Measuring the Absolute Accuracy of 10GbE Packet Timestamping

Arista designed and implemented a rigorous test methodology, described in this document, to measure the absolute accuracy of a 10 Gigabit Ethernet (10GbE) packet capture and timestamping solution. This methodology involved measuring the solution's packet timestamps relative to a PPS time synchronisation pulse disciplining the solution, to a precision of 50 picoseconds.

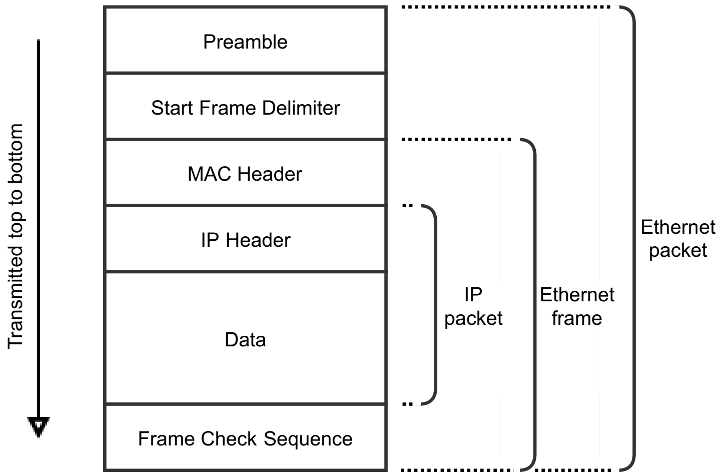
An independent specialist firm, the Securities Technology and Analysis Center (STAC) has adopted this methodology as part of their STAC-TS (Time Synchronisation) benchmarking suite. Arista measured the absolute timestamp accuracy of their nanosecond-resolution MetaWatch product using the above mentioned benchmark methodology and demonstrated that 92.3% of timestamps were accurate to ±1 nanosecond with 100% of timestamps within -3 and +2 nanoseconds of absolute accuracy.

Table of Contents

Nomenclature

Packet Capture

The term packet capture is used widely to refer to capturing network traffic. It is however, worth understanding precisely what the term packet refers to. The IEEE sets the standards for Ethernet and define an Ethernet packet at the physical layer as a unit of data wrapping a MAC frame at the data link layer (IEEE Std 802.3-2015 Chapter 3.1.1). An Ethernet packet is therefore an Ethernet frame, prefixed by preamble and start frame delimiter (SFD) bytes and optionally suffixed by an extension field (only used for half-duplex GbE networks). Moving up to the network layer, the IETF set the standards for the Internet Protocol (IP), and defines an IP packet as a unit of data including an IP header and data which may be wrapped by a data link layer header and trailer (IETF RFC 1122 Section 1.3.3). Therefore, for IP traffic over Ethernet, an Ethernet packet wraps an Ethernet frame which, in turn, wraps an IP packet. This can be illustrated in the following diagram:



Which of these is actually being captured? This really is up to the implementer however 10GbE capture processes Ethernet packets but, by general convention, actually captures and timestamps the Ethernet frame within the Ethernet packet by stripping off the preamble and start frame delimiter. Given that the purpose of these fields is to allow receiving devices to synchronise their receiver clocks at the bit and byte level and detect an incoming Ethernet frame at the physical layer, there is no practical loss to doing so. On most networks, the majority of Ethernet traffic is IP-based with Ethernet packets/frames encapsulating IP packets. The remainder consists of Ethernet packets/frames which encapsulate payloads not IP-based. Examples include Ethernet flow control, the Link Layer Discovery Protocol (LLDP), Spanning Tree Protocol (STP), Precision Time Protocol (PTP) over Ethernet, Fibre Channel over Ethernet (FCoE) and RDMA over Converged Ethernet (RoCE). Timestamps are therefore generally referenced to the start of the Ethernet frame. Arista defines this point in time as when the middle of the first bit of the incoming Ethernet frame, immediately following the start frame delimiter, as it leaves the receiving 10GbE SFP+ transceiver.

In light of the above, for the remainder of this document, we will use the label "packet" to refer to a unit data transported over 10GbE, in all cases except where Ethernet packet or frame-level processing is described where we will use the label Ethernet packet or frame. Conceptually, the terms are interchangeable however as we are digging into the actual mechanics of packet capture and timestamping, it is important that the relevant technical content be accurate and unambiguous.

1. Introduction

An increasing number of network devices such as switches, capture cards and appliances as well as network adapters offer the ability to timestamp incoming and event outgoing packets. This ability to timestamp Ethernet traffic precisely is used widely in applications such as compliance, troubleshooting, capacity planning, out-of-band networked application performance analysis and intrusion detection and prevention. Precise Ethernet timestamps are key to knowing exactly when a networked event occurred. These timestamps are delivered with the packet with which they are associated.

Though these timestamps may appear to have nanosecond resolution, when it comes to knowing what their actual resolution and absolute accuracy is, things are less clear. Most vendors quote an advertised timestamp resolution and have a pretty good idea of the accuracy of the timestamps they place on packets relative to each other. When it comes to their absolute accuracy however, the level of complexity increases; especially when taking into account the contribution of disciplining the oscillator generating the timestamp from an external time reference.

As an internal project, in 2016, Arista set out to design and implement a methodology to measure the absolute accuracy of GbE timestamps on its MetaWatch product. In early 2017, Arista adapted this methodology to measure the absolutely accuracy of MetaWatch's 10GbE timestamps and teamed up with the Securities Technology and Analysis Center (STAC) to implement it as part of an independently verified, new industry standard for measuring time synchronisation accuracy: the STAC-TS (time synchronisation) suite of benchmarks. These benchmarks were enhanced and formalized by the STAC Benchmark Council as part of the STAC-TS benchmark suite. The Council consists of industry experts from trading firms, exchanges, and vendors. STAC-TS is a large set of benchmark standards and software tools for measuring the accuracy and other important characteristics of components used for time synchronization, timestamping, and event capture.

This white paper covers:

- The concept of absolute timestamp accuracy
- The concepts around clock accuracy and synchronisation
- The logic behind the choice of methodology
- The conceptual details of the implementation of the methodology
- The timestamp absolute accuracy results obtained.

