Networking Terms, Protocols, and Standards

REVISION HISTORY

Revision	*Date*	*Change Description*
NET-AN100-R	08/06/04	Initial release.

Broadcom Corporation
P.O. Box 57013
16215 Alton Parkway
Irvine, CA 92619-7013

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

This document provides a high-level definition for several of the most commonly-used networking terms, protocols, and standards. The goal is to ensure that everyone has the same basic vocabulary foundation to promote better understanding during high-level feature discussions.

This is not meant to be all encompassing networking terms glossary, but rather a targeted list of definitions that specifically apply to Broadcom's current and future generation devices. Since Broadcom is continuously adding new features to their devices, this document will be updated periodically with pertinent information that pertains to the feature additions.

ACRONYMS AND ABBREVIATIONS

The following tables list the acronyms and abbreviations within this document.

Table 1: Acronyms and Abbreviations

Acronym/ Abbreviation	*Description*
4D-PAM5	Four-Dimensional/Pulse-Amplitude-Modulation
ADSL	Asymmetric Digital Subscriber Line
BGP	Border Gateway Protocol
BGP-4	BGP version 4
BPDU	Bridge Protocol Data Units
CBS	Committed Burst Size
CDP	Cisco Discovery Protocol
CE	Customer Edge
CFI	Canonical Format Indicator
CIR	Committed Information Rate
CoS	Class of Service
CR-LDP	Constraints-based Label Distribution Protocol
CVID	Customer VLAN ID
DDR	Double-data Rate
DiffServ	Differentiated Services
DSL	Digital Subscriber Line
DSLAM	Digital Subscriber Line Access Multiplexer
DVMRP	Distance Vector Multicasting Routing Protocol
EAPOL	Extensible Authentication Protocol over LAN
EAP	Extensible Authentication Protocol
EBS	Excess Burst Size
ECMP	Equal Cost Multipath
EFM	Ethernet in the First Mile
EPON	Ethernet Passive Optical Network
ETTB	Ethernet to the Business
ETTH	Ethernet to the home
ETTX	Ethernet to the X
FEC	Forwarding Equivalence Class
FCS	Frame Check Sequence
FFP	Fast Filter Processor
FIFO	First-in First-out
FTP	File Transfer Protocol
GARP	Generic Attribute Registration Protocol
GBIC	Gigabit Interface Converter
GMII	Gigabit Media Independent Interface
GRE	Generic Routing Encapsulation
GVRP	GARP VLAN Registration Protocol
HDLC	High-Level Data Link Control
HTLS	Hierarchal Transparent LAN Services
HTTP	Hyper Text Transfer Protocol
ICMP	Internet Control Message Protocol
IETF	Internet Engineering Task Force
IGMP	Internet Group Management Protocol
IHL	Internet Header Length
IP	Internet Protocol
IPv4	Internet Protocol Version 4
IPv6	Internet Protocol Version 6
ISN	Initial Sequence Number
LAN	Local Area Network
LACP	Link Aggregation Control Protocol

Acronym/ Abbreviation	Description
LDP	Label Distribution Protocol
LER	Label Edge Router
LLDP	Link Layer Discovery Protocol
LSP	Label-Switched Path
LSR	Label-Switched Router
MAC	Media Access Controller
MII	Media Independent Interface
MLT-3	Multi-Level Transmission-3
MSA	Multi-Source Agreement
MSTP	Multiple Spanning Tree Protocol
MTU	Multitenant Unit
MPLS	Multiprotocol Label Switching
NAT	Network Address Translation
NEXT	Near-End Cross-Talk
NRZ	Non-Return to Zero
NRZI	Non-Return to Zero Inverted
OSPF	Open Shortest PathFirst
PE	Provider Edge
PHY	Physical layer device
PIM	Protocol Independent Multicast
PIM-DM	Protocol Independent Multicast-Dense Mode
PIM-SM	Protocol Independent Multicast-Sparse Mode
PIM-SSM	Protocol Independent Multicast-Source Specific Multicast
PIR	Peak Information Rate
POP3	Post Office Protocol Version 3
POTS	Plain Old Telephone Service
PPP	Point-to-Point Protocol
QoS	Quality-of-Service
RGMII	Reduced Gigabit Media Independent Interface
RIPv2	Routing Information Protocol, Version 2
RMII	Reduced Media Independent Interface
RP	Rendezvous Point
RPT	Rendezvous Point Trees
RSTP	Rapid Spanning Tree Protocol
RSVP	Resource Reservation Protocol
RTBI	Reduced 10-Bit Interface
S3MII	Source-Synchronous Serial Media Independent Interface
SFF	Small Form Factor
SFP	Small Form Factor Pluggable
SGMII	Serial Gigabit Media Independent Interface
SMII	Serial Media Independent Interface
SMTP	Simple Mail Transfer Protocol
SNMP	Simple Network Management Protocol
SPT	Shortest Path Tree
SPVID	Service Provider VLAN ID
srTCM	Single-Rate Three-Color Marker
SSM	Source Specific Multicast
STP	Spanning Tree Protocol
TBI	10-Bit Interface
TCP	Transmission Control Protocol
TLS	Transparent LAN Services
TLV	Type-Length-Value
ToS	Type of Service
trTCM	Two-Rate Three-Color Marker
TTL	Time-to-Live
UDP	User Datagram Protocol
UTP	Unshielded Twisted Pair
VDSL	Very high Data Rate Digital Subscriber Line
VID	VLAN Identifier
VLAN	Virtual Local Area Network
WAN	Wide Area Network
VPLS	Virtual Private LAN Service
WCMP	Weighted Cost Multipath
XAUI	10-Gigabit Attachment Unit Interface
XFP	10-Gigabit Small Form Factor Pluggable
XGMII	10-Gigabit Media Independent Interface

STANDARDS AND PROTOCOLS

Because there are so many different terms throughout the networking industry, it can be difficult to remember which are standards and which are protocols. This section identifies some of the most common of each.

IEEE 802.1d—TRANSPARENT BRIDGING

802.1d is the industry standard for the transparent bridge designed to ensure interoperability between bridge venders. One of the most important aspects of 802.1d is the Spanning Tree Protocol (STP). STP is a link management protocol that provides path redundancy while preventing loops in the network, a common problem introduced by Local Area Network (LAN) bridging.

Only one active path can safely exist between to hosts. Multiple active paths between hosts are referred to as loops. These loops can confuse switches, resulting in various ill effects including duplicate packets and never-ending broadcasts.

To provide path redundancy, STP defines a tree that spans all switches in an extended network. STP forces certain redundant data paths into a standby or blocked state. If one network segment in the STP becomes unreachable, or if STP costs change, the Spanning Tree algorithm reconfigures the Spanning Tree topology and reestablishes the link by activating the standby path.

STP operation is transparent to end stations, which are unaware whether they are connected to a single LAN segment or a switched LAN of multiple segments. Spanning Tree enabled devices communicate to one another by exchanging Bridge Protocol Data Units (BPDUs).

The way STP controls equipment ports is by setting them to one of five port states. The following table summarizes the capabilities of each state.

Table 2: Spanning Tree Port States

Port State	*Receive BPDUs*	*Transmit BPDUs*	*Learn Addresses*	*Forward Packets*
Disabled				
Blocking	X			
Listening	X	X		
Learning	X	X	X	
Forwarding	X	X	X	X

All Broadcom switch devices support Spanning Tree by allowing the management software to set the port state. All devices also have some mechanism for detecting BPDUs and forwarding them to the management software.

IEEE 802.1q—Virtual LANs

The IEEE 802.1q standard was developed to address the problem of how to break large networks into smaller parts so broadcast and multicast traffic would not grab more bandwidth than necessary. The standard also helps provide a higher level of security between segments of internal networks. The 802.1q specification establishes a standard method for inserting Virtual Local Area Network (VLAN) membership information into Ethernet frames.

A VLAN is an administratively configured LAN or broadcast domain. Instead of going to the wiring closet to move a cable to a different LAN, network administrators can accomplish this task remotely by configuring a port on an 802.1q-compliant switch to belong to a different VLAN. The ability to move end stations to different broadcast domains by setting membership profiles for each port on centrally-managed switches is one of the main advantages of 802.1q VLANs.

The switch acts as an intelligent traffic forwarder and a network security device. Frames get sent only to the ports where the destination device is attached. Broadcast and multicast frames are constrained by VLAN boundaries, so only stations whose ports are members of the same VLAN see those frames. This way, bandwidth is optimized and network security is enhanced. 802.1q VLANs can span many switches and are not limited to one switch, and can span even across WAN links. Sharing VLANs between switches is achieved by inserting a tag with a VLAN Identifier (VID) between one and 4094 into each frame. A VID must be assigned for each VLAN. By assigning the same VID to VLANs on many switches, one or more VLAN (broadcast domain) can be extended across a large network.

It is important to understand the each network device's 802.1q capabilities when provisioning VLANs. 802.1q-compliant ports can be configured to transmit tagged or untagged frames. Because tagged frames can be used to carry VLAN membership information between switches if a port has an 802.1q-compliant device attached, a VLAN can span multiple switches. If an attached device does not support 802.1q, sending it tagged frames could result in an error and the packets being dropped.

With the exception of unmanaged devices, most Broadcom switch devices support 802.1q. Depending on the market for which the switch was designed, the devices support from 16 to all 4K VLANs.

IEEE 802.1p—Priority

Working in conjunction with 802.1q, 802.1p uses the priority field in the 802.1q tag (see the IEEE 802.1q header structure) to communicate the packets intended priority. The intent is to group traffic into one of eight possible priority classes, but IEEE stopped short of mandating the use of its recommended traffic class definitions. There is no requirement for network devices along the traffic route to enforce or adhere to these priorities. The sender can only tag traffic with its preferred priority and hope that devices in the network respect it.

While most vendors today agree that 802.1p is the mechanism to tag frames for prioritization, there is no single uniform approach to implementing the underlying queuing mechanisms that actually implement the priority flows. 802.1p establishes eight levels of priority, but it is common for switch vendors to implement fewer than eight priority queues (i.e. two or four). Since the eight priority levels must be mapped to the fewer queues, the number of effective priority classes is decreased.

Any Broadcom switch that supports 802.1q also supports 802.1p. The priority may be used to map packets into the preferred egress queue.

IEEE 802.1s—Multiple Spanning Tree

Multiple Spanning Tree and Multiple Spanning Tree Protocol (MSTP) extend Spanning Tree support to VLANs. Its used to prevent loops just as before, but now multiple active paths between end stations are allowed to exist as long as they utilize different VLANs. In a non-VLAN environment, Spanning Tree only has one Spanning Tree group. In Multiple Spanning Tree, there can be several simultaneous Spanning Tree groups, often referred to as a Spanning Tree forest. MSTP improves the operation of the spanning tree while maintaining backward compatibility with equipment that is based on the (original) 802.1d spanning tree. Most Broadcom switch devices that support 802.1q also support 802.1s.

IEEE 802.1w—Rapid Spanning Tree

An addition to the IEEE 802.1d specification, Rapid Spanning Tree and Rapid Spanning Tree Protocol (RSTP) are designed to allow the network to recover from a bridge failure more quickly than with regular Spanning Tree. This is accomplished by reducing the number of port states and allowing a port to transition to the forwarding state much quicker. The following table shows the relationship between the STP and RSTP port states.

Table 3: Rapid Spanning Tree Port States

STP Port State	***RSTP Port State***
Disabled	Discarding
Blocking	Discarding
Listening	Discarding
Learning	Learning
Forwarding	Forwarding

RSTP improves the operation of the spanning tree while maintaining backward compatibility with equipment that is based on the (original) 802.1d Spanning Tree.

From a hardware point of view, because RSTP only differs from STP in the port state transitions, all Broadcom switch devices that support STP should also support RSTP. This is because the management software is responsible for setting the port state, not the switch itself.

IEEE 802.1x—Port-Based Authentication and Security

802.1x adopts the Extensible Authentication Protocol (EAP) to port authentication and security. By detecting Extensible Authentication Protocol over LAN (EAPOL) packets and forwarding them to a authentication server, a switch can block access from a particular station until the station has been positively identified.

Several Broadcom switch devices support 802.1x by detecting EAPOL packets and forwarding them to the management software.

IEEE 802.3x—Full-Duplex Flow Control

802.3x is the portion of the Ethernet standard that defines the use of special Media Access Controller (MAC) control frames known as PAUSE frames to signal a transmitting MAC to go quiet for the period of time specified in the frame. This form of flow control is strictly MAC-to-MAC. The amount of time a MAC can be paused depends on two things:

- The speed at which the link is running.
- The count value specified in the PAUSE packet.

The following table shows the possible PAUSE timer ranges.

Table 4: PAUSE Timer Ranges

Speed	*Time Range*
10 Mbps	0 to 3.36s (in 51.2 μs increments)
100 Mbps	0 to 336 ms (in 5.12 μs increments)
1000 Mbps	0 to 33.6 ms (in 512 ns increments)

One may also use the Xon/Xoff style of flow control. This style also uses PAUSE packets, but only two timer values of the maximum and the minimum. For Xon, a MAC transmits a PAUSE frame with the timer set to the maximum. If this amount of time is not sufficient, more PAUSE frames with the maximum timer value may be transmitted. Once the MAC no longer needs to flow control its partner, it sends a PAUSE frame with the timer set to zero. This is known as Xoff. All Broadcom switch devices support 802.3x flow control.

IEEE 802.3ab—1000BASE-T

802.3ab is the portion of the Ethernet standard that defines 1000 Mbps Ethernet over Category 5e unshielded twisted pair (UTP). As with each new speed of Ethernet, 802.3ab builds upon the earlier specifications and strives to maintain consistence throughout the standard where ever possible. Unlike 100BASE-TX which only uses two pairs within the Cat 5 cable, 802.3ab uses all four pairs. In addition, 802.3ab transmits and receives on all four pairs simultaneously. All Broadcom 1000 Mbps copper Physical layer devices (PHYs) (stand alone as well as integrated) support 802.3ab.

IEEE 802.3ad—Link Aggregation

802.3ad is the link aggregation standard. Among other things, 802.3ad provides increased bandwidth by combining the capacity of multiple links into one logical link, increased availability, and load sharing between the links.

The concept of link aggregation has been around for a long time. There are several proprietary implementations that allow for link aggregation or trunking, but 802.3ad defines an industry standard way and even provides a protocol for automatically configuration and maintenance. This protocol is known as Link Aggregation Control Protocol (LACP).

Most Broadcom switch devices support some form of trunking. Several devices (such as those belonging to the StrataXGS® family) support 802.3ad style trunking.

IEEE 803.2ae—10GBASE-X

802.3ae is the portion of the Ethernet standard that defines 10-Gbps Ethernet LAN and Wide Area Network (WAN) interfaces over fiber. In includes the definition of the MAC-PHY interface, 10-Gigabit Media Independent Interface (XGMII), and its popular extender, the 10-Gigabit Attachment Unit Interface (XAUI).

Broadcom's BCM5673 and BCM5674 directly support 802.3ae. Electrically, all StrataXGS devices support the XAUI portion of the specification through the HiGig™ interface.

IEEE 803.2ah—Ethernet in the First Mile

Ethernet in the First Mile (EFM) is an industry initiative of over a hundred companies trying to solve the problem of delivering higher bandwidth to businesses and homes in a cost-effective way. Other commonly used terms used to describe this are Ethernet to the X (ETTX), Ethernet to the Business (ETTB), and Ethernet to the home (ETTH). Several technologies are under consideration for this including 1000BASE-LX, 100BASE-FX, and Ethernet Passive Optical Network (EPON).

High-Level Data Link Control

High-Level Data Link Control (HDLC) is one of the most widely used serial protocols in use today. In fact, several other popular protocols such as Point-to-Point Protocol (PPP) are based in HDLC.

The actual HDLC frame structure is simple. The beginning and end of each frame is marked by flag characters (0x7E). No flag character can appear within the data itself. To ensure this requirement, any byte in the frame that is actually 0x7E is transparently modified and marked by a process known as bit stuffing. At the end of the frame, a Frame Check Sequence (FCS) is used to verify the data integrity.

INTERNET PROTOCOL

Internet Protocol Version 4

Internet Protocol Version 4 (IPv4) is the most widely used routing layer datagram service. IPv4 is designed to be a connectionless protocol where each packet has a 32-bit source and 32-bit destination address.

IPv4 is the foundation of the vast majority of networking applications today. Almost all other TCP/IP functions are constructed by layering atop IPv4. IPv4 is a datagram-oriented protocol, treating each packet independently. Each packet contains source and destination addressing information to allow receive devices to properly forward the packet to its final destination, but IPv4 lacks error checking (except for the IPv4 header) and the means to determine if the packet actually arrived at its destination.

The following table describes the services that Internet Protocol (IP) provides.

