hp networking

june 2003

invent

technical white paper

hp ProLiant network adapter teaming

table of contents

introduction

This document addresses the Teaming technology behind Network Fault Tolerance (NFT), Transmit Load Balancing (TLB), and Switch-Assisted Load Balancing (SLB), including failure recovery methods, load balancing logic, and network scenario considerations.

executive summary

This document provides detailed information about the design, implementation, and configuration of HP's ProLiant Network Adapter Teaming, which includes network fault tolerance and load balancing technologies. Although this document specifically discusses the teaming of HP network adapters under Microsoft Windows 2000 and Windows Server 2003, many of the concepts of HP Network Adapter Teaming are applicable to other operating systems.

The design goal of HP's Network Adapter Teaming is to provide fault tolerance and load balancing across a Team of two or more network adapters. The term "Team" refers to the concept of multiple network adapters working together as a single network adapter, commonly referred to as a Virtual Network Adapter.

The purpose of this document is to assist networking specialists, systems engineers, and IM professionals in the design and troubleshooting of environments incorporating this technology in HP servers. This white paper assumes that the reader is familiar with the basics of IP, the OSI model, the use of network drivers, and the fundamentals of network switching. Additionally, the reader should be familiar with the terms found in the glossary of this white paper.

NOTE: Information in this document was derived from network driver version NCDE 7.30. Because this white paper is about technology, most of the information is generally applicable to future releases; however, specific features may differ slightly between revision levels.

overview of network addressing

Understanding the concepts of network addressing is the key to understanding how HP's Network Adapter Teaming works. This section provides a brief overview of network addressing as a baseline for explaining how HP's Network Adapter Teaming can create one logical network adapter from a Team of two or more adapters.

layer 2 vs. layer 3 addressing

Devices on a computer network use unique addresses, much like telephone numbers, to communicate with each other. Each device, depending on its function, will use one or more of these unique addresses. The addresses correspond to one or more layers of the OSI model. Most often, network devices use an address at Layer 2 (Data Link Layer) called a MAC address, and an address at Layer 3 (Network Layer) called a protocol address (e.g., IP, IPX, AppleTalk). One could say that a MAC address is one that is assigned to the hardware, whereas a protocol address is one that is assigned to the software.

MAC addresses are in the format of 00-00-00-00-00-00 (hexadecimal), IP addresses in the format of 0.0.0.0 (dotted decimal), and IPX addresses in the format of 000000.000000000000 (hexadecimal). Because multiple protocols can reside on the same network device, it is not uncommon for a single network device to use one MAC address and one or more protocol addresses.

Ethernet devices communicate directly using the MAC address, not the protocol address. For instance, when a PING is initiated for the address 1.1.1.1, the network device must find a corresponding MAC address for the IP address of 1.1.1.1. A frame is then built using the MAC address that corresponds to 1.1.1.1 and sent to the destination computer. The frame carries the sender's protocol address in its payload, which is how the destination network device knows to which device to respond. This means that

protocol addresses must be resolved to MAC addresses. For IP, this is done using ARP (refer to "Scenarios of Network Addressing and Communication"). For IPX, the MAC address is part of the IPX address, so no special mechanism is needed.

addresses: unicast vs. broadcast vs. multicast

There are three types of Layer 2 and Layer 3 addresses: unicast, broadcast, and multicast. A unicast address is one that corresponds to a single network device, either a single MAC address or a single IP address. A broadcast address is one that corresponds to all network devices. A multicast address is one that corresponds to many network devices, but not necessarily all network devices. When a station transmits a frame to a unicast address, the transmitting device intends for only a single network device to receive the frame. When a station transmits a frame to a broadcast MAC address or IP address, the station intends for all devices on a particular network to receive the frame. When a station transmits a frame to a multicast MAC or IP address, the station intends for a predefined group of network devices to receive the frame. A group, as used here, can be defined as more than one network device, but less than all the network devices on a particular network.

A multicast MAC address is used in HP Network Adapter Teaming for the purpose of transmitting and receiving heartbeat frames (refer to "Heartbeats"). Because the heartbeat frames are Layer 2 only frames (only use MAC addresses), HP Network Adapter Teams do not need a protocol address assigned to them (e.g., IP address) for heartbeat frames to function.

hp network adapter teaming and layer 2/layer 3 addresses

One of the most important concepts to understand when implementing HP Network Adapter Teaming is that of Layer 2 and Layer 3 addresses, and the way they are handled. When network adapters are teamed together, they function as a single virtual network adapter. Other network devices communicating with an HP Network Adapter Team cannot distinguish that they are communicating with more than one network adapter. HP Network Adapter Teaming must maintain strict IEEE standards compliance in its use of Layer 2 and Layer 3 addresses.

In order for an HP Network Adapter Team to appear as a single virtual network adapter, it is necessary for all networking devices to refer to the Team by a single Layer 2 address and a single Layer 3 address. In other words, when a device is communicating with a Team, regardless of the number of network adapters that make up the Team, the network device only "sees" one MAC address and one protocol address (e.g., IP, IPX). When communicating using IP, this means that a networking device will have only one entry in its ARP cache for an HP Network Adapter Team regardless of the number of network adapters that make up the Team.

When an HP Network Adapter Team initializes, the Teaming driver for each Team "reads" the BIA for each network adapter assigned to that particular Team. Essentially, the MAC addresses are decoupled from the network adapters and pooled together for use by the Teaming driver. The Teaming driver picks one MAC address as the Team's MAC address and assigns it to the Primary Adapter, unless the user has manually set the MAC address (Locally Administered Address) via the configuration GUI (HP Network Teaming and Configuration GUI). In addition, all ARP Replies from the server for this particular HP Network Adapter Team provide this same MAC address as the Team's MAC address. This address does not change unless the Team is reconfigured. The Teaming driver assigns the remaining MAC addresses to the Non-Primary Adapters.

When a failover event occurs, the MAC addresses of the current Primary Adapter and one of the Non-Primary Adapters are swapped. The former Non-Primary Adapter becomes the new Primary Adapter and the former Primary Adapter becomes a Non-Primary Adapter. By swapping the MAC addresses in this manner, the HP Network Adapter Team is always known by one MAC address and one protocol address. It is unnecessary for protocol addresses to swap during a failover event, because the protocol address is directly assigned to the Intermediate (Teaming) driver, and not to the Miniport driver.

When transmitting frames, the current Primary Adapter always transmits using the Team's MAC address as the Layer 2 address and the Team's Protocol address as the Layer 3 address. Non-Primary Adapters always transmit using the MAC address assigned to them by the Teaming driver and using the Team's protocol address as the Layer 3 address. For NFT and TLB, the MAC address used when transmitting is always different from the Primary Adapter's MAC address and is always unique from that of any other Non-Primary Adapter, to comply with IEEE standards. For SLB, the additional switch intelligence allows all teamed adapters to transmit using the same MAC address, the Team's MAC address.

A network device communicating with an HP Network Adapter Team may receive frames from more than one network adapter in the same Team. When this happens, the network device does not know that more than one MAC address is being used. The important issue is that all frames originating from the same HP Network Adapter Team use the same protocol address. The network device does not know that multiple MAC addresses are coming from the Team because MAC headers are stripped off before the frames are processed up the stack by the operating system of the network device. When the operating system receives the frames, they all appear as though they came from the same network adapter. In addition, ARP cache entries are not made by learning the MAC addresses from received frames. ARP cache entries are ONLY made from ARP Requests and ARP Replies or from static entries by hand. Because the Team always sends ARP Replies using the same MAC address, the Team is only known by one MAC address to all network entities.

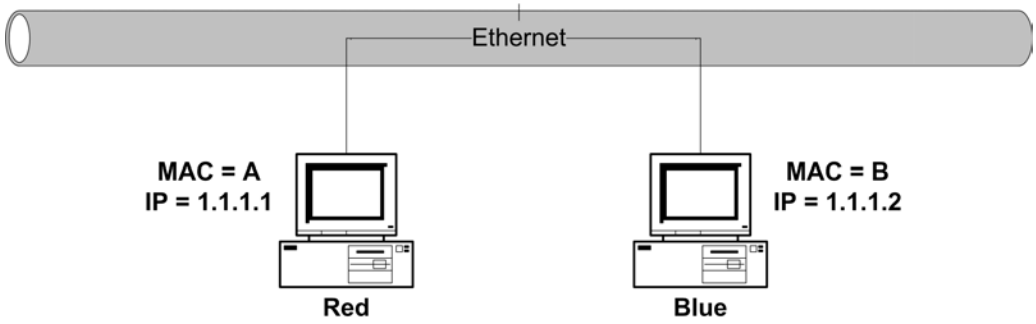
scenarios of network addressing and communication

As discussed earlier, protocol addresses (e.g., IP, IPX) must be resolved to hardware addresses (MAC) for network devices to communicate. What follows are two simple scenarios with one network device (named Red) PINGing another network device (named Blue). The first scenario cites one device PINGing another on the same Layer 2 network. The second scenario cites one device PINGing another on a different Layer 2 network, which requires the use of a router to effect communication.

These scenarios provide a baseline of typical network addressing and communication using IP. This baseline will be referred to later in this document to differentiate how HP Network Adapter Teaming functions in these same scenarios. By understanding the differences in simple examples such as these (without HP's Network Adapter Teaming technology involved), implementers will have a better understanding of how HP's Network Adapter Teaming technology may work in their environment.

scenario 1-A: one device PINGs another on the same layer 2 network

figure 1. One device PINGs Another on the Same Layer 2 Network



1. Red transmits a broadcast ARP Request asking for Blue's MAC address.

   A user on Red issues the command "ping 1.1.1.2" to initiate a PING to Blue. The number 1.1.1.2 refers to Blue's IP address, or protocol address. First, Red determines whether or not Blue is on the same Layer 2 network by running an algorithm (details of this algorithm are beyond the scope of this document) using its own IP address of 1.1.1.1, its own subnet mask (not shown), and Blue's IP address of 1.1.1.2. If Blue is on a different Layer 2 network, then Red will need to use its gateway, or router, to get to Blue.

   Once Red has determined that Blue is on the same Layer 2 network, Red must find out what Blue's MAC address is. First, Red checks its own ARP cache for a MAC address entry matching the IP address of 1.1.1.2. ARP is used to map protocol addresses to hardware addresses. If Red does not have a static entry or an entry cached from a previous conversation with Blue, then it must broadcast an ARP Request frame containing the IP address of Blue on the network asking Blue to respond and provide its MAC address. Red must broadcast this ARP Request because without knowing Blue's unique MAC address, it has no way of sending a frame directly (unicast) to Blue.

2. Blue transmits a unicast ARP Reply to Red, providing its MAC address (B).

   Blue sees the ARP Request containing its own IP address and responds with a unicast ARP Reply directly to Red. Blue also notes Red's MAC address (A) and IP address of 1.1.1.1, and enters them into its ARP cache. Red receives the ARP Reply and enters Blue's MAC address (B) and IP address (1.1.1.2) into its own ARP cache.

3. Red transmits a unicast PING Request to Blue using Blue's MAC address (B).

   Red can now create a PING Request frame using Blue's MAC address (B). Red sends the PING Request to Blue using Blue's MAC address (B). Blue receives the PING Request frame and notices that a station with an IP address of 1.1.1.1 is requesting that it respond.

4. Blue transmits a broadcast ARP Request asking for Red's MAC address.

   NOTE: This step may not occur if Blue's ARP table still contains an entry for Red as a result of steps 1 and 2.

   Blue checks its ARP cache for a MAC address entry that corresponds to 1.1.1.1. If Blue does not find one (i.e., ARP cache timed out since last communication with Red), then Blue broadcasts an ARP Request asking for Red's MAC address.

5. Red transmits a unicast ARP Reply to Blue providing its MAC address (A).

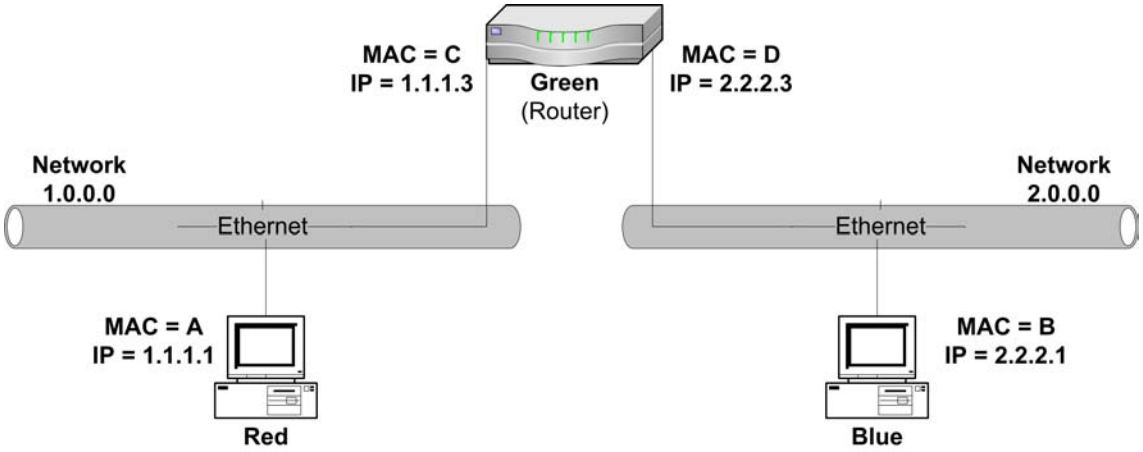
   NOTE: This step will not occur if step 4 does not take place.

   Red sees the ARP Request and transmits a unicast ARP Reply directly to Blue providing its MAC address (A). Blue receives the ARP Reply and puts Red's MAC address (A) and IP address (1.1.1.1) in its ARP cache.
6. Blue transmits a unicast PING Reply to Red using Red's destination MAC address (A).

Blue transmits a unicast PING Reply to Red using Red's MAC address (A) and the user sees the PING REPLY message printed on the screen. This completes the entire conversation.

scenario 2-A: one device PINGs another on a different layer 2 network

figure 2. One device PINGs another on a different Layer 2 network



1. Red transmits a broadcast ARP Request on Network 1.0.0.0 asking for Green's MAC address.

   A user on Red issues the command "ping 2.2.2.1" to initiate a PING to Blue. The number 2.2.2.1 refers to Blue's IP address, or protocol address. First, Red determines whether or not Blue is on the same Layer 2 network by running an algorithm (details of this algorithm are beyond the scope of this document) using its own IP address of 1.1.1.1, its own subnet mask (not shown), and Blue's IP address of 2.2.2.1. If Blue is on a different Layer 2 network, then Red will need to use its gateway or router (Green) to get to Blue.

   Once Red has determined that Blue is on a different Layer 2 network, Red must use Green as a gateway to get to Blue. Red communicates directly with Green at Layer 2 but communicates directly with Blue at Layer 3. This means that Red must transmit a frame with the Layer 2 address (MAC) of Green, but the same frame will have Blue's Layer 3 address (IP) in it. When Green receives the frame, it sees the Layer 3 data destined for Blue and forwards the frame onto Blue via Green's interface that is attached to Blue's Layer 2 network (Network 2.0.0.0). This means that Red must find out what Green's MAC address is. First, Red checks its own ARP cache for an entry that matches 1.1.1.3. If Red does not have an entry cached, then it must broadcast an ARP Request frame on network 1.0.0.0 asking Green to respond and provide its MAC address.

2. Green transmits a unicast ARP Reply to Red providing its MAC address (C).

   Green sees the ARP Request and responds with a unicast ARP Reply to Red. Also, Green enters Red's MAC address and IP address into its ARP cache. Red receives Green's ARP Reply and enters Green's MAC address (C) and IP address (1.1.1.3) into its ARP cache.

3. Red transmits a PING Request to Blue (2.2.2.1) using the destination MAC address (C) of Green's 1.1.1.3 interface, because Green is Red's gateway to Blue.

