Ethernet Jumbo Frames

Version 0.1

November 12, 2009

Contributors:

Dell:	*Robert Winter, Rich Hernandez, Gaurav Chawla*
Cisco:	*Anthony Faustini, Carl Solder, Thomas Scheibe*
3Com:	*David Law, Siamick Ayandeh*
Applied Micro:	*Brad Booth*
Ethernet Alliance:	*Blaine Kohl*
NetApp:	*Charlie Lavacchia*
Force 10:	*Subi Krishnamurthy, Raja Karthikeyan*
Intel:	*Eric Multanen*
Qlogic:	*Manoj Wadekar*

Ethernet Alliance | 3855 SW 153rd Drive | Beaverton, OR 97006

www.ethernetalliance.org

Motivation

This paper is intended as a resource to help better understand what Jumbo frames are, the implications of their use and the types of applications which drive their use. It is also recognition that Jumbo frames aren't going away….

What is a "Jumbo" Frame?

An Ethernet frame is defined as that unit of packetized formatted information which includes the Ethernet header, payload and Cyclic Redundancy Check (CRC) trailer. It is enclosed by the Start-of-Frame-Delimiter (SOFD) and the Inter-Frame-Gap (IPG) as shown in Figure 1. Note that the payload represents all the information enclosed by the Ethernet header and the CRC. The largest possible payload in a frame is called the Maximum Transmission Unit (MTU).

The original IEEE 802.3 specifications (Reference 2) defined a valid Ethernet frame size from 64 to 1518 bytes. The standard Ethernet header is 18 bytes in length and therefore the payload for a standard frame ranges in size from 46 to 1500 bytes. Since the original Ethernet specification was defined various IEEE standards have been developed that support additional, expanded frame types listed below and shown in Figure 1.

- The support for an additional 4 bytes in the Ethernet header for VLAN tagging (IEEE 802.1Q) increases the maximum Ethernet frame size to 1522 bytes.
- The Provider Bridge (802.1ad) Ethernet frame adds 8 bytes to the original frame to support service and customer tagging to increase the frame size to 1526 bytes.
- The work in 802.3as recognizes that due to various tag definitions in IEEE over the years the frame size should be increased and have called out a frame size value of 2000 bytes. The MTU size is still 1500 bytes.
- 802.3AE adds a prefix of 64 bytes to the frame increasing the standard frame size by that amount.
- T11 has agreed to an MTU of 2500 bytes for Fibre Channel over Ethernet (FCoE) frames.
- Multi-Protocol Label Switching (MPLS) increases the maximum Ethernet frame size to 1518 bytes + (n * 4 bytes), where n is the number of stacked labels.

When it comes to naming frame sizes Ethernet has a perplexing abundance of terms that serve to confuse those referring to non-standard sized MTU frames. Some of the terms used are:

- "Baby Giant" often refers to the frame type used with MPLS, 802.1Q, 802.1ad and 802.3AE.
- "Mini Jumbo" is often used to refer to an MTU size of 2500 bytes and has become specific to the frame size used by Fibre Channel over Ethernet (FCoE).
- "LINK MTU" is a term used by some vendors to indicate the total size of the Ethernet frame (headers and payload) and "PAYLOAD MTU" is a term sometimes used to indicate the total size of the payload within the frame.
- Ethernet Jumbo frames, also referred to as "Giants" or "Giant Frames" differ from "Baby Giants" but relate to "Mini Jumbos" in that they define an Ethernet Frame that carries more payload than the maximum specified by IEEE 802.3. See Figure 1. For standard Ethernet the MTU is 1500 bytes (References 1 and 2).

For clarity we define Jumbo frames as all frames that have MTUs larger than the standard, originally specified Ethernet payload size of 1500 bytes.

Jumbo frames have been around as long as Ethernet. There is no industry standard that defines the size of a Jumbo frame and this becomes evident when noting the different vendor viewpoints on just what a Jumbo frame is. Furthermore, IEEE has determined they will not support or define Jumbo frames due to concerns around vendor and equipment interoperability. This can make the use of Jumbo frames problematic especially in heterogeneous networks with equipment from multiple vendors. With that being said, there has been general consensus on what a standard Jumbo frame size should be. This is in part due to the decisions of some US Federal Government departments to accept a specific MTU as the maximum Jumbo frame size and in part due to de-facto usage scenarios that have developed over the last few years. One of the driving forces behind common Jumbo frame use is the stated requirement that FCoE frames have an MTU of 2500 bytes. Jumbo frames thus have thus become more relevant to any discussion of network architecture and deployment.