Table 5: IP Services

Service	*Description*
Addressing	IP headers contain 32-bit addresses that identify the sending and receiving hosts. These addresses are used by intermediate routers to select a path through the network for the packet.
Fragmentation	IP packets can be split or fragmented into smaller packets. This permits a large packet to travel across a network that can only handle smaller packets. IP transparently fragments and reassembles packets.
Packet timeouts	Each IP packet contains a Time-to-Live (TTL) field, which is decremented every time a router handles the packet. The packet is discarded if TTL reaches zero, preventing packets from running in circles forever and flooding a network.
Type of Service (ToS)	IP supports traffic prioritization by allowing packets to be labeled with an abstract ToS.
Options	IP provides several optional features, allowing a packet's sender to set requirements on the path it takes through the network (source routing), trace the route a packet takes (record route), and label packets with security features.

IP Version 6

IP Version 6 (IPv6) is designed to address the certain limitations inherent in IPv4. One of the most obvious enhancements is the increased size of the source and destination addresses to 128-bits each. This allows for more levels of network hierarchy as well as a much greater number of addressable end stations.

TRANSMISSION CONTROL PROTOCOL

Originally defined by RFC-793, Transmission Control Protocol (TCP) provides a reliable stream delivery and virtual connection service to applications through the use of sequenced acknowledgment, with retransmission of packets when necessary. TCP is known for its three-way handshaking that guarantees packet delivery.

TCP hosts and clients establish port connections for which to pass data. A client may have multiple simultaneous connections to the same host port, each from different source ports. The following table shows some of the more common TCP port numbers.

Table 6: Common TCP Port Numbers

Port Number	***Layer 5 Protocol***
20	File Transfer Protocol (FTP) Control
21	FTP Data
23	Telnet
25	Simple Mail Transfer Protocol (SMTP)
80	Hyper Text Transfer Protocol (HTTP)
110	Post Office Protocol Version 3 (POP3)
161	Simple Network Management Protocol (SNMP)
179	Border Gateway Protocol (BGP)

For a complete list of TCP/UDP port numbers as well as other useful information, go to the Internet Assigned Numbers Authority website at http://www.iana.org.

USER DATAGRAM PROTOCOL

User Datagram Protocol (UDP) provides a simple but unreliable message service for transaction-oriented services. Each segment has a port system similar to TCP, but lacks the multiple handshaking that guarantees packet delivery. UDP is defined by RFC-768.

INTERNET CONTROL MESSAGE PROTOCOL

Internet Control Message Protocol (ICMP) is a message protocol designed to convey information about the current conditions of IP network. ICMP is a required part of the IP protocol suite and is defined by RFC-792.

ICMP tasking includes:

- Announce network errors, such as a host or entire portion of the network being unreachable due to some type of failure. A TCP or UDP packet directed at a port number with no receiver attached is also reported via ICMP.
- Announce network congestion. When a router begins buffering too many packets due to an inability to transmit them as fast as they are being received, it generates ICMP Source Quench messages. Directed at the sender, these messages should slow the packet transmission rate. Source Quench messages are used sparingly, because generating too many would cause more network congestion.
- Assist troubleshooting. ICMP supports an Echo function, which just sends a packet on a round-trip between two hosts. Ping is a common network management tool and based on this feature. Ping transmits a series of packets, measuring average round-trip times and computing loss percentages.
- Announce timeouts. If an IP packet's TTL field drops to zero, the router discarding the packet often generates an ICMP packet announcing this fact. TraceRoute is a tool that maps network routes by sending packets with small TTL values and watching the ICMP timeout announcements.

INTERNET GROUP MANAGEMENT PROTOCOL

Internet Group Management Protocol (IGMP) is used by stations to report their group memberships to or to request group memberships from neighboring multicast-enabled routers. IGMP messages are encapsulated in IP datagrams, with an IP protocol number of 2. Common IGMP messages include group membership reports, group joins, and group leaves. IGMP is defined by RFC-1112.

GENERIC ATTRIBUTE REGISTRATION PROTOCOL

Generic Attribute Registration Protocol (GARP) provides a generic framework whereby devices in a bridged LAN (such as end stations and switches) can register and de-register attribute values (such as VLAN IDs) with each other. In doing so, the attributes are propagated to devices in the bridged LAN, and these devices form a reachability tree that is a subset of an active topology. For a bridged LAN, the active topology is normally created and maintained by the STP. GARP defines the architecture, rules of operation, state machines, and variables for the registration and de-registration of attribute values. By itself, GARP is not directly used by devices in a bridged LAN. It is the GARP applications that specify what the attribute represents and perform the meaningful actions.

GARP VLAN REGISTRATION PROTOCOL

GARP VLAN Registration Protocol (GVRP) allows a LAN device to signal other neighboring devices that it wants to receive packets for one or more VLANs. The main purpose of GVRP is to allow switches to automatically discover some of the VLAN information that would otherwise have to be manually configured in each switch. This is achieved by using GARP to propagate VID attributes across a bridged LAN. GVRP can also be run by network servers. These servers are usually configured to join several VLANs, and then signal the network switches of the VLANs they want to join.

MULTIPROTOCOL LABEL SWITCHING

Multiprotocol Label Switching (MPLS) is an Internet Engineering Task Force (IETF)-specified framework that provides for the designation, routing, forwarding, and switching of traffic flows through the network based on simple labels.

This framework performs the following tasks:

- Specifies mechanisms to manage traffic flows of various granularities, such as flows between different hardware, machines, or even flows between different applications.
- Remains independent of the layer-2 and layer-3 protocols.
- Provides a means to map IP addresses to simple, fixed-length labels used by different packet-forwarding and packet-switching technologies.
- Interfaces to existing routing protocols, such as Resource Reservation Protocol (RSVP) and Open Shortest PathFirst (OSPF).
- Supports IP, ATM, and Frame Relay layer-2 protocols.

In MPLS, data transmission occurs on Label-Switched Paths (LSPs). LSPs are a sequence of labels at each and every node along the path from the source to the destination. LSPs are established either prior to data transmission (control-driven) or upon detection of a certain flow of data (data-driven). The labels are underlying protocol-specific identifiers. There are several label distribution protocols used today (such as Label Distribution Protocol [LDP] or RSVP) or piggybacked on routing protocols like BGP and OSPF. Each data packet encapsulates and carries the labels during their journey from source to destination. High-speed switching of data is possible because the fixed-length labels are inserted at the beginning of the packet or cell, and can be used by hardware to switch packets quickly between links.

MPLS is a versatile solution to address the problems faced by present-day networks-speed, scalability, Quality-of-Service (QoS) management, and traffic engineering. MPLS has emerged as a solution to meet the bandwidth-management and service requirements for next-generation IP-based backbone networks.

LABEL DISTRIBUTION PROTOCOL

LDP is a protocol that defines a set of procedures and messages by which one Label Switched Router (LSR) informs another of the label bindings it has made.

A LSR may use LDP to establish LSPs through a network by mapping network layer routing information directly to data-link layer switched paths. These LSPs may have an endpoint at either a directly-attached neighbor or a network egress node, enabling switching via all intermediary nodes. A Forwarding Equivalence Class (FEC) is associated with each LSP created. This FEC specifies which packets are mapped to that LSP.

Two LSRs that use LDP to exchange label mapping information are known as LDP peers, and have an LDP session between them.

There are four kinds of LDP messages:

- Discovery messages
- Session messages
- Advertisement messages
- Notification messages

LDP messages have a common structure that uses a Type-Length-Value (TLV) to indicate the message type.

Constraints-Based Label Distribution Protocol

Constraints-based Label Distribution Protocol (CR-LDP) is an extension of LDP designed to enhance LDP's capabilities. The message structure and TLVs are the same between LDP and CR-LDP, except CR-LCP has additional TLVs that are not found in LDP.

Border Gateway Protocol Version 4

BGP version 4 (BGP-4) is a protocol for exchanging routing information between gateway hosts (each with its own router) in a network of autonomous systems. BGP is often the protocol used between gateway hosts on the Internet. The routing table contains a list of known routers, the addresses they can reach, and a cost metric associated with the path to each router so that the best available route is chosen.

Hosts send updated router table information only when one host has detected a change. Only the affected part of the routing table is sent. BGP-4 is the latest version, and allows administrators to configure route cost metrics based on their policy of choice. BGP-4 is defined by RFC-1771.

Open Shortest Path First

OSPF is an interior gateway routing protocol developed for IP networks based on a shortest path first or link-state algorithm. OSPF is defined by RFC-2328.

Routers use link-state algorithms to send routing information to all nodes in an interior network by calculating the shortest path to each node based on topography of the Internet constructed by each node. Each router sends that portion of the routing table (keeps track of routes to particular network destinations) that describes the state of its own links, and also sends the complete routing structure (topography).

The advantage of shortest path first algorithms is that they results in smaller more frequent updates everywhere. They converge quickly, preventing such problems as routing loops and count-to-infinity (when routers continuously increment the hop count to a particular network). This makes for a stable network.

Cisco Discovery Protocol

Cisco Discovery Protocol (CDP) is a media-independent and protocol-independent device-discovery protocol that runs on all Cisco-manufactured equipment including routers, access servers, bridges, and switches. Using CDP, a device can advertise its existence to other devices and receive information about other devices on the same LAN or on the remote side of a WAN. CDP runs on all media that support SNAP, including LANs, Frame Relay, and ATM media.

Link Layer Discovery Protocol

Link Layer Discovery Protocol (LLDP) is a new protocol defined in IEEE 802.1ab that lets neighboring devices notify one another of their existence. Each device on each port stores information defining itself, and sends updates to its directly connected neighbors as needed. These neighboring devices then store the information in standard SNMP MIBs.

The goal of LLDP is to be industry standard tool for link layer discovering that communicates the same sort of information previously available only through the use of vender proprietary protocols, such as CDP.

RESOURCE RESERVATION PROTOCOL

RSVP is a setup protocol designed for an integrated (also known as differentiated) services throughout the IP network. RSVP is used by a host on behalf of an application data stream to request a specific QoS from the network for particular data streams or flows. RSVP is also used by routers to deliver QoS control requests, and set up MPLS LSPs. RSVP is defined by RFC-2205.

PROTOCOL APPLICATIONS

This section details some of the current protocol application trends in the market today. These protocol applications are a practical use of the protocols discussed in this document to solve a particular problem or create a new service.

TRANSPARENT LAN SERVICES

Transparent LAN Services (TLS) is a somewhat broad term used to describe a variety of services that provide connectivity between geographically dispersed customer sites across a MAN/WAN network(s), as if they were connected using a LAN. TLS examples include, but are not limited to, Q-in-Q Virtual Private LAN Service (VPLS) and MPLS VPLS.

VIRTUAL PRIVATE LAN SERVICES USING MPLS

As mentioned earlier, VPLS is a means to connect geographically dispersed customer sites while providing the flexibility and control as if it were one single LAN. This is achieved by simulating an Ethernet virtual 802.1D bridge. It provides a L2 broadcast domain that is fully capable of learning and forwarding on Ethernet MAC addresses that is closed to a given set of users (VLAN).

An Ethernet port is used to connect a customer to the Provider Edge (PE) router acting as an MPLS Label Edge Router (LER). Customer traffic is subsequently mapped to a specific MPLS L2 VPN by configuring L2 MPLS LSPs based upon the input port ID or VLAN index, which depends on the preferred VPLS service.

Broadcast and multicast services are available over traditional LANs, but MPLS does not support such services. Sites that belong to the same broadcast domain and that are connected via an MPLS network expect broadcast, multicast, and unicast traffic to be forwarded to the proper location(s). This solved by creating a mesh of individual LSPs between PE routers belonging to a VPLS instance. This requires MAC address learning/aging on a per LSP basis, packet replication across LSPs for multicast/broadcast traffic, and for flooding of unknown unicast destination traffic.

The MPLS network provides a number of LSPs that form the basis for connections between LERs attached to the same MPLS network. The resulting set of interconnected LERs forms a private MPLS VPN where each LSP is uniquely identified at each MPLS interface by a label. An MPLS interface acting as a bridge must be able to flood, forward, and filter bridged frames.

The set of PE router devices interconnected via transport tunnels appears as a single 802.1D bridge/switch to customer. Each PE device learns:

- Remote MAC addresses to VC LSP associations
- Directly attached MAC addresses on customer facing ports

Many ISP deploy a PE router only in the central office and rely on a Multitenant Unit (MTU) switch to forward customer traffic to the PE. In this type of network, VPLS is deployed in a hierarchical fashion where the MTU has a spoke connection to the PE using a single Q-in-Q tunnel for each VPLS instance. The flows presented below describe packet processing steps in this model. It is possible for the customer to be connected directly to the PE. In that case, the PE performs the steps for both MTU and PE except that the Q-in-Q tunnel is not needed. The following figure illustrates the connections in a hierarchical VPLS model.

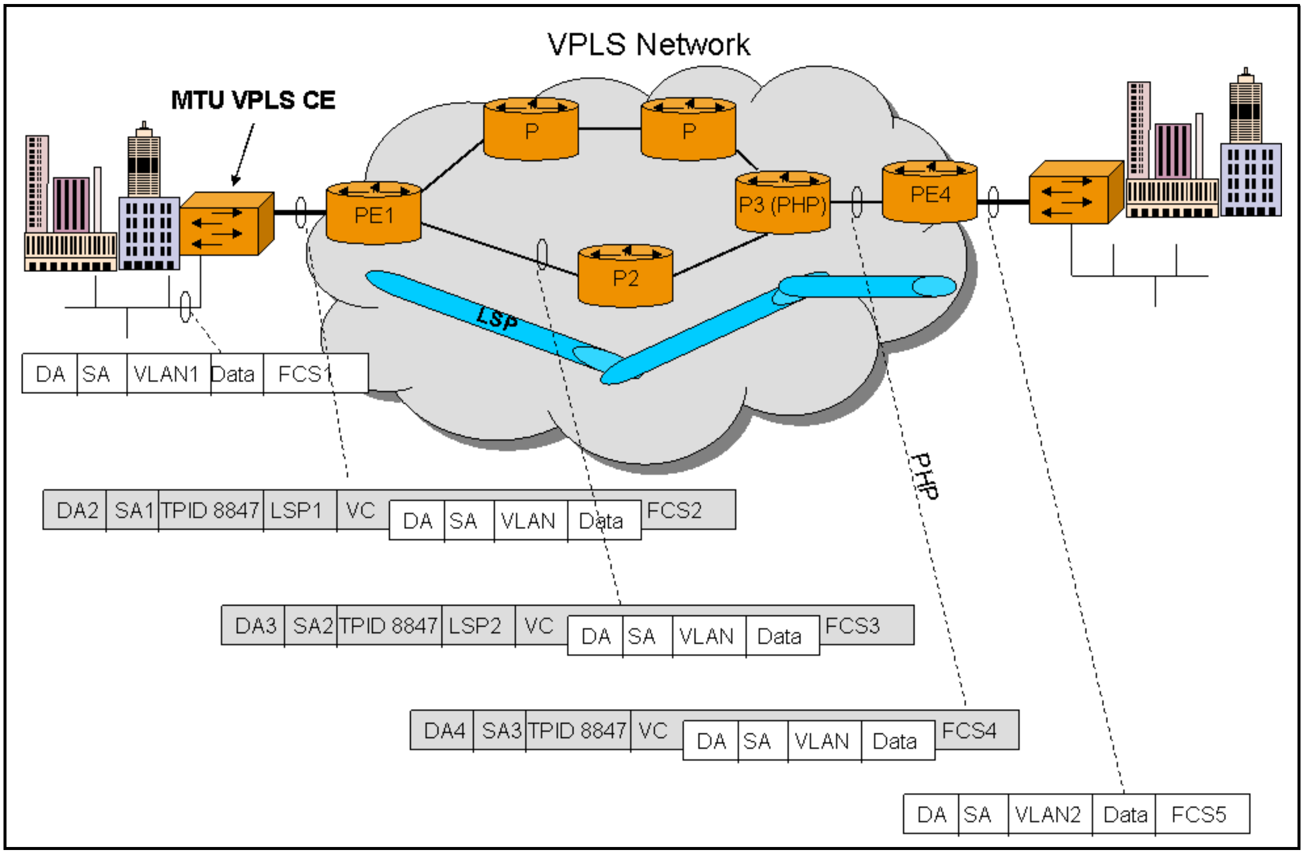
VPLS Network
MTU VPLS CE
P
P
PE1
P3 (PHP)
PE4
P2
LSP
PHP
DA | SA | VLAN1 | Data | FCS1
DA2 | SA1 | TPID 8847 | LSP1 | VC | DA | SA | VLAN | Data | FCS2
DA3 | SA2 | TPID 8847 | LSP2 | VC | DA | SA | VLAN | Data | FCS3
DA4 | SA3 | TPID 8847 | VC | DA | SA | VLAN | Data | FCS4
DA | SA | VLAN2 | Data | FCS5

Figure 1: MPLS VPLS Example

Broadcom supports MPLS VPLS in our Tucanna and EasyRider devices.

