



Gigabit Ethernet Jumbo Frames

And why you should care

Phil Dykstra
Chief Scientist
WareOnEarth Communications, Inc.
phil@wareonearth.com
20 December 1999

Whether or not Gigabit Ethernet (and beyond) should support frame sizes (i.e. packets) larger than 1500 bytes has been a topic of great debate. With the explosive growth of Gigabit ethernet, the impact of this decision is critically important and will affect Internet performance for years to come.

Most of the debate about jumbo frames has focused on local area network performance and the impact that frame size has on host processing requirements, interface cards, memory, etc. But what is less well known, and of critical concern for high performance computing, is *the impact that frame size has on wide area network performance*. This document discusses why you should care, and about the largely ignored but important impact that frame size has on the wide area performance of TCP.

How jumbo is a jumbo frame anyway?

Ethernet has used 1500 byte frame sizes since it was created (around 1980). To maintain backward compatibility, 100 Mbps ethernet used the same size, and today "standard" gigabit ethernet is also using 1500 byte frames. This is so a packet to/from any combination of 10/100/1000 Mbps ethernet devices can be handled without any layer two fragmentation or reassembly.

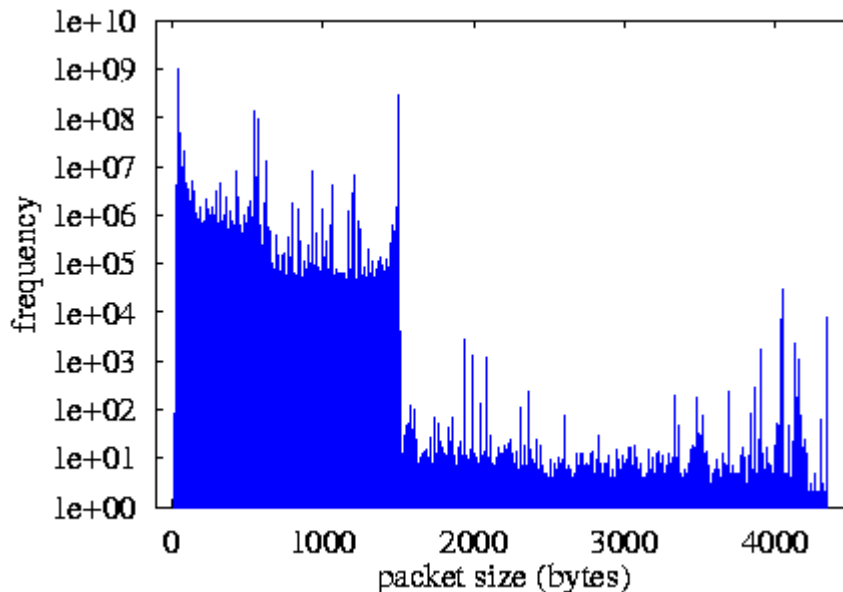
"Jumbo frames" extends ethernet to 9000 bytes. Why 9000? First because ethernet uses a 32 bit CRC that loses its effectiveness above about 12000 bytes. And secondly, 9000 was large enough to carry an 8 KB application datagram (e.g. NFS) plus packet header overhead. Is 9000 bytes enough? It's a lot better than 1500, but for pure performance reasons there is little reason to stop there. At 64 KB we reach the limit of an IPv4 datagram, while IPv6 allows for packets up to 4 GB in size. For ethernet however, the 32 bit CRC limit is hard to change, so don't expect to see ethernet frame sizes above 9000 bytes anytime soon.

How can jumbo frames and 1500 byte frames coexist?

Two basic approaches exist:

- On a port by port basis, where everything "downstream" from a given port is known to support jumbo frames.
- Using 802.1q Virtual LANs, where jumbo frame and non-jumbo frame devices are segregated to different VLANs.

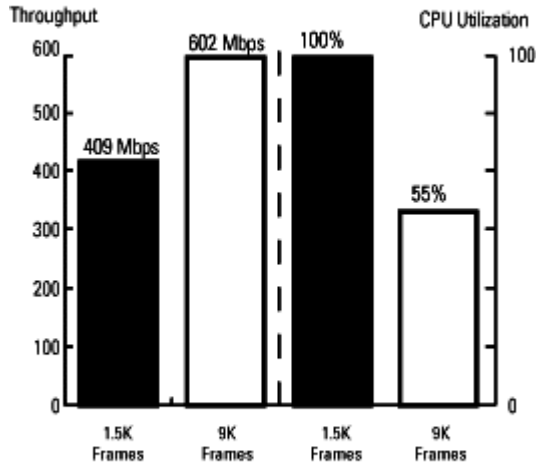
What frame sizes are actually being used?



