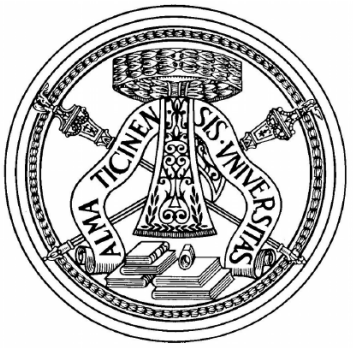


Università degli Studi di Pavia
Facoltà di Ingegneria
Dipartimento di Ingegneria Industriale e
dell'Informazione

Dottorato di Ricerca in Microelettronica
XXIV ciclo

Low-noise Design Techniques for High-speed Optical Receiver Front-ends

Tutore:
Chiar.mo Prof. Francesco Svelto

Coordinatore:
Chiar.mo Prof. Franco Maloberti

Tesi di Dottorato di
Dan Li

Contents

Chapter 1

Introduction

In the last 150 years, human beings have been constantly pursuing new means of communication: telegraph, radio, telephone, TV, cellphone and etc. Finally, at the end of 20th century, marked by the proliferation of personal computer and advent of the Internet, we entered the information age or WWW era and start to live a connected life.

Early Internet applications like web browsing, search, email, BBS (Bulletin Board System) and IM (instant message) ask for relatively low communication bandwidth. From the middle 2010s, Internet gradually evolves into so-called web 2.0 era featuring much more user-concentrated activity. Social network like Facebook, video-sharing site like YouTube suddenly became essential parts of daily life, which all generate huge amount of Internet data traffic. Meanwhile, thanks to the rapidness and richness of content, Internet also becomes the entertainment and media center. Example are that 1) TV system starts to be designed based on Internet rather than the traditional TV network; 2) Internet function is at least as equal as, if not more important than, the communication function of a mobile phone. In the business world, Internet has evolved from a pure propaganda tool (website and advertisement) into an essential part or even a means of business itself. Examples are like E-commerce (EBay, Amazon), video-conferencing (WebEx) and cloud computing (iCloud, Dropbox).

Fig. 1.1 Internet application evolution

All those new application essentially ask for one thing: higher communication bandwidth. Because of the aforementioned Internet centered media spread in recent years, data traffic in Internet grows exponentially. Meanwhile, large enterprises have to build their own data centers to provide the service and content and the interconnection bandwidth within the data center is also critical to provide satisfied user experience. Indeed, this trend can be seen in Fig. 1.2.

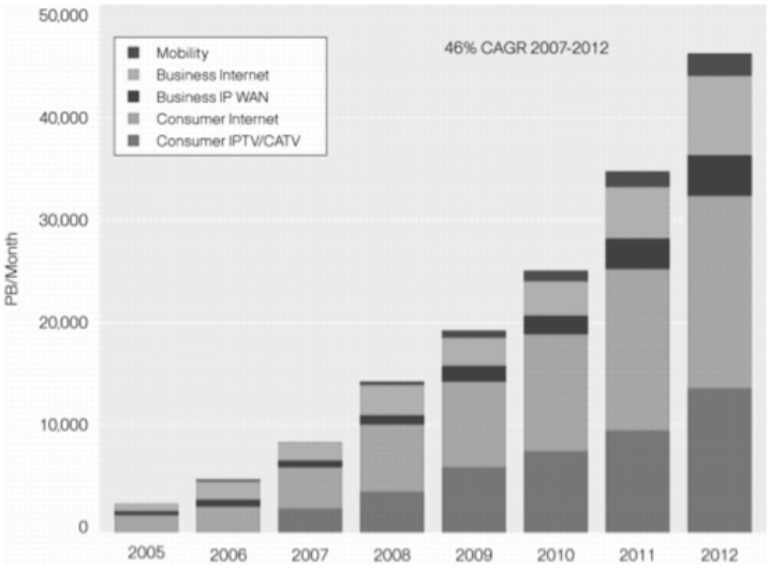


Fig. 1.2 Internet data traffic [1]

This ever-increasing bandwidth requirement is continually calling for faster means of data transmission. From the first transatlantic telegraph cable laid in 1858, electrical cable based transmission has well served for more than one and half century. However, electrical wire transmission is reaching limit with data rate approaching 10 Gb/s. Even with sophisticated pre-emphasis and equalization techniques, 10 Gb/s electrical wire signaling is limited to less than 1 meter backplane and 7 m copper cable [2], let alone the significant power consumption and system complexity. Therefore, electrical wiring transmission is a poor candidate for high-speed transmission in a reasonable physical distance, e.g. 100m.

While researchers are actively looking for new means of data transmission, optical fiber communication is by far the most promising one. Compared with electrical wiring, optical fiber has the following advantages: lighter, thinner and most importantly, high bandwidth and low loss [3]. What's more, high parallelism by means of wavelength-division multiplexing (WDM) can significantly scales up the aggregate bandwidth, making it several order of magnitude faster than electrical communication. The down side of optical communication is the cost, rendering it traditionally more suitable to build the backbone network. Yet, the advancements on silicon-based photonics, a.k.a. silicon photonics are changing this picture by promising much cheaper and scaled optics built on silicon. This will enable optical communication conquer more territory from electrical wiring communication and fundamentally solve the bandwidth problem.

Fiber optical communication became technically available in 1960s and began mass commercialization in early 1980s in telecommunication. Since Internet boomed in the 1990s, it had gained significant momentum, reflected by Telecom (telecommunication) standard SONET/SDH, from 50 Mb/s in OC-1 to 40 Gb/s in OC-768. In the mean time, Datacom (data communication in computer network) standards, like Ethernet, Infiniband, Fibre Channel and etc. had also undergone very fast expansion. Take a recent Ethernet standard, issued by IEEE 802.3 workgroup, for example: it has evolved from 10 Mb/s (10-BASE) to 100 Gb/s

(100G-BASE) by a factor of 10000. Both trends are reflected in Fig. 1.3, where core network stands for Telecom and Sever I/O stands for Datacom. More aggressive Telecom and Datacom standards are still underway to cope with the insatiable need for bandwidth and custom-designed optical transport systems have reached data rate of more than 100 Tb/s [40].

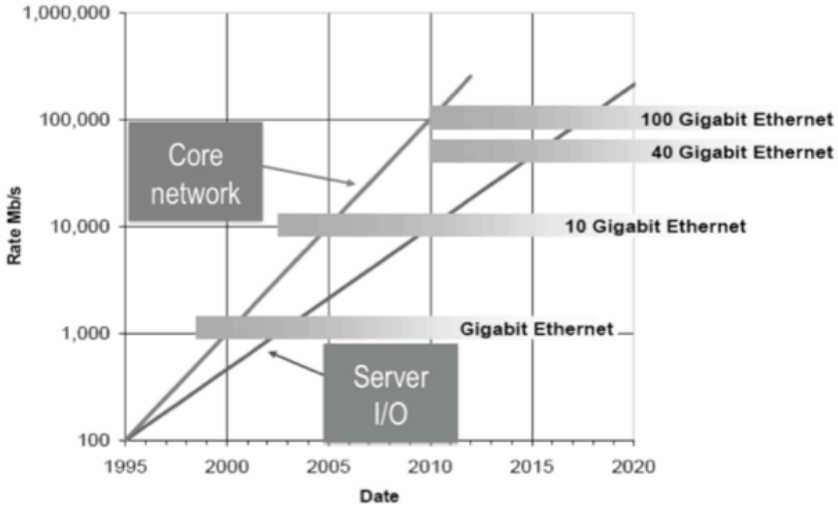


Fig. 1.3 Evolution of optical communication standard [4]

The rapid evolution of fiber optical communication draws thrilling blueprint for future ultra fast communication system. However, while data transmission within the optical network seems less a problem, the optical pulses bits eventually need to be translated into electrical bits since all the information processing units work in electrical domain. The bottleneck hence shifts from the link to the interface between optical domain and electrical domain, a.k.a. the optical transceiver (transmitter & receiver). This work focuses on the low-noise design techniques for high-speed optical receiver circuitry in highly scaled technology, the 65-nm Bulk CMOS. The design challenge comes from many aspects: low noise, large bandwidth, high gain, low power and large dynamics range.

Similar to wireless receiver front-end, noise is the prime design concern for optical receiver front-end, since the it would expect very small input signal, i.e. several μA_{pp} input current generated by the photodiode. Furthermore, low-noise receiver at higher speed is more difficult to design due to the large bandwidth and other inevitable trade-offs from gain and power consumption. This work investigates low-noise design techniques for optical receiver front-end circuit. A new two-stage low-noise front-end topology is proposed. Two implementation examples are given: a 25 Gb/s receiver for discrete photonics and another targeting silicon photonics.

Chapter 2 gives an overview of optical communication principle on both signal and device level. Silicon photonics, which is thought to be the future of optical communication, is also described briefly.