2. Packet Timestamping Accuracy

The absolute accuracy of packet timestamps has two main components:

1. How accurately the clock being used to apply the timestamp is synchronised to its reference
2. How accurately the timestamp can be applied to each packet

Synchronising a clock to a reference clock is a complex subject in itself; however, it essentially involves the clock periodically comparing its position against the reference and adjusting as necessary. Invariably the oscillators will run at different rates so the clock being synchronised will constantly have to correct itself to match the reference. To do so, it will need to see a difference between itself and the reference and apply the appropriate correction. If the reference clock is not very frequency stable, it makes it extremely difficult for the clock being synchronised to maintain its accuracy relative to the reference as it is constantly having to change its frequency to match that of the reference.

a. Choice of Oscillator/Clock Options

As there is a huge range of oscillator (clock) options available, varying widely in stability, the choice of oscillator is a key factor in how well a clock can synchronise to a reference. At the extremely accurate end of the scale, oscillators derive their frequency from an electronic transition frequency of an atom e.g. caesium and rubidium. At the far less accurate end of the scale, we have the uncompensated crystal oscillator that derives it frequency from the mechanical resonance of a vibrating crystal of piezoelectric material, typically quartz. Between them, there are various enhancements to the crystal oscillator such as electronically compensating for temperature changes (TCXO) or keeping its temperature constant by enclosing it in an insulated micro-oven (OCXO).

It is also extremely important that the clock being synchronised have a synchronisation time constant – essentially how many successive time updates it averages across before adjusting its time – that works well with the reference clock. A good example of this is synchronising to the GNSS satellite clocks. These clocks are extremely frequency stable when measured over long periods of time. However mainly due to tropospheric delay – the near-earth atmosphere adding a delay to their signal – their short-term timestamp accuracy can only be measured by the receiver by up to about 100 ns. A reference clock synchronised to GNSS satellites will generally set a long enough time constant in the synchronisation algorithm between its local oscillator and the GNSS receiver to average this jitter out.

Timestamp accuracy is ultimately determined by the underlying frequency of the timestamping clock which also defines its resolution. For example, a 156.25 MHz 10GbE clock implemented in an FPGA or a 10GbE network adapter will "tick" every 6.4 ns so timestamps are at best accurate to this granularity or resolution. This clock can also only synchronise itself to an external reference with this level of accuracy. Timestamp accuracy is also determined by clock jitter where a sample being timestamped arrives right on a clock boundary – which will have a frequency jitter component – and may fall into the current or next time quanta. The actual clock used to timestamp also plays a significant part in the ultimate accuracy and it is conceptually simplest to timestamp using the recovered Ethernet clock however its quality is unknown and the IEEE 802.3ae standard allows it to vary by ±100 ppm therefore potentially adding a not-insignificant amount of timestamp jitter.

b. Synchronising Clocks

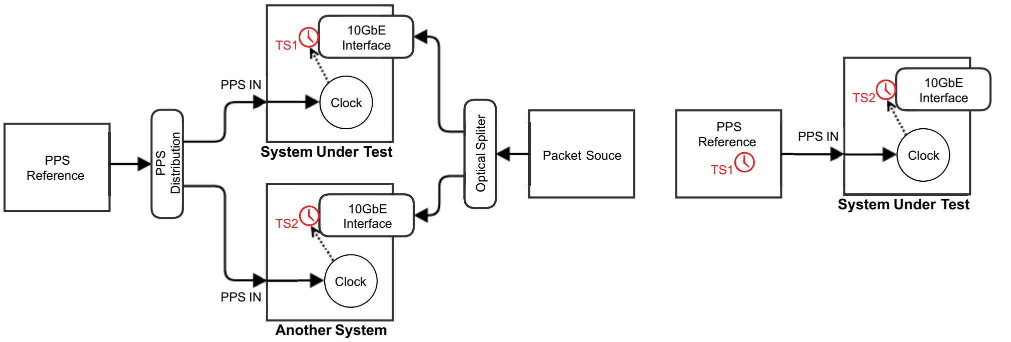
There are numerous ways to synchronise clocks. All of them involve measuring the difference between two or more clocks and attempting to reduce this difference as close to zero as possible by altering its frequency, current position or both. A precise time distribution standard in the networking industry is the use of transmitting a pulse down a 50 Ω coaxial cable every second with the middle of the rising edge of the pulse representing the start of a second. This is known as pulse-persecond (PPS). The sending clock generates the pulse precisely on the second and the receiving clock compares its arrival to its time and adjusts as necessary. Other standards exist, also transmitting pulses and waves down coaxial cables such as 10 MHz which is widely used for frequency synchronisation. There are also clock synchronisation protocols running over Ethernet such as Network Time Protocol (NTP) and Precision Time Protocol (PTP).

It is important to note that when PPS, or any kind of one-way time dissemination is used, the propagation delay through the coaxial cable and any distribution equipment must be compensated for. This is because it takes a finite amount of time for the signal to travel down the cable so the receiving clock will "see" the pulse later than it was sent. Most clocks with PPS inputs allow a compensation value to be set to account for this delay.

3. How to Measure Packet Timestamp Accuracy

An accuracy measurement involves that the quantity being measured be compared to a reference. In the case of packet timestamp accuracy, the requirement is that an asynchronous event; the arrival of a given packet, is timestamped concurrently by a system under test (SUT) and another system and the two generated timestamps compared. Unless the system being compared to the SUT is also the time reference for the SUT, any timestamp comparisons are relative rather than absolute.

There is certainly more than one way to achieve this, however Arista decided to leverage the fact that 10GbE ultimately gets delivered to the Ethernet physical layer from the transceiver as a differential signal pair. PPS is generally delivered down a coaxial cable as a pulse. It is therefore possible to compare the relative arrival of the electrical signals representing the start of an Ethernet frame to the rising edge of a PPS pulse, each in the time domain, and to calculate a timestamp with absolute accuracy (to the reference start of second pulse). The SUT timestamp is then compared to this reference timestamp which allows to determine its absolute accuracy.



