Perils of the Network

A Series on Network Idiosyncrasies and Degradations

Size Matters:

Using Jumbo Packets on Gigabit Ethernet

appareNet

First published • September 2003

Apparent Networks

Canadian Head Office • 400 - 321 Water Street • Vancouver, British Columbia, Canada, V6B 1B8 • http://www.apparentnetworks.com

Company

Apparent Networks is a leading innovator of network intelligence software. Apparent Networks' technology, AppareNet, a network intelligence system, operates non-intrusively on live networks, to and from any location worldwide. Without requiring specialized hardware or remote agents, AppareNet views the network from the application's perspective. In doing so, AppareNet rapidly identifies the locations and causes of network bottlenecks anywhere in the world so that companies can boost the performance of, and gain more value from, the network infrastructure they already have. Apparent Networks improves its customers' businesses by helping organizations reduce operational costs, increase IP availability, and protect revenues.

Contact Information

Canadian Head Office

The Hudson House
400 - 321 Water Street
Vancouver, BC
Canada V6B 1B8

Sales: 1.800.508.5233
Support: 1.800.664.4401
Telephone: 604.433.2333
Fax: 604.433.2311

http://www.apparentnetworks.com
marketing@apparentnetworks.com

Technical Series

This is the fifth in a technical series of white papers from Apparent Networks examining the Perils of the Network. This series explains network idiosyncrasies and degradations and how AppareNet is capable of identifying these network problems.

The first paper in this series, The Apparent Network, introduced the concept of the apparent network as a complete end-to-end view of a network. The Apparent Network and Maximum Transmission Unit (MTU) papers are recommended background reading for this paper, which discusses both the benefits and hidden perils of using jumbo packets on Gigabit Ethernet.

Introduction

Gigabit Ethernet (GigE) has rapidly gained prominence and acceptance as the next step in the evolution of corporate networks. Relatively low-cost, high-speed, and interoperable with today's de facto standard, 100 Mbps Fast Ethernet, are just a few of GigE's promises. For many network planners, it is really only a matter of time before they adopt GigE, not only for their core networks, but also to the desktop.

Although Gigabit Ethernet is interoperable with 10/100 Mbps, there are some important differences that bear careful consideration. One of the most important is the **absence of any standard Maximum Transmission Unit** or MTU. The 1500 byte standard MTU of 10 and 100 Mbps networks has been replaced with no standard at all. Packets on Gigabit Ethernet can be any size supported by network vendors, varying from 1500 bytes to over 16000 bytes. Vendors are constrained by the component manufacturers who typically limit largest supported frame size to around 9000 bytes.

The benefits of so-called **jumbo** packets are significant – **jumbo packets can more than double accessible bandwidth** on today's computers compared to using smaller 1500 bytes packets, and yet there are some hidden perils. Due to the lack of standard MTU values, MTU conflicts may hamper 100-Mbps-to-Gigabit transitions. For example, various forms of MTU conflict, such as **black holes, can devastate network performance**.

These conflicts may prove to be more devastating and elusive than duplex mismatches found on 100 Mbps Ethernet. However MTU-related network issues are entirely avoidable with some careful planning such as adopting a rigorous approach to interface MTU assignment and applying a **widely used MTU convention like 9000 bytes** on all jumbo LAN/WAN links.

Why Jumbo MTU?

The term "jumbo" has typically been applied to any network unit (frame, packet, MTU) that is greater than the 10/100 Mbps Ethernet standard – at Layer 3 (packets and MTU), the standard size is 1500 bytes; at Layer 2 (frames and frame size), it is 1518 bytes. Some network researchers refer to jumbo packets as **jumbograms.**

> In GigE, there is no standard MTU - vendors have subsequently chosen from a range of sizes, anywhere from 1500 to 16128 bytes and beyond.

Jumbo packets are one of the obvious differences between 100 Mbps and GigE. However, there is also a looming issue in that Gigabit Ethernet standard has no default Maximum Transmission Unit (MTU). MTU is a Layer 3 parameter that controls the maximum packet size allowed on the network. For 10 and 100 Mbps Ethernet, the standards (RFC 894, 895) clearly set the largest MTU to 1500 bytes and almost all Ethernet interface cards defaulted to it.

