Study Of 400GbE QSFP-DD Transceiver Module For OFC.

Mithuna kumar BM
2nd Sem, M Tech,
VLSI & ES, Electronics & Communication Dept,
East Point College of Engineering & Technology Bengaluru, India
mithunabm@gmail.com

Dr. Anita R
Assoc.Prof,
Electronics & Communication Dept,
East Point College of Engineering &Technology
anitar.ece@eastpoint.ac.in

Dr. Yogesh G S
HOD and Prof,
Electronics & Communication Dept,,
East Point College of Engineering &Technology
hodece.epcet@eastpoint.ac.in

***Abstract*— This Paper describes the concept and technology evolution deployed for optical interconnects and the trade-offs in the design of low complexity, low power DSP and implementation for direct detect and coherent, 400G pluggable optical modules for hyperscale datacenter, enterprise and telecom customers. The design trade-offs include the choice of modulation format, baud rate, optical link design, forward error correction, signal shaping and dispersion compensation.**

400 GbE implementations will be affected by cabling, new form factors and system architecture requirements. The session will explore some of the technology implications, challenges and requirements as well as provide deployment insights to help with the adoption of 100 GbE to 400 GbE in current and future architectures. The session in targeted to everyone interested in learning more about this technology space.

Keywords—400G, OFC(Optical Fiber Communication), QSFP-DD(Quad Small Form-factor Pluggable (QSFP) – Double Density (DD), Ethernet;

I. INTRODUCTION

The use of digital signal processing (DSP) in optical links has about a 20-year history [1]. To use the powerful DSP techniques, the analog optical signal had to be digitized first. This required a low power, high-speed ADC (analog to digital converter) at the receiver. This, together with the development of a high-speed DAC (digital to analog converter) at the transmitter for signal conditioning, enabled the growth of the modern-day high-speed optical interconnects. The first use of DSP was in MLSE (maximum likelihood sequence estimation) implementation for chromatic dispersion compensation in intensity modulated direct detect (IMDD) 10 Gbit/s systems [1]. This was not very power efficient and provided only limited mitigation of chromatic dispersion for long-haul applications; hence MLSE was not deployed widely in 10 Gbit/s IMDD systems. The development of high-speed coherent receivers enabled the linear detection of both signal amplitude and phase. The use of DSP in coherent applications to compensate for a variety of linear optical impairments, as well as enabling QAM (Quadrature Amplitude Modulation), was a very powerful innovation which led to the widespread

deployment of DSP hardware in optical links [1]. In 2015, Inphi was the first to develop and commercialize a PAM4 (4 level Pulse Amplitude Modulation) DSP for direct detect (DD), 100G, intra data center (DC) applications [3]. Based on this DSP, and a highly integrated silicon photonics platform, we successfully developed and deployed an inter data centre (DCI), 100G switch pluggable module for 80 km reach [3]. Recently, Inphi and Broadcom Developed a low power coherent DSP, in the 7nm CMOS node, and successfully demonstrated a silicon photonics based 400ZR coherent pluggable module for DCI applications [3,4].

With the current trends, optics cost is becoming a barrier to transition: the reduction of ASIC cost-per-bit is outpacing that of optics. In this blog we will make the case for using QSFP-DD (Quad Small Form Factor – Double Density) modules as the ideal connectivity option for all next

generation cloud and other hyperscale data centres. II.

