Setup team with broadcom nic assignment

Broadcom Gigabit Ethernet Teaming Services 4/20/2004 Version 1. 1 Dell Inc. One Dell Way Round Rock, Texas 78681 Table of Contents

Intermediate Driver Event Log Messages51 List of Graphs Graph 1. Teaming Performance Scalability18 Graph 2.

Backup Performance With no NIC Teaming38 Graph 3. Backup Performance40 THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND. Executive Summary This white paper describes the technology and implementation considerations when working with the network teaming services offered by the Broadcom software shipped with Dell’s servers and storage products. The goal of the Broadcom teaming services is to provide fault tolerance and link aggregation across a team of two or more adapters.

The information in this document will assist IT professionals during the deployment and troubleshooting of server applications that require network fault tolerance and load balancing. 1 Key Definitions and Acronyms | Term | Definition | | BACS | Broadcom Advanced Configuration Suite Configuration GUI | | BASP Broadcom Advanced Server Program Intermediate driver | | Smart Load Balancing (SLB) w/ one Standby port | Switch independent Failover team. Primary member handles all incoming | | | and outgoing traffic while the standby adapter is idle until a failover | | | event (for example, loss of link occurs).

Intermediate driver manages | | | incoming/outgoing traffic. | | Smart Load Balancing (SLB) | Switch independent Load balancing and failover team – Intermediate | | | driver manages outgoing/incoming traffic. | | LACP | Link

Aggregation Control Protocol | | Generic Trunking (FEC/GEC)- 802. ad Draft Static | Switch dependent load balancing and failover- Intermediate driver | | | manages outgoing traffic/ Switch manages incoming traffic | | Link Aggregation (802. 3ad) | Switch dependent load balancing and failover with LACP-Intermediate | | | driver manages outgoing traffic/Switch manages incoming traffic. | NDIS | Network Driver Interface Specification | | PXE | Pre-Execution Environment | | ARP | Address Resolution Protocol | | RAID | Redundant Array of Inexpensive Disks | | MAC | Media Access Control | | DNS | Domain Name Service | WINS | Windows Name Service | | TCP | Transmission Control Protocol | | UDP | User Datagram Protocol | | IP | Internet Protocol | | ICMP | Internet Control Message Protocol | | IGMP | Internet Group Management Protocol | | G-ARP | Gratuitous Address Resolution Protocol | | HSRP | Hot Standby Router Protocol | | LOM | LAN On Motherboard | | NLB | Network Load Balancing (from Microsoft) | | WLBS | Windows Load Balancing Services | Table 1. Glossary of Terms 2 Teaming Concepts The concept of grouping multiple physical devices to provide fault tolerance and load balancing is not new. It has been around for years. Storage devices use RAID technology to group individual hard drives. Switch ports can be grouped together using technologies such as Cisco Gigabit EtherChannel, IEEE 802. 3ad Link Aggregation, Bay Network Multilink Trunking, and Extreme Network Load Sharing. Network interfaces on Dell servers can be grouped together into a team of physical ports called a virtual adapter. 1 Network Addressing

To understand how teaming works, it is important to understand how node communications work in an Ethernet network. This paper assumes that the reader is familiar with the basics of IP and Ethernet network communications. The following information provides a high level overview of the concepts of network addressing used in an Ethernet network. Every Ethernet network interface in a host platform such as a server requires a globally unique Layer 2 address and at least one globally unique Layer 3 address. Layer 2 is the Data Link Layer, and Layer 3 is the Network layer as defined in the OSI model. The Layer 2 address is assigned to the hardware and is often referred to as the MAC address or physical address.

This address is pre-programmed at the factory and stored in NVRAM on a network interface card or on the system motherboard for an embedded LAN interface. The layer 3 addresses are referred to as the protocol or logical address assigned to the software stack. IP and IPX are examples of Layer 3 protocols. In addition, Layer 4 (Transport Layer) uses port numbers for each network upper level protocol such as Telnet or FTP. These port numbers are used to differentiate traffic flows across applications. Layer 4 protocols such as TCP or UDP are most commonly used in today’s networks. The combination of the IP address and the TCP port number is called a socket. Ethernet devices communicate with other Ethernet devices using the MAC address, not the IP address.

