

Gigabit Ethernet

Auto-Negotiation

By Rich Hernandez

The Auto-Negotiation standard allows devices based on several Ethernet standards, from 10BaseT to 1000BaseT, to coexist in the network by mitigating the risks of network disruption arising from incompatible technologies. This capability helps ensure a smooth migration path from Ethernet to Fast Ethernet and Gigabit Ethernet. This article provides an in-depth explanation of auto-negotiation and its functioning and also discusses special cases that may be encountered.

Today a number of technologies, such as 10BaseT, 100BaseTX, and 1000BaseT, use the same RJ-45 connector, creating the potential for connecting electrically incompatible components together and causing network disruption. In addition, with the advent of Gigabit Ethernet over copper, three-speed devices now support 10 Mbps, 100 Mbps, and 1000 Mbps operation. The Institute of Electrical and Electronics Engineers (IEEE®) developed a method known as *auto-negotiation* to eliminate the possibility of dissimilar technologies interfering with each other.

Gigabit transceivers at the physical layer (PHY) of the Open Systems Interconnection (OSI) model use auto-negotiation to advertise the following modes of operation: 1000BaseT in full or half duplex, 100BaseTX in full or half duplex, and 10BaseT in full or half duplex. Although auto-negotiation can be disabled for 100BaseTX or 10BaseT connectivity, it is always required for normal 1000BaseT operation.

Auto-negotiation enables an easy upgrade path to gigabit speeds by future proofing the server network connectivity with a three-speed network interface card (NIC) or LAN on motherboard (LOM). A server connected to a Fast Ethernet switch or hub can easily be upgraded to Gigabit Ethernet by connecting the NIC to a Gigabit Ethernet switch. If both the NIC and the switch are set to auto-negotiate, the interface will be automatically configured to run at 1000 Mbps.

The auto-negotiation algorithm (known as NWay) allows two devices at either end of a 10 Mbps, 100 Mbps, or 1000 Mbps link to advertise and negotiate the link operational mode—such as the

speed of the link and the duplex configuration of half or full duplex—to the highest common denominator.

In addition, for 1000BaseT, NWay determines the master-slave interlock between the PHYs at the ends of the link. This mode is necessary to establish the source of the timing control of each PHY. NWay is an enhancement of the 10BaseT link integrity test (LIT) signaling method and provides backward compatibility with link integrity.

Auto-negotiation is defined in Clause 28 of the 1998 edition of the IEEE Standard (Std) 802.3. Clause 28 defines a standard to address the following goals:

- ▶ Provide easy, plug-and-play upgrades from 10 Mbps, 100 Mbps, and 1000 Mbps as the network infrastructure is upgraded
- ▶ Prevent network disruptions when connecting mixed technologies such as 10BaseT, 100BaseTX, and 1000BaseT
- ▶ Accommodate future PHY (transceiver) solutions
- ▶ Allow manual override of auto-negotiation
- ▶ Support backward compatibility with 10BaseT
- ▶ Provide a parallel detection function to recognize 10BaseT, 100BaseTX, and 100BaseT4 non-NWay devices

In addition, the 1999 standard for Gigabit over copper cabling, IEEE Std 802.3ab, added the following enhancements to the Auto-Negotiation standard:

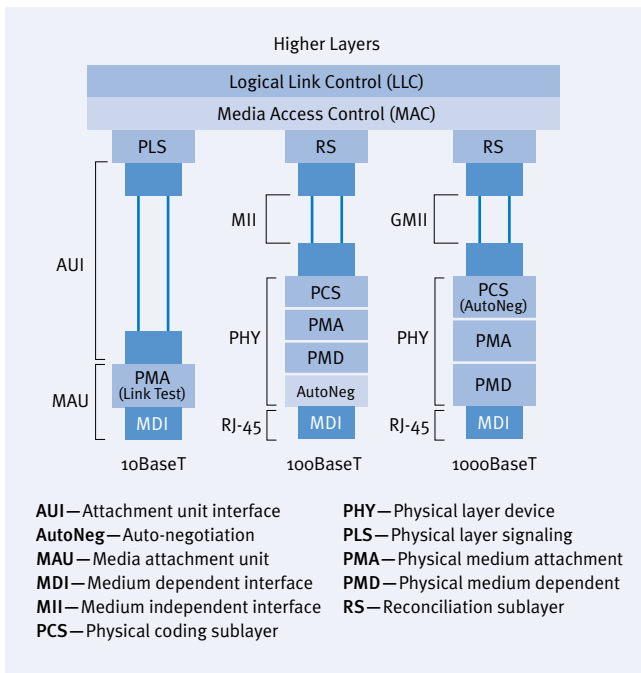


Figure 1. Data terminal equipment layer model (Redrawn from the IEEE Std 802.3, 1998 Edition)