So, what is the interest in using jumbo packets on Gigabit Ethernet?

There are two simple answers to this question:

1. In current implementations, GigE data transfer performance is strongly dependent on MTU – recent studies have shown that jumbo packets permit most hosts to send data at much higher transfer rates than the smaller 1500 byte packets.

2. Lack of standardized MTU in GigE networks can result in MTU conflicts, even in networks that ostensibly only use the 1500 byte 10/100 standard.

Technical Overview of MTU

The Maximum Transmission Unit (MTU) of a link is the largest packet a particular interface can accommodate without fragmentation. Fragmentation is the process of breaking large packets into smaller packets, usually to accommodate a particular maximum packet size. Each network interface has its own MTU assigned as part of its configuration. Any two interfaces that are connected together should negotiate use of the smallest MTU allowed between them. By extension, a particular network path between two hosts, consisting of one or more links, has a characteristic MTU referred to as **path MTU** – this is the lowest MTU of any network interface on that path.

Consider an example IP network with three segments connecting four nodes (Layer 3 devices such as routers or gateways), labeled A, B, C, and D. Nodes A and D only have one interface each, but B and C have at least two. Each interface has its own MTU setting with any two linked interfaces set to the same value.

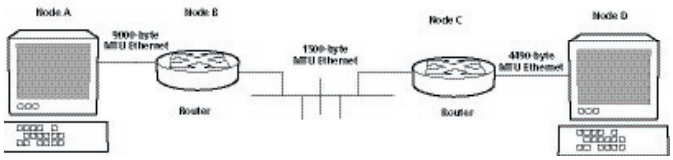


Figure 1

In Figure 1, the path MTU from A to D is 1500 bytes. Regardless of any larger MTU links, such as the jumbo MTU of the AB or CD links, the path MTU is the smallest of the three segments. The MTU of the BC link constrains all network traffic on the path from A to D to 1500 bytes. When a packet arrives at B that is larger than the 1500-byte MTU of the BC link, it can either fragment the packet into several smaller pieces or drop the packet completely.

Mid-Path Device Fragmentation

While fragmentation is considered a normal part of network design (RFC 791), it poses a significant burden on the mid-path device performing the fragmentation. It also increases network utilization without increasing the amount of application data transmitted. Further, it increases the processing requirements of downstream network devices - splitting a packet into two or more fragments is a processing intensive operation, especially for devices with limited processing power like smaller or older routers. Each fragment requires its own IP and lower layer headers and check-sums that the mid-path device is required to generate.

Another sometimes unexpected consequence for mid-path packet fragmentation is network border packet re-assembly. If an organization implements any type of connection tracking (e.g. a stateful firewall) at their network border, they will need to de-fragment all packets at the border point. The connection information is contained in layers 4 and above, which are missing from all but the first packet fragment. The connection tracking device must collect all the packet fragments making up the original packet and reassemble them into one packet in order to perform its connection tracking task.

Fragmentation, while conforming to the design standards, is considered undesirable and to be avoided wherever practical. Typically, only "ping" and router advertisements omit the IP Header "Do Not Fragment" flag, thereby allowing fragmentation. Hosts transmitting data typically try to determine the correct path MTU before sending packets to their intended destination.

RFC 1191 - Path MTU Discovery

RFC 1191[1] describes the standard process for hosts to discover the path MTU to some other host. It requires that any interface sending a packet set the IP Header "Do Not Fragment" flag, and that any interface receiving a packet that is too large for transmission, drop the packet and send an ICMP "Fragmentation needed but DF Set" message back to the originating host with the required MTU indicated. By sending oversize packets to the end-host and receiving these ICMP messages back from intermediate interfaces, a transmitting host can discover the path MTU to a specific end-host, and adjust its traffic patterns accordingly.

Mid-Path MTU Induced Intentional Packet Loss

Not all network devices properly implement RFC 1191. In many cases, such as with Layer 2 switches and bridges, packets are simply dropped. This is permitted by network design but it makes path MTU discovery unreliable.

