

Using NIC Teaming

to Achieve High Availability on Linux Platforms

Network interface card (NIC) teaming is one method for providing high availability and fault tolerance in servers. This article examines how Broadcom[®] Advanced Server Program (BASP) and Intel[®] Advanced Network Services (ANS) can be used to team NICs, and describes a study showing how NIC teaming was configured on Linux[®]-based nodes of an Oracle9i[™] Real Application Clusters Certified Configuration for Dell.

BY AMIT BHUTANI AND ZAFAR MAHMOOD

Maintaining high server availability is vital in enterprise IT environments. A significant contributor to that availability is the network connection to the server. Network interface card (NIC) teaming can help ensure such availability and provide other benefits to improve network performance.

NIC teaming combines two or more physical NICs into a single logical NIC, to which an administrator can then assign an IP address. If one of the physical NICs fails, the IP address remains accessible because it is bound to the logical NIC rather than to a single physical NIC. By installing multiple network cards on a server and load balancing traffic among them, administrators can gain fault tolerance and can achieve greater throughput, at least in the case of Fast Ethernet networks.

Tools for NIC teaming on the Linux[®] operating system (OS) include the Broadcom[®] Advanced Server Program (BASP) and the Intel[®] Advanced Network Services (ANS) software. This article explains how these two tools load balance network traffic and provide failover capabilities, and then describes a detailed study examining the configuration process for NIC teaming using BASP on Linux-based nodes of an Oracle9i[™] Real Application Clusters Certified Configuration for Dell.

Using NIC teaming for load balancing and failover

NIC teaming provides traffic load balancing and redundant NIC operation if a network connection fails. When multiple NICs are installed in the same server, they can be combined into teams. Each team must include at least two members, but can support up to eight members. The number of NICs that are installed limits the number of teams.

In a team of two or more NICs, the secondary NIC fills the primary, or active, role if the primary NIC fails. Failover will not occur unless the primary has been specified. When no priority has been specified, the NIC with the highest supported speed is chosen as the primary. If more than one NIC runs at the highest supported speed, the last of these NICs added to the team becomes the primary. A team can have as many as eight primary NICs.

If any team member fails because of problems in the NIC, cable, switch port, or switch (if the teamed NICs are attached to separate switches, as shown in Figure 1), the load distribution is failed over (reevaluated and reassigned) to the remaining team members. If all the active NICs are down, the standby NIC—if one was configured for the team—becomes active.

Examining Intel ANS and BASP teaming capabilities

To support NIC teaming, BASP and Intel ANS include intermediate drivers specific to each company's network controllers and base drivers. Intermediate drivers serve as a wrapper around one or more base drivers, providing an interface between the base driver and the network protocol stack. By doing so, the intermediate driver gains control over which packets are sent to which physical interface as well as control over other properties essential to teaming. Intermediate drivers also allow for the inclusion of other brands of NICs in their teaming configurations. Support for heterogeneous NICs—a feature known as Multi-Vendor Teaming (MVT) in the case of Intel ANS software—makes configuration very flexible. BASP supports this feature as well.

When forming a NIC team using the BASP or Intel ANS driver, at least one NIC must be from Broadcom (if using BASP) or Intel (if using Intel ANS). However, a single card cannot be used in more than one team at a time. In other words, a token NIC per module is required for teaming to work. For example, if implementing an Intel ANS team, at least one NIC should be from Intel.

Although the two intermediate drivers can exist concurrently on a system, using both BASP and Intel ANS software modules at the same time is not recommended, even if they are each used on different physical network controllers, because results can be unpredictable.

Intel ANS teaming modes

The Intel ANS teaming modes include the following:

Adapter Fault Tolerance (AFT). This mode provides automatic redundancy for an adapter and works with a hub or switch. Speed and duplex capabilities and settings can be mixed. AFT supports two to eight adapters per team. Only one active team member transmits and receives traffic. If this primary connection (cable, adapter, or port) fails, a secondary, or backup, adapter takes over. After a failover, if the connection to the primary adapter is restored, control passes automatically back to that primary adapter. AFT is the default mode when a team is created; this mode does not provide load balancing.

