Utah Scientific

A Utah Scientific White Paper

IT ESSENTIALS:
Unbundling Ethernet for Studio Video Over IP

By Scott Barella, CTO of Utah Scientific

Introduction

I remember when I was a junior studio broadcast engineer learning the intricacies of the analog waveform monitor and vectorscope back in the 1980s. For video engineers, these were the tools of the trade, and they were essential for a broadcast station to stay in compliance of very stringent analog signals. Fast-forward a few decades, and the signals that were once based on pulses have been replaced by digital SDI signals — and soon, those SDI signals will be replaced by Ethernet packets. With the new SMPTE ST 2110 standard for uncompressed IP video and audio about to come online, it's incumbent on engineers to understand all they can about the standard called Ethernet.

While Ethernet has existed in television studios for years, it has mainly been confined to configuration and control networks. More recently, Ethernet has also provided a vehicle to carry compressed transport streams, introducing a new way of thinking and a departure from the older pulse-based streams for carrying compressed video using Asynchronous Serial Interface (ASI).

This primer is intended to give a good overview of Studio Video over IP (SVIP) in the uncompressed domain using Ethernet, with the goal of making these concepts a bit easier for engineers to understand and, more importantly, to command.

Ethernet Basics

While it might be interesting to explore the origins of Ethernet with Bob Metcalf and David Bogg's Xerox project in 1973, we will focus instead on the basic structure beginning with subnetworks (subnets). These are defined using a four-octet address scheme, such as the private address of 192.168.1.10, followed by a network mask, or netmask, of 255.255.255.0 to further divide the network into smaller sizes in the address range. For this example, the addresses from 192.168.1.1 to 192.168.1.254 can be used to define a subnet called 192.168.1.0, also known as a class C private network. The private designation means that this address is not publically available in the same manner as an internet address, for example. Here are the common IP version 4 (IPv4) private network addresses set aside by the Internet Assigned Numbers Authority (IANA):

1. Class C Private Networks - 192.168.0.0 – 192.168.255.255, making 65,536 addresses possible
2. Class B Private Networks - 172.16.0.0 – 172.31.255.255, making 1,048,576 addresses possible
3. Class A Private Networks - 10.0.0.0 – 10.255.255.255, making 16,777,216 addresses possible

All of these addresses have been set aside for use in private networks beyond the purview of public networks such as the internet. The beauty of this idea is that it reserves a great deal of private addresses for use in broadcast television facilities, outside of the realm of the public networks.

The basics of any subnet are a common set of addresses like the before-mentioned 192.168.1.10. This address can "speak" with another address as long as the second address uses the first three octets; e.g. 192.168.1.X. Think of this like a city street where there are a number of house addresses along the street. All of the houses on a certain street, or in the subnet, can share information among each other. However, houses on a different street or in a different subnet cannot share information.

The "ping" command on the Computer Line Interface (CLI) operates by sending Internet Control Message Protocol (ICMP) Echo Request packets to the target host and waiting for an ICMP Echo Reply. If the computer with address 192.168.1.10 attempts to "ping" an address of 192.168.1.30, it will receive a response from that Ethernet interface. However, if it attempts to "ping" 192.168.2.30, it will not get a response because it is in a different subnet, but it might receive a response if another object on the Ethernet network acts as the gateway.

The gateway acts as the door to another subnet. In the previous example, the address of 192.168.1.1 might be the gateway address to route the data from 192.168.1.10 to 192.168.2.30. To use the street analogy, the gateway might have access to another street (like an alley) that can provide a route from one street to another. Gateway addresses are used in routers for this very purpose.

Most networks are configured as Local Area Networks (LANs), but they can also have larger scope in the form of Metro Area Networks (MANs) and Wide Area Networks (WANs). Some companies connect themselves using a WAN to carry data from different subnets within their private network. A station in Denver might have a need to deliver its Ethernet data to a station in New York, for example. The Denver LAN subnet might be 192.168.1.0 but the one in New York might be 192.168.20.0. The WAN would make it possible for these two subnets to communicate.

Video Over IP

In most office environments, it's possible to create a unicast in which video from a subnet address is sent to another address. This is very similar to the idea of an SDI router in which one source wants to send its video to a destination. In the SDI environment, the architecture ensures that the bandwidth for such a connection is dedicated. In the Ethernet unicast environment, the address acts as the sender and another address within the subnet acts as the listener. In our earlier example, the sender at 192.168.1.10 might send a video stream to a receiver at 192.168.1.50. In a unicast scenario, no other receiver in the subnet would be able to listen to that video stream. The requirement for a sender to connect with multiple listeners, much like a distribution amplifier in SDI, calls for multicast.