Packet loss consumes network bandwidth that does not translate into application data transmission. In situations where the sending host cannot differentiate MTU induced packet loss caused by congestion, the application experiences lower network performance[2]. TCP Slow-Start mechanisms interact with MTU conflict paths, allowing the occasional packet to survive. Therefore, paths with MTU conflicts typically experience a 5 to 10 fold increase in transfer times when transferring large data blocks. In extreme cases, the result is complete network failure for certain applications.

Standards

While network failure due to MTU conflicts may sound overly dramatic, the declining use of protocols like FDDI and Token Ring may be a related consequence. Their fate has been attributed by some veteran engineers to the then-growing dominance of Ethernet and constant problems interfacing FDDI and Token Ring to other networks. Without a doubt, many of those **problems** were MTU conflicts arising from the differing MTUs at the interfaces.

The decline of FDDI and Token Ring may be linked to MTU conflicts at the interfaces between differing MTUs.

Each networking implementation typically has a standard MTU associated with it. Table A shows many of the most familiar standard values and their related RFCs:

RFC #	Description	MTU
894	Minimally required	68
1051	ARCNet	508
879, 1356	X.24, ISDN	576
1055	Serial Line IP (SLIP)	1006
1042, 2516	IEEE 802.3/802.2, PPPoE	1492
894, 895	**Ethernet**	**1500**
1390	FDDI	4352
1042	4 Mbit Token Ring	4464
1042	802.4 Token Bus	8166
IBM	16 Mbit Token Ring	17914
1374	HIPPI	65535

Table A

Usually the Layer 3 devices interfacing between different technologies handle the difference in MTU. But if they are not configured properly or Layer 2 devices are used between MTU boundaries, MTU conflicts can occur, leading to severe performance degradation.

MTU-Dependent Network Performance

One of the primary reasons for considering jumbo MTU is the performance advantage. Typically, a workstation or server will NOT see the full value of a GigE connection if it is restricted to 1500 byte packets. This is due to several factors that are not immediately apparent. First, let's look at some test data that demonstrates the range of performance that can be expected.

In a series of experiments[3] that were conducted in mid-2002 and mid-2003 by members of the Advanced Test Engineering and Measurement (ATEAM), performance measurements were taken across the Abilene and CA*net4 backbone networks between a variety of GigE-equipped, jumbo-enabled hosts. They were located across North America (both Canada and the U.S.) and were equipped with a number of performance analysis tools. The goal of the experiments was to uncover any inherent dependencies between network performance and MTU.

MTU Performance Project Members:

- John Moore, NCSU/Centaur Labs
- Kevin Walsh, SDSC/CalNGI
- William Rutherford, BCIT/iEL
- Loki Jorgenson, Apparent Networks

The measurement systems used included SmartBits, iPerf and AppareNet – each system measured the network performance relative to MTU using different methodologies. SmartBits is a hardware tool that provides unparalleled control and precision in the transmission and reception of test packets – a unit is placed at each end of a end-to-end path and the path is flooded to measure the rates of transfer. iPerf is a widely used network performance tool available as free-use software distributed under copyright by the University of Illinois - it floods the network similar to SmartBits to get performance measurements in several protocols. AppareNet is also a software-based system that uses an active sampling technique with several protocols, coupled with a sophisticated mathematical modeling engine, to determine a range of performance characteristics.

In these tests, the broadest range of results was derived from AppareNet – the Maximum Achievable 2-way Bandwidth was measured at the IP layer as a function of packet size between 22 hosts. It was varied from 512 bytes all the way up to the maximum MTU supported by the path (typically 9000 bytes). Measurements were made between all test hosts, and also to other known GigE/jumbo-enabled hosts. Measurements were two-way, measuring the bandwidth in both directions – in other words, for an optimal full-duplex GigE path, the bandwidth should be almost 2000 Mbps. Figure 2 shows a distribution of the measurements.

