

White Paper

QSFP-DD Module Testing

400G and QSFP-DD

QSFP-DD optical modules are the mainstream form factor for 400G client interfaces. This white paper shares the key factors in successful test, troubleshooting and validation of QSFP-DD modules for module developers, network element manufactures and end users.

Client interface speeds have seen a steady increase with typical rate increases of at least tenfold every decade. We now see widespread deployment of 100GE via QSFP28 interfaces and we are at the early stages of 400G deployment. The IEEE¹ has developed the 400G Ethernet client interface standard as part of 802.3.bs which was formally standardized in December 2017. Early adopters used the CFP8² form-factor but the broader market is focused on adopting the QSFP-DD³ which allows a degree of backwards compatibility with the widely adopted QSFP28.

Since Ethernet has a broad range of applications and reaches a range of PMD (physical media dependent) choices – these allow one 'QSFP-DD' slot to support a huge range of applications and reaches from a few meters with a passive copper DAC cable through to 80 km coherent based ZR. There are also a smaller number of companies who are focusing on OSFP⁽⁴⁾ form factor. Although not as widespread and backwards compatible, it does offer some advantages in terms of electrical signal integrity and thermal management. Much of what is discussed below on the QSFP-DD is applicable to OSFP and the VIAVI ONT family which supports many OSFP based applications.⁴

400G relies on higher order (PAM-4) modulation for both the electrical module to host interfaces and the electrical or optical PMD. The PAM-4 modulation was adopted to maximize the data capacity for a given bandwidth, but it drives significant challenges in complexity and performance which also means the link requires forward error correcting (FEC) coding to allow reliable data transmission.

1. <http://www.ieee802.org/3/bs/>

2. <http://www.cfp-msa.org/documents.html>

3. <http://www.qsfp-dd.com/>

4. <https://osfpmsa.org/>

Why QSFP-DD?

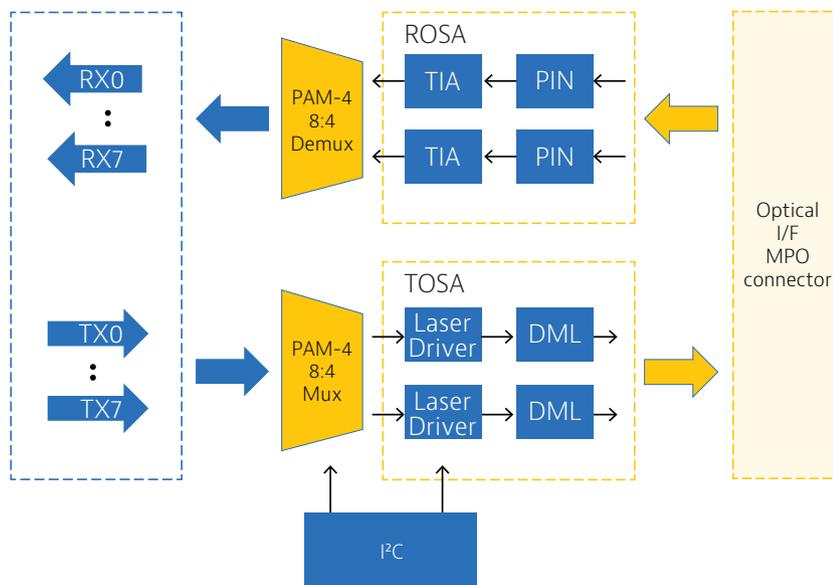
100G Ethernet started deployment in 2008 with early designs based on CFP pluggable modules. Second generation systems moved to CFP2 (or CPAK for one major equipment manufacturer) before settling on QSFP28 which drove widespread and cost-effective volume adoption. CFP4 was a slightly earlier challenge to QSFP28 but for multiple factors QSFP28 drove a huge ramp up of 100G. The industry was mindful of the 'form-factor' wars and wanted to minimize the additional complexity and cost challenges of multi-step form factor evolution at 400G. CFP8 allowed very early adopters to develop and validate 400G. However it did not meet the density, power, cost and 'compatibility' needs so the industry quickly focused on QSFP-DD as the target form factor. An alternative – the OSFP – was proposed. It offered a superior technical solution but could not meet the compelling need for legacy module support. In principle a QSFP-DD socket could support legacy QSFP-28 optics – this would allow vendors to ship '400G ready' network elements that could be shipped with 100G plugs today and the field-upgrade would be a simple module swap-out.

In order to meet the increased bandwidth, power and cooling needs driven by the step to 400G several enhancements were made to the existing QSFP28 concept. These included the doubling of high speed electrical lanes (from 4 lanes of NRZ at 25 Gbps to 8 lanes of PAM-4 at 56 Gbps) and a lengthening of the module 'nose' to give a greater internal volume and enhanced thermal performance. Further work has also been done to enhance the module control interface giving rise to the CMIS 4.0⁵ standard.