Relative: the difference between TS1 (SUT) and TS2(AS) is being measured

Absolute: the difference between TS1 (Ref) and TS2(SUT) is being measured

To minimise time synchronisation error when measuring absolute accuracy, it is desirable to have the most stable reference clock possible. Ideally, a caesium atomic clock would be used as it provides stability in the 5×10^{-13} range (500 fs/ sec jitter). Unfortunately, caesium atomic clocks are rather costly and a rubidium atomic clock is far more affordable. Rubidium clocks exhibit stability in the 1×10^{-11} range (10 ps/sec jitter). An oscillator backed by a GNSS receiver could also be used, however wall-time timestamps are not required for this measurement and depending upon implementation, PPS pulses output from GNSS receivers can exhibit short term jitter of up to 100 ns/s. A free-running rubidium atomic clock does not exhibit this level of short-term jitter and hence provides more than adequate stability to meet the precision criteria of this measurement.

Key to these measurements was an acquisition device capable of performing very accurate simultaneous timestamping of multiple electrical signals with extremely high resolution. Given that 10GbE at Layer 1 is a 5.15625 GHz differential carrier signal encoded with a 64b/66b line code, identifying the reference bit at the beginning of an Ethernet frame required that the frame be decoded into 66-bit blocks and the offset of the middle of the first bit of the frame from the block boundary calculated. A device capable of both capturing the fine details of the 10GbE differential signal and decoding it while preserving each block's precise location in the time domain was required. The ideal device for this task is an oscilloscope. Oscilloscopes are designed to acquire signals on multiple channels, are precisely synchronised in the time domain and are capable of extremely high sampling rates. They also optionally come with numerous protocol decodes including 64b/66b. Ultimately, the single shot resolution of an oscilloscope is defined by its sampling rate. As a 10GbE bit serialises in just under 97 ps, the Nyqvist rate dictates that 10GbE be sampled at a rate of at least 20.625 GSa/s. An oscilloscope capable of meeting, or ideally exceeding this sampling rate was therefore required.

By distributing the same Ethernet stream and start of second reference pulse concurrently to the SUT and oscilloscope, each device can independently reference one to the other. In the case of the SUT, the start of second reference pulse is used to correct the SUT's timestamping clock which timestamps each frame. The oscilloscope's acquisition buffer contains the precise temporal difference between the arrival of the start of second reference pulse and the arrival of the Ethernet frame (defined as the middle of the first bit immediately following the Ethernet preamble). The oscilloscope therefore provides an extremely accurate external reference with which to validate the SUT's packet timestamp's absolute accuracy; incorporating any SUT clock synchronisation error and local clock jitter on timestamps.

When an SUT is being synchronised once per second, it will typically make clock corrections immediately after receiving each synchronisation event. Given oscilloscope acquisition buffers fill up in the order of milliseconds (or less) at high sample rates, the temptation becomes to trigger acquisition on the PPS pulse and capture as many Ethernet frames as will fit into the acquisition buffer; typically frames arriving no more than 1 ms before the pulse and 1 ms after. Though representative of the instantaneous accuracy of the SUT's timestamping ability, it tells us very little about the consistency of the SUT's timestamping accuracy between PPS pulses. It is therefore important to structure the test so that packets can be timestamped throughout the second, ideally for multiple seconds. One way of doing this is to generate a higher frequency pulse, frequency-locked to the time reference which

4. Ethernet Clocks

Section 4, Clauses 51.6.2 and 51.7.2 of the IEEE Standard 802.3-2015 for 10 Gb/s Ethernet specify that the transmit (local) and receive (remote) clock specifications be within ±100 ppm of the 644.53125 MHz reference frequency (derived from the 10.3125 Gbps data rate). As Ethernet does not mandate any clock frequency synchronisation between Ethernet sender and receiver, it is important that all clocks be within this range to protect from buffer overruns (transmit clock is running faster than receive clock). From a timestamping perspective, there are three possible clock sources for the timestamp:

1. The clock recovered from the incoming Ethernet byte stream
2. The local Ethernet transmit (TX) clock
3. A clock outside the requirements of the Ethernet processing

The problem with using the recovered clock is that it is external to the timestamping device and its stability characteristics are essentially unknown and will vary across sending devices. The second two options use clocks local to the timestamping device which can be extremely stable however a mechanism needs to be in place to generate the timestamp as accurately as possible on an asynchronous (to the local clock) byte stream. When measuring timestamp accuracy, it is therefore important to explore the effects on the SUT's timestamping accuracy at both extremes of the sender's permitted Ethernet source clock frequency i.e. 5.15625 GHz ± 100 ppm.

5. An Introduction to MetaWatch

MetaWatch combined with a suitable Arista K-Series device is an application designed to capture, timestamp, buffer and aggregate up to 30 1/10GbE ports. MetaWatch offers an unmatched combination of features in the aggregation tap/packet broker space:

- Integrated 10GbE Layer 1 matrix switch offering port mirroring and pass-through adding the same latency as 1 m of fibre
- Full Ethernet per-port statistics
- 2x15:1 1/10GbE buffered aggregation via a 8 GB or 32 GB buffer
- 1 ns timestamp resolution capturing Ethernet packets
- Time synchronisation support for PTP, NTP, optionally coupled with PPS
- Support for IEEE 802.3x PAUSE frames on the aggregated egress ports allowing consuming devices to leverage the deep buffers to moderate incoming packet rate

MetaWatch supports NTP, PTP and PPS to synchronise to a timing reference. In theory, each can provide sub-nanosecond accuracy. In practice, however, currently 1 PPS provides the greatest accuracy. MetaWatch uses the device's local oscillator (VCXO) by default. Both OCXO and atomic (rubidium) clock modules are optionally available providing more stability, particularly over longer holdover periods (periods where the reference clock is unavailable, most often due to unavailability of the reference clock or a path to the reference clock).

6. The Benchmark Methodology

The purpose of this methodology was to characterise the accuracy of MetaWatch 10GbE 1 ns-resolution timestamps running on an Arista 7130 Series 32K device. The Arista 7130 32K is a 32-port network device comprising Layer 1+ switching and an FPGA with the optional Atomic Clock Module, synchronised via PPS. MetaWatch on an Arista 7130 32K was therefore the SUT.