   Red can now create a PING Request frame using Green's MAC address and Blue's IP address. Red sends the PING Request. Green receives the PING Request and determines that the frame is meant for Blue because of the Layer 3 address (IP).

4. Green transmits a broadcast ARP Request on Network 2.0.0.0 asking for Blue's MAC address.

   Green looks in its ARP cache for a MAC address for Blue. If one is not found, Green broadcasts an ARP Request frame on Blue's Layer 2 network asking for Blue's MAC address.

5. Blue transmits a unicast ARP Reply to Green providing its MAC address (B).

   Blue sees the ARP Request frame and responds with a unicast ARP Reply frame to Green. Also, Blue enters Green's MAC address and IP address into its ARP cache. Green receives the ARP Reply from Blue and enters Blue's MAC address (B) and IP address (2.2.2.1) into its ARP cache.

6. Green forwards Red's PING Request to Blue using Blue's destination MAC address (B).

   Green now transmits Red's original Ping Request frame onto Blue's network using Blue's MAC address and Blue's IP address as the destination MAC and destination IP address. The source MAC address is Green's MAC address (D) and the source IP address is Red's IP address (1.1.1.1). Blue receives the frame and notices that a station with an IP address of 1.1.1.1 is asking for it to respond to a PING. Before Blue can respond with a PING Reply, it must determine whether or not 1.1.1.1 is on the same Layer 2 network. Blue runs an algorithm (details of this algorithm are beyond the scope of this document) using its own IP address (2.2.2.1), its own subnet mask (not shown) and the IP address of Red (1.1.1.1). Blue then determines that Red is on a different network. Because of this, Blue must use its gateway (Green) to get the PING Reply back to Red.

7. Blue transmits a broadcast ARP Request on Network 2.0.0.0 asking for Green's MAC address.

   NOTE: This step may not occur if Blue's ARP table still contains an entry for Green resulting from steps 4 and 5.

   Blue checks its ARP cache for the MAC address that corresponds to the IP address of 2.2.2.3 (Blue's gateway). If an entry is not found, Blue must broadcast an ARP Request asking for Green's MAC address.

8. Green transmits a broadcast ARP Reply to Blue providing its MAC address (D).

   NOTE: This step will not occur if step 7 does not take place.

   Green sees the ARP Request and responds with a unicast ARP Reply directly to Blue. Also, Green enters Blue's MAC address and IP address into its ARP cache. Blue receives the ARP Reply and puts Green's MAC address (D) and IP address (2.2.2.3) in its ARP cache. Blue now has all the information it needs to send a PING Reply to Red.

9. Blue transmits a unicast PING Reply to Red (1.1.1.1) using the MAC address of Green's 2.2.2.3 interface (D).

   Blue transmits a unicast PING Reply to Red through Green by using Green's MAC address as the destination MAC address, Red's IP address as the destination IP address, Blue's MAC address as the source MAC address and Blue's IP address as the source IP address. Green receives the PING Reply and determines that the frame is meant for Red because of the Layer 3 address (IP).
10. Green transmits a broadcast ARP Request on Network 1.0.0.0 asking for Red's MAC address.

    NOTE: This step will not occur if Green's ARP table still contains an entry for Red resulting from steps 1 and 2.

    Green looks in its ARP cache for a MAC address for Red. If one is not found, Green broadcasts an ARP Request frame on network 1.0.0.0 asking for Red's MAC address.
11. Red transmits a unicast ARP Reply to Green providing its MAC address (A).

    NOTE: This step will not occur if step 10 does not take place.

    Red sees the ARP Request frame and responds with a unicast ARP Reply frame to Green. Also, Red enters Green's MAC address and IP address into its ARP cache. Green receives the ARP Reply from Red and enters Red's MAC address (A) and IP address (1.1.1.1) into its ARP cache.
12. Green forwards Blue's PING Reply to Red using the destination MAC address of Red (A).

    Green transmits Blue's Ping Reply frame onto Red's network using Red's MAC address (A) and Red's IP address (1.1.1.1) as the destination MAC and destination IP address. The source MAC address is Green's MAC address (C) and the source IP address is Blue's IP address (2.2.2.1). The user sees the PING REPLY message printed on the screen. This completes the entire conversation.

teaming mechanisms

architecture of hp network adapter teaming

Within an operating system (OS), a hierarchy of layers work together to enable one OS to communicate with another. Each of these layers performs a separate function and passes information between the layers above and below it. Within Windows 2000, there are four layers that are important to understand when discussing HP Network Adapter Teaming: the Miniport layer, Intermediate layer, NDIS layer, and Protocol layer.

- Miniport Layer

  The network adapter driver resides at the Miniport Layer. This driver is responsible for directly controlling the hardware. It is necessary for basic network adapter functionality and is used even when HP's Network Adapter Teaming is not deployed. Typically, this driver is written by the vendor of the network adapter hardware. HP network adapters drivers (e.g., Q57W2K.SYS, N1000NT5.SYS, N100NT5.SYS) are considered Miniport drivers.

- Intermediate Layer

  The Intermediate layer driver provides a network function, but is not considered a Miniport because it does not directly control a piece of hardware. The Intermediate layer driver performs a function that is between the Miniport layer and NDIS. The

networking function that is performed by the Intermediate layer is beyond the ability of a Miniport layer driver. In this case, HP Network Adapter Teaming is considered an Intermediate driver (i.e., CPQTEAM.SYS). It performs the function of making several Miniport drivers seamlessly work as a single network adapter that interfaces with NDIS. Another example of an Intermediate driver is the NLB (Network Load Balancing) feature in Microsoft ®Windows® 2000.

- NDIS

  NDIS, Microsoft's Network Driver Interface Specification, handles communications between the underlying layers, either Miniport drivers or Intermediate drivers, and the Protocol layer.

- Protocol Layer

  The Protocol layer is where IP, IPX, and AppleTalk, etc., interface with NDIS. This layer is responsible for Protocol addresses (e.g., IP or IPX addresses), and also for translating Layer 3 addresses (i.e., IP addresses) to Layer 2 addresses (i.e., MAC addresses).

  In the absence of an Intermediate driver, a protocol address is usually assigned to each individual Miniport driver. However, when utilizing HP's Network Adapter Teaming, the protocol address is assigned to a single HP Network Adapter Teaming instance that represents the underlying Miniports. If more than one HP Network Adapter Team exists in a single server, there will be more than one instance of the HP Network Adapter Team and an individual protocol address will be assigned to each instance.

teaming software components

HP Network Adapter Teaming consists of three components: the Miniport Driver, Intermediate Driver, and configuration GUI.

- Miniport Driver

  For Microsoft Windows 2000, the Miniport driver used with the HP network adapter will be Q57W2K.SYS, N100NT5.SYS, or N1000NT5.SYS depending on the adapter in use.

- Intermediate Driver

  For Microsoft Windows 2000, the Intermediate driver is CPQTEAM.SYS, and is used for all teaming functions involving HP NC series adapters.

- Configuration GUI

  For Microsoft Windows 2000, the configuration GUI is called the HP Network Teaming and Configuration GUI and the file name is CPQTEAM.EXE. The configuration GUI is accessible from Control Panel or from the Tray icon (if enabled).

These three components are designed to work as a single unit. When one is upgraded, it is advisable to upgrade all components to the current version. For driver updates to HP network adapters and HP Network Adapter Teaming, please visit the HP ProLiant Network Adapter Driver site.

HP ProLiant Networking Home:

http://www.hp.com/servers/networking

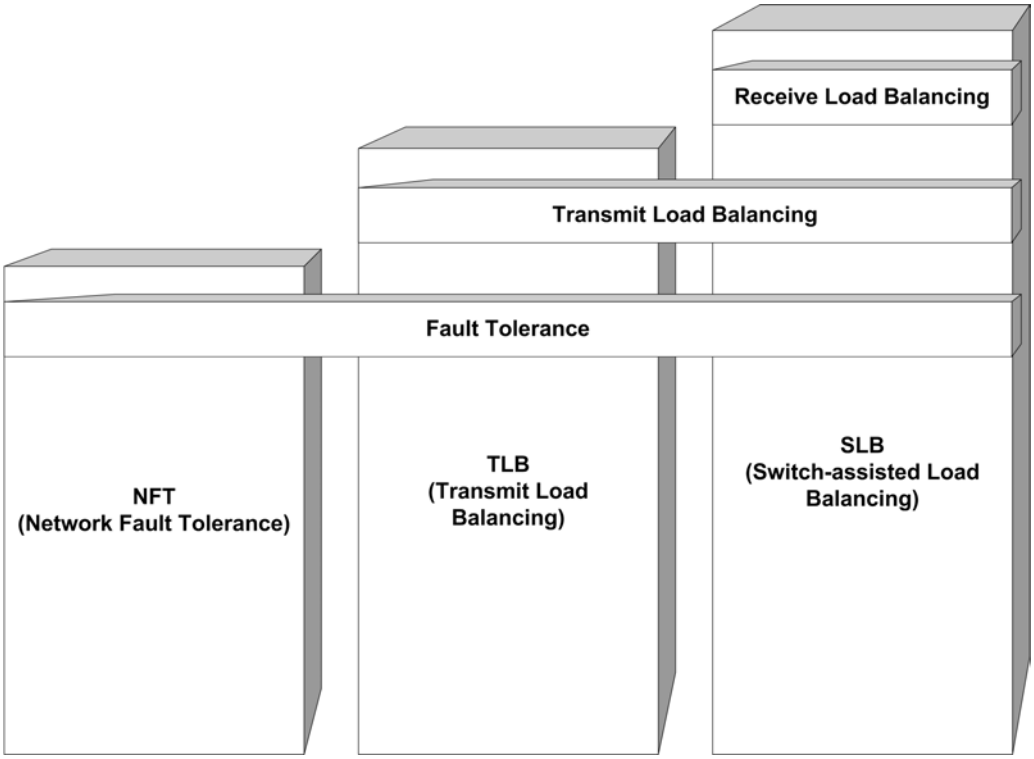
HP ProLiant Network Adapter Drivers:

http://h18000.www1.hp.com/support/files/networking/nics/index.html

types of hp network adapter teams

There are three teaming modes for HP network adapters: Network Fault Tolerance (NFT), Transmit Load Balancing (TLB), and Switch-assisted Load Balancing (SLB). Respectively, each mode gains in features and incorporates most features from the previous teaming mode (refer to figure 3). In other words, NFT is the simplest teaming mode, supporting only network adapter fault tolerance. TLB supports network adapter fault tolerance plus load balancing of IP traffic being transmitted from the server. SLB supports network adapter fault tolerance, load balancing of any traffic being transmitted from the server, plus load balancing of any traffic being received by the server.

figure 3. Teaming types and teaming functionality



Network Fault Tolerance (NFT)

Network Fault Tolerance is the foundation of HP Network Adapter Teaming. In NFT mode, two to eight adapters may be teamed together to operate as a single virtual network adapter. However, only one network adapter, the Primary Adapter, is used for both transmit and receive communication with the server. The remaining adapters are considered stand-by, or secondary, adapters, referred to as Non-Primary adapters, and remain idle unless the Primary adapter fails. All adapters may transmit and receive heartbeats, including Non-Primary adapters.

Network Addressing and Communication using NFT

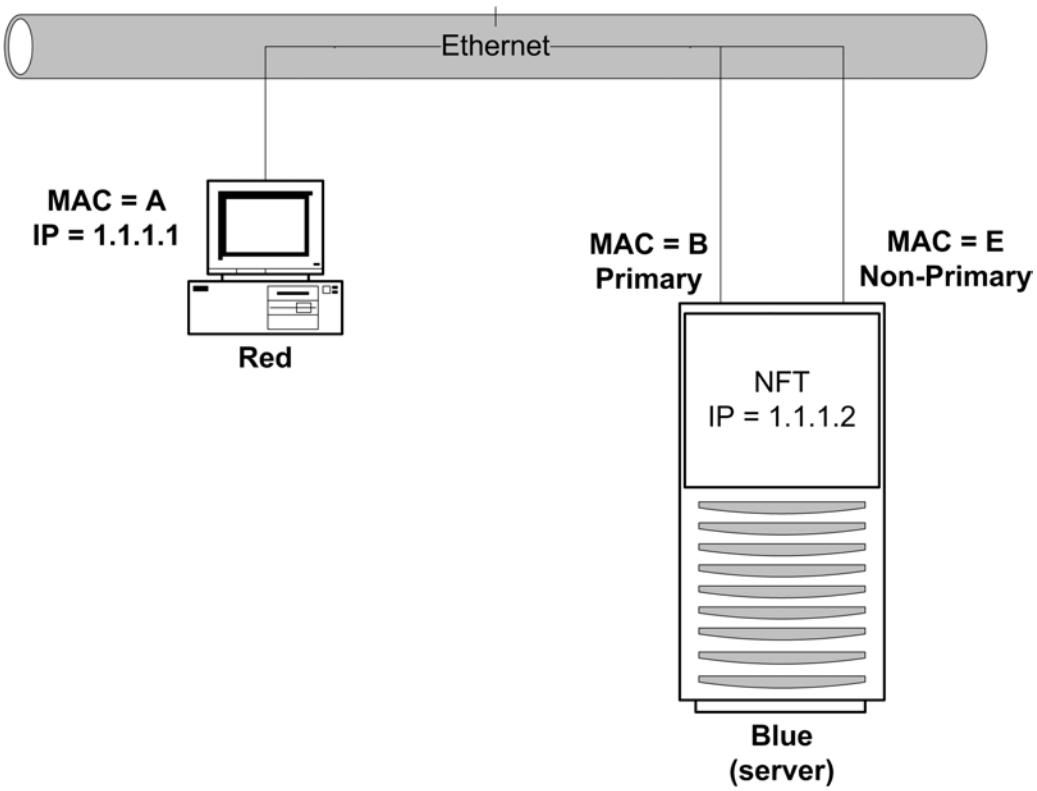
Before learning the specifics of NFT and how it communicates on the network, it is recommended that the section titled "HP Network Adapter Teaming and Layer 2/Layer 3 addresses" be thoroughly reviewed and understood.

Scenario 1-B: a device PINGs an NFT team on the same layer 2 network

This section builds on the concepts reviewed previously in the section titled, "Scenarios of Network Addressing and Communication - Scenario 1-A", and describes how NFT functions from the network addressing and communication perspective.

Utilizing a network diagram similar to figure 1, Blue has been modified to be a server utilizing an HP Network Adapter Team in NFT mode with two network adapters in a Team (refer to figure 4). The two network adapters have MAC addresses of "B" and "E", and are known by a single Layer 3 address of 1.1.1.2. Network adapter B has been designated as the Primary Adapter in this NFT Team.

figure 4. A device PINGs an NFT Team on the same Layer 2 network



1. Red transmits a broadcast ARP Request asking for Blue's MAC address.

   A user on Red issues the command "ping 1.1.1.2" to initiate a PING to Blue. First, Red determines whether or not Blue is on the same Layer 2 network. Once Red has determined that Blue is on the same Layer 2 network, Red must find out what Blue's MAC address is. First, Red checks its own ARP cache for a MAC address entry matching the IP address of 1.1.1.2. If Red does not have a static entry or an entry cached from a previous conversation with Blue, then it must broadcast an ARP Request frame on the network asking Blue to respond and provide its MAC address. Red must broadcast this ARP request because without knowing Blue's unique MAC address, it has no way of sending a frame directly (unicast) to Blue.