400G DR4 Module, 500 m, 4 Parallel SMF Wavelength range: 1304.5 to 1317.5 nm

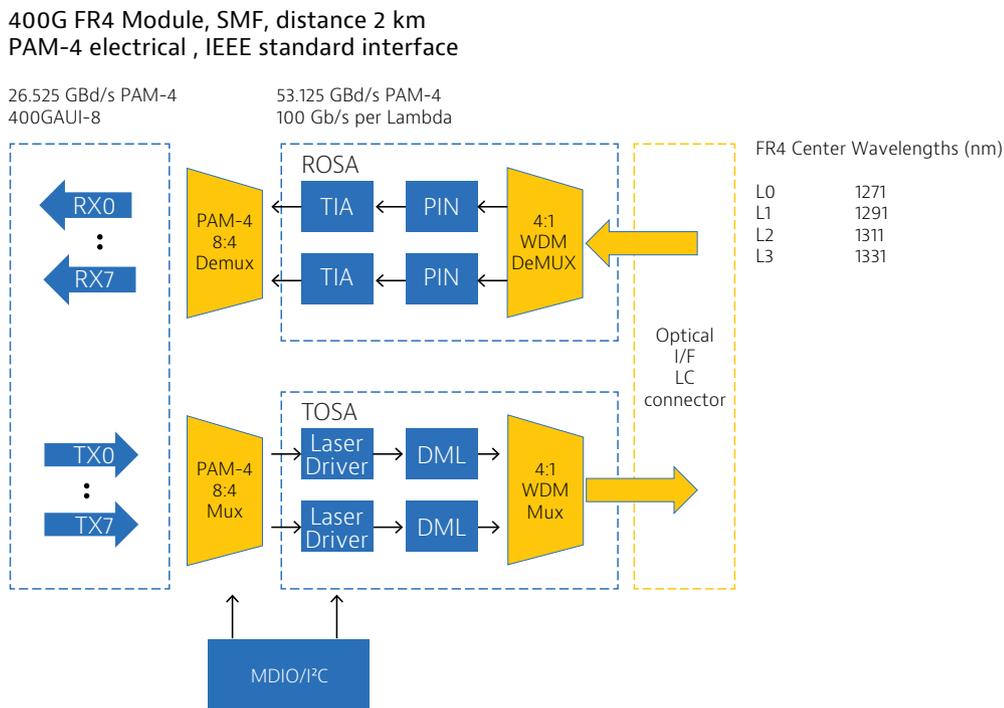
26.5625 GBd/s PAM-4

53.125 GBd/s PAM-4
100 Gb/s per Fiber

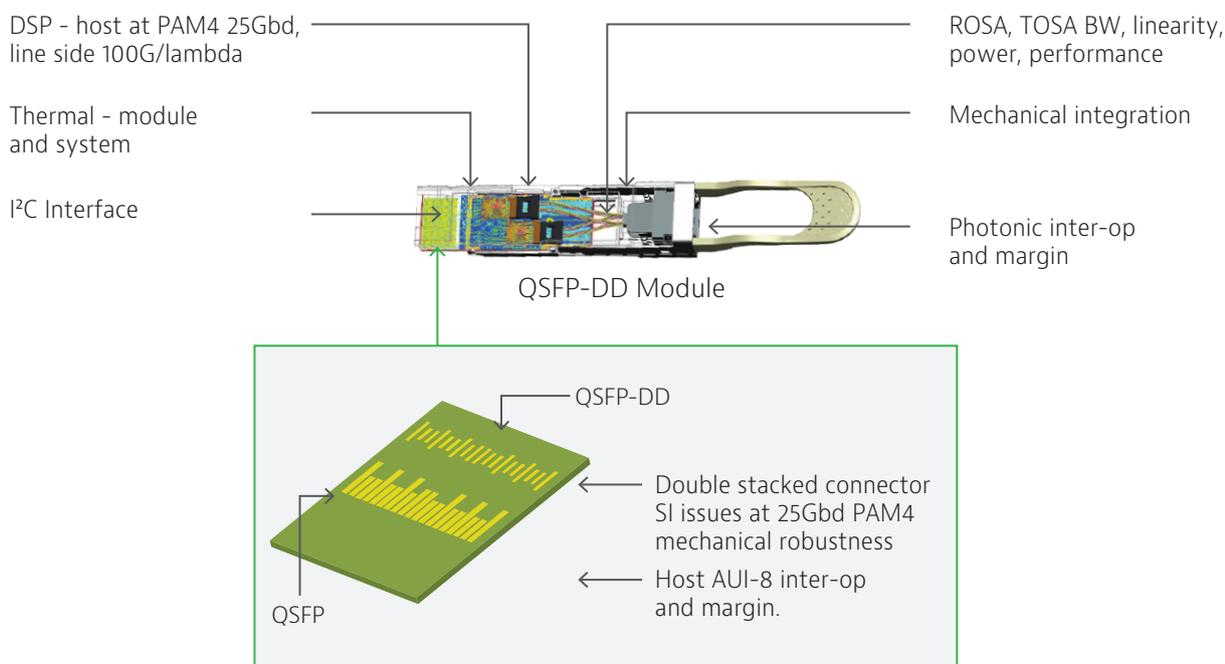


5. <http://www.qsfp-dd.com/qsfp-dd-msa-group-announces-updated-specificatio/>

DR4 will be one of the most common 400G client optical interfaces deployed in 2020. It carries 400G as four 100G signals on individual single mode fibers. It has widespread applications in enterprise. It supports 500 m reach and its ability to break out to individual 100G Ethernet links is attractive as a high density 100G solution, this can quadruple port count density.



The FR4 interface will also have widespread applications, including with telcos. It offers a longer link budget of 2 km over one single mode fiber. The 400G is carried on four 100G signals, each on a slightly different wavelength.



400G PMD Modules (Physical Medium Dependent)

PMD	Reach	Application	Technology
DAC	2 to 3 m	Intra-rack & server	Passive copper cable, 50G PAM-4 electrical
SR8	100 m	Enterprise	Parallel multi-mode, 50G/λ – PAM-4
DR4	500 m	Datacenter and enterprise	Parallel single-mode, 100G/λ – PAM-4
FR4	2 km	Large scale datacenter	Single-mode, 100G/λ, PAM-4
LR8	10 km	Telecom reach	Single-mode, 100G/λ, PAM-4
ZR	80 km	Metro and DCI	Single-mode/coherent, PAM-4

QSFP-DD module – standards and themes

Many standards and MSAs are applicable as we have seen from references above. It is also important to understand what the critical tests are at each stage of the development cycle – from basic IC evaluations through module H/W integration, S/W and firmware, to vendor selection and qualification. Production also has its own set of key test requirements.

A solid understanding of key documents including IEEE, CMIS, QSFP-DD, MSA and OIF are required to successfully design, test, validate, manufacture and deploy pluggable optics. The QSFP-DD is a magnificent integrated combination of electronics, optics, mechanics, thermal management and firmware. All must work together before modules can be successfully deployed.