However, most applications work with a host name that is translated to an IP address by a Naming Service such as WINS and DNS. Therefore, a method of identifying the MAC address assigned to the IP address is required. The Address Resolution Protocol for an IP network provides this mechanism. For IPX, the MAC address is part of the network address and ARP is not required. ARP is implemented using an ARP Request and ARP Reply frame. ARP

Requests are typically sent to a broadcast address while the ARP Reply is typically sent as unicast traffic. A unicast address corresponds to a single MAC address or a single IP address. A broadcast address is sent to all devices on a network. 2 Teaming and Network Addresses

A team of adapters function as a single virtual network interface and does not appear any different to other network devices than a non-teamed adapter. A virtual network adapter advertises a single layer 2 and one or more layer 3 addresses. When the teaming driver initializes, it selects one MAC address from one of the physical adapters that make up the team to be the Team MAC address. This address is typically taken from the first adapter that gets initialized by the driver. When the server hosting the team receives an ARP Request, it will select one MAC address from among the physical adapters in the team to use as the source MAC address in the ARP Reply.

In Windows operating systems, the IPCONFIG /all command shows the IP and MAC address of the virtual adapter and not the individual physical adapters. The protocol IP address is assigned to the virtual network interface and not to the individual physical adapters. For switch independent teaming modes, all physical adapters that make up a virtual adapter must use the unique MAC address assigned to them when transmitting data. That is, the frames that are sent by each of the physical adapters in the team must use a unique MAC address to be IEEE compliant. It is important to note that ARP cache entries are not learned from received frames, but only from ARP Requests and ARP Replies. 3 Description of teaming modes

There are three methods for classifying the supported teaming modes: one is based on whether the switch port configuration must also match the NIC teaming mode; the second is based on the functionality of the team, whether it supports load balancing and failover or just failover; and the third is based on whether the Link Aggregation Control Protocol is used or not. The following table shows a summary of the team modes and their classification. | Broadcom Teaming Selections | Switch Dependent- switch must | Link Aggregation Control | Load Balancing | Failover | | | support specific teaming mode Protocol support required on | | | | | | the switch | | | | | | | | | | Smart Load Balancing and Failover (SLB) | | | | | |(with 2-8 load balance members) | | | | | | Smart Load Balancing and Failover (SLB) | | | | | |(with 1 load balance member and 1 or more| | | | | | standby member) | | | | | | Link Aggregation (802. 3ad) | | | | | | Generic Trunking (FEC/GEC)/802. 3ad-Draft | | | | | | Static | | | | | Table 2.

Teaming Mode Selections Offered by Broadcom 1 Smart Load Balancing (SLB) Smart Load Balancing provides both load balancing and failover when configured for Load Balancing, and only failover when configured for fault tolerance. It works with any Ethernet switch and requires no trunking configuration on the switch. The team advertises multiple MAC addresses and one or more IP addresses (when using secondary IP addresses). The team MAC address is selected from the list of load balancing members. When the server receives an ARP Request, the software-networking stack will always send an ARP Reply with the team MAC address. To begin the load balancing process, he teaming driver will modify this ARP Reply by changing the source MAC address to match one of the physical adapters. Smart Load

Balancing enables both transmit and receive load balancing based on the Layer 3/Layer 4 IP address and TCP/UDP port number. In other words, the load balancing is not done at a byte or frame level but on a TCP/UDP session basis. This methodology is required to maintain in-order delivery of frames that belong to the same socket conversation. Load balancing is supported on 2-8 ports. These ports can include any combination of add-in adapters and LAN-on-Motherboard (LOM) devices. Transmit load balancing is achieved by creating a hashing table using the source and destination IP addresses and TCP/UDP port numbers.

The same combination of source and destination IP addresses and TCP/UDP port numbers will generally yield the same hash index and therefore point to the same port in the team. When a port is selected to carry all the frames of a given socket, the unique MAC address of the physical adapter is included in the frame, and not the team MAC address. This is required to comply with the IEEE 802. 3 standard. If two adapters transmit using the same MAC address, then a duplicate MAC address situation would occur that the switch could not handle. Receive Load Balancing is achieved through an intermediate driver by sending Gratuitous ARPs on a client by client basis using the unicast address of each client as the destination address of the ARP Request (also known as a Directed ARP).

This is considered client load balancing and not traffic load balancing. When the intermediate driver detects a significant load imbalance between the physical adapters in an SLB team, it will generate G-ARPs in an effort to redistribute incoming frames. The intermediate driver (BASP) does not answer ARP Requests; only the software protocol stack provides the required

ARP Reply. It is important to understand that receive load balancing is a function of the number of clients that are connecting to the server via the team interface. SLB Receive Load Balancing attempts to load balance incoming traffic for client machines across physical ports in the team.

