ETHERNET PACKET FILTERING for FTI – PART II

Ø. Holmeide[1], J-F. Gauvin[2]
1 OnTime Networks AS, Oslo, Norway
oeyvind@ontimenet.com
2 OnTime Networks AS, Oslo, Norway
jean.frederic@ontimenet.com

Abstract:

Network loads close to Ethernet wire speed and latency sensitive data in a Flight Test Instrumentation (FTI) system, represent challenging requirements for FTI network equipment. Loss of data due to network congestion, overflow on the end nodes, as well as packet latency above a few hundred microseconds, can be critical during a flight test. To avoid these problems, several advanced packet filtering and network optimization functions are required in order to achieve best possible performance and thus avoid loss of data.

This paper gives insight into how to properly engineer an Ethernet based FTI network and how to use advanced Ethernet switch techniques such as Quality of Service (QoS) and rate shaping.

Keywords: IP, UDP, filtering, QoS, traffic shaping, port trunking, LACP, FTI

Introduction

Ethernet in a Flight Test Instrumentation (FTI) system is a fairly new approach. Even though Ethernet is widely accepted for FTI systems, the use of Ethernet for FTI remains fairly simple. Advanced packet filtering, QoS, traffic shaping and port trunking are only used to a small extent today. These features can, if configured properly, increase the network switch throughput by more than 100% compared to legacy network equipment and guarantee a worst-case latency of less than hundred microseconds for latency sensitive data.

This paper targets two FTI network challenges:

1.) How to guarantee worst-case latency for latency sensitive data?
2.) How to achieve near wire speed network performance on recorder ports without packet loss?

Abbreviations:

CoS Class of Service
DSCP Differentiated Services Code

	Point
FCS	Frame Check Sequence
IED	Intelligent Electronic Device
IP	Internet Protocol
IPG	Inter-Packet Gap
LACP	Link Aggregation Control Protocol
MAC	Medium Access Control
QoS	Quality of Service
RTOS	Real Time Operating System
UDP	User Datagram Protocol
TCP	Transmission Control Protocol
ToS	Type of Service

How to guarantee worst-case latency for latency sensitive data?

Ethernet switches may have support for priority containing two or more output queues per port, where the high priority queue(s) are reserved for latency sensitive critical data offering best possible QoS for such data. Relevant packet scheduler schemes for an Ethernet switch with four priority queues can be:

1. Round-robin weighting. I.e. N-highest (Priority Level 3) packets are sent from the highest priority queue, before N-high (Priority Level 2) packets are sent from the high priority queue, before N-low (Priority Level 1) packets are sent from the low priority queue, before N-lowest (Priority Level 0) packets are sent from the lowest priority queue. The packet scheduler will move directly to the next priority queue in the chain if no packets are present in the given queue.
2. Strict priority scheduling. I.e. all available packets in the highest priority queue will be transmitted from the highest priority queue before any of the lower priority queues are served. Thus, packets from a queue will only be sent if all higher priority queues are empty.

Note that a high priority packet also will be delayed due to a low priority packet if the transmission of the low packet is started before the high priority packet enters the egress port. The high priority packet will then be delayed by the time it takes to flush the rest of the low priority packet. Worst case will be that the transmission of a low priority packet with maximum packet length (1518 bytes) is just started when a high priority packet arrives the given egress port. The extra switch queuing delay will then be 122µs in case of 100Mbps egress port speed, and 12µs in case of 1Gbps port speed.

A high priority packet may also be delayed through the switch due to other latency sensitive packets that are already queued for transmission in the same high priority queue for a given egress port. It is, however, often a straightforward job to calculate the worst-case switch latency such a packet may experience if the network load and traffic pattern of the latency sensitive applications using the high priority queues are known, and all other traffic in the network have lower priority. Typical worst-case switch latency for a high priority packet in such a system will be a few hundred µs through each network hop in case 100Mbps is used on the egress port and less than 50us case 1Gps is used on the egress port.

Example 1:

- 100 Mbps with full duplex connectivity is used on all drop links.

- The switch is a store-and-forward switch, with a minimum switch latency of 10µs.

- The switch uses strict priority scheduling.

- The latency sensitive packet has a length of 200 bytes including preamble, MAC, IP, UDP, payload, FCS and minimum IPG. The latency sensitive packets are treated as high priority packets, all other packets have less priority.

