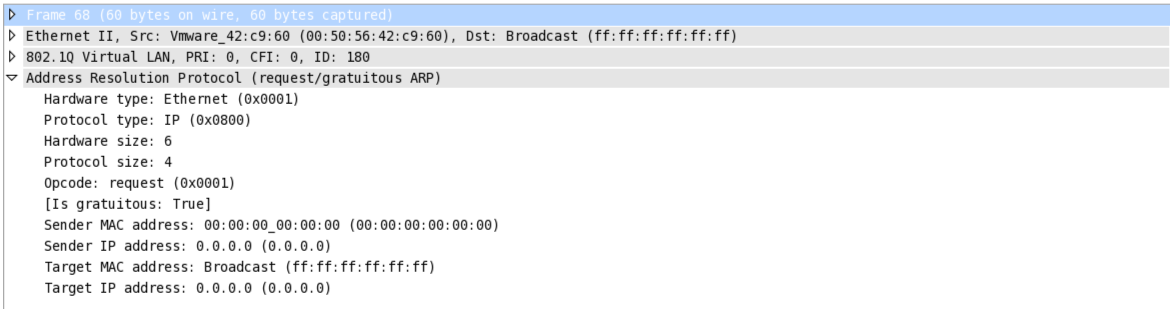


Figure 10 shows the gARP as sent by the vSwitch forwarded to the upstream network.

Figure 10. Sniffer Trace Showing gARP sent by the vSwitch when One Link Fails



```
▷ Frame 68 (60 bytes on wire, 60 bytes captured)
▷ Ethernet II, Src: Vmware_42:c9:60 (00:50:56:42:c9:60), Dst: Broadcast (ff:ff:ff:ff:ff:ff)
▷ 802.1Q Virtual LAN, PRI: 0, CFI: 0, ID: 180
▽ Address Resolution Protocol (request/gratuitous ARP)
    Hardware type: Ethernet (0x0001)
    Protocol type: IP (0x0800)
    Hardware size: 6
    Protocol size: 4
    Opcode: request (0x0001)
    [Is gratuitous: True]
    Sender MAC address: 00:00:00_00:00:00 (00:00:00:00:00:00)
    Sender IP address: 0.0.0.0 (0.0.0.0)
    Target MAC address: Broadcast (ff:ff:ff:ff:ff:ff)
    Target IP address: 0.0.0.0 (0.0.0.0)
```

For ESX Server running vSwitch/DVS/Nexus 1000v and using Cisco UCS Manager Version 1.3 and below, it is recommended that fabric failover not be enabled, as that will require a chatty server for predictable failover. Instead, create regular vNICs and let the soft switch send gARPs for VMs. vNICs should be assigned in pairs (Fabric A and B) so that both fabrics are utilized.

Behavior in Cisco UCS Manager Version 1.4 and Beyond Using the FabricSync Feature

