VoIP drives realtime Ethernet

Øyvind Holmeide and Tor Skeie

***Abstract*--The real time properties of traditional Ethernet are poor. The end-to-end latency through an Ethernet network based on thin Ethernet (coax - 10BASE-2) or hubs (10BASE or 100BASE) depends on the network load. The non deterministic property of the Carrier Sense Multiple Access – Collision Detection (CSMA/CD) scheme has been the main argument against Ethernet as the communication solution for applications with real time requirements. However, this has changed significantly with the introduction of Ethernet switches together with the new priority features of Ethernet switches. Real time applications based on switched Ethernet can take advantage of this technology driven by the fast growing Voice Over IP (VOIP) business.**

***Index Terms*--Deterministic Ethernet, IP ToS, Priority tagging, Voice over IP.**

I. TRADIONAL ETHERNET

Traditional Ethernet is not real time friendly. The CSMA/CD scheme of Ethernet makes access to the medium non-deterministic. An Ethernet controller connected to a thin Ethernet (coax - 10BASE-2) or a hub (10BASE or 100BASE) is not able to send a packet as long as the medium is busy sending another packet. The Ethernet controller is free to send its packet as soon as the Ethernet is idle. While transmitting, the station continues to listen on the wire to ensure successful communications. If two stations attempt to transmit information at the same time, the transmissions overlap causing a collision. If a collision occurs, the transmitting station recognises the interference on the network and transmits a bit sequence called jam. The jam helps to ensure that the other transmitting station recognises that a collision has occurred. After a random delay, the stations attempt to retransmit the information and the process is repeated. The probability for a collision depends on the collision domain, i.e. the range of the Ethernet, and the network load. A traditional CSMA/CD Ethernet with 20% utilisation has less than 0.1% collision, while as much as 5% of the packets will experience collisions if the network utilisation is above 40%. A CSMA/CD network with 40% utilisation is in trouble, and the net data rate will in fact decrease due to collisions if the load is further increased. However, bare in mind, those collisions are not errors. Collision is a normal part of Ethernet networks. The figures below show the principles of CSMA/CD.

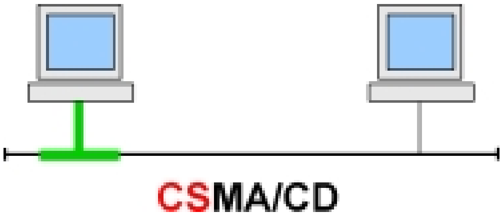

Figure 1, carrier sense

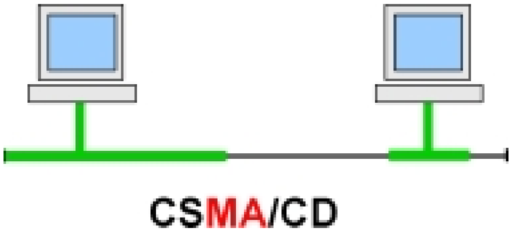

Figure 2, multiple access

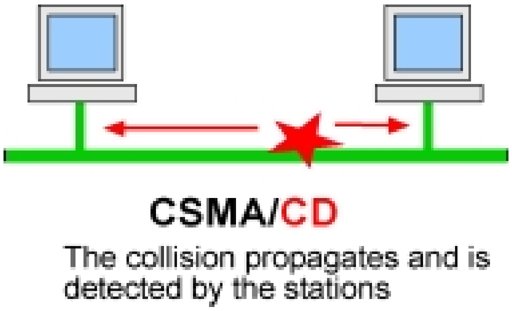

Figure 3, collision detection

II. SWITCHED ETHERNET

From a functional point of view, switching is exactly the same as bridging. However switches use specially designed hardware called Application Specific Integrated Circuits (ASICs) to perform the bridging and packet-forwarding functionality (opposed to implementations using a central CPU and special software). As a consequence switches are much faster than bridges.

Ethernet switches provide 10, 100, 1 Gbps or even 10 Gbps (under development) on each drop link. This represents a scalable and huge bandwidth increase compared to e.g. an

Øyvind Holmeide is president in OnTime Networks, a spin-off company from ABB that develops deterministic industrial Ethernet switches. Øyvind is still associated with ABB Corporate Research, and he can be contacted at oeyvind@ontimenet.com.

Tor Skeie is a senior scientist at ABB Corporate Research. Tor is responsible for an ABB research project, where a new communication solution based on deterministic Ethernet is developed together with OnTime Networks for HV/MV substation automation applications. He can be contacted at tor.skeie@no.abb.com.

Ethernet hub where the bandwidth is either 10 or 100 Mbps and shared between all users connected to the same network segment.

Ethernet switches also offer both half and full duplex connectivity. This means that an Ethernet controller never will see any collision if full duplex connectivity is used.