Inputs provided to MetaWatch were:

- A 10GbE SFP+ transceiver delivering Ethernet packets containing 64-byte Ethernet frames at line-rate each containing a 32-bit sequence number plugged into a single port on the Arista 7130 32K
- A PPS reference pulse delivered down a 50 Ω coaxial cable

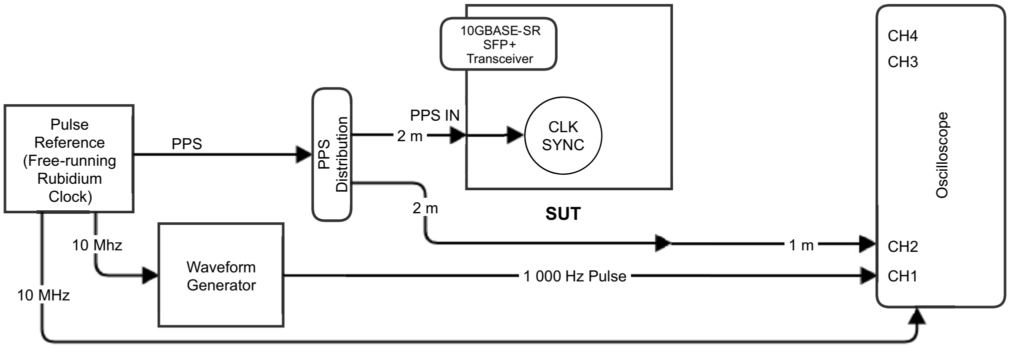
MetaWatch outputs a stream of packets on a 10GbE output port with Arista's industry standard trailer containing MetaWatch's nanosecond-precise timestamp applied to each Ethernet frame.

COMPONENT	KEY FUNCTIONALITY
Pulse reference	SRS rubidium frequency standard with <2 × 10-11 (1 s) stability, PPS and 10 MHz Out
Pulse distribution unit	TimeTech Pulse Distribution Unit - 1U 1:16 PPS with minimal skew and jitter
Waveform generator	Keysight AWG providing stable 1 000 Hz pulse frequency locked to 10 MHz from pulse reference
Oscilloscope	LeCroy 36 GHz, 80 GSa/s, 256 MS buffer with 64b/66b serial decode, timebase frequency-locked to 10 MHz from pulse reference
Packet source/capture	Simultaneously deliver and capture consistent line-rate packets without loss
Adjustable Ethernet clock source	Move packet source Ethernet clock frequency ±100 ppm from reference
Optical 10 GbE splitter	Send the same 10GbE stream to both the SUT and oscilloscope with minimal skew
SFP+ to SMA breakout board	Convert 10GBASE-SR SFP+ transceiver output to 2 × SMA 50 Ω coax
10GBASE-SR transceivers	Feed the optical splitter from the packet source and deliver the Ethernet stream to the SUT and the breakout board
1 m and 2 m SMA-SMA 50 Ω low loss test cables	Interconnect all above electrical components
LC-LC 50/125 OM4 duplex multimode fiber optic Cables	Interconnect all above optical components

a. **Pulse Distribution**

The pulse reference fed PPS to the pulse distribution unit which re-amplified the incoming pulse and sent it out on two ports with extremely low port-to-port skew and jitter (measured at 2 ps with the oscilloscope). It also provided a matched and isolated 50 Ω impedance to both MetaWatch and oscilloscope. To make the most of the acquisition buffer in the oscilloscope, a waveform generator was used to generate a frequency-stable 1 000 Hz pulse from the pulse reference's 10 Mhz output. This 1 000 Hz pulse would trigger the oscilloscope to acquire 1 000 acquisition buffer segments every second allowing accuracies to be measured over multiple seconds rather than over a few contiguous milliseconds imposed by the 80 GSa/s sampling rate of the oscilloscope.

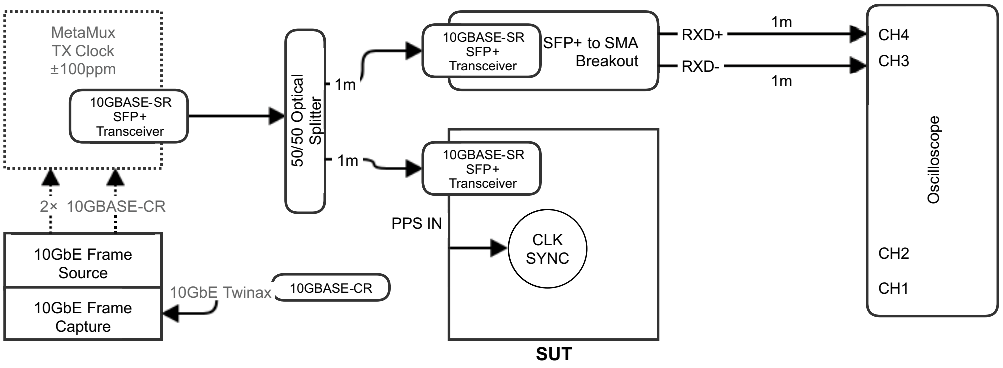
PPS reference pulse to SUT and oscilloscope (also showing 1000 Hz to oscilloscope and 10 MHz to oscilloscope and waveform generator timebase inputs)



b. **Ethernet Packet Distribution**

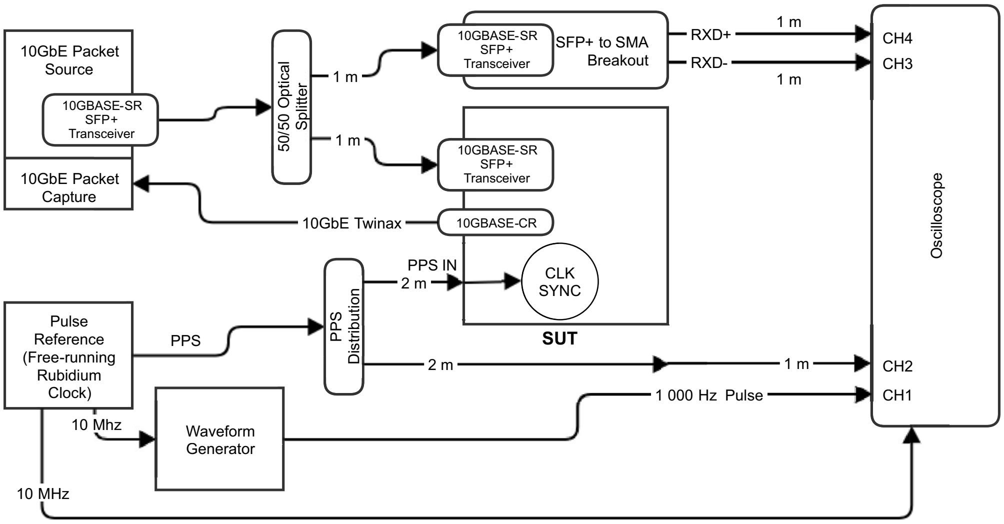
A two-way optical splitter was used to send a stream of Ethernet packets containing 64-byte Ethernet frames at line rate to MetaWatch and oscilloscope (via the SFP+ to SMA breakout board) from the packet source.