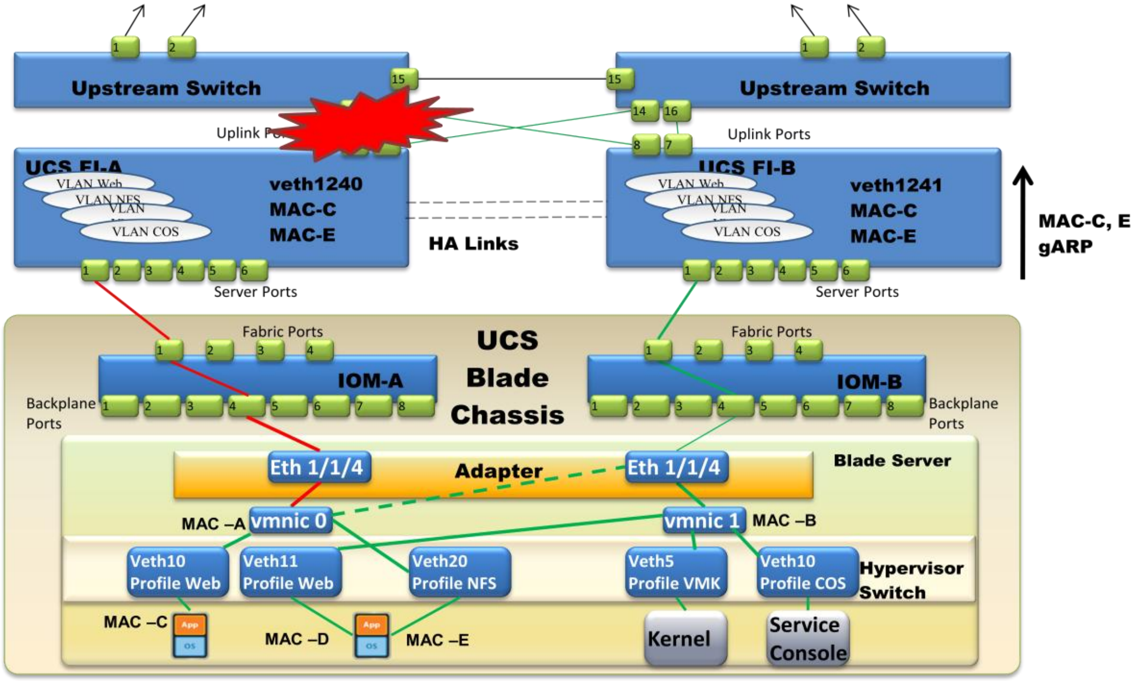
Cisco UCS Manager version 1.4 has introduced a feature called FabricSync that enables the dynamic MACs belonging to VMs to be synchronized or replicated to the peer FI for a fabric failover-enabled vNIC. This MAC replication is done using L1 and L2 peer links and the MACs are maintained in a “replmac” table.

The contents of the table can be viewed by the following command in the NX-OS plane:
TME-TestBed-A(nxos)# show platform fwm info replmac

There is no setting to enable this feature. Dynamic MACs from fabric failover-enabled vNICs are replicated to the other fabric automatically on defining a fabric failover vNIC. It is important to note that only dynamic MACs are replicated using the L1 and L2 links. Any issue with L1 and L2 links and the Fabric Sync feature will not work as the replication table will not be updated. Static vNIC failover will still work as that is done via the virtual interface card (VIC) protocol between the adapter and the FIs.

Due to the Fabric Sync feature in version 1.4, the dynamic MACs replicated to Fabric B and gARPs are now are sent out if the primary path fails by the FI as shown in Figure 11.

Figure 11. gARPs Sent by the Fabric Interconnect for VM MACs on Fabric A Failure



For ESX Server running vSwitch/DVS/Nexus 1000v on the Cisco UCS Manager Version 1.4 and beyond, it is recommended that, even with the Fabric Sync feature, the vNICs in pairs (two vNICs, one on each fabric) should be used for both fabrics.

The current recommendation is not to enable Fabric Failover for the vNICs given to the soft switch but instead to let the softswitch see vNIC go down and issue gARPs following this.