However, packets can still be lost if one of the following scenarios appears:

1. The total network load exceeds the switching capability of the switch engine. I.e. the switch is not able to handle full wire speed on each drop link.
2. The output buffer capacity is not sufficient. I.e. the amount of packets sent to an output port exceeds the bandwidth of this port for a time period that is longer than the output buffer is able to handle. Thus, packets from several input ports compete for the same output port causing a non-deterministic buffering delay.

Higher protocol layers at the stations must handle lost packets.

These two scenarios can be avoided by using the following Ethernet techniques:

- Back pressure; the switch can send a jam pattern simulating traffic on a port operating in half duplex mode if the amount of packets received on this port is more than the switch can handle.
- Flow control; the switch can send PAUSE packets according to IEEE802.3x on a port operating in full duplex mode if the amount of packets received on this port is more than the switch can handle.
- Priority; Ethernet packets that are identified as high priority packets are put in a high priority queue. Packets from a high priority queue are sent before the low priority packets. The low priority packets may still be lost. This is the most relevant technique with respect to optimal real time properties for latency sensitive real time data.

III. PRIORITY

Ethernet switches today may have support for priority containing two or more output queues per port, where the high priority queue(s) are reserved for real time critical data offering Quality of Service (QoS). How the switch alternates between the priority queues vary from vendor to vendor. Relevant alternating schemes for a switch with two priority queues could be:

1. Round-robin weighting.. I.e. N packets are sent from the high priority queue before one packet is sent from the low priority queue.
2. Strict priority. I.e. all packets will be transmitted from the high priority queue. Packets from the low priority queue will only be sent in case the high priority queue is empty.

Note that a high priority packet will be delayed due to a low priority packet if the transmission of this packet is started before the high priority packet enters the output port. The high priority packet will then be delayed by the time it takes to flush the rest of the packet. Worst case will be that the transmission of an Ethernet packet with maximum packet length (1518 bytes) is just started. The extra delay will then be 122 µs in case of 100 Mbps, and 1.22 ms in case of 10 Mbps.

A high priority packet may also be delayed through the switch due to other real time packets that are already queued for transmission in the same high priority queue. However, it is 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 real time application using the high priority queues are known, and all other traffic use lower priority. Typical worst-case switch latency for a high priority packet in such a system will be a few hundred µs in case of 100 Mbps on each drop link.

Example:

- 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 10 µs.
- The switch uses strict priority scheduling.
- The real time packet has a length of 200 bytes including preamble, MAC, IP, UDP, payload, Frame Check Sequence (FCS) and minimum Inter Packet Gap (IPG).
- The real time packets are treated as high priority packets, all other packets have less priority.
- Up to five other stations may generate similar real time packets of 200 bytes that may be in the same priority queue before the packet enter queue, and cause extra switch delay.
- All real time packets are generated in a cyclic manner.

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

- 16 µs, store-and-forwards.
- 10 µs, minimum switch latency.
- 122 µs, worst case latency due to flushing of a packet with maximum packet length.
- 80 µs, five real time packets already in the same priority queue.
- 228 µs, total.

This worst case latency for the real time packets is valid regardless of any other network load with less priority.

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:

1) Layer 2 switch

A layer 2 switch performs switching based on the Ethernet MAC destination addresses, see Figure 4. A layer 2 switch may provide priority identification based on:

- MAC addresses. Both the MAC source- and destination address can be used for priority identification, see Figure 4. The switch must be a managed switch in order for the user to set high priority MAC addresses. This is not a very flexible feature.
- Ethernet port. One or more of the 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. Switches that provide this function are in most cases managed. The advantage of this feature is limited, but this feature has been the only priority function available upto now.
- 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 5. This field contains a 3 bit priority field that is used for priority handling. Thus, the standard defines 8 different levels of priority. However, most Ethernet switches available on the marked that support priority queuing have only two or four queues. A switch with two priority queues will put Ethernet packets with priority tags set to four or higher in the high priority queue while all other packets will be put in the low priority queue. Both unmanaged and managed switches can support this feature. Thus, no switch configuration is needed. A disadvantage with this method is that most stations upto now do not support priority tagging. Configuring the switch to remove the tags after switching can solve this, and 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 there exist other Ethernet switches in the network that do not support priority tagging. I.e. the maximum packet size will due to the tag increase by four bytes to 1522 bytes, and some switches will not forward packets with a packet length larger than 1518 bytes.