Three sets of test runs were completed. The first set had the packet source connected directly to the optical splitter. For the second two sets, requiring the packet source's Ethernet transmit clock to be run at the extremes of allowable Ethernet clock tolerances, an identical Arista 7130 32C running MetaMux (an application offering ultra-low-latency packet multiplexing and aggregation) was added between the packet source and the optical splitter. In general, only network test equipment allows user-control over the frequency of oscillator(s) supplying Ethernet interface clocks. The packet source was no exception with its fixed-frequency oscillator supplying its Ethernet interface clocks. Given the Arista 7130 32K is an Arista product, Arista has complete control of the oscillator providing the Ethernet transmit clock thus allowing its frequency to be adjusted extremely finely. It was first adjusted to +100 ppm and then -100ppm from the Ethernet 5.15625 GHz base clock frequency as measured against the pulse reference. For the test runs at +100 ppm, a second "top up" Ethernet stream was also sent from the packet source into MetaMux to ensure that line-rate was maintained as with the 100 ppm increase in Ethernet clock TX frequency. With the Ethernet transmit clock running at this frequency extreme, receiving from a single Ethernet source, MetaMux would not have been able to send packets out at line-rate.



Complete test harness and SUT (MetaMux removed for clarity)

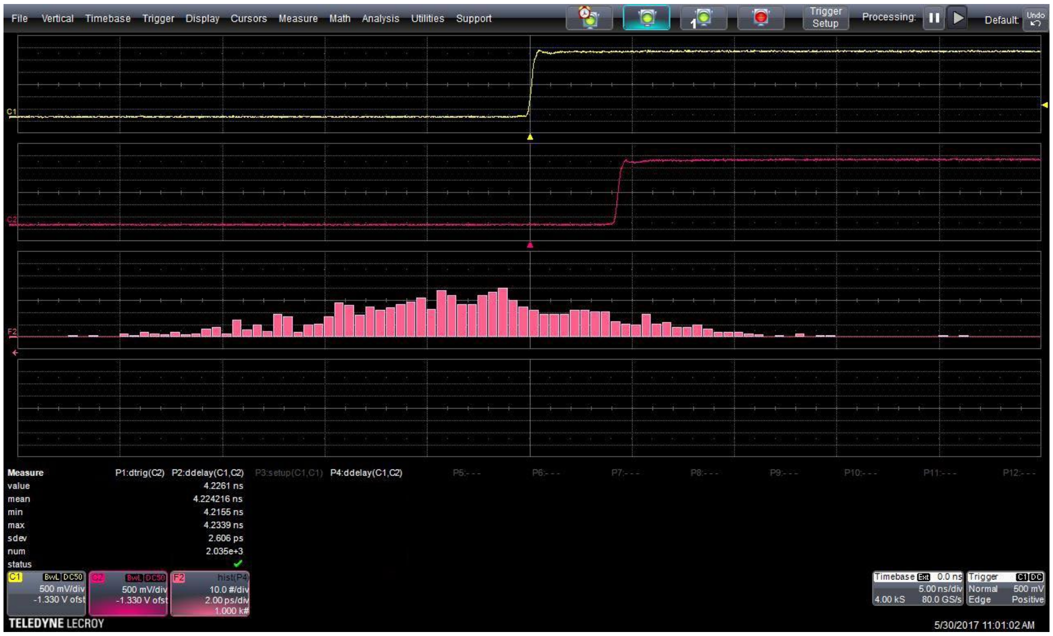
The pulse reference distributes pulses concurrently to both the SUT and the oscilloscope. The packet source/capture distributes Ethernet packets concurrently to both the SUT and the oscilloscope and captures the packets after they have been timestamped by the SUT.)



c. **Test Harness Calibration**

As the oscilloscope was capable of measuring acquisition samples to ±12.5 ps, it was possible to measure and account for any skew between a) Ethernet frame arrival and b) PPS reference pulse arrival at the SUT and oscilloscope extremely precisely.

Calibrating the relative arrival of the PPS from the outputs from the PPS distribution unit was fairly straight forward as it involved comparing them in the time domain on different oscilloscope channels. Though not strictly necessary, as all propagation delays were measured and accounted for. An additional 1 m coaxial test cable was added to the PPS output to the oscilloscope for the actual testing, via an SMA-SMA coupler. This was to compensate for the 1 m coaxial test cables between the SFP+ to SMA breakout board and the oscilloscope.



```
Measure   P1:dtrig(C2)  P2:ddelay(C1,C2)  P3:setup(C1,C1)  P4:ddelay(C1,C2)  P5:---  P6:---  P7:---  P8:---  P9:---  P10:---  P11:---  P12:---
value                       4.2261 ns
mean                      4.224216 ns
min                         4.2155 ns
max                         4.2339 ns
sdev                         2.606 ps
num                         2.035e+3
status                             ✓

C1  BwL DC50      C2  BwL DC50      F2       hist(P4)
    500 mV/div        500 mV/div             10.0 #/div
 -1.330 V ofst     -1.330 V ofst             2.00 ps/div
                                             1.000 k#

Timebase Ext  0.0 ns    Trigger     C1 DC
          5.00 ns/div   Normal     500 mV
4.00 kS   80.0 GS/s     Edge     Positive

5/30/2017 11:01:02 AM
```