Inter-operability

The great beauty of the Ethernet client interface ecosystem is that we have a set of robust and clear standards driven by IEEE and others that allow a multi-vendor ecosystem to inter-operate without resorting to 'engineered' links.

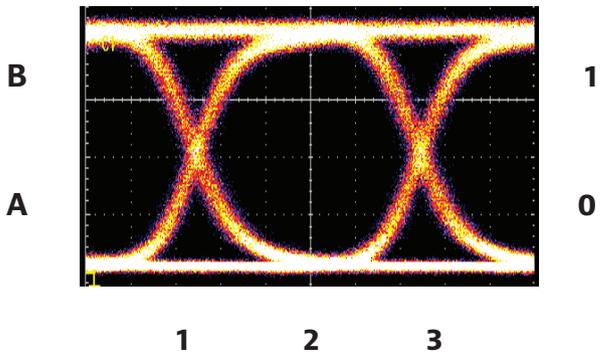
Both the module to host and optical interface are key to this inter-working. On the host to module interface we are primarily concerned with three areas:

- The high-speed data path (AUI) – built from chip to module (C2M) – has multiple challenges including signal integrity and signal equalization. Although a portion of the FEC budget is allocated for this part of the link, any issues with this interface can cause significant issues with the link. A badly 'tuned' link (in terms of equalizers and channel) can cause difficult to troubleshoot issues such as random bursts or the worst case of occasional bit slips.
- Module management– this I²C based interface has evolved from a basic memory mapped management of the SFF-8636 through to QSFP28 at 100G to a sophisticated and stateful CMIS 4.0. This evolution has been extremely challenging for the ecosystem and a solid working knowledge of the CMIS 4.0 documentation is key to robust and stable module management.
- Module power--The power demands of modules have crept from a few watts at 100G to potentially close to 20W for pluggable coherent (QSFP-DD ZR) modules for DCI applications. This places high demands on power supply robustness and stability. Furthermore, it must be able to supply the dynamic and transient nature of the power demands as modules wake up.

These areas are all closely intertwined and need to be treated as a whole (especially with regard to CMIS 4.0 module management) to ensure trouble free module operation.