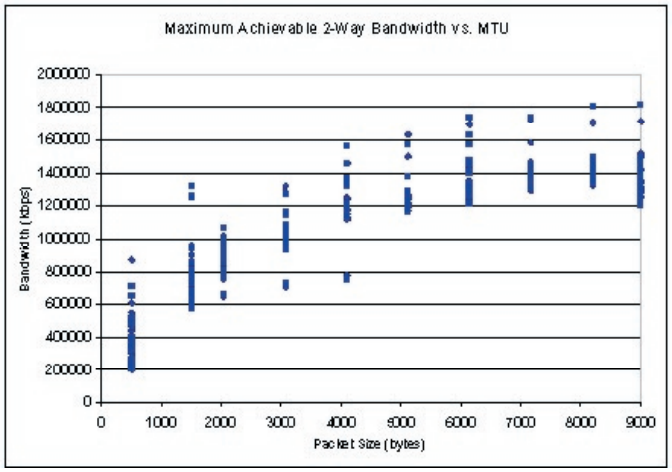
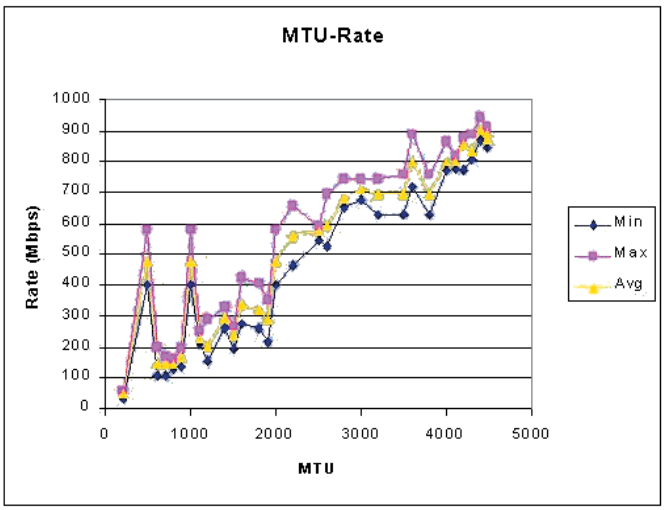


Figure 2: AppareNet Bandwidth measures between 20 different jumbo/GigE hosts

The range of values for each packet size show how each different host varied in its responsiveness. The best performance clearly occurs for the largest packets (around 9000 bytes). The performance at 1500 bytes is significantly poorer, somewhere around 40% of the rated capacity.

The Both TCP and IP performance show dramatic improvements using jumbo packets (9000 bytes) over 10/100 Mbps standard sized packets (1500 bytes), on end-to-end Gigabit Ethernet LAN and WAN.

Tests performed with iPerf show similar results. Another measurement project at Pittsburgh SuperComputing led by Raghu Reddy and Matt Mathis produced[4] 1-way TCP-specific measures using iPerf between PSC and Arlington, VA on a GigE path. The data in Figure 3 shows a very similar trend toward full saturation of the Gigabit link as the packet size approaches 4470 bytes. At 1500 bytes, accessible bandwidth was severely limited to around 30% of its capacity.



Reprinted with the permission of Raghu Reddy.

Figure 3: iPerf TCP data transfer measures between PCS and ISI-East

End-to-End Performance

Network performance is not simply defined by the intermediate devices such as hubs, switches and routers. End-to-end includes the influences of the networking elements of the host at each end. On a clean GigE path, the CPU, bus, and network interface card (NIC) of the hosts are the most likely limiting factors of performance.

Typically the intermediate devices can forward packets at the rated capacity since they were primarily designed for the task. But servers and workstations are not solely designed for sending and receiving packets and current hardware simply cannot keep up. Different CPUs, with different busses, NICs, drivers and OSes, offer significantly different responses. The range of apparent bandwidth in Figure 2 shows that even carefully configured high-end workstations can vary as much as 20%.

Choosing Jumbo MTU for Performance

The obvious result from these experiments is that network performance clearly depends on the size of the packets. The reason for that dependence rests with the end-host. A SmartBits or similarly dedicated hardware can put almost any size packets to the wire at full rate of transfer. However general-purpose computers, and even high-end servers, are limited by the rate they can process packets. Smaller packets mean more packets per second. When the machine has reached its processing limit, it cannot transmit or receive packets any faster even if there is additional unused capacity. So, using larger packets can give a host access to more bandwidth by avoiding that limitation.

> A network device is limited by the rate at which it can process packets, regardless of its data transfer capacity. If that limitation impacts performance, using larger packets can give a host access to more bandwidth.

