

Untangling Ethernet performance problems

It's the root cause behind the most common SOS we get related to IT infrastructure performance. An organization installs a new application, maybe some new servers and switches. Suddenly network performance takes a nosedive. It's not just bad; at peak times it's abysmal.

The fact is that it doesn't take a wholesale change in an Ethernet network to cause performance problems such as these. The reason is that autonegotiation in 10 and 100 Mb/s twisted-pair networks doesn't work as advertised. Autonegotiation, sometimes known as autosensing, is the mechanism by which each end of a copper Ethernet connection decides how to talk: at what speed, and at full or half duplex. The problem is that the standards for 10 and 100 Mb/s Ethernet evolved rapidly, different manufacturers interpreted the standards differently, and some made proprietary extensions that complicated interoperability further. The problem with autonegotiation is so significant that even two different products from the same vendor won't always negotiate connections properly if they are built with two different Ethernet chipsets. That's the bad news.

The good news is that you can follow some relatively straightforward guidelines that will keep you and your network out of trouble. This article focuses on the lessons we've learned in the real world actually trying to get network hardware from different vendors to work together. Don't expect to find these lessons in your owner's manual, as in theory everything works just fine. This article focuses exclusively on twisted-pair copper networks.

A brief history

Ethernet was invented in 1973 at the Xerox Palo Alto Research Center. An industry consortium established the first standard in 1980, and the IEEE published the 802.3 standard for thick Ethernet in 1985. This standard specified a coaxial cable that was about as big around as a garden hose, with each device connected to the network through a tap that physically penetrated the cable (Figure 1).

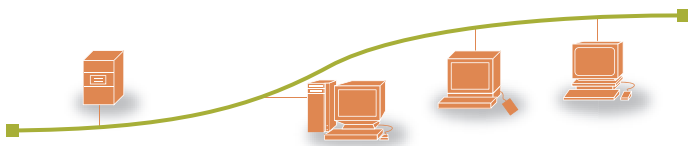


Figure 1: Ethernet circa 1985

Ethernet quickly grew in popularity. The limitations of coaxial-cable implementations were overcome with standards for star topologies that use multi-port switches and hubs connected to servers and workstations through twisted-pair cables (Figure 2).

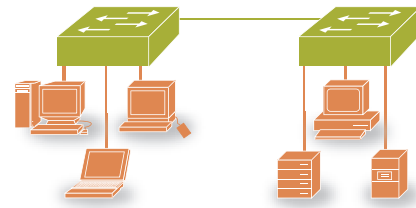


Figure 2: Switched copper Ethernet

Fiber standards were developed to overcome copper cable-length limitations. Speeds increased from 10 to 100 to 1000 megabits per second, and today, 10 gigabit Ethernet is beginning to catch on. Full-duplex standards were published in 1997 to allow devices to talk and listen simultaneously on separate twisted pairs. This allows two conversations at a time, a feat impossible with the original shared coaxial-cable media.

What's the problem?

Ethernet grew too quickly for its own good. Standards for twisted-pair media were not well thought out, leaving vendors to make their own interpretations and add their own "helpful" extensions. The problem is particularly acute with autonegotiation in 10 and 100 Mb/s networks, whereas the problem is almost non-existent in 1000 Mb/s networks.

When autonegotiation or manual settings result in speed mismatches, you won't even get a link light to display. Duplex mismatches — one side half duplex and the other side full duplex — cause significant packet-loss problems. Even if both sides of the link agree on their speed and duplex settings, one end of the connection can sometimes think it needs to renegotiate, bringing the link down at random times.

It's best to conceptualize today's switched networks as a collection of smaller networks. Each link between a workstation or server and a switch, and each link between two switches, is its own network. In a switched network, each end of each link can be set to half or full duplex, and at 10, 100, and 1000 Mb/s. Incompatible settings at different ends of the link, or substandard cabling, can cause problems. String a set of links together to make a path between two devices, and a problem at any point along the way can hamper performance.