ETHERNET ROADMAP FOR 400GBE DEVELOPMENT

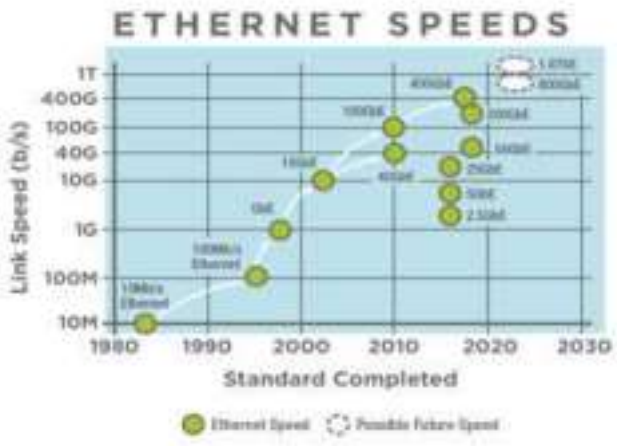

Figure 1: Ethernet road map

Giga Ethernet or GbE is Ethernet with speeds above 100 Gbit/s. 400 Gigabit Ethernet (400G, 400GbE) and 200 Gigabit Ethernet (200G, 200GbE)[1] standards developed by the IEEE P802.3bs Task Force using broadly similar technology to 100 Gigabit Ethernet were approved on December 6, 2017.[4] In 2016, several networking equipment suppliers were already offering proprietary solutions for 200G and 400G.

In the current development history we have

1. 6 new speeds in 1st 35 years of Ethernet
2. 6 new speeds in a 2 year span
   - 2.5 GbE – 2016
   - 5 GbE – 2016
   - 25 GbE – 2016
   - 50 GbE – late 2018
   - 200 GbE – 2017
   - 400 GbE – 2017

According to Pars Mukish, Business Unit Manager, Solid State Lighting (SSL) & Display at Yole: *"The state of the art of fiber-optic communication technologies has advanced dramatically over the past 25 years. The highest capacity of commercial fiber-optic links available in the 1990s was only 2.5-10 Gb/s while today they can carry up to 800 Gb/s. The last decade of developments have enabled higher efficiency digital communication systems and solved problems with degraded signals".*

III. EVOLUTION OF TRANSPORT TECHNOLOGY

Fig. 1 shows a bifurcated technology and product space for the 100G data center interconnect deployment. Intra (inside) data center (DC) optical interconnects were of the NRZ (NonReturn to Zero) format, based on analog CDR (Clock and Data Recovery). The intra DC deployment of 100G, in volume, started in 2015 [4]. Inter (between) DC interconnects, for up to 100 km, were a combination of IMDD PAM4 pluggable modules, which we introduced in 2017 [4], and a variety of coherent formats and

rates in compact modular boxes, from different of vendors [4]. Specialized hardware and lower complexity coherent formats, like 8QAM and QPSK, are commercially deployed for longer reaches.

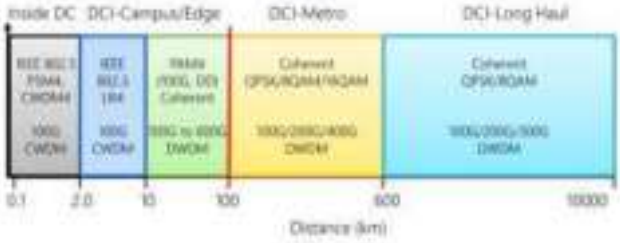

Figure 2: Reach and transmission formats for a variety of DC links at the time of 100G deployment. These are not strict demarcations. These have also evolved over time and vary between service providers and fiber plants

As shown in Fig 2, DSP-based optical interconnects are being deployed at 200G/400G intra DC applications. This transition was in tandem with the evolution of the port level data rates in the DC switches from 25G NRZ to 50/100G PAM4 (25/50 Gbaud) [4]. At 400G, for inter DC applications, there was an opportunity for a common coherent pluggable module hardware to span a variety of distances as shown in Fig. 2. We designed a multi-format, multi-rate, wide-reach DSP ASIC, with a common silicon photonics optical subsystem to accomplish this, without much compromise in performance or power consumption.

Figure 3: Evolution of the DCI reach and technology configuration for the 200G/400G deployments. The color grading shows the DSP complexity for the various applications and reach (with green being the least complex and red being the most).