i. Calibrating PPS reference pulse distribution

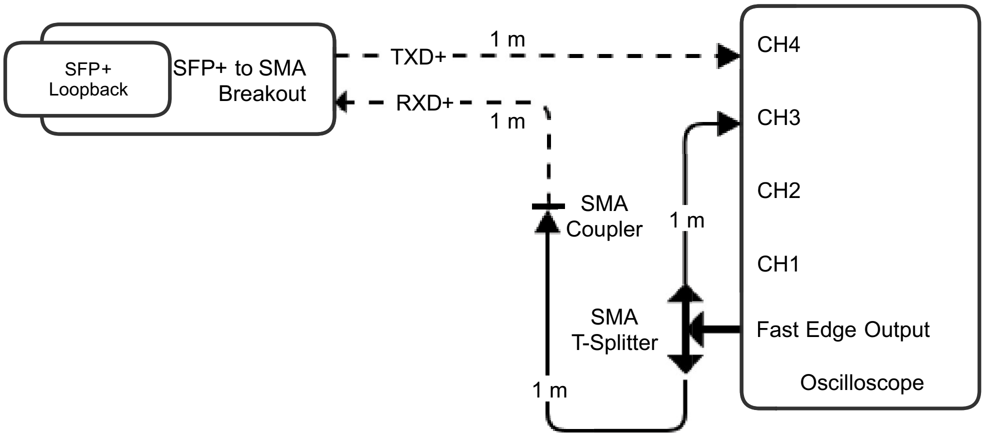
Channel 1 (C1) was the PPS reference pulse destined for the SUT. Channel 2 (C2) trails C1 by ~4.22 ns which is the propagation delay down the additional 1 m of coaxial destined for the oscilloscope combined with any skew in the pulse distribution unit.

The oscilloscope measurement statistics show that over 2 000 PPS pulses, the difference in their arrival times was within a range of 4 233.9 ps - 4 215.5 ps = 18.4 ps. The mean was 4 224.2 ps. Given the sampling resolution error of the oscilloscope was ±12.5 ps which is greater than jitter between the pulse arrivals, the SUT and the oscilloscope essentially receive PPS pulses coincident to the ultimate sampling resolution of the oscilloscope with 4 224 ps of skew which was to be accounted for when calculating the results.

ii. Calibrating Ethernet Packet Distribution

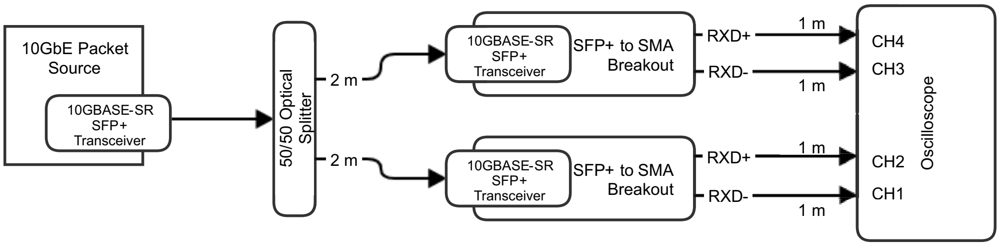
Packet distribution from the packet source required a physical media converter for the oscilloscope as the standard for Ethernet used by the SUT was the SFP+ cage. Therefore, an SFP+ to SMA breakout board was used. It was connected to the oscilloscope via a pair of 1 m coaxial test cables. The breakout board was a passive device connecting the SFP+ differential receive (RX) pins from the 10GBASE-SR SFP+ transceiver to the coaxial test cables. The combination of the breakout board and the coaxial test cables therefore added a fixed and measurable propagation delay between the 10GBASE-SR SFP+

transceiver and oscilloscope. This propagation delay was measured and accounted for as follows with the help of an SFP+ loopback transceiver:



The oscilloscope's fast edge output provides a constant stream of pulses with extremely short rise-times and a very low impedance allowing it to serve as a common reference for propagation delay measurement and probe de-skewing. In the above diagram, an SMA T-splitter is used to send the same pulse down two 1 m coaxial test cables. Initially each was connected to a separate oscilloscope channel input (channels 3 and 4) to ensure that there was no skew in the arrival of the pulses. Then, the SFP+ to SMA breakout board and pair of 1 m coaxial test cables were added between the coaxial test cable connected to C4 of the oscilloscope. In a similar way to the PPS calibration, the propagation delay through the coaxial test cables, breakout board and SFP+ loopback transceiver was measured. The propagation delay through the SFP+ loopback transceiver was then subtracted before dividing the remaining propagation delay by two, as signals from the transceiver only passed through the board and 1 m of copper coaxial once during the test rather than twice during this calibration procedure.

The next step was to measure and potentially correct for any skew in the optical splitter, the fibre patch cables and the 10GBASE-SR transceivers.



The above test harness was used. The propagation delay through each SFP+ to SMA breakout board and pair of 1 m test cables was measured as described above and any resultant skew corrected for. Packets were then passed through the splitter and the difference in arrival time at the oscilloscope of each pair of packets was calculated by post-processing the output from the oscilloscope's 64b/66b decode on each Ethernet stream.

d. The mechanics of the Test Runs

The goal was to compare oscilloscope versus SUT timestamps, both referenced to the PPS reference pulse from the pulse reference, intra-second and inter-second. The purpose of the 1 000 Hz pulse source was to trigger oscilloscope capture 1 000 times during each second across multiple seconds. It was determined that each run last three seconds. This duration was dictated by a combination of the acquisition buffer size of the oscilloscope and the time taken to copy the 64b/66b decoded result set from the oscilloscope for analysis. Essentially, 3 000 acquisition segments (3 seconds), each containing 6 to 7 Ethernet packets, were captured and correlated during each test run.