In much the same manner as reserving addresses for private networks, the IANA has also set aside addresses to be used for multicast. Typically, multicast addresses ranging from 239.0.0.0 to 239.255.255.255 are used, along with port numbers, to carry video data to multiple receivers. An address of 239.1.1.1:10000 could designate a certain address so that others could listen to it on the same subnet (in this example, the port number is 10,000). It's also important to note that the rules of the subnet are still intact — meaning that receivers and senders will need to be in the same IP address of the subnet in order to broadcast to those within that subnet. For example, a source IP address of 192.168.1.10 might emit a transmission on a multicast address of 239.1.1.1:10000, and the receiver would listen to that multicast transmission from 192.168.1.50 provided the receiver was tuned to the multicast address of 239.1.1.1:10000. If the receiver were not in that subnet, it would not be able to "hear" the multicast transmission because it's not in the same subnet. It could be routed using a gateway, but that would require specialized architecture and routing.

The method that joins the sender with the listener is called the Internet Group Management Protocol (IGMP), enabled in an Ethernet switch called IGMP snooping. IGMP allows the listener to request to be joined to the stream that the sender is sending. In the above example, the listener would send an IGMP request to listen to the multicast stream of 239.1.1.1:10000, and the switch would then connect the sender to the receiver. This method has two general versions: IGMPv2 (more common) and IGMPv3 (seeing growing use because it allows more multicast specificity than IGMPv2). In general, both versions can be accompanied by today's Ethernet switches.

Virtual Local Area Networks (VLANs) will also be used to keep traffic isolated and confined. Think of VLANs as small switches that can be used independently. For example, if there is a good reason to keep all of the video in one VLAN, this might be a way to group the video from the audio and data. It might also be important to keep certain video confined to studios; in this manner, an entire studio might be kept in a separate VLAN even though the switch might also serve another studio or production area. VLANs are generally managed within the switch and are generally required to be identified for each port on the switch. All the ports can be put in the same VLAN or they can be divided among a few VLANs.

Bandwidth Considerations

Ethernet, like analog video, started with rather small bandwidths carried on coaxial cable and twisted-pair wire technology that maxed out at 10 MB/s. As the demand for Ethernet gathered momentum in the IT industry, the bandwidth capacity grew from 10 to 100 MB/s, and then to 1 Gb/s, primarily using twisted-pair cabling.

Most video engineers are keenly aware of bandwidth requirements. In SDI we started with Standard Definition (SD), the digital equivalent of analog signals whose payload was approximately 270 MB/s. With the advent of High Definition (HD), the SDI payload increased to approximately 1.5 Gb/s. With this big improvement, coaxial cable could keep up with the payloads, albeit at somewhat shorter distances than SD. While there are new cables that address 4K signals at around 12 Gb/s, the limit is now in sight for SDI on coax.

When Ethernet took the next jump from 1 Gb/s to 10 Gb/s, fiber became more practical than copper twisted pairs. And while video has utilized fiber in the past, it has done so without Ethernet but instead as a raw electrical SDI-to-optical (fiber) conversion. One Ethernet-over-fiber tenet that confuses some to this day is that the actual signal is Ethernet, and it typically utilizes multimode fiber for runs under 350 feet (OM-3 or OM-4) and single-mode fiber for runs up to 10,000 feet (OS-1 or OS-2).