Adaptive Load Balancing (ALB). This mode provides load balancing of the transmission traffic and fault tolerance for the NICs. The transmission load is shared among all adapters in the team. This teaming mode also supports mixed-speed and mixed-duplex settings among team members, and works with any switch.

Cisco Fast EtherChannel trunking technology or Intel Link Aggregation. This mode provides increased transmission and reception throughput in a team of two to eight 10/100 adapters, and also includes AFT and load balancing (only for routed protocols). However, this mode requires a switch that supports Intel Link

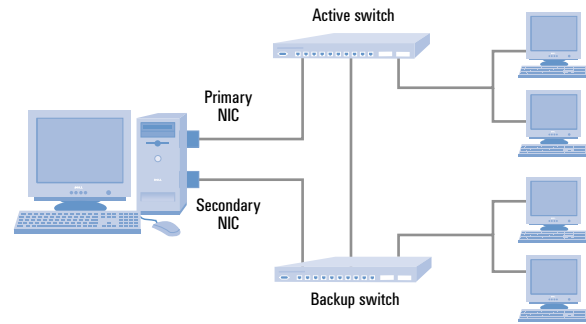


Figure 1. Topology for switch failover using NIC teaming

Aggregation, Cisco® Fast EtherChannel®, or static IEEE® 802.3ad capability. Also, speed and duplex settings on all team members must match, and Spanning Tree Protocol must be turned off.

Cisco Gigabit EtherChannel static link aggregation. This mode is the Gigabit Ethernet¹ extension of the Fast EtherChannel mode.

IEEE 802.3ad dynamic link aggregation. This mode creates one or more teams using dynamic link aggregation with mixed-speed adapters. Like the static link aggregation modes, dynamic 802.3ad teams increase transmission and reception throughput and provide fault tolerance. Link Aggregation Control Protocol (LACP) provides dynamic link aggregation capability. This protocol can automatically detect the presence and capabilities of other aggregation-capable devices; that is, LACP enables administrators to specify which links in a system can be aggregated.

If the link partner is not correctly configured for 802.3ad link aggregation, errors are detected and noted. When using this mode, administrators configure all team members to receive packets for the same Media Access Control (MAC) address. The Intel ANS driver determines the outbound load-balancing scheme, while the team's link partner determines the load-balancing scheme for inbound packets. This mode requires a switch that fully supports the IEEE 802.3ad standard.

BASP teaming modes

BASP teaming modes include the following:

Smart Load Balancing. This proprietary Broadcom technology provides fault tolerance and load balancing based on IP flow. This feature can balance IP traffic across as many as eight team members for both outbound and inbound traffic. In this mode, all adapters in the team have separate MAC addresses. Smart Load Balancing™ (SLB) provides automatic fault detection and dynamic failover to another team member or to a hot-standby member, and works with any switch or hub.

IEEE 802.3ad link aggregation. This mode supports link aggregation through static and dynamic configurations and conforms to the IEEE 802.3ad specification. Similar to the Intel implementation of this

¹Gigabit Ethernet indicates compliance with IEEE 802.3ab or IEEE 802.3z and does not connote speeds of 1 Gbps.

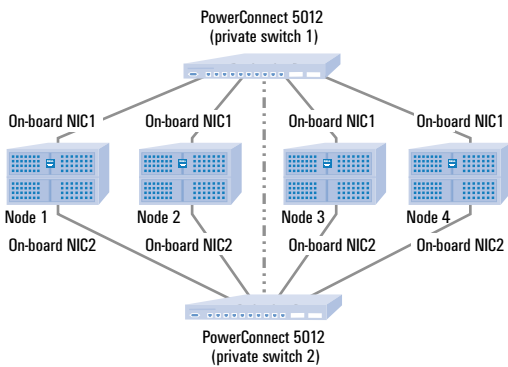


Figure 2. Oracle cluster topology

standard, the BASP driver determines the outbound load-balancing scheme; the team's link partner determines the load-balancing scheme for inbound packets.