2. Blue transmits a unicast ARP Reply to Red, providing its MAC address.

   Blue sees the ARP Request (the frame is received on both the Primary and Non-Primary Adapters in the Team) because the frame is broadcast on the network. However, the Team discards all non-heartbeat frames incoming on Non-Primary Adapters, and responds with a unicast ARP Reply to Red. The ARP Reply is transmitted by the Primary Adapter (B). In Blue's ARP Reply, Blue provides the MAC

address of its Teaming driver, which is the same as the current Primary Adapter's MAC address (B) (refer to "HP Network Adapter Teaming and Layer 2/Layer 3 addresses"). Blue also notes Red's MAC address (A) and IP address (1.1.1.1) and enters them into its ARP cache. Red receives the reply and enters the MAC address (B) and the IP address of Blue (1.1.1.2) into its own ARP cache.

3. Red transmits a unicast PING Request to Blue using Blue's destination MAC address.

   Red can now create a PING Request frame using Blue's MAC address (B). Red sends the PING Request to Blue. Blue receives the frame on its Primary Adapter (B) and notices that a station with an IP address of 1.1.1.1 is asking for it to respond.

4. Blue transmits a broadcast ARP Request asking for Red's MAC address.

   NOTE: The following step may not occur if Blue's ARP table still contains an entry for Red as a result of steps 1 and 2.

   Blue checks its ARP cache for a MAC address entry that matches 1.1.1.1. If Blue does not find one, then Blue broadcasts an ARP Request asking for Red's MAC address.

5. Red transmits a unicast ARP Reply to Blue providing its MAC address.

   NOTE: The following step will not occur if step 4 does not take place.

   Red sees the ARP Request and transmits a unicast ARP Reply directly to Blue providing its MAC address (A). Blue receives the ARP Reply and puts Red's MAC address (A) and IP address (1.1.1.1) in its ARP cache.

6. Blue transmits a unicast PING Reply to Red using Red's destination MAC address.

   Blue then transmits a unicast PING Reply to Red using Red's MAC address (A) and the user sees the PING Reply message printed on the screen. This completes the entire conversation.

NFT applications

NFT is deployed in environments that only require fault tolerance and do not require transmit or receive throughput greater than the capacity of the Primary Adapter (e.g., a server that requires fault tolerance in case of a network adapter malfunction, but does not have a demand for receiving or transmitting more than the capacity of the Primary adapter.)

recommended configurations for an NFT environment

HP recommends that:

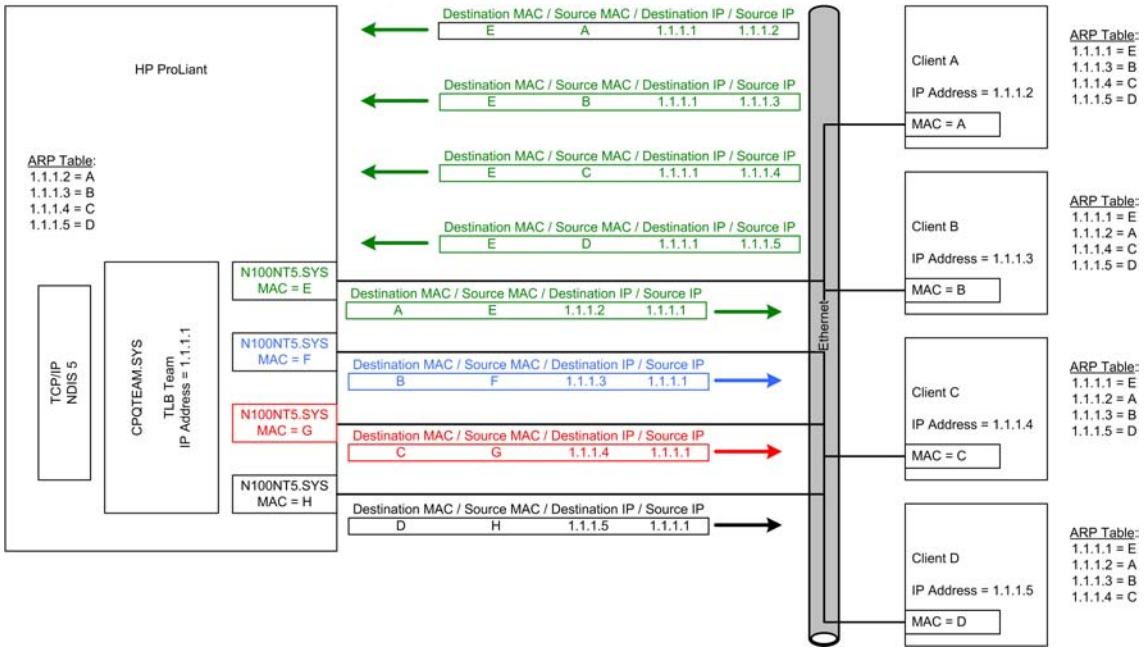
- Heartbeats be enabled (default)
- MAC addresses not be manually set to a locally administered address (LAA) via the Microsoft UI. A user should not implement LAAs on individual network adapters that are members of a Team; otherwise Teaming may not function correctly. Setting an LAA for the Team is permitted via the HP Network Adapter Teaming and Configuration GUI.
- Spanning Tree's blocking, listening, and learning stages be disabled, or bypassed, on all switch ports to which an HP Network Adapter Team port is attached. These stages are not needed when a non-switch networking device (e.g. server) is attached to the switch port. HP ProCurve switches have a feature called STP Fast Mode that is used to disable these Spanning Tree stages on a port-by-port basis. Cisco® switches have an equivalent feature called PortFast.
- Team members can be split across more than one switch in order to achieve switch redundancy. However, all switch ports that are attached to members of the same Team must comprise a single broadcast domain (i.e., same VLAN). Additionally, if

problems exist after deploying a Team across more than one switch, reattach all Team members to the same switch. If the problems disappear, then the cause of the problem resides in the configuration of the switches and not in the configuration of the Team. If switch redundancy is required (Team members are attached to two different switches), then HP recommends that the switches be deployed with redundant links between them and Spanning Tree be enabled (or other Layer 2 redundancy mechanisms) on the ports that connect the switches. This helps prevent switch uplink failure scenarios that leave Team members in separate broadcast domains.

Transmit Load Balancing (TLB)

Transmit Load Balancing mode, previously known as Adaptive Load Balancing (ALB), incorporates all the features of NFT, plus Transmit Load Balancing. In this mode, two to eight adapters may be teamed together to function as a single virtual network adapter. The load-balancing algorithm used in TLB allows the server to load balance traffic transmitted from the server. However, traffic received by the server is not load balanced, meaning the Primary Adapter is responsible for receiving all traffic destined for the server (refer to figure 5). In addition, only IP traffic is load balanced. As with NFT, there are two types of Team members, Primary and Non-Primary Adapters. The Primary Adapter transmits and receives frames and the Non-Primary Adapters only transmit frames.

figure 5. Overview of TLB communication



network addressing and communication using TLB

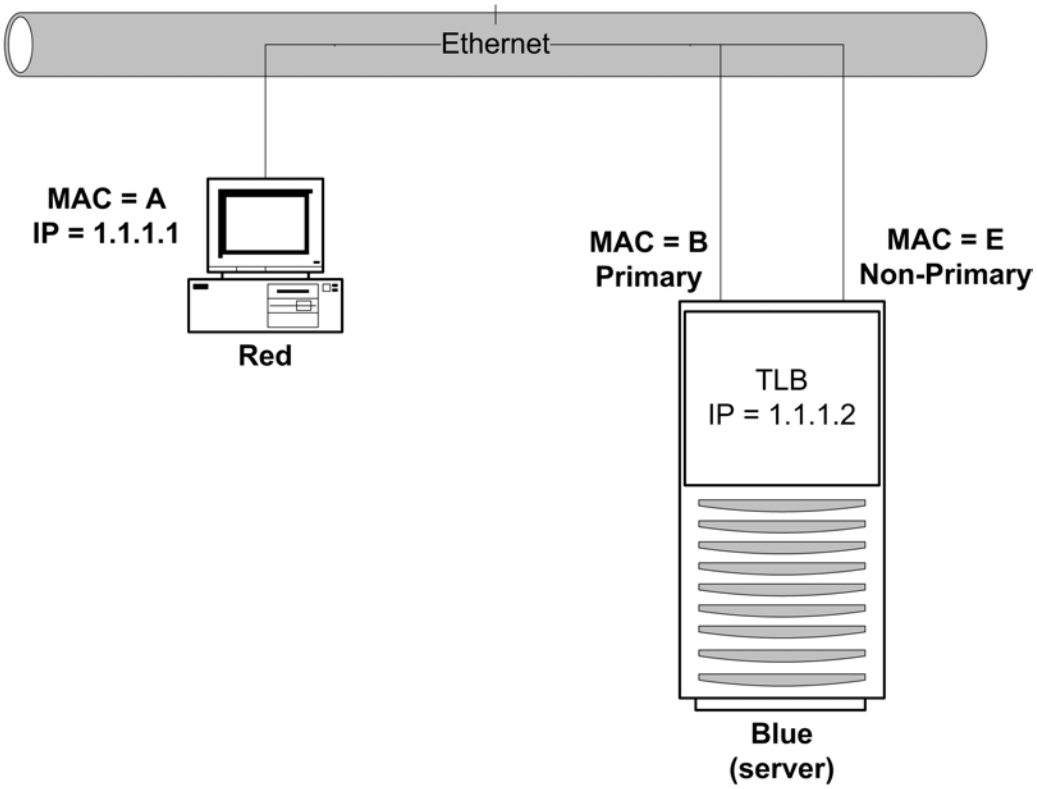
Before learning the specifics of TLB and how it communicates on the network, it is recommended that the section titled "HP Network Adapter Teaming and Layer 2/Layer 3 addresses" be thoroughly reviewed and understood.

scenario 1-C: a device PINGs a TLB team on the same layer 2 network

This section builds on the concepts reviewed previously in the section titled, "Scenarios of Network Addressing and Communication - Scenario 1-A", and describes how TLB functions from the network addressing and communication perspective.

Utilizing a network diagram similar to figure 1., Blue has been modified to be a server utilizing an HP Network Adapter Team in TLB mode with two network adapters in a Team (refer to figure 6). The two network adapters have Layer 2 addresses of MAC B and MAC E, respectively, and are known by a single Layer 3 address of 1.1.1.2. Network adapter B has been designated as the Primary Adapter in this NFT Team.

figure 6. A device PINGs a TLB Team on the same Layer 2 network



1. Red transmits a broadcast ARP Request asking for Blue's MAC address.

   A user on Red issues the command "ping 1.1.1.2" to initiate a PING to Blue. First, Red determines whether or not Blue is on the same Layer 2 network.

   Once Red has determined that Blue is on the same Layer 2 network, Red must find out what Blue's MAC address is. First, Red checks its own ARP cache for a MAC address entry matching the IP address of 1.1.1.2. If Red does not have a static entry or an entry cached from a previous conversation with Blue, then it must broadcast an ARP Request frame on the network asking Blue to respond and provide its MAC address. Red must broadcast this ARP request because without knowing Blue's unique MAC address, it has no way of sending a frame directly (unicast) to Blue.

2. Blue transmits a unicast ARP Reply to Red, providing its MAC address.

   Blue sees the ARP Request (the frame is received on both adapters the Primary and Non-Primary Adapters in the Team because the frame is broadcasted onto the network. However, the Team discards all non-heartbeat frames incoming on Non-Primary Adapters), and responds with a unicast ARP Reply to Red. The ARP Reply is transmitted by the Primary Adapter (B) because all non-IP frames are always transmitted by the current Primary Adapter (ARP has an EtherType of 0x0806 and IP has an EtherType of 0x0800).

   In Blue's ARP Reply, Blue provides the MAC address of its Teaming driver, which is the same as the current Primary Adapter's MAC address (B) (refer to, " HP Network Adapter Teaming and Layer 2/Layer 3 addresses"). Blue also takes note of Red's MAC address (A) and IP address (1.1.1.1) and enters them into its ARP cache. Red receives the reply and enters the MAC address (B) and the IP address of Blue (1.1.1.2) into its own ARP cache.

3. Red transmits a unicast PING Request to Blue using Blue's destination MAC address

   Red can now create a PING Request frame using Blue's MAC address (B). Red sends the PING Request to Blue. Blue receives the frame on its Primary Adapter (B) and notices that a station with an IP address of 1.1.1.1 is asking for it to respond.

4. Blue transmits a broadcast ARP Request asking for Red's MAC address.

   NOTE: The following step may not occur if Blue's ARP table still contains an entry for Red as a result of steps 1 and 2.

   Blue checks its ARP cache for a MAC address entry that matches 1.1.1.1. If Blue does not find one, then Blue broadcasts an ARP Request asking for Red's MAC address.

5. Red transmits a unicast ARP Reply to Blue providing its MAC address.

   NOTE: The following step will not occur if step 4 does not take place.

   Red sees the ARP Request and transmits a unicast ARP Reply directly to Blue providing its MAC address (A). Blue receives the ARP Reply and puts Red's MAC address (A) and IP address (1.1.1.1) in its ARP cache.

6. Blue transmits a unicast PING Reply to Red using Red's destination MAC address.

   The final step in the conversation is for Blue to transmit a PING Reply to Red. However, because Blue's Team is running TLB, it must make a load balancing decision before transmitting the PING Reply. The load balancing decision is made by using either Red's MAC address or Red's IP address. Once Blue decides which network adapter to use, it transmits a unicast PING Reply to Red using Red's MAC address (A).

   If Blue chooses to transmit from the Primary Adapter, Red will receive a PING Reply from Blue with a source MAC address of "B", destination MAC address of "A", a source IP address of "1.1.1.2" and a destination IP address of "1.1.1.1". However, if Blue chooses to transmit from the Non-Primary Adapter, Red will receive a PING Reply from Blue with a source MAC address of "E", destination MAC address of "A", a source IP address of "1.1.1.2" and a destination IP address of "1.1.1.1". Either way, Red only distinguishes that it received a PING Reply from the Layer 3 address, "1.1.1.2" (refer to "TLB Transmit Balancing Algorithm" for a complete discussion). The user sees the PING Reply message printed on the screen. This completes the entire conversation.

TLB transmit balancing algorithm

A Team in TLB mode attempts to load balance transmitted frames. In order to avoid frames being transmitted out of order when communicating with a single network device, the load-balancing algorithm assigns conversations to a particular adapter. In other words, load balancing is performed on a conversation-by-conversation basis rather than on a frame-by-frame basis. To accomplish this, when making a decision about which teamed adapter will transmit the frame, the algorithm uses the destination MAC address or the destination IP address of the frame to be transmitted.

It is very important to understand the differences the algorithm uses when deploying HP Network Adapter Teaming in an environment that requires load balancing of routed Layer 3 traffic. In addition, the algorithm provides for statistical load balancing, not absolute load balancing. Therefore, because load-balancing decisions are made on a conversation basis, it is possible that transmitted frames will not be equally distributed across all adapters in a Team.

Implementers of HP Network Adapter Teaming can choose the appropriate algorithm for load balancing via the HP Network Teaming and Configuration GUI.

TLB and layer 2 load balancing using MAC address

This algorithm makes load-balancing decisions based on the destination MAC address of the frame being transmitted by the Teaming driver. The destination MAC address of the frame is the MAC address that belongs to the next network device that will receive the frame. This next network device could be the ultimate destination for the frame or it could be an intermediate router used to get to the ultimate destination. The Teaming driver utilizes the last three bits of the destination MAC address and assigns the frame to an adapter for transmission.

