March 10, 2016

ENSURING INTEROPERABILITY FOR OPEN NETWORKING SYSTEMS

WHITE PAPER

PRESENTED BY: CHRISTINA DUBE (UNH-CRC), DANIEL GRAY (UNH-IOL), DAVID WOOLF (UNH-IOL)

21 MADBURY RD, SUITE 2
DURHAM, NH 03824
+1-603-862-0090
WWW.IOL.UNH.EDU

TABLE OF CONTENTS

ABSTRACT

Open Networking has created new opportunities and new challenges for Data Center designers and implementers. Many customers are used to deploying monolithic solutions from a single integrator. When implementing Open Networking solutions from multiple providers, a few key questions arise: Will it all work? Are there common interfaces I can use to monitoring and control the hardware? How do Data Center designers and implementers ensure that the components they've chosen will be interoperable? Will the Open Networking Solution work as well as or better than a traditional monolithic solution?

This paper will cover key projects related to Interoperability, which are underway in the OCP Networking group:

1. Interoperability Testing including:

   a) A brief history of the project at the University of New Hampshire InterOperability Lab (to test and validate combinations of whitebox switches, network operating systems and interface modules)
   b) An outline of the test procedures for ensuring basic interoperability of Open Networking components (including proper physical layer operation for cables, pluggable modules, and switches, as well as baseline packet loss expectations) plans for further expansion of the testing in 2016.

INTRODUCTION

The advent of Open Networking solutions, pioneered through the Open Compute Project, has brought new opportunities to the data center. The solutions allow for greater flexibility and customization while reducing upfront cost. As more companies explore "rolling their own" data center solutions, in effect acting as their own integrators, they quickly realize a key trade-off. While acting as your own integrator allows greater flexibility and customization, ensuring the interoperability of all the components used in building a data center network requires an immense amount of effort. Testing the combination of Network Operating Systems, whitebox switches, optical transceivers, and passive and active cables takes time and expertise, creating a mountain of work for for both greenfield and brownfield data center deployments. In order to give data center implementers a leg up on this work, the Open Networking community has come together to create an Open Networking Integrators List. The goal of this list is to show users what combinations of products have been reliably demonstrated to work together. These products can help jump start data center projects, as they can be counted on as starting candidates for a total data center solution.

In the tradition of Open Networking, the list itself and the testing behind it is completely open.

End users are able to see exactly what works and what doesn't, including:

- Test Plans
- Test Scripts and Configurations
- Test Results

The test results will allow end users to browse the many combinations of Open Networking products (network device, pluggable, and network operating system) that are proven to be interoperable.

The concept of Open Networking started due to operational efficiency needs of consumers and users who are building and running large data centers. Significant innovations have occurred in this emerging market such as various OCP Networking Projects like Open Network Install Environment (ONIE) and Switch Abstraction Interface (SAI). With these innovations, more and more adoption is occurring among the enterprises and small businesses that leverage Open Networking techniques. These technologies allow alternative networking solutions and lower price points for everyone.

The sections that follow will walk the reader through the various tests that are performed as part of Open Network Systems Interoperability testing. This discussion will provide better understanding of the test procedures, and the motivation behind why certain tests are performed. The section titles and information are drawn directly from the Open Network Systems Interoperability Test Plan. The latest versions of this test plan can be found here.

MANAGEMENT OF NOS THROUGH ONIE (GROUP 1 TESTS)

The Open Network Install Environment (ONIE) is a small operating system which is pre-installed as firmware on bare metal network switches. ONIE enables users to have flexibility in the Network Operating System (NOS) install as part of the data center provisioning process. ONIE was designed to discover Network Operating Systems (NOS) on a network, transfer the NOS to the switch, then install it. Once ONIE has installed a NOS, the switch will boot directly into the NOS and bypass ONIE. ONIE also provides the means to uninstall the NOS and can be accessed even when the NOS is installed.

The Group 1 tests are only applicable if the bare metal switch has ONIE pre-installed. The tests utilize ONIE to install and un-install a NOS which is located on the network.