With the test harness calibrated taking into account the propagation delays through each relevant component of the test harness, each test run was comprised of the following:

1. Starting 10GbE packet capture
2. Starting the packet source, causing it to produce Ethernet packets containing 64-byte Ethernet frames at line-rate, each containing a 32-bit unique sequence number
3. Immediately starting acquisition on the oscilloscope causing it to capture a segment containing 40 KSa on the rising edge of each 1 000 Hz pulse
4. Stopping capture once the three seconds had elapsed

The oscilloscope captured the following channels during the run:

1. The PPS reference pulses
2. The 1 000 Hz pulses
3. The differential 10GBASE-R Ethernet pair

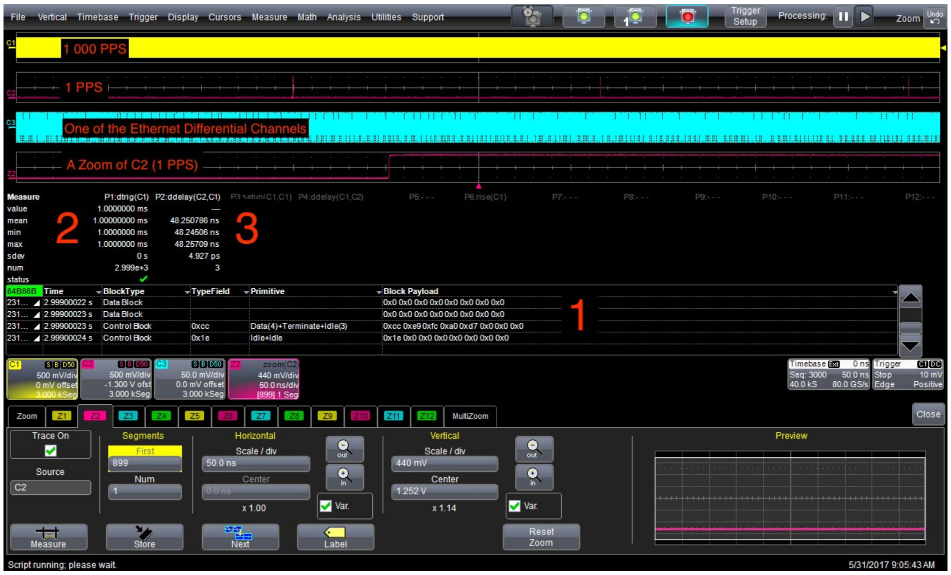
Packet capture captured every resulting Ethernet frame from the MetaWatch.

e. Post Processing

i. Oscilloscope

Immediately following acquisition, the oscilloscope was configured to:

1. Apply its 64b/66b decode to the Ethernet channels
2. Calculate and provide statistics on the elapsed acquisition times between each 1 000 Hz trigger pulse (validating the stability of the 1 000 Hz source)
3. Calculate and provide statistics on the elapsed acquisition times between the 1 000 Hz pulse triggering acquisition of the segment containing each PPS reference pulse



Measure	P1:dtrig(C1)	P2:ddelay(C2,C1)	P3:setup(C1,C1)	P4:ddelay(C1,C2)	P5:- - -	P6:rise(C1)	P7:- - -	P8:- - -	P9:- - -	P10:- - -	P11:- - -	P12:- - -
value	1.0000000 ms	---										
mean	1.00000000 ms	48.250786 ns										
min	1.0000000 ms	48.24506 ns										
max	1.0000000 ms	48.25709 ns										
sdev	0 s	4.927 ps										
num	2.999e+3	3										
status	✓											

64B66B	Time	BlockType	TypeField	Primitive	Block Payload
231...	2.99900022 s	Data Block			0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0
231...	2.99900023 s	Data Block			0x0 0x0 0x0 0x0 0x0 0x0 0x0 0x0
231...	2.99900023 s	Control Block	0xcc	Data(4)+Terminate+Idle(3)	0xcc 0xe9 0xfc 0xa0 0xd7 0x0 0x0 0x0
231...	2.99900024 s	Control Block	0x1e	Idle+Idle	0x1e 0x0 0x0 0x0 0x0 0x0 0x0 0x0

C1 S B D50 500 mV/div 0 mV offset 3.000 kSeg
C2 S B D50 500 mV/div -1.300 V ofst 3.000 kSeg
C3 S B D50 50.0 mV/div 0.0 mV offset 3.000 kSeg
Z2 zoom(C2) 440 mV/div 50.0 ns/div [899] 1 Seg

Timebase Ext 0 ns Seq: 3000 50.0 ns 40.0 kS 80.0 GS/s
Trigger C1 DC Stop 10 mV Edge Positive

Trace On ☑
Source: C2
Segments: First 899, Num 1
Horizontal: Scale / div 50.0 ns, Center 0.0 ns, x 1.00, ☑ Var.
Vertical: Scale / div 440 mV, Center 1.252 V, x 1.14, ☑ Var.

Script running; please wait.
5/31/2017 9:05:43 AM

64b/66b Decode Packet Reassembly

A custom extraction script was written that was executed on the oscilloscope that dumped the contents of the 64b/66b decode table with picosecond acquisition timestamps for each 66b-block before:

1. Extracting the precise oscilloscope acquisition time offset for each Ethernet frame.
2. Validating the Ethernet frame check sequence (FCS) of each Ethernet packet. If incorrect, the Ethernet packet was discarded to avoid the possibility of a corrupted sequence number generating a false positive match.
3. The sequence number in the Ethernet packet payload was extracted, coupled with the start-of-frame acquisition time offset and written to a file for post analysis.

ii. Packet Capture

Following the test run, an analysis script was run on all the Ethernet frames captured from MetaWatch which extracted each sequence number from the Ethernet frames' payload and their associated nanosecond-resolution timestamps from the packet trailer added by the MetaWatch and wrote them to a file.

iii. Correlating Oscilloscope Capture and Packet Capture

The previous two steps generated files containing pairs of timestamps and sequence numbers. The timestamps in the file from the packet capture device came from the MetaWatch with nanosecond resolution. The timestamps in the file from the oscilloscope were relative to the start of oscilloscope acquisition, that means the first 1 000 PPS pulse triggering acquisition of the first segment. To correlate them, the precise offset from the start of oscilloscope acquisition to the nearest PPS reference pulse was required. This offset would then be added to each oscilloscope start of Ethernet frame timestamp to align it in time with the nanosecond portion of the MetaWatch timestamp. It was obtained by having the oscilloscope measure the relative time between each PPS reference pulse and the 1 000 Hz pulse that triggered the segment containing it.

