Title
Energy-Efficient Multi-Band Signaling for High-Speed Memory Interface

Author
Cho, Wei-Han

Publication Date
2016

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Energy-Efficient Multi-Band Signaling for High-Speed Memory Interface

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Electrical Engineering

by

Wei-Han Cho

2016

ABSTRACT OF THE DISSERTATION

Energy-Efficient Multi-Band Signaling

for High-Speed Memory Interface

by

Wei-Han Cho

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2016

Professor Mau-Chung Frank Chang, Chair

The scaling of CMOS technology continues to improve the processor capability and memory capacity, requiring memory interface with higher bandwidth and better energy efficiency to enhance the overall system performance. Among those cutting-edge designs of memory interface, multi-band signaling has shown great potential for its high throughput along with low energy consumption. With spectrally divided signaling, the multi-band transceiver can be designed to avoid spectral notches and extend communication bandwidth on multi-drop buses. Also, the multi-band transceiver is immune to inter-symbol interference caused by channel attenuation because of unique self-equalization double-sideband signaling. In this dissertation, we will demonstrate a tri-band transceiver with four parallel lanes that achieved a total data rate of 40Gb/s with total power consumption of 38mW in 28nm CMOS technology.

To realize the total data rate of 40Gb/s, PAM-4 and 16-QAM are used at the baseband and 3/6GHz bands, respectively, to carry 10 parallel bit streams at 1GBaud via each lane of the transceivers. These ten parallel bit streams share the same physical channel to minimize the time skew among them. In view of this, the strobe signal, DQS, is assigned to one of the ten bits for data recovery at the receiving end. Under 6dB attenuation at 6GHz on a 2" dense FR-4 differential bus (line pitch of 6mil), the transmitting end consumes only 1.6mW/lane. Together with 4.7mW/lane of the receiving end and 13.4mW of the carrier generator to be shared among all lanes consumes, the total power consumption and average energy efficiency are 38mW and 0.95pJ/b. Compared with prior arts, the proposed design achieves not only better energy efficiency but also substantial size advantage ($0.01mm^2$/lane, including the carrier generator). This transceiver realizes a total data rate of 40Gb/s with BER < 10^{-12}. Moreover, this tri-band architecture can be scaled in the frequency domain for further increasing the data throughput without increasing the symbol rate, which enables a new design dimension with more compact size and significantly improved energy efficiency for future memory interfaces.

The dissertation of Wei-Han Cho is approved.

Chih-Kong Ken Yang

Sudhakar Pamarti

Pei-Yu Eric Chiou

Mau-Chung Frank Chang, Committee Chair

University of California, Los Angeles

2016

To my wife, Pi-Feng Chiu

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

"I heard that you were capable of three men's job back in Taiwan, but in here you need to do ten men's job since I paid you three times more." Professor Frank Chang told me when I attended my first group meeting of High-Speed Electronics Lab at UCLA. I truly hope that I had met Professor Chang's expectation during the past four years.

First I have to express my sincere gratitude to my lab mates who had been working on RFI with me, which eventually resulted in this dissertation, and the list includes Rod Kim, Yilei Li, Yuan Du, Ken Wong, Jiequiong Du, Bug Huang, Gabriel Virbila. I also need to thank the most important person of HSEL, Janet Lin, and those guys who have laughed with me in lab or cubicle, Adrian Tang, Li Du, Richard Al Hadi, Yan Zhao, Boyu Hu, Shawn Wang, Yan Zhang, Yen-Hsiang Wang, Hao Wu, Arash Mirhaj, Frank Hsiao, Ryan Shin, Yen-Cheng Kuan, Hugh Wu, and Joseph Chen. Thanks also to my parents who were 6000 kilometers away during my four year journey of Ph.D., but their support is always within a second.

Most importantly, I have to thank my wife, Pi-Feng Chiu, for everything.

VITA

2004-2008	Bachelors of Science in Electrical Engineering National Tsing Hua University, Hsinchu
2008-2010	Masters of Science in Electronic Engineering National Tsing Hua University, Hsinchu
2010-2011	Second Lieutenant in Military Police Army, Taiwan
2011-2012	Research Assistant in Electronic Engineering National Tsing Hua University, Hsinchu
2012-2016	Graduate Student Researcher in Electrical Engineering University of California, Los Angeles

CHAPTER 1 INTRODUCTION

1.1. Overview of Wireline Communication Systems

During the long history of wireline communication, wide range of application has been established on various channel media. From telephone network, cable television to worldwide internet based on twisted pair, coaxial, or fiber-optic cables, the exchanged information gradually shift from analog to digital. Digitalization of information allows wireline communication to be facilitated by more and more powerful computing systems nowadays. For example, digital voice service, like Voice over IP (VoIP), provides a low-cost replacement for telephone network and allows digital processing for quality improvement, bandwidth efficiency, content enhancement, security, privacy, etc. Also, digitalization of cable television not only greatly raises the image resolution and channel number but also enables on-demand programs and cable internet through Data over Cable Service Interface Specification (DOCSIS) that uses the most complex modulation and requires heavy digital processing. While the quality, capacity and capability of wireline communication are greatly improved by processors, techniques used in wireline communication can also enhance computing systems.

1.2. Wireline Communication in Computing Systems

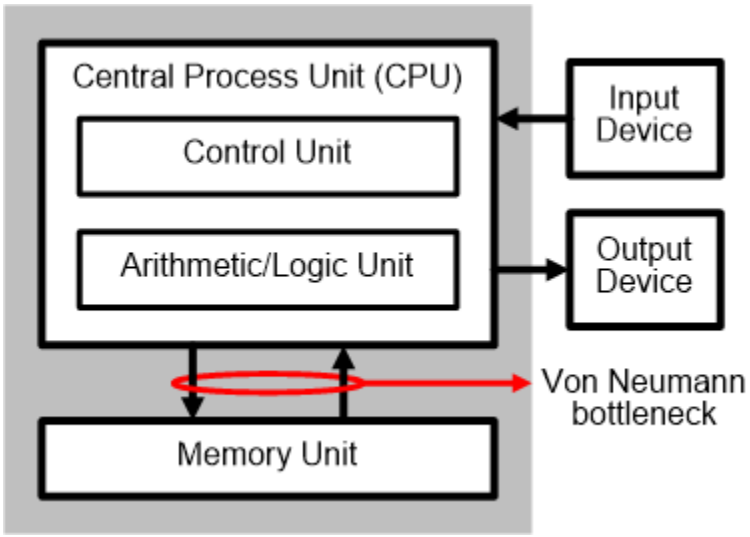


Figure 1.1 The von Neumann architecture and its performance bottleneck

It took 33 years for engineers to run into the von Neumann bottleneck since the famous architecture was first proposed in 1945 [1][2]. As described in the first draft written by John von Neumann, modern computing systems continue to fetch instructions and data from memory and then place the processed result back to memory. The performance of a computing system is then determined by its CPU speed and memory size. While the demand for higher performance keeps driving these two factors to grow, another factor emerged as the performance bottleneck in 1970s. At the time, the effective processing speed was once limited by the waiting time needed for data to transfer between CPU and main memory. Soon enough cache and memory hierarchy were introduced to hide the waiting time, but still memory interface start to play an important role in determining the performance of computing systems.

Non-return-to-zero (NRZ) signaling is the most commonly used form to transfer data in computing systems due to its simplicity. To transmit a binary bit stream, the easiest way is

sending high voltage to indicate logic 1 and low voltage to indicate logic 0 as in digital circuits, and that is exactly what NRZ signaling is. To recompose the binary bit stream at the receiving end, the time duration of each bit needs to be defined in order to tell two or more consecutive logic 1 from a long pulse of high voltage. People usually use "bit rate" to indicate the time duration. Besides bit rate, there are many other "rates." Symbol rate and data rate are the two most commonly referred rates in communication systems. With NRZ signaling, each symbol (either high or low voltage level) indicates only one binary bit and thus symbol rate always equals bit rate. With other signaling techniques, each symbol can contain more than one bit of information in one time frame and thus symbol rate can be lower than bit rate. Similarly, data rate can equal to or be lower than bit rate. If every bit that has been received is recognized as one bit of data, then the data rate equals the bit rate. However, sometimes redundant bits are inserted among data bits for various reasons. The most common reason is for DC balance and error bit detection or correction. With 8b/10b encoding for example, every 10 received bits will be translated to 8 bits of data. In that case, data rate will be 80% of bit rate. Nevertheless, since coding techniques will not be discussed in this dissertation, all data rate referred afterward equals to bit rate.

In frequency domain, NRZ signaling presents a very broad spectrum. As the transferred data are usually random, the NRZ signal is equally possible to be a constant voltage (when transferring endless logic 0 or 1) or a square wave with frequency equivalent to one half of the data rate (when transferring 101010…). That means the NRZ signal contains similar amounts of power at frequencies of zero and half data rate. To be more precise, the spectral power density of NRZ signals decreases very slowly from zero frequency to half data rate and then quickly drops to none at frequency of one data rate, which resembles the square of a sinc function. Beyond one

data rate, residual power, coming from the harmonics of square waves, resurges repeatedly as the second, third and more lobes. Usually only the first lobe is needed to reconstruct a healthy signal waveform and retrieve the data, and thus the frequency response of the channel medium is required to be flat from zero frequency to one data rate. For a twisted-pair cable used in telephone network, the flat response is only available within 3kHz, and thus it is simply impossible to achieved nowadays Asymmetric Digital Subscriber Line (ADSL)'s data rate of 24Mb/s with NRZ signaling. On the other hand, for memory interface made of Printed Circuit Board (PCB) traces that are only a few inches long, it was not too difficult to reach data rate of 66Mb/s with NRZ signaling when Synchronous Dynamic Random-Access Memory (SDRAM) was first introduced in 1990s. However, as the data rate of memory interface goes higher and higher, channel non-idealities that trouble those long cables in wireline communication systems start to appear on the short traces between CPU and main memory.

1.3. Demands for Energy-Efficient High-Speed Memory Interface

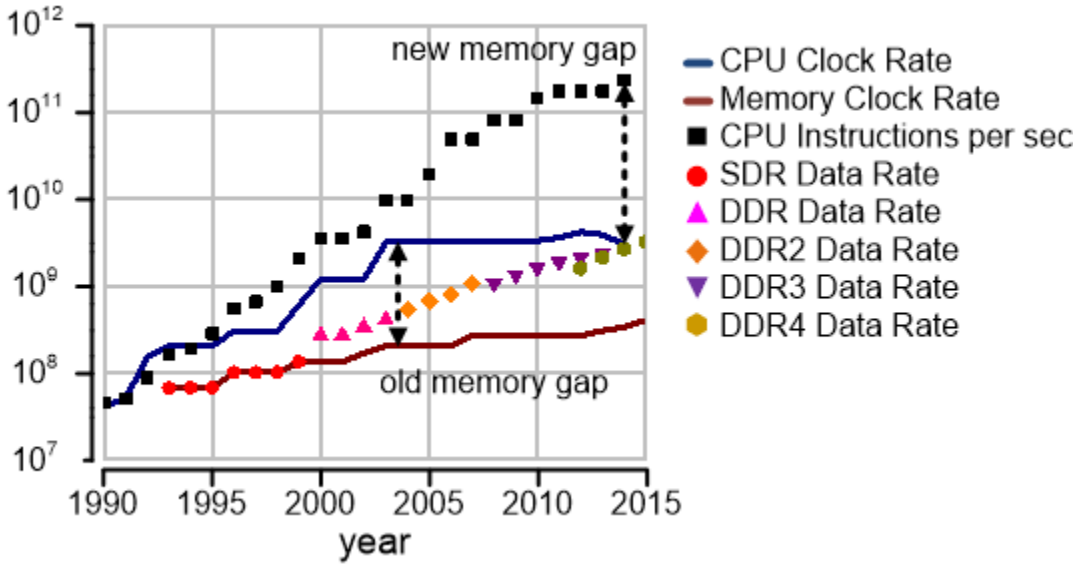


Figure 1.2 CPU-Memory performance gap over years

In the past 25 years, the semiconductor industry continues to follow Moore's law and reduce transistor size by half every two years. Reduction of transistor dimensions helps to enhance the CPU performance by decreasing the load capacitances and fitting more transistors within the same area. Before 2005, clock rate was a good indicator of CPU performance; however, after getting closer and closer to the power wall at 5GHz of clock rate, designers started to integrate more cores and more function-specific blocks in order to further improve computing capability without overdriving the clock frequency. As to memory, the scaling of CMOS transistors mostly contributes to the increase of memory capacity, and the clock rate of SDRAM merely increased from 66MHz to 400MHz, while the clock rate of CPU increased from 150MHz to 4.2GHz. The increasing difference of clock rate between CPU and memory was called the memory gap, as shown in Fig 1.2, which needs to be compensated by more and more complicated memory hierarchy. Double Data Rate (DDR) SDRAM and its successors were invented to make up the difference.

To increase the total throughput of memory interface, we can either increase the number of interconnects or increase the data rate of each interconnect. Due to limited I/O pin number of modern IC packaging, the data width of memory interface has reached the maximal 64 bits back in the age of Single Data Rate (SDR) SDRAM with 168-pin Dual-Inline Memory Module (DIMM). DDR technologies chose to focus on increasing the data rate of each I/O pin. Even in the latest DRR4, the data bus remained 64-bit wide despite the pin number of DIMM was raised to 288. With slow memory clock rate, DDR technologies serialize low-speed memory readouts into high-speed data bit streams. With 2:1 serialization, first generation of DDR achieved a data rate twice the memory clock rate. In 2015, with 8:1 serialization and 2:1 bank multiplexing, the data rate of DDR4 finally caught up with the CPU clock rate. However, if compared with CPU

instructions per second, the memory gap actually grew bigger than ever. In light of that, designers turned to the new 3DIC packaging technology.

Wide-I/O was proposed to take advance of a newly matured semiconductor process technique, called Through-Silicon Via (TSV), to stack memory directly on top of CPU [3][4]. Vertically integration of CPU and memory could greatly extend the data width of memory interface to 512 bits or more. Also, by stacking memory vertically, the footprint for the same capacity can be reduced. Therefore, it allows the memory to operate at higher clock rate with lower energy dissipation by reducing loading capacitance on bit lines and word lines. While everything about wide-I/O sounds just perfect, one thing about CPU inhibited it from wide adoption. In the 3D structure, the stacked memory stands right in the way of heat dissipation of CPU. As a result, with wide-I/O, the CPU needs to lower its clock rate during intense computation mode and thus limits its peak performance. On the other hand, High-Bandwidth Memory (HBM) took a step back and used TSV to stack only memory chips. As to the CPU-memory integration, HBM inserted an intermediate packaging layer that allows finer planar interconnection between CPU and memory before mounted onto the conventional Ball Grid Array (BGA) substrate. On the intermedia layer of silicon interposer, the memory interface made of micrometer-sized bumps and metal wires could be as wide as 1024-bit with data rate up to 1Gb/s each [5]. This innovation has been applied to graphical computation products, but the memory capacity is far from enough for general-purpose computing systems as HPM only supports up to 4GB per package. Especially for servers in data centers that requires hundreds of gigabytes, 4GB of HPM would be just another level of cache. Another technology called Hybrid Memory Cube (HMC) was proposed to interconnect CPU and memory on PCB with 64 lanes of high-speed serializer and deserializer (SERDES). At the bottom of stacked memory, HMC

integrated a layer of high-performance CMOS technology, with transistor performance better than DRAM technology, to implement the up-to-10Gb/s/lane SERDES array. With this heterogeneous architecture, the most advanced wireline techniques can then be incorporated into memory interface design.