The multiple terms used to define Ethernet frame sizes coupled with the lack of any official guidance creates a confusing landscape of packets with sometimes unknown and invariably inconsistent MTU sizes.

Figure 1. Ethernet Frame Sizes

The Pros

Larger MTUs allow greater efficiency in data transmission since each frame carries more user data (payload or MTU) while protocol overhead and underlying per-packet delay remain fixed. There are many applications that can benefit from the use of Jumbo frames such as, but not limited to, the following:

- Server Clustering
- Server Backups (larger MTUs permit faster backups)
- High Speed Supercomputer Interconnect (for data transfer, not messaging)
- Network File Server (NFS) Protocol (9000 byte MTU to carry an 8192 NFS data block)
- iSCSI SANs (9000 bytes to reduce the effect of TCP frame overhead)
- FCoE SANs (2500 bytes to enclose an FC frame of 2000 bytes)

Extensive studies have been performed to analyze the impact of how MTU affects the performance of TCP. A common, well-known result based on findings of these studies (LargeMTU Study, Reference 6) provide a general conclusion that doubling the MTU

size can double the throughput of the TCP session assuming packet loss and round trip times are constant.

Sending data in Jumbo frames results in fewer frames being sent across the network. Processing fewer frames generates conservation of CPU cycles and thus greater throughput. Figure 2, from a well-known though now somewhat dated Alteon Networks study (Alteon Paper, Reference 4) shows the improvements on a 1 Gbps Ethernet network utilizing Jumbo frames with throughput gains and CPU savings of nearly 50%. A Jumbo frame of 9000 bytes is large enough to encapsulate a standard NFS (Network File System) data block of 8192 bytes, yet not large enough to exceed the 11,455 byte limit of Ethernet's error checking CRC (cyclic redundancy check) algorithm (Alteon Paper, Reference 4).

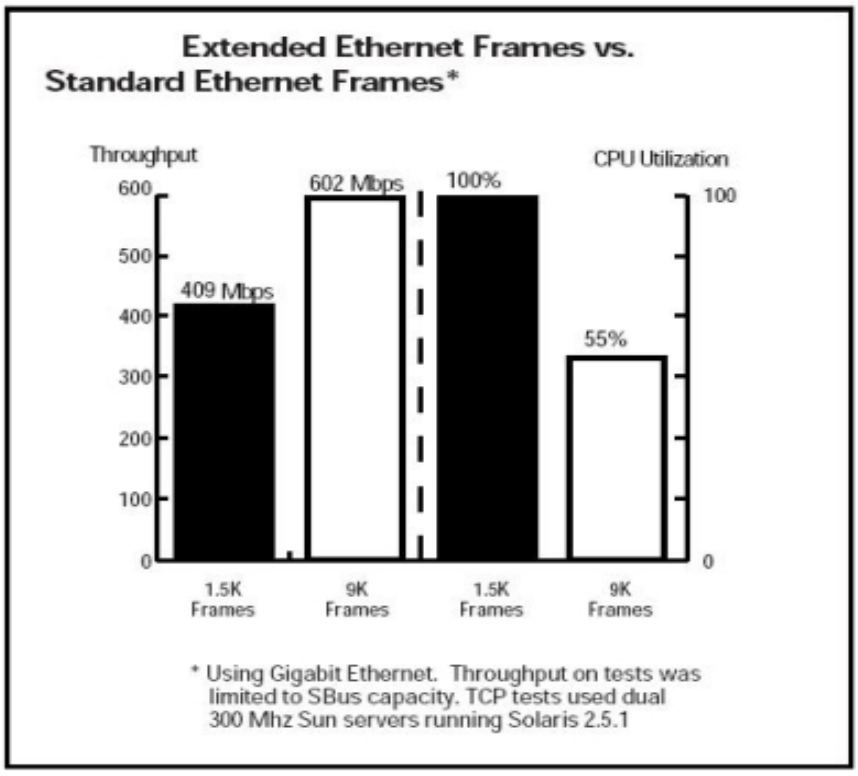

Figure 2. Alteon Analysis

The Alteon analysis states that:

"a single 9000 byte MTU jumbo frame replaces six 1500 byte MTU standard frames, producing a net reduction of five frames, with fewer CPU cycles consumed end to end. Further, only one TCP/IP header and Ethernet header is required instead of six, resulting in (5(40+18)) = 290 fewer bytes transmitted over the network."*

It takes over 80,000 standard Ethernet frames per second to fill a 1 Gbps Ethernet pipe, consuming significant CPU cycles and overhead. Sending the same data with 9000 byte MTU Jumbo frames, only 14,000 frames need to be generated, with the reduction in header bytes freeing up 4 Mbps of bandwidth. For 10 Gbps Ethernet the maximum frame rate is 812,744 frames per second using the standard 1500 byte MTU, while with 9000 byte MTU Jumbo Frames the maximum frame rate is 138,581 frames per second.