The commonly used pluggable module form factors, for 100G and above, are shown in Fig. 3. The QSFP-28 (Quad Small Form-Factor Pluggable, with 4 input electrical lanes) 100G PAM4 and QSFP-DD 2 (QSFP-Double Density, which has 8 input electrical lanes, and 2 in the name refers to the longer version of the QSFP-DD form factor) 400ZR coherent DCI modules, have the same width, hence the faceplate density in a switch, at 4x the data rate. The OSFP (Octal Small Factor Pluggable) module is also used for 400G inter and intra DC applications.

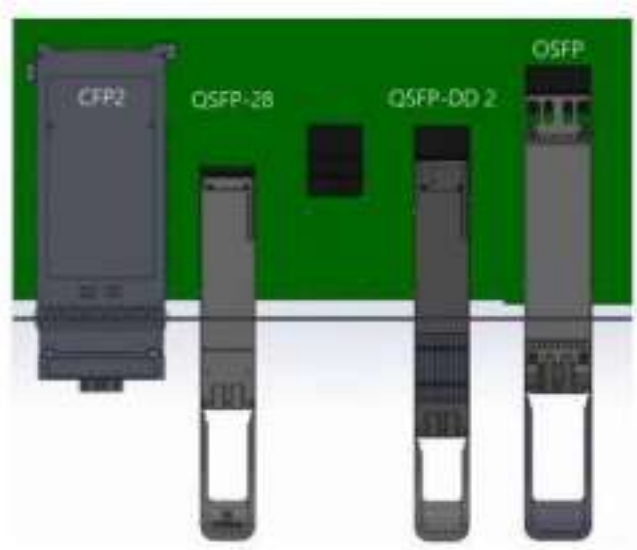

Figure 4: Commonly used pluggable module form factors for data center optical interconnects.

IV. Optical Module Technology

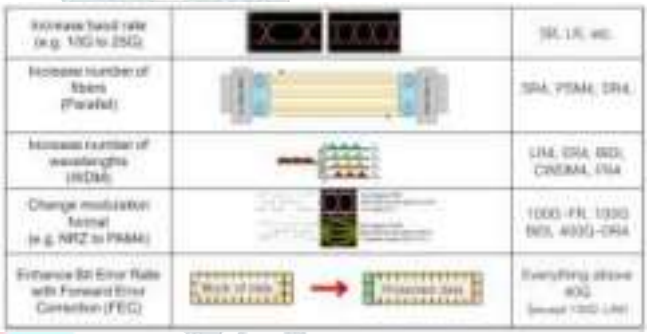

Figure 5: Different type of Optical Module types

An optical transceiver is a small yet powerful device that can both transmit and receive data. In fiber optics, this data is sent in the form of pulses of light over an optical fiber, at very high speeds and across long distances. The transceiver is an important part of a fiber optics network and is used to convert electrical signals to optical (light) signals and optical signals to electrical signals. It can be plugged into or embedded into another device within a data network that can send and receive a signal

V. Low Complexity DSP Design Methodology

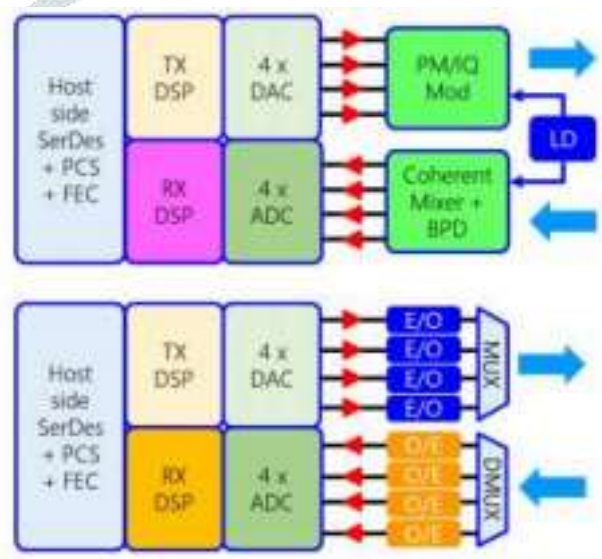

Figure 6: Similarities between a 16QAM coherent interconnect for inter DC and DR4/FR4 PAM4 interconnect for intra DC applications.