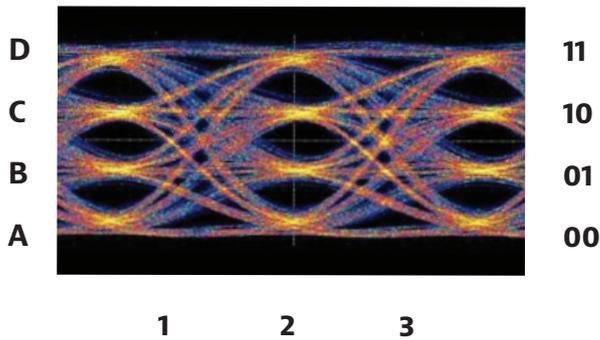
PAM-4

PAM-4 modulation was adopted for both electrical (module to host interface) and optical (electrical) links. This higher order modulation scheme allows a doubling of bits sent in a unit time. While NRZ technology is widespread and mature for high-speed, SERDES PAM-4 is a relatively new technology and is more complex and challenging. We have a wealth of experience in error analysis of NRZ links. But we still saw issues with the move from 10G to 25G NRZ lanes used at 100GE. So the move to PAM-4 is expected to be a significant and industry-wide challenge. This is compounded by the use of FEC based links (always a background error rate) and far more complex channel equalization. It would be fair to say PAM-4 is an order of magnitude more complex than the widespread 25G NRZ.



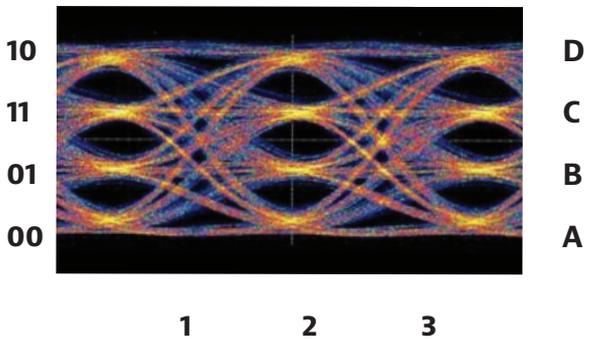
NRZ Modulation:

- ▶ 1 bit per clock cycle:
 - Voltage A = „0“
 - Voltage B = „1“



PAM-4 Modulation, linear (non-Gray) coding:

- ▶ 2 bits per clock cycle
 - Wrong Decision between B and C -> 2 errored bits!



PAM-4 Modulation, Gray coding:

- ▶ 2 bits per clock cycle
 - Wrong decision between B and C -> only 1 errored bit

Image showing NRZ and PAM-4 signaling

FEC

Since it would be extremely challenging to develop components that could offer error free PAM-4 transmission, the developers used a FEC which protects both the electrical module interface as well as the optical module to module interface. Great care was given to understand the error mechanisms in both the transmission channels and components, while setting balance in the 'cost' of the FEC logic (on both encoding and receiving) side. FEC 'costs' include additional circuitry which draws power and adds latency to any link.

DSP & Equalizers

It was decided in the adoption of 400G to use the concept of a 'powerful' electrical receiver equalizer to mitigate the performance of a 'worst case' transmitter coupled with a 'worst case' channel. This could lead to a closed PAM-4 eye at the input of a PAM-4 receiver, so the PAM-4 receiver needs a powerful and potentially complex receiver to equalize the TX and channel impact so a clear eye can be recovered to achieve correct decoding of a given symbol. The equalizer complexity means that in most cases a DSP based solution must be implemented, which can have an impact on power, latency, complexity, error performance and management/control. Although a DSP equalizer is powerful, the complexity of features can cause challenges in finding the best settings of taps etc. Furthermore, the equalizer is often hidden behind DSP firmware and a control API giving a high level of abstraction from the user. Further challenges occur with the measurement of TDECQ⁶ – this measurement is complex and can be inconsistent, which further adds to the challenges of a freely inter-operating, multi-vendor ecosystem.

Key points

There will always be errors – links now will always have a background error rate. The error statistics 'fingerprint' is critical. A true random error flow would be generally compatible with the FEC used to protect the link. But bursts, slips and other deterministic issues may seriously degrade the FEC error correction capability. In real links the errors could be a complex mix of electrical and optical channel noise, crosstalk, signal integrity issues, bursts, bit-slips and even error multiplication in incorrectly set equalizers.

Ultimately what matters is how the FEC performs when given a particular error fingerprint. What is the margin? How long before we get a dropped packet? Can we predict the long term performance to see link degradation? What is the error root cause?

Several tools can be used to investigate the error fingerprint, from error bias in individual PAM-4 symbols to analysis of the burst or bit-slip nature. Understanding error bias can be further enhanced by tools such as clock variation and skew.

The PAM-4 symbol analysis can be used to make sure that there is no 'level' bias in the error distribution. The stability of key photonic elements, such as the receiver photonic AGC, can be further validated by variation of optical power (via an attenuator) while observing PAM-4 error distribution.