Chapter 3 introduces the receiver at circuit level. The three main building blocks of an optical receiver: TIA (transimpedance amplifier), LA (limiting amplifier) and Output buffer are explained in detail plus the design challenge of

each block. Recent advancements on optical receiver front-end design are also introduced in brief.

Chapter 4 lays down the theoretical foundation of this thesis. Noise composition of optical receiver is analyzed first and then follows the improved noise optimization approach. Based on that, a new low-noise two-stage front-end topology is proposed. Finally, g_m-reuse technique is also investigated.

Chapter 5 and Chapter 6 provide experimental results of the optimization and design techniques based on the theoretical ground from chapter 4. Chapter 5 describes the design of a 25 Gb/s low noise receiver to interface discrete commercial photodiode. Chapter 6 studies the design of a 25 Gb/s optical receiver interfacing silicon photonics. By utilizing developed noise optimization and design techniques, low-noise, high gain, low power, large dynamic range are obtained with superior noise performance and state-of-the-art FOM (figure of merit).

Chapter 7 concludes the work and poses future research perspective.

Chapter 2

Optical communication system

In this chapter, basic concepts and backgrounds on optical communication system are presented. We begin with an overview of the optical communication system and then investigate aspects of optical signal processing. Data format used in typical high-speed optical communication system is explained. Signal integrity issues including ISI, noise, jitter and overshoot are addressed. Methods to inspect signal quality are also given. Finally, we examine at component level by exploring real-world photonic devices and their performance. Current status of silicon photonics is briefly introduced, which represents the future of optical communication system.

2.1 Optical communication system overview

Optical communication, also called fiber-optic communication, is a means of transmitting information by optical pulses via transmission medium, usually the optical fiber. Physically, since fiber is seen as a kind of wire, optical communication is often categorized into wireline communication. Functionally, because it is used to transmit high-speed serial data, e.g. Ethernet data, optical communication is also viewed as a kind of serial link communication. Most optical communication system are digital in nature. Fig. 2.1 shows the typical high-speed optical communication system, formed by three major blocks: optical fiber, optical transceiver and Serdes.

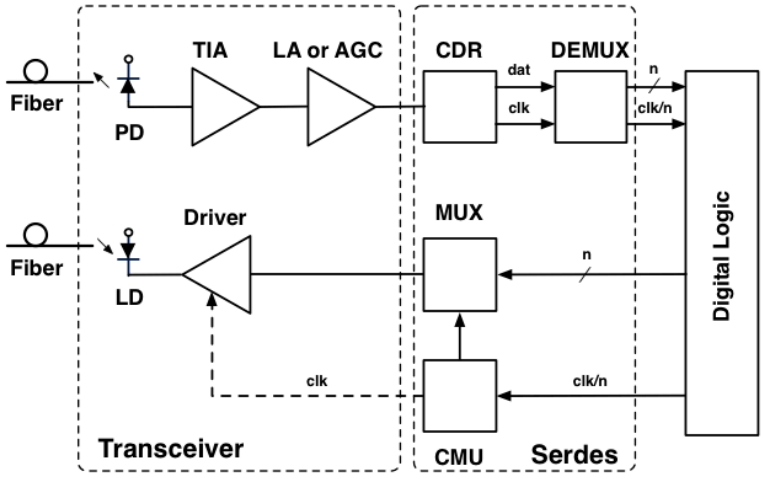


Fig. 2.1 Typical optical communication system [5]

Optical signal is transmitted within the optical fiber varying from several meters to several thousands kilometers. The optical transceiver that interfaces with the optical fiber comprises optical transmitter and optical receiver. Since

information bits need to be processed in electrical domain, the receiver performs O/E (optical to electrical) transformation while the transmitter makes E/O (electrical to optical) transformation. On the receiver (RX) side, the O/E device, a.k.a. the photodetector (PD) is coupled to the output of optical fiber, converting the optical pulses to electrical current. The transimpedance amplifier (TIA) then amplifies the small signal current to voltage, which usually needs to be further amplified by a limiting amplifier (LA) or automatic gain control amplifier (AGC), to generate voltage swing of several hundreds millivolts for post digital processing. On the transmitter (TX) side, there may be either the laser driver directly modulating the current of laser diode (LD) as shown in Fig. 2.1, or the light beam from continuous wave (CW) laser modulated by external modulator, where the laser driver is substituted by a modulator driver.

Since most optical communication systems are used in serial link transmission system, a clock and data recovery circuit (CDR), which extracts data and clock from receiver output, is usually present. Because the received data speed in the channel is much faster than core digital logic, a DEMUX following the CDR converts the high-speed serial data stream into several parallel low-speed data streams that can be easily processed by digital logic. For the same reason, a MUX is placed prior to the transmitter to merge the parallel data streams into a serial data stream. The CDR, DEMUX and MUX are collectively called Serdes, a contraction of the word serializer and deserializer.

2.2 Signal characteristic

2.2.1 Modulation methods

According to the physical span of the optical communication network, the computer network can be roughly divided into local area network (LAN), metropolitan area network (MAN) and wide area network (WAN). The counterparts in telecommunication network are usually named as short reach, medium reach and long-haul. For each kind of network, certain modulation scheme (in electrical domain) is utilized.

For short and medium reach, in most cases baseband modulation scheme, also called line coding, is adopted. The most commonly adopted data format in optical communication system is non-return-to-zero (NRZ) coding, to distinguish itself from return-to-zero (RZ) coding, both belong to OOK (on-off keying) modulation in the absence of carrier, as shown in Fig. 2.2. In the NRZ binary sequence a high signal represents a ONE bit and low signal represents a ZERO bit; while in the RZ binary sequence, a pulse represents a ONE bit and no pulse represents a ZERO bit.

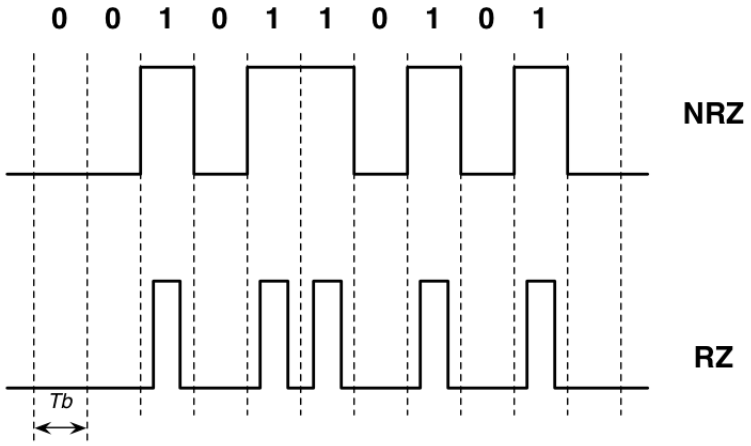


Fig. 2.2 NRZ and RZ coding

If T_b is the bit period then the bit rate R_b is $1/T_b$. It can be seen that for the same R_b, RZ pulses switch as much as twice the speed of NRZ pulses and thus require more bandwidth than NRZ signal, resulting more expensive transceiver components. In the mean time, RZ signal is more tolerant to pulse distortion and spread without disturbing its neighbor bits [5], and needs less signal-to-noise ratio (SNR) than NRZ. This means more immunity to fiber imperfections and non-linearity. As a result, NRZ is adopted more in short reach, e.g. LAN, while RZ is adopted more in longer reach, e.g. MAN.

It should be noticed that in RZ and NRZ, the bit rate equals symbol rate, where the latter is often called baud rate or simply baud. If one symbol is able to express more than 1 bit (2 states), say N states (either amplitude or phase), one symbol will have $B = \log_2 N$ bits, or in other words, bit rate is B times of baud.

For even longer distance, e.g. long-haul communication, the fiber imperfections and the non-linearity become more present and advanced modulation formats are needed to circumvent these detrimental effects. For example, the recent OIC-100G-DWDM standard [6] has mandated the DP-QPSK (dual-polarization quadrature phase shift keying) as the sole modulation scheme, where data are supposed to transmit 1000 to 1500 km long. The DP-QPSK has better spectral efficiency, meaning more tightly spaced channel and larger aggregate bandwidth. On the other hand, this modulation requires much more complex optics and electronic transceiver components.

The data format bases on several trade-offs of economics, fiber characteristics and transceiver capability. In this work, the most widely adopted NRZ is assumed as baseband format. To be transmitted optically, the electrical baseband data need to be expressed in optical domain. In a way similar to wireless communication, the baseband electrical data modulates the optical carrier to create the optical pulse train (E/O), the physical form of the information, to transmit in the optical link. Since the optical carrier frequency is extremely high, huge amount of spectrum resources are available. The concept of frequency-division multiplexing (FDM) in wireless communication also holds in optical communication corresponding to the wavelength-division multiplexing (WDM), since wavelength is a more poplar metric than frequency to characterize optical signal. When multiple optical carriers are multiplexed, the aggregate data rate can be more than several Tb/s.