- Up to five other end nodes may generate similar latency sensitive packets of 200 bytes that may be in the same priority queue before the packet enters the queue, and causes extra switch delay.

- All latency sensitive packets are generated in a cyclic manner.

The worst case switch latency of a latency sensitive packet will then be:

1.) 16µs, store-and-forwards.

2.) 10µs, minimum switch latency.

3.) 122µs, worst case latency due to flushing of a packet with maximum packet length.

4.) 80µs, five latency sensitive packets already in the same priority queue.

5.) 228µs, total.

Example 2:

Same as above, but with 1Gps rate on the egress port. The worst-case switch latency of a latency sensitive packet will then be:

1.) 16µs, store-and-forwards.

2.) 10µs, minimum switch latency.

3.) 12µs, worst-case latency due to flushing of a packet with maximum packet length.

4.) 8µs, five latency sensitive packets already in the same priority queue.

5.) 46µs, total.

Example 3:

Same as above, but with 1Gps rate and rate shaping set to 256Mbps on the egress port. The worst-case switch latency of a latency sensitive packet will then be:

1.) 16µs, store-and-forwards.

2.) 10µs, minimum switch latency.

3.) 48µs, worst-case latency due to flushing of a packet with maximum packet length.

4.) 31µs, five latency sensitive packets already in the same priority queue.

5.) 105µs, total.

These three examples represent worst-case latency for the latency sensitive packets identified as high priority packets. These estimations are valid regardless of any other network load with less

priority in the network. Several priority implementations exist with respect to how a packet is identified as a high priority packet. The priority handling depends on the switch functionality.

Layer 2 priority

A layer 2 switch performs switching based on the Ethernet MAC destination addresses, see Figure 1.

Figure 1, MAC header (layer 2), no VLAN tag

A layer 2 switch may provide priority information based on:

MAC addresses. Both the MAC source- and destination address can be used for priority identification, see Figure 1. This is not a very flexible feature.

Ethernet port. One or multiple ports of the switch can be configured for high priority. This means that all packets received on these ports will be treated as high priority packets. The technique requires a static setup and all packets received on a given port will be treated with the same priority.

Priority tagging. The IEEE 802.1p (and IEEE 802.1Q) standard specifies an extra field for the Ethernet MAC header. This field is called Tag Control Info (TCI) field, and is inserted between the source MAC address and the MAC Type/Length field of an Ethernet packet (see Figure 2). This field contains a three bit priority field that is used for priority handling. These three priority bits map to the priority queues of the switch. The mapping depends on the number of queues the switch supports. For example: priority field = 111 will map to priority queue 7 on a switch with 8 priority queues, while priority field = 111 and 110 will both map to priority queue 3 on a switch with four priority queues. Both unmanaged and managed switches can support this feature. Thus, no switch configuration is needed. A disadvantage with this method is that most end nodes do not support VLAN tagging. Configuring the switch to remove the tag after switching can solve this, and should be done before the packets are sent on the output ports, where stations without support for this feature are connected. This requires managed switch operation. Another problem could be that other existing Ethernet switches in the network do not support priority tagging. See [1] for more details on this topic. The maximum Ethernet packet size will, due to the VLAN tag, increase by four bytes to 1522.

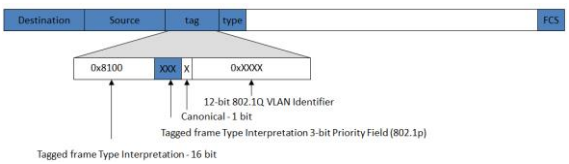

Figure 2, MAC header (layer 2) with VLAN tag

Layer 3 priority

A layer 3 switch can perform packet switching based on both the Ethernet MAC destination addresses and the layer 3. E.g. the header fields of IP packets.

A layer 3 switch may provide priority identification based on the same criteria's as a layer 2 switch. The following layer 3 field is also relevant:

IP ToS/Cos. Each IPv4 header contains a ToS/CoS field, see Figure 3. The RFC standards known as Differentiated Services see [2], partition the ToS/CoS field into two fields: DSCP (6 bit) and CU (2 bit). The DSCP field is used to determine the required priority. The 6 bit of the DSCP field represents 64 possible "code points" that is split in three pools:

-Pool 1 DSCP = [0 .. 31] reserved for standard actions (e.g. VOIP)

-Pool 2 DSCP = [32 .. 47] reserved for experimental or local use, but may be allocated for standard actions in the future.