It is important to fully investigate error bursts and validate that they are bursts rather than bit (or symbol) slips. Slips are often related to the DSP (and related firmware) and would not be correctable by the FEC. Normal test sets cannot differentiate between bursts that are caused by classical signal integrity or noise issues, and bursts that are related to clock and phase sensitivity. So, while investigating the nature and root causes of errors with QSFP-DD a host of new tools and techniques must be deployed.

6. <https://ieeexplore.ieee.org/document/7937468>

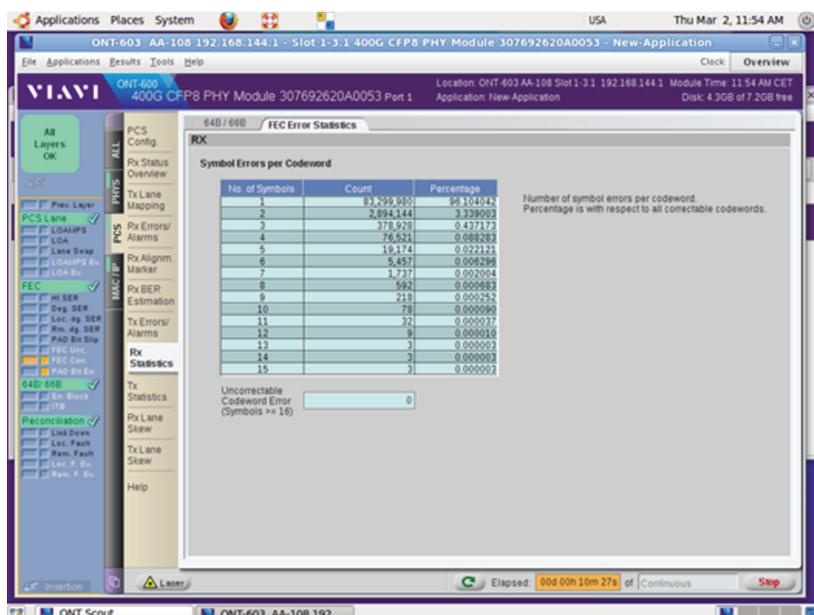
The simplest top-level view can be obtained by looking at the number of errored 10 bit symbols per 5440 bit FEC codeword (KP4 FEC). We would normally expect a monotonic distribution dropping approximately a decade per symbol count. That is, for each increase in errored symbol/codeword we expect the error count to drop a decade. Any long tail or isolated peaks suggest some non-random (system) cause. We also expect the number of symbols errored to increase x10 in measurement time. So, if we observed a count in 10 errored symbols per codeword after 10 seconds we would expect to see counts in the 11 errored symbol column after ~100 seconds.

Such a rule of thumb can be used to estimate the time to an uncorrectable error (occurring at 16 or more errored symbols per codeword). For example, after 100 hours of test time if we observed a maximum of 12 errored symbols/codeword we could then expect the following approximation:

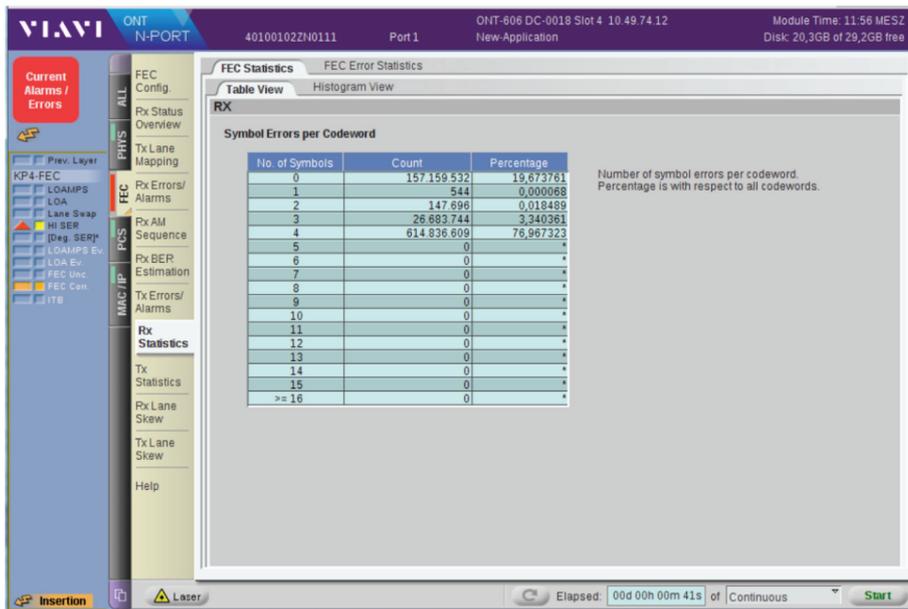
Errored symbols	Time	Notes
12	100 hours	Measurement
13	1000 hours	Estimate
14	~420 days	
15	~11 ½ years	
16 (uncorrectable error)	~114 years	First dropped packet after > century