The above graph is from a study[1] of traffic on the InternetMCI backbone in 1998. It shows the distribution of packet sizes flowing over a particular backbone OC3 link. There is clearly a wall at 1500 bytes (the ethernet limit), but there is also traffic up to the 4000 byte FDDI MTU. But here is a more surprising fact: while the number of *packets* larger than 1500 bytes appears small, more than 50% of the *bytes* were carried by such packets because of their larger size. Also, the above traffic was limited by FDDI interfaces (thus the 4000 byte limit). Many high performance flows have been achieved over ATM WAN's offering 9180 byte MTU paths.

Local performance issues

Extended Ethernet Frames vs. Standard Ethernet Frames*



* Using Gigabit Ethernet. Throughput on tests was limited to SBus capacity. TCP tests used dual 300 Mhz Sun servers running Solaris 2.5.1

Smaller frames usually mean more CPU interrupts and more processing overhead for a given data transfer size. Often the per-packet processing overhead sets the limit of TCP performance in the LAN environment. The above graph, from a white paper[2] by Alteon is an often cited study showing an example where jumbo frames provided 50% more throughput with 50% less CPU load than 1500 byte frames.

Such local overhead can be reduced by improved system design, offloading work to the NIC interface cards, etc. But however you feel about these often debated local performance issues, it is the WAN that we are most concerned about here.

WAN TCP performance issues

The performance of TCP over wide area networks (the Internet) has been extensively studied and modeled. One landmark paper by Matt Mathis et al.[3] explains how TCP throughput has an upper bound based on the following parameters:

$$\text{Throughput} \leq \sim 0.7 * \text{MSS} / (\text{rtt} * \text{sqrt}(\text{packet_loss}))$$

So maximum TCP throughput is directly proportional to the Maximum Segment Size (MSS, which is MTU minus TCP/IP headers). All other things being equal, you can double your throughput by doubling the packet size! This relationship seems to have escaped most of the arguments surrounding jumbo frames. [Packet_loss may also increase with MSS size, but does so at a sub-linear rate, and in any case has an inverse square effect on throughput, i.e. MSS size still dominates throughput.]

In the local area network or campus environment, rtt and packet loss are both usually small enough that factors other than the above equation set your performance limit (e.g. raw available link bandwidths, packet forwarding speeds, host CPU limitations, etc.). In the WAN however, rtt and packet loss are often rather large and **something that the end systems can not control**. Thus their only hope for improved performance in the wide area is to use larger packet sizes.

Let's take an example: New York to Los Angeles. Round Trip Time (rtt) is about 40 msec, and let's say packet loss is 0.1% (0.001). With an MTU of 1500 bytes (MSS of 1460), TCP throughput will have an upper bound of about 6.5 Mbps! And no, that is **not** a window size limitation, but rather one based on TCP's ability to detect and recover from congestion (loss). With 9000 byte frames, TCP throughput could reach about 40 Mbps.

Or let's look at that example in terms of packet loss rates. Same round trip time, but let's say we want to achieve a throughput of 500 Mbps (half a "gigabit"). To do that with 9000 byte frames, we would need a packet loss rate of no more than 1×10^{-5} . With 1500 byte frames, the required packet loss rate is down to 2.8×10^{-7} ! While the jumbo frame is only 6 times larger, it allows us the same throughput in the face of 36 times more packet loss.

But aren't jumbo frames bad for multimedia?

For applications that are sensitive to burst drops, delay jitter, etc., it can be argued that large frames are a bad idea. No application *has* to use large frames however, so the question is really whether other application's large frames will negatively impact your application's small ones. This is primarily an issue of slot time, i.e. how much will a large packet delay (or quantize) the time(s) available to transmit the small packets.

A 9000 byte GigE packet takes the same amount of time to transmit as a 900 byte fast ethernet packet or a 90 byte 10 Mbps ethernet packet. So jumbo frames on gigabit ethernet at worse add less delay variation than 1500 byte frames do on slower ethernets. And no one is suggesting that slower ethernets use 9000 byte frames. As for queuing delay concerns, that could happen whether packets are large or small. If delivery QoS is required, then the routers need to implement some kind of priority or expedited forwarding, regardless of the packet sizes. Tiny frames (including 53 byte ATM cells) may be helpful when multiplexing lower bit rate streams, but they become increasingly ridiculous on gigabit and beyond links.