- ▶▶ Mandatory auto-negotiation for 1000BaseT¹
- ▶▶ Configure master and slave modes for the PHY

The Auto-Negotiation specification includes reception, arbitration, and transmission of normal link pulses (NLPs). It also defines a receive LIT function for backward compatibility with 10BaseT devices. All of these functions are implemented as part of the physical layer transceiver as shown in Figure 1. The exchange of link information occurs between the PHY and the Medium Dependent Interface (MDI) or RJ-45 connector.

Gigabit Ethernet defines auto-negotiation as a functional block part of the physical coding sublayer (PCS) function, while in 100BaseT it is defined as a separate sublayer in the PHY. All auto-negotiation functions are implemented as part of the transceiver integrated circuit, which is part of a network interface card or integrated on the motherboard of a computer.

10BaseT link test pulses

The 10BaseT standard includes a link test mechanism to ensure network integrity. In the absence

The Auto-Negotiation specification includes reception, arbitration, and transmission of normal link pulses (NLPs).

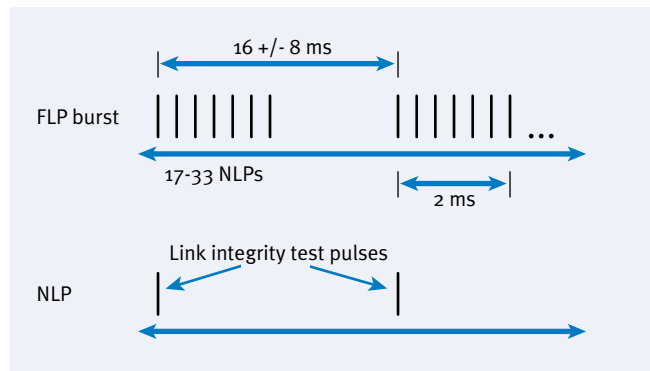


Figure 2. FLP and NLP comparison

of network traffic, a 100 nanosecond (ns) heartbeat unipolar (positive only) pulse is sent every 16 milliseconds (ms) within a range of +/- 8 ms. The link test pulse is sent by the transmitters of all 10BaseT media attachment units (MAUs) between the data terminal equipment (DTE) and the repeater.

A link fail condition is entered if the receiver does not receive a packet or a link test pulse within 50-150 ms. The link fail condition disables the data transmit, data receive, and loopback functions. The link test pulses continue to be transmitted and received during the link failure. The link is reestablished when two consecutive link test pulses or a single data packet have been received.

100BaseT/1000BaseT fast link pulses

The link information is encoded in a special pulse train known as the fast link pulse (FLP) burst. The FLP builds on the LIT pulse used by 10BaseT devices as a heartbeat pulse to the link partner at the opposite end of the link. The LIT was redefined as the normal link pulse (NLP). As shown in Figure 2, the NLP is the 10BaseT link integrity test pulse, and the FLP is a group of NLPs. Each pulse is 100 ns in width.

Auto-negotiation replaces the single 10BaseT link pulse with the FLP burst. Auto-negotiation stops the transmission of FLP bursts once the link configuration is established. The FLP burst looks the same as a single link test pulse from the perspective of 10BaseT devices. Consequently, a device that uses NWay must recognize the NLP sequence from a 10BaseT link partner, cease transmission of FLP bursts, and enable the 10BaseT physical medium attachment (PMA). Auto-negotiation does not generate NLP sequences—it only recognizes NLPs. Instead,

¹Auto-negotiation is optional for 100BaseT as defined in the original IEEE Std 802.3, Clause 28.1.1, 1998 Edition. However, all Dell server NICs implement this standard.

auto-negotiation passes control to the 10BaseT PMA to generate NLPs.

FLP bursts

Each FLP burst consists of 33 pulse positions that provide clock and data information. The 17 odd-numbered pulses are designated as clock pulses, while the 16 even-numbered pulse positions represent data information. A logic one is represented by the presence of a pulse, while the absence of a pulse is represented by a logic zero. Figure 3 shows the timing characteristics of the clock and data pulses.