Combining various types of equalization, including feedforward equalization (FFE), continuous-time linear equalization (CTLE) and decision-feedback equalization (DFE), nowadays serial link design can achieve data rate as high as 60Gb/s [6]. These equalization techniques are mainly used to compensate for severe channel attenuation caused by capacitive loading, skin effect and dielectric loss. These types of channel attenuation are usually predictable and thus can be compensated by pre-designed FFE and CTLE. For unpredictable channel characteristics, after trained by pre-defined sequence, DFE can adjust the compensation accordingly. However, DFE has its limitation. Impedance discontinuity and open stubs on multi-drop buses (MDB) of memory interface induce notches in channel frequency responses. With notch depth larger than 30dB, severe reflection and ringing appears in time domain and thus DFE requires many taps for the long-lasting pulse response. Also, for the same length of pulse response, the required DFE tap number increases with the data rate, and so does the power consumption of each tap. As a result, the energy efficiency of DFE degrades quickly as data rate increases. Around 10Gb/s, the lowest reported power is 3.8mW/Gbps for equalization up to 35dB [7]. This number is impressively low but still higher than 2.5pJ/b that have been achieved in DDR4 [8]-[9]. In the meanwhile, multi-band signaling has shown great potential because of its high data rate and low power consumption while being able to cleverly avoid the undesired effects caused by MDB and channel attenuation [10].

References:

[1] J. von Neumann, "First Draft of a Report on the EDVAC," *IEEE Annals of the History of Computing*, vol. 15, no. 4, pp. 27-75, Apr. 1993.

[2] John Backus, "Can programming be liberated from the von Neumann style?: a functional style and its algebra of programs," *Communications of the ACM*, vol. 21, no. 8, pp. 613-641, Aug. 1978.

[3] J.-S. Kim, et al.: 'A 1.2 V 12.8 GB/s 2 Gb Mobile Wide-I/O DRAM With 4 × 128 I/Os Using TSV Based Stacking', *IEEE J. Solid-State Circuits*, vol. 47, no. 1, pp. 107-116, Jan. 2012.

[4] S. Takaya, et al.: 'A 100GB/s wide I/O with 4096b TSVs through an active silicon interposer with in-place waveform capturing', *ISSCC Dig. Tech. Papers*, pp. 434-435, Feb. 2013.

[5] "A 1 Tbit/s bandwidth 1024 b PLL/DLL-Less eDRAM PHY using 0.3 V 0.105 mW/Gbps low-swing IO for CoWoS application," *IEEE J. Solid-State Circuits*, vol. 49, no. 4, pp. 1063-1074, Apr. 2014.

[6] J. Han, *et al.*, "Design techniques for a 60 Gb/s 173 mW wireline receiver frontend in 65 nm CMOS Technology" *IEEE J. Solid-State Circuits*, vol. 51, no. 4, pp. 871-880, Apr. 2016.

[7] N. Kocaman, *et al.*, "A 3.8 mW/Gbps quad-channel 8.5–13 Gbps serial link with a 5 Tap DFE and a 4 Tap Transmit FFE in 28 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 51, no. 4, pp. 881-892, Apr. 2016.

[8] T. C. Hsueh, *et al*. "A 25.6Gb/s differential and DDR4/GDDR5 dual-mode transmitter with digital clock calibration in 22nm CMOS", *ISSCC Dig. Tech. Papers*, pp. 444-445, Feb. 2014.

[9] Young-Chul Cho, *et al*., "A sub-1.0V 20nm 5Gb/s/pin post-LPDDR3 I/O interface with Low Voltage-Swing Terminated Logic and adaptive calibration scheme for mobile application," *Symposium on VLSI Circuits*, 2013, pp. 240-241.

[10] W. Cho, *et al*., "A 5.4-mW 4-Gb/s 5-Band QPSK Transceiver for Frequency-Division Multiplexing Memory Interface," in *IEEE CICC* Dig. Tech. Papers, Sept. 2015.

CHAPTER 2 MULTI-BAND SIGNALING

2.1. Introduction of Multi-Band Signaling

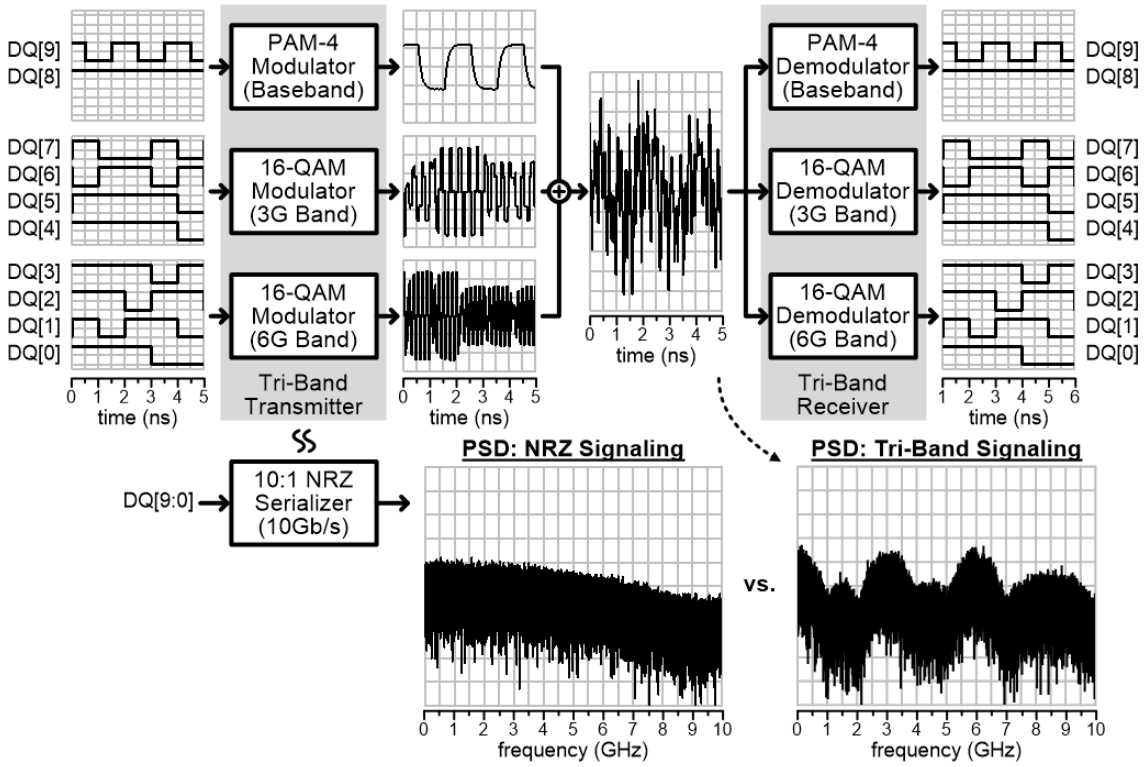


Figure 2.1 Illustration of multi-band signaling in time domain and frequency domain using PAM-4 / 16-QAM tri-band signaling and its comparison with NRZ signaling as example

Unlike NRZ's broad spectrum, multi-band signal has a divided and narrow spectrum. Here we take the tri-band PAM-4 / 16-QAM signaling used in Chapter 3 for example (Figure 2.1). Two random bit streams are modulated with PAM-4 and converted to spectral components inside the first lobe at baseband. Another four random bit streams are modulated at 3GHz with 16-QAM and converted to spectral components inside the second lobe centered at 3GHz. The other four input random bit streams are modulated at 6GHz with 16-QAM and converted to

spectral components inside the third lobe centered at 6GHz. In total, ten input bit streams are modulated simultaneously through the tri-band PAM-4 / 16-QAM signaling and thus an aggregate data rate of 10Gb/s can be achieved with a symbol rate of 1GBaud. With the much lower symbol rate, compared to 10GBaud of NRZ signaling, most channel quality requirements can be greatly relieved. Typically NRZ signaling requires an insertion loss ripple to be less than ±2dB (some protocols require ±1dB) and a group delay variance less than ±0.1UI within the signal bandwidth (60-90% of data rate). With the lower symbol rate, the frequency range of interest is much smaller and thus it is easier to meet the requirements. Also, multi-band signaling can handle many channel non-idealities that are very difficult to solve while using NRZ signaling.

2.2. Comparison with Non-Return-to-Zero (NRZ) signaling

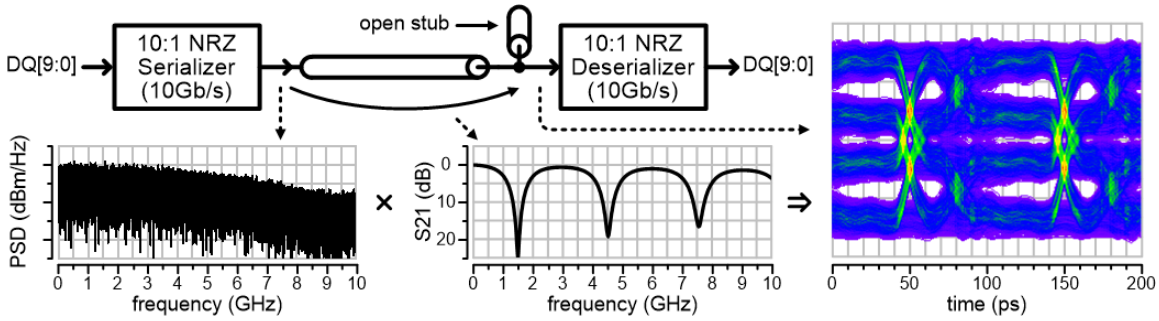


Figure 2.2 *Time responses (eye diagrams) of non-return-to-zero (NRZ) signaling with channel frequency notches*

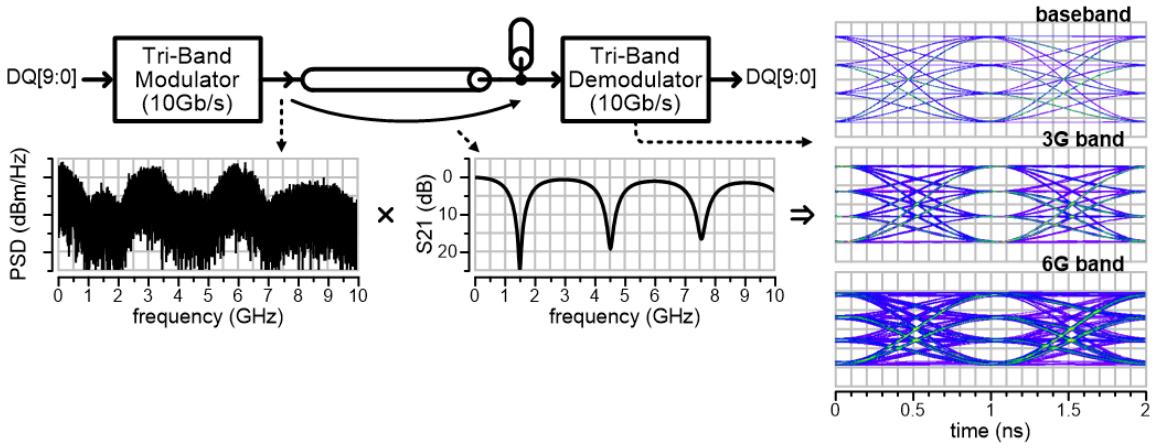


Figure 2.3 *Time responses (eye diagrams) of PAM-4 / 16-QAM tri-band signaling with channel frequency notches*

As mentioned before, open stubs on multi-drop memory buses can cause notches in the channel frequency response. At the notch frequencies, transmitted signal is entirely reflected and absent at the receiving end. As a result, the horizontal data eye opening reduces, and closes completely when the data rate exceeds twice the first notch frequency. Figure 2.2 shows one example when the data rate is 10Gb/s and the first notches frequency is located at 1.5GHz; the data eye is completely closed as mentioned. Using DFE to retrieve data from such signal can be very power hungry with a huge area overhead. As many as 18 DFE taps are required in some cases. With the same channel condition, a multi-band signal can be designed to bypass these frequency notches. As shown in Figure 2.3, the PAM-4 / 16-QAM tri-band signal utilizes three of the passbands (centered at baseband, 3GHz and 6GHz, respectively) on the channel with frequency notches. Since no significant signal energy is located at frequency notches, little reflection is induced. Also, the main lobes of each band are completely transmitted to and remain intact at the receiving end. The demodulated signals present wide horizontal eye opening, which greatly simplifies process of data recovery. The eye diagrams of 3GHz band and 6GHz band are superposed of both in-phase and quadrature demodulated signals. Note that with different

locations of these frequency notches, the carrier frequencies and symbol rate of the multi-band signal must be adjusted accordingly in order to preserve signal integrity.