MANAGEMENT OF OPTICAL MODULE (GROUP 2 TESTS)

It is very common in industry for a switch manufacturer to whitelist the optics and cables that their proprietary switches support. Generally, the whitelisted optic or cable is a standard off-the-shelf product. The only difference is that unique vendor information has been stored in the EEPROM to indicate to the proprietary switch that the device is supported. One goal in Open Networking is to move away from this model entirely, enabling the NOS to recognize and support any module that is installed in the bare metal switch. In order to accomplish this, the EEPROM data needs to be well understood and thus, the group 2 tests prove out that the EEPROM has data programmed according to spec, that the NOS has access to the data, and that the data the NOS reads is the same as when read by a traditional EEPROM reader.

During the initial plugfest in which this test plan was proven out, it was quickly discovered that new methods for obtaining EEPROM data and managing modules was necessary. This need became the spark for the Open Optical Monitoring (OOM) effort within OCP. This software allow for remote management of optical modules from multiple vendors. Future revisions of the Open Networking Interoperability Test Plan will likely include requirements for NOS to support OOM.

PASSIVE CABLE TESTING (GROUP 3 TESTS)

Direct Attach Cables (DAC) are used in network solutions as short, passive channels generally at the Top of Rack (TOR). Over a short distance, DACs are usually less expensive than a fiber solution. However, even when limited to short distances, given the high data rates, these cables must operate properly at range to ensure the proper transmission of data. Group 3 addresses DACs by testing the return loss and insertion loss for both 10G DACs and 40G DACs using SFP+ technology and QSFP+ technology, respectively.

In this section of tests, DACs are characterized for input and output return loss, as well as insertion loss. Return loss examines the signal reflected at the port, while insertion loss examines the signal passed through the cable. A vector network analyzer (VNA) is used to take S-parameter measurements.

The VNA is used in concert with two compliance boards placed on each end of the cable. These boards are ensured to be compliant with either SFF-8431 (for 10GbE DACs) or IEEE Std. 802.3 Clause 85 (for 40GbE DACs) before use in the test setup.

Return loss characteristics for both 10GbE SFP+ DACs and 40GbE QSFP+ DACs are taken by measuring attenuation of reflected signals throughout the frequency band of 10MHz to 11.1GHz, as required by the standards SFF-8431 and IEEE Std. 802.3-2012 Clause 85. The measurements taken generate a .s4p file which is then exported to MATLAB. Next, a graphical analysis of return loss characteristics throughout the 10MHz to 11.1GHz range with respect to the governing equations provided by the standard is produced. Any violations of this governing equation result in a failing DAC, which will be prohibited from addition to the Open Networking Integrator's List.

The insertion loss characteristics of a passive cable can help diagnose issues in data transmission. For 10GbE, SFP+ Direct Attach Cables, there is no value for insertion loss limits included in the SFF-8431 specification, Appendix E. A certain amount of insertion loss in 40GBASE-CR4 cabling is allowed by IEEE Std. 802.3-2012 Clause 85.10 for compliance. Due to the similarities between SFP+ Direct Attach Cables and QSFP+ Direct Attach Cables (40GBASE-CR4), the limits for QSFP+ Direct Attached Cables is used in the testing. This measurement is taken using a similar procedure as described above for return loss. The loss that is measured at the Nyquist rate (5.15625 GHz) must fall between 3dB and 17.04dB as defined by IEEE Std. 802.3-2012 Clause 85.10. Any DAC which does not fall within that range at Nyquist fails this test and will be prohibited from inclusion to the Open Networking Integrator's List.

These tests provide a detailed, comprehensive view into the health and performance of both 10GbE SFP+ DACs and 40GbE QSFP+ DACs. They also allow passing DACs to be included on the Open Networking Integrator's List with a great deal of confidence.

ACTIVE OPTICAL CABLE AND MODULE COMPLIANCE TESTING (GROUP 4 TESTS)

Active Optical Cables and Active Modules are a significant portion of Data Center installs due to increase bandwidth, lower power consumption and longer reach capabilities by these module types. Fiber used with these devices also allows for better turn radii, which allows it to be very versatile to the installer. As expected the signal provided by Active Optical Cables (AOC) and Modules must be of high quality to ensure proper operation of any network.

This group of tests ensures that 10G Ethernet (10GbE) and 40G Ethernet (40GbE) AOCs and Modules are compliant to their respective standards. IEEE Standard 802.3-2012 provides values for 40G Ethernet devices, where SFF-8431 defines many values necessary for 10G Ethernet. The equipment used to test this compliance is carefully calibrated for each test case to provide reliable and repeatable results. In these tests, signal quality for AOCs and Modules is examined at the electrical output, where the device may interface with a host device's receiver. Modules are further scrutinized to ensure that their optical transmitters are compliant with the standard.