FLP burst encoding

The data pulses in the FLP burst encode a 16-bit link code word (LCW). A device capable of auto-negotiation transmits and receives the FLP. The receiver must identify three identical LCWs before the information is authenticated and used in the arbitration process. The devices decode the base LCW and select capabilities of the highest common denominator supported by both devices. Once the LCWs are properly received, each device transmits a FLP burst with an acknowledge bit. At this point, both devices enable the mode that is the highest common mode negotiated.

The clock pulses are used for timing and recovery of the data pulses. The 17 clock pulses are always present in the FLP burst. The first pulse on the wire is a clock pulse. The 16 data pulses may or may not be present. If the data pulse is present, it represents a value of one in the LCW for that position. The lack of a data pulse indicates a zero in the LCW for that position, as shown in Figure 4.

Base link code word

The base LCW is transmitted within an FLP burst after power-on, reset, or renegotiation. The DTE and its link partner communicate

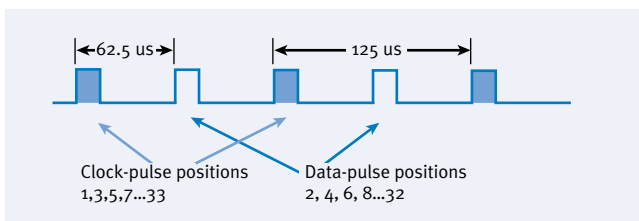


Figure 3. FLP burst timing

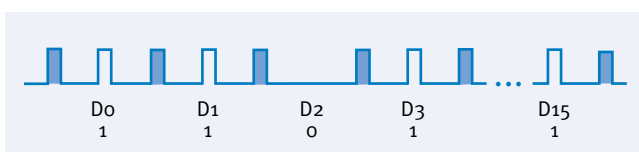


Figure 4. FLP burst encoding

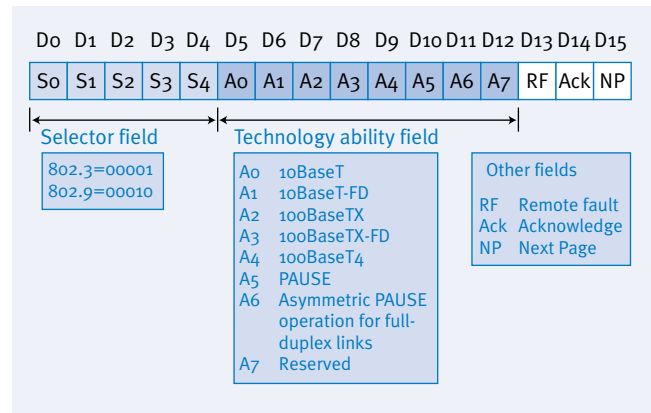


Figure 5. Base link code word definition

their capabilities by exchanging LCWs. Figure 5 defines the bit positions of the base LCW. These bit positions map directly to the data pulses in the FLP burst—bits D0 through D15.

Technology ability field

The technology ability field (TAF), which is encoded in bits D5 through D12 of the FLP burst, is shown for the IEEE 802.3 Base Page as defined in the Selector field (00001) for IEEE 802.3 Ethernet. The order of the bits within the TAF does not correspond to the relative priority of the technologies during the arbitration process. Each device capable of auto-negotiation maintains a prioritization table used to determine the highest common denominator ability. These priorities were updated to include Gigabit Ethernet over copper, as shown in Figure 6.

Other fields

A local device may use the Remote Fault (RF) bit to indicate the presence of a fault detected by the remote link partner. For example, the RF bit is set to a logic one when the device enters

| Priority | Technology |
|-------------|-------------------------|
| 1 (highest) | 1000BaseT — Full duplex |
| 2 | 1000BaseT — Half duplex |
| 3 | 100BaseT2 — Full duplex |
| 4 | 100BaseTX — Full duplex |
| 5 | 100BaseT2 — Half duplex |
| 6 | 100BaseT4 |
| 7 | 100BaseTX — Half duplex |
| 8 | 10BaseT — Full duplex |
| 9 (lowest) | 10BaseT — Half duplex |