-Pool 3 DSCP = [48 .. 63] reserved for experimental or local use.

Any subset of the 64 possible code points can be used as a high priority identification criterion in the switch. The high priority code points should preferably be user configurable. The code points from Pool 3 are the preferred alternative for a given nonstandard IP based real time application. F.ex. an FTI UDP stream.

High priority setting of the IP ToS field of real time critical packets must be set in the IP protocol of the sending station. This can be done on TCP/UDP socket level by a setsockopt( ) command both on the client and server socket side in most Operating Systems (OS).

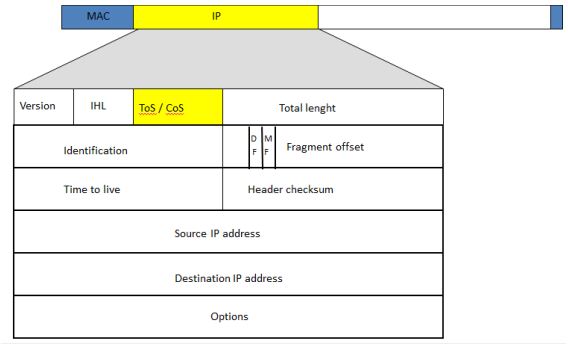

Figure 3, IPv4 header (layer 3)

An IPv6 header contains a corresponding field called Traffic Class. This field has the same function as the ToS/Cos field of IPv4. The Traffic Class octet has the same location in the IPv6 header as the ToS field has in the IPv4 header.

How to achieve near wire speed network performance on recorder ports without packet loss?

The following FTI application example (see Figure 4 and 5) will be used to demonstrate the different Ethernet packet filtering and network engineering techniques that can be used to maximize network performance:

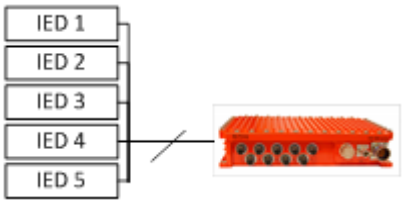

Figure 4, FTI data acquisition cluster

Figure 5**,** Central switch and recorder connecting six clusters

The FTI application example, shown in Figure 4 and 5, is composed of six data acquisition clusters, where each cluster is connected to a central switch via a GigE link. The central switch forwards the FTI data received from all six clusters to a recording unit over a GigE link, while each cluster contains Intelligent Electronic Devices (IED) that send data over 100Mbps links.

The worst-case load scenario will occur when all IEDs in all clusters send their data at the same time. This means that the total IED load that will hit the central switch will be several Gbps for a short time period even if the average IED data rate is a few tenths of Mbps. This load exceeds the data rate of the recorder link. This means that the central switch must buffer incoming packets from the IEDs. If the total incoming IED data rate exceeds 1Gbps for several ms, then the switch packet buffers will go full and the switch will start to drop packets. Low priority packets will be dropped first.

The higher the total average IED data rate is, the higher is the probability for packet loss on the central switch due to network congestion. Note, however, that packets can be lost on the central switch in the example above even with an average total IED data rate of only a few hundred Mbps.

The property of some IEDs is that data originating from each IED is cyclic. This is both a good and a bad property from a network engineering point of view. It is good that an IED data stream from an IED is spread out in time. The risk for long packet burst on the central switch in the example above, can then be reduced, while it is bad if the cyclic streams from the IEDs are synchronized. Thus, each cluster switch will then generate a burst to the central switch of at least five IED packets from the five IEDS connected to the cluster switch if the IEDs send their data more or less at the same time.

Ethernet drivers and TCP/IP stacks of most RTOS are far from deterministic. This means that so-called true cyclic IEDs also may send packets with a variable Inter-Packet Gap (IPG). One should also have in mind that if IEDs from different vendors (where some are cyclic and others are not) are combined in the same network including high-speed cameras (that for sure will send data in packet burst), then network engineering could only to a small extent utilize the assumption that some of the data sources are cyclic.

A robust and future proof FTI system should be able to handle packet bursts originating from switches as well as IEDs. The FTI example described above contains only two network hops with ports speed of only 100Mbps on all the IED drop links. One should note that the bursty behavior in a network load will increase:

- If the number of network hops increases
- If the same port speed (e.g. 1GBps) is used on all or the majority of drop links in the network
- If the packet sizes are longer

So, we will assume that the data sources from time to time can generate packet bursts. The idea is to shape the traffic egressing from the cluster switches.