As the starting point for reducing the DSP complexity, we have compared the block diagrams for the implementation of the PAM4 and 16QAM (which is essentially PAM4 in the I and Q dimensions) modulation formats in Fig. 6. At 100 Gbit/s per lane (/per dimension), the baud rates for the two systems are similar, although 16QAM typically requires slightly higher overhead for pilots and 400ZR framing. The host side SerDes (serializer/de-serializer), FEC and PCS (physical coding sublayer) interface to the switch or some other host system is identical for both implementations. The

line or optical output of the transmit (TX) path are 4 streams of PAM4 data that are either individually encoded using Electrical to Optical converters and then optical multiplexed or optically converted to generate a single carrier, 16QAM signal using a polarization multiplexed IQ (PM/IQ) modulator structure. The TX DSP path, for all practical purposes, is identical.

Broadcom's Tomahawk 3 (TH3) and Tomahawk 4 (TH4), are running at 50Gbps lane rate driven by 50G PAM-4 technology. When these switch chips are used in chassis based systems, a 16:16 retimer PHY is typically needed in the line card to ensure robust delivery of 16x50G PAM-4 data from the switch to the optical interface. At the optical module, a 400G 8:4 gearbox PHY like Broadcom's BCM8740x can be used to convert 50G PAM-4 electrical I/Os to 100G per lambda optical I/Os in order to connect to 100G optics. Depending on the fiber link distance, certain specialized laser components such as VCSEL and EML are needed to enable single-lambda 100G transmission over multi-mode fiber (MMF) and single-mode fiber (SMF) optical cables. Following are some examples of 400G module implementation supporting various 400G interfaces: 400G SR4, 400G DR4, 400G FR4 and 400G LR4.

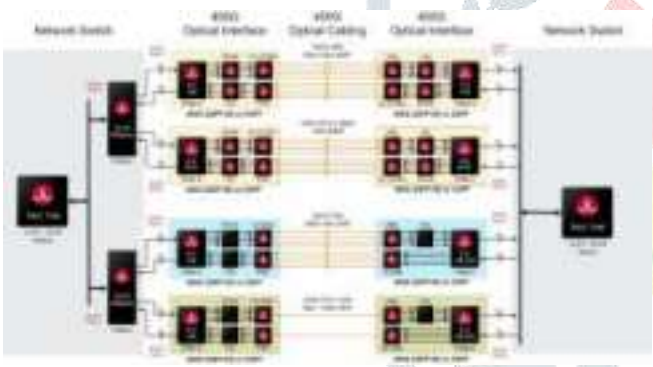

Figure 7:Broadcom Phy

VI. Silicon Photonics for 400G Applications

Fig. 8 shows coherent silicon photonics chip block diagram. The Transmit and Receive blocks are integrated onto the same chip. The modulator driver, at the input, and the TIA, at the output, are adjacent (to the left) of the silicon photonics chip.

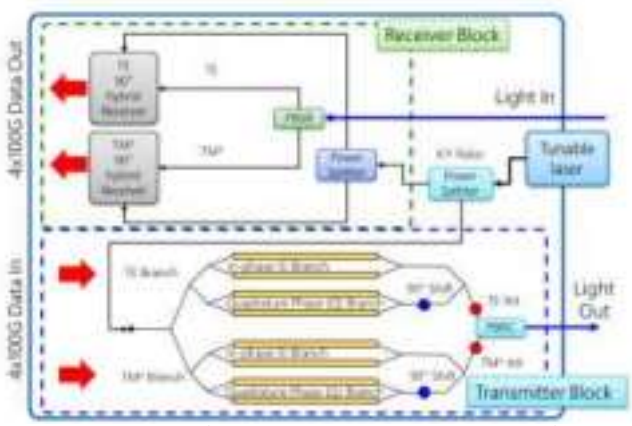

Figure 8: Block diagram for the transmitter and receiver blocks of the integrated, coherent silicon photonics chip