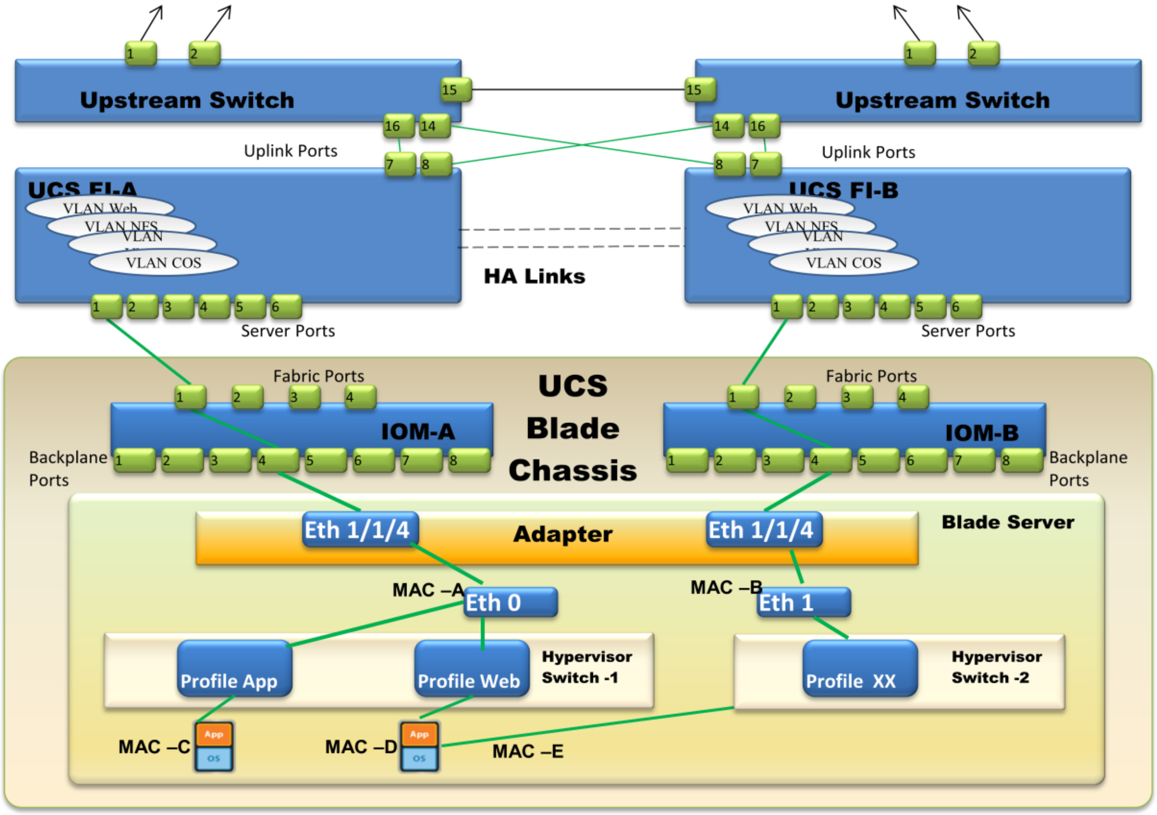
The above recommendation is based on the following:

- VMware DVS and Nexus 1000v now support queuing at the soft switch level for various traffic types. For effective queuing, it is important for the soft switch to see the link down. In case of fabric failover vNICs, the link always stays up.
- When a fabric failover vNIC fails over to the other fabric, for every failed vNIC, a “path enable” message is sent to the FI by the adapter and then gARPs are sent out by the FI. If dynamic MACs are also learned on that vNIC, with the Fabric Sync feature the FI will send out gARPs for them too. This is a software process and is taxing on the FI CPU. For hundreds or thousands of VMs this will be a slow process. If the vNIC is not fabric failover enabled, the gARP for VMs becomes a responsibility shared between the ESX hosts where each host will send out gARPs for VMs on it only and hence becomes distributed in nature, which the FI will pass in hardware. So depending on the scale, non fabric failover vNICs will always result in quicker failover than using Fabric Sync.

The Hyper-V Virtual Network Switch on UCS

Unlike ESX, Hyper-V does not support multiple adapters on the same Virtual Network Switch. Each Virtual Network Switch can only be bound to a single adapter. Figure 12 shows the connectivity when two Virtual Network Switches are configurered in Hyper-V.