How to know when it's your problem

You'll know for sure that you have a network problem when you encounter a scenario like the one that begins this article, and almost every network has a problem lurking somewhere. We've seen network file operations that usually take 3-4 seconds suddenly take 7-8 minutes to complete. When the problem is this severe, you'll hear users screaming in the hallways. There

are, however, more subtle symptoms that you might notice beforehand:

1. Poor performance on the local network, including dropped packets. If you have packet-loss problems, your network staff might notice higher-than-usual switch error rates, or your IT people might notice application performance glitches. Local networks should drop packets only rarely.
2. Peak-time performance issues. Sometimes the problems are only noticeable when the network is heavily loaded, such as peak times when everyone is in the office.
3. Inability to transmit large packets. Every peer on the network should be able to exchange packets up to the Maximum Transmit Unit (MTU) for the network, usually 1500 bytes.
4. Switch-to-switch connections fail. You might find links between switches resetting frequently, causing brief outages that affect many users. You also might find links failing completely.
5. Switch ports reset. Some switches have options like “port monitoring” or “port security,” which only allow certain Media Access Controller (MAC) addresses to connect. These switches will sometimes automatically disable ports, shutting off specific hosts or parts of the network, indicating a problem.

Diagnostic steps

There are several ways to diagnose whether you have a speed or duplex mismatch problem:

6. Ping with large payloads. Use the ping command between hosts where you’re noticing performance problems. Specify a large payload close to the network’s MTU, typically 1500 bytes. You should not see any packet loss. Even if you see only a 10-20% loss, you have probably found a bad link.
7. Measure throughput. Use `ttcp` or `Iperf` (see <http://dast.nlanr.net/Projects/Iperf/>) to measure throughput between the same links you’ve tested with ping. A healthy 100 Mb/s switched link should give you 70 Mb/s or greater.
8. Check your logs. Monitor your switch logs for ports that change speed or duplex frequently. Watch your `syslog` to see if workstations or servers are doing the same.
9. Beware of misconfigured firewalls. This problem can cause some of the same symptoms as speed/duplex mismatches, and is another of the top 10 most-common problems we diagnose. Overzealous administrators sometimes deny Internet Control Message Protocol (ICMP) packets that are needed to support MTU discovery (ICMP Type 3, Code 4, or “packet too big”). With this mechanism disabled for packets crossing the firewall, some large packets can be improperly fragmented or dropped. When blocked ICMP packets are the problem, you’ll find that 99% of hosts have no problem using the network, but a very few have significant problems.

Clean house

If you’ve determined that you have a problem, or you want to be proactive about problems with speed/duplex mismatch, the first step is to take stock of what’s connected to your network:

10. If you have switches or NICs that don’t allow speed and duplex to be set manually, get rid of them. This means eliminating all unmanageable switches.
11. Consider each link and make sure that speed and duplex is either set manually or that it is set in such a way that both ends will successfully negotiate the right settings — and not reset at random times.
12. Take stock of your cabling. Premade drop cables are usually pretty good, provided that they are adequate for the connection speed. Make sure these cables are good quality and are the ones that you’ve provided. Cables brought from home might not meet your standards.
13. Make sure you know how your offices and your datacenters are wired; often they are different. Category 5 cabling is fine for 100 Mb/s Ethernet, except if you’re trying to handle full-duplex connections. For 100 Mb/s full-duplex connections, crosstalk can be an issue even with short cable lengths. Category 5e cabling is often sufficient for 1000 Mb/s Ethernet, but the standard requires you to validate each link’s line quality with a network analyzer to confirm that it meets more stringent requirements than the original Category 5e specification. Often you’ll find datacenter cabling to have acceptable quality, while office cabling quality is insufficient to support gigabit speeds.
14. Remember that patch cables (including wall-to-workstation cables), patch panels, and wall jacks must also be certified for Category 5e. Insist on seeing the manufacturer’s specifications when installing new jacks and panels to ensure that unwanted substitutions do not occur.

Once you have a network that allows you to configure each link manually, check each and every one of them. Configure each end of each link manually, or in such a way that both ends will successfully negotiate the right speed and duplex settings. Autonegotiation in 10/100 Mb/s networks has its place, but it is limited to conference rooms and hoteling situations where systems come and go and where autonegotiation is the lesser of many evils.

The following sections will help you make the right choice depending on whether the least common denominator speed is 10, 100, or 1000 Mb/s. Although the discussion focuses on NIC-to-switch links, the same rules of thumb apply to switch-to-switch links. The sidebars give instructions for setting in the Linux, Solaris,™ and Microsoft® Windows® XP operating systems.

Setting gigabit Ethernet links

Wouldn’t it be nice to have an all-new network with switches, NICs, and cabling all up to 1000 Mb/s standards? Unfortunately,

most of us don't have the luxury. The standards for gigabit Ethernet were designed with the problems of 10/100 Mb/s networks in mind, so speed and duplex problems have nearly been eliminated. Indeed, autonegotiation is different, and full duplex is standard for 1000 Mb/s links, vastly simplifying network configuration.