FEC – errored symbols/codeword

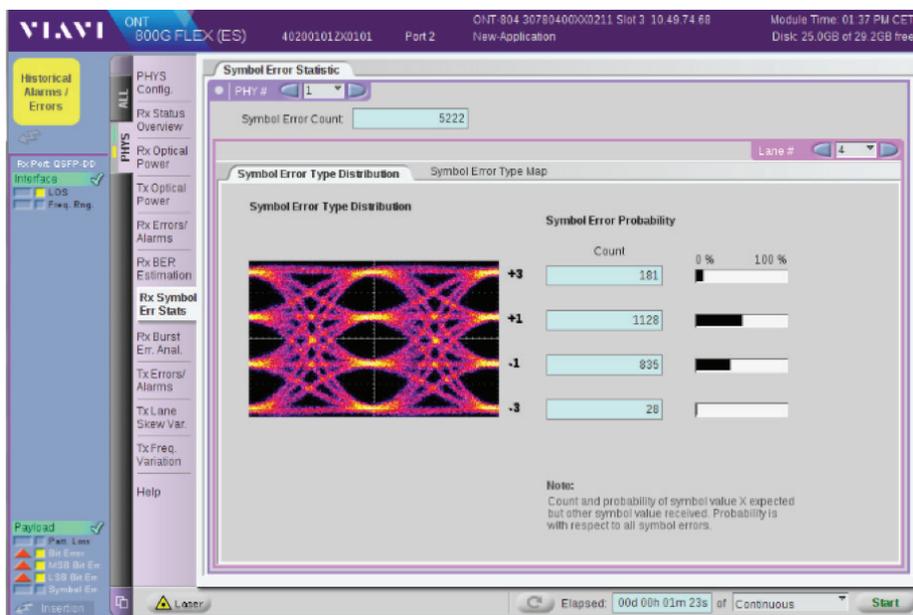
In the case below, an ONT was left running with a 400G optical link that was heavily attenuated so that over a 10 min interval significant errors occurred. Which is about what would be expected with a compliant link. As you can see, the distribution is generally monotonic. The count per errored symbols drops, but it does display a slightly longer tail from 12 errored symbols/codeword. In this case the link is extremely likely to drop a packet due to an uncorrected codeword.



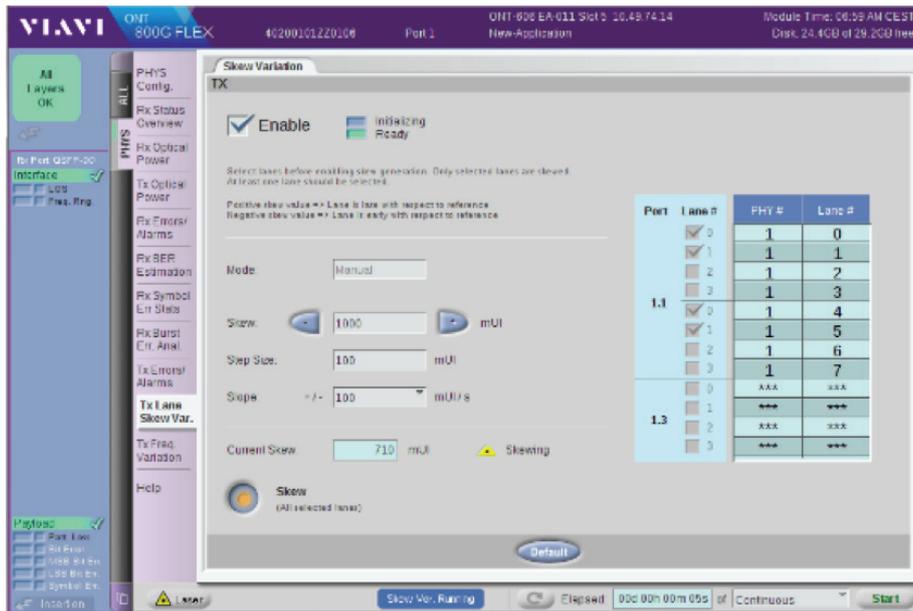
The screen shot below shows the case where a serious issue is occurring. Although the FEC has significant margin (we see a maximum of 4 error symbols in a codeword) the distribution is not monotonic and shows that this system has an underlying error source at work. Note that this example for a 100G link was generated by a special VIAVI ONT application which can create a wide range of FEC error distributions to stress and validate FEC logic and power integrity.



The ONT has the ability to both analyze error distributions and patterns across a whole sequence and track the error profile on a per PAM-4 symbol basis.