Does GigE have a place in a NAP?

Not if it reduces the available MTU! Network Access Points (NAPs) are at the very "core" of the internet. They are where multiple wide area networks come together. A great deal of internet paths traverse at least one NAP. If NAPs put a limitation on MTU, then all WANs, LANs, and end systems that traverse that NAP are subject to that limitation. There is nothing the end systems could do to lift the performance limit imposed by the NAP's MTU. Because of their critically important place in the internet,

NAPs should be doing everything they can to remove performance bottlenecks. They should be among the most permissive nodes in the network as far as the parameter space they make available to network applications.

The economic and bandwidth arguments for GigE NAPs however are compelling. Several NAPs today are based on switched FDDI (100 Mbps, 4 KB MTU) and are running out of steam. An upgrade to OC3 ATM (155 Mbps, 9 KB MTU) is hard to justify since it only provides a 50% increase in bandwidth. And trying to install a switch that could support 50+ ports of OC12 ATM is prohibitively expensive! A 64 port GigE switch however can be had for about \$100k and delivers 50% more bandwidth per port at about 1/3 the cost of OC12 ATM. The problem however is 1500 byte frames, but GigE with jumbo frames would permit full FDDI MTU's and only slightly reduce a full Classical IP over ATM MTU (9180 bytes).

A recent example comes from the Pacific Northwest Gigapop in Seattle which is based on a collection of Foundry gigabit ethernet switches. At Supercomputing '99, Microsoft and NCSA demonstrated HDTV over TCP at over 1.2 Gbps from Redmond to Portland. In order to achieve that performance they used 9000 byte packets and thus had to *bypass* the switches at the NAP! Let's hope that in the future NAPs don't place 1500 byte packet limitations on applications.

What about GigE on the campus?

The Gartner Group predicts that 95% of all large-enterprise LAN backbones will be based on high-speed ethernet technology by 2002. Cost, bandwidth, compatibility, and easy administration are all driving this. So the technology will be there, but it shouldn't come at the cost of our future wide area performance.

If you want high performance, the best network design advice that I can give for a campus, regardless of the networking technology being used is this: *Every host in the campus should have a path between it and the wide area network that,*

1. *does not reduce the link bandwidth, and*
2. *does not reduce the MTU.*

So if, e.g. a host has an OC3 ATM interface running Classical IP, there should be an end to end path between it and the WAN of at least OC3 speed and that supports at least a 9000 byte MTU; every FDDI host should have at least a 100 Mbps 4000 byte MTU path, etc. Wherever you have installed jumbo frame GigE, you should have a jumbo frame path to (and through) the Internet.

Summary: Gigabit Ethernet needs Jumbo Frames

If you intend to leave the local area network at high speed, the dynamics of TCP will require you to use large frame sizes. Without them, the packet loss rate over a high

bandwidth-delay product path would have to be extraordinarily low. Core internet infrastructure, from campus backbones to Network Access Points (NAPs), should be particularly careful not to limit the permitted MTU to 1500 bytes. In the long run there is no reason to stop at 9000 byte frames, but given the current ethernet CRC limitation it is a good evolutionary step for gigabit data rates.

References

- [1] [the nature of the beast: recent traffic measurements from an Internet backbone](#)
- [2] [Extended Frame Sized for Next Generation Ethernets](#) - a white paper by Alteon
- [3] [The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm](#) - the performance equation used above
- [4] [Jumbo Frames? Yes!](#)
- [Matt Mathis on Raising the Internet MTU](#)

About the Author

Phil Dykstra is the Chief Scientist of WareOnEarth Communications, Inc., and heads the San Diego office which is focused on the measurement and analysis of high performance networks. Prior to WareOnEarth, he was the head of Advanced Development in the Army Research Laboratory High Performance Computing Division. With over 20 years of Internet and HPC experience, he fostered US Federal Networking for many years, and until recently, co-chaired the Joint Engineering Team (JET) which coordinates Next Generation Internet (NGI) and Internet2 engineering and plans. He has taught Computer Science courses at Johns Hopkins University and the University of Delaware.