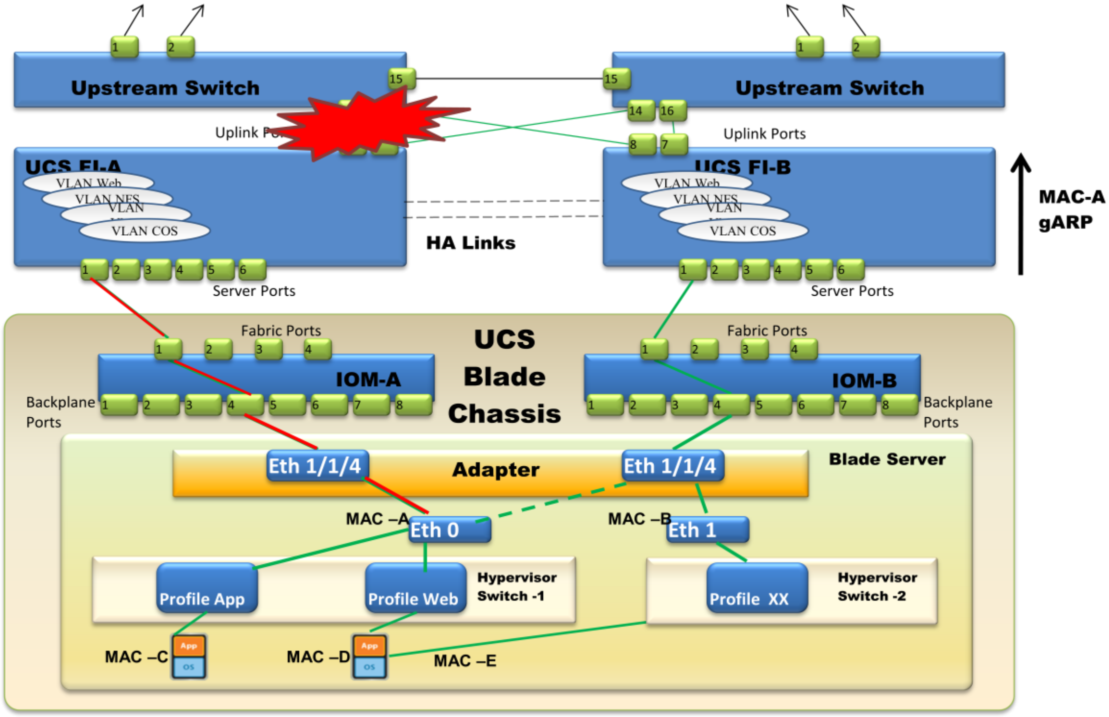
Figure 12. Each Hyper-V soft switch binds to one adapter



Behavior in UCS Version 1.3 and Below

If fabric failover is configured for the NICs (eth0 and eth1), as seen in Figure 13 on Fabric-A failure for example, gARPs are sent out by the FI-B for MAC-A in UCS version 1.3 and below. No gARPs are sent for MAC C/D which makes network connectivity to the VMs unpredictable until the time those VMs send out traffic (another example of the “silent server” issue).