Dynamic skew variation is a power tool to stress and validate QSFP-DD modules. It can be used to validate compliance to IEEE 802.3 as well as general stability of the DSP and associated firmware. This is especially important in DR4 modules where individual electrical and optical lane pairs could be on completely different clock domains!



The above screen shot shows the dynamic skew application for PAM-4. The ability to precisely control the relative timing of transmitted lanes to a fraction of a UI and still retain 'hitless' phase shifting is key to resolving challenging problems including crosstalk and DSP based firmware timing issues.

Dynamic skew (or skew variation) is a critical test for any parallel lane communication system. It has application in signal integrity testing and validation (crosstalk) and can also be used to stress and validate the performance of the FIFO & CDR inside the PAM-4 SERDES.

Varying rates of skew can also be used to investigate issues with signal integrity and crosstalk, which has wide applications with H/W and SI teams. The lane timing can be adjusted to ensure that the aggressor lane transitions occur in the middle of the victim lane's PAM-4 eye.

PAM-4 signaling (because of the lower signal margin) is far more susceptible to crosstalk than classic NRZ. In the closely packed confines of a QSFP-DD (especially around the host connector) high speed PAM-4 lanes are routed in close proximity and unless care is taken issues may occur with signal crosstalk. Normally BER test sets run the parallel lanes with a fixed phase so it may be that 'worst case' alignment does not occur for SI stressing. With dynamic skew the aggressor lane can be swept in relative phase to fully validate that issues do not occur, even under worst case phase shift. The end user simply needs to observe if errors occur at a particular phase offset (typically when the aggressor lane has a level transition in the middle of the victim 'eye').

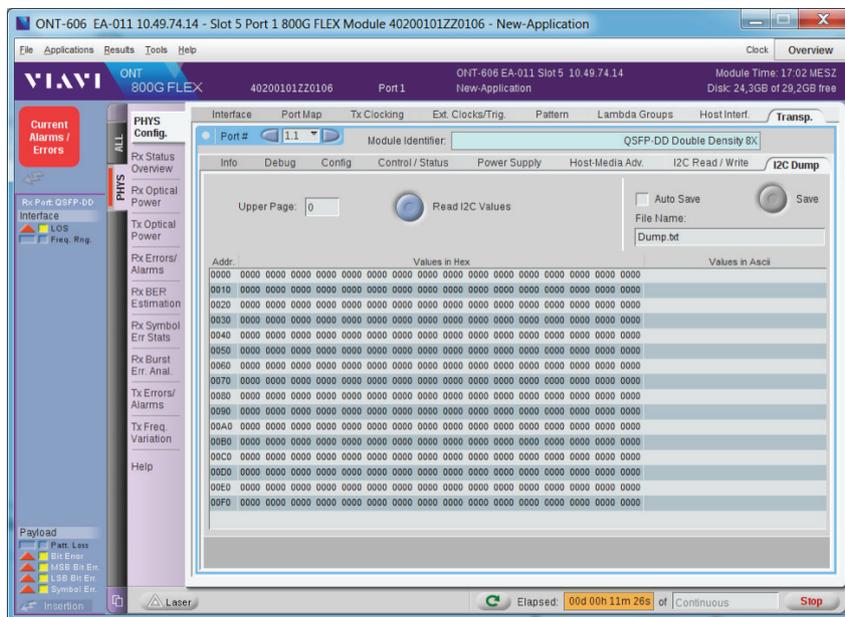
Modern SERDES use a range of FIFO buffers to re-time and re-align signals before further processing within the IC fabric. The realignment uses a series of FIFO buffers that are re-clocked from a master clock source (often a master lane via a CDR).

If the system is not correctly designed or implemented it can be that phase variation and change between the master (CDR reference lane) and other lanes causes misalignment or even a slip in the FIFO. This would manifest itself as a bit slip, which the ONT advanced error analysis could track as a bit slip rather than an error burst as it would be seen by conventional test equipment. With the dynamic skew application, the ONT can deliberately stress the performance of the CDR/FIFO in the SERDES and try to force failure by skew (range and rate). This, combined with the ONT advanced error analysis, makes for a very powerful and complete test system for SERDES test AND can be used to quickly resolve the very challenging issues which cause occasional bit slips in 400GE links. The ONT PAM-4 dynamic skew can force these errors to help diagnosis and root cause resolution.