It uses a modified Gratuitous ARP to advertise a different MAC address for the team IP Address in the sender physical and protocol address. This G-ARP is unicast with the MAC and IP Address of a client machine in the target physical and protocol address respectively. This causes the target client to update its ARP cache with a new MAC address map to the team IP address. G-ARPs are not broadcast because this would cause all clients to send their traffic to the same port. As a result, the benefits achieved through client load balancing would be eliminated, and could cause out of order frame delivery. This receive load balancing scheme works as long as all clients and the teamed server are on the same subnet or broadcast domain.

When the clients and the server are on different subnets, and incoming traffic has to traverse a router, the received traffic destined for the server is not load balanced. The physical adapter that the intermediate driver has selected to carry the IP flow will carry all of the traffic. When the router needs to send a frame to the team IP address, it will broadcast an ARP Request (if not in the ARP cache). The server software stack will generate an ARP Reply with the team MAC address, but the intermediate driver will modify the ARP Reply and send it over a particular physical adapter, establishing the flow for that session. The reason is that ARP is not a routable protocol. It does not have an IP header and therefore is not sent to the router or default gateway.

ARP is only a local subnet protocol. In addition, since the G-ARP is not a broadcast packet, the router will not process it and will not update its own ARP cache. The only way that the router would process an ARP that is intended for another network device is if it has Proxy ARP enabled and the host has no default gateway. This is very rare and not recommended for most applications. Transmit traffic through a router will be load balanced as transmit load balancing is based on the source and destination IP address and TCP/UDP port number. Since routers do not alter the source and destination IP address, the load balancing algorithm works as intended.

Configuring routers for Hot Standby Routing Protocol (HSRP) does not allow for receive load balancing to occur in the NIC team. In general, HSRP allows for two routers to act as one router, advertising a virtual IP and virtual MAC address. One physical router is the active interface while the other is standby. Although HSRP can also load share nodes (using different default gateways on the host nodes) across multiple routers in HSRP groups, it always points to the primary MAC address of the team. 2 Generic Trunking Generic Trunking is a switch-assisted teaming mode and requires configuring ports at both ends of the link: server interfaces and switch ports.

This is often referred to as Cisco Fast EtherChannel or Gigabit EtherChannel. In addition, generic trunking supports similar implementations by other switch OEMs such as Extreme Networks Load Sharing and Bay Networks or IEEE 802. 3ad Link Aggregation static mode. In this mode, the team advertises one MAC Address and one IP Address when the protocol stack responds to ARP Requests. In addition, each physical adapter in the team uses the same team MAC address when transmitting frames. This is possible since the switch at the other end of the link is aware of the teaming mode and will handle the use of a single MAC address by every port in the team.

The forwarding table in the switch will reflect the trunk as a single virtual port. In this teaming mode, the intermediate driver controls load balancing and failover for outgoing traffic only, while incoming traffic is controlled by the switch firmware and hardware. As is the case for Smart Load Balancing, the BASP intermediate driver uses the IP/TCP/UDP source and destination addresses to load balance the transmit traffic from the server. Most switches implement an XOR hashing of the source and destination MAC address. 3 Link Aggregation (IEEE 802. 3ad LACP) Link Aggregation is similar to Generic Trunking except that it uses the Link Aggregation Control Protocol to negotiate the ports that will make up the team.

LACP must be enabled at both ends of the link for the team to be operational. If LACP is not available at both ends of the link, 802. 3ad provides a manual aggregation that only requires both ends of the link to be in a link up state. Because manual aggregation provides for the activation of a member link without performing the LACP message exchanges, it should not be considered as reliable and robust as an LACP negotiated link. LACP automatically determines which member links can be aggregated and then aggregates them. It provides for the controlled addition and removal of physical links for the link aggregation so that no frames are lost or duplicated.

The removal of aggregate link members is provided by the marker protocol that can be optionally enabled for Link Aggregation Control Protocol (LACP) enabled aggregate links. The Link Aggregation group advertises a single MAC address for all the ports in the trunk. The MAC address of the Aggregator can be the MAC addresses of one of the MACs that make up the group. LACP and marker protocols use a multicast destination address. The Link Aggregation control function determines which links may be aggregated and then binds the ports to an Aggregator function in the system and monitors conditions to determine if a change in the aggregation group is required.

Link aggregation combines the individual capacity of multiple links to form a high performance virtual link. The failure or replacement of a link in an LACP trunk will not cause loss of connectivity. The traffic will simply be failed over to the remaining links in the trunk. 3 Software Components Teaming is implemented via an NDIS intermediate driver in the Windows Operating System environment. This software component works with the miniport driver, the NDIS layer, and the protocol stack to enable the teaming architecture (see Figure 3). The miniport driver controls the host LAN controller directly to enable functions such as sends, receives, and interrupt processing.