Generic link aggregation (trunking). This mode is very similar to 802.3ad in that all team members are configured to receive packets for the same MAC address. This mode supports a variety of environments where the link partners of the NIC are statically configured to support a proprietary trunking mechanism. For instance, this mode could be used to support Lucent® OpenTrunk or Cisco Fast EtherChannel. Basically, this mode is a lighter version of the 802.3ad link aggregation because it does not use a formalized protocol for link aggregation; it is protocol-independent. Trunking supports load balancing, thereby boosting bandwidth, and also provides failover for both outbound and inbound traffic.

Using BASP on an Oracle Certified Configuration for Dell

Dell examined the use of NIC teaming to offer fault tolerance at the NIC and switch levels, which should provide no single point of failure within the cluster. The Dell™ team performed tests using both Intel ANS and BASP teaming software in an Oracle9i Real Application Clusters Certified Configuration. To avoid redundancy, this article discusses the detailed installation and configuration scenario only for BASP; the Intel ANS scenario is very similar, differing mainly in configuration file and command syntax.

The study used four Dell PowerEdge™ 6650 servers as the cluster nodes, two Dell PowerConnect™ 5012 gigabit switches, and two on-board Broadcom BCM5700 gigabit NICs per node. Figure 2 shows the hardware connections for the multiswitch, multi-NIC, high-availability configuration used in the study.

The nodes ran the Red Hat® Linux Advanced Server 2.1 OS, kernel version 2.4-9.e3, and were installed with Oracle9i Real Application Clusters, version 2.1. The nodes used the BASP teaming software (installed from the `basplnx-3.0.9-1.i386.rpm` package) and the BCM5700 driver (installed from the `bcm5700-2.2.22-1.i386.rpm` package).

Manual installation and configuration

Administrators can configure the BASP software manually, which is recommended for experienced users, or with the BASP configuration tool. The BASP tool, `baspcfg`, is a command-line tool for configuring BASP teams, adding and removing NICs, and adding and removing virtual devices. This tool can be used in custom initialization scripts.

For this study, the Dell team installed and configured the BASP software manually. Before starting this procedure, administrators should be logged in as root on all the nodes where teaming is to be implemented. Otherwise, some of the commands may result in errors.

1. Complete the cabling as shown in Figure 2.
2. Before installing the intermediate driver, bring down the NICs that will be teamed and unload any existing linked Broadcom driver modules:


```
% ifdown eth0
% ifdown eth1
% rmmod bcm5700
```
3. On each node, install the latest BCM5700 series NIC driver for Red Hat Linux Advanced Server 2.1. The study used the `bcm5700-2.2.22-1.i386.rpm` package:


```
% rpm -i bcm5700-2.2.22-1.i386.rpm
```
4. Load the driver module into the kernel:


```
% insmod bcm5700
```
5. Uninstall any previous versions of BASP:


```
% rpm -e basplnx
```
6. Install the latest BASP package that is qualified for the specific operating system version and NIC version required by each node. The study used the `basplnx-3.0.9-1.i386.rpm` package:


```
% rpm -i basplnx-3.0.9-1.i386.rpm
```
7. Load the BASP module into the kernel:


```
% insmod basp
```
8. Before configuring the team, first configure the physical NICs that will take part in the team. For example, if configuring the first NIC `eth0`, create a file named `icfg-eth0` in the `/etc/sysconfig/network-scripts` directory. The file should resemble the following:

```
DEVICE=eth0
BOOTPROTO=static
ONBOOT=no
```

Do not assign IP addresses, subnet masks, or network IDs to individual NICs. Instead, assign this information to the virtual adapter (see step 10).