Because MAC addresses are in hexadecimal format, it is necessary to convert them to binary format. For example (refer to Table 1), a MAC address of 01-02-03-04-05-06 (hexadecimal) would be 0000 0001 – 0000 0010 – 0000 0011 – 0000 0100 – 0000 0101 – 0000 0110 in binary format. The Teaming driver load balances based upon the last three bits (110) of the least significant byte (0000 0110 = 06) of the MAC address. Utilizing these three bits, the Teaming driver will consecutively assign destination MAC addresses to each functional network adapter in its Team starting with 000 being assigned to network adapter 1, 001 being assigned to network adapter 2, and so on. Of course, how the MAC addresses are assigned depends on the number of network adapters in the TLB Team and how many of those adapters are in a functional state.

table 1. Load Balancing Based on Destination MAC Address*

Two Port Team		**Three Port Team**	
Destination MAC	Transmitting Adapter	Destination MAC	Transmitting Adapter
000	network adapter 1	000	network adapter 1
001	network adapter 2	001	network adapter 2
010	network adapter 1	010	network adapter 3
011	network adapter 2	011	network adapter 1
100	network adapter 1	100	network adapter 2
101	network adapter 2	101	network adapter 3
110	network adapter 1	110	network adapter 1
111	network adapter 2	111	network adapter 2
Four Port Team		**Five Port Team**	
Destination MAC	Transmitting Adapter	Destination MAC	Transmitting Adapter
000	network adapter 1	000	network adapter 1
001	network adapter 2	001	network adapter 2
010	network adapter 3	010	network adapter 3
011	network adapter 4	011	network adapter 4
100	network adapter 1	100	network adapter 5
101	network adapter 2	101	network adapter 1
110	network adapter 3	110	network adapter 2
111	network adapter 4	111	network adapter 3

* Destination MAC represents only the last three bits of the least significant byte of the address

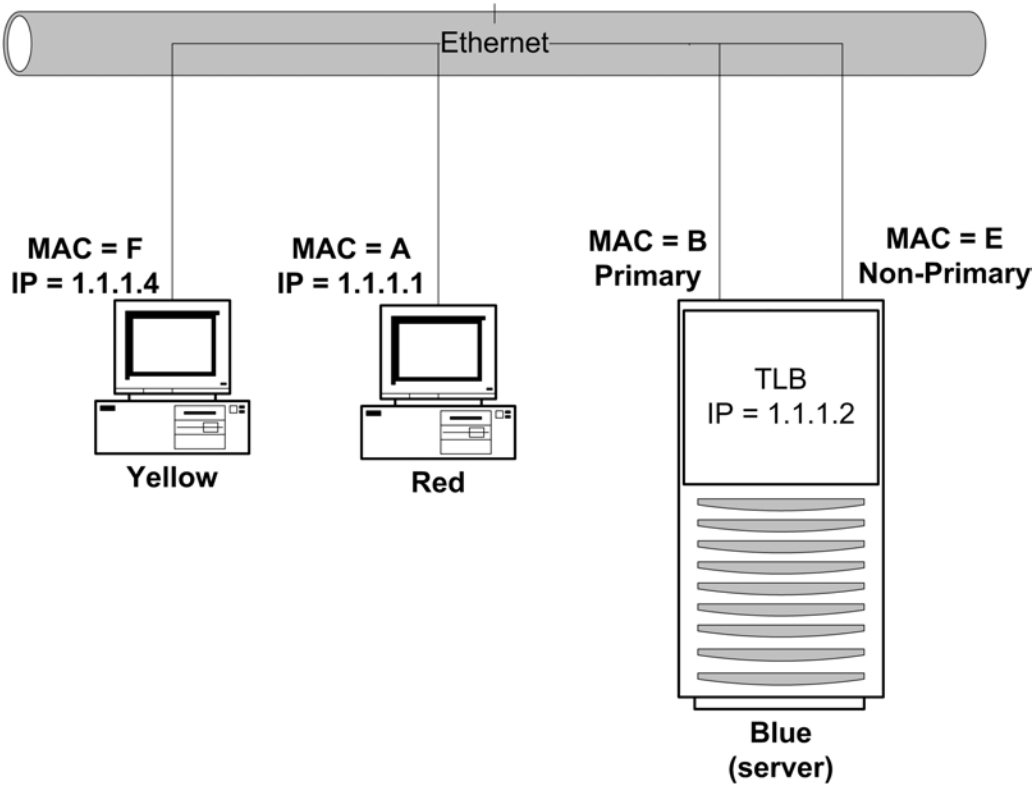
scenario 2-B: a TLB team using MAC address based load balancing

Taking the concepts reviewed in the section titled, "Scenarios of Network Addressing and Communication – Scenario 2-A", this section describes how TLB MAC addressed based load-balancing functions.

Beginning at the point in Scenario 1-A where Blue/1.1.1.2 transmits the PING Reply to Red/1.1.1.1, Blue must decide whether to use network adapter B or E. Blue's Teaming driver calculates using the MAC address of Red (A) because Red is the frame's destination. Because a hexadecimal "A" is equal to "1010" in binary, and the last three bits (010) are used to determine the transmitting network adapter (refer to Table 1 – Two Port Team), "010" is assigned to network adapter 1 (or the Primary Adapter). Therefore, when communicating with Red, Blue will always use the Primary Adapter to transmit frames.

If Blue transmits a frame to Yellow, the same calculation must be made. Yellow's MAC address is hexadecimal "F," which is equal to "1111" in binary. Blue's Teaming driver will again use the last three bits to determine which network adapter will transmit the frame. Referring to Table 1 for a Team with two network adapters, "111" is assigned to network adapter 2 (or the Non-Primary Adapter). Therefore, when communicating with Yellow, Blue will always use the Non-Primary Adapter to transmit frames.

figure 7. TLB Team using MAC address for Load Balancing algorithm



TLB and layer 3 load balancing using IP address

This algorithm makes load-balancing decisions based on the destination IP address of the frame being transmitted by the Teaming driver. The frame's destination IP address is that which belongs to the network device that will ultimately receive the frame. The Teaming driver utilizes the last three bits of the destination IP address to assign the frame to an adapter for transmission.

Because IP addresses are in decimal format, it is necessary to convert them to binary format. For example, an IP address of 1.2.3.4 (dotted decimal) would be "0000 0001 . 0000 0010 . 0000 0011 . 0000 0100" in binary format. The Teaming driver only uses the last three bits (100) of the least significant byte (0000 0100 = 4) of the IP address. Utilizing these three bits, the Teaming driver will consecutively assign destination IP addresses to each functional network adapter in its Team starting with 000 being assigned to network adapter 1, 001 being assigned to network adapter 2, and so on. Of course, how the IP addresses are assigned depends on the number of network adapters in the TLB Team and how many of those adapters are in a functional state (refer to Table 2).

table 2. Load Balancing Based on Destination IP Address*

Two Port Team		**Three Port Team**	
Destination IP	Transmitting Adapter	Destination IP	Transmitting Adapter
000	network adapter 1	000	network adapter 1
001	network adapter 2	001	network adapter 2
010	network adapter 1	010	network adapter 3
011	network adapter 2	011	network adapter 1
100	network adapter 1	100	network adapter 2
101	network adapter 2	101	network adapter 3
110	network adapter 1	110	network adapter 1
111	network adapter 2	111	network adapter 2
Four Port Team		**Five Port Team**	
Destination IP	Transmitting Adapter	Destination IP	Transmitting Adapter
000	network adapter 1	000	network adapter 1
001	network adapter 2	001	network adapter 2
010	network adapter 3	010	network adapter 3
011	network adapter 4	011	network adapter 4
100	network adapter 1	100	network adapter 5
101	network adapter 2	101	network adapter 1
110	network adapter 3	110	network adapter 2
111	network adapter 4	111	network adapter 3

* Destination IP represents only the last three bits of the least significant byte of the address

scenario 2-C: a TLB team using IP address based load balancing

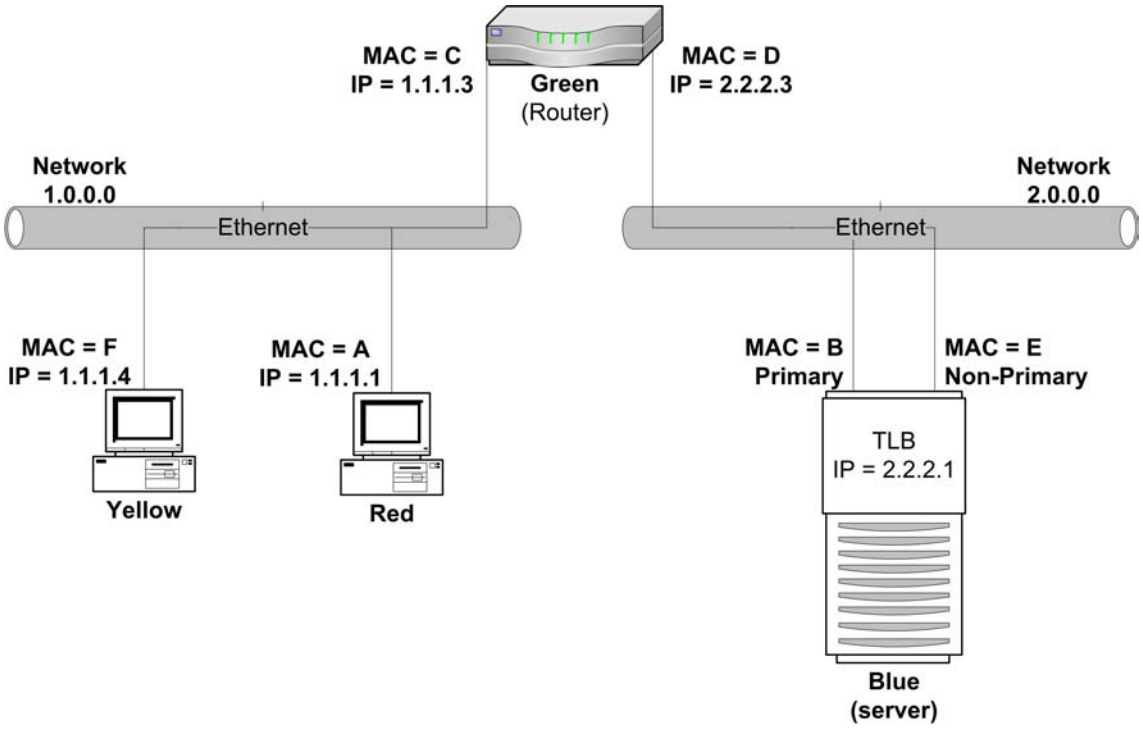
Taking the concepts previously reviewed in Scenario 2-A of the section titled, "Scenarios of Network Addressing and Communication", and figure 8, this section describes how TLB IP addressed- based load-balancing functions.

Beginning at the point in Scenario 2-A where Blue/2.2.2.1 transmits the PING Reply to Red/1.1.1.1, Blue must decide whether to use network adapter B or E. Blue's Teaming driver calculates using the IP address of Red (1.1.1.1) because Red is the frame's destination. Because a dotted decimal "1.1.1.1" is equal to "0000 0001 . 0000 0001 . 0000 0001 . 0000 0001" in binary, and the last three bits (001) are used to determine the transmitting network adapter (refer to Table 2 - Two Port Team), "001" is assigned to network adapter 2 (or the Non-Primary Adapter). Therefore, when communicating with Red, Blue will always use the Non-Primary Adapter to transmit frames.

If Blue transmits a frame to Yellow, the same calculation must be made. Yellow's IP address in dotted decimal is "1.1.1.4" and equal to "0000 0001 . 0000 0001 . 0000 0001 . 0000 0100" in binary. Blue's Teaming driver will again use the last three bits to determine which network adapter will transmit the frame. Referring to Table 2 - Two Port Team, "100" is assigned to network adapter 1 (or the Primary Adapter). Therefore, when communicating with Yellow, Blue will always use the Primary Adapter to transmit frames.

It is important to note that if an implementer uses the MAC address load balancing algorithm for the network in figure 8, load balancing will not function as expected, and traffic will not be load balanced using all Teamed network adapters. Because Blue transmits all frames destined for Red and Yellow via Green (Blue's Gateway), Blue uses Green's Layer 2 address (MAC) as the frame's DESTINATION MAC ADDRESS but uses Red's and Yellow's Layer 3 addresses (IP) as the frame's DESTINATION IP ADDRESS. Blue never transmits frames directly to Red's or Yellow's MAC address because Blue is on a different Layer 2 network. Because Blue always transmits to Red and Yellow using Green's MAC address, the Teaming driver will assign all conversations with clients on Network 1.0.0.0 to the same network adapter. When an HP Network Adapter Team needs to load balance traffic that traverses a Layer 3 device (Router), IP address based load balancing should be used.

figure 8. TLB Team using IP address for Load Balancing algorithm



TLB applications

TLB is deployed in environments that require fault tolerance and additional transmit throughput greater than the capacity of the Primary Adapter. TLB environments do not require receive throughput greater than the capacity of the Primary Adapter. For example, in a database server whose primary role is that of transmitting data to clients, receive throughput requirements may be much smaller than the transmit requirements because database requests require less bandwidth than transmitting database content.

recommended configurations for a TLB environment

HP recommends that:

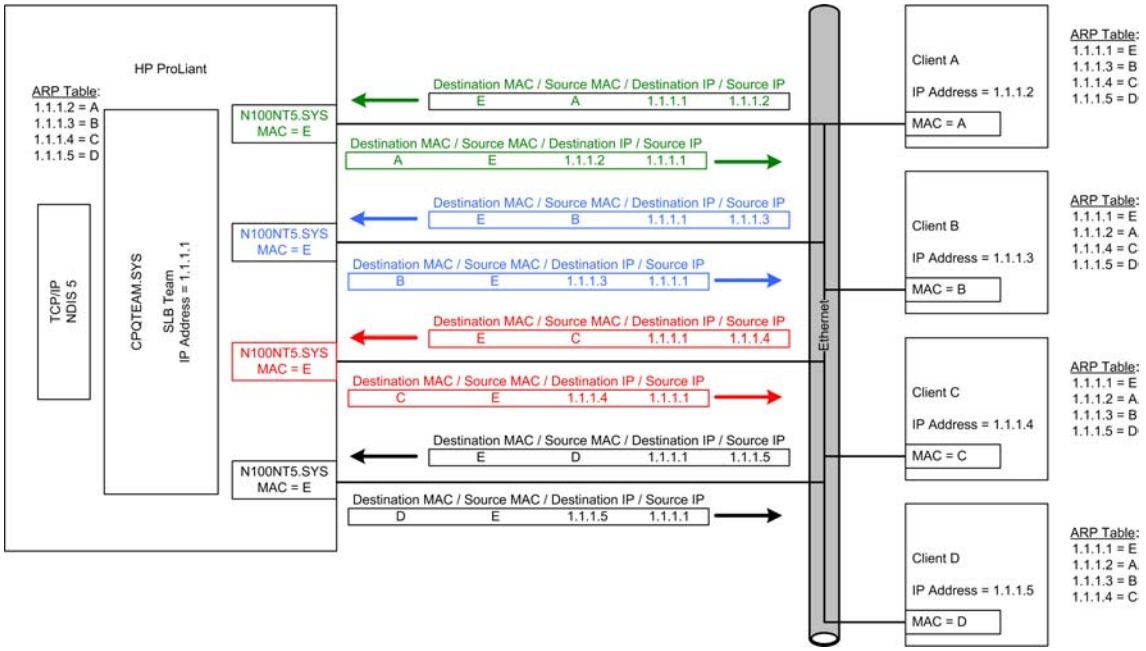
- Heartbeats be enabled (default)
- MAC addresses not be manually set to a locally administered address (LAA) via the Microsoft UI. A user should not implement LAAs on individual network adapters that are members of a Team, otherwise Teaming may not function correctly. Setting an LAA for the Team is permitted via the HP Network Adapter Teaming and Configuration GUI.
- Spanning Tree's blocking, listening, and learning stages be disabled, or bypassed, on all switch ports to which an HP Network Adapter Team port is attached. These stages are not needed when a non-switch networking device (e.g. server) is attached to the switch port. HP ProCurve switches have a feature called STP Fast Mode that is used to disable these Spanning Tree stages on a port-by-port basis. Cisco switches have an equivalent feature called PortFast.
- Team members can be split across more than one switch in order to achieve switch redundancy. However, all switch ports that are attached to members of the same Team must comprise a single broadcast domain (i.e., same VLAN). Additionally, if problems exist after deploying a Team across more than one switch, reattach all Team members to the same switch. If the problems disappear, then the cause of the problem resides in the configuration of the switches and not in the configuration of the Team. If switch redundancy is required (Team members are attached to two different switches), then HP recommends that the switches be deployed with redundant links between them and Spanning Tree be enabled (or other Layer 2 redundancy mechanisms) on the ports that connect the switches. This helps prevent switch uplink failure scenarios that leave Team members in separate broadcast domains.
- TLB Teams that communicate with TCP/IP network devices via a router should use the IP address-based load balancing algorithm (configured via the HP Network Teaming and Configuration GUI).