The intermediate driver fits between the miniport driver and the protocol layer multiplexing several miniport driver instances, and creating a virtual adapter that looks like a single adapter to the NDIS layer. NDIS provides a set of library functions to enable the communications between either miniport drivers or intermediate drivers and the protocol stack. The protocol stack implements IP, IPX and ARP. A protocol address such as an IP address is assigned to each miniport device instance, but when an Intermediate driver is installed, the protocol address is assigned to the virtual team adapter and not to the individual miniport devices that make up the team.

The Broadcom supplied teaming support is provided by three individual software components that work together and are supported as a package. When one component is upgraded, all the other components must be upgraded to the supported versions. The following table describes the three software components and their associated files for supported operating systems. | Software Component | Broadcom Name | Windows | Linux | NetWare | | Miniport Driver | Broadcom Base Driver | B57xp32. sys | Bcm5700 | B57. lan | | | | B57w2k. ys | | | | Intermediate Driver | Broadcom Advanced Server | Baspxp32. sys | Basp | Basp. lan | | | Program (BASP) | Baspw2k. sys | | | | Configuration GUI | Broadcom Advanced Control | Bacs | N/A | N/A | | | Suite (BACS) | | | | Table 3. Broadcom Teaming Software Component 4 Hardware Requirements

The various teaming modes described in this document place certain restrictions on the networking equipment used to connect clients to teamed servers. Each type of network interconnect technology has an effect on teaming as described below. 1 Repeater Hub A Repeater Hub allows a network administrator to extend an Ethernet network beyond the limits of an individual segment. The repeater regenerates the input signal received on one port onto all other connected ports, forming a single collision domain. This means that when a station attached to a repeater sends an Ethernet frame to another station, every station within the same collision domain will also receive that message.

If two stations begin transmitting at the same time, a collision will occur, and each transmitting station will need to retransmit its data after waiting a random amount of time. The use of a repeater requires that each station participating within the collision domain operate in half-duplex mode. Though half-duplex mode is supported for Gigabit Ethernet devices in the IEEE 802. 3 specification, it is not supported by the majority of Gigabit Ethernet controller manufacturers and will not be considered here. Teaming across Hubs is supported for troubleshooting purposes such as connecting a Network Analyzer for SLB teams only. 4 Switching Hub Unlike a Repeater Hub, a Switching Hub (or more simply a Switch) allows an Ethernet network to be broken into multiple collision domains.

The switch is responsible for forwarding Ethernet packets between hosts based solely on Ethernet MAC addresses. A physical network adapter that is attached to a switch may operate in half-duplex or full-duplex mode. In order to support Generic Trunking and 802. 3ad Link Aggregation, a switch must specifically support such functionality. If the switch does not support these protocols, it may still be used for Smart Load Balancing. 7 Router A router is designed to route network traffic based on Layer 3 or higher protocols, although it will often also work as a Layer 2 device with switching capabilities. Teaming ports connected directly to a router is not supported. 6 Supported Teaming by OS

All teaming modes are supported for the IA-32 server operating systems as shown in Table 4. | Teaming Mode | Windows | Linux | NetWare | | Smart Load Balancing and Failover | | | | | Generic Trunking | | | | | Link Aggregation | | | | Table 4. Teaming Support by Operating System 7 Utilities for

Configuring Teaming by OS Table 5 lists the tools used to configure teaming in the supported operating system environments. Operating System | Configuration Tool | | Windows 2000 | BACS | | Windows 2003 | BACS | | NetWare 5/6 | Autoexec. ncf and Basp. lan | | Linux | Baspcfg | Table 5. Operating System Configuration Tools The Broadcom Advanced Control Suite (BACS) (see Figure 1) is designed to run in one of the following 32-bit Windows operating systems: Microsoft® Windows® 2000 and Windows Server 2003. BACS is used to configure load balancing and fault tolerance teaming, and VLANs. In addition, it displays the MAC address, driver version, and status information. The BACS also includes a number of diagnostics tools such as hardware diagnostics, cable testing, and a network topology test. [pic] Figure 1.