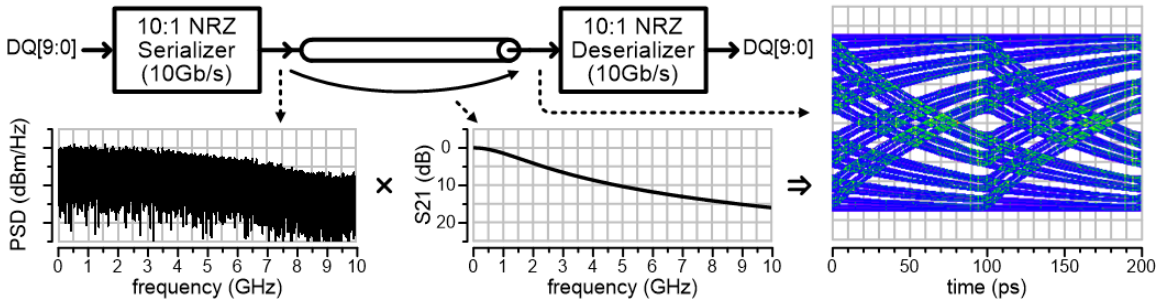


Figure 2.4 *Time responses (eye diagrams) of Non-return-to-zero (NRZ) signaling with monotonic channel attenuation*

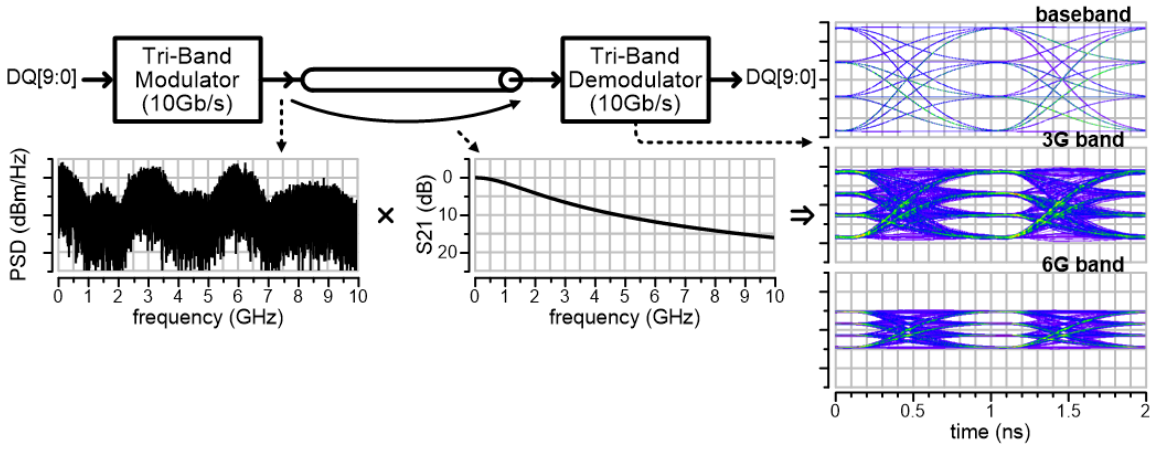


Figure 2.5 *Time responses (eye diagrams) of PAM-4 / 16-QAM tri-band signaling with monotonic channel attenuation*

Another non-ideality that can be cleverly handled with multi-band signaling is the channel attenuation. In most cases, channel attenuation is monotonic and increases with frequency. A small ripple could be induced by impedance mismatch but it can be easily reduced to an insignificant level with reasonable matching conditions. To have a ripple smaller than ±1dB, either one-ended matching of S_{11} < -24dB or both-ended matching of S_{11} < -12dB is

required. Impedance mismatch also causes a ripple on group delay but it is less significant for channels with length less than 4 inches on FR-4. More detail can be found in section 3.1.3. With a well-matched channel, the most common sources of channel attenuation include capacitive loading, skin effect and dielectric loss. All three of them present similar trends of attenuation increasing with frequency, except each at a different rate. With capacitive loading, the channel attenuation increases at a rate of -20dB/dec. With skin effect, the channel attenuation increases at a rate of -10dB/dec. With dielectric loss, the channel attenuation also increases at a rate of -20dB/dec. At frequencies between 1 and 10GHz, skin effect usually dominates, and then dielectric loss starts to kick in beyond 10GHz. The effective frequency range of capacitive loading depends on the value of capacitance. Regardless of the exact increasing rate, channel attenuation will increase monotonically. With this monotonic channel attenuation, the input signal at the receiving end toggles less rapidly than the output signal at the transmitting end. Consequently, the received signal presents reduced either horizontal or vertical eye opening. When the channel attenuation at Nyquist frequency is 12dB larger than that at DC, the data eye will completely close up, giving no chance for correct data recovery. Figure 2.4 shows one example when the channel attenuation at Nyquist frequency is about 10dB larger than that at DC. FFE at the transmitting end and CTLE at the receiving end can help to restore sufficient eye opening. Even though these two types of equalization are less power hungry with smaller area overheads compared to DFE, their contribution to total power consumption and chip area is still significant in designs of high-speed interconnect. On the other hand, multi-band signaling requires less, or none in most cases, equalization circuitry. As shown in Figure 2.5, with the same channel condition and without any equalizer, the demodulated signals at the receiving end of the PAM-4 / 16-QAM tri-band signaling again remain intact and preserve wide eye opening.

Two reasons involve the ineffectiveness of channel attenuation. The first reason is because each band of the tri-band signal occupies a much smaller bandwidth, and thus the insertion loss variation within its bandwidth is much smaller. The baseband signal takes a bandwidth smaller than the channel 3dB bandwidth and the insertion loss variation within the signal bandwidth is about only 1dB. The second reason is because of self-equalization of double-sideband (DSB) signals. For the other two bands centered at 3GHz and 6GHz, even with the smaller signal bandwidth, the insertion loss variation is still larger than 4dB. However, the eye diagrams of 3GHz and 6GHz bands are still horizontally wide open due to self-equalization. The vertical eye opening is still reduced but can be easily fixed with plain amplification at either the transmitting end or receiving end. In the next section, more detail and limitations about self-equalization of DSB signals will be discussed.

2.3. Self-Equalization of Double-Sideband (DSB) Signaling

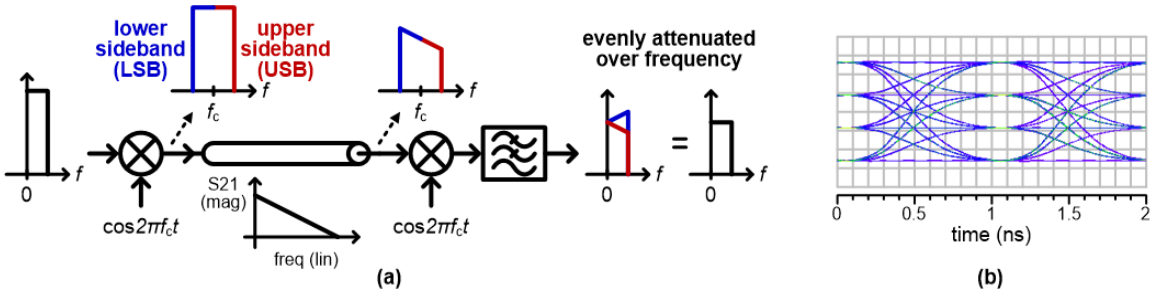


Figure 2.6 *(a) simple explanation of self-equalization and (b) the time response (eye diagram) of a self-equalized DSB signal*

A DSB signal can be obtained by modulating a baseband signal (Figure 2.1(a)). After frequency up-conversion, the DSB signal is composed of two copies of the original baseband signal, mirrored to each other and centered at the carrier frequency (f_c) side by side. The copy below f_c is called lower sideband (LSB) and the other beyond f_c is upper sideband (USB).

Passing through the channel with straight downward frequency response, the DSB signal attenuates less at LSB and more at USB. After frequency down-conversion, both LSB and USB are converted back to baseband and then LSB compensates for USB. As a result, the demodulated signal at baseband is evenly attenuated over frequencies; this means, to the demodulated signal, the effective channel frequency response is flat with constant attenuation thus and zero insertion loss variation. In such an ideal case, the demodulated signal presents an ideal eye diagram (Figure 2.6(b)). This ideal situation happens only when the channel frequency response is straight in linear scale (not log scale). However, channel frequency response is usually not straight and thus insertion loss variation is usually not zero but still greatly reduced compared to that of NRZ signals without self-equalization. Before we discuss the exact value of insertion loss variation after reduction, another non-ideality of DSB signaling needs to be mentioned.

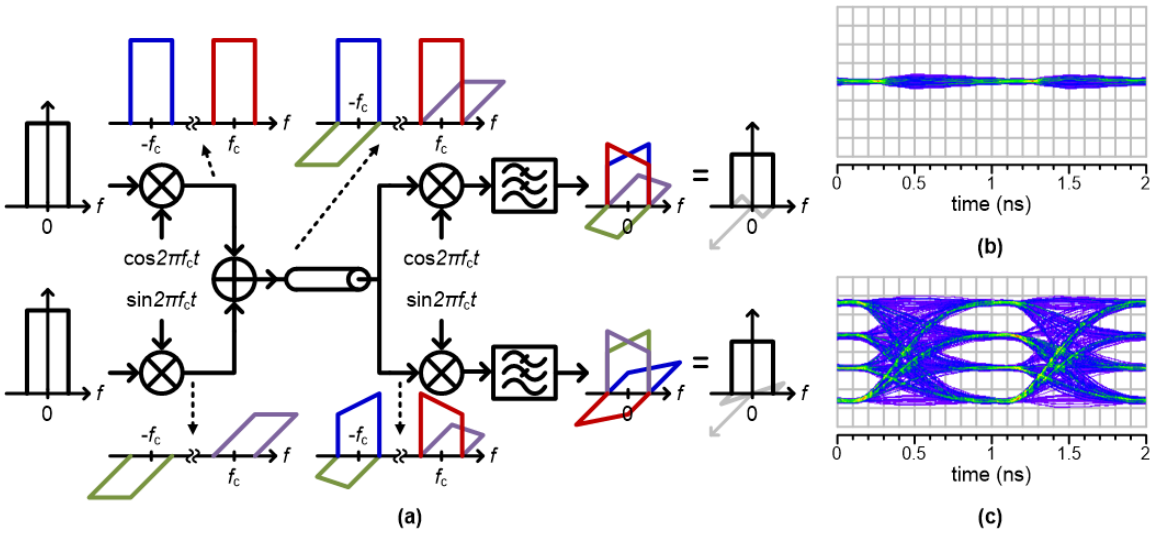


Figure 2.7 *(a) Illustration of I/Q interference of quadrature modulation due to uneven channel attenuation, (b) the time response (eye diagram) of I/Q interference in time-domain, and (c) the degraded time response (eye diagram) due to I/Q interference*

With a quadrature (90°) phase difference, two carriers at the same frequency are mathematically orthogonal. Therefore, two baseband signals that are separately modulated by the two orthogonal carriers ideally can be demodulated without interference from each other, which is called quadrature modulation and the two carriers are referred to as in-phase (I) and quadrature (Q). In that case, the two modulated signals share the same frequency band and double the aggregate data rate without any penalty. In reality, phase noise of carrier generators causes vibration of the phase difference, compromises the orthogonality and induces I/Q interference, which increases probability of error bits during data recovery. Besides phase noise, uneven channel attenuation also brings about I/Q interference. To explain this, we need to introduce the concept of negative frequency and reexamine the orthogonality of quadrature modulation. With negative frequency, a baseband signal is a DSB signal itself, which is centered at 0Hz in frequency domain, and frequency up-conversion is simply shifting the center of the DSB signal to the carrier frequency (Figure 2.7(a)). With the in-phase carrier ($\cos 2\pi f_c t$), the baseband signal is shifted to both $+f_c$ and $-f_c$. With the quadrature carrier ($\sin 2\pi f_c t$), another baseband signal is also shifted to both $+f_c$ and $-f_c$ but multiplied by $-j$ and $+j$, respectively (assume j is square root of -1). Combining the two modulated signals, we have a complex signal at the output of the transmitting end. At the receiving end, during frequency down-conversion, the in-phase carrier again shifts the complex signal by $+f_c$ and $-f_c$. Ignore the components located at $2f_c$ and $-2f_c$, which can be greatly attenuated with a low-pass filter, and focus on the components that have been shifted to baseband. Two real components that were modulated by the in-phase carrier share the same sign and thus are constructive to each other. The other two imaginary components that were modulated by the quadrature carrier have the opposite sign and thus are destructive to each other. Furthermore, since the two imaginary components share the exact same shape, they

perfectly cancel each other. As a result, only signals that are modulated by the in-phase carrier will remain after demodulation by the in-phase carrier. With similar procedure, we can prove that only signals that are modulated by the quadrature carrier will remain after demodulation by the quadrature carrier. This happens when the channel frequency response is flat. With uneven channel attenuation, the two imaginary components are still destructive to each other but their shapes become different. Without perfect cancellation, there is a remaining imaginary component at the output of the low-pass filter that interferes with the desired real component. With the I/Q interference (Figure 2.7(b)), the final output eye diagram (Figure 2.7(c)) is slightly degraded from the ideal case with self-equalization (Figure 2.6(b)).

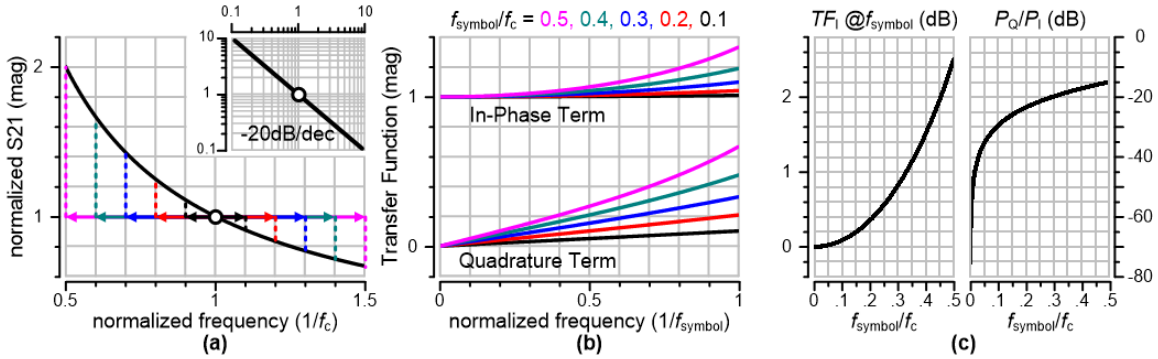


Figure 2.8 *(a) Example channel frequency response with a slope of -20dB/dec, (b) effective I/Q transfer functions derived from the example, and (c) peaking / interference of the transfer functions*

The degree of degradation depends on the degree of unevenness of channel attenuation. The straightness of channel attenuation also determines the insertion loss variation, which exacerbates the data eye degradation as we mentioned previously. Therefore, even though the multi-band signaling can handle worse channel condition than NRZ signaling, but it still has its limitation. To quantify the limitation, we first examine the case with a slope of -20dB/dec due to capacitive loading or dielectric loss (Figure 2.8(a)). The example channel frequency response is

straight in log scale (-20dB/dec) but concave in linear scale. When the symbol rate, f_{symbol}, is much smaller than the carrier frequency, f_c, the frequency response can be approximated as a straight line in linear scale, and thus the effective transfer functions of the in-phase component is pretty flat (the upper black line in Figure 2.8(b)), which causes little insertion loss variation. Also, the difference of the channel frequency response within $\pm 1 \times f_{symbol}$ is still small, and thus the effective transfer functions of the quadrature component is near zero (the lower black line in Figure 2.8(b)), which induces few I/Q interference. When f_{symbol} goes up, the channel frequency response looks more curvy and uneven. Consequently, the insertion loss variation and I/Q interference both worsen. To manage the degradation of output eye diagram, we require the insertion loss variation to be less than 1dB and the I/Q interference to be less than -20dB. From Figure 2.8(c), we find that f_{symbol} needs to be smaller than $f_c/3$. With different channel conditions, f_{symbol} limitation will be different. With a less steep channel frequency response, -10dB/sec for example, f_{symbol} can be higher while maintaining the same quality of output eye diagram. With a steeper channel frequency response, for example -30dB/sec, f_{symbol} needs to be lower to sustain the same quality of output eye diagram. With different modulation, the requirement will also be different. The insertion loss variation of <1dB and I/Q interference of <-20dB might be a little overdesigned for 16-QAM, but definitely not enough for 1024-QAM. The exact requirement for different situations can be found using similar analysis procedure.