Hidden Perils in Gigabit Ethernet

Whether choosing to use jumbo MTU or staying with legacy 1500 byte MTU, the proliferation of GigE will require that careful attention be paid to MTU. In the days of 10/100 Mbps networks, the default 1500 bytes could be assumed and MTU was rarely a topic of concern. However, with no standard/default in the GigE world and not even any agreed industry conventions to rely on, MTUs of GigE-capable interfaces may be set to almost anything. This opens the door to a variety of problems related to MTU conflicts.

A uniform approach to path MTU discovery would take care of most of those problems. Unfortunately, RFC 1911 is often not properly considered in the design of jumbo-frame networks. If not properly configured, or if ICMP is indiscriminately blocked on Layer 3 interfaces, necessary ICMP will not find their way to a source interface. These devices are referred to as **black holes** for the obvious reasons. Layer 2 devices (e.g. switches) do not generate ICMP at all and, if used at the boundary between differing MTUs, can introduce this black hole effect. Older, poorly implemented VPNs are another common source of black hole behavior, since they lower the native MTU by introducing additional header information on the standard IP packet.

Another phenomenon regularly encountered is a type of interface that reports its MTU incorrectly – this is referred to as a **grey hole**. By responding properly but with a misleading MTU, the interface can do even more damage than a black hole. Certain older VPNs and some implementations of X25 networks are typically responsible for this behavior.

Older VPNs, Layer 2 devices like switches and bridges, and ICMP blocking can generate MTU conflicts such as Black Holes and Grey Holes on 10/100 Mbps and GigE networks. Greater use of jumbo packets could make these problems more common.

For more information on MTU conflicts, see the whitepaper available from Apparent Networks entitled "Maximum Transmission Unit: Hidden Restrictions on High Bandwidth Networks"[5].

Best Practices for Deploying GigE

MTU conflicts can be avoided. And where desired, the full benefit of jumbo packets can be optimally achieved. Effective deployment of GigE must be addressed end-to-end with careful consideration of all the potential issues. For example, core networks that are composed of non-GigE technology (e.g. POS, ATM, MUXes) must be able to handle the larger packets as well.

With this awareness and commonsense, implementation of appropriate network policies can ensure an effective deployment. Typical policies could include:

- choose a "standard" jumbo MTU – preferably one that is consistent with others who have already deployed – 9000 byte MTU (9018 frame size) is recommended
- require that boundaries between differing MTUs be handled by Layer 3 devices – never allow switches to support paths between different MTUs
- specify an MTU requirement in network RFI/RFQs
- associate a single MTU value for each Layer 3 subnet, and ensure that all interfaces within a subnet has the same MTU value
- ensure that MTU is explicitly labeled on all network diagrams just as netmasks and broadcast masks are, including Layer 2 devices like switches - it might also be recommended that hardware be physically labelled as well
- mandate the creation of Layer 3 logical network diagrams, separate from physical diagrams - for each subnet include information regarding router addresses, IP address ranges, subnet masks, routes and MTUs.
- define guaranteed end-to-end MTU for each MTU domain as opposed to an interface-by-interface basis
- ensure that intermediate devices support somewhat larger packets to allow for seamless VPN integration (e.g. 9000 bytes end-to-end but 9180 supported in the core)
- regularly monitor the path MTU on critical paths
- be specific in promoting a "preferred" end-to-end MTU to users and customers - stress that end-hosts should be either at 1500 or some specific jumbo size such as 9000 bytes
- do not indiscriminately block ICMP packets
- encourage manufacturers toward even larger MTUs (>9000 bytes) in future hardware releases

Models for Deployment

The academic networks continue to be at the forefront of network deployment. It is worthwhile to examine their experience closely to see how new technologies may work in your networks. Jumbo packets are now widely supported in many major academic backbones including Abilene (Internet2) in the United States, CA*net4 (CANARIE) in Canada, SURFnet (Stichting SURF) in Holland, and soon (2004) AARnet in Australia. Each has carefully considered how to deploy jumbo MTU most effectively and developed appropriate policies and deployment plans. Best of all, as public networks they offer access to their planning and deployment information.

The jumbo MTU of 9000 bytes has surfaced as a general convention. It has been cited for its flexibility (8*1024 = 8192 plus some extra room for Layer 4-7 headers), being easy to remember, and common to the greatest range of vendors' hardware.