Figure 6. Priority resolution table

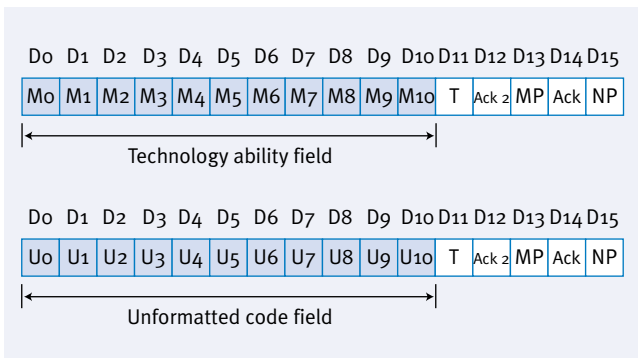


Figure 7. Message and unformatted pages

| Data pulse number | Bit position | Definition | Value |
|---------------------------------------|--------------|---|---|
| BASE PAGE | | | |
| D15 | NP | Next Page | 1=Next Page follows |
| D14:D1 | S4:S0, A7:A0 | Selector field Technology field | As defined in the 802.3 Base Page |
| PAGE 0 (Message Next Page) | | | |
| D10:D0 | M10:M0 | 8 | |
| D10:D5 | U10:u5 | Reserved | 0 |
| D4 | U4 | 1000BaseT half duplex | 0=No half-duplex capability 1=Half-duplex capability |
| D2 | U2 | 1000BaseT port type | 0=Single port device 1=Multiport device |
| D1 | U1 | 1000BaseT Master-Slave manual configuration value | 0=Slave 1=Master |
| D0 | U0 | 1000BaseT Master-Slave manual configuration value | 1=Enable 2=Disable |
| PAGE 2 (Unformatted Next Page) | | | |
| D10 | U10 | 1000BaseT Master-Slave seed bit 10 (SB10) MSB | Master-Slave seed value (10:0) |
| D9 | U9 | 1000BaseT Master-Slave seed bit 9 (SB9) MSB | |
| D8 | U8 | 1000BaseT Master-Slave seed bit 8 (SB8) MSB | |
| D7 | U7 | 1000BaseT Master-Slave seed bit 7 (SB7) MSB | |
| D6 | U6 | 1000BaseT Master-Slave seed bit 6 (SB6) MSB | |
| D5 | U5 | 1000BaseT Master-Slave seed bit 5 (SB5) MSB | |
| D4 | U4 | 1000BaseT Master-Slave seed bit 4 (SB4) MSB | |
| D3 | U3 | 1000BaseT Master-Slave seed bit 3 (SB3) MSB | |
| D2 | U2 | 1000BaseT Master-Slave seed bit 2 (SB2) MSB | |
| D1 | U1 | 1000BaseT Master-Slave seed bit 1 (SB1) MSB | |
| D0 | U0 | 1000BaseT Master-Slave seed bit 0 (SB0) MSB | |

Figure 8. 1000BaseT Base and Next Pages

the link fail state. The RF bit is reset to zero once the LCW is successfully negotiated. After the successful receipt of the three identical LCWs, the device will transmit an LCW with the Ack bit set to one. This LCW must be transmitted a minimum of six to eight times.

Additional pages, other than the base LCW, can be sent if both devices on the link set the Next Page bit in the base LCW. The Next Page protocol consists of a message page (MP) and one or more unformatted pages. These pages are LCW encoded in the FLP like the Base Page. Figure 7 shows the message and unformatted LCW bit definition.

If the MP bit is set to one, the LCW will be treated as a message page and interpreted to carry a message as defined in IEEE Std 802.3, Annex 28C, 1998 Edition. The unformatted pages include the specific message information, indicated by an MP bit that is set to zero.

Gigabit auto-negotiation

1000BaseT devices use auto-negotiation to set up the link configuration by advertising the PHY capabilities, including speed, duplex, and master-slave mode. Gigabit Ethernet over copper relies on the exchange of a Next Page LCW that describes the Gigabit extended capabilities. These capabilities are documented via one Base Page, one 1000BaseT message Next Page, and two 1000BaseT unformatted Next Pages, as shown in Figure 8. A message code with a value of 8—M10:M0=00000001000 indicates that the 1000BaseT technology message code will be transmitted.

A 1000BaseT PHY can operate as a master or slave. A prioritization scheme determines which device will be the master and which the slave. The IEEE supplement to Std 802.3ab, 1999 Edition defines a resolution function to handle any conflicts. Multiport devices have higher priority to become master than single port devices. If both devices are multiport devices, the one with higher seed bits becomes the master.