Traffic shaping

Traffic shaping is also known as "egress rate shaping". The switch will delay some or all of the incoming packets before the packets are forwarded to an egress port. As the name suggests this network engineering technique will "shape" the traffic by using a rate limiter. The rate limiter will increase the minimum IPG on the packets egressing on the GigE port connected to the central switch, see Figure 6. Modern switches also have the support for shaping traffic for each priority queue individually.

Figure 6, Traffic shaping principle

Example:

- Rate shaping is set to 150Mbps on the switch port connected to the central switch for each of the cluster switches

The IEDs may send burst data with the following worst-case pattern:

- Packet length of 220 bytes
- 35 packet per burst with minimum IPG
- Inter-burst gap of 2ms.

This setting means an average IED data rate load of 29.9Mb/s and an average load to the central switch of 149Mb/s. (The worst-case data rate load egressing from a cluster switch is, however, 700Mbps unless rate shaping is used.) The total average load on the central switch and recorder will be 894Mbps.

The total number of packets sent in each burst is: 210,000. The packet bursts sent from the IEDs are not synchronized. The packet pattern before and after the central switch will be as shown in Figure 7.

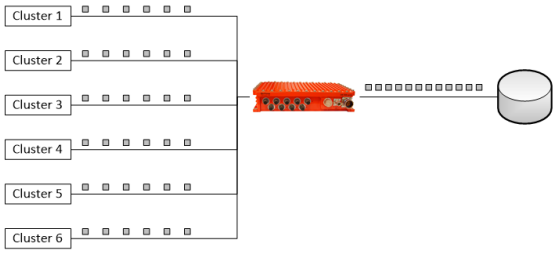

Figure 7, Traffic shaping before central switch

Traffic shaping - test result

A lab test setup based on the following components was established.

- 30 x IED simulators (Smartbit cards)
- 6 x CM1600 cluster switches
- 1 x CM1600 central switch
- 1 x recorder simulator

Figure 8 shows the lab setup:

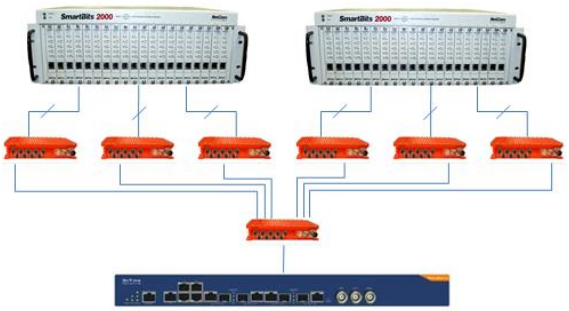

Figure 8, Lab setup

Two tests were performed:

1.) No rate shaping enabled on the cluster switches
2.) Rate shaping enabled on each of the cluster switches. Rate shaping level set to 150Mbps

The first test, 1.), showed that 9,043 packets out of the 210,000 packets were lost, while no packets were lost for the second test. The packet loss statistics would be even higher in case of no rate shaping and if the IEDs burst sending were synchronized.

Link Aggregation Control Protocol (LACP)

The next step on the migration path to 10GBE is to combine two or more one GigE ports in a logical trunk by using the Link Aggregation Control Protocol.

LACP combines (aggregate) multiple network connections in parallel to increase throughput beyond what a single connection could sustain, and to provide redundancy in case one of the links fail.

LACP is defined in IEEE 802.1ax or the previous IEEE 802.3ad standard.

This can be relevant for a recorder having support for more than one GigE ports.

Conclusion

This paper has demonstrated that worse-case latency through an Ethernet switch with ingress data rate of 100Mbps and egress data rate of 1Gps can be far less than 100µs for latency sensitive data if this data is identified as high priority data. This means that we can guarantee worse-case latency through a FTI network based on using standard QoS techniques.

Near wire speed data rate on the recorder drop links can be achieved if rate shaping is used on Ethernet switches.

Port trunking according to LACP can also be considered on the switches in order to be able to go beyond 1Gps on the recorder links.

References

[1] Holmeide, Gauvin: "ETC2014-Ethernet packet filtering for FTI - part I"
[2] Diffserv, RFC 2474, http://www.ietf.org/rfc/rfc2474.txt