Switch-Assisted Load Balancing (SLB)

Switch-assisted Load Balancing mode, formerly known as Fast EtherChannel mode (FEC) or Gigabit EtherChannel mode (GEC), incorporates all the features of NFT and TLB, and adds the feature of load balancing of received traffic. In this mode, two to eight adapters may be teamed together as a single virtual network adapter. The load-balancing algorithm used in SLB allows for the load balancing of the server's transmit and receive traffic (refer to figure 9). Unlike TLB, which only load balances IP traffic, SLB load balances all traffic regardless of the Protocol.

Switch-assisted Load Balancing (SLB) is an HP term that refers to an industry standard technology for grouping multiple network adapters into one virtual network adapter and multiple switch ports one virtual switch port. HP's SLB technology works with multiple switch vendors' technologies. Other compatible technologies include: HP ProCurve Port Trunking, Cisco Fast EtherChannel (FEC)/Gigabit EtherChannel (GEC) (Static Mode Only – no PAgP), IEEE 802.3ad Link Aggregation (Static Mode only – no LACP), Bay Network MultiLink Trunking, and Extreme Network® Load Sharing. Switch-assisted Load Balancing (SLB) is not the same thing as Server Load Balancing (SLB) as used by some switch vendors. Switch-assisted Load Balancing operates independently of, and in conjunction with, Server Load Balancing.

figure 9. Overview of SLB communication



Unlike NFT and TLB, SLB does not incorporate the concepts of Primary and Non-Primary Adapters within a Team. All adapters within a Team are considered equal and perform identical functions as long as the particular adapter is in a functioning state. The algorithm for load balancing transmit traffic used by SLB is identical to the algorithm used by TLB. Unlike TLB, SLB load balances all traffic regardless of the protocol being used.

SLB and layer 3 load balancing using IP address

The algorithm for load balancing transmit traffic used by SLB is identical to the algorithm used by TLB (refer to "TLB and Layer 3 load balancing using IP address").

Switch-assisted load balancing receive balancing algorithm

The switch determines which load balancing algorithm is used to load balance receive traffic for an SLB Team. An SLB Team does not control which adapter in the Team receives the incoming traffic. Only the switch can choose which adapter to use to send the traffic to the server. Therefore, please consult the switch manufacturer to determine the algorithm the switch uses.

Switch-assisted load balancing and Cisco EtherChannel® technology

Teamed Cisco's Fast EtherChannel (FEC) and Gigabit EtherChannel (GEC) technology is a MAC layer (Layer 2) load balancing technology using two to eight network adapters grouped together as one logical network adapter. Depending on the specific load-balancing algorithm used, FEC/GEC may not efficiently load balance traffic to network adapters.

FEC/GEC was originally designed as a switch-to-switch technology allowing two switches to increase the bandwidth between each other by aggregating multiple ports together as a single logical port for both transmits and receives. This is in contrast to Transmit Load Balancing (TLB) that only balances transmit traffic. An algorithm had to be used that could statistically divide the traffic over each port in the FEC/GEC group in an attempt to divide it evenly.

There have been at least three algorithms that have been developed: source-based, destination-based, and XOR (refer to Table 3). The Source-based algorithm utilizes the last one or two bits (depending on the number of ports in the FEC/GEC group) of the source address in the packet. If the bit is 0, the first port is used. If the bit is 1, the second port is used. The Destination-based algorithm utilizes the last one or two bits (depending on the number of ports in the FEC/GEC group) of the destination address in the packet. If the bit is 0, the first port is used. If the bit is 1, the second port is used. The XOR algorithm utilizes the last one or two bits (depending on the number of ports in the FEC/GEC group) of the destination AND source addresses in the packet. The algorithm XORs the bits. If the result is 0, then the first port is used. If the result is 1, then the second port is used.

FEC/GEC has developed into not only a switch-to-switch technology but also a switch-to-node technology. In most cases, the node is a multi-homed server with network adapter drivers that support FEC/GEC. Problems can arise with switches using the destination-based algorithm when switch-to-node FEC/GEC is used. Because the destination address of the FEC/GEC node is always the same, the switch always sends traffic to that server on the same port. Because of this, receive traffic is not evenly distributed across all ports in the FEC/GEC group.

table 3. Example of Load Balancing Algorithms

Preliminary Information:		
MAC address of a two Port FEC/GEC group	00-00-00-00-00-01	Hexadecimal
Last Byte (01) Conversion to Binary	0000 0001	Binary
MAC address of Client1	00-00-00-00-00-02	Hexadecimal
Last Byte (02) Conversion to Binary	0000 0010	Binary
MAC address of Client2	00-00-00-00-00-03	Hexadecimal
Last Byte (03) Conversion to Binary	0000 0011	Binary
Packet 1 is a frame transmitted from Client 2 to the two port FEC/GEC group. Packet 2 is a frame transmitted from Client 1 to the two port FEC/GEC group.		
Method 1: Destination-based Algorithm		
1. Packet 1 – Destination MAC address: 00-00-00-00-00-01 Last binary bit = 1 so frame is transmitted on port 2.		
2. Packet 2 – Destination MAC address: 00-00-00-00-00-01 Last binary bit = 1 so frame is transmitted on port 2.		
Method 2: Source-Based Algorithm		
1. Packet 1 – Source MAC address: 00-00-00-00-00-03 Last binary bit = 1 so frame is transmitted on port 2.		
2. Packet 2 – Source MAC address: 00-00-00-00-00-02 Last binary bit = 0 so frame is transmitted on port 1.		
Method 3: XOR Algorithm (best method to use on switch)		
1. Packet 1 – Source MAC address: 00-00-00-00-00-03 Last binary bit = 1. Packet 1 – Destination MAC address: 00-00-00-00-00-01 Last binary bit = 1. XOR result of binary bits 1 & 1 = 0, so frame is transmitted on port 1.		
2. Packet 2 – Source MAC address: 00-00-00-00-00-02 Last binary bit = 0. Packet 2 – Destination MAC address: 00-00-00-00-00-01 Last binary bit = 1. XOR result of binary bits 0 & 1 = 1, so frame is transmitted on port 2.		

The effects of the destination-based algorithm do not indicate a fault in the network adapter drivers nor on the switch. Destination-based load balancing is considered a functional FEC/GEC algorithm because packets between switches may not always use the same destination or source addresses. Only single-node-to-switch FEC/GEC uses the same destination address.

The algorithm used for load balancing has no effect on fault tolerance and fault tolerance will function the same in any implementation.

Some switches have the option to change the load-balancing algorithm. In such cases, HP advises using the algorithms in this order of preference: XOR, source-based, destination-based.

network addressing and communication using SLB

SLB functions identically to TLB (refer to "Network Addressing and Communication using TLB" and the scenarios described in that section) except in its use of MAC addresses. SLB requires a switch capable of grouping multiple switch ports as a single switch port and SLB Teaming uses the same MAC address on all network adapters in the same Team. This does not violate IEEE standards because the switch is fully aware of the port groupings and expects that all network adapters will transmit using the same MAC address.

SLB applications

Switch-assisted Load Balancing is deployed in environments that require fault tolerance and additional transmit and receive throughput greater than the capacity of the Primary Adapter and that have a switch capable of providing, and configured to provide load balancing assistance (e.g., a backup server that requires additional receive throughput for backing up other servers and clients).

recommended configurations for an SLB environment

HP recommends that:

- Transmit Validation Heartbeats be enabled (default)
- MAC addresses not be manually set to a locally administered address (LAA) via the Microsoft UI. A user should not implement LAAs on individual network adapters that are members of a Team, otherwise Teaming may not function correctly. Setting an LAA for the Team is permitted via the HP Network Adapter Teaming and Configuration GUI.
- Spanning Tree's blocking, listening, and learning stages be disabled, or bypassed, on all switch ports to which an HP Network Adapter Team port is attached. These stages are not needed when a non-switch networking device (e.g. server) is attached to the switch port. HP ProCurve switches have a feature called STP Fast Mode that is used to disable these Spanning Tree stages on a port-by-port basis. Cisco switches have an equivalent feature called PortFast.
- SLB Teams that communicate with TCP/IP network devices via a router, that the IP address-based load-balancing algorithm be used (configured via the HP Network Teaming and Configuration GUI).
- Implementers thoroughly understand the configuration guidelines set by the switch vendor because SLB is dependent on the switch being configured in a compatible mode. HP's SLB technology has been designed to allow for flexibility. Therefore, the HP Network Teaming and Configuration GUI may allow configuration of an SLB Team that will not work correctly with a particular vendor's switch.
- The switch's load balancing algorithm be set to XOR or SOURCE-BASED but not DESTINATION-BASED (refer to "Switch-assisted Load Balancing and Cisco's EtherChannel Technology").
- The switch's load balancing algorithm be set to balance by IP address if most traffic destined for the server originates on a different network and must traverse a router.

network adapter failover

NFT and network adapter failure recovery

There are three operating modes available for NFT Teams: Manual, Fail On Fault, and Preferred Primary.

- Manual Mode

  This mode for NFT is used for user-initiated failovers (manual failovers). When set, Manual mode allows an NFT Team to automatically failover during events that normally cause a failover (e.g., a cable is unplugged on the Primary Adapter of an NFT Team), however Manual mode also allows the Team to manually failover with the click of a button. Manual mode is normally used for troubleshooting purposes (e.g., using an analyzer to take an inline network trace).

- Fail On Fault Mode

  The second mode available for NFT is Fail On Fault. In this mode, an NFT Team will initiate a failover from the Primary Adapter to an operational Non-Primary Adapter whenever a failover event occurs (refer to "Failover Events") on the Primary Adapter. When the failover occurs, the two adapters swap MAC addresses so the Team remains known to the network by the same MAC address. The new Primary Adapter is considered just as functional as the old Primary Adapter. If the old Primary Adapter is restored, it becomes a Non-Primary Adapter for the Team but no MAC address changes are made unless there is another failover event on the Primary Adapter.

- Preferred Primary Mode

  The last mode available for NFT is Preferred Primary mode. When choosing Preferred Primary mode, the operator is presented with a drop down box to select the "Preferred Primary Adapter". The operator should choose the adapter that, for a particular reason, is best suited to be the Primary Adapter. For instance, if an NFT Team were to be created using a Gigabit adapter and a 10/100 adapter, an operator would choose the Gigabit adapter as the Preferred Primary Adapter, because of the increased bandwidth available with a Gigabit adapter.

  When an adapter is chosen as the Preferred Primary Adapter, it will be used as the Primary Adapter whenever it is in an operational state. If the Preferred Primary Adapter experiences a failover event, the NFT Team fails over to a Non-Primary Adapter. If the Preferred Primary Adapter is restored, the Team will then initiate a failback to the Preferred Primary Adapter. Essentially, the Team will initiate a failover twice even though only one error occurred. The second failover is to make the restored Preferred Primary Adapter the Team's Primary Adapter once again. A failover back to the Preferred Primary is more specifically referred to as a failback.

  NOTE: Failures of Non-Primary Adapters do not trigger any type of recovery because Non-Primary Adapters are already in standby mode. The only consequence of a failed Non-Primary Adapter is the possibility of the Primary Adapter failing and the Team becoming unavailable to the network because both adapters in the Team are in a failed state. If there are three or more network adapters in a Team and two adapters fail, the Team is still available via the third adapter.

TLB and network adapter failure recovery

With TLB, the recovery mechanism provided is very similar to the NFT failover mode discussed in section titled, "Fail On Fault". In a two port TLB Team, the primary adapter receives all data frames, while the Non-Primary Adapter receives only heartbeat frames. Both adapters are capable of transmitting data frames. In the event of a failover, the Non-Primary Adapter becomes the Primary Adapter and assumes the MAC address of the Team. In effect, the two adapters swap MAC addresses. The new Primary Adapter now receives and transmits all data frames. If the old Primary Adapter is restored, it becomes a Non-Primary Adapter for the Team. It will now only receive heartbeat frames and be capable of transmitting data frames. If a Non-Primary Adapter fails in a two-port Team, the data frames being load balanced by the adapter are transmitted by the Primary Adapter. If a Non-Primary Adapter is restored, it remains Non-Primary, and the Team will resume load balancing data frames on that adapter. No MAC address changes are made when a Non-Primary Adapter fails or is restored.

SLB and network adapter failure recovery

With SLB, the recovery mechanism is somewhat different than those discussed previously. All members of the Team transmit and receive frames with the same MAC Address and there is no concept of a Primary or Non-Primary Adapter as there is in NFT and TLB. With SLB, there are no heartbeat frames, and consequently no heartbeat failovers. In a two-port SLB Team, all members are capable of receiving data frames (based on the switch's load balancing algorithm), and transmitting data frames (based on the Teaming Driver's load balancing algorithm). In the event of a failover, all transmit traffic is redistributed among the working adapters. After a failover event in a two-port Team, only one adapter is currently working, so all transmit traffic is sent using it. All receive traffic is determined by the switch, which should detect that only one adapter is working. If a failed adapter is restored, all transmit and receive traffic is once again load balanced among all adapters.

All receive traffic is still determined by the switch algorithm, which should detect that both adapters are functional. If the switch sends a frame destined for the Team MAC address to any of the "operational" adapters in the Team, the adapter will receive it. The HP Network Adapter Teaming driver does not control frames received, but only load balances the transmit traffic. All protocols are load balanced, not just IP. Remember that the Teaming driver load balances network conversations per teamed adapter, and therefore it is possible that transmit traffic being sent out a particular adapter to a particular device could change ports after a failure on another adapter in the Team. This can occur because the Teaming driver algorithm redistributes transmit traffic among working adapters every time the state of any member of the Team changes.

Unlike NFT and TLB, a failure on any adapter within the same SLB Team has the same ramifications as a failure on any other adapter because all adapters are considered equal.

failover events

link loss

When a network adapter is a member of a Team and loses physical link (i.e., wire fault, link light is lost), the Teaming driver disables that adapter in the Team. If this adapter is in use by the Team, the Team recovers from the failure based on the adapter's role in the Team (Primary or Non-Primary) and the Team's mode (NFT, TLB, or SLB), unless this was the last available Team member in the Team. In that case, the Team would fail.

heartbeat failures

The use of Heartbeat frames for network adapter failovers was designed to detect network adapter communication failure even in cases when it maintained physical link. Special Heartbeat frames (refer to "Heartbeats") are transmitted and received by teamed network adapters to validate the transmit and receive paths of each adapter. When a Heartbeat frame is transmitted by one teamed adapter but not received by another teamed adapter, the Teaming driver assumes that one of the adapters is having a communications problem. If the Primary Adapter in an NFT or TLB Team experiences a failover event, a failover may occur to a functional non-Primary adapter.

Heartbeat frames were not designed to monitor for other networking failures such as loss of connectivity between switches, or loss of client connectivity.

heartbeats

Heartbeats are special frames that HP's Network Adapter Teaming uses for validating team member network connections and for notifying other network equipment of MAC address changes as a result of failover events (refer to "Heartbeat Functionality and Timers" for a complete description of the different heartbeat types).