Virtual Private LAN Services Using Q-in-Q

Q-in-Q tagging (also know as double-tagging) is another form of TLS and uses 802.1q tags in a stacked fashion similar to a MPLS label stack. With this approach, customers and service providers can use all available VLAN tags independently. The customer tags are referred to as Customer VLAN IDs (CVIDs). Service provider tags are referred to as Service Provider VLAN IDs (SPVIDs).

When customer packets reach the service PE, the SPVID addition tag is added to the packet in front of the CVID. The service provider can then forward the customer's traffic based on the service provider's policies without limiting VLAN usage or performing VLAN ID translation.

When the double-tagged packet reaches the other side of the service provider's network, the SPVID is stripped from the packet.

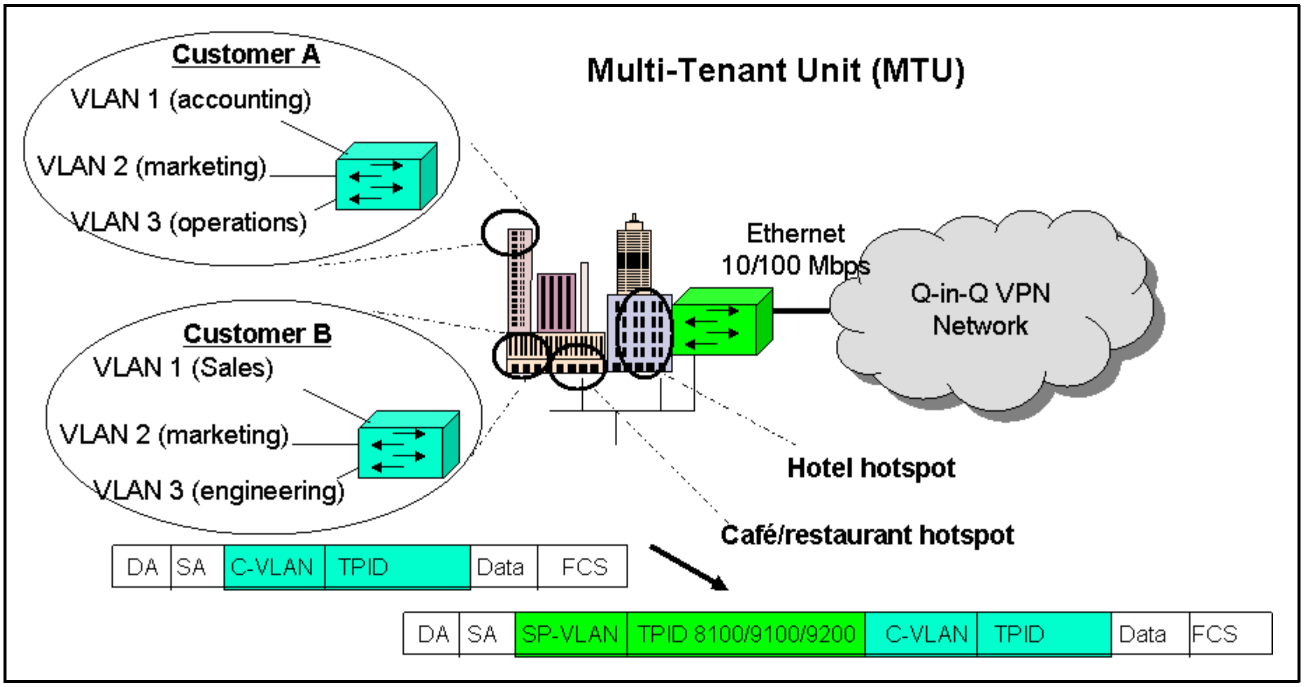

Figure 2: Q-in-Q VPLS Example

Broadcom supports Q-in-Q VPLS in our Tucanna and EasyRider devices.

Hierarchal Transparent LAN Services

Since both Q-in-Q VPLS and MPLS VPLS both allow for a hierarchical service structure (i.e. stacked tags or labels, with the outmost tag or label having the greatest significance at any point in the network), both implementations are often referred to as Hierarchal Transparent LAN Services (HTLS).

This is even truer in the case of MPLS. A customer may already differentiate services on their private network by using different MPLS labels for different services. In this case, the service PE devices no longer need to map an L2 or L3 address to an LSP. Instead, the PE can add an additional label representing the customer's traffic. Different labels could be added to the same customer's traffic to provide different classes of service for the different traffic types.

DIFFERENTIATED SERVICES

Differentiated Services (DiffServ) is a way to break traffic types into classes so that points along a forwarding path may treat the different classes in a preferred manner. Instead of creating a completely new protocol for signaling the preferred class or behavior, the ToS field in the IPv4 header is used. The 8 bits provided by the ToS field are replaced with a 6-bit DSCP (Differentiated Services Codepoint) that signifies the traffic class or the preferred behavior. The remaining two bits are currently unused. For more information on DiffServ, please see RFC-2474 and RFC-2475.

All Broadcom devices that have Fast Filter Processors (FFP) have the ability to remap or alter the DSCP in DiffServ packets. In addition, some devices can use the DSCP for egress queue mapping.

RFC-2547—BGP/MPLS VPNS

RFC-2547 describes how a service provider can provide IP-VPN services to customers by carrying the customer's IP traffic across the service provider's network using MPLS.

Each PE router actually acts as several virtual routers, each maintaining route information for one VPN. BGP sessions are used between the Customer Edge (CE) router and the PE, and between each PE within the provider's network. With this notion of a virtual router per VPN (in this case, virtual router per MPLS label or group of labels), a customer's traffic is isolated to their assigned VPN. A customer's traffic cannot be forwarded to another VPN within the provider's network.

The following process simplifies the packet flow into and out of the IP-VPN.

1. The customer's packet arrives at the PE. The PE:
   - Uses the ingress port to identify the correct VPN and virtual router.
   - Looks up the destination IP address in the virtual router's forwarding table to create a two-deep label stack (a BGP label and a transport label).
   - Sends the packet to the next hop.
2. The transport label is swapped if there are intermediate LSRs in the provider's network, but the BGP label is unchanged.
3. When the packet arrives at the destination PE, the labels are stripped and the packet is forwarded to the port identified by the BGP label.

The EasyRider family of devices support RFC-2527.

RFC-2697—Single-Rate Three-Color Marker

RFC-2697 defines Single-Rate Three-Color Marker (srTCM), which can be used as a component in a Diffserv traffic conditioner (see RFC2475 and RFC2474). The srTCM meters an IP packet stream and marks its packets either green, yellow, or red. Marking is based on a Committed Information Rate (CIR) and two associated burst sizes, a Committed Burst Size (CBS) and an Excess Burst Size (EBS). A packet is marked green if it does not exceed the CBS, yellow if it exceeds the CBS but not the EBS, and red otherwise.

Example: The srTCM is useful for ingress policing of a service, where only the length (not the peak rate) of the burst determines service eligibility.

The meter operates in one of two modes. In the color-blind mode, the meter assumes that the packet stream is uncolored. In the color-aware mode, the meter assumes that some preceding entity has precolored the incoming packet stream so that each packet is either green, yellow, or red.

The marker (re)colors an IP packet according to the results of the meter. Both the FireBlade and EasyRider families of devices support RFC-2697.

RFC-2698—Two Rate Three-Color Marker

RFC-2698 is a companion document to RFC-2697. RFC-2698 describes the Two-Rate Three-Color Marker (trTCM) where packets are marked based on two rates, and their associated burst sizes. A packet is marked red if it exceeds the Peak Information Rate (PIR), otherwise it is marked either yellow or green depending on whether it does or does not exceed the CIR. Both the FireBlade and EasyRider families of devices support RFC-2698.

RFC-3519—Mobile IP Traversal of NAT Devices

RFC-3519 attempts to solve the problem of mobile IP traversal of Network Address Translation (NAT) devices. Mobile IP relies on sending traffic from the home network to the mobile node or foreign agent through IP-in-IP tunneling. IP nodes that communicate from behind a NAT are reachable only through the NAT's public address(es). IP-in-IP tunneling does not generally contain enough information to permit unique translation from the common public address(es) to the particular care-of address of a mobile node or foreign agent that resides behind the NAT. More specifically, there are no TCP/UDP port numbers available for a NAT to use in the mapping of translations. For this reason, IP-in-IP tunnels generally cannot pass-through a NAT, and Mobile IP does not work across a NAT.

RFC-3519 defines an UDP tunneling between the mobile node and the home network, thereby providing NAT with enough information to properly map address translations.

RFC-2890—Key and Sequence Number Extensions to GRE

RFC-2890 describes enhancements to the Generic Routing Encapsulation (GRE) specification. The current GRE specification defines GRE as a protocol for encapsulation of an arbitrary protocol over another arbitrary network layer protocol. RFC-2890 takes this further by adding two new fields to GRE, the Key and the Sequence Number. The Key field is intended to be used for identifying an individual traffic flow within a tunnel. The Sequence Number field is used to maintain sequence of packets within the GRE tunnel.

Distance Vector Multicasting Routing Protocol

Described in RFC 1075, Distance Vector Multicasting Routing Protocol (DVMRP) is a multicast routing protocol that is based on Routing Information Protocol, Version 2 (RIPv2). DVMRP uses distance vector techniques to determine the next-best-hop among its neighboring routers. DVMRP uses messages called floods, prunes, and grafts to build a multicast spanning tree, which is often referred to as the dense mode source distribution tree, a Shortest Path Tree (SPT), or a truncated broadcast tree.

- Floods are a flood of multicast information to all outgoing interfaces.
- Prunes are sent up the spanning tree (or upstream) to communicate that nothing downstream from that point is interested in the multicast traffic.
- Grafts are the opposite of prunes. Grafts are sent upstream when a router wants to re-join the multicasts spanning tree.

To learn all its neighbors, a DVMRP router sends Hello messages with the All DVMRP router multicast address of 224.0.0.4. When neighbors respond, the initiator adds those neighbors to its multicast route table.

The whole purpose of a multicast routing protocol is to communicate the multicast membership between routers throughout the entire network, or at least throughout the entire multicast path. Since DVMRP is local in nature, it does not lead itself to large scale deployment.

Protocol Independent Multicast

The Protocol Independent Multicast (PIM) is multicast routing that can operate regardless of the unicast routing protocol being used. PIM uses the existing routing tables instead of creating new, separate tables like those used by DVMRP. The two main PIM modes are sparse and dense.

Protocol Independent Multicast-Sparse Mode

PIM-Sparse Mode (PIM-SM) provides a more efficient mechanism for multicasting when only a small percentage of end-users need to listen to the group. PIM-SM uses Rendezvous Point Trees (RPT) as its primary spanning tree, making use of single Rendezvous Point (RP) between sources and recipients. The RP is a network manager-specified, multicast-enabled router that is usually close to the multicast sources. Because end-users are downstream from the RP-based distribution tree, the designation for a particular multicast group is (*,G). All multicast groups are sourced from the RP, making the existence of multiple RPs possible, each being responsible for some subset of all required multicast groups.

The method of discovery of new group users involves a standard SPT from the last-hop router back to the RP. Standard unicast routing tables are used and a PIM Shared Tree Join is performed. The only necessity is for each router along the way and all the way up to the RP itself, is to add the (*,G) entry for the required multicast group. In this way, the end-user has joined the RPT for subsequent multicast packets for this group. When a given last-hop router discovers (via IGMP) that there are no more end-users for a given multicast group, a Shared Tree Prune message can be sent up the SPT towards the RP so that timeouts are not needed to prune the proper branches.

One feature of PIM-SM that is not available in other sparse mode protocols is the ability for a given last-hop router to ask for a direct SPT back to a given multicast source without requiring the source to link to the shared RP tree. This feature allows a given sourcing node the option of providing service directly to a set of end-users without routing its multicast payloads through the RP.

A multicast source provides its traffic to the RP via another SPT between the two. The RP knows that the source exists due to a unicast packet that is sent from the source directly to the RP's IP address using a special PIM message called a Source Registration. Once the unicast packet is received, the RP can now make the reverse connection back to the sourcing node.

PROTOCOL INDEPENDENT MULTICAST-DENSE MODE

While DVMRP (also a dense mode multicast protocol) uses an MRT and MFT to calculate which ports to transmit to for a given (S,G) combination, PIM-Dense Mode (PIM-DM) blindly transmits the multicast packet to all interfaces as long as that interface has not been pruned. PIM-DM accepts this additional packet duplication to operate independently of the unicast routing tables and their resultant topology. In addition, no parent/child databases need be created using this model. PIM-DM should only be used when a large percentage of the end-users require multicast traffic and bandwidth is plentiful.

PROTOCOL INDEPENDENT MULTICAST-SOURCE SPECIFIC MULTICAST

PIM-Source Specific Multicast (PIM-SSM) is similar to PIM-SM, except it solves one of the biggest problems with sparse mode operation—multicast source discovery. PIM-SSM builds SPTs rooted at the source immediately, because the router closest to the interested receiver host in SSM is informed of the unicast IP address of the source for the multicast traffic. With this approach, PIM-SSM bypasses the RP connection stage through shared distribution trees and goes directly to the source-based distribution tree.

Even though the messages are identical between PIM-SSM and PIM-SM, Source-Specific Multicast (SSM) introduces new terms to describe its new, source-specific, nature. The processes of joins and leave are known as subscribe and unsubscribe in SSM. The address identifiers in SM are known as (G), and in SSM they are more specifically known as (S,G). Lastly, what is referred to in SM as a group is referred to in SSM as a channel.

ROUTING CONCEPTS

This section highlights two routing concepts that are new to the StrataXGS family.

EQUAL COST MULTIPATH ROUTING

Equal Cost Multipath (ECMP) routing is a technique for routing packets along multiple paths of equal cost. If multiple equal cost routes exist to the same destination, ECMP can be used to provide load balancing among the redundant paths.

WEIGHTED COST MULTIPATH ROUTING

Weighted Cost Multipath (WCMP) routing is a technique for routing packets along multiple paths of weighted cost. If multiple routes exist to the same destination, weights can be assigned to certain routes so that more of the traffic is carried by that route than a route of lesser weight.

MAC/PHY INTERCONNECTS

This section attempts to describe all currently used MAC to PHY connections. Pin descriptions as well as connection diagrams are provided. The management interface (MDC/MDIO) is shown for all interfaces that support it. For signal integrity and EMI performance, series impedance matching resisters are shown in diagrams that should implement them. The value of Rsm may vary from implementation to implementation, but is usually between 20Ω and 39Ω.

MEDIA INDEPENDENT INTERFACE

IEEE 802.3u (Fast Ethernet Specification) first defined the Media Independent Interface (MII), which is a nibble-wide data interface with control and clock signals.

Pin Descriptions

The following table summarizes the MII signals, and Figure 3 on page 21 shows the proper signal connections. All signals are conveyed with positive logic.