2.2.2 Power spectrum

The binary sequence used for optical communication is random in nature, and it is useful to gain some insights on the characteristics of the signal spectrum. Fourier transform is applied to the random NRZ data in Fig. 2.2 to get its spectral characteristics. If the pulse is rectangular, a SINC function $S_x(f)$ can be obtained, expressed as

$$S_x(f) = T_b \left[\frac{sin(\pi f T_b)}{\pi f T_b}\right]^2 \quad (2.1)$$

and power spectrum is given in Fig. 2.3, nulls at frequency of n times of $1/T_b$.

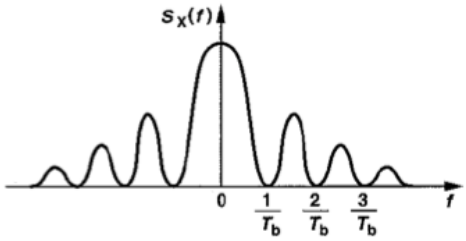


Fig. 2.3 Spectral characteristic of random NRZ binary data [7]

It is interesting to observe the cumulative power with respect to frequency, shown in Table 2.1, that 0.75 times of data rate gives 93.6% of total power and further frequency growth only gains marginal increase of total power. In addition to that, wide bandwidth not only makes circuits design more difficult, but also make receiver captures more electrical noise. These explain why the receiver bandwidth (high-frequency cutoff) is usually set around 70% of bit rate.

Table 2.1 Cumulative spectral power of random NRZ [8]

Frequency (Hz) Data rate (bits/s)	Cumulative % of Total Power
0.5	81.4%
0.75	93.6%
0.8	94.3%
0.9	95.0%
1.0	95.1%
1.1	95.2%
1.2	95.5%
1.3	96.2%
1.4	97.1%
1.5	98.1%
1.6	98.9%
1.7	99.5%
1.8	99.9%

2.3 Signal Integrity

Though the binary bits are digital conceptually, they are physically analog pulses that suffer from many signal distortion such as ISI (intersymbol interference), noise, jitter, ringing, etc. We first review the impact of those effects, and then show some common ways used to inspect the signal quality.

2.3.1 ISI

Low-pass filtering. Having mentioned before that there should be an upper limit for receiver bandwidth, alternatively there also exits a lower limit for the high-frequency cutoff, due to the signal distortion effect from ISI, shown in Fig. 2.4.

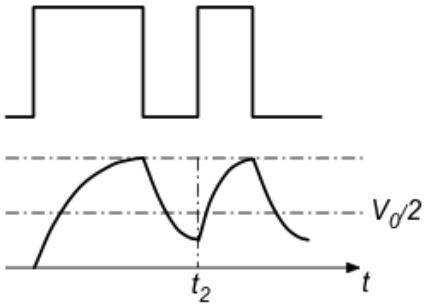


Fig. 2.4 Intersymbol interference: low pass filtering effect

Since most components in optical transceiver is bandwidth limited, signal has a finite rise and fall time. In Fig. 2.4, due to that the signal needs to rise again after t_2 even when the previous fall is incomplete, a much higher zero is generated, leading to a lower signal-to-noise ratio at this moment and probable detection error in the decision circuit.

High-pass filtering. The baseband pulse has a low-frequency cutoff not very close to zero, in the range from several kHz to several MHz. For instance, in many cases, AC coupling is frequently used to block different DC component to ease circuit design. Although the transmitted signals are usually encoded to be DC balanced (the total bits contain equal number of ONEs and ZEROs), the presence of many consecutive ONEs or ZEROs lead to signal drift and potential detection error. This effect is known as baseline wander (DC wander), as depicted in Fig. 2.5. Accordingly, for each standard the low-frequency cutoff is defined based on the data block coding method (e.g. 64B/66B in 10G-Ethernet).

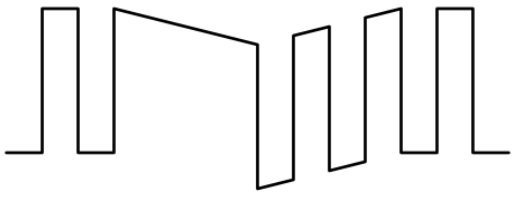


Fig. 2.5 Baseline wander

2.3.2 Noise and BER

Since optical communication usually targets a fair long distance of transmission, the signals may experience substantial attenuation, resulting a small signal as low as several μA seen by the receiver. The presence of noise from both photodetector and receiver circuits thus may lead to detection error, characterized by the bit error rate (BER). Though the error occurrence is a statistic event, communication theory shows that there is a relationship between noise and BER. The complete explanation can be found in [5],[7] and a brief illustration is given here.

Noise impacts signal both vertically (amplitude noise) and horizontally (timing noise/jitter), where the former is the main contributor of BER. In this section we deal with amplitude noise only, and its impact shown in Fig. 2.6.

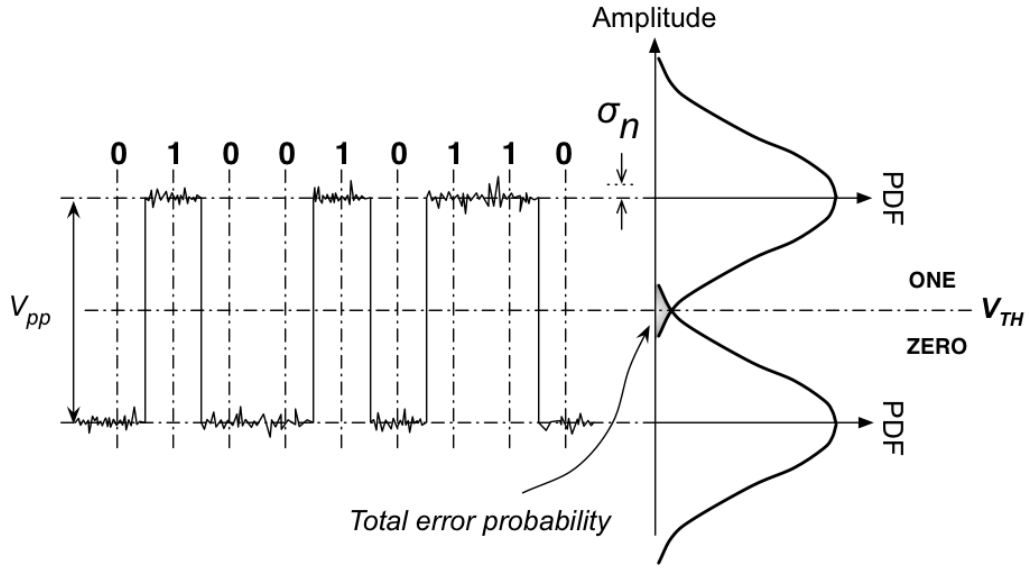


Fig. 2.6 Error mechanism due to amplitude noise

At the presence of noise, the actual signals become jagged as shown in the left side of the Fig. 2.6. The decision circuit judges the signal by the threshold (V_{TH}), which is usually set in the middle of signal swing if ONE and ZERO have equal occurrence probability. Assume the noise amplitude distribution is Gaussian, when added to signal, the overall amplitude probability distribution function (PDF) is shown on the right side (by convolution of their PDFs): vertical axis corresponds to amplitude the same as the signal and horizontal axis indicates the probability of occurrence. Thus, the cross section of the two Gaussian functions (shaded region) represents total error probability: the region higher than V_{TH} means a ZERO is erroneously recognized as ONE while the region lower than V_{TH} the vice versa. The metric to quantify the occurrence of an error is called bit error rate (BER): the probability of misjudged bits. From the PDF, BER is just the shaded region.

Assume the PDF has a standard deviation of σ_n, which equals the RMS value of noise voltage. Introducing the Q-function, defined as

$$Q(x) = \int_x^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du \tag{2.2}$$

Thus the BER the Gaussian function can be conveniently expressed as

$$BER = Q\left(\frac{V_{pp}}{2\sigma_n}\right) \tag{2.3}$$

where $V_{pp}/(2\sigma_n)$ is known as signal-to-noise-ratio (SNR). Since Q function doesn't contain closed form, the numerical values are reported in Table 2.2 with several values of SNR and corresponding BER.

Table 2.2 Q-function

SNR	BER	SNR	BER
0	0.5	5.998	10^{-9}
3.090	10^{-3}	6.361	10^{-10}
3.719	10^{-4}	6.706	10^{-11}
4.265	10^{-5}	7.035	10^{-12}
4.753	10^{-6}	7.349	10^{-13}
5.199	10^{-7}	7.651	10^{-14}
5.612	10^{-8}	7.942	10^{-15}

From above analysis, to have enough SNR we are prone to lower the bandwidth for lower noise; but this on the other hand will increase ISI, which also leads to smaller SNR. The trade-offs between noise and bandwidth becomes one fundamental concern for receiver circuits design.

2.3.3 Jitter

While ISI and noise may result amplitude error, they also impact the signal timing and the effect is called jitter. More specifically, ISI in the time domain is known as data-dependent jitter (DDJ) and noise in time domain is known as random jitter (RJ). DDJ is mainly due to insufficient bandwidth and phase non-linearity. It belongs to a larger category called deterministic jitter (DJ) which is bounded it nature. On the other hand since RJ comes from noise, it's unbounded. The overall effects of DJ and RJ are called total jitter (TJ).