Parallel detection

A device determines that a link partner can use auto-negotiation by detecting the FLP burst. However, some devices may not have implemented the auto-negotiation function. For devices that support 100BaseTX, 100BaseT4, or 10BaseT, a parallel detection function allows detection of the link.

The receiver device passes the signals in parallel to both the NLP/FLP detector functional block and to any 100BaseTX or 100BaseT4 PMAs or the NLP receive LIT functions in the physical transceiver. If the native signal causes the TX/T4 PMA to enter a link good condition, then the auto-negotiation function is bypassed.

Examples of auto-negotiation interoperability

These three examples illustrate the flexibility of the 10/100/1000 802.3ab Auto-Negotiation standard. Each has a server with a triple-speed Gigabit over copper network device (NIC or LAN on motherboard) connected to a 10/100 Ethernet switch.

Case one: Interoperability with 10BaseT devices

The link partner can be a port on a 10 Mbps hub or a 10/100 Mbps switch configured for 10 Mbps operation only. The NIC in the server is configured for auto-negotiation.

Communication between a 10BaseT device and the NIC proceeds as follows (see Figure 9):

- ▶▶ The DTE powers up in link fail mode and transmits FLPs.
- ▶▶ The link partner transmits NLPs.
- ▶▶ The link partner goes into link fail initially because it has not yet received any NLPs.
- ▶▶ The DTE parallel detection function detects the NLPs, passes control to the 10BaseT PMA, and starts transmitting NLPs.
- ▶▶ A link is established at 10 Mbps half duplex. (*Note: When the auto-negotiation parallel detection function detects the link, it defaults to half duplex; therefore, the 10BaseT legacy device must be set to half duplex.*)

Case two: Interoperability with non-auto-negotiation 100BaseT

The link partner can be a port on a 100 Mbps hub or a 10/100 Mbps switch configured for 100 Mbps operation only. The NIC in the server is configured for auto-negotiation.

Communication between a non-auto-negotiation 100BaseT device and the NIC follows these steps, illustrated in Figure 10:

- ▶▶ The DTE powers up in link fail mode and transmits FLPs.
- ▶▶ The 100BaseTX link partner powers up and sends idle symbols.
- ▶▶ The DTE parallel detection function detects the idle symbol, bypasses the auto-negotiation function, passes control to the 100BaseTX PMA, and transmits idle.
- ▶▶ A link is established at 100 Mbps half duplex.

Case three: Interoperability with auto-negotiation 100BaseT (10/100)

The link partner is a port on a 10/100 switch configured for auto-negotiation. The NIC in the server is configured for auto-negotiation and capable of 10/100/1000 Mbps operation.

Figure 11 illustrates the following communication steps between a 100BaseT device and the NIC:

- ▶▶ Both devices power up in link fail mode and transmit FLPs.