Figure 3 below illustrates the relative frame efficiencies (payload vs. overhead) of various protocols. Note that all values are very close but it is clear that larger MTUs provide a measureable benefit.

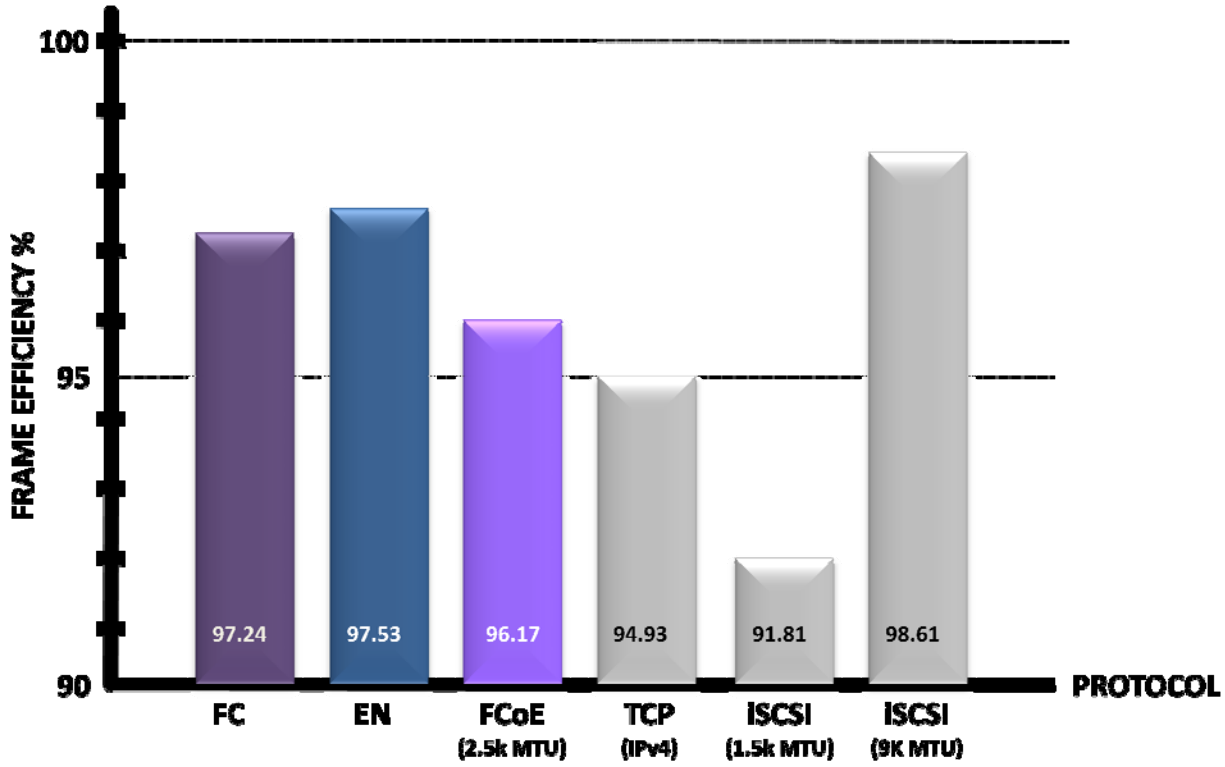

Figure 3. Frame Efficiencies (Payload to Protocol Overhead)

Figure 4, from a Dell Inc. study (Reference 8), illustrates performance benefits from using 9000 byte MTUs over 10 Gbps Ethernet. The system used was a Dell R910, dual CPU 6 core 1.73Ghz with Hyper-Threading, 16 GigaBytes of Ram and PCI-e Gen 2 slots. The Operating System used was Windows Server W2k8 R2. The performance benchmark used was Chariot.

This shows that MTU considerations are link speed agnostic as the benefits can be realized even as link speed increases, especially so for larger I/O block sizes (those above 1024 bytes) that are important for storage protocol traffic such as iSCSI and FCoE. MTU size selection thus remains important and relevant.