A single, external tunable laser is split between the transmit and receive blocks. The power split ratio is a trade off between the module output power and receiver LO (local oscillator) power. The dual polarization, together with the IQ modulation, gives the 16QAM the 4x increase in density compared to PAM4.

VII. Interface Requirement

The main component driving interface cost and power is the SerDes. For Long Reach (LR) SerDes, power was reduced from ~7pj/bit in 25G NRZ, to ~4pj/bit in 56Gbps PAM4, due to innovations in modulation and circuit design; however, going to 112Gbps speed, the SerDes power increased to ~7pj/bit as the fundamental transistor performance, and wiring did not improve significantly with process technology. The industry is now considering 224Gbps as the next speed upgrade, mainly driven by switch ASIC package limitations to achieve 100Tbps bandwidth and potentially cheaper optics with 200G lambda. Different modulation techniques (PAM4/6/8 and QAM) are under consideration. It is also expected that a stronger FEC (such as BCH FEC) is required to achieve acceptable BER, and that could imply a 2X increase in FEC area in the switch.

Co-packaged optics will help reduce overall system power and enable an increase in interface bandwidth. However, the challenge will be in manufacturing the solution at high volume and low cost at the package and system level. Multi-source solutions become more challenging compared to pluggable optics.

Whether co-packaged optics or better modulation techniques, lower loss material, and better FEC, the industry needs to solve the interface bandwidth problem to enable scaling of bandwidth for generations to come.

VIII.Overall Scalling and Extening the Challanges.

Power is emerging as a vital issue in scaling network switch bandwidth. Both the power of the switch silicon and the power of the switch system are essential. The system

power budget, as well as cooling, determine the maximum switch power possible. Multi-die solutions for higher switch bandwidth incur higher power as the interface between dies adds power to the package (e.g., PHY Tile approaches). Lower reach SerDes (VSR) does reduce switch chip power, but it adds system power as we need to account for re-timers. Direct drive co-packaged optics enables system power reduction and probably offers the best overall cost and power solutions in the future.

In terms of process technology, the biggest challenge is the wire scaling or lack thereof. As the process technology goes from 16nm to 7nm to 5nm, the so-called back end or metallization is not scaling, and in some instances, it is remaining the same (same resistance per unit). While the semiconductor industry is finding ways to pack more transistors by shrinking the transistor and developing 2.1D, 2.5D, and even 3D solution, the wiring remains a sore point preventing the full power of this density increase. Wiring density scaling has to be addressed by the silicon industry to gain efficiencies going forward.

To address scaling switch silicon challenges, Broadcom continues to invest in technologies to develop and build optimized products for the market segments with better economics, silicon power, system power, and high performance.

IX. PLUGGABLE 400-XX MODULES

A pluggable 400-XX module is shown in Fig. 9. The block diagram for the module construction is shown in Fig. 10. The coherent DSP chip has the host side interface to the switch and the line side interface to the silicon photonics chip. Our 100G PAM4 modules have a similar block level implementation.