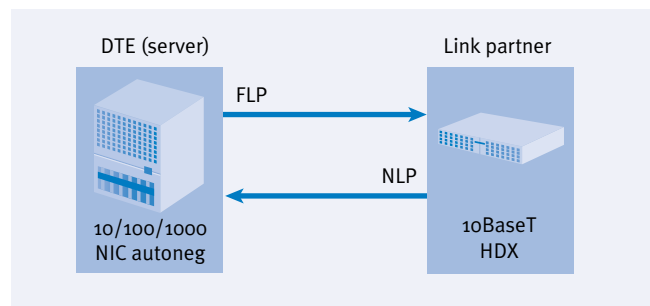


Figure 9. 10BaseT-only legacy device

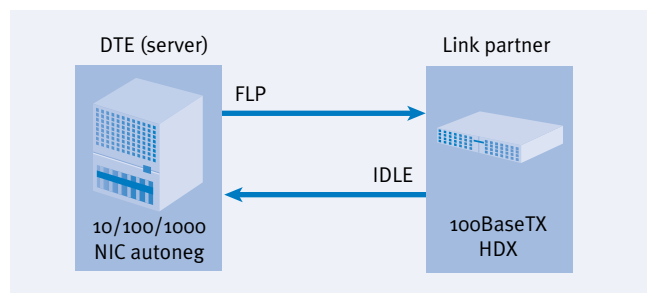


Figure 10. Non-auto-negotiation 100BaseT

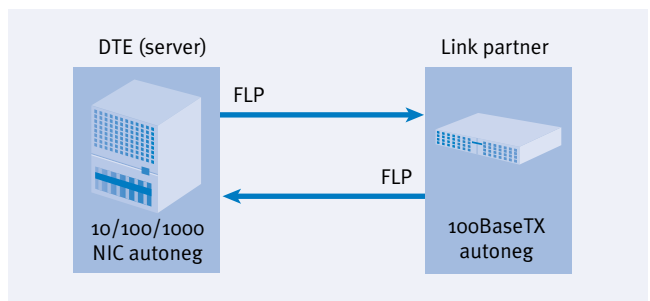


Figure 11. Auto-negotiation 10/100 device

- ▶▶ Each device receives and decodes the capabilities of the other.
- ▶▶ A link is established at 100 Mbps full duplex.

Duplex mismatching

A link is degraded when a device and its link partner, such as a server NIC and a switch, are configured so that the duplex settings do not match; that is, one is set to half duplex and the other to full duplex. When both devices send frames simultaneously on the link in this configuration, the following occurs:

- ▶▶ The half-duplex link detects a collision, which corrupts its outgoing frame and discards the incoming frame. The half-duplex link will attempt to retransmit the frame.
- ▶▶ The full-duplex link will not resend its frame. It determines that the incoming frame is bad and flags cyclical redundancy check (CRC) errors.

| 10/100/1000 BaseT switch configuration | | | | | | | | |
|--|--------|-----------------|--------|-----------------|---------|-----------------|----------|-----------------|
| Speed | Auto | 10 | 10 | 100 | 100 | 1000 | 1000 | |
| Duplex | | Half | Full | Half | Full | Half | Full | |
| NIC configuration | | | | | | | | |
| Speed | Duplex | | | | | | | |
| Auto | | 1000 FDX | 10 HDX | Duplex conflict | 100 HDX | Duplex conflict | 1000 HDX | Duplex conflict |
| 10 | Half | 10 HDX | 10 HDX | | | | | |
| 10 | Full | Duplex conflict | | 10 FDX | | | | |
| 100 | Half | 100 HDX | | | 100 HDX | | | |
| 100 | Full | Duplex conflict | | | | 1000 FDX | | |
| 1000 | Half | 1000 HDX | | | | | 1000 HDX | |
| 1000 | Full | Duplex conflict | | | | | | 1000 FDX |


Figure 12. Configuration table for 10/100/1000 BaseT devices

- ▶ Applications will timeout and retransmit continuously, causing a very slow connection.

Figure 12 summarizes all possible combinations of speed and duplex settings, both on 10/100/1000-capable switch ports and on NICs. For example, connecting a 10/100/1000-capable NIC configured for auto-negotiation with a 10/100/1000 switch also configured for auto-negotiation results in the ports for both the NIC and the switch being configured at 1000 Mbps full duplex. Figure 12 also shows combinations that would yield no link or link fail conditions, as well as combinations that would yield a duplex mismatch.

Enabling migration to gigabit speeds

The IEEE standard for auto-negotiation ensures easy migration from 10 Mbps to 100 Mbps and 1000 Mbps speeds. A 10/100 BaseT NIC installed on a server connected to a 10/100 switch port set for auto-negotiation can be upgraded to a 10/100/1000 BaseT connection by simply replacing the NIC in the server. The new NIC will auto-negotiate to 100 Mbps full duplex automatically. The same is true with a LAN-on-motherboard (LOM) interface. A new server with a 10/100/1000 BaseT LOM interface can be substituted with no configuration changes in a switch or cable plant.

When a 10/100 Ethernet switch is upgraded to a Gigabit over copper switch, the NIC in the server will negotiate to operate at 1000 Mbps automatically, without stopping or rebooting the server. The Auto-Negotiation Standard allows newer, faster devices to be incorporated into the network without disruption, by negotiating capabilities of the highest common denominator between the two ends of a link. 

Rich Hernandez (rich_hernandez@dell.com) is a senior engineer with the Server Networking and Communications Group at Dell. Rich has been in the computer and data networking industry for over 16 years. He has a B.S. in Electrical Engineering from the University of Houston and has pursued postgraduate studies at Colorado Technical University.

FOR MORE INFORMATION

For more information about the IEEE Standard 802.3, visit www.ieee.org