2.4. Frequency Notches on Multi-Drop Buses (MDB)

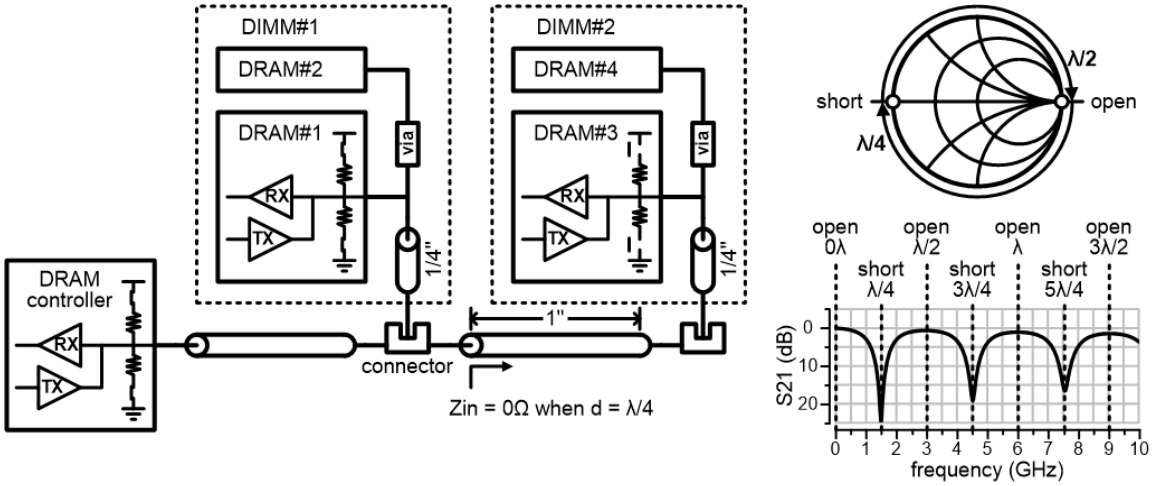


Figure 2.9 A dual-DIMM multi-drop memory bus and the analysis of its induced frequency notches

Now knowing how to determine the symbol rate with a certain carrier frequency, the remaining question is what determines the carrier frequency. As we mentioned, the multi-band signaling can bypass notches in the channel frequency response. For best signal quality, the carrier should be placed at passbands in the middle of two consecutive notches and thus the carrier frequency should be determined by the notch frequencies. With a dual-DIMM multi-drop memory bus (Figure 2.9), the worst case is when data is exchanged between the controller and first DIMM with the second DIMM turning off and then the transmission line between the first and second DIMM become a long open stub. Assume the length of the open stub is l. The loading impedance of the open stub is circling on the Smith chart with increasing frequency and decreasing wave length, λ. When $l = \lambda/4$, the loading impedance become near zero, which means the entire transmitted signal will be short to ground and none will be received. That is when the first notch is formed. When $l = \lambda/2$, the loading impedance returns to high and the entire transmitted signal can be received again. This cycle continues and notches are located at

frequencies when the length of the open stub equals an odd multiple of λ/4, l = λ/4, 3λ/4, 5λ/4, etc. Also, the passbands can be found at frequencies when the length of the open stub equals an odd multiple of λ/4, l = λ/2, λ, 3λ/2, etc. That means passbands can be found at every even multiple of the first notch frequency.

When the distance between the two DIMMs is one inch, the first frequency is located at 1.5GHz and thus passbands are at 3GH, 6GHz, 9GHz, etc. However, while a signal is modulated at 3GHz, the harmonics will be located at 6GHz, 9GHz, etc., which becomes severe interference if other signals are modulated at 6GHz 9GHz, etc. The 2nd-order harmonic located at 6GHz can be greatly suppressed with fully differential signaling, which means the second passband is now available. The 3rd-order harmonics can be suppressed with filters or harmonic-rejection mixers but these are of no interest to this work due to excessive circuit overhead. Therefore, three frequency bands are used in Chapter 3, located at baseband, 3GHz, and 6GHz. To ensure $f_{symbol} < f_c/3$, the symbol rate is set to 1GBaud.

CHAPTER 3 FOUR-LANE TRI-BAND PAM-4 / 16-QAM TRANSCEIVER

3.1. Transceiver System Overview and Analysis

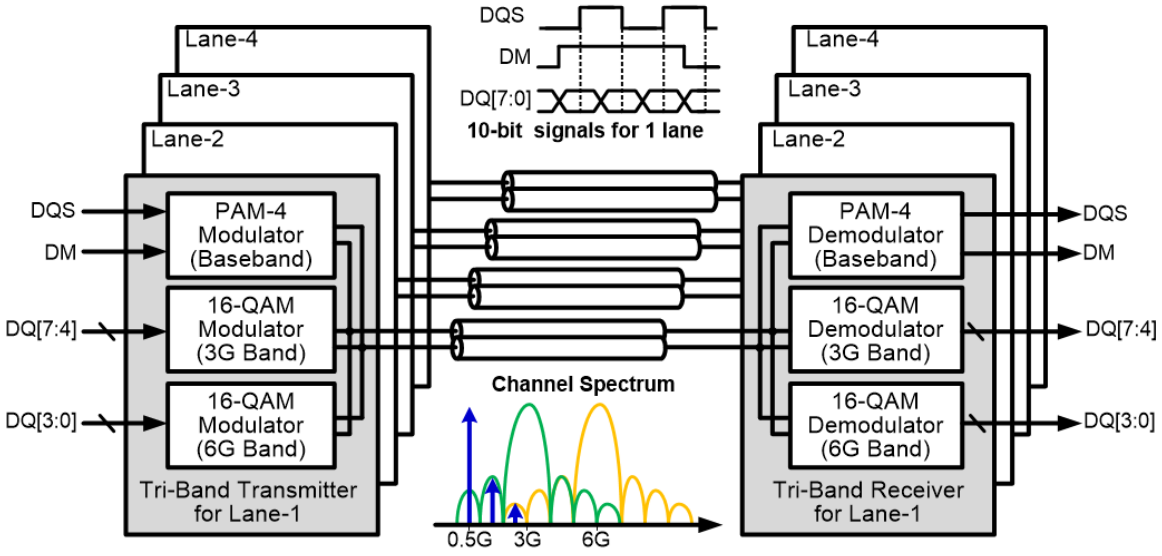


Figure 3.1 System architecture of the 4-lane tri-band transceiver with PAM-4 at baseband and 16-QAM at 3 and 6GHz

In this chapter, we will go through a 4-lane tri-band transceiver implemented in TSMC 28nm HPC technology. With PAM-4 at baseband and 16-QAM at 3GHz and 6GHz bands, the transceiver achieves an aggregate data rate of 10Gb/s/lane and 40Gb/s in total while operating at symbol rate of 1Gbaud. Using multi-band signaling, this transceiver can bypass and avoid reflection caused by notches in the channel frequency response with depth of greater than 30dB. Also, due to self-equalization of double-sideband (DSB) signal, the transceiver can easily handle more than 10dB attenuation at Nyquist frequency without any equalization circuitry. Including a

dual-band I/Q carrier generator, this transceiver takes up an area of only 0.01m m^2/lane and consumes only 38mW. With total data rate of 40Gb/s, the energy efficiency is 0.95pJ/b. A 32-bit built-in self BER tester is integrated with the transceiver and the measured BER is less than 10^{-12}. The overall experimental results show that the multi-band RF interconnect technology can scale the data rate in frequency domain and provide an energy/area-efficient method to tolerate channel non-ideality other than conventional wireline equalization techniques.

3.1.1 Adjacent-Band Interference

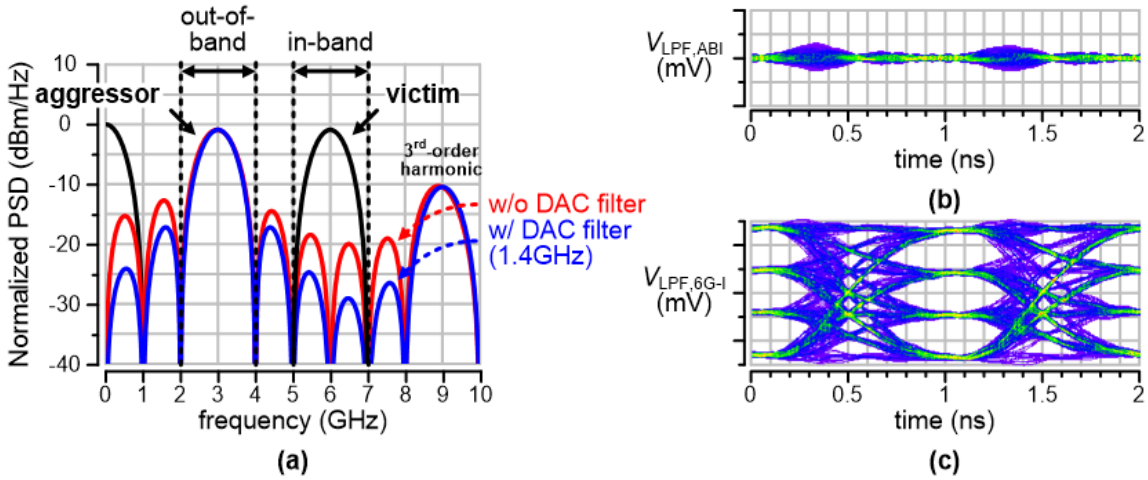


Figure 3.2 *(a) Adjacent-band interference analysis, (b) folded waveform of the remaining interference, and (c) the eye diagram of the demodulated signal from 6GHz band*

Three signals modulated at three different frequencies are combined at the output of the transmitting end. At the receiving end, after frequency down-conversion, the low-pass filter needs to suppress not only the up-converted components from the desired signal but also undesired components from other frequency bands. While demodulating the 6GHz band, the 3GHz band is the major source of adjacent-band interference (Figure 3.2(a)). The main lobe of the 3GHz band will remain centered at 3GHz after mixing. Therefore, to suppress the main lobe

sufficiently, the low-pass filter needs to provide 30dB rejection at offset frequency of 3GHz. Besides the main lobe, the side lobes centered at 5.5GHz and 6.5 GHz are also problematic. Unlike the main lobe and the other side lobe, the two side lobes cannot be suppressed by the low-pass filter because they are located within the main lobe of the desired frequency band. Therefore, the two side lobes are referred to as in-band interference. The in-band interference can be suppressed by a pulse-shaping filter at the transmitting end. The pulse-shaping filter can either be implemented digitally together with the digital-to-analog converter (DAC) or it can be simply an analog low-pass filter inserted at the output of the DAC. In this work, a single capacitor is inserted at the output of the DAC to suppress the in-band interference to be 30dB lower than the main lobe of the desired frequency band. With the remaining in-band and out-of-band interference (Figure 3.2(b)), the output eye diagram of the in-phase signal at 6GHz is slightly degraded but still wide open (Figure 3.2(c)). Similar to I/Q interference, the requirement for adjacent-band interference will be more stringent if more complex modulation is adopted, e.g. 1024-QAM. In such cases, more complex filters are required at both the transmitting and receiving end.

Finally, the PAM-4 / 16-QAM signaling is examined with a real channel model. The channel model is built based on a 2” FR-4 multi-drop memory bus with 1” open stub. The frequency response of the real channel model has notches where the first notch frequency is located at 1.5GHz, and the channel attenuation at 6GHz is about 6dB. The output eyed diagrams of signals at baseband and 3GHz band has similar eye opening, which is slightly smaller than that of signals at 6GHz band (Figure 3.3). Based on numbers of signal-to-interference ratio (SIR), the baseband signal is about the same as the 3GHz band signals and about 3dB worse than the 6GHz band signals. The reason is because, for the baseband and 3GHz band signals, the

adjacent-band interference comes from both sides (upper and lower frequency), but for 6GHz band signal, the adjacent-band interference comes from only one side. The constellation plots of 3GHz and 6GHz band signals show the same result. The error vector magnitude at 3GHz band is 3dB worse than that at 6GHz band. However, this does not mean that the 6GHz band will always have a better bit error rate (BER).

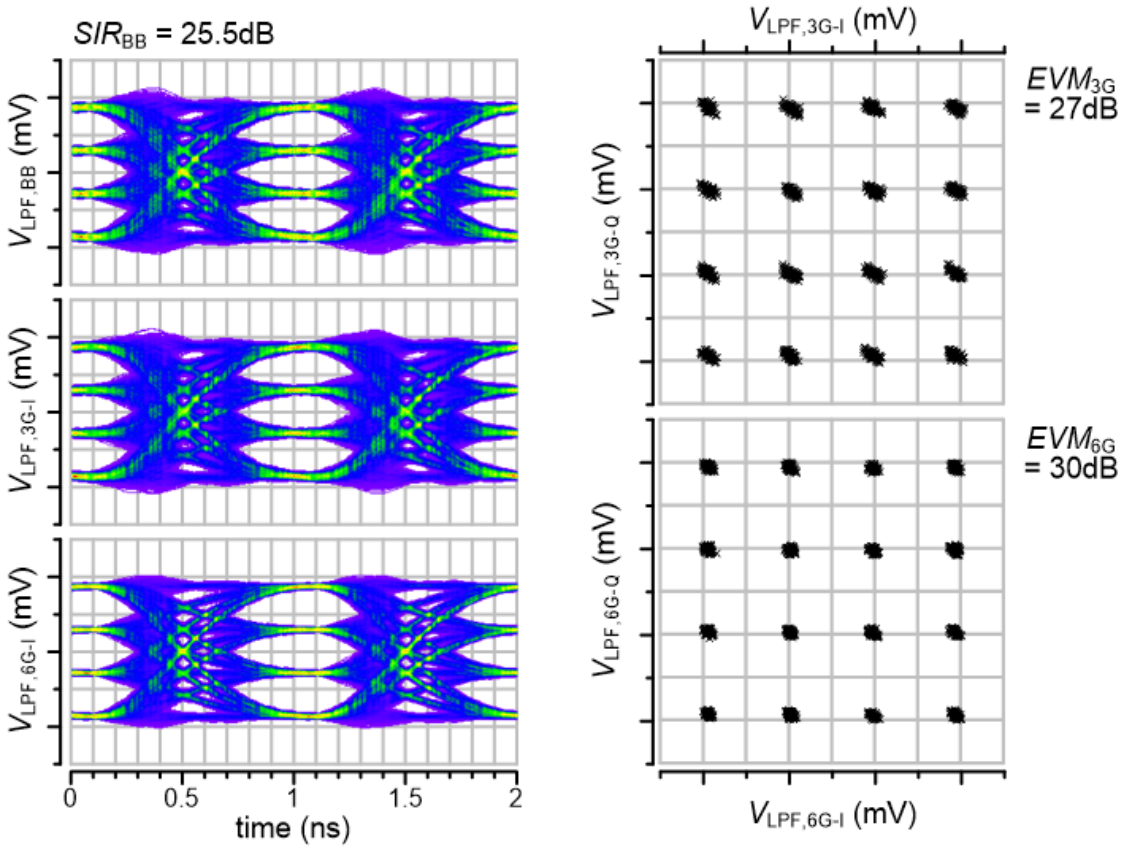


Figure 3.3 *Time responses (eye diagrams) of the PAM-4 / 16-QAM tri-band signaling and the constellations of 3GHz and 6GHz bands*