Not only the data has jitter, the decision circuit also has sampling jitter from the sampling clock, known as clock jitter. Since both DJ and clock jitter contributes to BER if they are uncorrelated, phase lock loop (PLL) can be used to let clock jitter track DJ. Clock jitter in frequency domain is called phase noise, which mainly comes from oscillator. Generally, in [5] it is shown that jitter has much less impact on BER than amplitude noise.

2.3.4 Overshoot

Another commonly seen signal distortion is overshoot, often followed by ringing, as shown in Fig. 2.6.

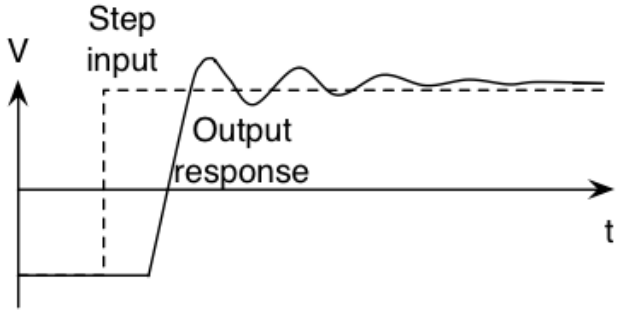


Fig. 2.7 Overshoot and ringing

Overshoot is an undesired parasitic oscillation, which lowers signal-to-noise ratio and leads to extra ISI. Frequency domain gain peaking will lead to overshoot and ringing and thus it is important to maintain gain flatness of transfer function in addition to the bandwidth concern.

2.3.5 Signal inspection

Eye diagram. Due to the several effects described above (ISI, noise, jitter, overshoot, ringing), it becomes difficult to judge the quality of transmitted signal in time domain. As a consequence, it is possible to fold the time axis by integer time of bits, generating an overlapping of time-domain signal known as the eye diagram since its appearance looks like a human eye.

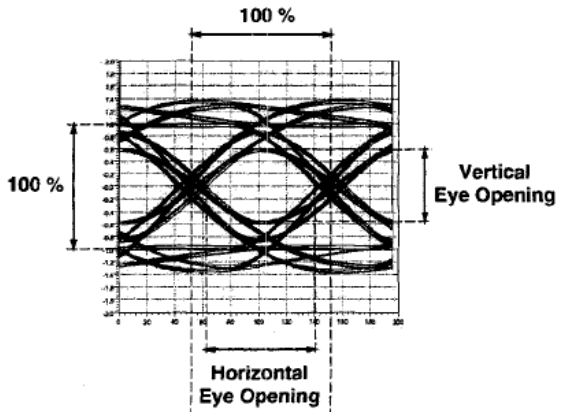


Fig. 2.8 Eye diagram in OOK system [5]

Intuitively, the larger and clearer the eye is opened, the better the signal quality. More rigorously, vertical and horizontal eye opening/closure is used to characterize the eye. The vertical eye opening (closure) is the percentage of minimal opening to full opening (distance between mean ONE level and mean ZERO level) in the sampling point, usually in the middle of the pattern. Similarly, the horizontal eye opening (closure) is the percentage taken at the slice level.

Eye diagram can be obtained by measurement from sampling oscilloscope. On the other hand, simulation can only provide noiseless eye and data dependent signal distortions. In a noiseless eye, the vertical eye closure is mainly due to ISI while horizontal eye closure is mainly due to deterministic jitter. In the eye diagram from oscilloscope, amplitude noise contributes to vertical eye closure

and random jitter contributes to horizontal eye closure. Since Gaussian noise is unbounded, eye closure needs to be replaced by the statistical metric eye margin, basically a constant-BER contour plot [5]. Here margin means how much detector related error could be tolerate for a certain BER. It should also be noted that eye diagram is data pattern dependent: generally longer PRBS (pseudo random bit sequence) pattern will close the eye more.

BERT. Since BER is the direct measure of the transmission quality at the presence of noise, ISI, distortion and other effects, it is thus necessary to make this measurement, known as bit error rate test (BERT). This concept of test setup is shown in Fig. 2.9.

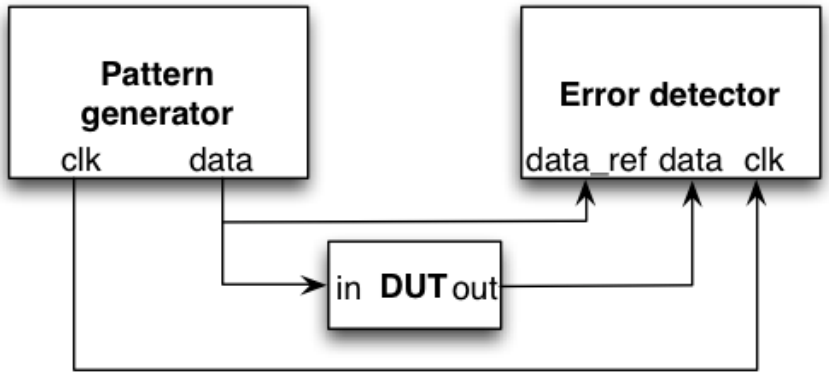


Fig. 2.9 BERT test setup

The pattern generator (PG) transmits test pattern to the DUT, usually the receiver. The error detector (ED), synchronized by the clock signal from the PG, checks the input bits from the DUT and count error number. In optical link test, E/and O/E device are necessary to test the optical communication signal.

2.4 Optical devices

Since the data bits will be finally transmitted in optical domain, it is substantial to understand the characteristic of optical device/component and their impact on signal transmission.

2.4.1 Optical Fiber

Optical pulses needs transmitted in a well-defined communication channel, i.e. the optical fiber. The characteristic of optical fiber essentially defines the transceiver and signal format thus deserves some description.

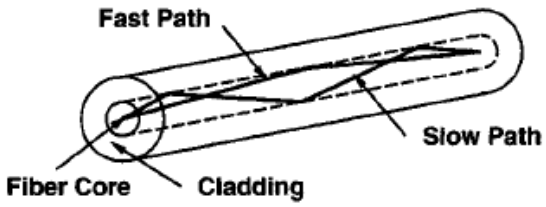


Fig. 2.10 Optical fiber and light transmission [5]

A typical optical fiber for optical communication is shown in Fig. 2.10, comprising three major layers: core, cladding and jacket (not shown). The core and cladding are made of silica glass, where the cladding has a lower refractive index. In such a way the light's propagation path is confined within the core, known as total internal reflection. The jacket layer protects the fiber. There are two types of optical fiber: the single-mode fiber (SMF) and the multimode fiber (MMF). Physically, SMF has a smaller core diameter (< 10 μm) while MMF has a larger core diameter (> 50 μm). The larger core in MMF not only simplifies the fiber connection and alignment, but also allows low-cost interface optics and electronics. However, since multiple optical transmission modes exist in MMF, the signal experiences multimode distortion that limits the bandwidth and distance. On the contrary, only one mode in SMF offers better bandwidth and transmission distance, at the cost of more expensive interfacing components. More specifically, dispersion and attenuation, known as fiber non-linearity, are used to characterize the optical fiber.

Absorption, scattering and fiber impurities cause attenuation, measured by dB/km. The attenuation profile of silica fiber is shown in Fig. 2.11 [9], where the three popular transmission windows for optical communication at 850 nm, 1310 nm, and 1550 nm are marked. The second and third windows obviously come from their low attenuation while the first window is chosen due to that low cost laser and detector are available in this wavelength, though with higher attenuation (2.5 dB/km).

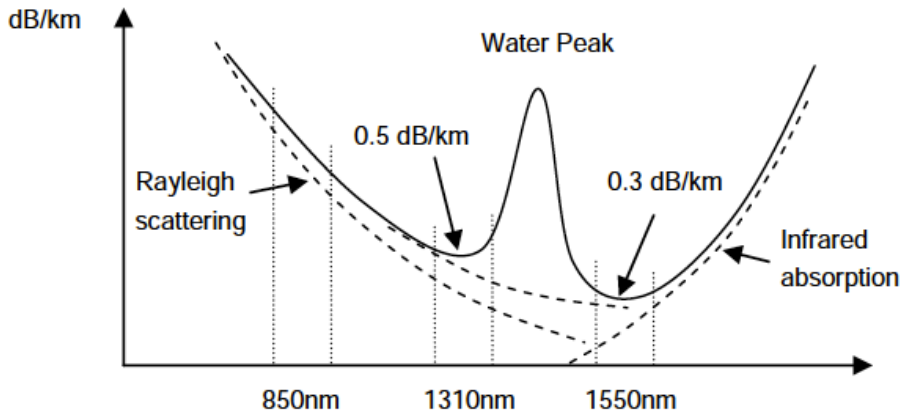


Fig. 2.11 Attenuation profile of silica fibers [9]

Dispersion is the fact that the optical pulses spread gradually as they travel in the optical fiber and three important dispersions needs to be taken account: modal dispersion, chromatic dispersion and polarization-mode dispersion (PMD). Dispersion is usually measured by ps/nm/km.