In order to ensure consistent, repeatable results, the pattern generator used for the testing is calibrated according to parameters and requirements in SFF-8431, the standard that 10GbE AOCs and Modules are compliant with. The pattern generator is used to emulate a worst-case compliant signal presented by a host device. Proper amounts of emphasis and jitter are introduced to the pattern and are measured using a high bandwidth oscilloscope. The eye diagram is examined to ensure that a hit ratio is no more than 5×10^{-5}. In addition to the eye diagram, values for specific types of jitter, rise and fall times, data dependent pulse width shrinkage (DDPWS) and amplitude are calibrated to meet the requirements for a calibrated signal as defined by SFF-8431. Throughout this process, Module Compliance Boards (MCB) and Host Compliance Boards (HCB) are used. These Compliance Boards are designed to be compliant to return and insertion loss characteristics defined in SFF-8431. A counter-propagating signal is used to provide a source of crosstalk onto the stressed signal being calibrated.

Once the pattern generator has been calibrated, this signal is used to test the optical transmitter of a "golden module" for testing. This module must demonstrate compliance to SFF-8431 at the optical transmitter. This verification is done by inserting a candidate module into an MCB, applying the calibrated input signal and verifying that the optical eye diagram has a hit ratio of no more than than 5×10^{-5}. Once this golden module has been verified, it is used to test the electrical output of a given Module under test.

At this point, all 10GbE calibration has been completed, and testing can begin for either AOCs or Modules.

To test the electrical output for AOCs, an MCB is connected to each end of the cable. PRBS31 is sourced by the calibrated pattern generator using appropriate amounts of emphasis and jitter. This stressed signal is passed through the cable, with a counter-propagating signal providing a crosstalk source. The oscilloscope captures an eye diagram over 1000 waveforms, and the diagram must show a hit ratio of less than than 5×10^{-5} when compared to the eye mask provided by SFF-8431.

The electrical output of the Modules are tested in the same way with one change; the pattern is sourced in the same way, however the golden module and a small fiber patch cable is used to stimulate the optical receiver of the module under test. The electrical output is then examined in the same way AOCs are tested. As with AOC testing, counter-propagating signals provide a source of crosstalk as is required by the standard.

Modules undergo testing beyond what is done for AOCs, since the module-fiber-module link is not a closed system like an AOC and therefore the optical transmitter needs verification. Using the stressed signal provided by the calibrated pattern generator, the resulting waveform is captured 1,000 times to build the optical eye diagram. This eye diagram should never violate the hit ratio when compared to the optical eye mask provided by IEEE 802.3 Clause 52.

40GbE devices undergo identical tests to ensure their compliance with the standards. However, the values required for 40GbE are provided by IEEE Std. 802.3-2012 and the 40GbE signaling is passed through a QSFP+ fixture, rather than an SFP+. Crosstalk is provided on each counter-propagating lane throughout the test procedure. Also, only one lane is tested for compliance; if the device, either an AOC or Module, fails any other test throughout the entire test plan, that device undergoes this Group 4 testing on each lane to further identify possible issues that it may exhibit.

10GbE and 40GbE Modules and AOCs are added to the Open Networking Integrator's List if and only if they pass each test applicable to their technology.

HOST MODULE ELECTRICAL VERIFICATION (GROUP 5 TESTS)

Host devices are an integral part in any network solution. In the case of Open Compute, a host device is comprised of two parts, the bare-metal switch and the NOS. In most cases, these parts are produced by two different vendors. The electrical characteristics of the hosts are reliant on the NOS understanding the hardware that is present and programming chipsets on bare-metal switch properly. The electrical characteristics of these hosts must exhibit compliant behavior in order to be included on the Open Networking Integrator's List.

10GbE and 40GbE switches are tested with a given NOS that is installed to be compliant to either SFF-8431 (for 10GbE) or IEEE Std. 802.3-2012 Clause 86 (for 40GbE). Access to emphasis settings and pattern generation on the device is required to perform the tests included in Group 5.

For 10GbE hosts, signal integrity is examined with respect to rise and fall times, total jitter output, and an overall transmitter eye mask verification of the device. This process provides an abundance of information regarding the health of the host's transmitter. The device is instructed to source PRBS31 through a HCB. A high-bandwidth oscilloscope is set to capture 1,000 sourced waveforms which are measured for compliance. For 40GbE hosts, similar testing is done; however, the test for jitter output is a bit more in depth, including an examination of J2 and J9 values exhibited by the host's transmitter. One lane is tested, and unused lanes are terminated at the output of the HCB.