The previous two steps generated files containing pairs of timestamps and sequence numbers. The timestamps in the file from the packet capture device came from MetaWatch with nanosecond resolution. The timestamps in the file from the oscilloscope were relative to the start of oscilloscope acquisition i.e. the first 1 000 PPS pulse triggering acquisition of the first segment. To correlate them, the precise offset from the start of oscilloscope acquisition to the nearest PPS start-of-second reference pulse was required. This offset would then be added to each oscilloscope start-of-Ethernet-frame timestamp to align it in time with the nanosecond portion of the MetaWatch timestamp. It was obtained by having the oscilloscope measure the relative time between each PPS reference pulse and the 1 000 Hz pulse that triggered the segment containing it. For example, in the above oscilloscope screenshot, the three PPS reference pulses were an average of 48.251 ns (with 12 ps of jitter) before the 1 000 Hz trigger. By identifying in which segment the first PPS reference pulse occurred, the number of 1 ms segments prior to it could be added to it to obtain the time difference between the start of acquisition and the start of second. Continuing this example, the first PPS reference pulse occurred in the 899th segment with the PPS reference pulse occurring 48.250 ns before the segment trigger hence the oscilloscope acquisition started 898 ms - 48.250 ns before the PPS reference pulse. By adding this offset of 897 999 951.750 ns to each Ethernet frame's oscilloscope acquisition time, timestamps were correlated to the 1 PPS reference pulse and could now be compared to the SUT packet timestamps.

The final step involved comparing each Ethernet packet acquired by the oscilloscope's PPS reference pulse correlated timestamp to the corresponding MetaWatch Ethernet frame's timestamp. This was done by matching their sequence numbers.

7. Absolute Accuracy Results

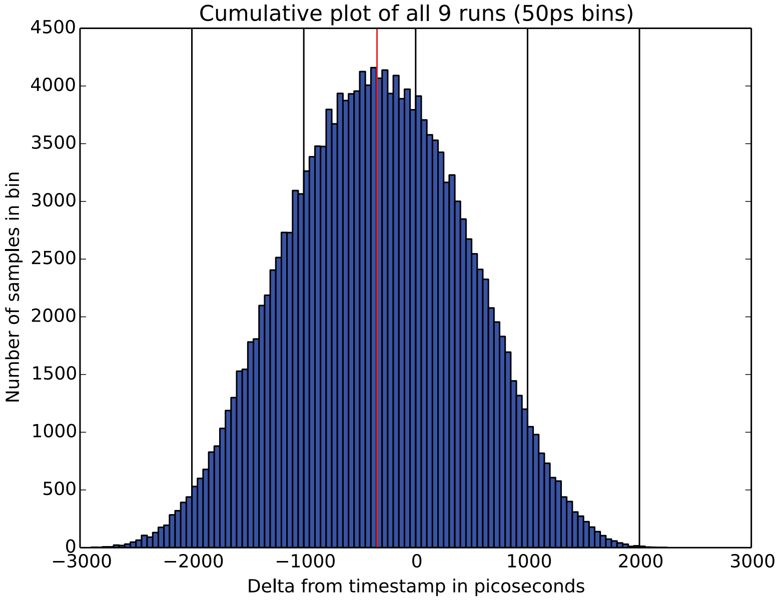
As stated previously, three groups of three runs were performed:

1. Three runs with the packet source's Ethernet clock frequency unmodified
2. Three runs via an Arista 7130 32C running MetaMux with its Ethernet clock frequency increased by 100 ppm (as measured against the rubidium frequency reference)
3. Three runs via an Arista 7130 32K running MetaMux with its Ethernet clock frequency reduced by 100 ppm (as measured against the rubidium frequency reference)

In each run, the oscilloscope acquired a segment containing 6 to 7 Ethernet packets every millisecond for three seconds resulting in just over 19 000 Ethernet packets acquired per run.

In analysing the results of these runs, no statistically valid differences were observed across the groups of runs confirming that MetaWatch's ability to timestamp is not influenced by the sender's Ethernet clock frequency.

Plotting 50 ps buckets of the deltas between the SUT timestamp (ns) and the oscilloscope timestamps (ps) across all nine runs (172 500 oscilloscope-acquired Ethernet packets) yielded the following results:



IN NUMBERS:	
Min	-3 ns
Mean	-0.345 ns
Max	2 ns

50 ps bins were chosen in calculating the latency distribution because this represents the measurement error of ±25 ps obtained from the test harness calibration.

a. Results summary

MetaWatch 0.5.2 running on MOS-0.14.0alpha3 on a Arista 7130 32K with the Atomic Clock Module option, when disciplined from a PPS source, achieved:

- 1 ns timestamp resolution
- Absolute timestamp accuracy (100% of timestamps) of -3 ns/+2 ns for the specified input port
- 42.4% of timestamps accurate to the nanosecond
- 92.3% of timestamps accurate to ±1 ns

8. Conclusion

Arista set out to design and implement a test methodology to measure the absolute accuracy of the MetaWatch 10GbE capture and timestamping application on the Arista 7130 32K device to within double-digit picosecond precision. To achieve this goal, Arista proposed test specifications to the STAC Benchmark Council, which made key enhancements and formalised the specifications as part of its STAC-TS suite.

Arista built a test harness based on the final specifications, then conducted calibration and testing. STAC confirmed that the uncertainty in the measurement of the results was 50 picoseconds, thus achieving Arista's accuracy goal. They also confirmed that for the capture port tested, 100% of samples were between -3 ns and +2 ns of the time reference and 92.3% of timestamps were accurate to ±1 ns.

From Arista's perspective, these are excellent results that more than justify the hard work put into designing the Arista 7130 32K hardware and the MetaWatch application. Clients can leverage MetaWatch to timestamp their packets using this solution and be confident that the timestamps are accurate. The STAC Report containing these and other STAC-TS results is available for download.

Santa Clara—Corporate Headquarters
5453 Great America Parkway,
Santa Clara, CA 95054

Phone: +1-408-547-5500
Fax: +1-408-538-8920
Email: info@arista.com

Ireland—International Headquarters
3130 Atlantic Avenue
Westpark Business Campus
Shannon, Co. Clare
Ireland

Vancouver—R&D Office
9200 Glenlyon Pkwy, Unit 300
Burnaby, British Columbia
Canada V5J 5J8

San Francisco—R&D and Sales Office 1390
Market Street, Suite 800
San Francisco, CA 94102

India—R&D Office
Global Tech Park, Tower A & B, 11th Floor

Marathahalli Outer Ring Road
Devarabeesanahalli Village, Varthur Hobli
Bangalore, India 560103

Singapore—APAC Administrative Office
9 Temasek Boulevard

#29-01, Suntec Tower Two
Singapore 038989

Nashua—R&D Office
10 Tara Boulevard
Nashua, NH 03062