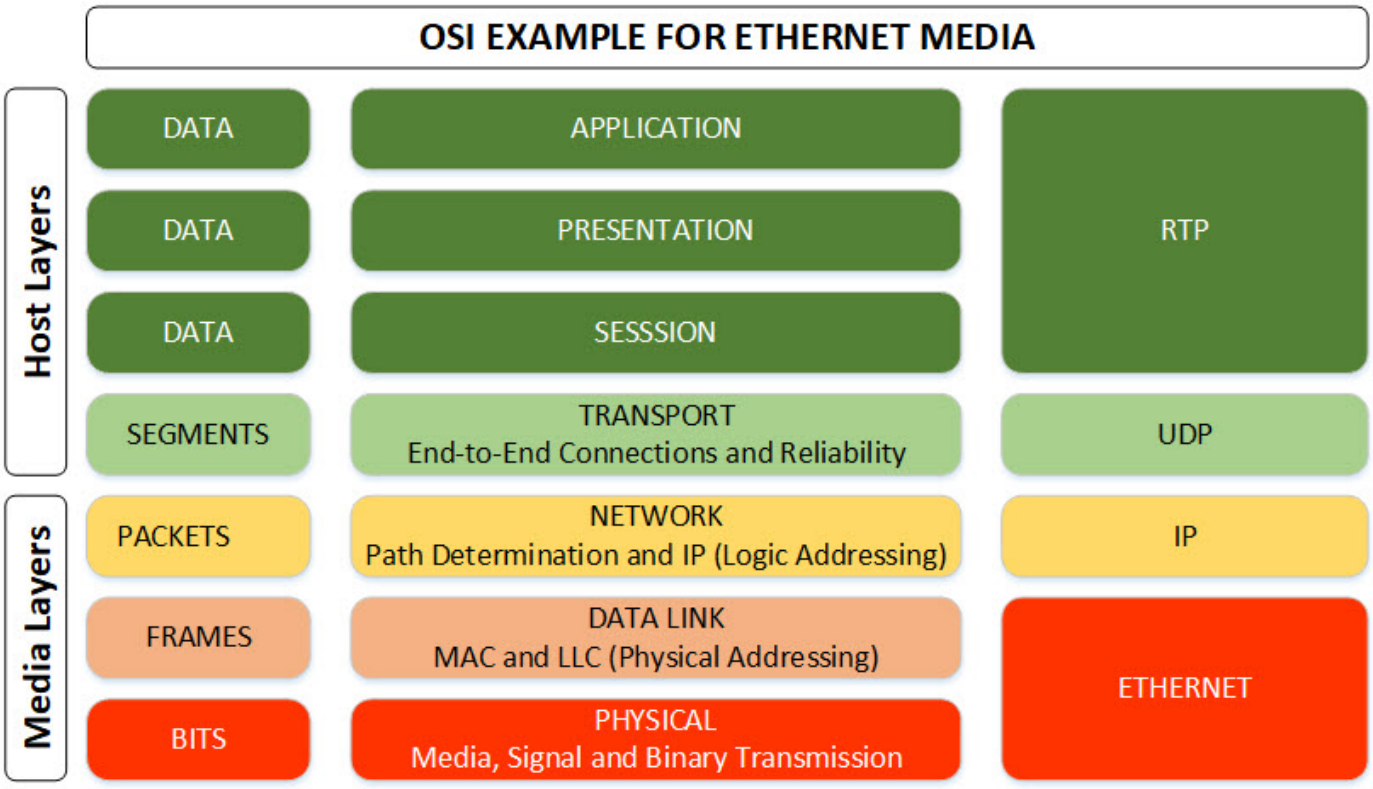
While Ethernet has been used on 1 Gb/s switches, it has carried compressed video in formats such as MPEG-2, H.264/265, and JPEG 2000. These compression technologies can keep the payload well under 1 Gb/s. In the new and emerging world of Studio Video over IP (SVIP), the payloads are the same as their coaxial counterparts — 1.5 Gb/s for HD-SDI and 3.0 Gb/s for 3G-SDI — and are therefore driving the need for large Ethernet ports.

If you think of payloads in regard to the size of the pipe for carrying the signal, it's obvious that if HD signals are to be used, then the pipe must be able to accommodate payloads larger than 1.5 Gb/s. In the case of an Ethernet switch, the port size needs to be at least 10 Gb/s. At this capacity the port can accommodate more than one HD signal; in fact, it can handle up to six HD signals (1.5 Gb/s x 6 = 9 Gb/s) or three 3G signals (3.0 Gb/s x 3 = 9 Gb/s).

On enterprise-grade 10 Gb/s Ethernet switches, the backbone is rated in terabytes so that the switch has enough bandwidth to carry plenty of SD, HD, and 3G signals. As 4K continues to emerge, many 4K broadcasters will be using the new 25 Gb/s Ethernet ports to accommodate payloads of 12 Gb/s, especially if the UHD signal uses larger resolutions and/or frame rates.

The Foundation: OSI

The Open Systems Interconnection (OSI) model is the foundation for all Ethernet technologies; therefore, it's important to explain how OSI is used in the application of Studio Video over IP. The first layer is the Physical (PHY) layer and is comprised of the raw emission of electrical pulses converted to light, the essence of fiber. The second layer is Ethernet with its formal addressing scheme, followed by the IP layer that defines all the rules that apply to the Internet Protocol.