Broadcom Advanced Control Suite (Release 6. 6 and 6. 7) When an adapter configuration is saved in NetWare, the NetWare install program adds load and bind statements to the Autoexec. ncf file. By accessing this file, you can verify the parameters configured for each adapter, add or delete parameters, or modify parameters. BASP Configuration (baspcfg) is a command line tool for Linux to configure the BASP teams, add/remove NICs, and add/remove virtual devices. This tool can be used in custom initialization scripts. Please read your distribution-specific documentation for more information on your distributors startup procedures. 8 Supported Features by Team Type

Table 6 provides a feature comparison across the teaming modes supported by Dell. Use this table to determine the best teaming mode for your application. The teaming software supports up to 8 ports in a single team and up to 4 teams in a single system. The four teams can be any combination of the supported teaming modes but must be on separate networks or subnets. | Teaming Mode-> | Fault Tolerance | Load Balancing | Switch Dependent Static | Switch Dependent Dynamic Link | | | | | Trunking | Aggregation (IEEE 802. ad) | | Function | SLB w/ Standby* | SLB | Generic Trunking | Link Aggregation | | Number of ports per team (Same | 2-8 | 2-8 | 2-8 | 2-8 | | Broadcast domain) | | | | | | Number of teams | 4 | 4 | 4 | 4 | | NIC Fault Tolerance | Yes | Yes | Yes | Yes | | Switch Link Fault Tolerance | Yes | Yes | Switch Dependent | Switch Dependent | |(same Broadcast domain) | | | | | | TX Load Balance | No | Yes | Yes | Yes | | RX Load Balance | No | Yes | Yes (Performed by Switch) | Yes (Performed by Switch) | | Requires Compatible Switch | No | No | Yes | Yes | | Heartbeats to check connectivity| No | No | No | No | | Mixed Media-NICs with different | Yes | Yes | Switch Dependent | Switch Dependent | | media | | | | | | Mixed Speed-NICs that do not | Yes | Yes | No | No | | support a common speed, but can | | | | | | operate at different speeds | | | | | | Mixed Speed-NICs that support | Yes | Yes | No (must be the same speed) | Yes | | common speed(s), but can operate| | | | | | at different speeds. | | | | | Load balances TCP/IP | No | Yes | Yes | Yes | | Mixed Vendor Teaming | Yes** | Yes** | Yes** | Yes** | | Load balances non-IP protocols | No | Yes (IPX outbound traffic | Yes | Yes | | | | only) | | | | Same MAC Address for all team | No | No | Yes | Yes | | members | | | | | | Same IP Address for all team | Yes | Yes | Yes | Yes | | members | | | | | | Load Balancing by IP Address | No | Yes | Yes | Yes | | Load Balancing by MAC Address | No | Yes (used for non-IP/IPX | Yes | Yes | | | | protocols) | | | *- SLB with one primary and one standby member **- Requires at least one Broadcom adapter in the team Table 6. Comparison of Teaming Modes 9 Selecting a team type

The following flowchart provides the decision flow when planning for teaming. The primary rationale for teaming is the need for additional network bandwidth and fault tolerance. Teaming offers link aggregation and fault tolerance to meet both of these requirements. Preference teaming should be selected in the following order: IEEE 802. 3ad as the first choice, Generic Trunking as the second choice, and SLB teaming as the third choice when using unmanaged switches or switches that do not support the first two options. However, if switch fault tolerance is a requirement then SLB is the only choice (see Figure 2). [pic] Figure 2. Process for Selecting a Team Type Teaming Mechanisms

This section provides an overview on how the Broadcom BASP intermediate driver is implemented and how it performs load balancing and failover. 1 Architecture The Broadcom Advanced Server Program is implemented as an NDIS intermediate driver (see Figure 3). It operates below protocol stacks such as TCP/IP and IPX and appears as a virtual adapter. This virtual adapter inherits the MAC Address of the first port initialized in the team. A Layer 3 address must also be configured for the virtual adapter. The primary function of BASP is to balance inbound (for SLB) and outbound traffic (for all teaming modes) among the physical adapters installed on the system selected for teaming. The inbound and outbound algorithms are independent and orthogonal to each other.

The outbound traffic for a particular session can be assigned to a given port while its corresponding inbound traffic can be assigned to a different port. [pic] Figure 3. Intermediate Driver 1 Outbound Traffic Flow The Broadcom Intermediate Driver manages the outbound traffic flow for all teaming modes. For outbound traffic, every packet is first classified into a flow, and then distributed to the selected physical adapter for transmission. The flow classification involves an efficient hash computation over known protocol fields. The resulting hash value is used to index into an Outbound Flow Hash Table. The selected Outbound Flow Hash Entry contains the index of the selected physical adapter responsible for transmitting this flow.