The rule of thumb for 1000 Mb/s links is to let autonegotiation work by setting the NIC and the switch to autonegotiate (see Table 1). If you have Cisco switches you can manually set each side of the link to full duplex and the link will work, but this technically violates the standard.

| | NIC Settings | Switch Settings | Comments |
|----------|--------------|-----------------|---|
| Setting: | 1000/Auto | 1000/Auto | Yes. This works because autonegotiation for gigabit Ethernet is implemented correctly. |
| Result: | 1000/Full | 1000/Full | |
| Setting: | 1000/Full | 1000/Full | Yes. Setting both sides to full duplex works for Cisco switches, but violates the specification. |
| Result: | 1000/Full | 1000/Full | |
| Setting: | 1000/Full | 1000/Auto | No. Although autonegotiation results in both sides of the link set to full duplex, the "auto" side of the link is likely to reset and renegotiate periodically, causing performance problems. |
| Result: | 1000/Full | 1000/Full | |
| Setting: | 1000/Auto | 1000/Full | |
| Result: | 1000/Full | 1000/Full | |

Table 1: Gigabit Ethernet settings

If you set the NIC or the switch manually, and you set the other side to autonegotiate, you'll see a problem common to this combination regardless of link speed. When one side is set manually, it does not participate in the autonegotiation process. The "auto" side, seeing no response to its negotiation requests, makes an assumption about the link speed and duplex. In the case of gigabit Ethernet, the settings assumed by the "auto" side are the same as the manual setting, and each side will be set to full duplex — usually. Sometimes the "auto" side will try again to autonegotiate the link, bringing it down from time to time and causing performance glitches.

Cabling is important for 1000 Mb/s links. At a minimum, no cable can exceed 100 m in length, it must be Category 5, 5e, or 6, and it must meet 1000Base-T standards for FEXT and return loss. Note that the standards for Category 5 and 5e cabling are not adequate for gigabit Ethernet, but cabling built to these standards often works by chance. Nevertheless, each link must be qualified using a network analyzer to be sure that it meets the ANSI/EIA/TIA-TSB-67 requirements. For any link not meeting these requirements, select a slower speed and duplex setting manually. Gigabit Ethernet will not negotiate lower speeds based on connection quality.

Setting 100 Mb/s Ethernet links

The most important thing to know about setting 100 Mb/s Ethernet links is that autonegotiation often does not work as

advertised. The only safe thing to do is to set each side of the connection to half duplex manually. As Table 2 illustrates, most other settings can lead to trouble:

- Setting both ends of a link to "auto" is the most risky combination of settings, other than manually creating speed and duplex mismatches that are sure to fail.
- If you set one side to "auto" and the other side to half duplex, the "auto" side will assume half duplex when it doesn't hear back from its peer during autonegotiation. Both sides will set to half duplex, but the "auto" side may periodically reset and renegotiate, dropping the link and affecting performance.
- If you set one side to full duplex and one side to "auto," you'll end up with a duplex mismatch that will severely affect performance. In these cases, the "auto" side, not hearing from its peer, sets to the half-duplex default, while the other side of the link remains at full duplex.

The second most important thing to know about 100 Mb/s Ethernet is that full duplex is overrated and almost always not worth the trouble, hence the warning in Table 2. The only place where 100 Mb/s full duplex is a good idea is on very short switch-to-switch links that use high-quality cable. These conditions are usually found only in datacenter environments.

| | NIC Settings | Switch Settings | Comments |
|----------|--------------|-----------------|---|
| Setting: | 100/Half | 100/Half | Yes. Setting both sides of the connection to half duplex is the safest combination of settings. |
| Result: | 100/Half | 100/Half | |
| Setting: | 100/Full | 100/Full | Yes. But this combination is highly sensitive to crosstalk from substandard cable quality or longer cable runs. Measure your full duplex performance and compare it to half duplex performance, because full duplex is not always better. |
| Result: | 100/Full | 100/Full | |
| Setting: | 100/Auto | 100/Auto | No. This is a risky combination of settings because you can't be sure what the outcome of the autonegotiation will be. |
| Result: | Unknown | Unknown | |
| Setting: | 100/Half | 100/Auto | No. Although autonegotiation results in both sides of the link set to half duplex, the "auto" side of the link is likely to reset and renegotiate periodically, causing performance problems. |
| Result: | 100/Half | 100/Half | |
| Setting: | 100/Auto | 100/Half | |
| Result: | 100/Half | 100/Half | |
| Setting: | 100/Full | 100/Auto | No. Duplex mismatch. Auto settings don't work in these cases because the manually-configured side of the link does not participate in auto-renegotiation, so the "auto" side assumes that the other side is half duplex. |
| Result: | 100/Full | 100/Half | |
| Setting: | 100/Auto | 100/Full | |
| Result: | 100/Half | 100/Full | |