Table 7: MII Signal Definitions

Signal Name	*Source*	*Description*
TXC	PHY	**Transmit Clock.** This clock is sourced by the PHY and is 25 MHz for 100 Mbps mode and 2.5 MHz for 10 Mbps mode. The PHY samples TXD on the rising edge of TXC.
TXEN	Switch/MAC	**Transmit Enable.** Active high single indicating that the current nibble on TXD is valid.
TXD[3:0]	Switch/MAC	**Transmit Data.** Nibble-wide transmit data.
TXER	Switch/MAC	**Transmit Error.** Active high signal instructing the PHY to transmit an errored symbol.
RXC	PHY	**Receive Clock.** This clock is sourced by the PHY and is 25 MHz for 100 Mbps mode and 2.5 MHz for 10 Mbps mode. The MAC samples RXD on the rising edge of RXC.
RXDV	PHY	**Receive Data Valid.** Active high signal indicating that the current nibble on RXD is valid.
RXD[3:0]	PHY	**Receive Data.** Nibble-wide receive data.
RXER	PHY	**Receive Error.** Active high signal instructing the MAC that a packet error has been detected.
CRS	PHY	**Carrier Sense.** Active high signal indicating activity on the media.
COL	PHY	**Collision.** Active high signal indicating that a collision has occurred on the media.
MDC	Switch/MAC	**Management Data Clock.**
MDIO	Bidirectional	**Management Data.**

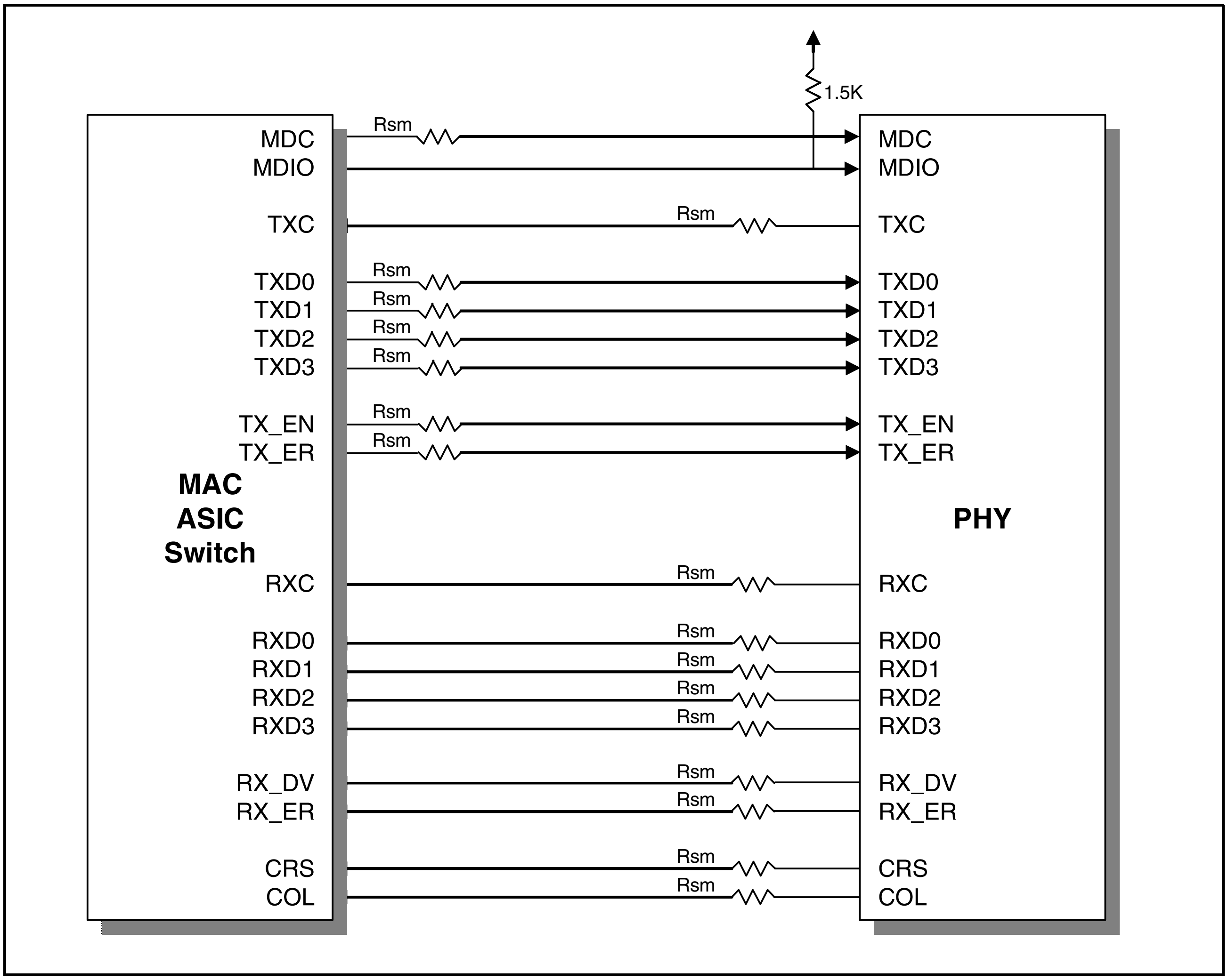

Figure 3: MII Signal Connections

REDUCED MEDIA INDEPENDENT INTERFACE

Reduced Media Independent Interface (RMII) is a reduced pin version of MII. RMII is comprised of a 2-bit-wide transmit and receive interfaces, and clocked by a single reference clock.

Pin Descriptions

The following table summarizes the RMII signals, and Figure 4 shows the proper signal connections. All signals are conveyed with positive logic.

Table 8: RMII Signal Definitions

Signal Name	*Source*	*Description*
REF_CLK	Switch/MAC or External	**Reference Clock**. This clock is sourced externally or by the MAC and is always 50 MHz. Both the MAC and the PHY have the same data and control signals on the rising edge of this clock.
TXEN	Switch/MAC	**Transmit Enable**. Active high single indicating that the current 2-bit data on TXD[1:0] is valid.
TXD[1:0]	Switch/MAC	**Transmit Data**. 2-bit transmit data bus.
CRS_DV	PHY	**Carrier Sense/Receive Data Valid**. Active high signal indicating that the media is non-idle and that the current data on the RXD[1:0] is valid.
RXD[1:0]	PHY	**Receive Data**. 2-bit receive data bus.
RXER	PHY	**Receive Error**. Active high signal instructing the MAC that a packet error has been detected.
MDC	Switch/MAC	**Management Data Clock**.
MDIO	Bidirectional	**Management Data**.

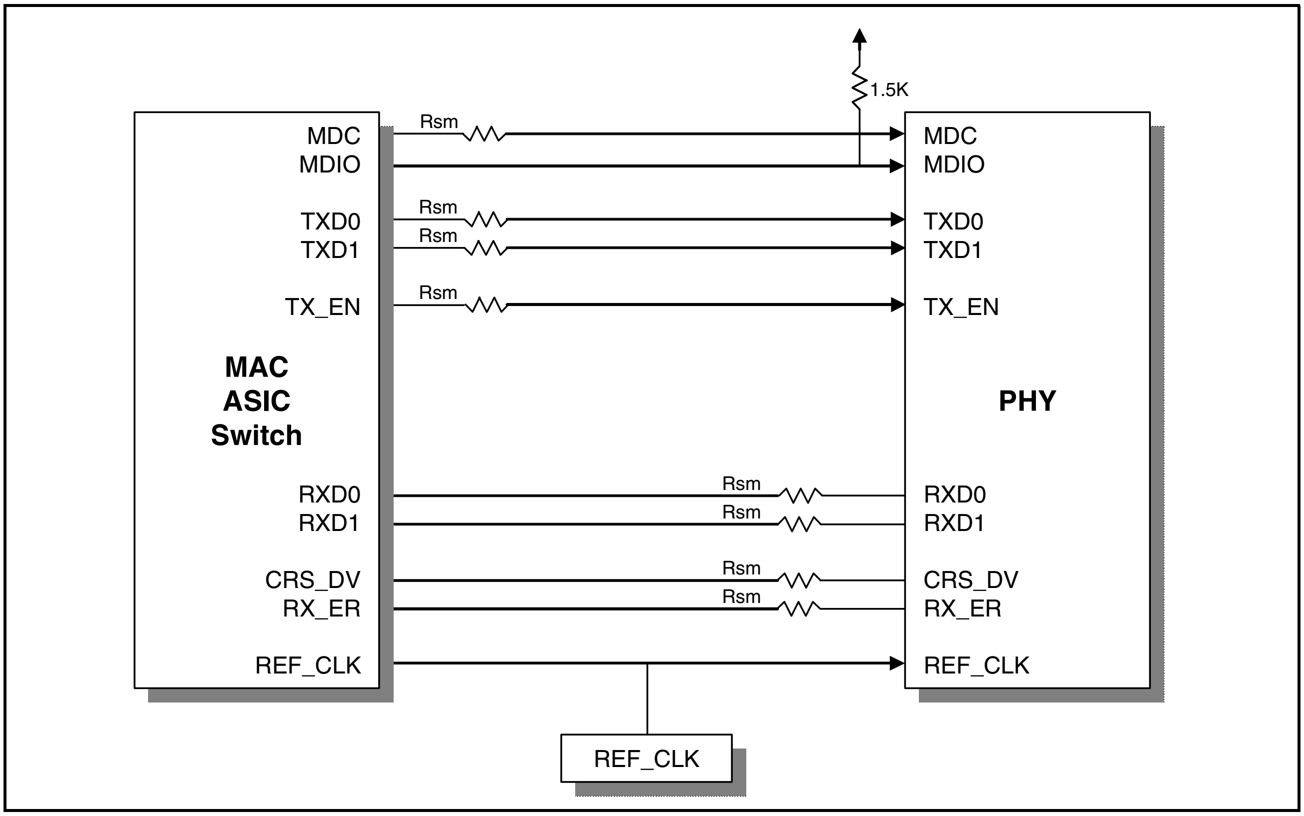

Figure 4: RMII Signal Connections

SERIAL MEDIA INDEPENDENT INTERFACE

The objective of Serial Media Independent Interface (SMII) is to reduce the number of pins required to interconnect the MAC and the PHY. This is accomplished by clocking data and control signals in and out of each PHY on a pair of pins at a rate of 125 MHz. SMII is most commonly used with multi-port PHYs (octals).

Data and control signals passing from the MAC to the PHY use the serial transmit (TXD) line, and data and control signals passing from the PHY to the MAC use the serial receive (RXD) line. All bit transfers are synchronous with clock (REF_CLK) at 125 MHz. Frame sync is provided by a fourth line (SYNC) that is asserted at the beginning of each segment, which occurs every 10 cycles of REF_CLK. Each PHY is provided with a TXD and an RXD pair.

Receive data and control information are passed from the PHY to the MAC in 10-bit frames. In 100 Mbps mode, each frame represents a new byte of data. In 10 Mbps mode, each byte of data is repeated 10 times and the MAC can sample any one of every 10 frames. Because the timing of data coming from a remote transmitter is not synchronized with the local SCLK or SYNC lines and may contain errors in frequency, a First-in First-out (FIFO) capable of storing 28 bits is provided in each receive path. The received data bits and the RX_DV signal are passed through the FIFO, but the CRS bit is not. The CRS bit is asserted for the time the wire is receiving a frame. If the remote transmitter is idle and no data needs to be passed from the receiver, status information becomes available.

Transmit data and control information are passed from the MAC to the PHY in 10-bit frames, as in the receive path. In 100 Mbit mode, each frame represents a new byte of data. In 10 Mbps mode, each byte of data is repeated 10 times and the PHY can transmit any one of every 10 frames.

Note: SMII is a Cisco Systems specification, not an IEEE specification.

Pin Descriptions

The following table summarizes the SMII signals, and Figure 5 shows the proper signal connections. All signals are conveyed with positive logic.

Table 9: SMII Signal Definitions

Signal Name	*Source*	*Description*
REF_CLK	External	**Reference Clock.** This clock is sourced externally and is always 125 MHz.
TXD	Switch/MAC	**Transmit Data.** Serial transmit data from the MAC to the PHY.
RXD	PHY	**Receive Data.** Serial receive data from the PHY to the MAC.
SYNC	Switch/MAC	**Synchronization.** Synchronization signal issued every 10 REF_CLK cycles that marks the beginning of each SMII segment.
MDC	Switch/MAC	**Management Data Clock**.
MDIO	Bidirectional	**Management Data**.

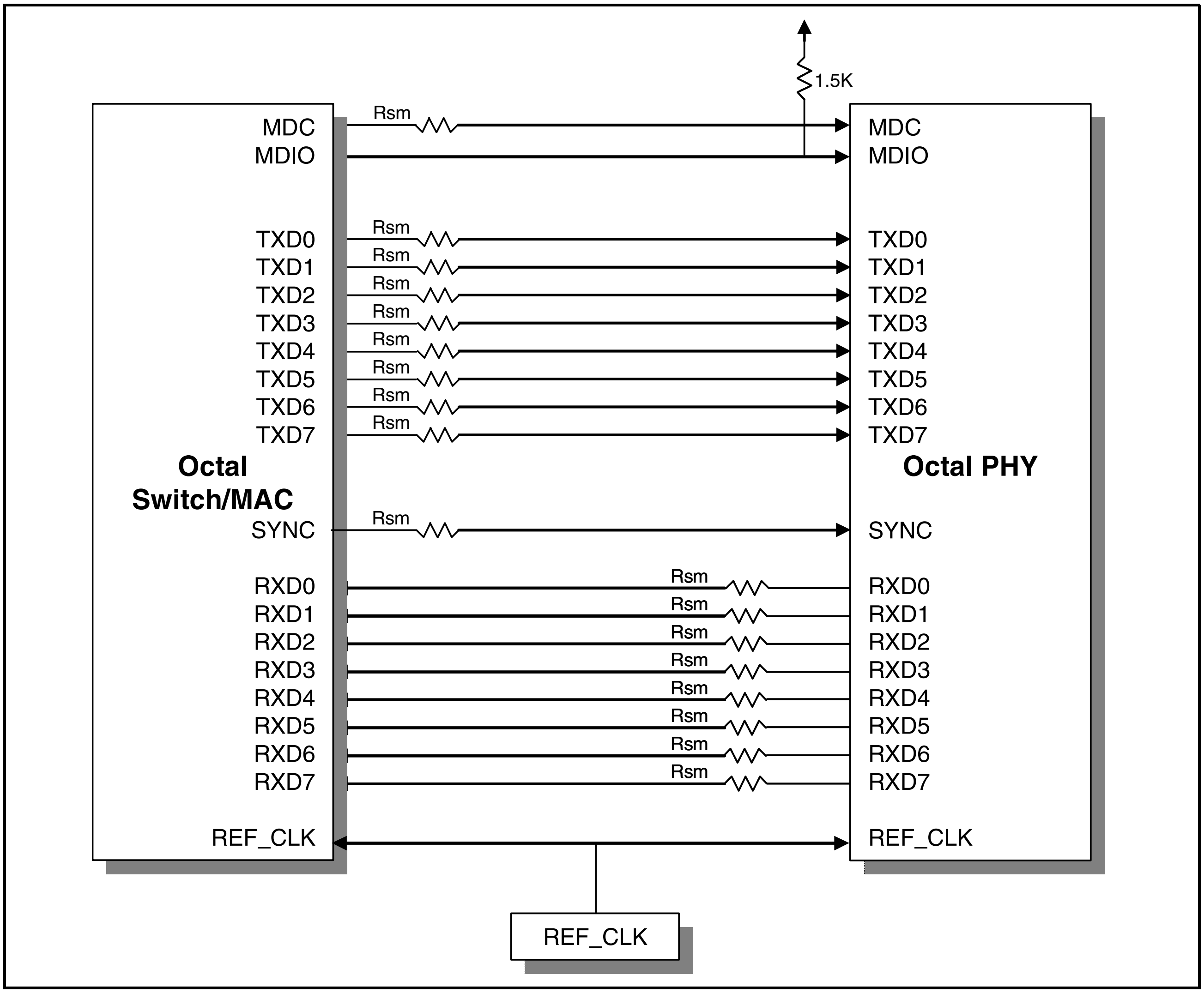

Figure 5: SMII Signal Connections

SOURCE-SYNCHRONOUS SERIAL MEDIA INDEPENDENT INTERFACE

From a data signaling standpoint, the Source-Synchronous SMII (S3MII) is essentially identical to standard SMII. The only difference is that source-synchronous employs specific 125-MHz clocks and the TXD and RXD SYNC signals that travel in the same direction as the data and are synchronous to the data. Therefore, a source-synchronous capable MAC that sends TXD to the PHY must also send a source-synchronous 125-MHz clock and SYNC signal to the PHY. The PHY uses this clock and SYNC to latch-in and delineate the TXD0 data stream.

Similarly, a PHY in S3MII mode drives RXD to the MAC along with a 125-MHz clock and SYNC signals. The MAC should use this clock and SYNC to latch-in and delineate RXD data streams. By using these separate clock and SYNC signals, SMII timing constraints are significantly eased.

Pin Descriptions

The following table summarizes the S3MII signals, and Figure 6 on page 26 shows the proper signal connections. All signals are conveyed with positive logic.

Table 10: S3MII Signal Definitions

Signal Name	*Source*	*Description*
REF_CLK	External	**Reference Clock**. This clock is sourced externally and is always 125 MHz.
TXC	Switch/MAC	**Transmit Clock.** 125-MHz transmit clock.
TXD	Switch/MAC	**Transmit Data.** Serial transmit data from the MAC to the PHY.
TSYNC	Switch/MAC	**Transmit Synchronization.** Synchronization signal issued every 10 TXC cycles that marks the beginning of each transmit S3MII segment.
RXC	PHY	**Receive Clock.** 125-MHz receive clock.
RXD	PHY	**Receive Data.** Serial receive data from the PHY to the MAC.
RSYNC	PHY	**Receive Synchronization.** Synchronization signal issued every 10 RXC cycles that marks the beginning of each receive S3MII segment.
MDC	Switch/MAC	**Management Data Clock**.
MDIO	Bidirectional	**Management Data**.