The source MAC address of the packets will then be modified to the MAC address of the selected physical adapter. The modified packet is then passed to the selected physical adapter for transmission. The outbound TCP and UDP packets are classified using Layer 3 and Layer 4 header information. This scheme improves the load distributions for popular Internet protocol services using well-known ports such as HTTP and FTP. Therefore, BASP performs load balancing on a TCP session basis and not on a packet-by-packet basis. In the Outbound Flow Hash Entries, statistics counters are also updated after classification. The load-balancing engine uses these counters to periodically distribute the flows across teamed ports.

The outbound code path has been designed to achieve best possible concurrency where multiple concurrent accesses to the Outbound Flow Hash Table are allowed. For protocols other than TCP/IP, the first physical adapter will always be selected for outbound packets. The exception is Address Resolution Protocol (ARP), which is handled differently to achieve inbound load balancing. 2 Inbound Traffic Flow (SLB Only) The Broadcom Intermediate Driver manages the inbound traffic flow for the SLB teaming mode. Unlike outbound load balancing, inbound load balancing can only be applied to IP addresses that are located in the same subnet as the load-balancing server.

Inbound load balancing exploits a unique characteristic of Address Resolution Protocol (RFC0826), in which each IP host uses its own ARP cache to encapsulate the IP Datagram into an Ethernet frame. BASP carefully manipulates the ARP response to direct each IP host to send the inbound IP packet to the desired physical adapter. Therefore, inbound load balancing is a plan-ahead scheme based on statistical history of the inbound flows. New connections from a client to the server will always occur over the primary physical adapter (because the ARP Reply generated by the operating system protocol stack will always associate the logical IP address with the MAC address of the primary physical adapter).

Like the outbound case, there is an Inbound Flow Head Hash Table. Each entry inside this table has a singly linked list and each link (Inbound Flow Entries) represents an IP host located in the same subnet. When an inbound IP Datagram arrives, the appropriate Inbound Flow Head Entry is located by hashing the source IP address of the IP Datagram. Two statistics counters stored in the selected entry are also updated. These counters are used in the same fashion as the outbound counters by the load-balancing engine periodically to reassign the flows to the physical adapter. On the inbound code path, the Inbound Flow Head Hash Table is also designed to allow concurrent access.

The link lists of Inbound Flow Entries are only referenced in the event of processing ARP packets and the periodic load balancing. There is no per packet reference to the Inbound Flow Entries. Even though the link lists are not bounded; the overhead in processing each non-ARP packet is always a constant. The processing of ARP packets, both inbound and outbound, however, depends on the number of links inside the corresponding link list. On the inbound processing path, filtering is also employed to prevent broadcast packets from looping back through the system from other physical adapters. 3 Protocol support ARP and IP/TCP/UDP flows are load balanced.

If the packet is an IP protocol only, such as ICMP or IGMP, then all data flowing to a particular IP address will go out through the same physical adapter. If the packet uses TCP or UDP for the L4 protocol, then the port number is added to the hashing algorithm, so two separate L4 flows can go out through two separate physical adapters to the same IP address. For example, assume the client has an IP address of 10. 0. 0. 1. All IGMP and ICMP traffic will go out the same physical adapter because only the IP address is used for the hash. The flow would look something like this: IGMP ——> PhysAdapter1 ——> 10. 0. 0. 1 ICMP ——> PhysAdapter1 ——> 10. 0. 0. 1

If the server also sends an TCP and UDP flow to the same 10. 0. 0. 1 address, they can be on the same physical adapter as IGMP and ICMP, or on completely different physical adapters from ICMP and IGMP. The stream may look like this: IGMP ——> PhysAdapter1 ——> 10. 0. 0. 1 ICMP ——> PhysAdapter1 ——> 10. 0. 0. 1 TCP ——> PhysAdapter1 ——> 10. 0. 0. 1 UDP ——> PhysAdatper1 ——> 10. 0. 0. 1 Or the streams may look like this: IGMP ——> PhysAdapter1 ——> 10. 0. 0. 1 ICMP ——> PhysAdapter1 ——>

10. 0. 0. 1 TCP ——> PhysAdapter2 ——> 10. 0. 0. 1 UDP ——> PhysAdatper3 ——> 10. 0. 0. 1

The actual assignment between adapters may change over time, but any protocol that is not TCP/UDP based will go over the same physical adapter because only the IP address is used in the hash. 4 Performance Modern network interface cards provide many hardware features that reduce CPU utilization by offloading certain CPU intensive operations (see Teaming and Other Advanced Networking Features). In contrast, the BASP intermediate driver is a purely software function that must examine every packet received from the protocol stacks and react to its contents before sending it out through a particular physical interface. Though the BASP driver can process each outgoing packet in near constant time, some applications that may already be CPU bound may suffer if operated over a teamed interface.