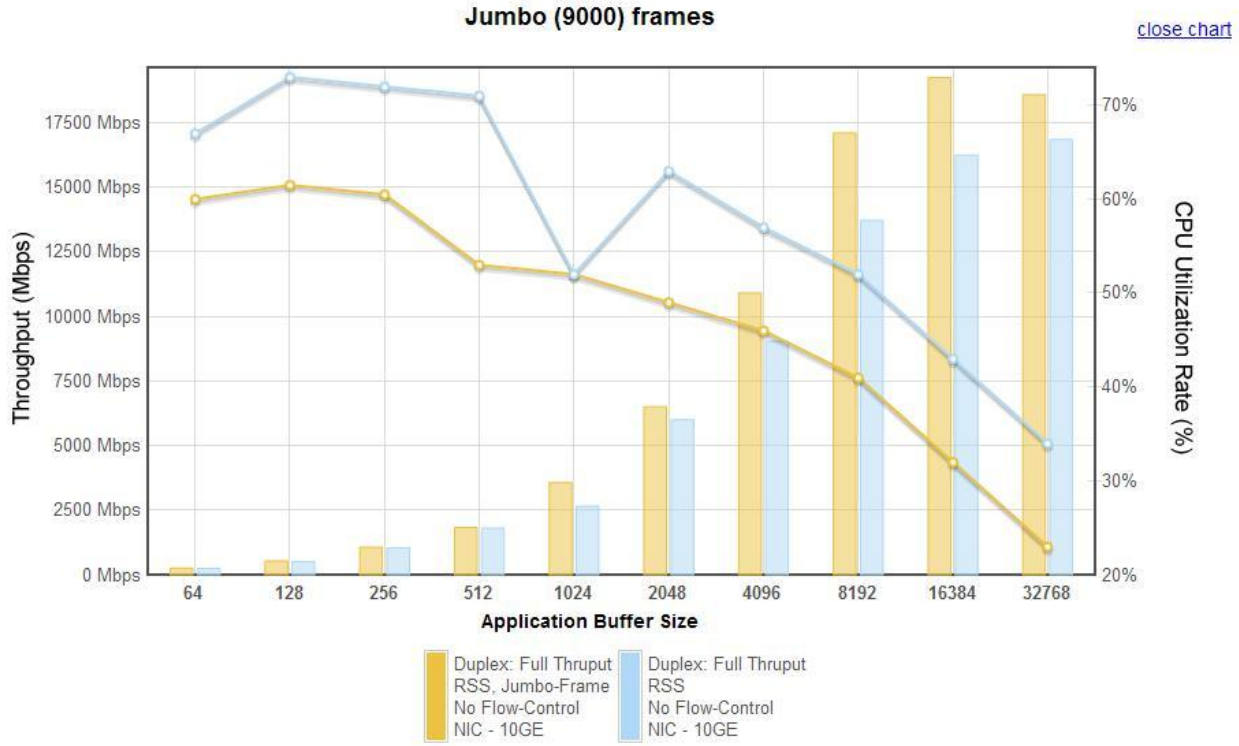

Figure 4. Dell Inc. Jumbo Frame Analysis

The Cons

Larger frames consume more Ethernet link transmission time, causing greater delays for those packets that follow and thus increasing lag time and latency. This could have a negative consequence for applications that require low latency and consist of smaller packet sizes such as Voice over IP or Inter-Process Communication (IPC). Frame transmission times are shown below in Table 2.

Transmission Time per Frame in Microseconds		
Link Speed, Gigabits per second (Gbps)	**1500 byte MTU frame**	**9000 byte MTU frame**
1 Gbps Ethernet	**12.00**	**72.00**
10 Gbps Ethernet	**1.20**	**7.20**
40 Gbps Ethernet	**0.30**	**1.80**
100 Gbps Ethernet	**0.12**	**0.72**

Table 2. Transmission Times

Larger frames may also fill available network equipment buffer queue memory at a faster rate and may interfere with the proper operation of such gear.

For example, an 802.1Qbb capable switch that supports per priority flow control on 8 queues requires that each queue buffer frames when it sends a PAUSE frame for a specific priority. This insures that there are enough buffers to store the incoming frames for the PFC enabled priority until the port at the other end of the link receives and processes the PAUSE frame. The following calculations assume that there may be one MTU sized buffer stored in the port queue and one MTU sized frame on the link that must be stored before the PAUSE is received at the other end of the congested link.

Reserved_Buffers = 2 * MTU_Size + link_delay * link_speed

For simplicity let's assume link_delay to be 0.

Since most switches have a common buffer pool shared across all ports,

Switch_Reserved_Buffers = Num_Ports * Num_PFC_Enabled_Queues * Reserved_Buffers

For a 24 port switch with 8 PFC capable queues and a 2500 byte MTU:

Switch_Reserved_Buffers = 24*8*(2*2500) = 960 KB

For a 24 port switch with 8 PFC capable queues and a 9000 byte MTU

Switch_Reserved_Buffers = 24*8*(2*9000) = 3456 KB

It is important to note that this is wasted memory in the switch, as it cannot be used to buffer the frames when there is no congestion. The switch has to reserve this amount of buffer space to be able to process the incoming frames when it sees congestion (sends a PAUSE frame).