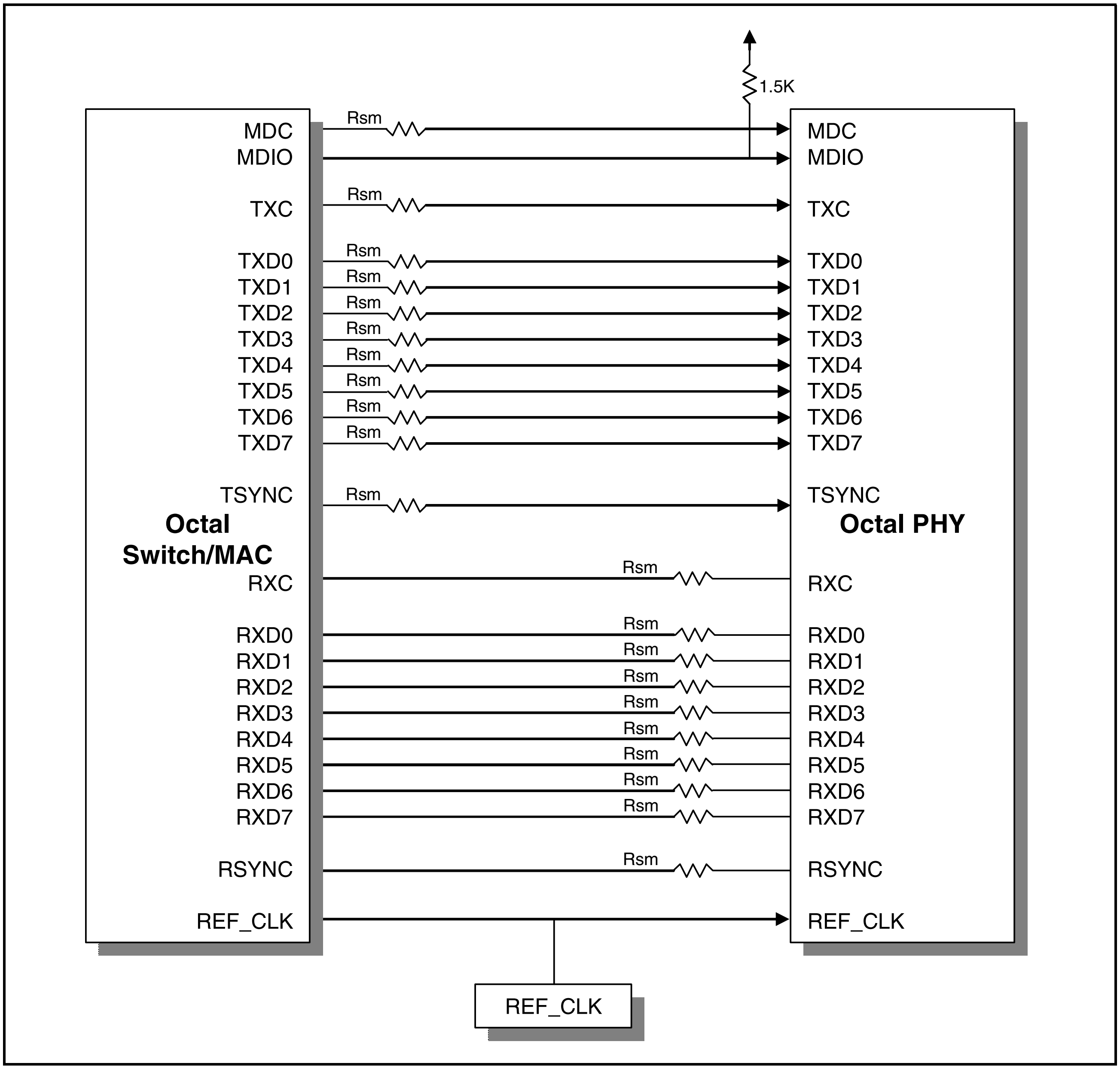

Figure 6: S3MII Signal Connections

Gigabit Media Independent Interface

IEEE802.3z introduced the Gigabit Media Independent Interface (GMII) as a wider, faster, source-synchronous version of MII.

Pin Descriptions

The following table summarizes the GMII signals, and Figure 7 on page 28 shows the proper signal connections. All signals are conveyed with positive logic.

Table 11: GMII Signal Definitions

Signal Name	*Direction (MAC)*	*Description*
TXC	Input	**Transmit Clock**. This clock is sourced by the PHY and is 25 MHz for 100 Mbps mode and 2.5 MHz for 10 Mbps mode. The PHY samples TXD on the rising edge of TXC. This clock is completely unused while in Gigabit mode.
GTXCLK	Output	**Gigabit Transmit Clock.** This 125-MHz clock is used when in Gigabit mode, and completely unused during 10/100 mode.
TXEN	Output	**Transmit Enable**. Active high single indicating that the current byte on TXD is valid.
TXD[7:0]	Output	**Transmit Data**. Byte-wide transmit data.
TXER	Output	**Transmit Error**. Active high signal instructing the PHY to transmit an errored symbol.
RXC	Input	**Receive Clock**. This clock is sourced by the PHY and is 125 MHz for 1000 Mbps mode, 25 MHz for 100 Mbps mode, and 2.5 MHz for 10 Mbps mode. The MAC samples RXD on the rising edge of RXC.
RXDV	Input	**Receive Data Valid**. Active high signal indicating that the current byte on RXD is valid.
RXD[3:0]	Input	**Receive Data**. Byte-wide receive data.
RXER	Input	**Receive Error**. Active high signal instructing the MAC that a packet error has been detected.
CRS	Input	**Carrier Sense**. Active high signal indicating activity on the media.
COL	Input	**Collision**. Active high signal indicating that a collision has occurred on the media.
MDC	Output	**Management Data Clock**.
MDIO	Bidirectional	**Management Data**.

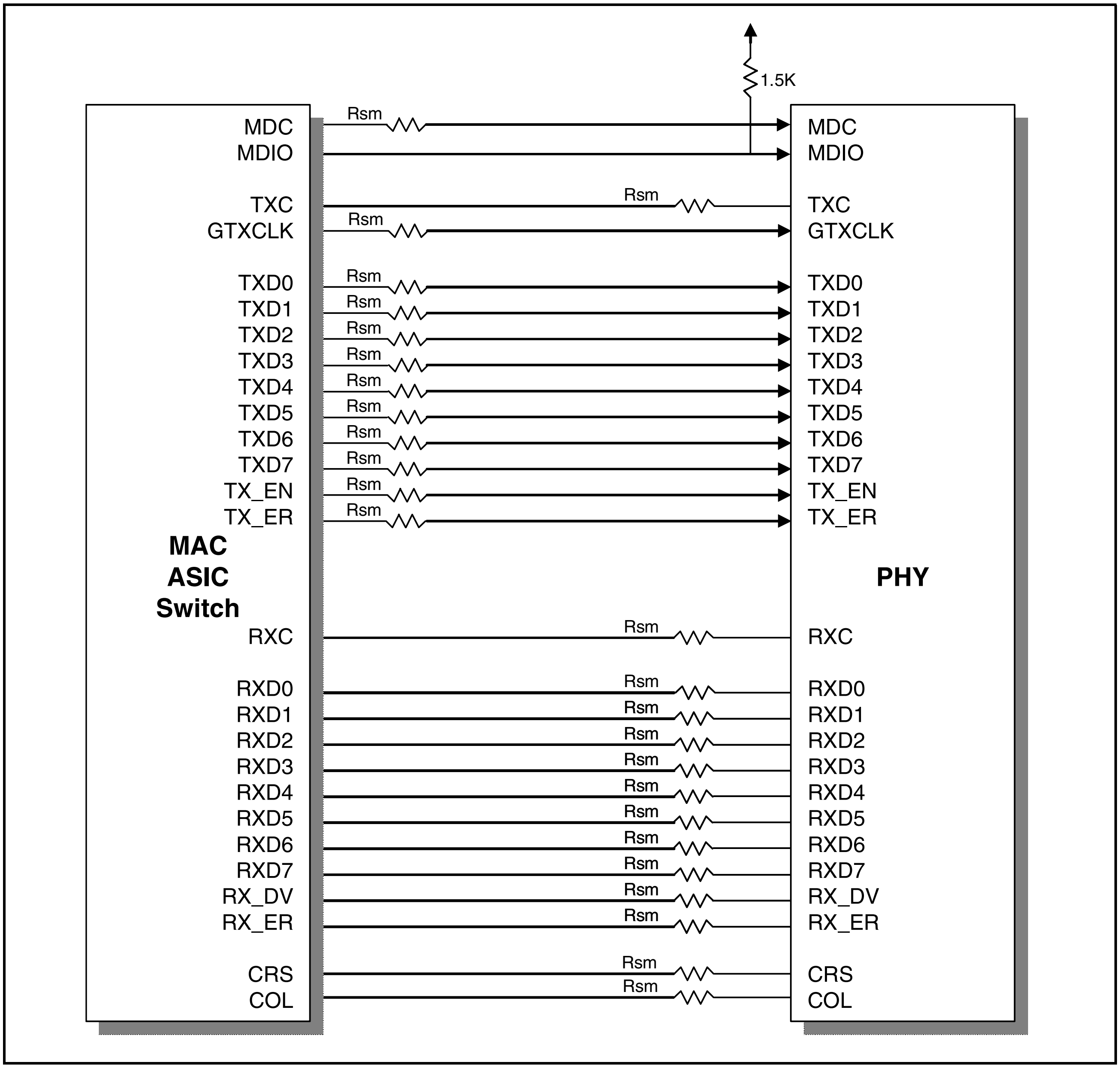

Figure 7: GMII Signal Connections

REDUCED GIGABIT MEDIA INDEPENDENT INTERFACE

The Reduced Gigabit Media Independent Interface (RGMII) interface is intended to be an alternative to the IEEE 802.3u MII, IEEE 802.3z GMII, or 10-Bit Interface (TBI). The principle objective is to reduce the number of pins required to interconnect the Switch/MAC and the PHY from a maximum of 27 pins (TBI) to 12 pins in a cost-effective and technology-independent manner. To accomplish this, the data path widths have been reduced, the control signals are multiplexed together, and both edges of the 125-MHz clock are used.

Pin Descriptions

Table 12 summarizes the RGMII signals, and Figure 8 on page 30 shows the proper signal connections. Many signals have two functions provided on the same pin. One function is provided on the rising edge of the clock, while the other function is provided on the falling edge of the clock. All signals are conveyed with positive logic. Data is sent least significant nibble first. Bytes are reconstructed by sampling with the rising edge first followed by the subsequent falling edge of the clock. For more details on the RGMII specification, refer to the *Reduced Gigabit Media Independent Interface (RGMII) Version 2.0* document (this document can be found on the HP website at http://www.hp.com/rnd/library/a-z_index.htm).

Table 12: RGMII Signal Definitions

Signal Name	*Source*	*Description*
GTXCLK	Switch/MAC	**Transmit Clock**. This clock is 125 MHz in 1000BASE-T mode, 25 MHz in 100BASE-TX mode, and 2.5 MHz in 10BASE-T mode.
TXEN (TXEN/TXER)	Switch/MAC	**Transmit Control**. This signal has a number of functions multiplexed onto it. • TXEN = Transmit Enable. This information is presented on the rising edge of the transmit clock in RGMII mode. • TXER = Transmit Error. This information is presented on the falling edge of the transmit clock in RGMII mode.
TXD[3:0]	Switch/MAC	**Transmit Data.** Data bits TXD[3:0] are presented on the rising edge of the transmit clock in both RGMII. Data bits TXD[7:4] are presented on the falling edge of the transmit clock in RGMII mode.
RXC	PHY	**Receive Clock**. This clock is 125 MHz in 1000BASE-T mode, 25 MHz in 100BASE-TX mode, and 2.5 MHz in 10BASE-T mode.
RXDV (RXDV/RXER)	PHY	**Receive Control.** This signal has a number of functions multiplexed onto it. • RXDV = Receive Data Valid. This information is presented on the rising edge of the receive clock in RGMII mode. • RXER = Receive Error. This information is presented on the falling edge of the receive clock in RGMII mode.
RXD[3:0]	PHY	**Receive Data.** Data bits RXD[3:0] are presented on the rising edge of the receive and data bits RXD[7:4] are presented on the falling edge of the receive clock.
MDC	Switch/MAC	**Management Data Clock**.
MDIO	Both	**Management Data**.

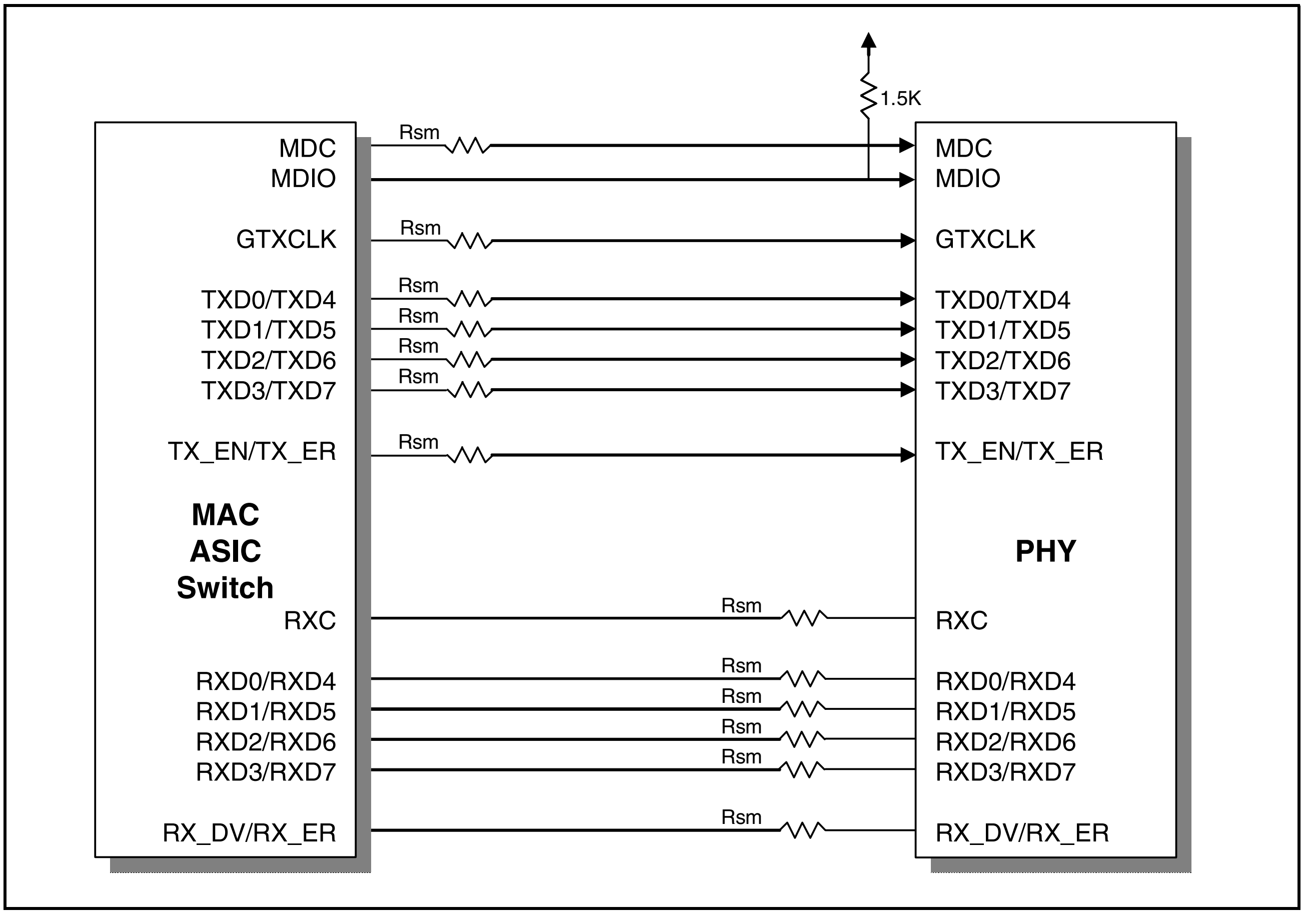

Figure 8: RGMII Signal Connections

Serial Gigabit Media Independent Interface

The Serial Gigabit Media Independent Interface (SGMII) is intended to be an alternative to the IEEE 802.3u MII, IEEE 802.3z GMII, or TBIs. The principle objective is to reduce the number of pins required to interconnect the Switch/MAC and the PHY from a maximum of 27 pins (TBI) to 6 pins in a cost-effective and technology independent manner. To accomplish this, the data path widths have been reduced, the control signals are encoded in the data stream, and both edges of the differential 625-MHz clock are used.

The SGMII uses two data signals and one clock signal to convey frame data and link rate information between the PHY and the Switch/MAC. The data signals operate at 1.25 Gbd, and the clocks operate as a 625-MHz double-data rate (DDR) interface. Each of these signals is realized as a differential pair because of the speed of operation, providing signal integrity while minimizing system noise.