The packeting structure completes the next layer up in the form of the Universal Datagram Protocol (UDP), with packets that fit within UDP defined by the Real Time Protocol (RTP) used to both sequence and time-stamp the audio, video, or data streams. This the last protocol that should receive special attention, since it's how we insure that the packets that comprise a given audio, video, or data stream are held in the correct order and have some time context with regard to each of the sequential packets. RTP packets are the essence of these new IP video standards, and tools such as "Wireshark," which helps to dissect a particular stream to the OSI stack, make troubleshooting Ethernet a lot less daunting given its extraordinary payloads when carrying video.

Compressed vs. Uncompressed

Using IP to carry video isn't entirely new; in fact, it's been going on for well over a decade using compression. MPEG-2, H.264, HEVC (H.265), and JPEG 2000 compression schemes have used the MPEG-2 transport stream method to carry video, audio, and data over networks primarily because it's the most practical means to move video over private and public WANs.

For playout of recorded programs, there have been attempts to use compressed video in file-based formats since there is less concern for latency. But in live production, real-time signals have to do the work that video production has been doing for nearly 70 years going back as far as the earliest analog days. The ability to switch the signal as close to real time as possible is critical for live production events such as sports and news shows, but the low latency thresholds required by live broadcasting demand the use of uncompressed video. When the industry moves from SDI to IP, the use of uncompressed video over IP networks will be critical for achieving the close-to-real-time latency that we're all used to with SDI.

The first formats for uncompressed IP emerged from the SMPTE ST 2022 family of standards describing the use of compressed video with transport streams. This standard defines two compressed formats, 2022-1 and 2022-2. When the desire to use uncompressed video came about, we discovered that these formats could also be used to encapsulate the uncompressed SDI stream in nearly the same fashion as the compressed ASI stream. This works by taking the entire SDI signal and placing it into standard Ethernet-sized packets and sending them out over an Ethernet network. At the receive end, the packets are collected and then de-encapsulated to bring them back to their native SDI form. This type of signal is referred to as 2022-6, and the 2022-7 derivation simply provides two streams simultaneously for "hitless" performance in the vein of a primary and backup stream.

Meanwhile, the Audio Video Bridging (AVB) platform from the IEEE was also being developed to address demand for an entirely new signal using Ethernet hardware, but utilizing a pulse-based sync to arrange the packets. The idea proved workable but required a special Ethernet switch to accommodate the new signal — driving market demand for a signal that could utilize common-off-the-shelf (COTS) Ethernet switches.

Both 2022-6/7 and AVB relied on carrying the full complement of audio, video, data and sync in an "embedded" signal, meaning that individual components like audio would have to be de-embedded and then subsequently re-embedded as required.

The Emergence of SMPTE ST 2110

Against this backdrop, the Joint Task Force for Networked Media (JT-NM) was a consortium of SMPTE, EBU, and the Video Services Forum (VSF) that put together a roadmap for SVIP, with the express purpose of defining an industry standard for uncompressed IP video and audio. While other companies were working on a proprietary means to help their customers, the work of JT-NM had an industry-wide focus.

The VSF took up the call to describe a signal that would be completely based on Ethernet standards, releasing Technical Recommendations 3 and 4 (TR-03 and TR-04). These recommendations call for a signal with three separate multicast addresses, one each for audio, video, and data, to make it much easier to manipulate each of the three elements. Furthermore, TR-03 and TR-04 use Precision Time Protocol (PTP) to synchronize the three signals together (more on this in the next section), a key capability for solving the lip synchronization issues that have plagued our industry for years.

Shortly thereafter, a consortium of users and vendors calling itself the Alliance for IP Media Solutions (AIMS) came together to implement TR-03/04. AIMS would later adopt the work of the SMPTE 2110 Draft Group, whose members began the task of detailing the specifications of a signal, now referred to as SMPTE ST 2110. This IP-based signal will be used in uncompressed studio environments in which HD-SDI and 3G-SDI have been the mainstays for live and playout conditions.

2110 Components

There are several sections to the 2110 standard, but they all interrelate to comprise a complete package of signals that have been used for the past several decades.

2110-10 System Timing and Reference. This section addresses the use of SMPTE 2059 1/2 for time-stamping each of the three streams and also providing a time reference for genlocked signals. We mention 2110-10 first because it's perhaps the most important, and we'll delve into it further down in this paper.