Heartbeat frames contain only Layer 2 addresses (refer to "Heartbeat Frame Format" for the complete frame format of a heartbeat) and do not contain any Layer 3 addresses. This means that heartbeat frames are not routable. In other words, heartbeat frames will not be routed (by a router) between Team members if the Team members are on two different Layer 2 networks joined by a Layer 3 device (router). If the heartbeat frames are not delivered between Team members, then erroneous failovers may occur.

heartbeat frame format

Heartbeat frame size is 84 Bytes including the FCS. Heartbeat frames are LLC Test Frames with 63 bytes of data. The destination address of all Heartbeat frames is a layer 2 multicast address of 03-00-C7-00-00-EE. This is a HP registered multicast address and used only for Heartbeat frames. Below is an example of a Heartbeat frame.

802.2 Frame Format	Value	84 Bytes
Destination MAC Address	03-00-C7-00-00-EE	6 Bytes
Source MAC Address	Varies	6 Bytes
Length	66	2 Bytes
DSAP	0xAA	1 Byte
SSAP	0xAA	1 Byte
SNAP Type	Unnumbered, TEST	1 Byte
Data	Insignificant data	63 Bytes
FCS	Varies	4 Bytes

Heartbeat frames on a tagged VLAN include an extra 4 byte field use for VLAN identification. The frame size for these Heartbeats is 88 Bytes. If multiple VLANs are configured for the Team, Heartbeats are sent out using the lowest VLAN ID configured for the Team. For example, if a Team of two adapters are configured for VLAN 20, 40, and 60, Heartbeat frames would only be sent on VLAN 20. Below is an example of what the Heartbeat frame would look like.

802.2 Frame Format	Value	88 Bytes
Destination MAC Address	03-00-C7-00-00-EE	6 Bytes
Source MAC Address	Varies	6 Bytes
802.1Q Tag	Varies	4 Bytes
Length	66	2 Bytes
DSAP	0xAA	1 Byte
SSAP	0xAA	1 Byte
SNAP Type	Unnumbered, TEST	1 Byte
Data	Insignificant data	63 Bytes
FCS	Varies	4 Bytes

heartbeat functionality and timers

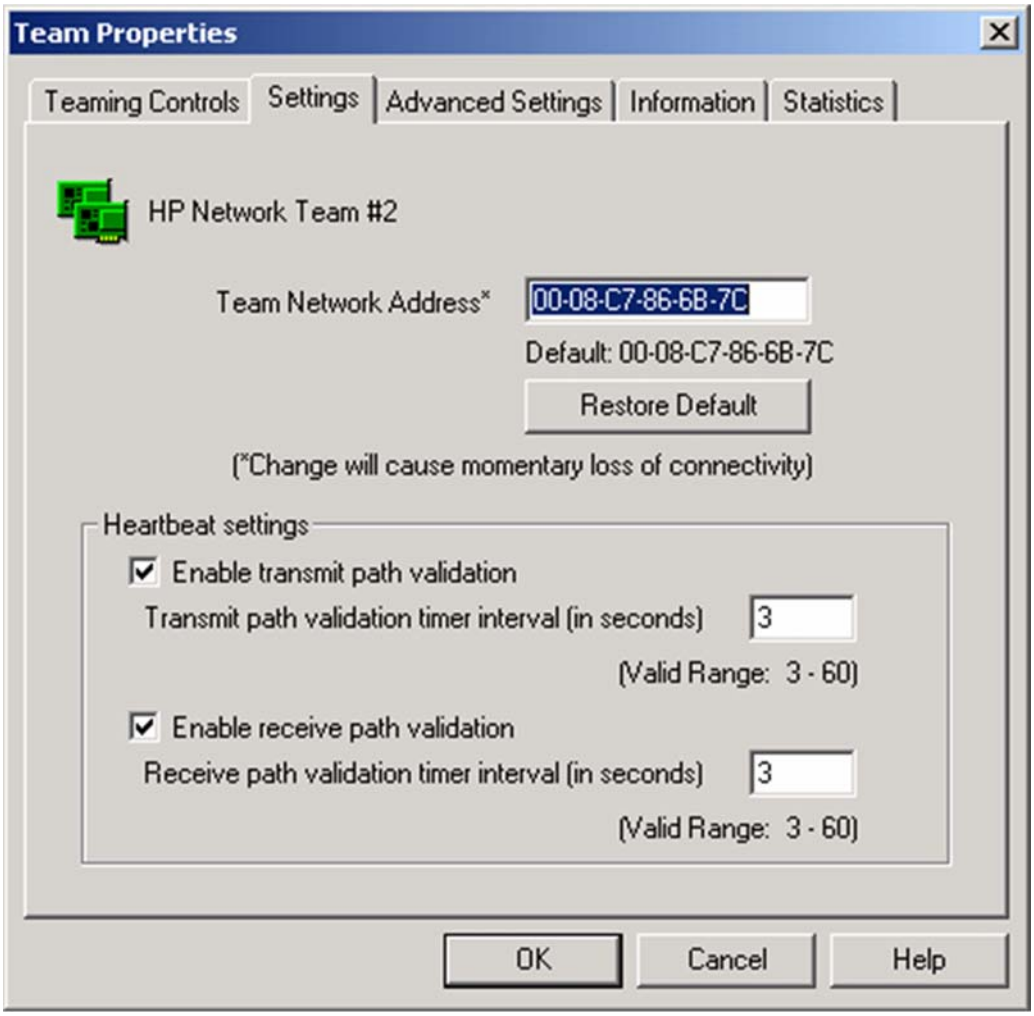
There are three main functions of the heartbeat frame used for HP Network Adapter Teaming: Receive Validation, Transmit Validation, and Switch MAC Table Updates. The same Heartbeat frame format is used for all heartbeat functions and they are indistinguishable from the network's perspective.

HP Network Adapter Teaming utilizes a mechanism of timers for checking network connectivity with Heartbeats. Two timers are available for custom tuning on the Settings Tab of the Properties Menu of a Team in the HP Network Teaming and Configuration GUI (refer to figure 10).

- Transmit Path Validation
- Receive Path Validation

NOTE: Transmit and Receive path validation is enabled by default with a default value for these timers of 3 seconds. If manual configuration is desired, the valid range for both timers is 3 to 60 seconds.

figure 10. Screenshot of Team Properties/Settings tab containing the Heartbeat settings



transmit path validation

Transmit Path Validation is used for checking the transmit path for all teamed adapters (Primary and Non-Primary. At the Transmit Path Validation timer interval (Default = 3 seconds), the Team checks to determine if each adapter has successfully transmitted any frame since the last timer interval. If not, then the adapter's internal status is incremented by 1 (starts at 0). After the adapter's internal status has degraded to 3 (4 transitions) without transmitting something, that adapter will transmit a Heartbeat frame. Therefore, a Transmit Path Validation Heartbeat frame may be transmitted once every 12 seconds (if Transmit Path Validation interval timer is set to default of 3 seconds) on an adapter that has not transmitted anything.

receive path validation

Receive Path Validation is used for checking the receive path for all teamed adapters (Primary and Non-Primary). At the Receive Path Validation timer interval (Default = 3 seconds), the Team checks to determine if each adapter has successfully received any frame since the last timer interval. If not, then the adapter's internal status increases by 1. After a Primary Adapter's internal status has degraded to 3 without receiving something, all Non-Primary Adapters will transmit a Heartbeat frame with the intention that the Primary will receive at least one of the transmitted Heartbeat frames. If the Primary Adapter does receive one of these Heartbeats, its internal status is reset to 0. If the Primary Adapter still does not receive a frame, then the Primary Adapter is placed in a failed state and the Team fails over to a functional Non-Primary Adapter.

If one of the Non-Primary Adapters has degraded to 3, then the Primary Adapter transmits a Heartbeat frame. 3 seconds multiplied by 3 state transitions = 9 seconds. Therefore, a Receive Path Validation Heartbeat will be transmitted from the Primary Adapter to a Non-Primary Adapter or from all Non-Primary Adapters (note that this is plural in a Team of three or more members) to the Primary Adapter every nine (9) seconds if there is no receive activity on a particular Team member.

switch MAC table update with team address heartbeat

The Primary Adapter must ensure that the switch has the Team's MAC address on the correct port, and that the switch keeps the Team's MAC address in its MAC table (or CAM table). The Switch MAC table update Heartbeat is used with NFT/TLB Teams when a failover occurs, and a Non-Primary adapter takes over the Primary role. This allows the new Primary Adapter to update the Switch MAC table immediately so traffic destined for the Team will make smooth transition and not experience timeouts or failures. This feature cannot be disabled and is always active on a NFT/TLB Team regardless whether receive or transmit path validation heartbeats are enabled or disabled.

team status and icons

adapter's teamed status

The HP Teaming and Configuration GUI reports a "Teamed Status" on the Information Tab of each adapter in the system. This status allows the user to understand the condition of the adapter's teamed status and take corrective action if necessary.

Listed below are the possible status conditions with definitions.

- Not Teamed – The selected adapter is not part of a Team
- Available – The selected adapter is part of a Team and is currently fulfilling its role
- Wire Fault – The selected adapter is part of a Team and cannot fulfill its role because it does not have a valid link
- Heartbeat Failure – The selected adapter is part of a Team and cannot fulfill its role because it has failed to receive a Heartbeat during the previous Heartbeat cycle
- Transmit Failure – The selected adapter is part of a Team and cannot fulfill its role because it has failed to transmit a packet

- Not Joined – The selected adapter is configured as a Team member, but was not allowed to join because it has a property setting that differs from that of the adapters which were already joined in the Team. This could also happen if a user changes a property setting via the Microsoft UI (user interface) for a teamed adapter, and it now differs from other Team members.

  NOTE: Certain "offloading features" are an example of property settings that might differ between adapters and would cause a "Not Joined" status.

Refer to the HELP file for HP Network Adapter Teaming for additional information and examples of adapter icons that represent each of these "Teamed Status" states.

team state

The HP Teaming and Configuration GUI reports a "Team State" on the Information Tab of each configured Team. This status allows the user to understand the overall condition of the Team and take corrective action if necessary.

Listed below are the possible state conditions with definitions.

- Ok – The Team is functioning properly. (Green)
- Degraded – The Team is functioning, but one or more members is not available to fulfill his role. (Yellow)
- Failed – The Team has failed and connectivity has been lost. (Red)
- Unknown – The Team has been created, or the membership has changed, however the GUI must be closed for changes to take effect. (White)
- Disabled – The Team has been disabled either through Device Manager or the Microsoft UI.

team icons

Team icons are available to allow the user to easily recognize status and conditions of Teams. Refer to figure 11 for Team State Icons. Refer to the HELP file for HP Network Adapter Teaming for additional information.

figure 11. Team State Icons

hp network adapter teaming and advanced networking features

When adapters are part of a Team, some advanced features are no longer configurable on the adapter's "Advanced Settings" Tab. These advanced features or either promoted to the Team's "Advanced Settings" Tab because this feature is compatible and supported by all Team members, or not promoted to the Team's Advanced Settings" Tab because one or more adapters do not support this feature. There are also special cases where a feature is supported by all Team members but the implementation is not compatible between the Team members because they use different miniport drivers, and in this case it is not promoted to the Team's "Advanced Settings" Tab.

checksum offloading

There are up to four features that are supported on certain HP NC Series network adapters; Receive TCP Checksum Offloading, Transmit TCP Checksum Offloading, Transmit IP Checksum Offloading, and Receive IP Checksum Offloading. HP Network Adapter Teaming supports the Offloading advanced feature when it is compatible and supported by all the adapters in the Team. If one or more adapter(s) in the Team has the feature enabled, and the feature is compatible and supported by all adapters in the Team, the feature is promoted to the Team's "Advanced Settings" Tab as ENABLED and the Teaming driver automatically enables this feature on all adapters. If all adapters in the Team have the feature disabled, and the feature is compatible and supported by all adapters in the Team, the feature is promoted to the Team's "Advanced Settings" tab as DISABLED.

- NC61xx and NC71xx adapters have compatible offloading features
- NC67xx and NC77xx adapters have compatible offloading features.

NOTE: If a Team contains NC61xx/71xx and NC67xx/NC77xx adapters, the offloading features are not promoted to the Team's "Advanced Settings" Tab because they are not compatible.

802.1p QoS tagging

HP Network Adapter Teaming supports the advanced feature 802.1p QoS, when supported by all adapters in the Team. This feature is used to mark frames with a priority level for transmission across an 802.1p QOS-aware network. If all adapter(s) in the Team have the feature enabled, the feature is promoted to the Team's "Advanced Settings" Tab as ENABLED. If one or more adapters in the Team have the feature disabled, and the feature is supported by all adapters in the Team, the feature is promoted to the Team's "Advanced Settings" tab as DISABLED. If any one Team member does not support 802.1p QOS, then the feature will not be promoted to the Team's "Advanced Settings" Tab. Most HP NC Series network adapters support this advanced feature.

Large Send Offload (LSO)

HP Network Adapter Teaming supports the advanced feature Large Send Offload (sometimes referred to as TCP Segmentation Offload) when it is compatible and supported by all the adapters in the Team. This feature is only available in Windows Server 2003 and allows adapters in the Team to offload large TCP packets for segmentation in order to improve performance. If one or more adapter(s) in the Team has the feature enabled, and the feature is compatible and supported by all adapters in the Team, the feature is promoted to the Team's "Advanced Settings" Tab as ENABLED and the teaming driver automatically enables this feature on all adapters. If all adapters in the Team have the feature disabled, and the feature is compatible and supported by all adapters in the Team, the feature is promoted to the Team's "Advanced Settings" tab as DISABLED.

- NC6170 and NC7170 adapters have a compatible Large Send Offload feature
- NC67xx and NC77xx adapters have a compatible Large Send Offload feature

NOTE: If a Team contains NC6170/7170 and NC67xx/NC77xx adapters, the Large Send Offload feature is not promoted to the Team's "Advanced Settings" Tab because they are not compatible.

maximum frame size (jumbo frames)

HP Network Adapter Teaming supports the advanced feature Maximum Frame Size (Jumbo Frames) when supported by all adapters in the Team. This feature allows adapters in the Team to increase maximum frame size for TCP/IP packets transmitted and received in order to improve performance. If all adapter(s) in the Team support Jumbo Frames, the feature is promoted to the Team's "Advanced Settings" Tab with the lowest size configured for the Team members. For example, if there are two adapters, one configured for 4088 Bytes and the other for 9014 Bytes, and they are teamed, the Maximum Frame Size feature is set to 4088 Bytes. A setting of 1514 Bytes is equal to Jumbo Frames being disabled. Note that Maximum Frame Size greater than 1514 would constitute Jumbo Frames being enabled. Jumbo Frames are supported on NC Series Gigabit Adapters only. In addition, the specified Maximum Frame Size in HP Network Adapter Teaming does not include the four byte Cyclic Redundancy Check (CRC) portion of an Ethernet frame. Some switch settings do include the CRC portion in their Jumbo Frame Size configuration. Therefore, it may be necessary to increase the switches Jumbo Frame Size setting by four bytes in order for Jumbo Frames to work correctly.

802.1Q Virtual Local Area Networks (VLANs)

The HP Network Teaming and Configuration GUI supports the configuration of VLANs on standalone HP network adapters and on HP Network Adapter Teams. This allows a network adapter or a Team to belong to more than one VLAN at the same time. When multiple VLANs are configured, 802.1Q VLAN Tagging is used to mark every transmitted frame with a VLAN identifier (number between 1 and 4094). The use of VLAN Tagging requires the support of the network infrastructure and/or the receiving network device.

The use of VLANs has only one effect on the operation of the Team, which is that heartbeats are transmitted only on the numerically lowest configured VLAN on the Team. This means that if four VLANs are configured on the Team and the numerically lowest VLAN is "20", the Teaming driver will use VLAN 20 for transmitting heartbeats between Team members.