Modal dispersion takes place in MMF due to different propagation delay in multiple pathways for light to travel, resulting distorted pulse and limited transmission distance within several hundred meters. In SMF there is only one mode is available and this effect is suppressed.

Chromatic dispersion is owing to the difference propagation delay in difference wavelength. Chromatic dispersion is more concerned in SMF than MMF since much longer transmission distance is usually required. For SMF, chromatic dispersion goes to minimum at around 1.3 μm, but 1.55 μm is adopted in long-haul due to its lower attenuation and the optimum *erbium-doped fiber amplifier* (EDFA, to amplify optical signal) performance. For shorter distance, 1.3 μm is preferred like in IEEE 802.3ba 40G/100G standard.

PMD is mainly in SMF due to fiber asymmetry, e.g. elliptic rather than circular. Different polarization modes (horizontal and vertical) result in different speeds. Both chromatic dispersion and PMD can be compensated optically by special fiber to counteract the effects or by equalizer in electrical domain.

From the above analysis, SMF is chosen for high speed, long distance optical communication while MMF servers the opposite. In another way, SMF has a larger bandwidth-distance product than MMF. Thus, MMF is usually used in data communication spanning several hundred meters with channel speed less than 10 Gb/s, while SMF is adopted in telecommunication spanning several km for channel speed of as high as 40 Gb/s. SOA (semiconductor optical amplifiers) can be used to regenerate the optical signal in SMF, extending the total transmission distance to even several thousand km.

2.4.2 Optical source and modulator

As mentioned before, the optical carrier is modulated by electrical baseband data to generate optical pulses for transmission and there are two different methods: direct modulation and external modulation. Direct modulation means the laser driver directly modulates the current of laser diode. External modulation means the laser is used as a continuous wave source and its light beam is modulated in the external modulator which driven by modulator driver. Direct modulation is more cost effective, easy to implement and integrate while external modulation generates better optical pulses, extending the transmission length and data rate, but is usually bulky and expensive. Optical modulation not only changes the light's amplitude as it supposed to do, but also modulates the light's frequency, known as Chirp. Direct modulation generates much more chirp than external modulation.

An important property of laser is the extinction ratio (ER), defined as

$$ER = \frac{P_1}{P_0} \tag{2.4}$$

where P_1 is the optical power for ONE and P_0 for ZERO. Finite ER incurs a power penalty to maintain the same eye opening. ER from direct modulation is between 9~14 dB while in external modulation it could exceed 15 dB [5].

Optical source. The semiconductor optical sources belong to two categories: laser (light amplification by stimulated emission of radiation) and LED (light-emitting diode). Light beam from the former comes from stimulated emission

and the latter from spontaneous emission, forming the basic difference of the two. Both the laser diode and LED contain a sandwich-like structure: a forward-biased PN junction wraps an active region (cavity) in between, but the laser is further covered by mirror on left and right side.

In LED, electrical current injects electrons and holes into active region where they recombine and emit photons of corresponding bandgap energy. Physically, the transition of electrons from higher energy to lower energy is spontaneous, thus this phenomenon is called spontaneous emission. Because the photons produced have arbitrary phase and direction, the light created is "incoherent". Thus not only the light has large spectral linewidth, but also little of the light can be coupled into fiber. On the other hand, since LED is low cost and reliable than laser, it is used for short-reach low-data rate communication based on MMF.

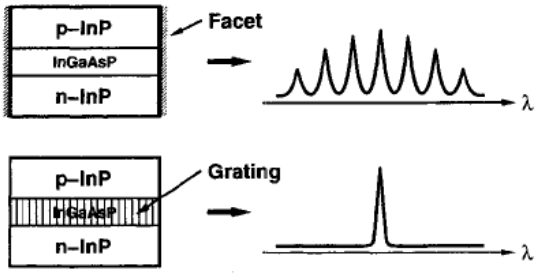


Fig. 2.12 FP laser (up) and DFB laser (down) [5]

On the contrary, the transition of electron from high energy to low energy is stimulated by incoming photons with proper energy, generating new photon in phase with the stimulus photon, which is known as "coherent". They together continue this procedure, called "stimulated emission", as long as the high-energy electrons are provided continually. In such a way, optical gain is obtained. To select certain light wavelength, the laser also needs to behave like an optical oscillator. The mirrors and the cavity length together determine the wavelength, forming Fabry-Perot laser (FP laser), as shown in Fig 2.12. Since the cavity length contains multiple times of the wavelength, the light spectrum thus have several peaks (sidemodes), where spectrum length is large, e.g. 3 nm. FP laser hence is used in 1.3 μm SMF where dispersion is low.

To get better spectral purity, a grating structure that provides distributed feedback is added to select only one wavelength. This kind of laser is called DFB laser (shown in Fig.2.12) and provides excellent spectrum linewidth, e.g. < 0.001 nm. DFB laser offers the best performance in terms of data rate and distance, but is sensitive to temperature variation. Depending on the requirement, there is cooled and uncooled lasers, where the latter contains a feedback loop to control the laser temperature for reliable operation and thus more bulky. FP laser and DFB laser are collectively known as edge emitting-laser since light emit to the edge.

Another type of laser is the VCSEL (vertical-cavity surface-emitting laser). Similar to LED, this kind laser emits light perpendicularly from the surface since the mirrors are placed above and below the active region. Though only one wavelength fits the cavity, there exits many traverse modes that widen the spectral length width to around 1 nm. VCSEL is low cost and popular in the 850 nm wavelength communication in MMF.

Modulator. External modulation is better than direct modulation in terms of turn-on delay, relaxation oscillation and chirp [7]. When data rate goes higher than 10 Gb/s, external modulation that separates the optical carrier generation and modulation is desired. Two types of modulators are commonly seen in optical transmission system: the Electroabsorption modulator (EAM) and the Mach-Zehnder modulator (MZM).

A DFB laser combined with EAM as the source is known as an electro-absorption modulated (EML). EAM has a similar structure like LED, where active region is sandwiched by reverse-biased PN junction. It works based on Franz-Keldysh effect: the bias voltage across PN junction controls the electric filed, which changes the effective bandgap to be opaque (absorbing photons) or transparent to the light wavelength. EML enables laser and modulator integrate in the same substrate, but produces some chirp.

The structure of MZM is configured as an interferometer, show in Fig. 2.13: optical signal are split into two arms and experiences phase modulation, and then the two arms are recombined. If their phase difference is 180°, the interference is destructive and the light intensity is ZERO; if their phase difference is 0°, the interference is constructive and the light intensity is ONE. The phase modulation is based on the electro-optic effect: the electric field changes refractive index, thus affecting the prorogation velocity of the light. Lithium niobate (LiNbOs) is usually used to make the waveguide where light is phase modulated. Dual-drive MZM enables differential signaling and lower the voltage swing to modulate each arm. MZM gives the lowest amount of chirp and highest ER, but is more bulky and hard to integrate.

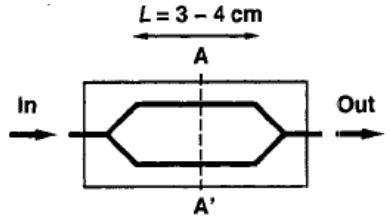


Fig. 2.13 Mach-Zehnder interferometer [5]

2.4.3 Photodetector

Optical signal needs to be transformed into electrical current by photodetector, which in semiconductor is usually realized by reverse-biased photodiode (PD). The structure of PIN photodiode is similar to LED: an intrinsic material sandwiched by PN junction, shown in Fig. 2.14 (a). Photons going into intrinsic region with proper energy generate electron-hole pairs, which then form electrical current at the presence of electric filed. The intrinsic region is GaAs or silicon for 0.85 μm and InGaAs for larger wavelength.

Due to reflection and absorption as heat, not every photon stimulates an electron-hole pair, which is quantified by the "quantum efficiency". Electrically, the photodiode is measured by "responsivity", defined as

$$\mathcal{R} = P/I \tag{2.5}$$

where P is the optical power and I is the generated electrical current.

The design of the intrinsic region poses a trade-off of bandwidth and responsivity: longer width (W in Fig. 2.14 (a)) increases responsivity, but the electrons and holes will need more time to transit this region. On the other hand, smaller width not only degrades responsivity but may also limit the bandwidth because more junction capacitance exhibits. The overall bandwidth is thus given by [5]:

$$BW = \frac{1}{2\pi(W/v_n + R_{PD}C_{PD})} \tag{2.6}$$

where v_n is the transit time, R_{PD} and C_{PD} are the parasitic resistance and capacitance. Since the resistance is usually small, a photodiode is usually modeled as a current source in parallel with the capacitance electrically.