In addition to these signal quality measurements, input and output return loss measurements are also taken. Using a vector network analyzer (VNA), S-parameter measurements are taken and the results are compared to the device's appropriate standard for the port's technology. For 10GbE ports, measurements are compared to governing equations provided by SFF-8431; for 40GbE ports, measurements are compared to governing equations provided by IEEE Std. 802.3-2012 Clause 86A. These equations provide a maximum limit for return loss values between 10MHz and 11.1GHz. The measurements are taken through a HCB, which exhibits compliant return and insertion loss characteristics from 10MHz to 11.1GHz. Measurements are taken by sweeping across the frequency range on the instrument and then saving the data in a .s4p file. This file is then processed to generate a plot for both input and output return loss. Any violations of the equations provided by the standard result in a failure and prohibit a host from earning a spot on the Open Networking Integrator's List.

It should be noted that most host devices support both 10GbE and 40GbE technologies, with ports dedicated for each speed. In the event of this occurrence, the host must pass Group 5 tests for both technologies to be included in the Open Networking Integrator's List. A failure of any port tested will result in the device not being included.

LINK FUNCTIONALITY (GROUP 6 TESTS)

The first five groups of tests focus on proving conformance to existing standards, with the expectation that the conformance is a solid foundation for interoperability. The Group 6 tests focus on interoperability, using the components tested in the previous sections to build a functioning Open Networking system.

The three tested components are; a bare-metal/whitebox switch, a Network Operating System, and connectivity in the form of AOCs, DACs and Modules. From these three components, a variety of combined systems are built and then subjected to a series of link verification tests in an attempt to prove that the multi-vendor system functions properly.

The initial test (6.1) in this group establishes a baseline performance metric. This baseline assessment is performed using known compliant, anonymous Modules or Cable Assemblies. The test is designed to create an understanding of performance that can translate through the remaining tests in this section. The baseline is established by determining what throughput rate the configuration can handle without dropping packets. This baseline is then used for all subsequent tests on that configuration.

The second interoperability test (6.2) establishes that the configuration has the ability to detect and establish a link under various power on conditions. The following three conditions are checked, as there may be different signals on the line during the boot up sequences of the hosts, such as remote and local fault, that could cause the system to improperly not establish a link.

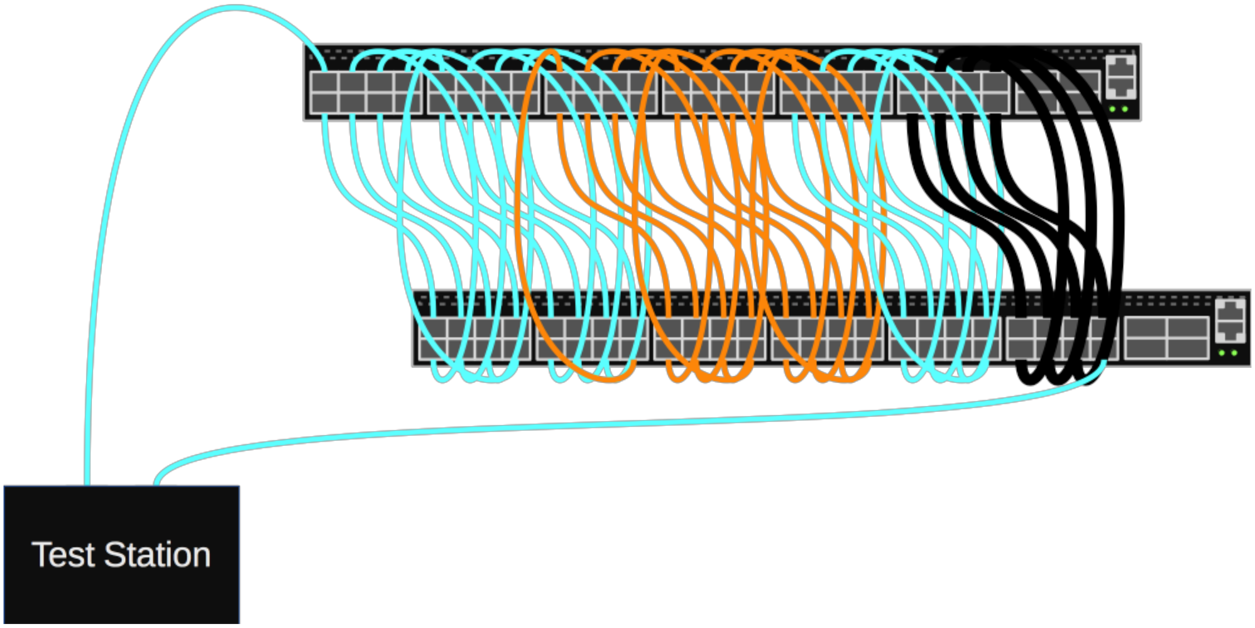
1. The Host is powered and has fully booted, and then receives the Module or Cable Assembly with and without a link partner connected to the Module or Cable Assembly during insertion.
2. The Host is powered on with the Module or Cable Assembly inserted and connected to a link partner that has not yet powered on.
3. The Host is powered on with the Module or Cable Assembly inserted and connected to a link partner that has already been powered on.