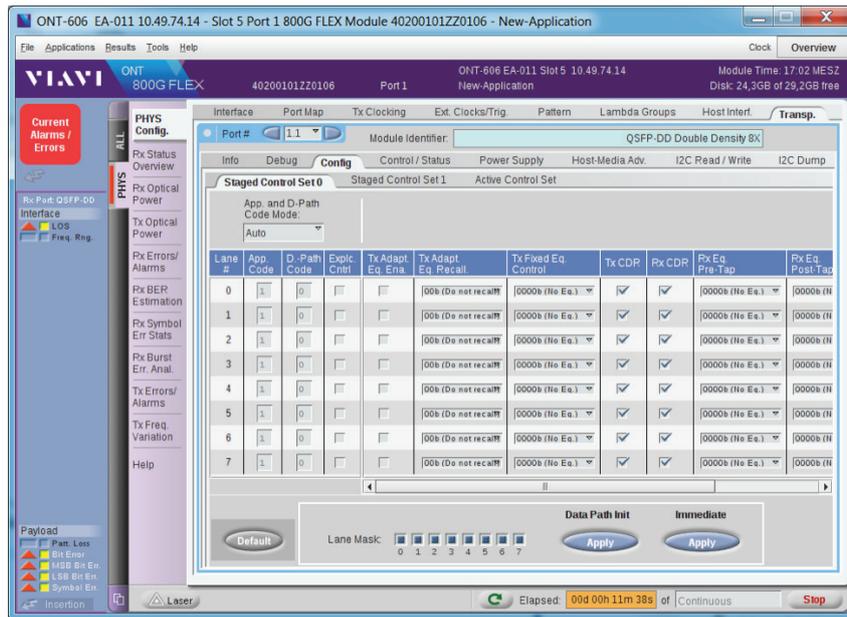
General QSFP-DD control screen

Module management has evolved over time from a very basic register-based system SFF 8636 to CMIS 4.0, which is a comprehensive and stateful module management system designed for the needs of more complex modules at 400GE and above.

The tight interaction between the module via the I²C control interface, the power and control pins and the data path is critical for robust and stable module operation. The greater complexity of the module, especially around the data path equalization in the module DSP, requires a far more comprehensive insight into the control set up and execution between the host and module. The correct order of commands, operations and the slot behavior must be closely orchestrated under CMIS 4.0. Without care a module may seem to operate in one host slot without issue but another (with subtle differences in timing around the commands, power and data path) may operate erratically. Or worse still, have an increased error rate with rare and difficult to resolve issues, likely bit slips. Tools such as the ONT which integrate the CMIS commands over the I²C, as well as the module power control and data path state, are very helpful, not only to debug and resolve issues, but also to stress and validate a module's robustness in different hosts.



The above screen shows the memory dump of the first page of memory. This can quickly check if the correct values have been stored in the QSFP-DD EEPROM. Blank or random data may indicate a device has not yet been initialized.



Some of the more advanced applications in the module management application allow precise and exact control of the module electrical parameters in a clear and unambiguous manner.

Summary

The QSFP-DD module is a marvel of electronic, photonic, mechanical and thermal engineering held together with complex firmware. A healthy multi-vendor QSFP-DD ecosystem is critical to the widespread deployment of the 400G network technologies. It represents both an evolution and revolution in technology over legacy 100G modules, with new challenges posed by PAM-4 signaling, both electrical and optical, the use of FEC for link error control and the new complexity of CMIS 4.0

What makes these challenges greater is the baked in price expectations driven by the scale and deployment needs of hyperscale users. Production must meet the volume and throughput to meet price expectations and yet have the coverage and analysis to meet the new challenges of PAM-4.

The VIAVI ONT family has over two decades of module validation and test applications in its DNA. It quickly established itself as the reference for development, validation and deployment of 400G optics through the pedigree of classic 100G applications like advanced error analysis and dynamic skew coupled with recent VIAVI innovations such as CMIS 4.0 debug and PAM-4 symbol analysis.

When seeking complete coverage for all the challenges of 400G QSFP-DD, whether the needs of classic client interfaces or the emerging novel coherent interfaces, the ONT has the right applications.

PAM-4 is still an emerging signaling for both 50Gbps channels, and for single-lambda 100G Ethernet. Use of the mature and low cost NRZ for 10Gbps and 25Gbps channels certainly will not disappear overnight. But the maturation of PAM-4 technology (analog and digital instantiations) has placed the new signaling method at the forefront of the newest higher-speed Ethernet implementations. In fact, the Institute of Electrical and Electronics Engineers (IEEE) has approved PAM-4 as the preferred signaling for all 50Gbit, 100Gbit, 200Gbit and 400Gbit Ethernet standards within the umbrella 802.3bs, 802.3cd and 802.3ck standards. Additionally, MSAs such as the 100G per Lambda group are organizing to fill out PAM-4 signaling for the entire data center ecosystem.



Contact Us **+1 844 GO VIAVI**
(+1 844 468 4284)

To reach the VIAVI office nearest you,
visit [viavisolutions.com/contact](https://www.viavisolutions.com/contact)

© 2020 VIAVI Solutions Inc.
Product specifications and descriptions in this
document are subject to change without notice.
qsfp-dd-moduletesting-wp-opt-nse-ae
30191044 901 0420