(a) (b)

Fig. 2.14 (a) PIN photodiode and (b) avalanche photodiode [5]

To overcome the responsivity-speed trade-off, very high-speed PD is illuminated from the side. Horizontal dimension now can be made large enough for sufficient quantum efficiency, while width of intrinsic region can still remain small for high speed. This type of photodiode belongs to waveguide photodiode (WPD).

Another kind of photodiode used in optical communication (lower-speed) is the avalanche photodiode (APD), as shown in Fig. 2.14 (b). An extra P layer is inserted to provide optical gain by avalanche multiplication of the electron-hole pairs. APD has much larger responsivity than PIN photodiode, but also presents itself more optical noise. Also, APD needs a much higher bias voltage to sustain the avalanche multiplication.

2.5 Silicon photonics

Motivation. In view of the tremendous success in electronic integrated circuits (EIC), photonic integrated circuits (PIC) also draw a lot of investigation. Major stimulus comes from Internet data center, high performance computing and storage network, where a large amount of high-speed I/O are needed. The integration enables optical go from rack-to-rack to board-to-board and chip-to-chip, and will eventually become on-chip high-speed routes.

Some initial success has been demonstrated on InP-based PIC. InP (indium phosphide) is lattice matched with GaAs and most III-V materials, lending itself superior advantage to produce active photonic device. Furthermore, InP has

been used traditionally to make high-speed Telecom electronic circuits due to its superior electron velocity. Thus, InP is an advantageous host material for PIC. For example, Infinera has demonstrated a 10 x 100 Gb/s DWDM InP transceiver IC with an aggregate bandwidth of 1 Tb/s [10].

On the other hand, Si-based photonics (silicon photonics), which has the natural compatibility with CMOS technology, allows higher degree of integration, lower cost and utilization of existing sophisticated silicon fab processing. Silicon photonics draws more attention and stands for the future of optical communication. From 2006, Luxtera [11], Intel [12] and IBM [13] have independently demonstrated several integrated optical transceivers based on silicon photonics, indicating the maturation of the silicon photonics and the upcoming massive commercialization. We briefly introduce silicon photonics here.

Passives. Silicon not only is transparent to near-Infrared (IR) light, but also has high refractive index (3.5) compared with silica (SiO_2: 1.44). Thus, optical waveguide could be realized in SOI CMOS, where the BOX (Buried Oxide) layer intervenes photonics from silicon substrate, as shown in Fig. 2.15. The high index contrast enables low loss and small bend radii optical waveguide realized for compact optical routing. Since the waveguide is small enough to allow only single mode propagation, modal dispersion is eliminated. Other passives like AWG (arrayed waveguide grating) and grating coupler have also implemented on silicon, allowing sophisticated manipulation of light on-chip.

Photodetector. Silicon doesn't detect IR light well and Ge-epitaxy extends the absorption of light up to 1.6 μm [14], where Ge is strained to overcome Si-Ge lattice mismatch. Since light coupled into the PD from on-chip waveguide, Ge-on-Si PD belongs to waveguide photo detector and features much lower capacitance than III-V counterparts [16]. Ge-on-Si PD also enables wafer scale test, compared with III-V-on-Si approaches, and improves yield and decreases bonding cost. LETI has reported Ge-on-Si PD with 42 GHz bandwidth, 1 A/W responsivity and a dark current of 20 nA at 4V bias [15].

Modulator. Two types of modulators exist on silicon: MZM and ring modulator. On-chip high-speed MZM operates on the same principle described in 2.4.2, while the phase modulation is based on the plasma dispersion effect. The reverse-biased diode causes carrier depletion and extract carrier from waveguide, resulting a change of the refractive index and thus the optical phase [26]. To get better phase modulation and lower the modulation voltage and parasitics, distributed MZM structure is preferred while it's bulky and introduces a high insertion loss of around 5~7 dB [38]. The ring modulator on the other hand is much smaller, based on forward-biased diode injecting carrier that changes refractive index. However, ring modulator faces the problem of thermal stabilization so that the optical parameters are sensitive to temperature [37].

Light source. Laser is the least integrated photonics on-chip and the hybrid integration approach is adopted in most situations. Since silicon is not able to lase effectively, III-V compound is indispensable. Three different bonding are available between III-V and silicon: 1) die-to-die (Luxtera); 2) die-to-wafer (LETI,

IMEC); 3) wafer-to-wafer (USCB, Intel). The automation level increases and thus bonding cost decreases.

Integration. Although silicon photonics lends itself the possibility to integrate EIC and PIC on the same substrate, in reality hybrid integration or 3D integration may provide higher performance, more flexibility and lower cost. Both approaches are reviewed in brief as follows.

Though monolithic approach gives higher degree of integration, only CMOS SOI wafer rather than standard CMOS could be utilized. Furthermore, co-inhabitancy of EIC and PIC will lower yield and flexibility. Third, CMOS advancement couldn't be enjoyed timely due to extra process tuning time for optics. An example of this approach is shown in Fig. 2.15.

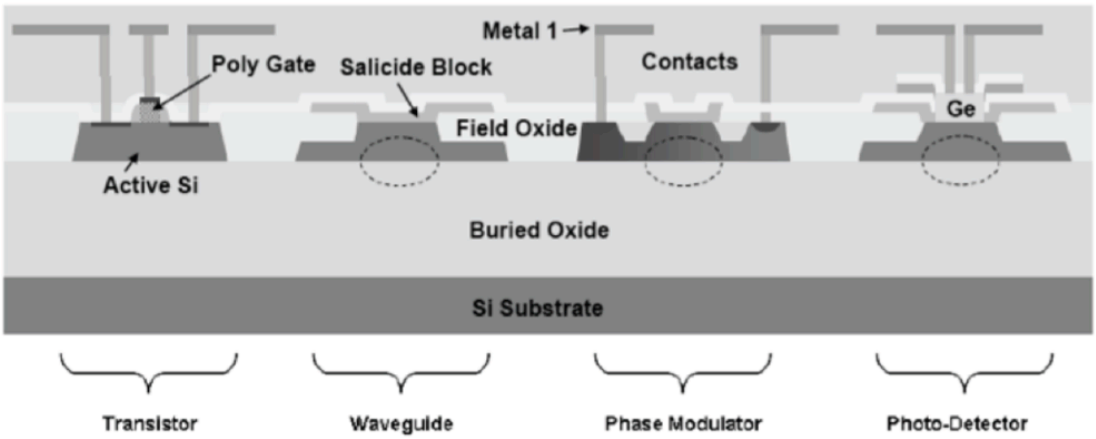


Fig. 2.15 Luxtera silicon photonics technology [16]

Hybrid approach stacks PIC and EIC in a 3D manner, bounded by 3D through silicon via (TSV) or Cu-pillar. The process development of PIC and EIC are thus separated, offering better flexibility and yield. Another potential advantage is the possibility of stack other chip, like MEMS, Memory, etc. Last, the advancement of bonding like Cu-pillar continues to decease the parasitics. Fig. 2.16 shows the concept of 3D integration of EIC and PIC [13].

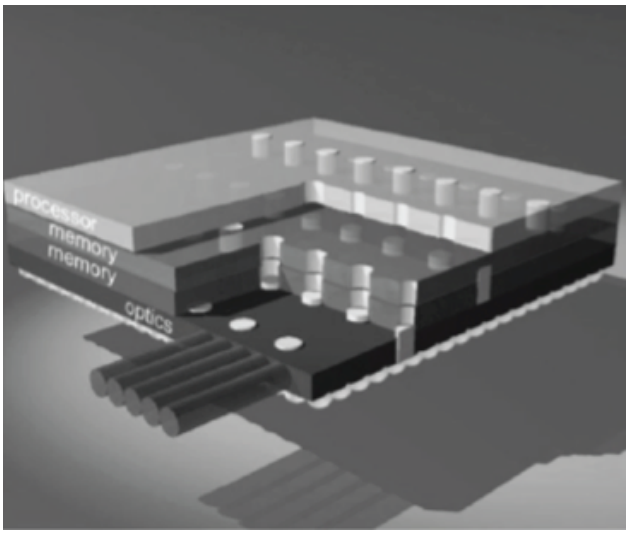


Fig. 2.16 Conceptual system-on-stack exascale computing node in 2020 [13]

2.6 Conclusion

In this chapter, we have presented basic concepts related to optical communication. Typical optical system has been introduced to give a system perspective. Then, data formats related to optical communication have been analyzed. Based on that, typical signal integrity issues like ISI, noise, jitter and overshoot, have been studied to gain insights of their impact on system performance. In particular, the relationship between noise and BER establishes the fundamental for low-noise receiver design. Signal inspection techniques including eye diagram and BERT have also been described. Then, key optical devices including fiber, laser source, modulator and photodiode have been explained from concept to structure. Finally, advancement on silicon photonics has been presented. With the knowledge from system to signal and device, we are now more comfortable to move to receiver circuits design in the next chapter.