Much like deploying SLB, the use of VLANs requires the switch or switches to be configured properly. Every Team member's switch port must be configured with the same VLAN configuration. This means that if a Team is to operate on four different VLANs, every Team member must have all four VLANs configured on their respective switch port.

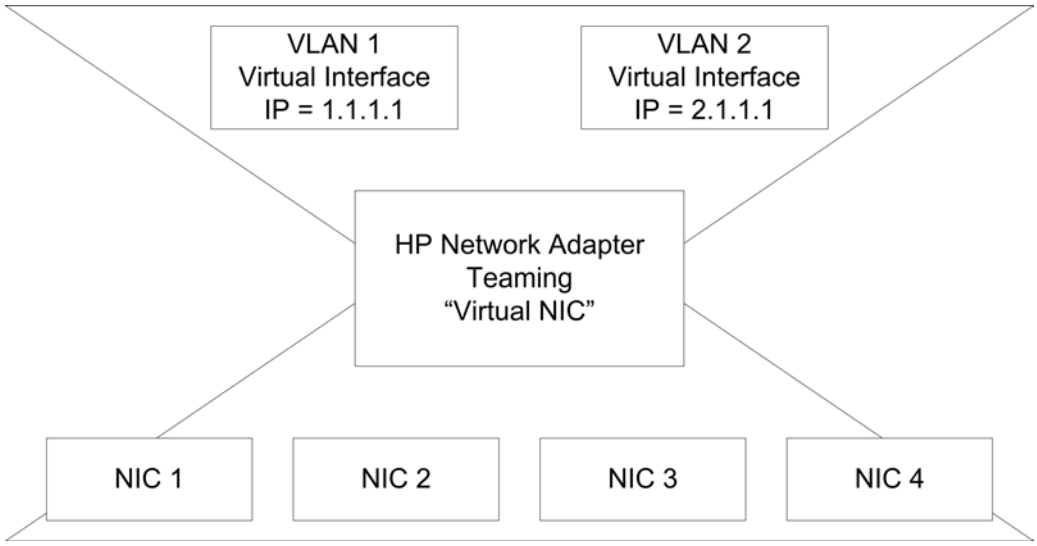
The maximum configurable VLANs per Team is 64. The valid VLAN number range per VLAN is 1 to 4094.

Example of VLAN Tagging used with HP Network Adapter Teaming (refer to figure 12):

- Four NICs teamed together as a single virtual NIC using HP Network Adapter Teaming
- Two VLANs configured on top of the virtual NIC to create two virtual interfaces

- Provides the same functionality to the OS as having two NICs installed but provides for fault tolerance and load balancing across four NICs.

figure 12. VLAN Tagging used with HP Network Adapter Teaming



Internet Group Messaging Protocol (IGMP) snooping

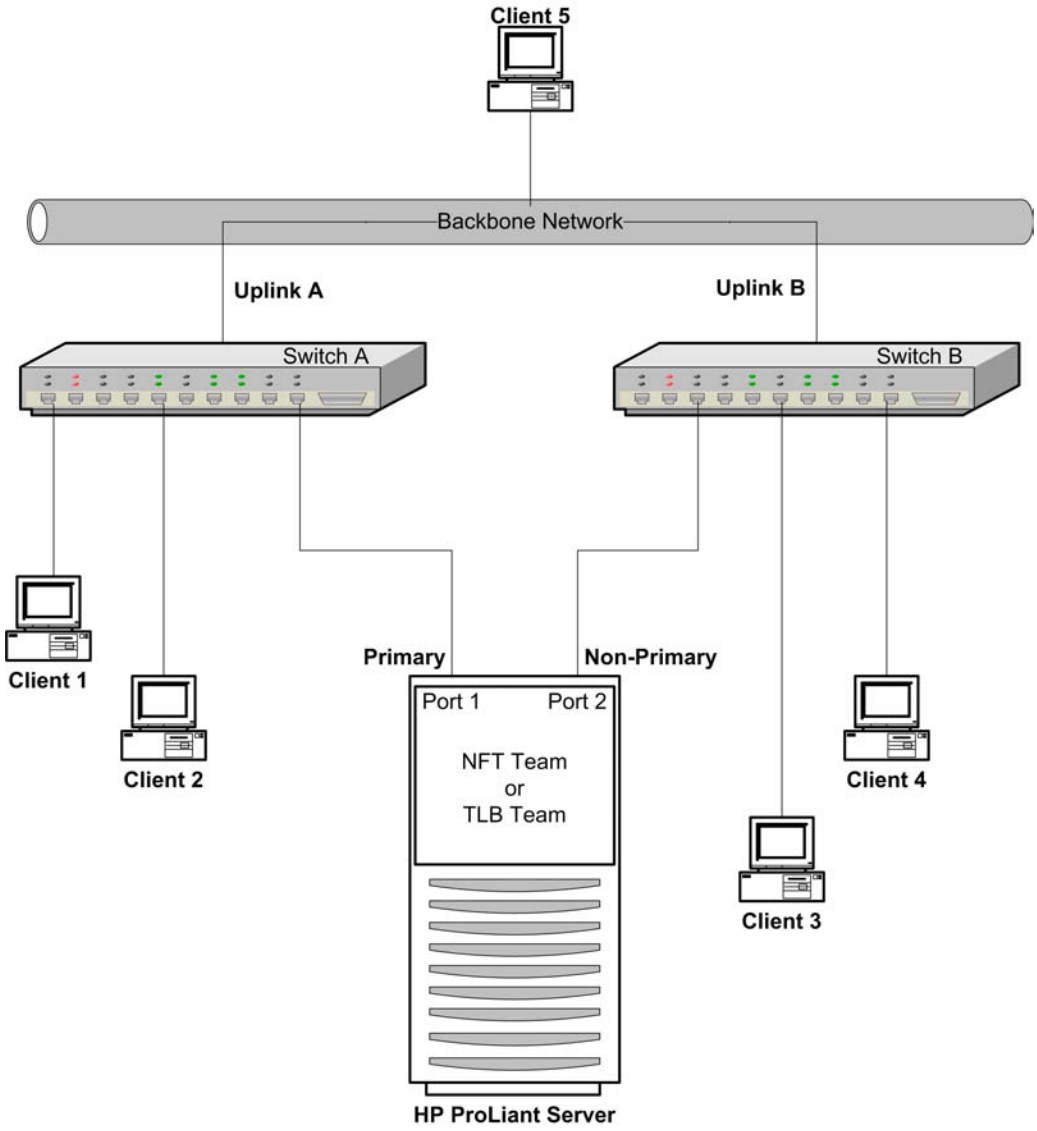
The HP Network Adapter Teaming supports the use of IGMP on HP Network Adapter Teams. Operating Systems supporting IGMP will send IGMP messages to upstream routers in order to receive certain IP multicast streams. Some switches support a feature called IGMP Snooping. This feature allows a switch to isolate IP multicast traffic to network devices that wish to receive it. Without IGMP Snooping, switches treat the IP Multicast traffic as broadcast traffic, forcing all devices in the broadcast domain to receive it. Because HP Network Adapter Teaming must manage more than one network adapter connection to a switch or switches, the Teaming driver will make sure that appropriate team members transmit the IGMP messages so that attached switches will register the appropriate Teamed ports for IP Multicast traffic.

network scenario considerations

NFT/TLB team split across switches

HP Network Adapter Teaming is designed for network adapter fault tolerance. However, some System Administrators or Network Administrators may desire to deploy HP Network Adapter Teaming with switch fault tolerance in mind (refer to figure 13). A thorough understanding of the heartbeat process and network adapter failover mechanisms is necessary before implementing NFT or TLB in this type of environment.

figure 13. NFT/TLB Team Split Across Switches



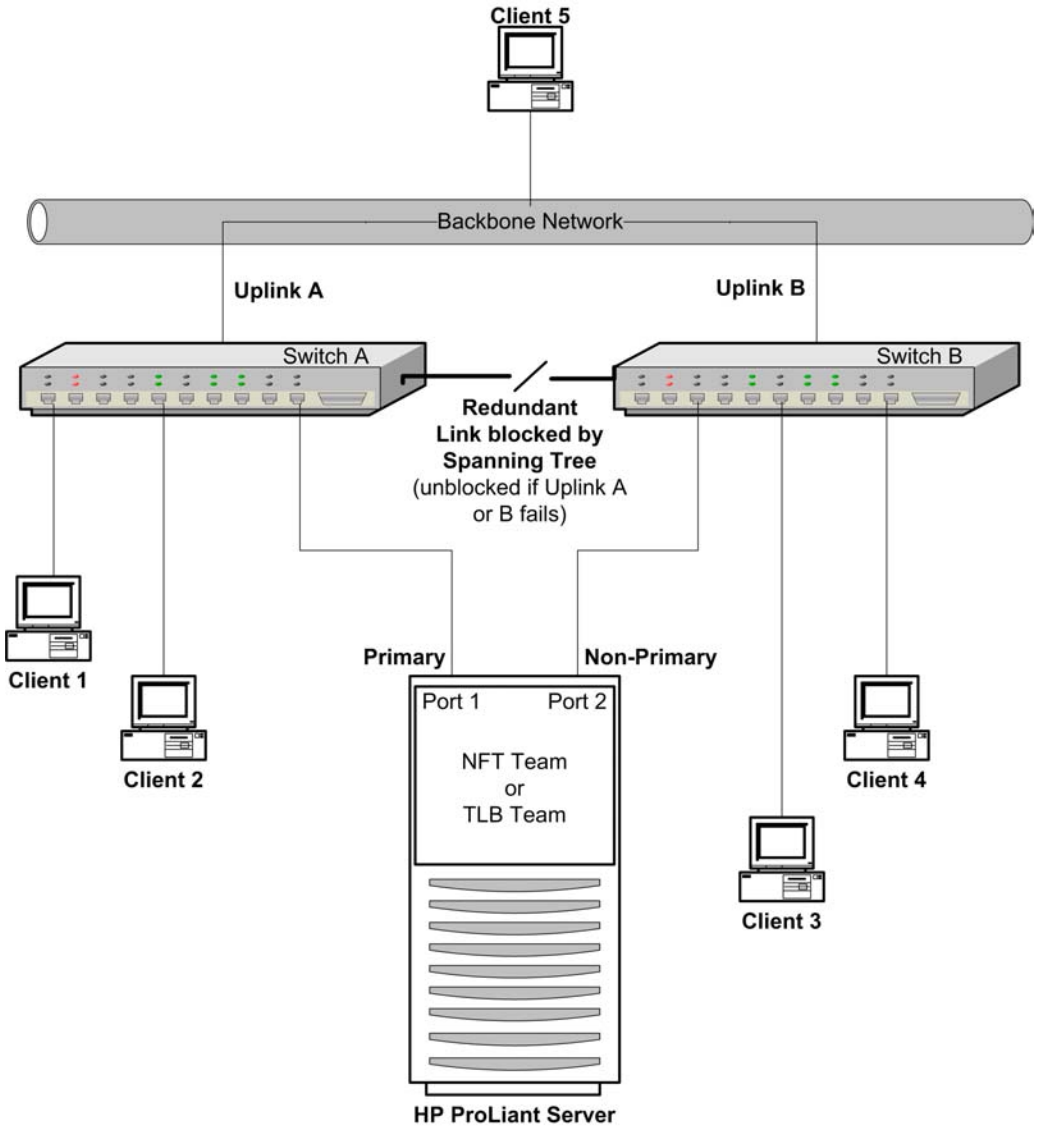
For example, in figure 13, the HP Network Adapter Team is attached to two switches, A and B. Both switches are connected to a core/backbone network via uplink A and uplink B, respectively. Network adapter 1 (Port 1) is the Primary Adapter in this NFT/TLB Team. Network adapter 2 (Port 2) is the Non-Primary Adapter. A typical Team member failure is caused by link loss on that Team member. If the link loss is on the Primary Adapter (Port 1), then the Teaming driver will failover to Port 2 if Port 2 is has link and all clients will maintain connectivity with the server.

Another type of failure can also occur in this scenario. Instead of one of the Team members losing link, the link from Switch A or B to the backbone network could be lost (Uplink A or Uplink B). This type of failure would isolate either switch from the backbone network and from the other switch. It would also cause the HP Network Adapter Team to enter Heartbeat failure mode because Heartbeat packets would not be successful transmitted between Port 1 and Port 2. This is caused by the Team members being isolated into two different networks that cannot communicate with each other (caused by an uplink failure on one of the switches). A problem can arise when the Teaming driver makes a failover decision at this point because the Teaming driver has no way to

determine where the failure occurred in the network. If the failure occurred at Uplink A, the best decision the Teaming driver could make is to failover to Port 2 and let Port 2 be the new Primary Adapter. This decision would give server access to the most clients, Clients 5, 4 and 3; and would isolate the fewest number of clients, Clients 1 and 2. However, because the Teaming driver cannot determine which switch uplink failed, the Teaming driver will initiate a failover to the Non-Primary Adapter every time a heartbeat failure occurs on the Primary Adapter as long as the Non-Primary Adapter is in an operational state. This means that if Uplink B were to fail, the Teaming driver would failover to Port 2 (the Non-Primary Adapter). This is not the optimal decision because it isolates more clients. Clients 3 and 4 have access but Clients 1, 2, and 5 are isolated.

One can avoid this issue by implementing redundant links between the switches and deploying a Layer 2 redundancy mechanism like Spanning Tree Protocol (STP) on the switched LAN (refer to figure 14). In this diagram, Switch A and Switch B have been redundantly connected, both directly and via the backbone network. In addition, STP has been implemented and has put the redundant link in standby mode (STP Block) until a failure occurs on the network that segments Switch A from Switch B. When the failure occurs, the link between the switches is made operational and the HP Network Adapter Team is able to communicate with all clients without isolating any.

figure 14. NFT/TLB Team Split Across Switches deployed with Spanning Tree



NFT (preferred primary) team split across switches

Using the same scenario as in figure 13, an NFT Team in Preferred Primary mode will behave differently than a TLB Team or NFT Team in Fail On Fault mode. In Preferred Primary mode, an NFT Team will use the adapter designated as Preferred Primary as long as the adapter is functioning. If Uplink A or Uplink B fails in the above network scenario, loss of Heartbeat frames may cause an adapter fail over and Port 2 will take over as the new Primary Adapter. However, Port 1 is elevated to an operational state if it receives any frame (unicast, multicast or broadcast). Therefore, even though a heartbeat failure just caused a failover to Port 2, the Teaming driver will initiate a failback to Port 1 as soon as it receives a frame. Once Port 1 becomes the Primary Adapter again, the Team may immediately enter a heartbeat failure state and failover to Port 2. Once again, as soon as Port 1 receives any type of frame, the Team will initiate a failback to Port 1. This can cause an endless loop of failovers. This behavior can be avoided by ensuring that all Team members remain in the same network. The best way to ensure this is to connect all Team members to the same switch. However, if switch redundancy is required and Team members are connected to more than one switch, it is HP's recommendation to implement some form of Layer 2 redundancy (e.g. Spanning

Tree with redundant links between switches) to recover from link failure between switches.

layer 3 routing of load balanced traffic

Special consideration should be given when choosing between the MAC Address-based and IP Address-based load balancing algorithms in an environment where the server and clients are separated by a Layer 3 device, such as a router. In such an environment, the server must communicate with the clients via the router (set as its default gateway). When communicating with the clients, the server sends all traffic to the router, which then sends the traffic to the clients. If MAC Address- based load balancing is selected, all traffic destined for clients is transmitted using the same network adapter in the load balancing Team and is not load balanced. This occurs because the server must use the router's MAC address as the Layer 2 address in every frame while it uses the client's IP address as the Layer 3 address in the same frame. Because the router's MAC address is used in every frame, the MAC address-based load-balancing algorithm chooses the same adapter for all traffic. Instead, choose the IP address-based load balancing algorithm, and load balancing will be based on the address of the clients (which varies) and not on the router (which is the same). In hybrid environments in which the server is communicating with clients on its own network, as well as clients on the other side of a router, HP recommends using IP-based load balancing (refer to "TLB and Layer 3 load balancing using IP address" and "SLB and Layer 3 Load Balancing using IP Address" for a more detailed explanation).

load balancing of non-IP traffic

When using HP Network Adapter Teaming in Transmit Load Balancing (TLB) mode, frames destined for clients will be load balanced over the Team members when transmitting. A unique Layer 2 address is used for each Team member but the same Layer 3 address is used by all Team members. Most protocols work perfectly in this environment; however, some non-IP (e.g. AppleTalk, IPX, SNA) clients require that the Layer 2 address remain constant for any host using a particular Layer 3 address. Because of this, TLB mode does not load balance non-IP traffic. Non-IP traffic is always transmitted out of the Primary Adapter and, in order to avoid any potential problems with other non-IP clients, always uses only one Layer 2 address.