2110-20 Video. Video is the largest of all the three payloads, and this section offers a fair amount of detail on how the video is generated in packet form.

2110-21 Timing Model. While this section is currently in the draft committee, the idea is to create a model for both Hardware (N for narrow) and Software (W for Wide). The intent is to provide a firm basis of packet evenness as the packets are emitted from the sender, through the switch and to the receivers. Narrow models that use FPGAs are naturally tighter in tolerance, and those using CPU cores are somewhat looser, but many believe that both hardware and software platforms will eventually be relatively similar in performance as software development improves.

2110-30 Audio. The audio described by this section so closely models AES 67 as to be nearly identical. Just as with AES 67, audio is also time-stamped, making it ideal as a near-perfect "drop in" to the SMPTE 2110 family.

2110-31 Audio. While many users are working with AES3, not so much for its current iteration but rather its compressed formats such as Dolby AC-3 (aka Dolby Digital) and Dolby E, this section in 2110 is meant to preserve the use of "other" audio types that will most likely be used in the future. This describes how to use these types of audio in the 2110 environment.

211-40 ANC Data. Closed captions, AFD, and timecode are just a few VANC data examples, and here the intent is to separate this data so that it can be manipulated as a discreet data stream, of course using time stamps so that it can be correlated to the audio and video elements.

A Closer Look at PTP

As we've noted, SMPTE 2110-10 is probably the most compelling section of the new standard, as the other sections simply address elementary audio, video, and data streams. 2110-10 describes the system timing and how PTP packets are to be used, as well as how each of the streams will be carried in the network.

The key to SMPTE 2110 is the ability to timestamp each of the three elements. When each as a timestamp, they can be used to compare other signals against the time marks that are in each separate stream. Not only can they be used for timing the composite streams for switching or other timed events, but they can also be used within the streams such as the relationship of video to audio and video to data in the case of closed captioning. It's a very elegant solution for solving the problem of lip-sync.

If you are familiar with Network Time Protocol (NTP), you have most of what you need to understand PTP. This signal is a separate stream of packets that contains nothing more than precise timing information. That timing is used to stamp the audio, video and data packets for each of the three elementary streams. This capability alone sets the PTP signal apart from the industry's pulse-based audio and video signal heritage. Never before has it been possible to stamp the time on video, as enabled by the PTP signal.

Because it's possible to stamp time using the actual moment in time that the event occurs, PTP can be used for genlock as well as interformat time relationship. It's shaping up to address lip-synchronization issues once and for all.

Switch Types

This is a topic that is still being actively discussed, since there are several approaches to switching video and audio in the Ethernet environment. Because Ethernet uses a packet structure, it's not possible to switch on frames or lines; rather, the switching must be done on boundaries between packets. While this may seem problematic, remember that each packet can stream quite quickly in Ethernet. If it's possible to time financial trading within nanoseconds, why shouldn't it be possible to switch video with the same accuracy?

Today's Ethernet switches fall into two main categories: Break Before Make (BBM) and Make Before Break (MBB). With a BBM switch, for a brief moment in time when a receiver is tuned to a particular signal and needs to be switched to another, the two signals need to be present to the receiver. In this fashion, a receiver tuned to one signal has to break the old signal and tune to a second signal when it appears. The downside of this idea is that it requires double the bandwidth — even if it's for a split second.

The MBB method simply breaks the old connection and then tunes to the new signal. The downside of this idea is that there is a small glitch in the decoded video during the transition from the former video stream to the new video stream, but no additional bandwidth is needed.

Another method is to leave the endpoint receiver's multicast address the same while the switch changes the signal upstream; in other words, it changes the source multicast address to the multicast address of the receiver. This one may be a good resolution to the requirement of preserving port bandwidth, and it also ensures a clean switch from one video source to another.

Conclusion

While there are many facets to the use of Ethernet, only a few key pieces — IP, UDP, and the RTP timestamped packets —relate to elementary audio, video, and data streams. Understanding this construction takes a good share of the sting out of learning about the new Video and Audio Over IP topologies.

The larger and more complicated piece is the network architecture that must consider aggregated payloads, SDN, signal management, VLANs, and host of best practices currently being used in the Ethernet environment. Given its importance in the broadcast infrastructure, this network architecture will be the biggest consideration as media organizations proceed in their inevitable migration to all IP-based operations.