The 1.25 Gbd transfer rate of the SGMII is greater than required for the PHY operating at 10 Mbps or 100 Mbps. When these situations occur, the PHY elongates the frame by replicating each frame byte 10 times for 100 Mbps and 100 times for 10 Mbps. This frame elongation takes place above the 802.3z PCS layer, making the start frame delimiter appear only once per frame.

Pin Descriptions

The following table summarizes the six SGMII signals, and Figure 9 shows the proper signal connections.

Table 13: SGMII Signal Definitions

Signal Name	*Source*	*Description*
SGIN±	Switch/MAC	**Transmit data**. Differential 1.25-Gbd transmit data from the Switch/MAC to the PHY.
SGOUt±	PHY	**Receive data**. Differential 1.25-Gbd receive data from the PHY to the Switch/MAC.
SCLK±	PHY	**Transmit clock.** Differential 625-MHz clock synchronized to the SGOUT± data. Some Switch/MACs can recover the clock from the SGOUT± data and do not need the SCLK±. If this is the case, then these pins can be left floating.

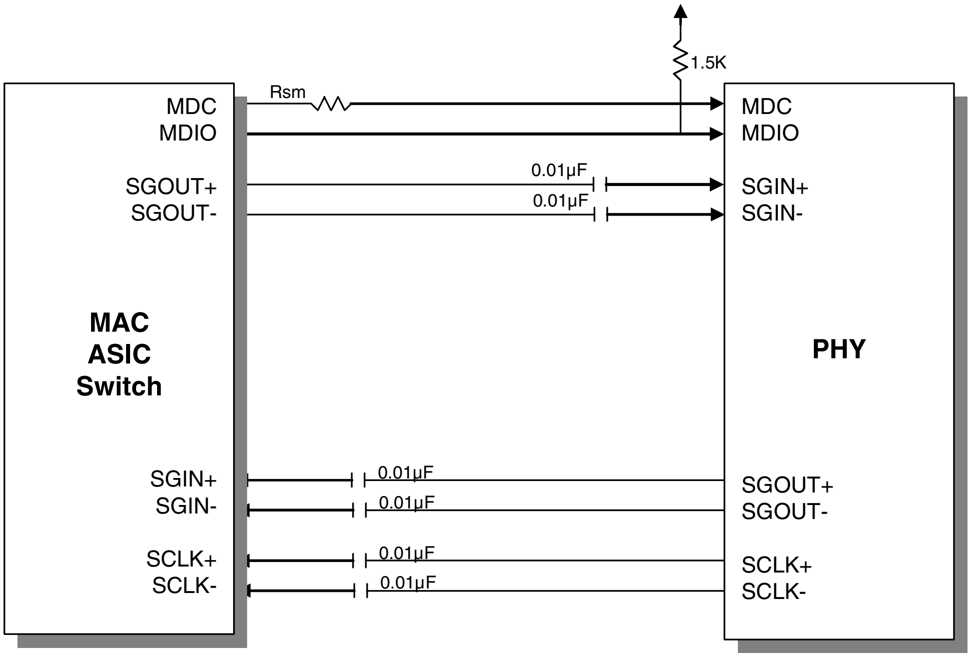

Figure 9: SGMII Signal Connections

10-BIT INTERFACE

This section describes the TBI, based on the industry-standard SerDes interface, usually used to transfer 10-bit encoded data between the MAC and a Gigabit Ethernet optical-fiber device in 1000BASE-SX/LX applications. When operating in TBI mode, the PHY speed is limited to 1000 Mbps. An MII is not available when operating in TBI mode.

Pin Descriptions

The following table summaries the TBI signals, and Figure 10 shows the proper signal connections.

Table 14: TBI Signal Definitions

Signal Name	*Source*	*Description*
GTXCLK	Switch/MAC	**Transmit Clock.** This 125-MHz clock is used to clock data from the Switch/MAC into the PHY.
TXD[9:0]	Switch/MAC	**Transmit Data.** 10-bit wide transmit data valid in the rising edge of GTXCLK.
RBC[1:0]	PHY	**Receive Clocks.** The receive clocks are 62.5 MHz and are derived from the received data stream. The two clocks are complements.
RXD[9:0]	PHY	**Receive Data.** Receive data bits [9:0] are clocked on the rising edge of both RBC0 and RBC1.

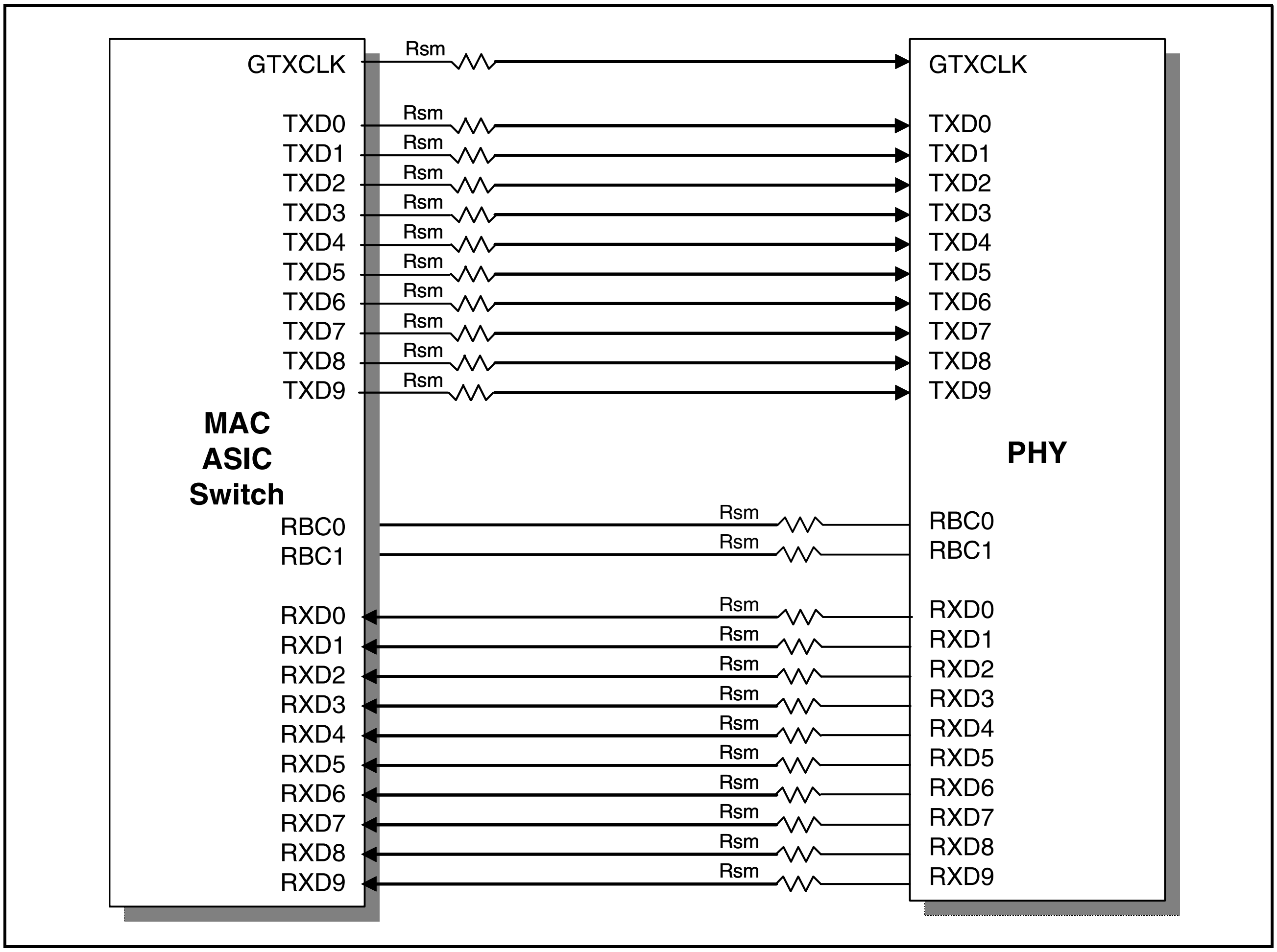

Figure 10: TBI Signal Connections

Reduced 10-Bit Interface

The Reduced 10-Bit Interface (RTBI) is intended to be an alternative to the TBI specification. The RTBI is derived from the RGMII specification. The RTBI shares four data path signals with the RGMII and control functionality with the fifth data signal. With the inclusion of the MDIO/MDC serial management signals, the RTBI does not require independent control signals like LK_REF, BYTE_EN, and so on. The principle objective is to reduce the number of pins required to interconnect the Switch/MAC and the PHY from a maximum of 27 pins (TBI) to 12 pins in a cost-effective and technologically independent manner. To accomplish this, the data path widths have been reduced, the control signals are multiplexed together, and both edges of the 125-MHz clock are used.

Pin Descriptions

Table 15 summarizes the RTBI signals, and Figure 11 on page 34 shows the proper signal connections. Many signals have two functions provided on the same pin—one function on the rising edge of the clock and the other function on the falling edge of the clock.

All signals are conveyed with positive logic. Data is sent least significant nibble first. Bytes are reconstructed by sampling with the rising edge first, followed by the subsequent falling edge of the clock.

Table 15: RTBI Signal Definitions

Signal Name	*Source*	*Description*
GTXCLK	Switch/MAC	**Transmit Clock.** This double-data clock is 125 MHz in 1000BASE-T mode.
TXD[3:0] TXD[8:5]	Switch/MAC	**Transmit Data.** Transmit data bits [3:0] are clocked on the rising edge of GTXCLK and transmit data bits [8:5] are clocked on the falling edge of GTXCLK.
TXD[4] TXD[9]	Switch/MAC	**Transmit Control.** Contains the fifth transmit data bit on the rising edge of GTXCLK and the tenth receive data bit on the falling edge of GTXCLK.
RXC	PHY	**Receive Clock.** The double data receive clock is 125 MHz for 1000BASE-T mode. This clock is derived from the received data stream.
RXD[4] RXD[9]	PHY	**Receive Control.** Contains the fifth receive data bit on the rising edge of RXC and the tenth receive data bit on the falling edge of RXC.
RXD[3:0] RXD[8:5]	PHY	**Receive Data.** Receive data bits [3:0] are clocked on the rising edge of RXC, and received data bits [8:5] are clocked on the falling edge of RXC.
MDC	Switch/MAC	**Management Data Clock.**
MDIO	Both	**Management Data.**

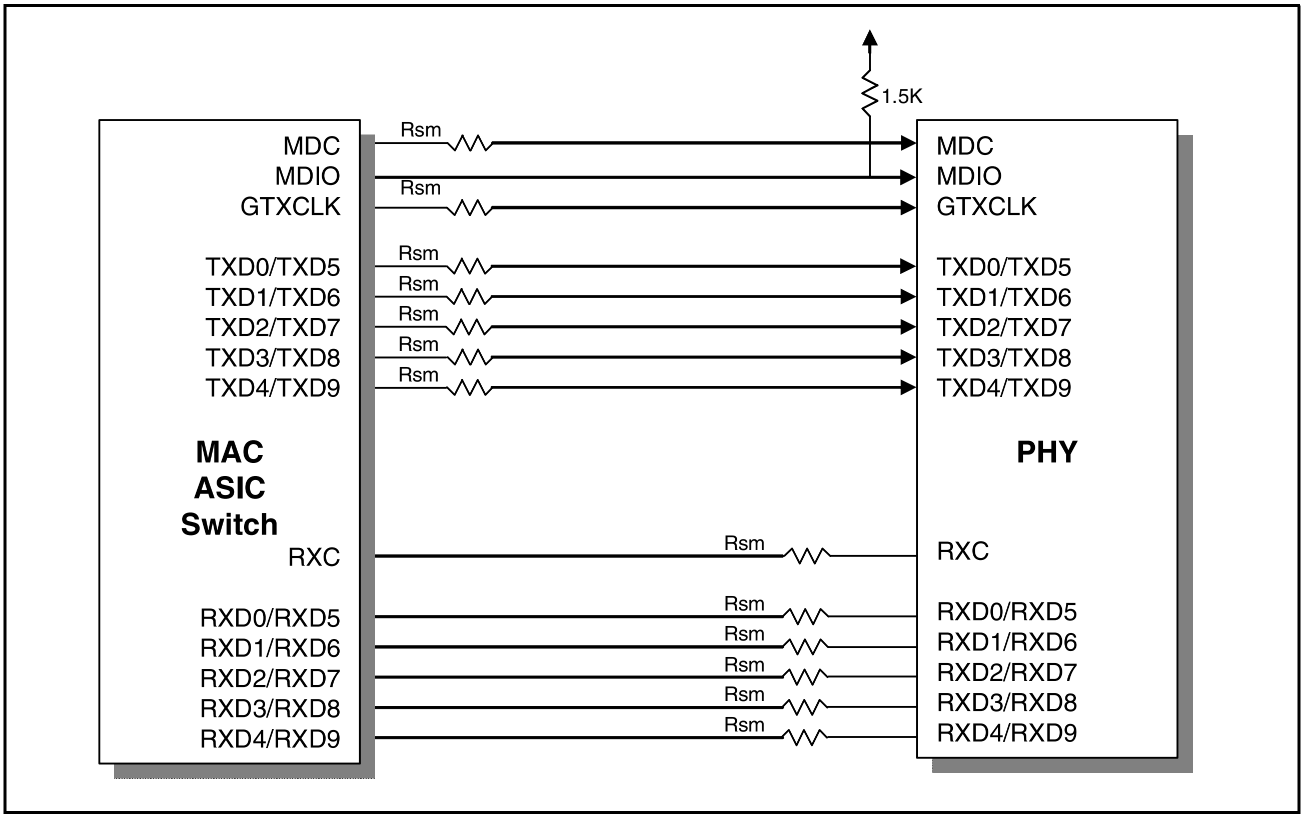

Figure 11: RTBI Signal Connections

10-Gigabit Media Independent Interface

The 10-Gigabit Media Independent Interface (XGMII) is defined by the IEEE802.3ae (10-Gigabit Ethernet) specification. The purpose of the XGMII is to provide a simple, inexpensive, and easy-to-implement interconnection between the MAC and the PHY. The XGMII is a source-synchronous, parallel data, and control interface capable of 10 Gbps in both the transmit and receive directions simultaneously (full-duplex).

Pin Descriptions

Table 16 summarizes the XGMII signals, and Figure 12 on page 36 shows the proper XGMII signal connections. Data is organized into byte lanes and octets are assigned to each lane with a round robin approach. Both transmit data, receive data, and their respective control data are sampled on both edges of their respective clock.