With the 8 PFC queues set for 1 port of 2500 byte MTU, 1 port of 9000 byte MTU and 6 ports of 1500 byte MTU

Switch_Reserved_Buffers = 24*1*(2*2500) + 24*1*(2*9000) + 24*6*(2*1500)

= 120 + 432 + 432 = 984 KB

This leads to a savings of 3456 KB – 984 KB = 2472 KB compared to having 8 PFC queues per port supporting a 9000 byte MTU.

Therefore appropriate design guidance for PFC capable queues in a multi-port Ethernet switch is to assign MTU size per priority queue per port and not assign static MTUs for all ports and queues on the switch.

There also may be difficulties that relate to network stack, operating system and driver behavior when larger MTUs are used. Over several years protocol stacks have been tuned to expect 1500 byte MTUs. While CPU usage might be decreased in the network stacks if they were tuned for larger MTUs most stacks simply aren't and awkwardness in buffer allocation overhead might limit potential efficiencies that could be seen with larger MTUs.

The effective use of Jumbo frames requires that every link along the network path support the same Jumbo frame MTU or sets of Jumbo frame MTUs. There is no Layer 2 discovery mechanism commonly defined that accomplishes this. Therefore it is important to specify an MTU size across an organization for well-known types of traffic. Without such specification erratic network behavior may result. More specifically, fragmentation of the packet may occur when a given interface in the path of the frame is not capable of sending the full size Jumbo frame. Fragmentation often invokes a CPU burden which has a corresponding impact on other application data that is processed on the device.

The technique called out in RFC 1191, "Path MTU Discovery" (Reference 5) only works for ICMP enabled devices and requires modification of the TCP MSS (Maximum Segment Size) value to work. The technique called out in T11/09-251v1 only works for FCoE networks.

So, to summarize, the three negatives to using larger MTUs are:

- Increased Latency
- Switch Inefficiencies
- Operating System, Network Stack and Driver Inefficiencies
- Jumbo Frame Discovery is required

While all of these with the exception of latency may be overcome it does require MTU aware equipment and software for enablement as discussed in the following section.

Jumbo Frame Enablement

For the convergence and unification of fabrics it is necessary to support Jumbo frames in the network. It is entirely possible that a mixed storage area network (SAN) with

NFS, FCoE and iSCSI packets will need to support the standard 1500 byte MTU, the FCoE specified 2500 byte MTU and an iSCSI 9000 byte MTU within the same physical link at the same time.

The following areas of work need to be addressed in order to obtain the full benefit of Jumbo frame usage:

- Assignment of MTU size per priority queue in a DCB enabled switch would be an efficient design for network equipment supporting mixed MTU traffic. This would probably require DCBx negotiation to determine MTU size per priority on the link.
- Operating systems, network stacks and drivers should be MTU aware in that they can tune themselves to accept MTU sizes different from the standard size of 1500 bytes to drive their buffer allocation schemes.
- MTU discovery along Layer 2 pathways must be supported. While the method called out in T11/09-251v1 (Reference 5) works for FCoE and Path MTU discovery as outlined in RFC 1191 (Reference 7) works for IP enabled devices a purely Layer 2 mechanism has not been standardized. This needs to be done.

About Ethernet Alliance

The Ethernet Alliance is a community of Ethernet end users, system and component vendors, industry experts and university and government professionals who are committed to the continued success and expansion of Ethernet. The Ethernet Alliance brings Ethernet standards to life by supporting activities that span from incubation of new Ethernet technologies to interoperability demonstrations, certification and education.

References

1. "A Standard for the Transmission of IP Datagrams over Ethernet Networks", IETF STD0041, Charles Hornig, Symbolics Cambridge Research Center, 1984
2. "IEEE 802.1D MAC Bridges, 2002", section 3.2.7
3. "Extended Frame Sizes for Next Generation Ethernets", Alteon Networks White Paper
4. http://staff.psc.edu/mathis/MTU/AlteonExtendedFrames_W0601.pdf
5. From www.t11.org discussion of "FCoE Max Size" generated from T11/09-251v1, 04/27/2009, "FCoE frame or FCoE PDU"
6. http://www2.rad.com/networks/2003/largemtu/tcperf.htm
7. RFC 1191, "Path MTU Discovery", 1990
8. Internal Dell Inc. study, 2009