3.1.2 Carrier Phase Noise and Jitter

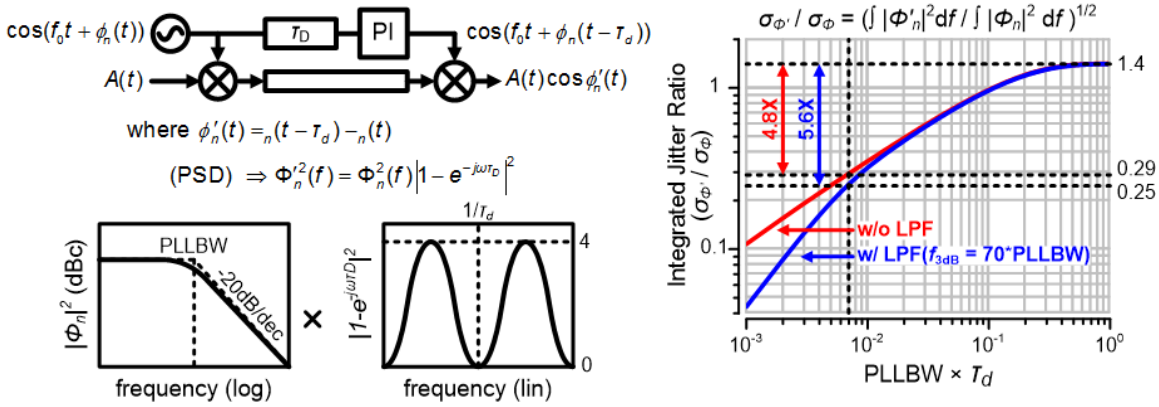


Figure 3.4 Phase noise shaping of synchronous signaling and its effect on carrier jitter

Phase noise is another key factor in determining the BER. To sustain a certain BER, the phase noise requirement of 6GHz band is more stringent than that of 3GHz band. Therefore, it is possible for the 6GHz band to have a worse BER even with less interference. To determine the requirement, we first need to look at the clock distribution of memory interface. As the clock of memory circuits is usually provided by memory controllers, a reference clock signal can be transmitted along with data signals on the memory bus. In that case, some of the phase noise can be canceled or reduced. Figure 3.4 shows one example that explains this phenomenon. Assume the clock and data signals have different delays from the transmitting to receiving end and the difference is τ_d. Then the effective phase noise ($\phi'_n(t)$) at the output of low-pass filter will be $\phi_n(t - \tau_d) - \phi_n(t)$, where $\phi_n(t)$ is the carrier jitter at the transmitting end. When τ_d is zero, which means the clock and data signals share exactly the same delay, the effective phase noise is zero. When τ_d is infinitely large, which mean $\phi_n(t - \tau_d)$ and $\phi_n(t)$ are two identical but independent random processes, the effective phase noise is then a random process that resembles $1.4\times\phi_n(t)$. With a finite but non-zero τ_d, phase noise at different frequency will respond differently. Phase noise at

frequencies of $1/\tau_d$ and its multiples is perfectly cancelled and phase noise at frequencies of $1/2\tau_d$ and its odd multiples is doubled in magnitude. Therefore, this phase noise shaping effect responds differently with different phase noise spectrum. If the phase noise is white or very broadband, the integrated jitter will remain unchanged. If the phase noise is concentrated at low frequencies, the integrated jitter will greatly reduce. In most cases, the reference clock is generated with a phase-locked loop (PLL) and the phase noise spectrum looks like a 1st-order low-pass response, which is flat within the loop bandwidth and decreasing at a rate of -20dB/dec beyond the loop bandwidth, where the loop bandwidth is usually around 10MHz. Some phase noise spectrums can have peaking or damping around the loop bandwidth frequency, depending on the percentage of phase noise contribution from the oscillator, but here we focus on only the general case to simplify the derivation. Multiplying the squares of the phase noise spectrum and shaping frequency response, and then integrating in frequency domain, the square root of the result indicates the standard deviation of the effective integrated jitter ($\sigma_{\phi'}$). On a 2" memory bus, the maximum delay difference is about 0.7ns when the data and clock signals are transmitted in the opposite directions. That means, in the worst case, the loop bandwidth is about $1/140\tau_d$ and the effective jitter reduces by 71% compared to the integrated jitter without shaping (σ_ϕ). Adding 3rd-order low-pass filtering with 3dB bandwidth of 700MHz to the frequency response of phase noise shaping, the effective jitter reduces even more ($\sigma_{\phi'} = \sigma_\phi/4$).

Knowing the exact reduction ratio of the effective integrated jitter, we can now start calculating the BER. As phase noise shifts the in-phase and quadrature signals by $\cos\phi'_n(t)$ and $\sin\phi'_n(t)$, respectively, the corresponding signal dot rotates on the I/Q constellation plot (Figure 3.5). Error bits occur when the dot rotates out of the decision boundary, which gives us an allowance of phase error in degree. With the phase error allowance, we can find the BER by

comparing the error allowance and standard deviation of carrier jitter. Assume the distribution of carrier jitter is Gaussian. If the ratio of the error allowance to standard deviation is larger than 7, the expected BER is lower than 10^{-12}. While transmitting different signals, the corresponding dot locations and error allowances will be different. Also adjacent-band interference and carrier phase error could shift the dots and shrink the error allowances. Including all these factors, the BER equation is shown in Figure 3.5, where $\Delta\theta$ is the carrier phase error and Δv is the adjacent-band interference in amplitude. The phase interpolation used in this work provides maximum step size of 1.2ps, which is ±1.3° of carrier phase error at 6GHz and the 30dB error vector magnitude (EVM) of 6GHz band is equivalent to 3.1%. With these two numbers, we find that the jitter requirement for BER < 10^{-12} is about 2° or $3.8ps_{rms}$ at 6GHz if counting the phase noise shaping effect. The integrated jitter requirement of $3.8ps_{rms}$ or equivalently $53.2ps_{p-p}$ for BER < 10^{-12} is comparable to that of 10Gb/s NRZ signaling.

$$BER < \frac{1}{2}\left[Q\left(\frac{16.9^\circ - \Delta\theta - \Delta v(268^\circ/2\pi)}{\sigma_\phi}\right) + Q\left(\frac{18.4^\circ - \Delta\theta - \Delta v(360^\circ/2\pi)}{\sigma_\phi}\right) + Q\left(\frac{20.8^\circ - \Delta\theta - \Delta v(360^\circ/2\pi)}{\sigma_\phi}\right) + Q\left(\frac{45^\circ - \Delta\theta - \Delta v(805^\circ/2\pi)}{\sigma_\phi}\right)\right]$$

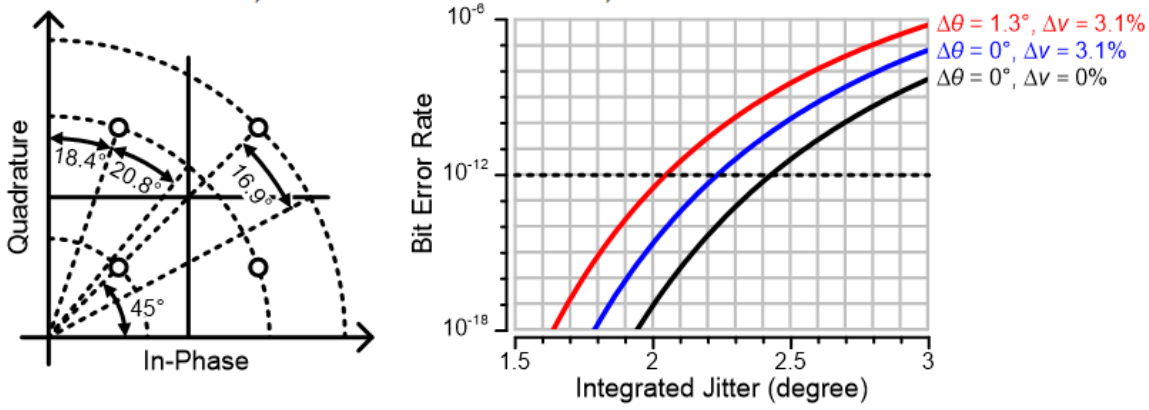


Figure 3.5 *Bit error rate (BER) equation together with phase error tolerance shown on the constellation plot deriving the corresponding jitter requirement calculation of 16-QAM*

3.1.3 Other Interferences

Other than the jitter requirement, there are a couple of other requirements that need to be specified. Previously, we mentioned that the second passband is available because differential signaling is used to suppress the 2nd-order harmonic from the first passband. Ideally with 50% duty cycle of the 3GHz carriers, the 2nd-order harmonic will be completely eliminated. In reality, the duty cycle could deviate from 50% and induces additional adjacent-band interference from the 2nd-order harmonic. Therefore, the carrier duty cycle error needs to be within ±1% to have the additional interference 30dB smaller than the desired signal. Another requirement is the co-band interference, which is also known as crosstalk. Again, we wish the interference to be 30dB less than the desired signal, and thus we need the crosstalk at 6GHz to be smaller than -30dB. For memory interface, far-end crosstalk (FEXT) is of more concern than near-end crosstalk (NEXT) because data signals are always transmitting in the same direction during either the reading or writing stage. On a 2" FR-4 memory bus with line pitch of 6mil, the FEXT is below -30dB at 6GHz, which meets our requirement.

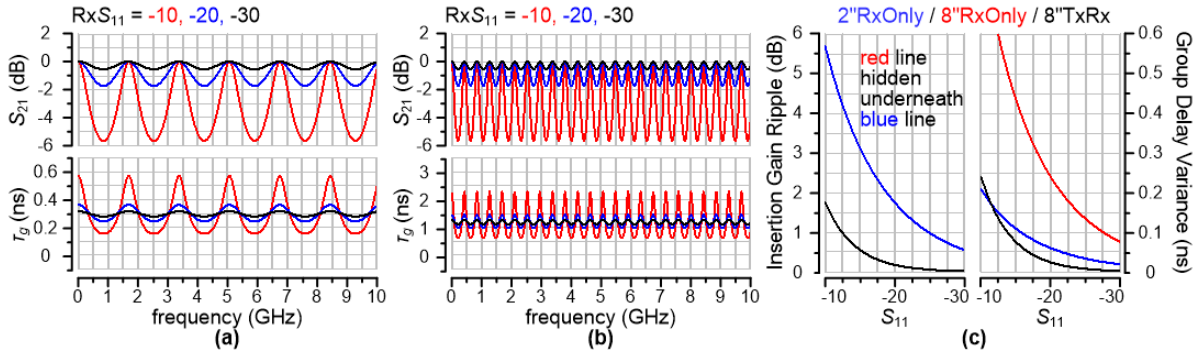


Figure 3.6 *Insertion gain (S_{21}) and group delay (τ_g) of an ideal (a) 2" (b) 8" FR-4 transmission line vs. frequency; (c) insertion gain ripple and group delay variance vs. channel impedance matching in terms of return loss (S_{11})*

Additional inter-symbol interference could be induced by impedance mismatching. Given non-perfectly matched input impedance at receiving end (RxS_{11} = -10/-20/-30 dB) and unmatched output impedance at transmitting end (~5 KΩ), the insertion gain and group delay vary periodically over frequency on 2'' FR-4 transmission line as shown in Figure 3.6(a). As mentioned before, the insertion gain variance should be less than 1 dB to ensure the signal integrity. Also, group delay variance is required to be less than ±0.1 UI, which is ±100 ps with symbol rate of 1 GHz). Therefore, Rx S_{11} of -20 dB is necessary for 2'' FR-4 transmission lines. From 2" to 8" ideal FR-4 transmission line, insertion gain variance remains the same but group delay variance increases about 4 times (Figure 3.6(b)). As a result, it becomes very difficult to meet the group delay variance requirement with one-end matching (Rx S_{11} < -30 dB), and thus both Tx and Rx needs to be matched for 8'' FR-4 transmission line. As shown in Figure 3.6(b), with Tx and Rx S_{11} both less than -15 dB (or return loss >15 dB), both group delay variance and insertion gain variance requirements are satisfied.