Table 16: XGMII Signal Definitions

Signal Name	*Source*	*Description*
TX_CLK	Switch/MAC	**Transmit Clock.** 156.25-MHz transmit clock.
TXC[3:0]	Switch/MAC	**Transmit Control.** 4-bit transmit control bus (one bit per lane) that indicates whether the current character on the transmit data bus is a control or data character. Transmit control is sampled on both the rising and falling edge of TX_CLK.
TXD[31:0]	Switch/MAC	**Transmit Data.** 32-bit transmit data bus. This 32-bit bus is actually four 8-bit buses with each of the 8-bits representing one octet lane. Transmit data is sampled on both the rising and falling edge of TX_CLK.
RX_CLK	PHY	**Receive Clock.** 156.25-MHz receive clock.
RXC[3:0]	PHY	**Receive Control.** 4-bit receive control bus (one bit per lane) that indicates whether the current character on the receive data bus is a control or data character. Receive control is sampled on both the rising and falling edge of RX_CLK.
RXD[31:0]	PHY	**Receive Data.** 32-bit receive data bus. This 32-bit bus is actually four 8-bit buses with each of the 8-bits representing one octet lane. Receive data is sampled on both the rising and falling edge of RX_CLK.
MDC	Switch/MAC	**Management Data Clock.**
MDIO	Both	**Management Data.**

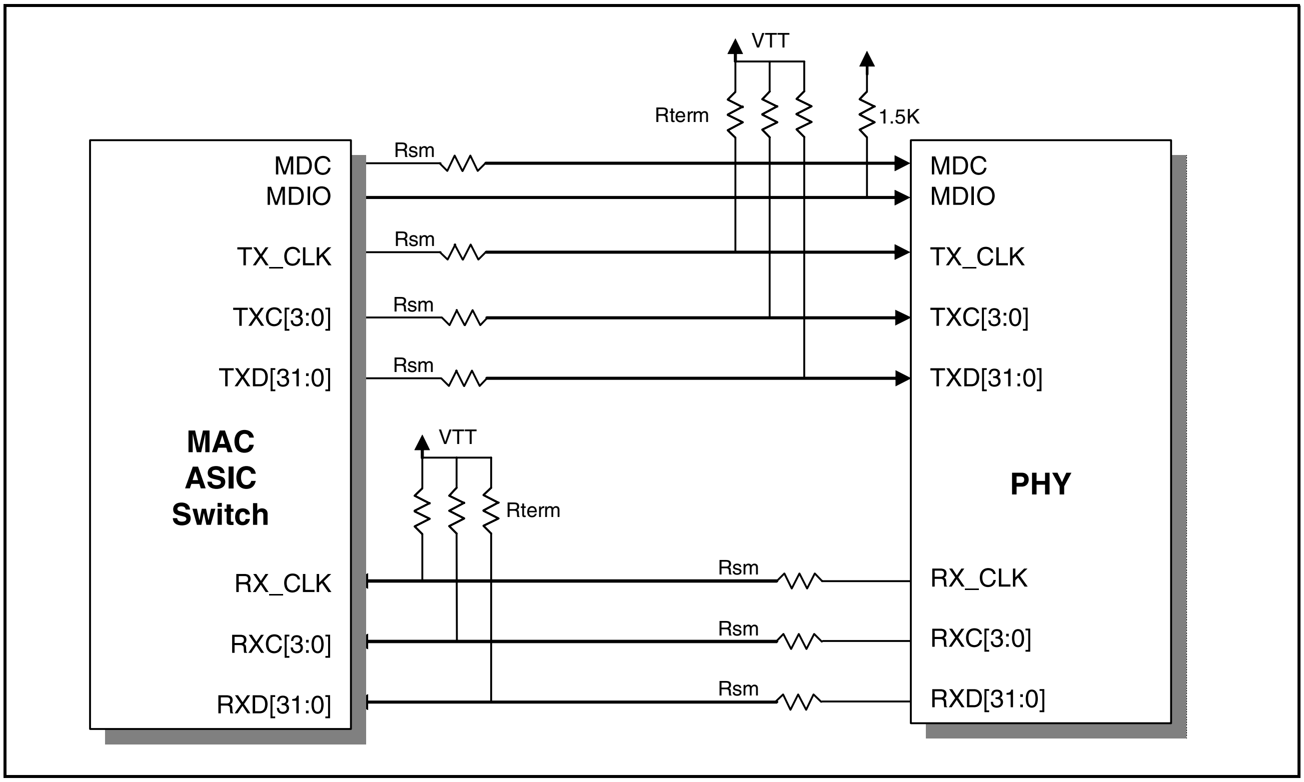

Figure 12: XGMII Signal Connections

10-Gigabit Attachment Unit Interface

The 10-Gigabit Attachment Unit Interface (XAUI, pronounced “zowie”) is defined by IEEE802.3ae specification as an interface extender for XGMII. XAUI supports 10 Gbps using four transmit and four receive lanes. Each lane encodes data with an 8B/10B code for differential serial transmission and operating at 3.125 Gbd. XAUI reduces 10-Gigabit Ethernet's 72-pin XGMII to 16 pins, enabling higher density and lower cost switch chips and optical transceivers. The lower pin count and longer trace lengths allow a single chip to support multiple 10-Gigabit Ethernet ports.

In addition to the reduced pin counts, XAUI provides other system benefits such as its inherently low EMI, compensation for multi-bit bus skew, fault isolation capabilities, and relatively low-power consumption.

Pin Descriptions

The following table summarizes the XAUI signals, and Figure 13 shows the proper signal connections.

Table 17: XAUI Signal Definitions

Signal Name	*Source*	*Description*
TXD0±	Switch/MAC	**Lane 0 Transmit Data.** Differential transmit lane running at 3.125 Gbps.
TXD1±	Switch/MAC	**Lane 1 Transmit Data.** Differential transmit lane running at 3.125 Gbps.
TXD2±	Switch/MAC	**Lane 2 Transmit Data.** Differential transmit lane running at 3.125 Gbps.
TXD3±	Switch/MAC	**Lane 3 Transmit Data.** Differential transmit lane running at 3.125 Gbps.
RXD0±	Switch/MAC	**Lane 0 Receive Data.** Differential receive lane running at 3.125 Gbps.
RXD1±	Switch/MAC	**Lane 1 Receive Data.** Differential receive lane running at 3.125 Gbps.
RXD2±	Switch/MAC	**Lane 2 Receive Data.** Differential receive lane running at 3.125 Gbps.
RXD3±	Switch/MAC	**Lane 3 Receive Data.** Differential receive lane running at 3.125 Gbps.

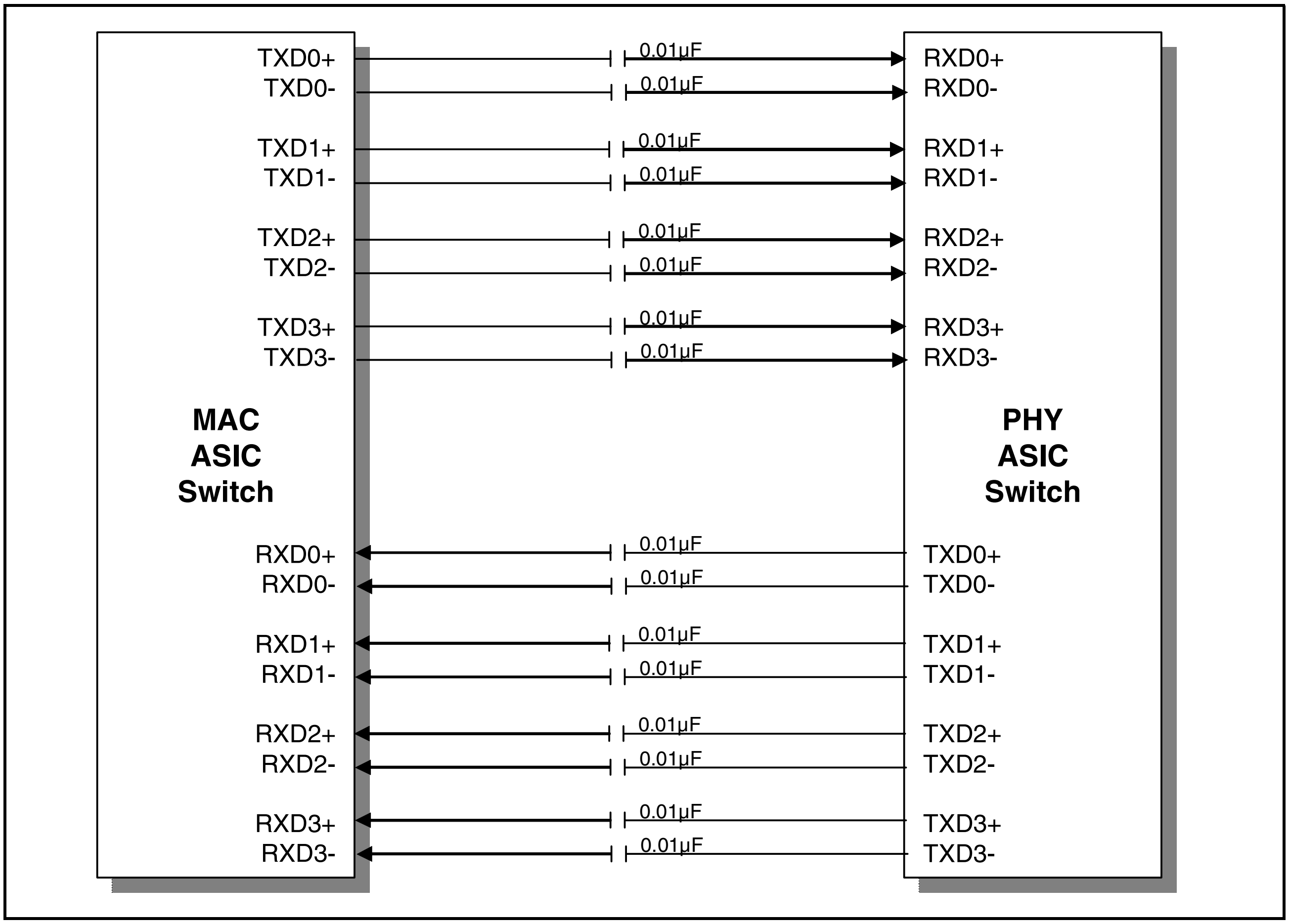

Figure 13: XAUI Signal Connections

HiGig—Broadcom 10-Gigabit Stacking Interface

The HiGig interface is electrically identical to standard IEEE802.3ae XAUI previously discussed, but HiGig does not directly support 10-Gigabit Ethernet due to its use of a proprietary header inserted into the preamble and inter-packet gap of a standard 10-Gigabit Ethernet packet. This header allows StrataXGS devices to communicate stacking and other important information across modules or individual switch chips. For signal descriptions and a connection diagram, see “10-Gigabit Attachment Unit Interface” on page 37.

HiGig+—Broadcom 12-Gigabit Stacking Interface

HiGig+ is the exact same concept as HiGig, with the exception that HiGig+ is also capable of 12-Gbps data rates in addition to 10-Gbps data rates. While in HiGig+ mode, data is sent using four transmit and four receive lanes. Each lane encodes data with an 8B/10B code for differential serial transmission and operating at 3.75 Gbd.

Media Technologies

This section describes several of the common Ethernet media technologies in the industry today.

100BASE-TX

100BASE-TX is defined by IEEE802.3u as 100 Mbps, full-duplex capable Ethernet transmission over two pairs of Category 5 UTP cable. The other two pairs of the cable are typically unused. 100BASE-TX uses 4b/5b encoding to guarantee clock recovery and data integrity and Multi-Level Transmission-3 (MLT-3) line encoding. Maximum cable length is 100m.

100BASE-FX

100BASE-FX is defined by IEEE802.3u as 100 Mbps, full-duplex capable Ethernet transmission over multimode fiber. 100BASE-FX uses 4b/5b encoding to guarantee clock recovery and data integrity and Non-Return to Zero Inverted (NRZI) line encoding. Maximum cable length is 100m.

1000BASE-SX

1000BASE-SX is defined by IEEE802.3z as 1000 Mbps, full-duplex capable Ethernet transmission over multimode fiber. 1000BASE-SX uses 8b/10b encoding to guarantee clock recovery and data integrity and Non-Return to Zero (NRZ) line encoding. Maximum cable length is 550m (depends on type of fiber used).

1000BASE-LX

1000BASE-LX is defined by IEEE802.3z as 1000 Mbps, full-duplex capable Ethernet transmission over single-mode fiber. 1000BASE-SX uses 8b/10b encoding to guarantee clock recovery and data integrity and NRZ line encoding. Maximum cable length is 5Km (depends on the type of fiber used).

1000BASE-CX

1000BASE-CX is defined by IEEE802.3z as 1000 Mbps, full-duplex capable Ethernet transmission over twinax (shielded and balanced copper cable). 1000BASE-CX uses 8b/10b encoding to guarantee clock recovery and data integrity and NRZ line encoding. Maximum cable length is 25m.

1000BASE-T

1000BASE-T is defined by IEEE802.3ab as 1000 Mbps, full-duplex capable Ethernet transmission over four pairs of Category 5e UTP cable. 1000BASE-T transmits and receives on all four pairs simultaneously, forcing the PHY to implement Near-End Cross-Talk (NEXT) cancellation and echo cancellation. 1000BASE-T uses Four-Dimensional/Pulse-Amplitude-Modulation (4D-PAM5) encoding. The maximum cable length is 100m.

ANSI 1000BASE-TX

The ANSI/TIA-854 standard provides a data rate of 1000 Mbps similar to the IEEE 802.3ab Gigabit Ethernet standard. The main difference between the two standards is that 1000BASE-TX requires Category 6 cabling instead of Category 5e cabling. Because of the improved performance of Category 6 cabling, the ANSI/TIA-854 standard also does not implement NEXT cancellation or echo cancellation. Since Category 5e installation is so widespread, it is unlikely that this specification will ever be as popular as IEEE802.3ab 1000BASE-T.

CX-4

CX-4 (also know as 10GBASE-CX4) is defined by IEEE803.2ak as 10-Gbps, full-duplex Ethernet transmission over 4x Infiniband cable. Basically, CX-4 is XAUI over shielded copper cable. The maximum cable length is 15m. CX-4 stands to gain popularity as a cost-effective short-reach interconnect between networking elements.

GIGABIT INTERFACE CONVERTER

A Gigabit Interface Converter (GBIC) is a transceiver that converts serial electric signals to serial optical signals. These became popular in the Fibre Channel market and were later adapted into the Gigabit Ethernet market space. Because they are pluggable and hot-swappable, they make media changes quick and easy.

Dimensions	*Inches*
Height	0.47
Width	1.2
Length	2.2

Figure 14: Typical GBIC

SMALL FORM FACTOR TRANSCEIVER

The Small Form Factor (SFF) Multi-Source Agreement (MSA) defines a standard, 2 row by 5-pin configuration for reduced size, non-changeable fiber optic modules. These are commonly used in all sorts of fiber optic networking applications including Gigabit Ethernet and Fibre Channel.

Dimensions	*Inches*
Height	0.39
Width	0.54
Length	2

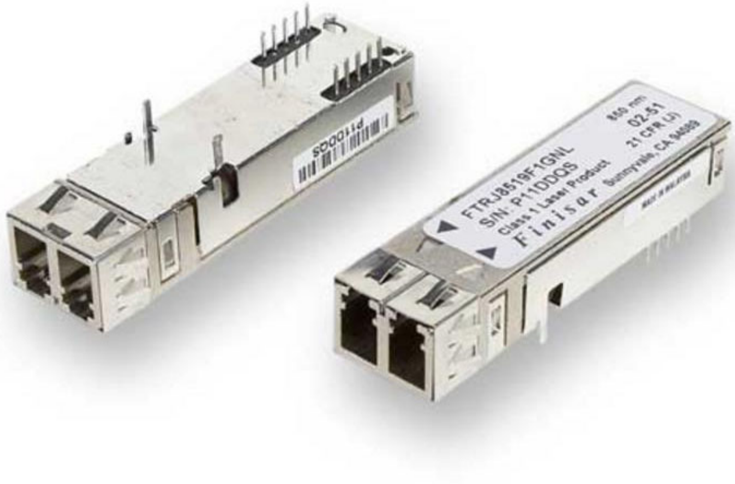

Figure 15: Typical SFF

SFF PLUGGABLE TRANSCEIVER

The SFF Pluggable (SFP) MSA takes the SFF one step further by defining a hot-pluggable version. When using an SFP, a small SFP cage and connector are soldered to the PCB. The transceiver itself then plugs into the cage, similar to the way a GBIC works. This allows the media type to be field changeable between 1000BASE-SX, 1000BASE-LX, and 1000BASE-T.

Dimensions	*Inches*
Height	0.3
Width	0.65
Length	2

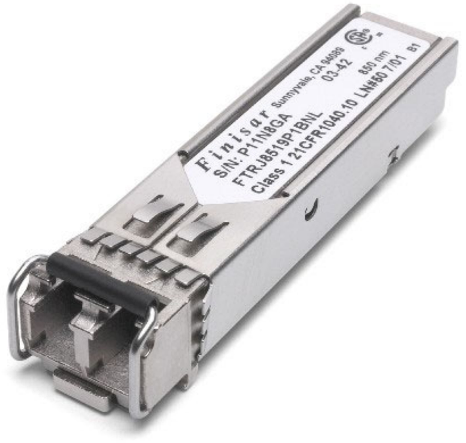

Figure 16: Typical SFP

XENPAK

The XENPAK MSA is defined as a fiber-optic transceiver module that conforms to the IEEE 802.3ae specification. The XENPAK module has a XAUI, is hot pluggable, and has an industry standard connector and foot print.

Dimensions	*Inches*
Height	0.88
Width	1.6
Length	5

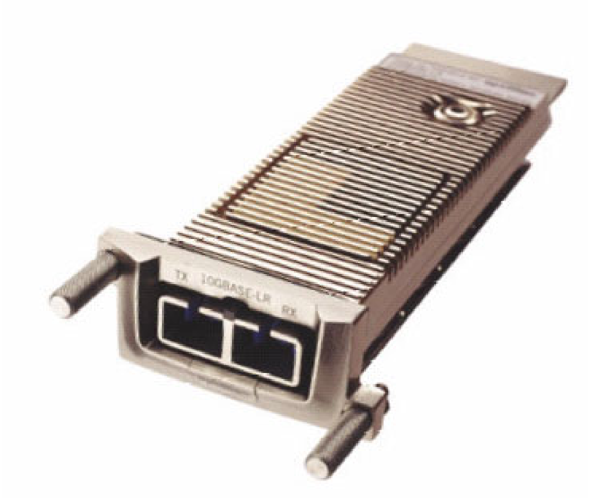

Figure 17: Typical XENPAK

10-Gigabit SFF Pluggable

The 10-Gigabit SFF Pluggable (XFP) MSA is defined as an industry standard, small, hot pluggable 10-Gbps transceiver. The technology is intended to be flexible enough to support OC192/STM-64, 10 G Fibre Channel, G.709, and 10 G Ethernet, usually with the same module. XFP modules are currently the most cost-effective solution for 10-Gbps fiber applications, but producing one with long haul optics has proven to be a challenge due to the small size of the module.

