Measuring the latency of a 4ns switch

Today, Arista's MetaWatch application provides picosecond level timestamping on packets. Prior to this being available, it was still important to us to give our clients a measure of the latency of our Layer 1 switches, the 7130 Connect Series of our Layer 1 switches of the 7130 Connect Series. It needed to be reliable and transparent as to its calculation.

While MetaWatch provides a much more accurate methodology for testing the latency through any component, in this blog we have outlined our previous methodology, both for historical reference and to document an independent method of testing latency.

Having enjoyed a healthy scepticism for others' benchmarks in the past, we hold ourselves to the highest standards with respect to this. We understood, based on the design of the 7130 Connect Series, that the Layer 1 switches should have a latency of approx. 5 nanoseconds and wanted to verify that. We devised an independent methodology which gives repeatable results that the industry can place its trust in.

Sources of error

As the latency of network technology has decreased, the technology required to measure it accurately has had to improve. There are a number of challenges with measuring low-latency devices.

Latency measurements are usually based on a timestamped packet capture – comparing the arrival times of two packets. If the timestamps are applied in software there is inevitably a non-determinism that comes when sharing resources (processors, caches, busses, etc). Network cards from several manufacturers include **hardware timestamping** and time synchronization hardware (dedicated packet capture cards like those from other third party companies) provide a line-rate capture with a hardware timestamping feature. These solutions solve the packet capture problem, but not the problems with experimental error that often occurs. Hardware-based timestamping, implemented well, should give accurate and deterministic results.

The next problem is one of **resolution** – hardware timestamping cards usually have a resolution of around 4ns – that's about the same as the maximum useful clock speed in the FPGAs generally used to implement them. The accuracy of the cards is generally much worse than the precision – you can't necessarily believe what the cards tell you.

Lastly, the **synchronization** accuracy is usually much worse again. Light travels through fiber (and copper) at roughly 5ns per meter. That means that variation in the amount of fiber in the test system can change the test results.

The problem for us in measuring the 7130 Layer 1 latency is that the latency we're measuring (around 4ns) is smaller than the precision of the packet capture cards, and is also about the same as the fibers used to connect it to the test equipment.

Test configuration

To overcome these problems we used the following setup:

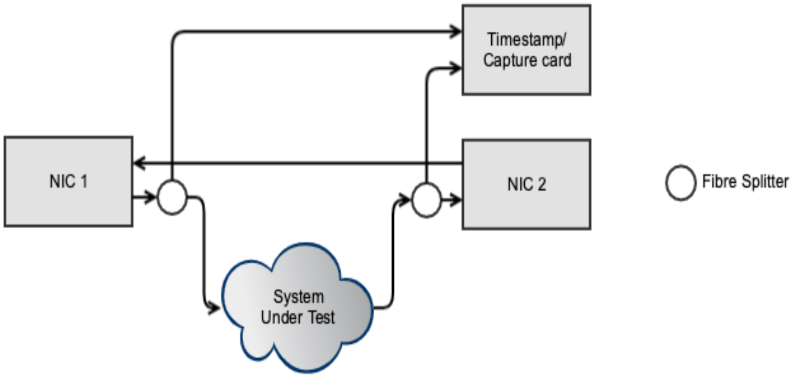


Figure 1

Two machines with two network cards (NIC 1 and NIC 2) are connected to each other directly in one direction, and via the system under test in the other. OM1 multimode fiber is used for connections throughout. The input and output of the system under test are tapped using optical taps, and the result is fed into a timestamp/capture card. The capture card has two ports on it which can timestamp using the same clock – i.e. there is no synchronization error. The published accuracy for the timestamp/capture card in use was 7.5ns.

In these tests, NIC 1 sends ping packets – ICMP requests – to NIC 2, which responds with ICMP replies. The rate and size of the ping packets can be adjusted but for the purposes of this test we have simply used a ping flood. Other test patterns could be generated, but for the purposes of the 7130 tests, the nature of the packets is not overly important (since the 7130 Connect devices forward at the bit level, the contents at Layer 2 – packets – are not a relevant performance metric).

We pass over 1 million of these packets through the system under test, and capture the packets and their timestamps using the DAG card. Comparing the timestamp from the packet going into the SUT to the timestamp on the packet coming out of the SUT gives us the latency of the SUT for the test configuration. We run an analysis over the million packets passed through the SUT for each test in order to obtain statistics about that test run.