3.2. Transmitter Circuit Design

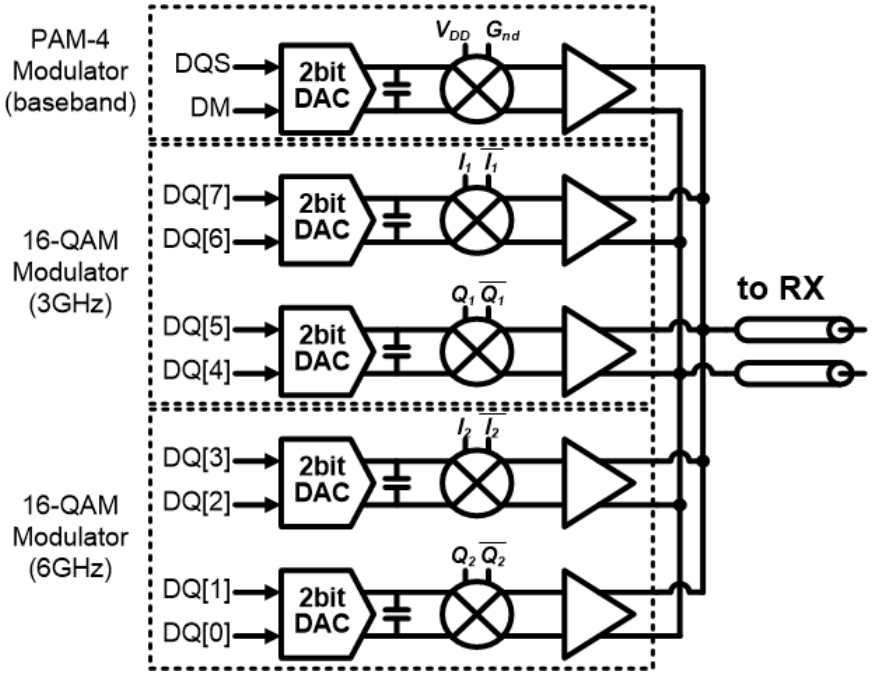


Figure 3.7 Block diagram of the tri-band PAM-4 / 16-QAM transmitter

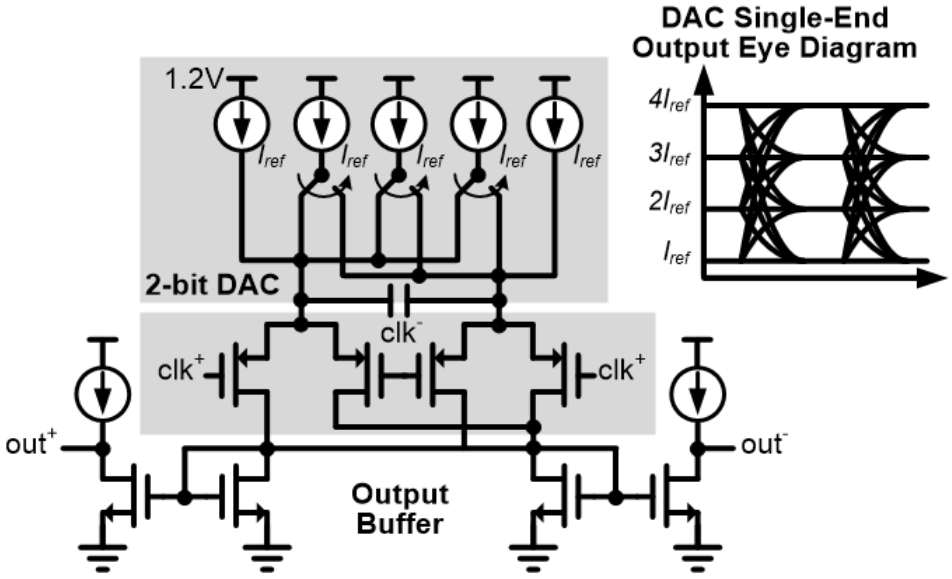


Figure 3.8 Circuit schematic of one modulation path in the tri-band transceiver

The transmitter of tri-band PAM-4 / 16-QAM signaling is composed of five identical modulation paths, each with one 2-bit DAC, one mixer, and one output buffer (Figure 3.7). Since differential signaling is adopted to suppress 2^{nd}-order harmonic, all the circuits are fully differential (Figure 3.8). The 2-bit DAC is designed with a minimum current flow of I_{ref} at each end to ensure the output buffer can operate up to 6GHz. Also, a capacitor is inserted at the output of the DAC to slow down signal transition and suppress in-band interference. The clock inputs of the baseband mixer are tied to logic high and logic low so that the output signal remains at the baseband. The baseband mixer is not necessary but added to match latency of each frequency band. If the latency from modulation to demodulation at each frequency band is the same, transmitted data will remain synchronous at the receiving end and thus de-skew circuitry (e.g. DLL) will not be required for data recovery. For common channel media used for memory interface (e.g. FR-4, silicon interposer, TSV, InFO), group delay variance is negligible ($\ll$0.1UI) over the three frequency bands. Therefore, as long as the latency of modulation and demodulation paths matches, the total latency matches. The clock inputs of the other four mixers are separately connected to four carriers generated from the dual-band carrier generator, which will be discussed in the next section. Finally, the output buffer is simply a current mirror with feedforward bias circuit to subtract the common mode and output only the differential mode. For output impedance matching, an optional matching circuit is inserted, which can be turned on and off according to channel condition. For short-reach application with less stringent impedance matching requirement, the circuit can be turned off to reduce power consumption and improve energy efficiency. Detail of the matching circuit will be discussed with the input buffer design at the receiving end.

3.3. Dual-Band Carrier Generator Circuit Design

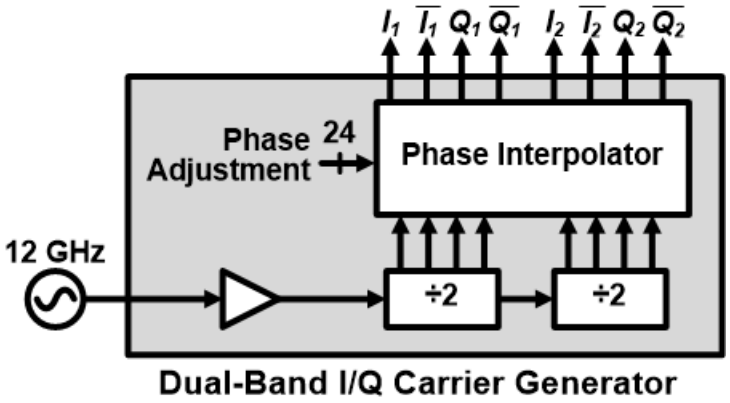


Figure 3.9 Block diagram of the dual-band I/Q carrier generator

The carrier generator, which is shared among four lanes of the tri-band transceiver, provides the in-phase and quadrature carriers at 3GHz and 6GHz. In order to maintain orthogonality after demodulation, the carriers at the receiving end must stay synchronized with the propagating signal, and thus the carrier generator must be able to adjust the carrier delay. The carrier generator is composed of one 12GHz clock buffer, two dividers (÷2), and four phase interpolators (Figure 3.9). The clock buffer first amplifies a 12GHz clock from either a phase-locked loop or an off-chip clock sources. The 12GHz clock buffer is then followed by the first divider. The first divide-by-2 circuit generates both in-phase and quadrature carriers phase at 6GHz. These in-phase and quadrature carriers at 6GHz are then buffered to drive the transmitter mixer. To drive the receiver mixer, the carriers are delayed to synchronize with the received signal. The carrier delay is imposed by the phase interpolator and thus two independent phase interpolator are used for the two in-phase and quadrature carriers. One of the two outputs of the first divider is applied to the input of another divider that generates the in-phase and quadrature carriers at 3GHz. Similarly to the 6GHz carriers, the 3GHz carriers are buffered to drive the

transmitter mixers and phase interpolated by another pair of phase interpolators to drives the receiver mixers. The carrier generator adopts current-mode logic (CML) topology, which provides better supply noise rejection, better duty cycle accuracy and less I/Q mismatches compared to CMOS logic topology. The CML topology also can provide appropriate dc bias for mixers at both the transmitting and receiving ends. The divider has two CML D latches in a negative feedback loop. Ideally, that will provide 50% duty cycle carrier and zero I/Q mismatch. However, layout and random mismatches lead to duty cycle error and I/Q mismatch. Therefore, the circuits are laid out carefully to reduce systematic mismatches, and the random mismatch caused by local variation is well controlled by device sizing.

The adjustable delay required for carriers at the receiving end is realized by interpolating the in-phase and quadrature carriers with a tail-current summation phase interpolator. The phase interpolator produces a weighted sum of two input carriers with quadrature phase difference in this case. The phase interpolator interpolates between the in-phase and quadrature carriers and provides a clock phase in between. A total of 90 degree phase rotation can be achieved. This is equivalent to 41.6ps delay range for 6GHz carriers and 83.2ps delay range for 3GHz carriers. By controlling the tail current weight, the output clock phase, and thus delay, can be controlled. In this design, forty identical tail current units and 6 control pins are used so that a resolution of 1.2ps for 6GHz and a resolution of 2.4ps for 3GHz can be achieved. In-phase and quadrature clocks are delayed separately by two identical phase interpolator but with inputs with swapped polarities. In order to improve the linearity of the phase interpolator, the input and output time constant (slew rate) of the phase interpolator needs to be carefully controlled. The time constant should not be too fast for phase mixing quality.

3.4. Receiver Circuit Design

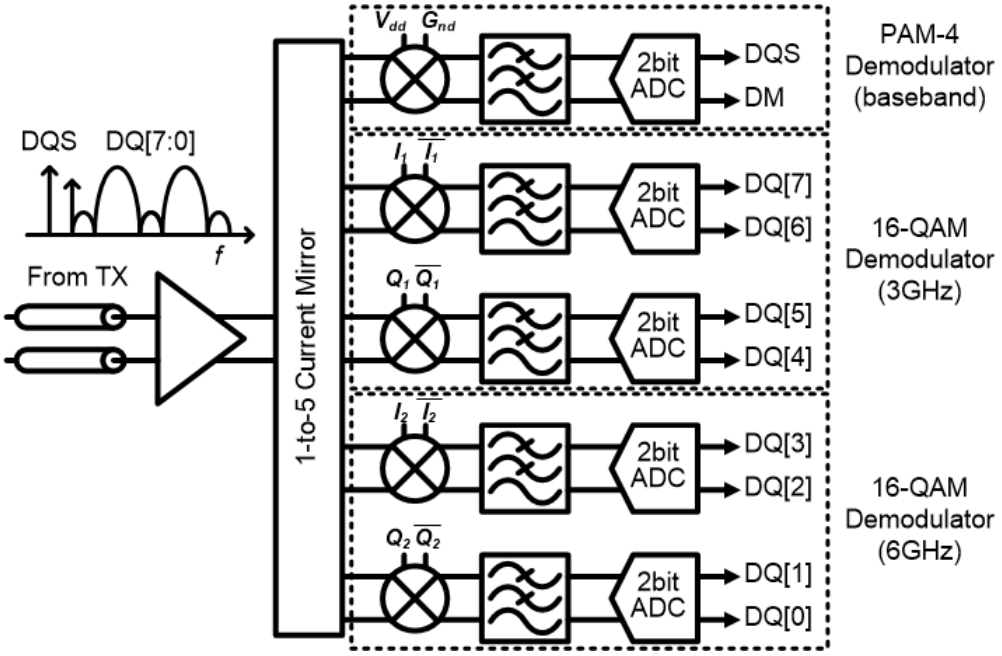


Figure 3.10 Block diagram of the tri-band PAM-4 / 16-QAM receiver

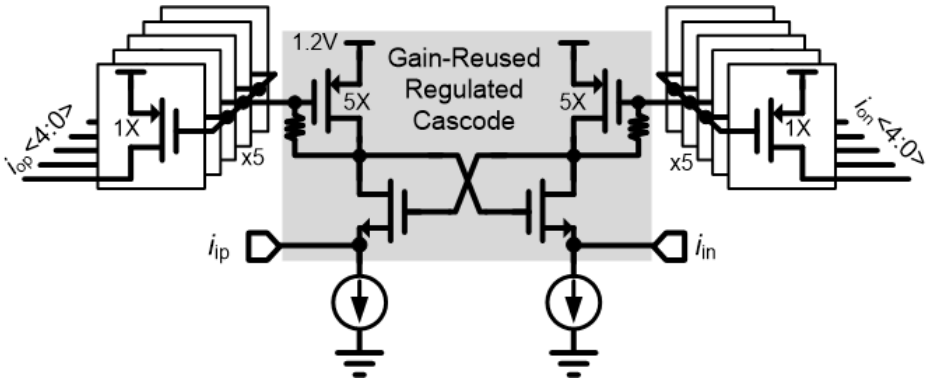


Figure 3.11 Circuit schematic of the gain-reused regulated cascode input buffer

Similar to the transmitter, we can find five identical demodulation paths in the receiver of tri-band PAM-4 / 16-QAM signaling (Figure 3.10). Before the demodulation paths, there is an input buffer that includes a 1-to-5 current mirror to distribute the received signal. The input

buffer also provides impedance matching for both the transmitter and receiver. Within the input buffer, a gain-reduced regulated cascode structure is used and its differential input impedance is determined by transconductance difference of the NMOS and PMOS, which is equivalent to ($2/g_{mn} - 2/g_{mp}$) at low frequency (Figure 3.11). Therefore, with proper sizing of those transistors, the differential input impedance can be as low as 100Ω even using a very small bias current. However, the circuit could oscillate when g_{mn} is larger than g_{mp} and induce a negative input impedance. That is possible when the bias current and the transconductances are too small. Let us say the transconductance variation of is ±2.5%. Then, negative input impedance is possible when the design values of g_{mn} and g_{mp} are lower than 1/1050 and 1/1000, respectively. Besides the stability problem, a small bias current could also cause impedance mismatch at high frequency. With parasitic capacitance (C_p) at the gate of PMOS, the equation can be modified as ($2/g_{mn} - 2/(g_{mp}+j\omega C_p)$) and thus the input impedance will start increasing beyond a corner frequency. With a larger bias current and hence larger transconductances, the corner frequency can be higher and the impedance matching condition can sustain within a wider frequency range. There is a circuit technique that can help to extend the corner frequency without increasing the bias current. By inserting a small resistor between the gate and drain of PMOS, the effective g_{mp} will reduce at high frequency, which forms an inductance to balance the parasitic capacitance. In this work, the resistor helps to improve the input return loss, S_{11}, by 12dB at 6GHz (Figure 3.12). Note that the inductance and capacitance could resonate and destabilize the input buffer, and thus the variation of the resistor also needs to be well controlled. An additional switch transistor is inserted between the NMOS and the bias current source at each side so that we can turn off the matching circuit if not used. While turned on, the switch transistors have a small but not zero

resistance, so the transconductance difference needs to be smaller in order to maintain differential input impedance of 100Ω.

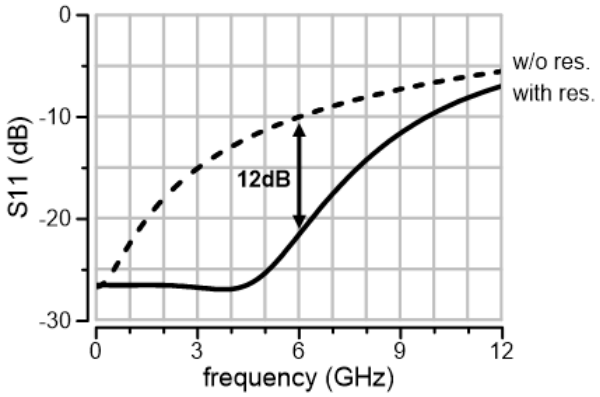


Figure 3.12 Simulated frequency response (S_{11}) of the gain-reused regulated cascode input buffer

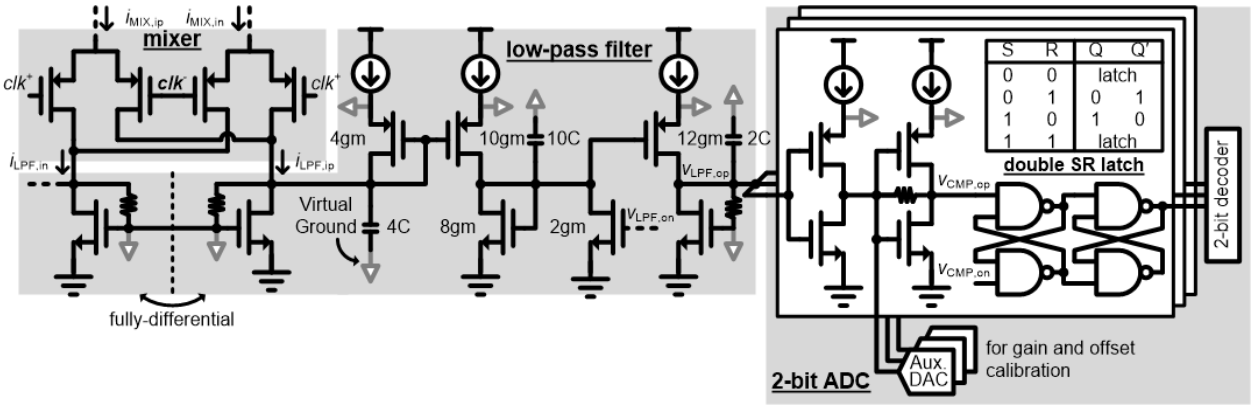


Figure 3.13 Circuit schematic of one demodulation path in the tri-band transceiver

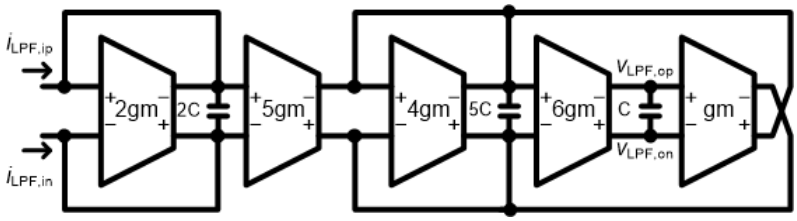


Figure 3.14 Block diagram of the 3rd-order Bessel Gm-C low-pass filter

With the 1-to-5 current mirror inside the input buffer, each demodulation path received one copy of the input signal. Then the signal is down-converted with a mixer, reconstructed with a low-pass filter and finally digitized with a 2-bit ADC. Again, all the circuits are fully differential to suppress 2^{nd}-order harmonic (Figure 3.13) and the baseband mixer is retained for latency matching. According to the system analysis performed in Section 3.1.1, the low-pass filter is built as 3^{rd}-order structure with 30dB rejection at 3GHz offset. To maximize the output eye diagram, the transfer function of the low-pass filter is designed as Bessel function with linear phase and maximally flat group delay, which has no ringing or peaking in step response. For a Bessel function with 30dB rejection at 3GHz offset, the 3dB bandwidth is about 700MHz. To implement such a high bandwidth filter, Gm-C architecture is adopted for its low power consumption while compromising on linearity (Figure 3.14). Also, the three Gm stages in the middle share one bias current source in order to further reduce power consumption. Finally, the 2-bit ADC is composed of three parallel comparators. Inside each comparator, the first two stages are used as a cherry hopper preamplifier. Between the first and the second stage, a reference current is injected from an auxiliary DAC, which is used for threshold adjustment and offset calibration. After amplification, two cascaded SR latches convert the differential analog signal into a single-ended digital bit stream. With the two cascaded SR latches, the output state change only when the differential input signal crosses threshold at both sides, which avoid change of duty cycle due to common-mode mismatch between analog and digital stages. Then, the three latch output bits are mapped back to two bits with a 2-bit decoder.