Dimensions	*Inches*
Height	0.33
Width	0.68
Length	2.3

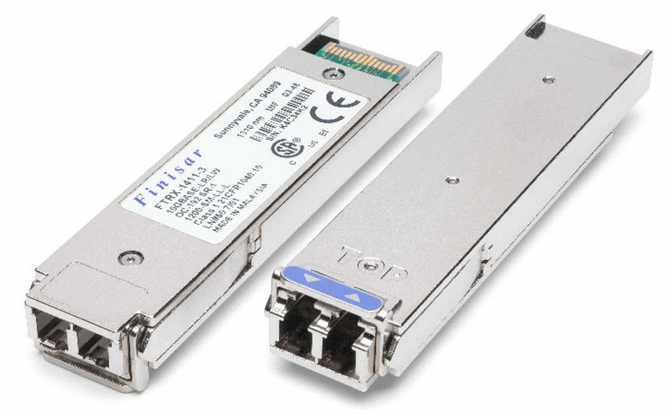

Figure 18: Typical XFP

OTHER NETWORKING TERMS

ASYMMETRIC DIGITAL SUBSCRIBER LINE

Asymmetric Digital Subscriber Line (ADSL) is a technology that allows more data to be sent over existing copper telephone lines, and is therefore intended for the last leg into a customer's premises. As its name implies, ADSL transmits an asymmetric data stream, with the downstream direction being the higher speed connection.

The following table shows the ADSL range of downstream speeds, which depends on the distance from the central office Digital Subscriber Line Access Multiplexer (DSLAM).

Table 18: ADSL Downstream Speeds

Distance from DSLAM	Data Rate	Equivalent WAN Interface Speed
Up to 18,000 feet	1.544 Mbps	T1
16,000 feet	2.048 Mbps	E1
12,000 feet	6.312 Mbps	DS2
9,000 feet	8.448 Mbps	N/A

Upstream speeds range from 16 kbps to 640 kbps. Individual products today incorporate a variety of speed arrangements, from a minimum set of 1.544/2.048 Mbps down and 16 kbps up, to a maximum set of 9 Mbps down and 640 kbps up. All of these arrangements operate in a frequency band above plain old telephone service (POTS), leaving POTS service independent and undisturbed, even if a premises ADSL modem fails.

On the surface, it may seem that ADSL has little to do with Ethernet. As more and more DSLAM venders abandon ATM on the uplink and transition to IP, Ethernet has made the transition as well. It is common for Ethernet to exist as the link between the Digital Subscriber Line (DSL) line cards and the DSLAM switch and as the uplink.

VERY HIGH DATA RATE DIGITAL SUBSCRIBER LINE

Very high data rate Digital Subscriber Line (VDSL) is essentially ADSL, but over shorter lines at higher rates. While no general standards exist yet for VDSL, the following table discusses some downstream speeds.

Table 19: VDSL Downstream Speeds

Distance from DSLAM	Data Rate	Equivalent WAN Interface Speed
4,500 feet	12.96 Mbps	1/4 STS-1
3,000 feet	25.82 Mbps	1/2 STS-1
1,000 feet	51.84 Mbps	STS-1

Upstream rates fall within a suggested range from 1.6 Mbps to 2.3 Mbps. In many ways VDSL is simpler than ADSL. Shorter lines impose far fewer transmission constraints, so the basic transceiver technology is much less complex, even though it is 10 times faster. VDSL only targets ATM network architectures, obviating channelization and packet handling requirements imposed on ADSL. And VDSL admits passive network terminations, enabling more than one VDSL modem to be connected to the same line at a customer site, in much the same way as extension phones connect to home wiring for POTS.

PROTOCOL HEADER FORMATS

IPv4

4 Bits	*8 Bits*	*16 Bits*	*32 Bits*	
Version	IHL	ToS	Total Length	
Identification			Flags	Fragment Offset
TTL		Protocol	Header Checksum	
Source Address				
Destination Address				
Option + Padding				
Data				

Version

Indicates the version of the IP header. In this case, the version is 4.

Internet Header Length

The Internet Header Length (IHL) defines the total length (in 32-bit words) of the IP header. The minimum value for a IPv4 header without options is 5.

Type of Service

An 8-bit QoS indication.

Bits 0-2: Precedence

111	Network control.
110	Internetwork control.
101	CRITIC/ECP.
100	Flash override.
011	Flash.
010	Immediate.
001	Priority.
000	Routine.

Bit 3: Delay

0	Normal delay.
1	Low delay.

Bit 4: Throughput

0	Normal throughput.
1	High throughput.

Bit 5: Reliability

0	Normal reliability.
1	High reliability.

Bits 6-7: Reserved

Reserved for future use.

Total Length

Total length of the datagram in bytes. This count includes everything from the first byte of the IP header to the last byte of the IP datagram, regardless of the datagram content. The maximum length is 65,535.

Identification

The ID assigned by the sender to aid in assembling the fragments of a datagram.

Flags

3 bits. Control flags.

Bit 0: Reserved

Bit 0 is reserved and must be zero.

Bit 1: Don't Fragment

0	May fragment.
1	Don't fragment.

Bit 2: More Fragments Bit

0	Last fragment.
1	More fragments.

Fragment Offset

This 13-bit field indicates where this fragment belongs in the datagram. The fragment offset is measured in units of 8 bytes (64 bits). The first fragment has offset zero.

Time-to-Live

Maximum amount of time (in seconds) the datagram can live on the network. If this field reaches zero, the datagram must be removed from the network. The TTL is decremented at each hop allow the datagram route.

Protocol

Indicates the next protocol to follow the IP header.

Example: If protocol = 0x06, then TCP is the next protocol in the packet.

Header Checksum

Checksum (*not* a CRC) of the current IP header only. Each time the header is modified (i.e. TTL updated), the checksum must be recalculated before transmission. During the checksum calculation, the checksum field itself is set to zero.

Source Address

32-bit source address.

Destination Address

32-bit destination address.

Options

Option transmission is optional and often not used, but every IP network entity must support and parse options if they are present. The option field is variable in length. There may be zero or more options. There are two possible formats for an option:

- Single octet option type
- Option type octet, an option length octet, and the actual option data octets.

The length octet includes the option type octet and the actual option data octets. The option type octet has three fields:

- 1 bit = Copied flag. Indicates that this option is copied into all fragments during fragmentation.

0	Copied.
1	Not copied.

- 2 bits = Option class.

0	Control.
1	Reserved for future use.
2	Debugging and measurement.
3	Reserved for future use.

- 5 bits = Option number.

Data

The rest of packet data that is encapsulated in the IP datagram.

IPv6

4 Bits	8 Bits	16 Bits	24 Bits	32 Bits
Version	Priority	Flow Label		
Payload Length			Next header	Hop limit
128-bit Source Address				
128-bit Destination Address				
Data				

Version

Indicates the version of the IP header.

Priority

Indicates the intended delivery priority of the packets.

Flow Label

Used by a source to label those packets for which it requests special handling by the IPv6 router. The flow is uniquely identified by the combination of a source address and a non-zero flow label.

Payload Length

Length of payload (in octets).

Next Header

Identifies the type of header immediately following the IPv6 header. This is similar to the Protocol field in IPv4.

Hop Limit

8-bit max hop count. Hop limit is decremented by each hop along the packet route. If it reaches zero, the packet must be removed from the network. This is similar to the IPv4 TTL.

Source Address

128-bit source address.

Destination address

128-bit destination address.

TCP

<table>
<tr><th colspan="3">16 Bits</th><th colspan="6">32 Bits</th></tr>
<tr><td colspan="3">Source Port</td><td colspan="6">Destination Port</td></tr>
<tr><td colspan="9">Sequence Number</td></tr>
<tr><td colspan="9">Acknowledgement Number</td></tr>
<tr><td>Offset</td><td>Reserved</td><td>U</td><td>A</td><td>P</td><td>R</td><td>S</td><td>F</td><td>Window</td></tr>
<tr><td colspan="3">Checksum</td><td colspan="6">Urgent Pointer</td></tr>
<tr><td colspan="9">Option + Padding</td></tr>
<tr><td colspan="9">Data</td></tr>
</table>

Source Port

Source port number. See the common port numbers list.

Destination Port

Destination port number. See the common port numbers list.

Sequence Number

Sequence numbers are used to indicate the current progress of a multiple segment transmission. The first packet in a transmission (denoted by SYN flag) has an initial sequence number (ISN). The first octet of the initial segment is numbered ISN+1. Each subsequent segment's sequence number is the sum of the ISN and the number of octets previously transmitted in this transmission.

Acknowledgment Number

If the ACK flag is set, this field contains the value of the next sequence number that the sender of the segment is expecting to receive.

Data Offset

This 4-bit field indicates the number of 32-bit words in the TCP header only.

Reserved

This 6-bit field is reserved for future use, and must be set to zero.

Control Flags

This 6-bit field contains the control flags (one bit per flag):

U (URG) Urgent pointer field significant.
A (ACK) Acknowledgment field significant.
P (PSH) Push function.
R (RST) Reset the connection.
S (SYN) Synchronize sequence numbers.
F (FIN) No more data from sender.

Window

This 16-bit field contains the number of octets which the sender of this segment is willing to accept, beginning with the octet indicated in the acknowledgment field. This value typically starts small and grows as the transmission of the datagram is underway.

Checksum

Checksum (*not* a CRC) of the current TCP segment, including the header and data. During the checksum calculation, the checksum field itself is set to zero.

Urgent Pointer

If the URG flag is set, this 16-bit field identifies the offset from the sequence number in this segment where the urgent data ends. In other words, it points to the offset where regular data begins.

Options

If options are present, they are transmitted at the end of the TCP header. Options have a length that is a multiple of 8 bits.

There are two possible formats for an option:

- A single octet of option type.
- An octet of option type, an octet of option length, and the actual option data octets.

The option length includes the option type and option length, as well as the option data octets.

Data

The rest of packet data that is encapsulated in the TCP segment.

UDP

16 Bits	***32 Bits***
Source Port	Destination Port
Length	Checksum
Data	

Source Port

Similar to the TCP source port in concept, but is actually optional. If not used, this field must be set to zero.

Destination Port

Similar to the TCP destination port in concept.

Length

Length (in octets) of this datagram, including the header and the data. The minimum legal length is 8, which would be a UDP header without any additional data.

Checksum

Checksum of the current UDP segment, including the header and data. During the checksum calculation, the checksum field itself is set to zero.

Data

The rest of packet data that is encapsulated in the UDP segment.

MPLS

20 Bits	***3 Bits***	***1 Bit***	***8 Bits***
Label	EXP	S	TTL
IP Header			

Label

20-bit label identifying the flow or path for which the packet belongs.

EXP

3-bit experimental field. This field is often referred to as the Class of Service (CoS) field. These 3-bit field provides for eight classes of service and can be mapped to other protocol CoS mechanisms such as IPv4 ToS or DSCP.

S

1-bit stack bit to indicate the last label in a stack. If a packet only has one MPLS label (one level of hierarchy), this bit is set. If the packet has multiple MPLS labels (multiple levels of hierarchy), the bottom-most or last label has this bit set.

TTL

8-bit TTL field. This field has similar meaning to the IPv4 TTL field.

ICMP

8 Bits	***16 Bits***	***32 Bits***
Type	Code	Checksum
Identifier		Sequence Number
Address Mask		

Type

8-bit field indicating the type of ICMP message.

Code

8-bit field indicating the code for the particular type of ICMP message.

Table 20: ICMP Type/Code Combinations

Type	*Code*	*Description*
0		Echo reply. The *ping* reply.
3		Destination unreachable.
3	0	Net unreachable.
3	1	Host unreachable.
3	2	Protocol unreachable.
3	3	Port unreachable.
3	4	Fragmentation needed and DF set.
3	5	Source route failed.
4		Source quench.
5		Redirect.
5	0	Redirect datagrams for the network.
5	1	Redirect datagrams for the host.
5	2	Redirect datagrams for the type of service and network.
5	3	Redirect datagrams for the type of service and host.
8		Echo. The *ping* request.
11		Time exceeded.
11	0	Time to live exceeded in transit.
11	1	Fragment reassemble time exceeded.
12		Parameter problem.
13		Timestamp.
14		Timestamp reply.
15		Information request.
16		Information reply.

Checksum

Checksum of the current ICMP header. During the checksum calculation, the checksum field itself is set to zero.

Identifier

Number used to identify or match requests to replies. This is optional and may be set to zero.

Sequence Number

Number used to identify or match requests to replies. Used in the same fashion as the Identifier. This is optional and may be set to zero.

Address Mask

32-bit network address mask. When replying network address mask requests, the sender includes its network mask (i.e. 255.255.255.0) here.

IGMP

8 Bits	16 Bits	32 Bits
Type	Max Response Time	Checksum
Group Address		

Version

IGMP protocol version.

Type

Message type:

0x11	Membership query, either general or group-specific.
0x16	Version 2 membership report.
0x17	Leave group.
0x12	Version 1 membership report.
0x22	Version 3 membership report.

Max Response Time

Used only in membership query messages, this value specifies the maximum time (in 100-ms units) allowed before sending a responding report. This field is set to zero for all other messages.

Checksum

Checksum of the IGMP header.

Group Address

When sending a group specific query or group-and-source-specific query, this field contains the group address being queried. In a membership report message, it contains the group address for which membership is being reported. In a leave message, it contains the group address for the group the host wishes to leave. This field is set to zero for general queries.

IGMP Version Membership Report Messages

The following table shows the IGMP Version Membership Report message structure.

8 Bits	16 Bits	32 Bits
Type	Reserved	Checksum
Reserved		Number of Group Record
Group Record		
Group Record		
Group Record		

Type

The IGMP Version Membership Report message has a type value of 21.

Reserved

Reserved. Set to zero on transmission, and ignored on reception.

Checksum

Checksum of the whole IGMP message.

Number of Group Records

The number of group records present in this report.

Group Record

Indicates the sender's membership in a single multicast group.

Group Record Type

The following are group record types that may be included in a Report message:

- Current-state Record, which has the following record types:
  - Include mode
  - Exclude mode
- Filter-mode-change Record, which has the following two record types:
  - Change to Include Mode
  - Change to Exclude Mode
- Source-list-change Record, which has the following two record types:
  - Allow New Sources
  - Block Old Sources

RSVP

4 Bits	*8 Bits*	*16 Bits*	*32 Bits*
Version	Flags	Message Type	RSVP Checksum
Send TTL		Reserved	RSVP Length

Version

Version number.

Flags

Currently, no flag defined.

Message Type

The message type is one of the following:

1	Path.
2	Resv.
3	PathErr.
4	ResvErr.
5	PathTear.
6	ResvTear.
7	ResvConf.

RSVP Checksum

Checksum of the RVSP header.

Send TTL

IP TTL value with which the message was sent.

RSVP Length

Total length (in bytes) of the RSVP message. This includes header and the variable length objects that follow.

IEEE 802.1q

16 Bits			*32 Bits*
Destination MAC Address			
Destination MAC Address	Source MAC Address		
Source MAC Address			
EtherType = 0x8100	Priority	CFI	VID
Ethertype	Data		
Data			
FCS			

EtherType

The EtherType for 802.1Q is 0x8100.

Priority

3-bit 802.1P priority field used to communicate the packet's priority. Can be 0–7.

CFI

When the Canonical Format Indicator (CFI) is set, an addition field called E-RIF is present immediately following the VID. This is only used in with certain Ethernet encapsulations and is less common.

VID

A 12-bit ID that identifies the VLAN to which the frame belongs. There are a total of 4096 VLAN IDs.

0	Null VLAN ID. Indicates that the tag header contains only user priority information, no VLAN ID.
1	Default PVID value used for classifying frames on ingress through a bridge port.
FFF	Reserved for implementation use.

All other values are available for general use as VIDs.

Broadcom Corporation

16215 Alton Parkway
P.O. Box 57013
Irvine, CA 92619-7013
Phone: 949-450-8700
Fax: 949-450-8710

Broadcom® Corporation reserves the right to make changes without further notice to any products or data herein to improve reliability, function, or design. Information furnished by Broadcom Corporation is believed to be accurate and reliable. However, Broadcom Corporation does not assume any liability arising out of the application or use of this information, nor the application or use of any product or circuit described herein, neither does it convey any license under its patent rights nor the rights of others.