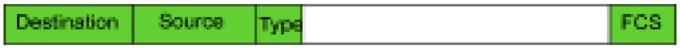

Figure 4, MAC header (layer 2)

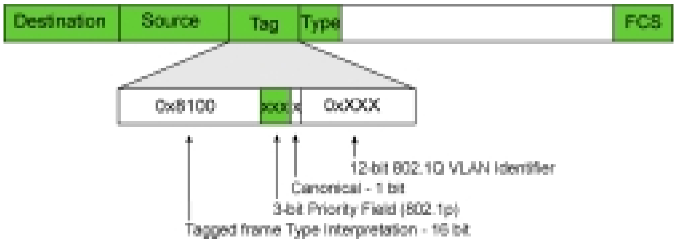
Destination
Source
Tag
Type
FCS
0x8100
xxx
x
0xXXX
12-bit 802.1Q VLAN Identifier
Canonical - 1 bit
3-bit Priority Field (802.1p)
Tagged frame Type Interpretation - 16 bit

Figure 5, MAC header (layer 2) with tag

2) Layer 3 switch

A layer 3 switch can perform switching based on both the Ethernet MAC destination addresses and the layer 3 contents (i.e. router functionality). 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 Type of Service (ToS). Each IPv4 header contains a ToS field, see Figure 6. Recent standards known as Differentiated Services (Diffserv, see RFC 2474), partition the ToS 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. A switch that has support for IP ToS priority can either be unmanaged or managed. The high priority code points will in most cases be user configurable in case the switch is managed, while the corresponding high priority code points for an unmanaged switch will be pre-configured. No switch configuration is needed in case of pre-configuration. The code points from Pool 3 are the preferred alternative for a given non standard IP based real time application.

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).

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

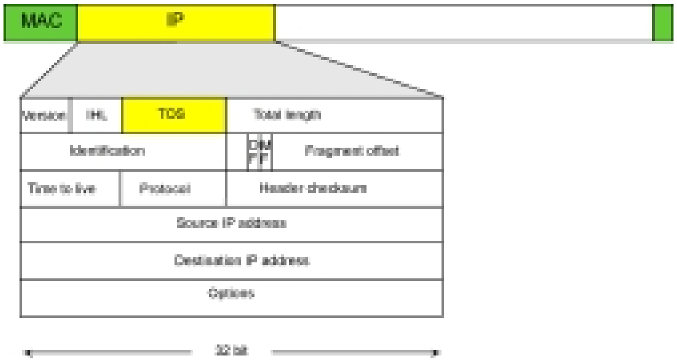

Figure 6, IP header (layer 3)

3) Layer 4 switch

A layer 4 switch can perform switching based on both the Ethernet MAC destination addresses and the layer 3 and layer 4 contents. E.g. the IP and TCP headers of TCP packets. A layer 4 switch may provide priority identification based on the same criteria's as a layer 3 switch. The following layer 4 fields are also relevant:

UDP or TCP destination port numbers. The destination port of an UDP or a TCP header can also be used in the switch as a high priority criterion. Figure 7 shows this field for the UDP header. Most layer 4 switches are and will be managed. Thus, this priority function requires switch configuration by the user. Each switch between the client and server in a real time application must be configured for the chosen socket port number of the server. However, the user should bear in mind that the socket port number of the client would be different from the corresponding server socket port number. I.e. the destination port and source port number in an UDP packet will be different. The switches should be configured for both port numbers if real time critical data also is sent from the server to the client. This could be a problem because the client socket port number in most cases is dynamically allocated.

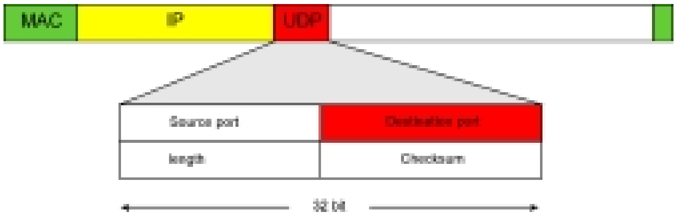

Figure 7, UDP header (layer 4)

IV. Conclusions

Deterministic Ethernet is achieved by using priority. Worst case latency for real time critical data through a switched Ethernet infrastructure can be guaranteed if this data is protected by priority. Typical worst case switch latency for priority protected packets are in the order of a few hundred μs.

Switch priority handling based on IP ToS is the preferred choice. A switch that is pre-configured for a high priority subset of possible ToS code points does not require any user configuration of the switches. Such switches will in most cases be unmanaged, and the most cost efficient alternative. Configuration of priority based on IP ToS is then a matter of the real time application. Configuration is easily performed on socket layer for both the client and server. Note also that priority based on IP ToS does not represent any conflict with respect to other switches or stations that do not support this feature. This is not the case for priority based on tagging (IEEE 802.1p) as long as some switches and stations do not support tagging. However, priority tagging may be a relevant alternative in the future, when most switches, Network Interface Cards (NIC's) and OS'es support this feature.