Chapter 3

CMOS optical receiver design fundamentals

Having reviewed the principles of optical communication system, we come to the main focus of this work: (low-noise) CMOS optical receiver design. We first introduce important receiver design parameters, and then explain general circuits design consideration for each receiver building block based on the system level investigation we have carried out in previous chapter. Finally, some state-of-the-art CMOS optical receivers are reviewed and compared.

3.1 Receiver design parameters

The architecture of monolithic optical receiver is shown in Fig. 3.1, where an Output buffer is required to drive off-chip loads. This is the case for receiver chip not integrated with Serdes circuits. As explained before, the basic function for receiver is to amplify the small photo-current generated by photodiode to a few hundred mV, large enough to drive CDR. In this sense, the receiver as a whole functions as a transimpedance amplifier with several kilo-ohm transimpedance gain. The limiting amplifier (LA) adds extra gain to TIA to ensure sufficient overall gain for signal amplification. As explained in previous chapter, due to several kinds of impairments (ISI, noise, jitter, overshoot, etc.) co-exiting in the amplifying process, stringent design parameters should be met for the receiver, explained as follows.

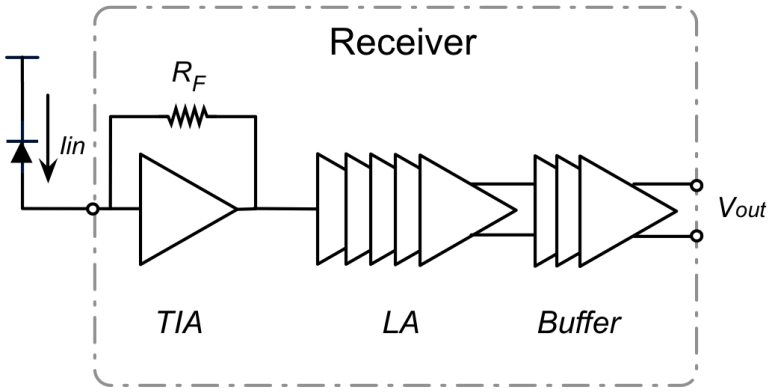


Fig. 3.1 General monolithic optical receiver architecture

Gain. From input photo-current of tens of μA to output voltages of hundreds mV means a transimpedance gain of tens of kilo-Ohms. The transimpedance is defined as

$$Z_T = \frac{Vout}{Iin} \tag{3.1}$$

where *Vout* and *Iin* are shown in Fig. 3.1, indicating small signal parameters. The Midband value of $|Z_T(f)|$ is called transresistance (R_T). The wideband bandwidth requirement makes high gain difficult to obtain, since gain and bandwidth trade with each other. Because the Output buffer usually contributes little gain, TIA and LA together determine the overall gain.

Noise. The small photo-current renders the receiver noise performance critical; or in another way, low noise (high sensitivity) receiver can extend the communication distance and tolerate more non-idealities in transmission. As will be seen, when data rate goes higher, low noise becomes more and more difficult to achieve. Generally, TIA and LA both impact the noise performance, where TIA has a key role because it's closer to the signal source. In practice, the input-referred noise is used as the metric for noise performance. Since the photodiode usually exhibits a high shunt resistance in parallel with the parasitic photodiode capacitance, input-referred noise current is sufficient as input-referred noise source. However, the input-referred noise current should be quoted together with the photodiode capacitance, which determines the source impedance. More rigorously, the input-referred noise current can be expressed in three ways [5].

- Input-referred noise current spectrum, $\overline{I_{n,in}^2(f)}$, is obtained from $\overline{V_{n,out}^2(f)}$ divided by the transimpedance gain square $|Z_T(f)|^2$, expressed as

$$\overline{I_{n,in}^2(f)} = \frac{\overline{V_{n,out}^2(f)}}{|Z_T(f)|^2} \tag{3.2}$$

  This power spectrum is measured in pA^2/Hz or its square root $pA/\sqrt{Hz}$.
- Input-referred RMS noise current, $I_{n,in}^{rms}$, is defined by the output RMS noise voltage $V_{n,out}^{rms}$ divided by Midband transimpedance gain R_T.

$$I_{n,in}^{rms} = \frac{V_{n,out}^{rms}}{R_T} \tag{3.3}$$

  With $I_{n,in}^{rms}$, input sensitivity can be easily calculated by

$$I_{sen}^{pp} = 2Q * I_{n,in}^{rms} \tag{3.4}$$

  For a BER of 10^{-12}, Q = 7.035, thus I_{sen}^{pp} is about 14 times of $I_{n,in}^{rms}$.
- Average input-referred noise current density, $I_{n,in}^{avg}$, is defined by $I_{n,in}^{rms}$ divided by square root of bandwidth.

$$I_{n,in}^{avg} = \frac{I_{n,in}^{rms}}{\sqrt{BW}} \tag{3.5}$$

  It should be noted that this value doesn't equal to the input-referred noise current spectrum averaged in band, because the noise bandwidth BW_n is usually different from *BW*. They are equal only when the transfer function has a rectangle shape.

Bandwidth. (I) High-frequency cutoff. This parameter, known as the receiver bandwidth, has dual impact on the two most important parameters: ISI and noise. This is because wide bandwidth not only decreases ISI, but also allows more noise to be captured. As a result, it is usually suggested the overall receiver bandwidth should be made about 70% [7] of data rate when NRZ format is utilized. In high-speed receiver, usually the TIA sets the overall receiver bandwidth due to the numerous trade-off. (II) Low-frequency cutoff. This frequency is closely related to baseline wander. In specific, it is affected by the maximum run length of continual ZERO/ONE, set by different communication standards.

Input dynamic range. The receiver input signal strength may vary for various fiber lengths. In the extreme case, when very large photo-current (e.g. on the order of mA) feeds into the receiver, its working condition may change completely. Though the binary transmission and limiting nature of receiver suggest that certain degree of distortion can be tolerated, cares should still be taken so that the pulse distortion and jitter don't impact the transmission quality seriously. To tolerate a large input dynamic range, both TIA and LA may need some kind of gain control mechanism.

Group-delay variation. The group-delay (τ) is calculated by the derivation of phase Φ:

$$\tau = -\frac{d\Phi}{d\omega} \tag{3.6}$$

and the group-delay variation (GDV) is an indicator of phase linearity. Large GDV will result ISI and data-dependent jitter that degrade transmission quality. Inductive peaking, which is popular in high-speed circuit design, while effectively boosts bandwidth, may have a detrimental effect on group-delay variation due to the introduction of high-Q complex poles.

***Power consumption*.** Since optical communication is to be massively utilized, for example, in data center where the transceiver number could reach more than million [17], it's important to keep the receiver power low to tens of mW.

***Offset correction*.** Both the DC component from photo-current and amplifier itself can give offset to the receiver. The low-frequency offset not only unbalances the receiver from its optimal working condition, but also lowers its sensitivity, since the decision threshold in decision circuit is no longer centered to the middle of receiver output eye. Consequently, offset correction function is needed. As will be explained, the low-frequency offset correction loop will introduce low-frequency cutoff to the receiver thus needs careful design.

In the following, we dive into each block to explore the circuit design considerations that meet the aforementioned requirements where each block has different focus. Generally, the front-end amplifier, in this context the TIA, plays the most critical role since it faces more trade-offs than other blocks and thus deserves the most attention.

3.2 Transimpedance amplifier

The transimpedance amplifier (TIA) has the decisive role on the overall input-referred noise because it is the frond-end amplifier, similar to the role of LNA in wireless receiver. Furthermore, we will see that its noise directly trade with receiver bandwidth in every situation, rendering it very difficult to go low in high data rate communication. Also, more transimpedance gain from TIA is highly desired since it can lower the noise contribution from latter stages, e.g. limiting amplifier. Consequently, noise, bandwidth and transimpedance gain are the most critical metrics for TIA design and we will examine them for each TIA topology with an emphasis on high-speed (wide bandwidth) communication.

3.2.1 Resistor TIA

A resistor is the simplest way to convert current into voltage, so it is the simplest form of TIA, where the two-port inter-stage network (transimpedance) retrogresses to a one-port inter-stage network, while the current to voltage conversion is still carried out, as shown in Fig. 3.2. Other stages following TIA are not drawn in the figure.

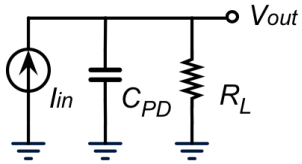


Fig. 3.2 Resistor TIA

The transimpedance gain, input-referred noise current can be easily obtained using the definitions given in the last section. The resistor value R_L is the transresistance. The -3-dB bandwidth is set by the first-order RC low-pass, formed by the photodiode capacitance C_{PD} (suppose it dominates) and transresistance R_L. The input-referred noise current equals the thermal noise current from the resistor. Thus we have

$$R_T = R_L \tag{3.7}$$

$$BW = \frac{1}{2\pi R_L C_{PD}} \tag{3.8}$$

$$\overline{I_{n,in}^2} = \frac{4kT}{R_L} \tag{3.9}$$

Here, we observe a direct trade-off between bandwidth and noise. Wideband width necessitates a small resistor, which gives large noise to the input. Furthermore, small resistor value also has less suppression on the noise from latter stage, resulting a much larger noise in the input. Thus, the resistor TIA is ill-suited for high-speed application.