9. Restart network services to verify that the new drivers are loaded:

```
% service network restart
```

10. On each node, create a team configuration file called `team-private` in the `/etc/basp` directory. The following is a sample team configuration file:

```
TEAM_ID=0
TEAM_TYPE=0
TEAM_NAME=team-private

# 1st physical interface in the team
TEAM_PAO_NAME=eth0
TEAM_PAO_ROLE=0

# 2nd physical interface in the team
TEAM_PAI_NAME=eth1
TEAM_PAI_ROLE=0

# 1st virtual interface in the team
TEAM_VAO_NAME=private
TEAM_VAO_VLAN=0
```

This file contains the following parameters:

- `TEAM_ID`: A number that uniquely identifies a team
- `TEAM_TYPE`: 0 = SLB; 1 = generic trunking, Fast EtherChannel, or Gigabit EtherChannel; 2 = 802.3ad
- `TEAM_NAME`: ASCII name of the team
- `TEAM_PAx_NAME`: ASCII name of the physical interface `x`, where `x` can be 0 to 7
- `TEAM_PAx_ROLE`: Role of the physical interface `x`, where 0 = primary and 1 = hot standby; the parameter value must be 0 for a generic trunking, Fast EtherChannel, or Gigabit EtherChannel team
- `TEAM_VAx_NAME`: ASCII name of the virtual interface `x`, where `x` can be 0 to 63
- `TEAM_VAx_VLAN`: 802.1Q VLAN ID of the virtual interface `x`; the valid VLAN ID can be 0 to 4094 (for an untagged virtual interface—that is, one without VLAN enabled—set the ID to 0)

The sample file sets the roles of both NICs to primary; thus, the NICs provide both load balancing and failover.

11. On all nodes, manually configure the virtual interface network script file in `/etc/sysconfig/network-scripts` because this file is not automatically generated. The script file for the virtual interface named “private” in this study was saved with the filename `ifcfg-private` and contained the following:

```
DEVICE=private
BOOTPROTO=static
IPADDR=192.168.0.1
NETMASK=255.255.255.0
NETWORK=192.168.0.0
BROADCAST=192.168.0.255
ONBOOT=yes
```

Test	Status	Comments
Unplugging the network cables from the first active interface	PASS	Instantaneous failover
Unplugging the network cables from the second active interface	PASS	Instantaneous failover
Unplugging the power cable from one of the two switches for failover	PASS	Delay of 15 to 45 seconds

Figure 3. Testing failover capability for NIC teaming configuration

12. Edit the `/etc/sysconfig/network` file to ensure the following lines are present:

```
NETWORKING=yes
HOSTNAME=racnode1
```

13. After creating the virtual interface network configuration file and making the changes listed, ensure that the BASP service is on at boot time and has been started on all nodes:


```
% chkconfig --level 35 basp on
% service basp start
```

Failover tests

Having completed the manual configuration steps, the team conducted several failover tests (see Figure 3). In the third test, “Unplugging the power cable from one of the two switches for failover,” if the team manually refreshed the Address Resolution Protocol (ARP) cache using a simple script in the background, the failover appeared to occur faster. However, because using such a script could have other unknown side effects, implementing one to accomplish faster failovers is not generally recommended.

The study showed that, from a network perspective, eliminating a single point of failure within the Oracle® cluster was indeed possible. Both BASP and Intel ANS proved to be simple to use, robust, and had failover times acceptable for the study.

Increasing network reliability through NIC teaming

High availability of servers is becoming much more critical as the expectation of uptime keeps rising. Because NIC teaming can help improve return on investment (ROI) for server purchases, it is a cost-effective method to quickly and easily increase network reliability and scalability. Most importantly, NIC teaming can ensure non-interruption of critical services to end users. BASP and Intel ANS give administrators helpful tools for creating NIC teaming. 

Amit Bhutani (amit_bhutani@dell.com) is a software engineer in the networking area within the Linux Development Group of the Dell Product Group. Amit has a master’s degree in Computer Engineering with a specialization in Software Engineering from West Virginia University.

Zafar Mahmood (zafar_mahmood@dell.com) is a software engineer in the Dell Product Group, focusing on Oracle® Real Application Clusters. Zafar has a master’s degree in Electrical Engineering with a specialization in Computer Communications from the City University of New York.