Methodology

To overcome the effects of resolution and quantization error, we tested a large number of different configurations, where the only change in the latency of each configuration was the number of times the signal through the SUT passed through an Arista 7130-16 device (a 16 port, layer 1 device). To do this we set up a base-line system like this:

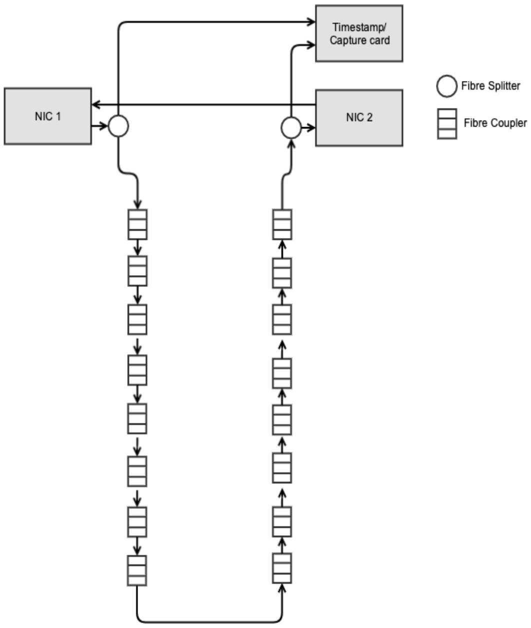


Figure 2

Here we have used fiber couplers to join a number of separate fibers together. We use 16 joiners in this configuration, and they are connected by 17 fibers.

To determine the latency of the Arista 7130 device, the 16 joiners are replaced by connections to the 7130 – the 7130 is configured to transmit anything received on a port back to the same port – a loopback. (Figure 3) We have no reason to expect that the loopback latency would be different to the latency to any other port – the data is taking the same path through the matrix switch (i.e. the loopback is not a shortcut).

For this test, we used 10G SFP+ modules in the Arista 7130 and NIC cards running at 10G. Note that for the purposes of this test we have no way of measuring the latency through the SFP+ modules. We would expect this to be on the order of a few hundred picoseconds for the round-trip.

By replacing the joiners, one-by-one, with a "hop" through the Arista 7130, we are able to determine the difference in latency for one hop, two hops, three hops, four hops, etc. Up to sixteen hops through the 7130. This diagram shows two of the couplers having been replaced with hops through the 7130.

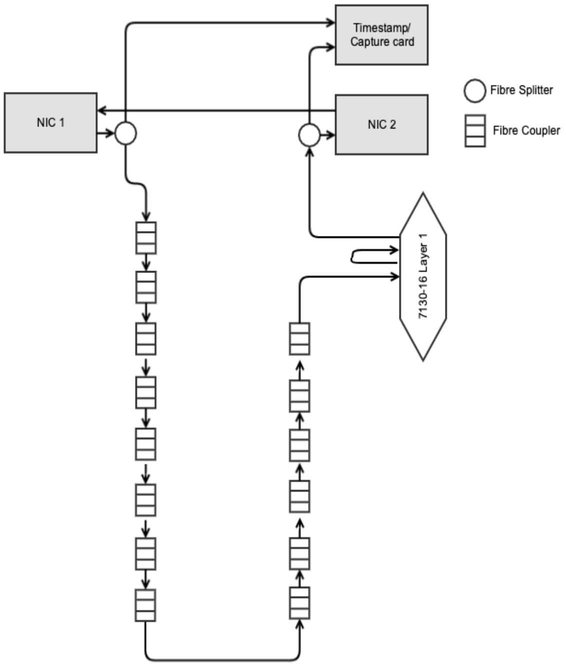
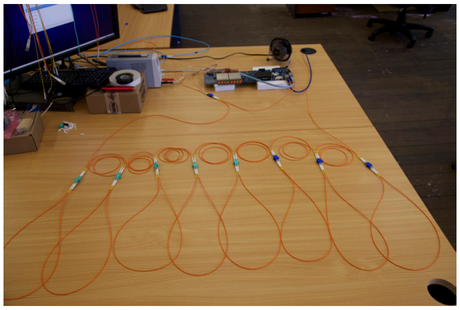


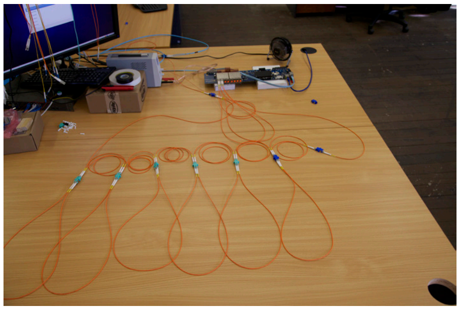
Figure 3

To get an idea of what this involved, see some of these pictures taken during this series of tests:

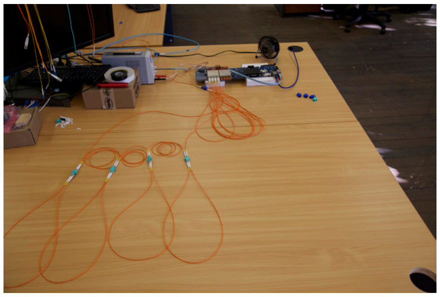
1 hop



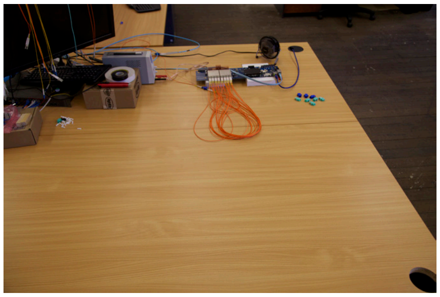
3 hops



8 hops



16 hops



Results

The mean results for each run are recorded and plotted here:

Ports & Latency (ns)	
3	110.779880797
4	114.873343863
5	118.690429716
6	125.849926528
7	129.397969418
8	133.460608952
9	137.394684852
10	140.994296039
11	144.996054107
12	148.986578401
13	153.195699513
14	157.251650876
15	158.195721974

Figure 4

Plotting this data gives us the following:

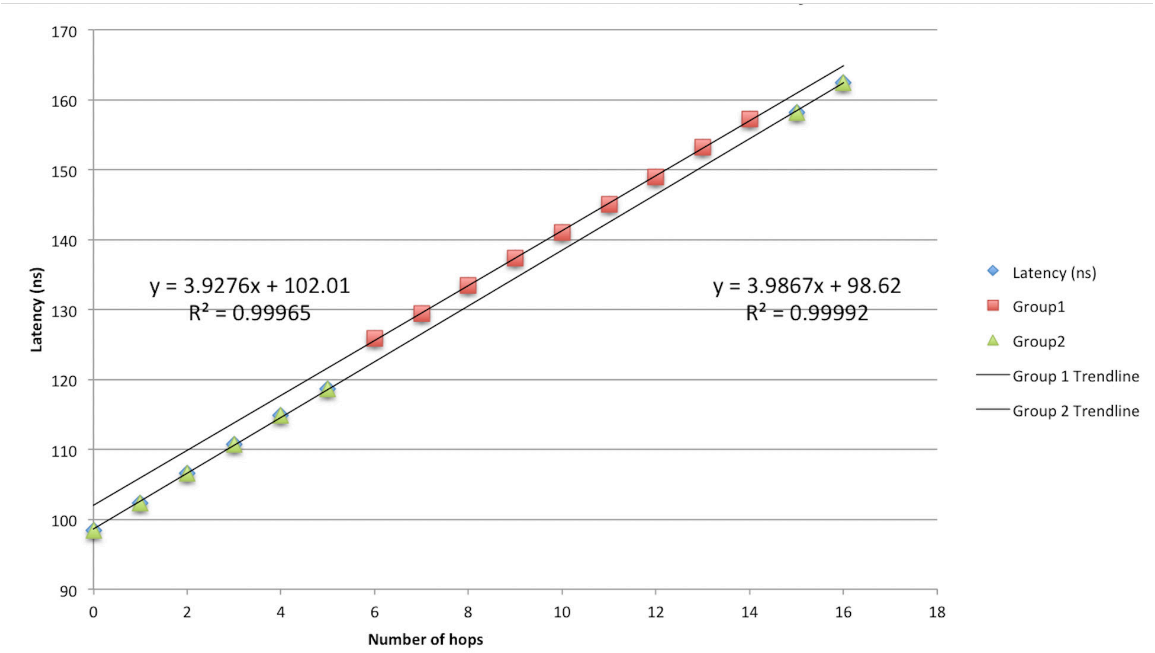


Figure 5

This is interesting!

We have fitted two lines to the data to show that it is bi-modal. Two different groups of measurements, each with a very similar gradient but a different offset. The R^2 is high for each linear model, and the gradient of each line is very similar – a bit over 3.9ns per hop. Our explanation for these results is counter-intuitive. We believe this is caused by aliasing – because there is very little jitter, we are affected by the quantization error in the timestamp/capture card. i.e. because the DAG has an accuracy of 7ns and the timestamps are quantised to this, the timestamps will be rounded. For different latencies, they will either be rounded up or down – i.e. it is bi-modal. We are seeing the network-latency equivalent of Moire Patterns.

So, by fitting two linear functions, observing that the coefficient of determination (R2) is high, and that the two fits have an extremely similar gradient, we can be confident in our measurements.

The gradients above indicate the latency per hop for that dataset – i.e. Arista 7130-16 has a latency per hop of 3.95ns.

Conclusion

We have discussed a methodology for measuring the latency through an extremely low latency device – the Arista 7130-16 Connect, and which can also be used to test any layer 1 device under test.

By taking a number of measurements and fitting a curve to those measurements we have been able to identify and compensate for several sources of error.

We have conclusively shown that the latency of the Arista 7130-16 is very close to 3.95ns per hop.

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

Copyright © 2019 Arista Networks, Inc. All rights reserved. CloudVision, and EOS are registered trademarks and Arista Networks is a trademark of Arista Networks, Inc. All other company names are trademarks of their respective holders. Information in this document is subject to change without notice. Certain features may not yet be available. Arista Networks, Inc. assumes no responsibility for any errors that may appear in this document.   July 2019