Such an application may be better suited to take advantage of the failover capabilities of the intermediate driver rather than the load balancing features, or it may operate more efficiently over a single physical adapter that provides a particular hardware feature such as Large Send Offload. 1 Performance Benchmarks Table 7 provides an example of the performance benefit that teaming offers by listing the throughput and CPU metrics for an LACP team as a function of the number of member ports. Chariot Benchmark throughput scales with the number of ports in the team with a modest increase in CPU utilization. The benchmark configuration consisted of 16 Windows 2000 clients with a TCP Window Size of 64KB used to generate traffic. The test server was running Windows Server 2003 with Large Send Offload. | Mode |# Of Ports | Receive Only Transmit Only | Bi-directional | | | |

CPU (%) | Mbps | CPU (%) | Mbps | CPU (%) | Mbps | | No Team | 1 | 22 | 936 | 21 | 949 | 29 | 1800 | | LACP Team | 2 | 34 | 1419 | 30 | 1885 | 35 | 2297 | | | 3 | 36 | 1428 | 38 | 2834 | 37 | 2375 | | | 4 | 31 | 1681 | 43 | 3770 | 44 | 3066 | Note: This is not a guarantee on performance. Performance will vary based on number of configuration factors and type of benchmark. It does indicate that link aggregation does provide a positive performance improvement as the number of ports increase in a team. Table 7. LACP Teaming Performance Large Send Offload enables an almost linear scalability of transmit throughput as a function of the number of team members as shown in Graph 1. [pic] Graph 1. Teaming Performance Scalability 2 Teaming Modes 1 Switch Independent The Broadcom Smart Load Balancing team allows 2 to 8 physical adapters to operate as a single virtual adapter.

The greatest benefit of SLB is that it will operate on any IEEE compliant switches and requires no special configuration. 1 Smart Load Balancing and Failover SLB provides for switch-independent, bi-directional, fault-tolerant teaming and load balancing. Switch independence implies that there is no specific support for this function required in the switch, allowing SLB to be compatible with all switches. Under SLB, all adapters in the team have separate MAC addresses. The load-balancing algorithm operates on Layer 3 addresses of the source and destination nodes, which enables SLB to load balance both incoming and outgoing traffic The BASP intermediate driver continually monitors the physical ports in a team for link loss.

In the event of link loss on any port, traffic is automatically diverted to other ports in the team. The SLB teaming mode supports switch fault tolerance by allowing teaming across different switches- provided the switches are on the same physical network or broadcast domain. 1 Network Communications The following are the key attributes of SLB: ? Failover mechanism – Link loss detection ? Load Balancing Algorithm – Inbound and outbound traffic are balanced through a Broadcom proprietary mechanism based on L4 flows ? Outbound Load Balancing using MAC Address – No ? Outbound Load Balancing using IP Address – Yes ? Multi-vendor Teaming – Supported (Must include at least 1 Broadcom Ethernet controller as a team member) 2 Applications

The SLB algorithm is most appropriate in SOHO and small business environments where cost is a concern or with commodity switching equipment. SLB teaming works with unmanaged Layer 2 switches and is a cost-effective way of getting redundancy and link aggregation at the server. Smart Load Balancing also supports teaming physical adapters with differing link capabilities. In addition, SLB is recommended when switch fault tolerance with teaming is required. 3 Configuration recommendations SLB supports connecting the teamed ports to hubs and switches if they are on the same broadcast domain. It does not support connecting to a router or layer 3 switches because the ports must be on the same subnet. 2 Switch Dependent 1 Generic Static Trunking

This mode supports a variety of environments where the NIC’s link partners are statically configured to support a proprietary trunking mechanism. This mode could be used to support Lucent’s “ Open Trunk,” Cisco’s Fast EtherChannel (FEC), and Cisco’s Gigabit EtherChannel (GEC). In the static mode, as in generic link aggregation, the switch administrator needs to assign the ports to the team, and this assignment cannot be altered by the

BASP, as there is no exchange of the Link Aggregation Control Protocol (LACP) frame. With this mode, all adapters in the team are configured to receive packets for the same MAC address. Trunking operates on Layer 2 addresses and supports load balancing and failover for both inbound and outbound traffic.

The BASP driver determines the load-balancing scheme for outbound packets, using layer 4 protocols previously discussed, whereas the team’s link partner determines the load-balancing scheme for inbound packets. The attached switch must support the appropriate trunking scheme for this mode of operation. Both the BASP and the switch continually monitor their ports for link loss. In the event of link loss on any port, traffic is automatically diverted to other ports in the team. 1 Network Communications The following are the key attributes of Generic Static Trunking: ? Failover mechanism – Link loss detection ? Load Balancing Algorithm – Outbound traffic is balanced through Broadcom proprietary mechanism based L4 flows.