Figure 13. Fabric Failover Feature with Multiple Hyper-V vSwitches

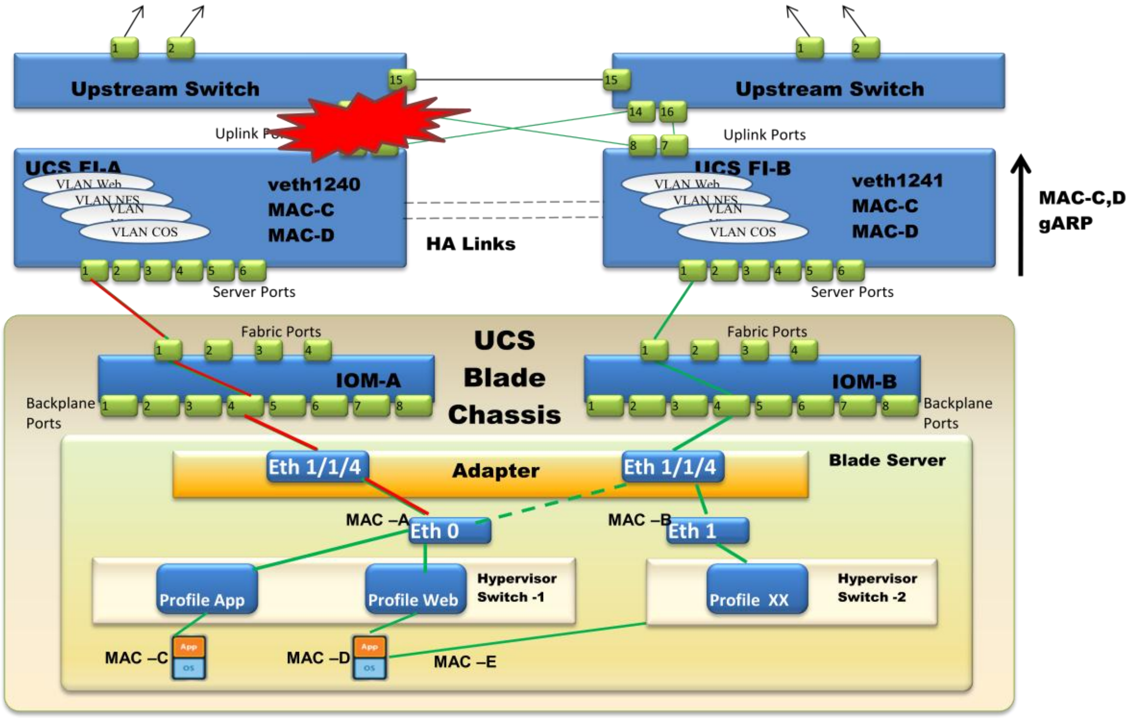


Upstream Switch
Upstream Switch
1
2
15
15
14
16
8
7
Uplink Po
Uplink Ports
UCS FI-A
UCS FI-B
VLAN Web
VLAN NFS
VLAN
VLAN COS
HA Links
MAC-A gARP
Server Ports
Server Ports
1 2 3 4 5 6
Fabric Ports
Fabric Ports
UCS Blade Chassis
IOM-A
IOM-B
Backplane Ports
Backplane Ports
1 2 3 4 5 6 7 8
Eth 1/1/4
Adapter
Eth 1/1/4
Blade Server
MAC –A
Eth 0
MAC –B
Eth 1
Profile App
Profile Web
Hypervisor Switch -1
Profile XX
Hypervisor Switch -2
MAC –C
MAC –D
MAC –E
App
OS

Behavior in Version 1.4 and beyond using the Fabric Sync feature

A failure scenario with the Hyper-V virtual network switch with FabricSync is shown in Figure 15.

Figure 15. FI-B sends out gARPs with the Fabric Sync Feature



With UCS version 1.4 and the Fabric Sync feature, gARPs for the VMs now are sent as part of the failover as seen in Figure 15. Hence with this feature redundancy can now be achieved in a Hyper-V environment under UCS in End Host Mode operations.

It is highly recommended to use Fabric Failover vNICs for the Hyper-V Virtual Network Switch for predictable failover.

Summary

Unique to Cisco UCS, fabric failover enables system failover between virtual network interface cards (vNICs) in the fabric instead of implementing it on the host as it is typically done with a teaming driver. It is available in End Host Ethernet switching mode for Ethernet traffic only.

For OS running on bare metal (non hypervisor environments), Fabric Failover is highly recommended as it provides seamless failover capability without any teaming driver requirement on the host.

Two vNICs (one on each fabric) given to the soft switch in ESX provide a load sharing capability. In UCS version 1.3 and below, the recommendation is to have the vNICs with fabric failover disabled due to the gARPs not being sent by the standby FI on failure for VM MACs.

Cisco UCS version 1.4 has introduced the Fabric Sync feature, which enhances the fabric failover functionality for hypervisors as gARPs for VMs are sent out by the standby FI on failover. It does not necessarily reduce the number of vNICs as load sharing among the fabric is highly recommended. Also recommended is to keep the vNICs with fabric failover disabled, avoiding the use of the Fabric Sync feature in 1.4 for ESX based soft switches for quicker failover.

For a Hyper-V environment, the Fabric Sync feature is the only way to provide predictable redundancy in End Host mode operations and is highly recommended.

For More Information

Cisco Unified Computing White Papers

http://www.cisco.com/en/US/netsol/ns944/networking_solutions_white_papers_list.html