3.5. Transceiver Integration with Built-In Self-Test (BIST)

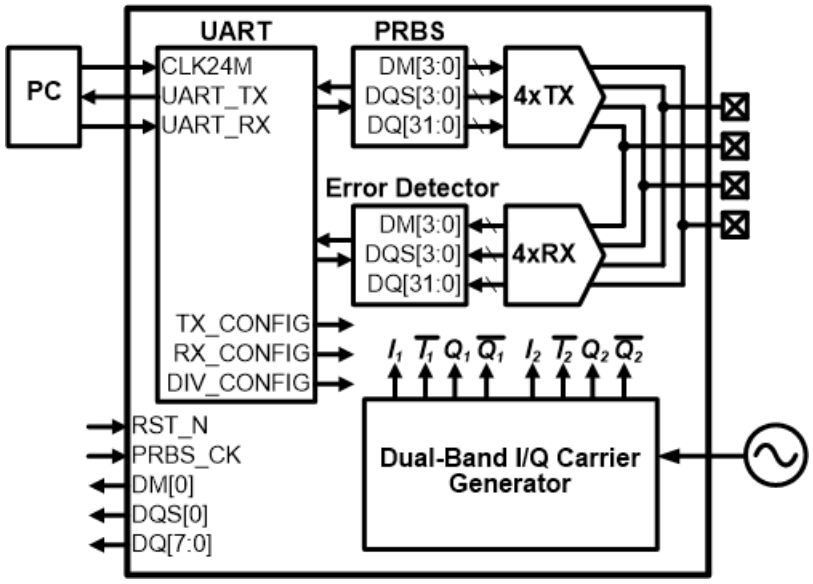


Figure 3.15 *Illustration of the 4-lane transceiver testing environment with built-in self-tester (BIST) and UART interface*

Combining four transmitters, four receivers, and one carrier generator, we obtain a 4-lane transceiver that achieves a total data rate of 40Gb/s (Figure 3.15). However, during measurement, the 4-lane transceiver requires a testing pattern of 40 bits with symbol rate of 1GBaud, which is very difficult to generate from regular testing instruments or general-purpose FPGA boards. Therefore, we choose to implement a built-in self-testing (BIST) machine integrated with the 4-lane transceiver. The BIST is composed of a 32-bit PRBS generator and a 32-bit error detector. PRBS generators are usually implemented with linear-feedback shift registers (LFSRs). In order to verify BER less than 10^{-12}, the LFSR's repeat cycle needs to be larger than 10^{12}, which is close to 2^{40}, which means the length of the LFSR needs to be at least 40. Also, for each of the 32 independent PRBSs, we need one primitive feedback polynomial but we cannot find 32 primitive

feedback polynomials with a length of 40, which means some of the 32 LFSRs need to have lengths longer than 40. Therefore, it is not efficient in terms of power and area to implement a 32-bit PRBS generator with conventional LFSR. In this work, the 32-bit PRBS generator is composed of only two reservedly combined LFSRs each with lengths of 32 and 33, respectively (Figure 3.16). Reverse combination and length difference are two keys to efficient multi-bit PRBS implementation. If the two LFSR are combined in the same direction, then the output PRBSs will not be independent but identical with time shift. If the two LFSR have the same length of 40, the repeat cycle will be $(2^{40} - 1)$, which is enough for BER < 10^{-12} but much smaller than $(2^{32}\text{-}1) \times (2^{33}\text{-}1)$ when having lengths of 32 and 33. Two LFSR with different lengths of 32 and 33 are apparently more area efficient than those with the same length of 40.

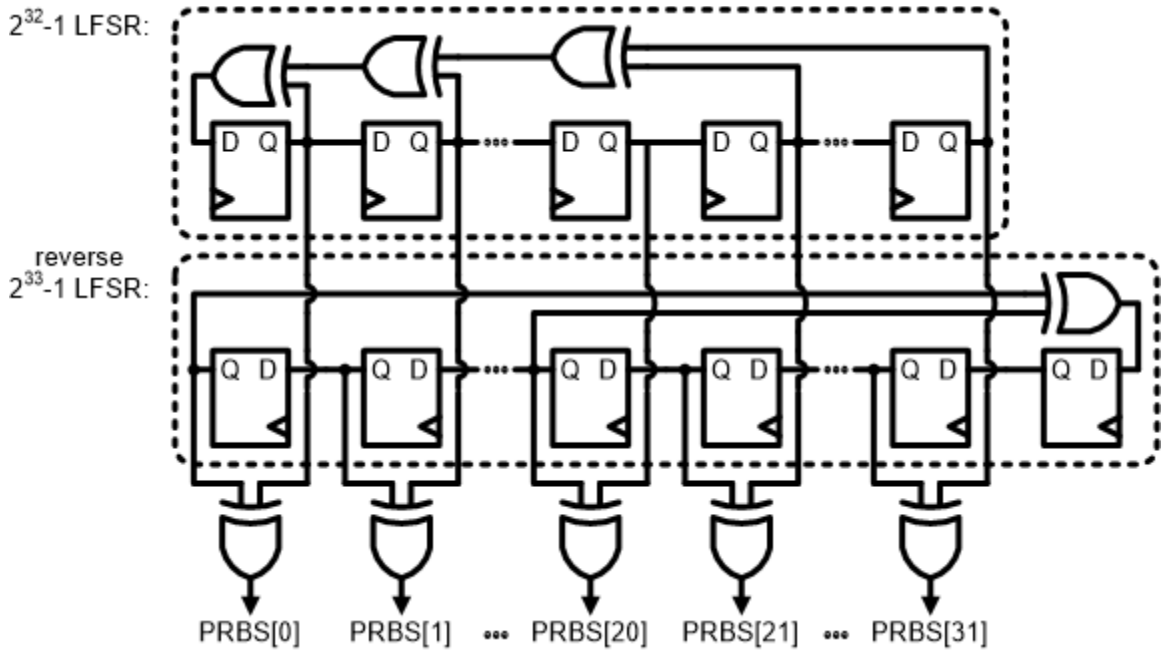


Figure 3.16 *32-bit PRBS generator implemented with reversely combined linear feedback shift registers (LFSRs)*

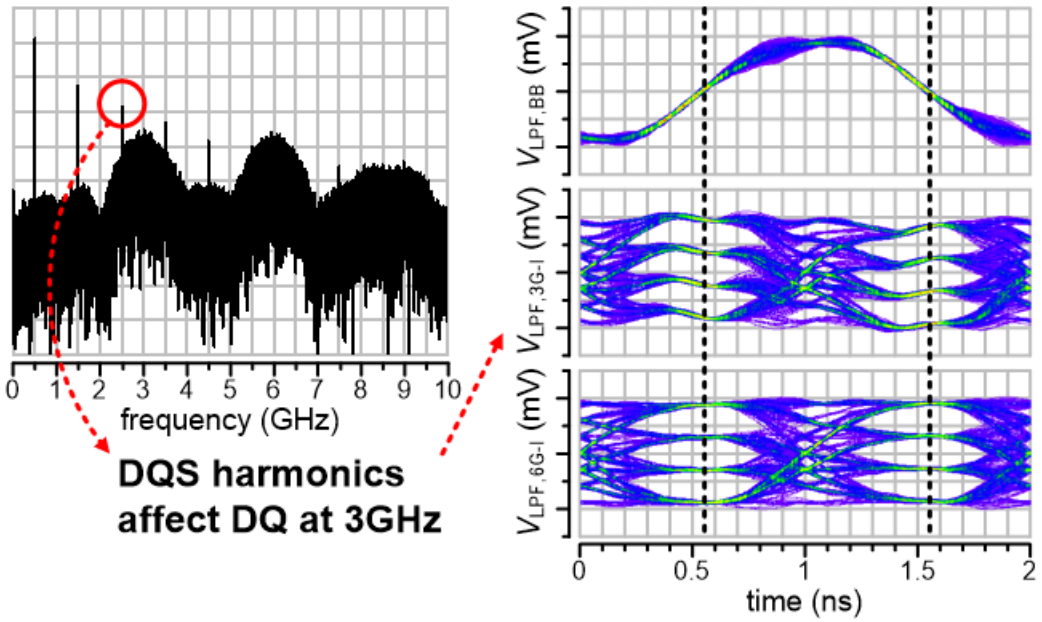


Figure 3.17 *Simulated output spectrum of the tri-band transmitter and eye diagrams of demodulated output signals of the tri-band receiver indicating strong harmonics from baseband degrade 3GHz band slightly*

Similar to the 32-bit PRBS generators, the 32-bit error detector also has its own design difficulty. Since DDR3 memory interface, a retiming technique using delay-lock loop (DLL) is adopted to synchronize multiple bits of received data. Physical channel difference due to PCB routing and PVT condition induces delay variation between data and clock signals. Before DDR3, people simply bundle every 8 bits of data signal (8 DQs) and assign one clock signal (1 DQS) so that the delay variation within each bundle can be tolerable. However, since DDR3 achieves a data rate up to 2.133GHz, even the delay variation within each bundle could cause error bit during data recovery. As a result, DLL is used to adjust the delay of and synchronize every DQ within a bundle so that the assigned DQS can correctly recover received data. Nevertheless, the introduction DLL creates circuit overhead and limits the reduction of power and area efficiency. In this work, we utilize a characteristic of multi-band signaling to avoid the necessity of DLL. As mentioned before, the delay variation within the 10 modulated bit streams of each lane is

negligible because they share the same physical channel. Therefore, if we simply assign one of the 10 bit streams to be the DQS, then we can directly use the demodulated DQS as the clock for data recovery. Here we modulate the DQS at baseband together with the data mask (DM), a low-speed signal bundled with 8 DQs and 1 DQS in DDR series memory. However, since DQS is a clock signal and its harmonics are more concentrated in spectrum compared to those of random data, the adjacent-band interference is more severe in time domain (Figure 3.17). As a result, the baseband signal is slightly turned down in order to reduce interference to 3GHz signal. Finally, the 32-bit error detector consists of 4 sets of 8-bit error detector and each is triggered with its assigned DQS. Also, each 8-bit error detector required one 32-bit PRBS generator triggered by the DQS so that we can compare the received data with the PRBS output. The comparison result can be accessed by personal computer or notebook via an integrated UART interface. The UART interface can operate at speeds up to 3MBaud and it has extra register file that is assigned to control pins for the transmitter, receiver and carrier generator.

CHAPTER 4 SILICON RESULTS AND CONCLUSION

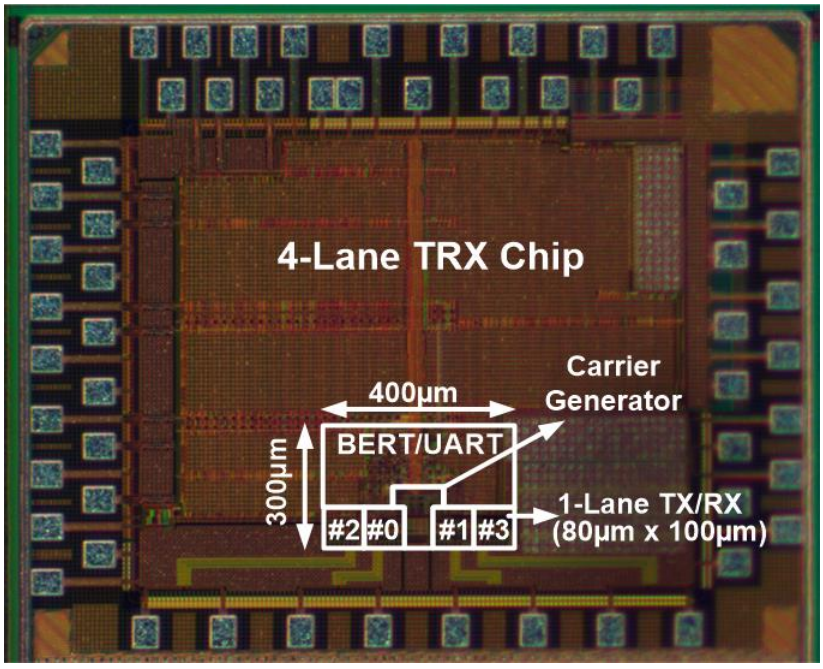


Figure 4.1 Silicon die photo of the 4-lane tri-band transceiver implemented in TSMC 28nm CMOS technology

The 4-lane tri-band transceiver with built-in self-tester is implemented in TSMC 28nm HPC technology (Figure 4.1). The entire design is pad limited and thus, even though the chip size is as large as 1.7×1.5mm^2, the core circuit takes only 400×300µm2. Splitting the chip area taken up by the shared carrier generator, the transceiver occupies 100×100µm^2/lane, and the BER tester including UART interface takes 400×200µm^2. Using chip-on-board (COB) packaging with wire bonding, two of the 4-lane transceivers are installed on a test board and interconnected with a 2” dense FR-4 differential bus of 4 lanes (Figure 4.2). The line pitch of the bus is 6mil and the

channel attenuation at 6GHz is about 6dB. With the channel condition, we first need to perform phase and gain calibration in order to correct received signal for data recovery. After calibration, we can see the measured output eye diagrams remain wide open because of self-equalization and stay aligned with negligible delay difference (Figure 4.3(a)). Putting the transmitted and received signals together on an oscilloscope, we find the delay from transmitter to receiver is about 1ns (Figure 4.3(b)). Connecting the output of transmitter to a spectrum analyzer, we can identify one tone of DQS and two lobes of DQ at 3 and 6GHz from the measured output spectrum (Figure 4.4(a)). Due to channel attenuation, the signal at 6GHz is at lower power level and thus needs to be strengthened at the transmitting end in order to maintain BER < 10^{-12}. Eventually, the transmitter power consumption increases to 6.4mW or 0.16pJ/b for 6dB attenuation at 6GHz (Figure 4.4(b)). At the other end, the 4-lane receiver totally consumes 18.8mW. Including 13.4mW from the carrier generation, the total power consumption of the 4-lane transceiver is 38mW and the energy efficiency is 0.95pJ/b considering the total data rate is 40Gb/s (Figure 4.5).