If load balancing of non-IP traffic is required, Switch-assisted Load Balancing (SLB) should be used. With SLB, all traffic is transmitted from each Team member using the same Layer 2 and Layer 3 addresses. Because SLB uses the same Layer 2 address, non-IP traffic will be load balanced without affecting the clients.

teaming feature matrix

Teaming Type	NFT	TLB	SLB
Number of adapters supported per Team	2-8	2-8	2-8
Supports network adapter Fault Tolerance	X	X	X
Supports Transmit Load Balancing		X	X
Supports Receive Load Balancing			X
Requires a switch that supports a compatible form of load balancing. (i.e. requires configuration on the switch.)			X
Can connect a single Team to more than one switch for switch fault tolerance (all ports must be in the same broadcast domain)	X	X	Switch dependent
Can utilize heartbeats for network integrity checks	X	X	
Can team adapters that do not support a common speed	X		
Can team adapters operating at different speeds as long as the adapters support a common speed	X	X	X
Can team adapters of different media	X	X	X
Maximum theoretical transmit/receive throughput (in Mbps) with maximum number of 100 Mbps adapters	100/100	800/100	800/800
Maximum theoretical transmit/receive throughput (in Mbps) with maximum number of 1000 Mbps adapters	1000/1000	8000/1000	8000/8000
Load balances TCP/IP		X	X
Load balances non-IP traffic			X
Supports load balancing by destination IP address as an alternative to destination MAC address		X	X
All adapters within a Team utilize the same MAC address on the network			X
All adapters within a Team utilize the same IP address on the network	X	X	X

frequently asked questions (FAQ)

Q1 Why is traffic not being load balanced out of my server?
- or - Why is traffic not being load balanced during backups?

A1 Either TLB or SLB is required for load balancing of transmit traffic. NFT will not provide for any type of load balancing.

HP NIC Teaming uses either the MAC address or the IP address of the destination to make its load balancing decisions. If the destination always has the same MAC and IP address (another server or client), no load balancing will result. If the destination has the same MAC address but the IP address is different (e.g. several clients on the other side of a router), then HP NIC Teaming needs to be configured to load balance by IP address instead of by MAC address.

Q2 Why is traffic not being load balanced into my server?
- or - Why isn't traffic being load balanced during backups?

A2 A Team type of SLB and a supporting switch are needed to achieve receive load balancing.

Receive load balancing is determined by the switch connected to the HP Team in SLB mode. If the HP NIC Team is configured for SLB, then receive load balancing should occur if the switch is configured properly. Please consult the technical resources provided by the switch manufacturer.

Q3 I am trying to team two NICs for load balancing but it will not let me. Why?
- or - I have an HP NC series Fast Ethernet adapter and an HP NC series copper Gigabit Ethernet adapter in a TLB or SLB Team but I can not add an HP NC series fiber Gigabit Ethernet adapter to the Team. Why?

A3 To team multiple adapters together for load balancing, all adapters must be capable of supporting a common speed. For instance, any 10/100 adapter can be teamed for load balancing with a 100/1000 adapter because both adapter support a common speed. The adapters don't have to be operating at the common speed. Teaming a 1000 Fiber adapter with a 10/100 adapter is not supported for load balancing because the adapters don't support a common speed.

Q4 Can I team HP NC Series fiber Gigabit Ethernet adapters with HP NC Series copper Gigabit Ethernet adapters?

A4 Yes, any team type.

Q5 What is the difference between HP's Load Balancing Teams and Microsoft's Network Load Balancing (NLB) or Window's Load Balancing Service (WLBS) features?

A5 HP Teaming provides for fault tolerance and load balancing across network adapters and is aimed at server resilience. Microsoft's NLB and WLBS are for fault tolerance and load balancing across servers and are aimed at application resilience.

Q6 Can I use HP Network Adapter Teaming with Microsoft's NLB or WLBS features?

A6 Yes, however, some special configuration may be required. Support is limited to NLB and WLBS in Multicast mode only, not Unicast mode.

Q7 What effect will teaming have on the use of Cisco's Hot Swap Router Protocol (HSRP) or the IETF's Virtual Router Redundancy Protocol (VRRP) in my environment?

A7 None. HSRP and VRRP operate independently of teaming.

Q8 Can I use HP Network Adapter Teaming in conjunction with Cisco Local Director?

A8 Yes, teaming will work correctly with Cisco Local Director.

Q9 I want to force a Locally Administered MAC address on my HP Network Adapter Team. How should I do it?

A9 Open the HP Network Teaming and Configuration Properties GUI. Click on the appropriate Team in the GUI interface and select PROPERTIES. Go to the SETTINGS tab and type the LAA address in the "Team Network Address" field.

Q10 How do I uninstall HP Network Adapter Teaming?

A10 HP Network Adapter Teaming can be uninstalled by opening the properties page of any network interface under "Network and Dial-up Connections" (Microsoft UI). Select "HP Network Teaming and Configuration" and click the UNINSTALL button.

Q11 Is teaming multiple Fast Ethernet network adapters better than upgrading to a Gigabit Ethernet network adapter?

A11 Teaming multiple adapters does provide for additional fault tolerance over using a single adapter. However, the throughput of several fast Ethernet adapters will not usually be better than a single gigabit adapter.

Q12 Why does having Spanning Tree turned on for the HP Network Adapter Team switch ports cause a problem sometimes?

A12 When link is lost on a port that has Spanning Tree enabled on it, Spanning Tree will isolate the port from communicating with the rest of the network for a certain time period. This time period can sometimes exceed a minute. This isolation period can cause communication loss, heartbeat failures, and undesired teaming failovers under certain conditions.

Q13 Is HP Network Adapter Teaming an industry standard technology?

A13 The core of HP Network Adapter Teaming technology is an industry standard technology used for grouping network ports together for fault tolerance and load balancing. However, some of the special mechanisms that HP uses to enhance network adapter teaming are unique to HP Teaming technology.

Q14 Can I use third party/non-HP network adapters with HP Network Adapter Teaming?

A14 No, only HP branded network adapters may be used.

Q15 What does the Network Infrastructure Group need to do to help me deploy HP Network Adapter Teaming correctly?

A15 For all team types, the Network Infrastructure Group needs to know the following:

- The VLAN IDs used on a particular Team.
- Which group of ports constitutes a Team. For each group, the following must be done:
    - All ports configured for the same VLANs, if any.
    - All ports must belong to the same broadcast domain/s.

For SLB teams, they also need to know the following:

- Which group of ports constitutes a Team. For each group, the following must be done:
    - All ports in each team must be configured as a single port trunk/Multilink Trunk/EtherChannel group.

Q16 Can I use HP Network Adapter Teaming in conjunction with Microsoft Cluster Server?

A16 Yes, however, Microsoft may request that teaming be disabled before technical assistance is provided.

Q17 My switch is setup for Switch-assisted Load Balancing (port Trunking) but my network adapters are not. Why am I having communication problems?

A17 The switch assumes that all ports are capable of receiving on the same MAC address/es and will randomly transmit frames for any of the NICs down any of the links for any of the NICs. If the NICs aren't configured for SLB Teaming, they will drop any frame meant for another NIC. Because of this, the switch should only be configured for port trunking after the SLB Team has been created.

If a single server with multiple NICs is connected to a switch configured port trunking and PXE is being used to deploy the server, communication problems will most likely prevent PXE from completing a successful install on the server. To avoid such problems, disconnect all NICs except for the NIC providing PXE support or remove the port trunking configuration on the switch.

Q18 What is the maximum number of network adapters that can be in a single Team?

A18 8 adapters

Q19 What is the maximum number of Teams that can be configured on a single HP Server?

A19 16 Teams

Q20 What is the maximum number of VLANs that can be configured on a single network adapter or a single Team?

A20 64 VLANs

Q21 Why does my Team loose connectivity for the first 30 to 90 seconds after the Preferred Primary port's link is restored?

A21 This may be caused by Spanning Tree. Disable Spanning Tree on the port or enable the Spanning Tree bypass feature if available (e.g., PortFast, bypass)

Q22 Is there a particular Windows 2000 Service Pack level that is required for HP Network Adapter Teaming to work correctly?

A22 No, not for Windows 2000. Windows NT4 requires Service Pack 5.

Q23 If I make an Altiris image of a server with a Team, can I deploy that image onto other servers?

A23 Yes, but the Team's MAC address registry entry will have to be restored individually on all servers the image was deployed on.

Q24 Why do I still see heartbeat frames after disabling them?

A24 Even when heartbeats are disabled, an HP Team must make sure that the Team's MAC address is known by the switch to prevent flooding of traffic being sent to the Team. To achieve this, the Team needs to transmit a frame every so often. The frame used for this purpose is a heartbeat frame. Heartbeat frames are also used during a failover to notify the switch of MAC address changes on the teamed adapters.

Q25 When should I increase the heartbeat timers for a Team?

A25 The heartbeat timers should be increase when heartbeat failures are caused by latency in the network infrastructure.

Q26 Is Unattended Installation of HP Network Adapter Teaming supported?

A26 Yes.

Q27 I need NIC redundancy, switch redundancy and load balancing. What Teaming type should I use?

A27 When NIC redundancy, switch redundancy and load balancing are all required, the only option is Transmit Load Balancing (TLB). TLB allows for an HP NIC Team to be connected to more than one switch (switch redundancy), if a NIC Fails then traffic is transmitted/received on another Team NIC (NIC redundancy) and all NICs in the Team share the load of transmitting traffic onto the network (load balancing). The only caveat is that receive load balancing is not supported.

An alternative is to use switch vendor redundancy mechanisms to make a single switch highly redundant. For example, Cisco provides an option called High Availability on some switches. This option allows a Cisco switch to have redundant Supervisor modules. An HP customer using a Cisco switch with redundant Supervisor modules and redundant power supplies can create an SLB team of several adapters, connect the team to two modules inside the same Cisco switch and enable High Availability. This provides transmit/receive load balancing, network adapter fault tolerance, switch power supply fault tolerance, Supervisor fault tolerance, and switch module fault tolerance.

Q28 What load balancing methods is SLB (Switch-assisted Load Balancing compatible with?

A28 FEC/GEC, Load Sharing, MLT, IEEE 802.3ad, etc.

Q29 Can I connect Teamed adapters to more than one switch?

A29 Yes, with NFT and TLB teams only. Also, all switch ports that have Team members connected to them must belong to the same broadcast domain. This means that the broadcast domain must span between the switches.

Q30 Who in HP is responsible for development and support of HP ProLiant Network Adapter Teaming?

A30 The Austin Development Group (ADG) provides all development and Level 3 support for HP Network Adapter Teaming technology, as well as all ProLiant networking products (e.g., BL switches, ProLiant network adapters). ADG is one of many groups that constitute HP's Industry Standard Servers division.

Q31 What is the limit for the number of Teams I can create in one server and what is the limit for the number of network adapters that I can include in one Team?

A31 The theoretical limit is 16 Teams of 8 network adapter ports per server. This is defined as a "theoretical" limit because not all servers will allow the installation of enough network adapters to create 16 Teams of 8 network adapter ports.

Q32 How do I upgrade HP Network Adapter Teaming drivers?

A32 HP provides the HP Network Adapter Teaming driver with an installation utility. Download the appropriate HP Network Adapter Teaming driver from http://h18000.www1.hp.com/support/files/networking/nics/index.html

glossary

ALB	Adaptive Load Balancing. Refer to Transmit Load Balancing (TLB).
ARP	Address Resolution Protocol. A protocol used to determine a MAC address from an IP address.
BIA	Burned In Address. The Layer 2 address that is permanently assigned to a piece of hardware by the vendor. Referred to as a MAC address.
Broadcast domain	Set of all devices that will receive Layer 2 broadcast frames originating from any device within the set. Broadcast domains are typically bounded by routers because routers do not typically forward Layer 2 broadcast frames.
Byte	Eight bits
Collision domain	A single Ethernet network segment in which there will be a collision if two computers attached to the system transmit simultaneously.
FEC	Fast EtherChannel. A method of load balancing that both transmits and receives traffic across multiple Fast Ethernet connections (100 Mbps) between two devices. Developed by Cisco Systems. Refer to SLB.
GEC	Gigabit EtherChannel. A method of load balancing that both transmits and receives traffic across multiple Gigabit Ethernet (1000 Mbps) connections between two devices. Developed by Cisco Systems. Refer to SLB.
GUI	Graphical User Interface.
IEEE	Institute of Electrical and Electronics Engineers. A standards body for, among other things, network communications and protocols.
LAA	Locally Administered Address. A temporary Layer 2 address that is manually assigned to a piece of hardware.
Layer 2	The second layer of the OSI model, the Data Link Layer. A Layer 2 address is the same as a MAC (Media Access Control) address.
Layer 3	The third layer of the OSI model, the Network Layer. A Layer 3 address refers to a protocol address such as an IP or IPX address.

MAC address	Media Access Control address. With Ethernet, this refers to the 6 byte (48 bit) address that is unique to every Ethernet device
Multi-homed	A device that is redundantly attached to a network or networks.
NDIS	Network Driver Interface Specification. Simplified, it is the interface between a network adapter and Microsoft's protocol stack.
NFT	Network Fault Tolerance. A Team of network adapters that transmits and receives on only one adapter with all other adapters in standby.
OSI Model	Open Systems Interconnect Model. The seven layer model developed by the International Standards Organization that outlines the mechanisms used by networked devices to communicate with each other.
PING	A type of packet used to validate connectivity with another network device. The packet asks another network device to respond to validate connectivity, a kind of "echo." PING packets for IP are accomplished using the ICMP protocol.
SLB	Switch-assisted Load Balancing. Also known as FEC/GEC. A Team of network adapters that load balances transmits and receives on all adapters.
STA	Spanning Tree Algorithm (IEEE 802.1D)
Switch MAC Table	A list of MAC addresses and associated ports that are used by a switch to transfer frames between attached devices. Also referred to as a CAM Table.
LLC	Logical Link Control
TLB	Transmit Load Balancing. Also known as Adaptive Load Balancing (ALB). A Team of network adapters that receives on one adapter but load balances transmitted IP traffic on all adapters. Other protocol traffic is transmitted by a single adapter.

technical support

To contact an HP Technical Support engineer regarding issues with HP Network Adapter Teaming, please call 1-800-652-6672. To speak with the appropriate support group, select the call routing options for HP Server Networking, or HP ProLiant Networking.

For online assistance, HP's ProLiant Web based Support Forum is located at

http://forums.itrc.hp.com/cm/CategoryHome/1,,264,00.html

For driver updates to HP Network Adapter Teaming, please visit HP's Network Adapter Driver site at

http://h18007.www1.hp.com/support/files/networking/nics/index.html

5981-8481EN