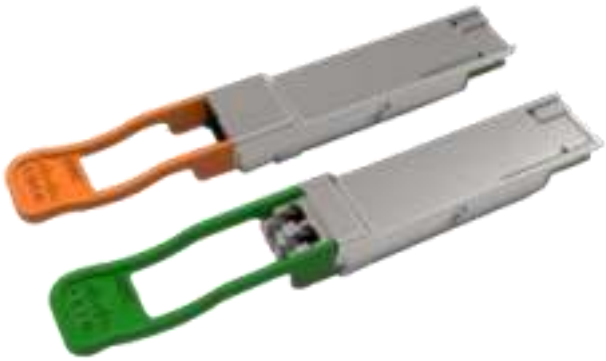

Figure 9: 400G Optic Module

The DSP chip is designed with the advanced 7nm CMOS node. It has 8 electrical inputs at 26.5625 Gbaud (53.125 Gbit/s) PAM4 signal (IEEE 802.3bs). The signal is then mapped into the 400-XX framing structure and encoded with CFEC. The output DAC feeds the driver IC which then drives the MZM on the silicon photonics chip.

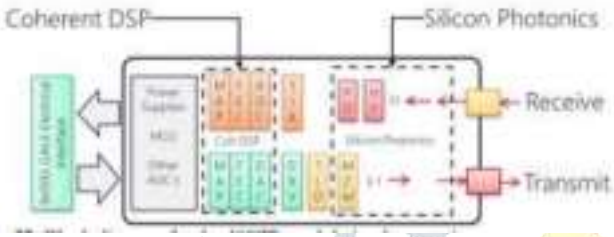

Figure 10 Block diagram for the 400ZR module implementation

On the receive side, the output of the TIA is sampled using the front-end ADC of the DSP. After signal processing and FEC decoding, the data is remapped into the ethernet frame and sent to the switch. A variety of power supplies, a micro-controller and control circuits fill rest of the module.

CONCLUSION

There are a whole array of constraints and trade-offs when designing low power, DSP based pluggable transceivers for data center applications. Total power utilization is critical for successful data center operations. Data center infrastructures have power supplied to them by the utilities. As the data rates grow, the growth in the overall services provided by the data center is limited by the maximum power available, which does not scale over time for an existing building. This imposes an upper limit on the power that can be allocated to the optical interconnects. The goal of the tutorial was to present the trade-offs that have been implemented and are being considered for data center optical interconnects that increasingly use digital signal processing for both direct detect and coherent transmission formats. With increasing data rates, the coherent transmission formats become more relevant to short distance applications. In the not too distant future, coherent links may displace direct detect formats even for intra data center applications.

A. Abbreviations and Acronyms

DSP: Digital Signal Processing
OFC: Optical Fibre communication
QSFP-DD-Quad Small Form-factor Pluggable (QSFP) – Double Density (DD),
DAC- Digital to Analog converter
Gbit- Giga Bit
QAM- Quadrature Amplitude Modulation
PAM4- 4 level Pulse Amplitude Modulation
GbE- Giga Bit Ethernet
ADC- Analog Digital Converter

B. Units.

Below Table describes the units of data measurement

Name	Equal To	Size(In Bytes)
Bit 1 Bit 01-Aug Nibble 4 Bits 1/2 (rare) Byte 8 Bits 1 Kilobyte 1024 Bytes 1024 Megabyte 1, 024 Kilobytes 1, 048, 576 Gigabyte 1, 024 Megabytes 1, 073, 741, 824 Terrabyte 1, 024 Gigabytes 1, 099, 511, 627, 776		

Table 1: Data measure units

REFERENCES

[1] S.Faruk and S.J.Savory, "Digital Signal Processing for Coherent Transceivers Using Multilevel Formats," J. Lightw. Technol., vol. 35, no. 5, pp. 1125-1141, Mar. 2017. [2] Broadcom 50G-800G Optical Module PHY Product Selection Guide- BC-0605EN
[3] R. Nagarajan, and I. Lyubomirsky, "Next-Gen Data Center Interconnects: The Race to 800G," Consortium for On-Board Optics webcast, Jan. 2021, https://www.onboardoptics.org/the-race-to-800ginphi
[4] Tony Antony, QSFP-DD Optical form factor: An evolutionary approach to 400GbE Interconnect, https://blogs.cisco.com/tag/cisco-400g-qsfp-dd