Table 2: Fast 100 Mb/s Ethernet settings

Remember that full duplex means the NIC and the switch can send and receive data simultaneously. This means that a number

of stars have to be aligned just right to give you a performance advantage: the NIC must be able to send and receive at the same time, the driver and the operating system must support it, and the application itself must be multi-threaded and have two threads running simultaneously. 100 Mb/s full-duplex Ethernet is very sensitive to crosstalk, and full-duplex cable lengths are limited to 25 m. A cable that works for one full-duplex link may not work with another, because even different chipsets from the same manufacturer can have different transmitter characteristics that affect the link.

If you really need the performance, consider upgrading to 1000 Mb/s, where you'll see a benefit from the high data rate if not from having full-duplex data paths. If you still think you need a full-duplex link, then set each end of the link manually as in Table 2.

Setting 10 Mb/s Ethernet links

10 Mb/s Ethernet has pretty much gone the way of vinyl records, but some organizations have a few old relics they must maintain. Besides the fact that it is slower, 10 Mb/s Ethernet suffers from even more autonegotiation problems than 100 Mb/s Ethernet.

The best way to manage 10 Mb/s Ethernet links is to set each side of the connection to half duplex (Table 3). Few manufacturers implemented full duplex in 10 Mb/s Ethernet correctly, so it's a good setting to avoid. All other link-setting combinations have the same issues as 100 Mb/sec Ethernet. The "auto" and half-duplex combination can result in switch or NIC resets. The "auto" and full-duplex combination results in a duplex mismatch. Both setting combinations degrade performance.

| | NIC Settings | Switch Settings | Comments |
|----------|--------------------------------------|-----------------|--|
| Setting: | 10/Half | 10/Half | Yes. This works because both sides of the link are set manually. |
| Result: | 10/Half | 10/Half | |
| Setting: | 10/Full | 10/Full | No. Although this should work, few manufacturers have implemented full duplex correctly. |
| Result: | 10/Full | 10/Full | |
| Setting: | All other combinations | | No. While autonegotiation results in both sides of the link set to full duplex, the "auto" side of the link is likely to reset and renegotiate periodically, causing performance problems. |
| Result: | Duplex mismatch or NIC/switch resets | | |

Table 3: 10 Mb/s Ethernet settings

Speed mismatch

One way to tell a speed mismatch is by a link light that won't illuminate. If you follow the rules of thumb and set both ends of each link in your network manually, you won't have to worry about speed mismatches from a failed autonegotiation. That said, remember that autonegotiation in 1000 Mb/s links — the recommended setting — will not fall back automatically to a slower speed based on cable quality. If you need to use a slower speed because your cable does not pass muster, change the

settings manually.

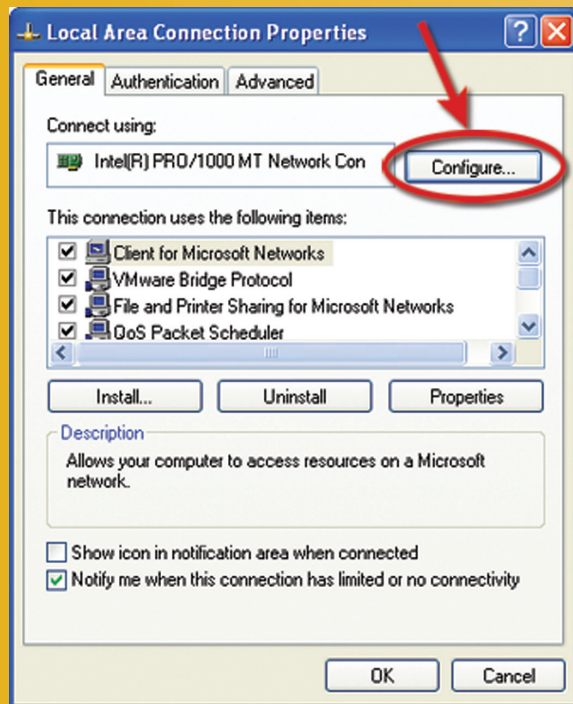
Summary

Autonegotiation in 10 and 100 Mb/s Ethernet does not work as advertised, especially when connecting switches and NICs from different manufacturers. The auto/auto setting that manufacturers recommend is actually one of the settings most fraught with obscure problems — unless you're working with 1000 Mb/s links.