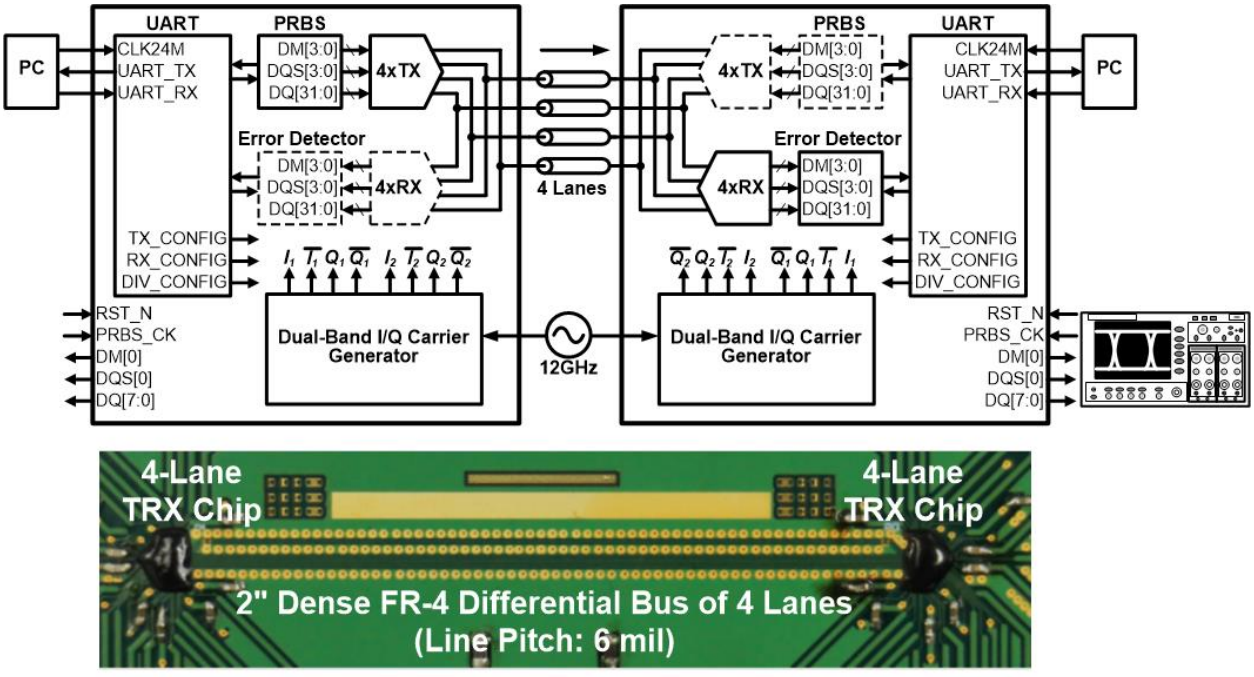


Figure 4.2 Illustration of the transceiver testing environment and a picture of the test board

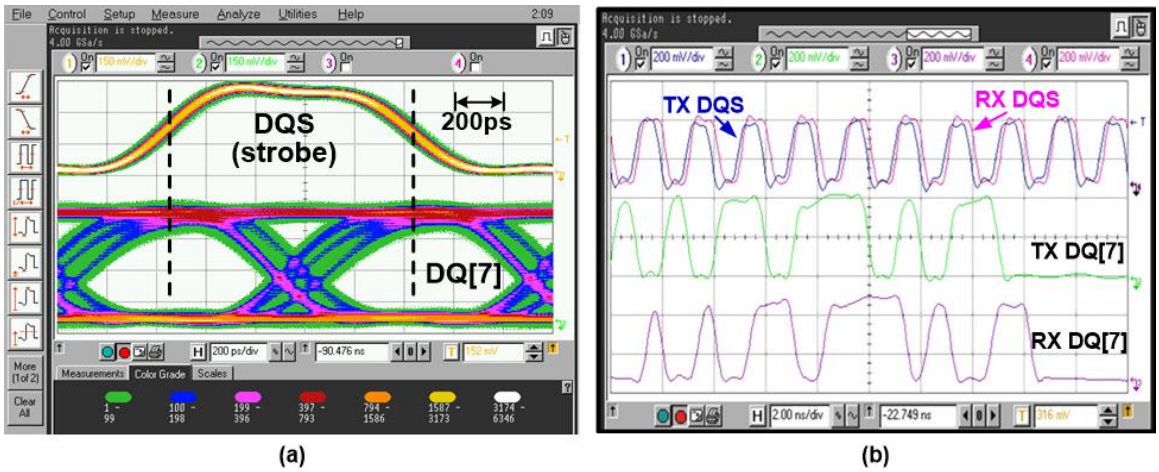


Figure 4.3 *Measured (a) eye diagrams and (b) real-time waveforms of demodulated signals*

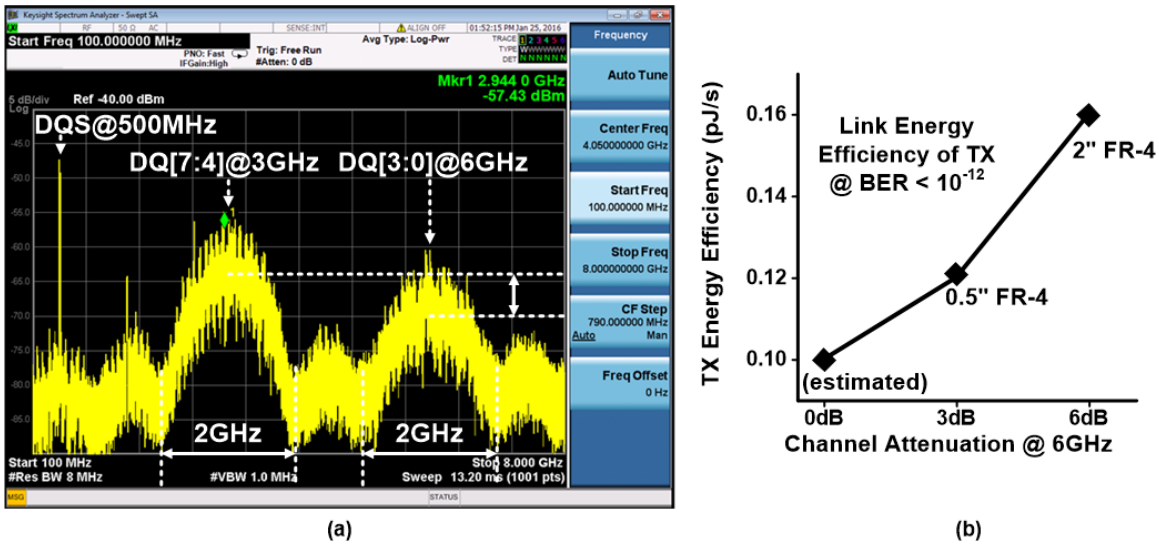
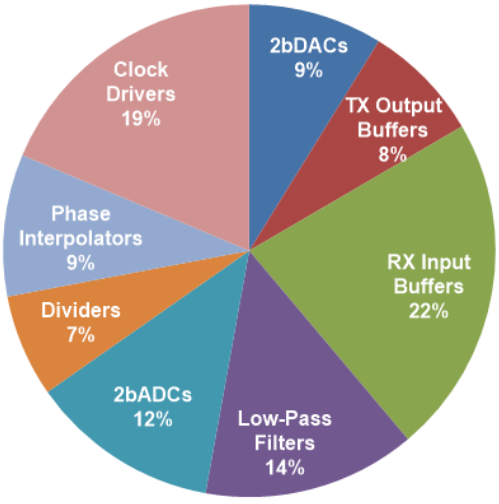


Figure 4.4 *Measured (a) channel spectrum of tri-band signaling and (b) the plot of Tx energy efficiency vs. channel attenuation*

In summary, we have implemented a tri-band transceiver with four parallel lanes in 28nm CMOS technology. The tri-band transceiver is tolerant to spectral notches of multi-drop buses by spectrally divided signaling and further extends communication bandwidth. Additionally, this transceiver is also immune to inter-symbol interference caused by channel attenuation without additional equalization circuitry as to the self-equalized double sideband signaling. To realize the

total data rate of 40Gb/s, PAM-4 and 16-QAM are used at the baseband and 3/6GHz bands, respectively, to carry 10 parallel bit streams at 1GHz symbol rate via each lane of the transceivers. These ten parallel bit streams share the same physical channel to minimize the time skew among them. In view of this, the strobe signal, DQS, is assigned to one of the ten bits for data recovery at the receiving end without any de-skew circuitry. Under 6dB attenuation at 6GHz on a 2" dense FR-4 differential bus (line pitch of 6mil), the Tx consumes only 1.6mW/lane. Together with 4.7mW/lane of the Rx and 13.4mW of the carrier generator to be shared among all lanes, the total power consumption is 38mW and the average energy efficiency of the 40Gb/s bus is 0.95pJ/b. Compared with prior arts, the proposed design achieves not only better energy efficiency but also substantial size advantage (0.01mm^2/lane including the carrier generator). This transceiver realizes a total data rate of 40Gb/s with BER < 10^{-12}. Moreover, this tri-band architecture can be scaled in the frequency domain for further increasing the data throughput without increasing the symbol rate, which enables a new design dimension with more compact size and significantly improved energy efficiency for future memory interfaces.



	Avg. Power	#	Total Power
TX	**1.6mW**	**4**	**6.4mW**
2bDAC	0.17mW	20	3.4mW
TX Mixer	-	20	-
Output Buffer	0.15mW	20	3mW

	Avg. Power	#	Total Power
RX	**4.7mW**	**4**	**18.8mW**
Input Buffer	2.1mW	4	8.6mW
RX Mixer	-	20	-
Low-Pass Filter	0.27mW	20	5.4mW
2bADC	0.24mW	20	4.8mW

	Avg. Power	#	Total Power
Carrier Generator	**13.4mW**	**1**	**13.4mW**
Divider	1.3mW	2	2.6mW
Phase Interpolator	0.9mW	4	3.6mW
Clock Driver	1.8mW	4	7.2mW

Figure 4.5 Power breakdown of the 40Gb/s 4-lane tri-band transceiver

Table 4.1 Benchmarking with state-of-the-art

	JSSC'12 [1]	*ISSCC'12* [2]	*JSSC'12* [3]	*ISSCC'15* [4]	**This Work**
Technology	40nm CMOS	90nm CMOS	65nm CMOS	40nm CMOS	28nm CMOS
Supply	1.0V	1 .25V	1.0 V	0.9 V	1.2V
Data Rate	12.8 Gb/s/pin	8 Gb/s/pin	8.4 Gb/s/pin	7.5 Gb/s/lane	**10 Gb/s/lane**
Total Power	64 mW/pin	32mW	21mW	7.5 mW	**9.5 mW/lane**
Energy/Bit	5pJ	4pJ	2.5pJ	1pJ	**0.95pJ**
Area	0.17 mm^2 (per pin)	0.23mm^2	0.15mm^2	0.015 mm^2	**0.01mm^2 (per lane)**
Channel	3" FR-4	2" FR-4	4" FR-4	12" FR-4	2" FR-4
Signaling	NRZ / CTLE + 1-tap DFE	NRZ/CTLE	BB+RF	NRZ/QPSK	PAM-4 / 16-QAM
BER	< 10^{-12}	< 10^{-12}	< 10^{-12}	< 10^{-12}	< 10^{-12}

In summary, we have implemented a tri-band transceiver with four parallel lanes in 28nm CMOS technology. The tri-band transceiver is tolerant to spectral notches of multi-drop buses by spectrally divided signaling and further extends communication bandwidth. Additionally, this transceiver is also immune to inter-symbol interference caused by channel attenuation without additional equalization circuitry as to the self-equalized double sideband signaling. To realize the total data rate of 40Gb/s, PAM-4 and 16-QAM are used at the baseband and 3/6GHz bands, respectively, to carry 10 parallel bit streams at 1GHz symbol rate via each lane of the transceivers. These ten parallel bit streams share the same physical channel to minimize the time skew among them. In view of this, the strobe signal, DQS, is assigned to one of the ten bits for data recovery at the receiving end without any de-skew circuitry. Under 6dB attenuation at 6GHz on a 2" dense FR-4 differential bus (line pitch of 6mil), the Tx consumes only 1.6mW/lane. Together with 4.7mW/lane of the Rx and 13.4mW of the carrier generator to be shared among all lanes, the total power consumption is 38mW and the average energy efficiency of the 40Gb/s bus is 0.95pJ/b. Compared with prior arts, the proposed design achieves not only better energy

efficiency but also substantial size advantage (0.01mm^2/lane including the carrier generator). This transceiver realizes a total data rate of 40Gb/s with BER < 10^{-12}. Moreover, this tri-band architecture can be scaled in the frequency domain for further increasing the data throughput without increasing the symbol rate, which enables a new design dimension with more compact size and significantly improved energy efficiency for future memory interfaces.

References:

[1] A. Amirkhany, *et al*., "A 12.8-Gb/s/link Tri-Modal Single-Ended Memory Interface," *IEEE J. Solid-State Circuits*, vol. 47, no. 4, pp. 911-915, Apr. 2012.

[2] Y. Kim, et al., "An 8Gb/s Quad-Skew-Cancelling Parallel Transceiver in 90nm CMOS for High-Speed DRAM Interface," ISSCC *Dig. Tech. Papers*, pp. 50-51, Feb. 2012.

[3] G. Byun, *et al*., "An Energy-Efficient and High-Speed Mobile Memory I/O Interface Using Simultaneous Bi-Directional Dual (Base+RF)-Band Signaling," *IEEE J. Solid-State Circuits*, vol. 47, no. 1, pp. 117-130, Jan. 2012.

[4] K. Gharibdoust, *et al*., "A 7.5mW 7.5Gb/s Mixed NRZ/Multi-Tone Serial-Data Transceiver for Multi-Drop Memory Interfaces in 40nm CMOS," *ISSCC Dig. Tech. Papers*, pp. 180-181, Feb. 2015.