Inbound traffic is balanced according to a switch specific mechanism. ? Outbound Load Balancing using MAC Address – No ? Outbound Load Balancing using IP Address – Yes ? Multi-vendor teaming – Supported (Must include at least 1 Broadcom Ethernet controller as a team member) 2 Applications Generic trunking works with switches that support Cisco Fast EtherChannel, Cisco Gigabit EtherChannel, Extreme Networks Load Sharing and Bay Networks or IEEE 802. 3ad Link Aggregation static mode. Since load balancing is implemented on Layer 2 addresses, all higher protocols such as IP, IPX, and NetBEUI are supported. Therefore, this is the recommended teaming mode when the switch supports generic trunking modes over SLB.

Configuration Recommendations Static trunking supports connecting the teamed ports to switches if they are on the same broadcast domain and support generic trunking. It does not support connecting to a router or layer 3 switches since the ports must be on the same subnet. 2 Dynamic Trunking (IEEE 802. 3ad Link Aggregation) This mode supports link aggregation through static and dynamic configuration via the Link Aggregation Control Protocol (LACP). With this mode, all adapters in the team are configured to receive packets for the same MAC address. The MAC address of the first NIC in the team is used and cannot be substituted for a different MAC address.

The BASP driver determines the load-balancing scheme for outbound packets, using layer 4 protocols previously discussed, whereas the team's link partner determines the load-balancing scheme for inbound packets. Because the load balancing is implemented on Layer 2, all higher protocols such as IP, IPX, and NetBEUI are supported. The attached switch must support the 802. 3ad Link Aggregation standard for this mode of operation. The switch will manage the inbound traffic to the NIC while the BASP manages the outbound traffic. Both the BASP and the switch continually monitor their ports for link loss. In the event of link loss on any port, traffic is automatically diverted to other ports in the team. 1 Network Communications The following are the key attributes of Dynamic Trunking: ? Failover mechanism – Link loss detection Load Balancing Algorithm – Outbound traffic is balanced through a Broadcom proprietary mechanism based on L4 flows. Inbound traffic is balanced according to a switch specific mechanism. ? Outbound Load Balancing using MAC Address – No ? Outbound Load Balancing using IP Address – Yes ? Multi-vendor teaming – Supported

(Must include at least 1 Broadcom Ethernet controller as a team member) 2 Applications Dynamic trunking works with switches that support IEEE 802. 3ad Link Aggregation dynamic mode using LACP. Inbound load balancing is switch dependent. In general, the switch traffic is load balanced based on L2 addresses. In this case, all network protocols such as IP, IPX, and NetBEUI are load balanced.

Therefore, this is the recommended teaming mode when the switch supports LACP, except when switch fault tolerance is required. SLB is the only teaming mode that supports switch fault tolerance. 3 Configuration recommendations Dynamic trunking supports connecting the teamed ports to switches as long as they are on the same broadcast domain and supports IEEE 802. 3ad LACP trunking. It does not support connecting to a router or layer 3 switches since the ports must be on the same subnet. 3 Driver Support by Operating System As previously noted, the BASP is supported in the Windows 2000 Server, Windows Server 2003, Netware, and Linux operating system environments.

In a Netware environment, NESL support is required because BASP relies on the NIC drivers to generate NESL events during link changes and other failure events. For Linux environments, Broadcom's Network Interface Card Extension (NICE) support is required. NICE is an extension provided by Broadcom to standard Linux drivers, and supports monitoring of Address Resolution Protocol (ARP) requests, link detection, and VLANs. The following table summarizes the various teaming mode features for each operating system. | Features | Windows | NetWare | Red Hat® Linux (AS 2. 1, EL3. 0) | | |(W2K/ Server 2003) |(5. 1, 6.) | | | Smart Load Balancing (SLB) | | User interfaces | BACS[1] | Command line | Command line | | Number of teams | 4 | 4 | 4 | | Number of adapters per team | 8 | 8 | 8 | | Hot replace | Yes | Yes | No | | Hot add | Yes Yes | No | | Hot remove | Yes | Yes | No | | Link speeds support | Different speeds | Different speeds | Different speeds | | Frame protocol | IP | IP/IPX | IP | | Incoming packet management | BASP | BASP | BASP | | Outgoing packet management | BASP | BASP | BASP | | Failover event | Loss of link | Loss of link | Loss of link | | Failover time | Set