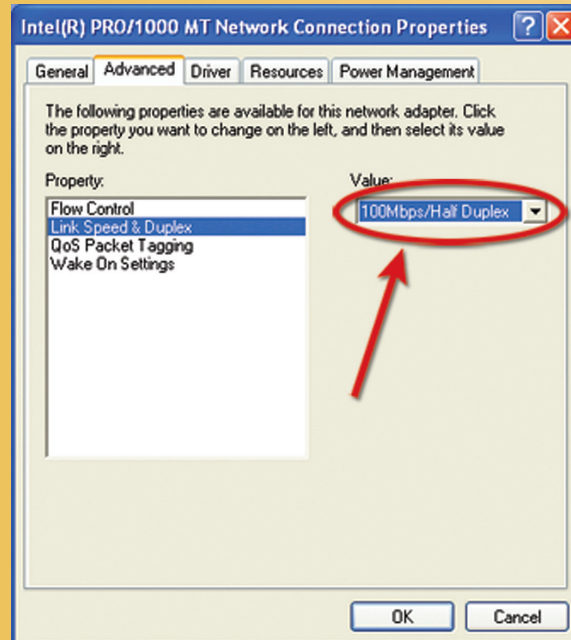
Let autonegotiation work on gigabit Ethernet links, and pay close attention to cable quality and length. For everything else, don't risk speed and duplex mismatches: manually set the speed and half duplex at each side of each link in your network. If you think you need full duplex, upgrade connection speed instead. Paying attention to these straightforward rules of thumb will keep you and your networks out of trouble — and it may save you from sending out your own network SOS. ■

SETTING 100 MB/S HALF DUPLEX IN MICROSOFT WINDOWS SYSTEMS

In Microsoft Windows operating systems, you can set link speed and duplex by configuring the driver properties. Click on “Network” in the Control Panel, and then click on “Configure...” next to the network card driver name.



Depending on the driver provided by your Ethernet NIC vendor, you should be able to click on the “Advanced” tab and select a property similar to the “Link Speed & Duplex” illustrated below. Select 100 Mb/s half duplex, click “OK,” and your NIC should be set.



SETTING 100 MB/S HALF DUPLEX IN SOLARIS

To set speed and duplex in the Solaris operating system, use the enigmatic `ndd` command as the root user to set options in the running driver. First use the `ifconfig -a` command to find out the names of your Ethernet interfaces, which give clues to the names of your Ethernet drivers. If `ifconfig` gives you interface names like `/dev/eri0` and `/dev/eri1`, you have two interfaces and your driver name is `/dev/eri`. You can use the following sequence of commands to find out how interface 1 is currently set:

```
ndd -set /dev/eri instance 1
ndd /dev/eri link_status
ndd /dev/eri link_mode
ndd /dev/eri link_speed
```

Link status is 1 if up, 0 if down; link mode is 1 if full duplex, 0 if half duplex; link speed is 1 if 100 10 Mb/s and 0 if 10 Mb/s.

To set the link speed manually, you must turn off autonegotiation and then restrict the driver's options in setting the link characteristics. The following sequence of commands

allows the driver to select only 100 Mb/s half duplex:

```
ndd -set /dev/eri instance 1
ndd -set /dev/eri adv_autoneg_cap 0
ndd -set /dev/eri adv_100fdx_cap 0
ndd -set /dev/eri adv_100hdx_cap 1
ndd -set /dev/eri adv_10fdx_cap 0
ndd -set /dev/eri adv_10hdx_cap 0
```

The parameters are self explanatory, and if you want to see the entire list of options that the `/dev/eri` driver allows, use the command: `ndd /dev/eri \?`.

SETTING 100 MB/S HALF DUPLEX IN LINUX

Setting speed and duplex in Linux is the easiest of all. Running as the root user, the `mii-tool` command with no arguments reports on all of your network interfaces and their current status. The `mii-tool -v interface-name` command gives tells you all of the capabilities of the interface. You can use the `-F` option to set an interface to 100Mb/s half duplex, as in:

```
mii-tool -F 100baseTx-HD eth0
```

This article was reprinted from the Q4 2005 issue of *The Barking Seal*, a publication of Applied Trust. You can subscribe to *The Barking Seal* online at <http://www.appliedtrust.com/barkingseal>. “Applied Trust” and the “Seal on a Rock” Applied Trust Engineering logo are registered service marks of Applied Trust. All other trademarks are registered to their respective owners.