A variety of on-line documents are available from each organization and from researchers working on MTU-related topics:

- http://aarnet.edu.au/network/mtu
- http://www.ncne.org/jumbogram
- http://darkwing.uoregon.edu/~joe/jumbo-clean-gear.html
- http://www.abilene.iu.edu/JumboMTU.html
- http://sd.wareonearth.com/~phil/jumbo.html
- http://www.psc.edu/~mathis/MTU/
- http://darkwing.uoregon.edu/~joe/jumbos/

In addition, there are examples of distributed path MTU monitoring and diagnostics tools available:

- http://pathmtu.apparenet.com:8282
- http://www.ncne.org/jumbogram/mtu_discovery.php

Jumbo Jumbo

Not only are jumbo MTU here to stay, network planners can expect that jumbo packets will only get bigger. Internet2 researcher Matt Mathis has predicted that, with the advent of 10, 40, and eventually 100 Gbps in the near future, packets will need to be even larger than 9000 bytes. He points at **time**, not size, as the critical consideration. His view is that the time between packets should be kept constant.

> Jumbo packets will continue to get bigger. Packets will need to be on the order of Mbytes to keep up with emerging high-speed networks.

As network speeds increase, the time between packets of the same size gets smaller and smaller. Interfaces handling the packets are forced to deal with more packets, and more headers, in shorter and shorter time frames. Each jump in Ethernet technology, from 10 to 100 to 1000 and now 10000 Mbps, has decreased packet times by an order of magnitude. In order to keep the packet times constant, MTU should have increased to 12,000 bytes for 100 Mbps and now be 96,000 bytes for GigE. This trend has MTUs increasing to **over 50 Mbytes** for the anticipated Terabyte standards.

Actual				Proposed		Alternate	
Rate	Year	MTU	Packet Time	MTU	Packet Time	MTU	Packet Time
10 Mbps	1982	1.5 KBytes	1200 µsec	1.5 KBytes	1200 µsec		
100 Mbps	1995	1.5 KBytes	120 µsec	12 KBytes	960 µsec		
1 Gbps	1998	1.5 KBytes	12 µsec	96 KBytes	768 µsec	64 KBytes	525 µsec
10 Gbps	2002	1.5 KBytes	1.2 µsec	750 KBytes	600 µsec	625 KBytes	500 µsec
100 Gbps				6 MBytes	480 µsec	6.25 MBytes	500 µsec
1 Tbps				50 MBytes	400 µsec	62.5 MBytes	500 µsec

Reprinted by permission from Matt Mathis

However, as it stands today, there are few interfaces that offer much greater than 9,000 bytes. And the market hype has almost entirely ignored MTU as a consideration, preferring to assume the 1500 byte standard for GigE. But for how much longer can this trend persist?

Conclusion

When, not if, you deploy Gigabit Ethernet, you will want to give careful consideration to performance issues. Your biggest instant improvement in performance comes from using jumbo MTU. If you use jumbo packets, choose your MTU carefully and apply it uniformly – 9000 bytes is recommended. And where you don't apply jumbo MTU with your Gigabit networks, be particularly fastidious about locking interfaces to 1500 bytes. Separate your MTU boundaries with RFC 1191 compliant routers, and ensure that ICMP messages are enabled. Finally, monitor the path MTU of your critical network paths on an ongoing basis and remember to check MTU when troubleshooting.

For further information on AppareNet, or to see it live, please contact us at marketing@apparentnetworks.com or toll free at 1 800 508 5233 or visit our website at www.apparentnetworks.com.

Endnotes

1 As of this publication, IETF has established a new working group to develop a new path MTU discovery specification.

2 See Apparent Networks whitepaper: *Traffic and Everything Like Traffic: Dealing with Network Performance Degradation;* http://www.apparentnetworks.com/main/whitepaper_traffic_may2003.pdf

3 http://www.ncne.nlanr.net/training/techs/2003/0803/presentations/0803-moore1_files/v3_document.htm

4 http://www.psc.edu/~rreddy/networking/mtu.html

5 See Apparent Networks whitepaper: *Maximum Transmission Unit: Hidden Restrictions on High Bandwidth Networks;* http://www.apparentnetworks.com/main/whitepaper_mtu_aug2002.pdf