In each case, packets are transmitted through the link in order to verify the link is properly established.

The next test (6.3) is designed to verify the ability of a Host to exchange packets with a Module or Cable Assembly. The exchange of packets must perform such that the bit error rate, as specified in the IEEE Std 802.3ae-2012, is 10^{-12}. In order to accomplish this, two Modules or Cable Assemblies are inserted into Host, and then the other ends of their connections are connected to an Ixia Packet generator. A learned conversation is established between the two Ixia ports and then $2.47*10^{8}$ 1518-byte frames are transmitted through the device. This applies to all Module and Cable Assemblies types. If more than 7 packets are lost during the exchange, then the BER criteria has not been met and the test fails. In addition to packets lost, local management information may make it possible to isolate the packet loss to either the transmit side or the receive side of the test channel relative to the Host and Module/Cable Assembly. If more than 7 packets are lost in either side of the channel, then the Host, NOS and Module/Cable Assembly combination has violated the BER and the result is considered a failure.

In addition to this bit error rate, packets are exchanged through the link for an hour for longevity testing. Again, if the expected BER is not maintained through this hour long test, then the Host, NOS and Module/Cable Assembly combination will fail this test.

Finally, a stress test (6.4) is performed. In this test, traffic is cycled through a 'snake' of modules and cables that passes through every port on two switches. See diagram below.



This test ensures the system will operate properly when both switches are fully populated with modules, AOCs, or DACs. Again, a desired BER is verified.

CONCLUSION

As with any new technology, Open Networking opens new doors for flexibility and customization that were simply not possible before, by providing freedom from vendor lock-in. For those opportunities to bear fruit, the ecosystem needs to be in place to ensure interoperability. The Open Networking Integrators List program, and the testing outlined in this whitepaper are key components enabling whitebox switch manufacturers, NOS vendors, and module and cable suppliers to publically prove that systems built from their products are interoperable and reliable.

Further information on the Open Networking Integrators List and the Open Network Systems Interoperability Test Plan can be found below.

OPEN NETWORKING INTEGRATORS LIST

https://www.iol.unh.edu/registry/opennetworking

OPEN NETWORK SYSTEMS INTEROPERABILITY TEST PLAN

https://www.iol.unh.edu/testing/open-networking/ocp/test-suites

ABOUT THE UNH-IOL

Founded in 1988, the UNH-IOL provides independent, broad-based interoperability and standards conformance testing for data, telecommunications and storage networking products and technologies. Combining extensive staff experience and standards-bodies participation, the UNH-IOL helps companies efficiently and cost effectively deliver products to the market. For more information, visit iol.unh.edu.

The UNH-IOL hosts multi-vendor group tests (often called "plugfests") as often as four times a month. These group test events compliment over 20 year-round standards-based testing programs that are managed and operated by the UNH-IOL. Each of the testing groups, called "consortiums", represents a collaboration of industry forums, service providers, test equipment vendors and otherwise competing companies who benefit each other by:

- Distributing the cost of testing
- Lowering R&D and QA expenses
- Reducing product time to market
- Obtaining trusted vendor-neutral verification

The laboratory maintains a strong reputation for independent, vendor-neutral testing with a focus on quality assurance. The confidential test reports the UNH-IOL provides to its members are recognized throughout the data communications industry as evidence of interoperability and conformance to technical standards.