3.2.2 Shunt-feedback TIA

3.2.2.1 First-order shunt-feedback TIA

From last section, input-referred noise can be reduced only when the resistor is sufficiently large, which then limits the bandwidth. Shunt-feedback (a.k.a. shunt-shunt feedback) shown in Fig. 3.3, on the other hand, enables both large resistance (low noise) and wide bandwidth.

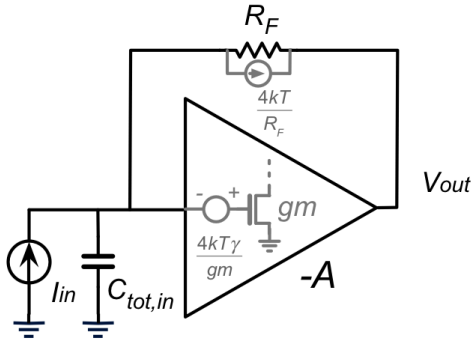


Fig. 3.3 Shunt-feedback TIA (first-order)

In the figure, $C_{tot,in}$ stands for the total input capacitance, the sum of photodiode capacitance C_{PD} and MOS capacitance from TIA $C_{in,tia}$. Only the thermal noise from the front-end transconductance is considered for the core amplifier.

From feedback theory, at low frequency, shunt-feedback lower input impedance by a factor of *A+1*, we have

$$R_{in} = \frac{R_F}{A+1} \tag{3.10}$$

Neglecting the output open-loop pole for a moment and the -3-dB bandwidth thus equals the closed-loop input pole frequency:

$$BW_{-3dB} = \frac{A+1}{2\pi R_F C_{tot,in}} \tag{3.11}$$

The Midband transimpedance gain is

$$R_T = \frac{A}{A+1} R_F \tag{3.12}$$

which roughly equals R_F if loop gain A is sufficiently large. The low-frequency input-referred noise current is given by

$$\overline{I_{n,in}^2} = \overline{I_{n,R_F}^2} + \overline{I_{n,amp}^2} \tag{3.13}$$

$$= \overline{I_{n,R_F}^2} + \frac{\overline{V_{n,amp}^2}}{R_F^2} \tag{3.14}$$

$$= \frac{4kT}{R_F} + \frac{4kT\gamma}{g_m R_F^2} \tag{3.15}$$

From Eq. 3.15 we can see that the input-referred amplifier noise is referred to the input by a factor of R_F^2, making it trivial in low frequency and the overall input-referred noise current roughly equals $4kT/R_F$.

Now let's compare shunt-feedback TIA with resistor TIA. Let $R_L = R_F/(A+1)$ and assume same total input capacitance, the two have equal bandwidth but shunt-feedback TIA exhibits A+1 times larger transimpedance gain and $A+1$ times smaller input-referred noise current. In another way, if we let $R_L = R_F$, the two have roughly equal transimpedance gain and input-referred noise current but shunt-feedback TIA gains a factor of A+1 in bandwidth.

Introducing the amplifier and shunt-feedback breaks the noise-bandwidth trade-off as long as the core amplifier provides high gain. This makes feedback-TIA very popular for a long time in CMOS TIA and it is still the workhorse for Bipolar TIA nowadays.

3.2.2.2 Second-order shunt-feedback TIA

In previous first-order analysis, the amplifier pole is neglected assuming that it is much higher than closed-loop input pole. However, if the overall TIA speed goes quite high, this assumption will no longer be true and the amplifier pole is going to impact the closed-loop behavior and eventually limit the bandwidth.

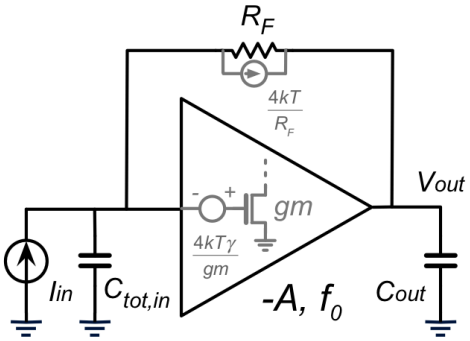


Fig. 3.4 Shunt-feedback TIA (2nd-order)

In Fig. 3.4, assuming the amplifier has a single pole f_0 due to the presence of C_{out}, which stands for the total output capacitance, the system becomes second-order. Intuitively, if input pole and output pole in open-loop are close enough, the system will face stability problem. To appreciate this, the open-loop frequency response is plotted in Fig. 3.5.

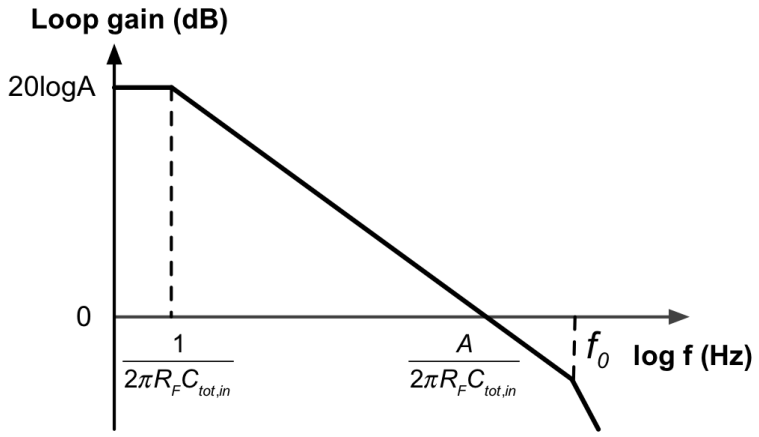


Fig. 3.5 Open-loop frequency response of 2nd-order feedback TIA

According to stability theory, by setting f_0 to $2A/(2\pi R_F C_{tot,in})$, we achieve a phase margin of about 60° and the transient response has negligible overshoot. More rigorously, the closed-loop frequency response is expressed as follows [5].

$$Z_T(s) = \frac{\frac{A}{A+1}R_F}{\frac{s^2}{\omega_n^2} + \frac{s}{\omega_n Q} + 1} \tag{3.16}$$

$$\omega_n = \sqrt{\frac{(A+1)\omega_0}{R_F C_{tot,in}}} \tag{3.17}$$

$$Q = \frac{\sqrt{\frac{(A+1)R_F C_{tot,in}}{\omega_0}}}{R_F C_{tot,in} + {}^{1}/_{\omega_0}} \approx \sqrt{\frac{(A+1)}{R_F C_{tot,in}\omega_0}} \tag{3.18}$$

When Q is $1/\sqrt{3}$ it is Bessel response, which gives maximally flat group-delay characteristics; when Q is $1/\sqrt{2}$ it is Butterworth response, which gives maximally flat frequency response and no peaking in frequency response. Larger Q will give rise to peaking in frequency response, which degrades the output eye. Thus, usually Butterworth response is the design target. Let Eq.3.18 equal $1/\sqrt{2}$ and we have:

$$f_0 \approx \frac{2A}{2\pi R_F C_{tot,in}} \tag{3.19}$$

which is in agreement with what we obtained from phase margin analysis before.

At this time, the -3-dB bandwidth can be easily obtained by setting the denominator of Eq. 3.16 to $\sqrt{2}$, and we get [7]

$$\omega_{-3dB} = \omega_n \tag{3.20}$$

or written as

$$BW_{-3dB} \approx \frac{\sqrt{2}A}{2\pi R_F C_{tot,in}} \tag{3.21}$$

Interestingly, as f_0 goes from infinity to lower frequency (Butterworth response), the overall -3-dB bandwidth is boosted by a factor of $\sqrt{2}-1=41\%$ due to the fact that the real poles become complex poles and Q goes higher. Smaller f_0 will reduce this bandwidth boost effect quickly, and lead to peaking in frequency response as we have stated before.

Transimpedance limit. Previous analysis has shown the impact of output open-loop pole f_0. Recall that the product of gain and bandwidth of core amplifier, $A*f_0$, is roughly a constant in a given technology. As a result, for high-speed TIA where f_0 needs to be correspondingly high, A may be small and achievable R_F is rather limited. In fact, this relationship is best characterized by the famous transimpedance limit [5], [18], expressed as

$$R_T \approx R_F \leq \frac{Af_0}{2\pi C_{tot,in} BW_{-3dB}^2} \tag{3.22}$$

where the equal holds for Butterworth response. Consequently, R_F trades with bandwidth square, which renders shunt-feedback TIA a bit dismal for high-speed application. For example, from 10 Gb/s to 25Gb/s, bandwidth grows 2.5 times and R_F decreases by a factor of 6.25. Having realized the critical role to maintain a decent R_F for low noise, the transimpedance limit predicts a rapid noise growth in wider bandwidth, making the shunt-feedback TIA less attractive for high-speed application (>10Gb/s).

In advanced CMOS technology, the low V_{DD} further limits the attainable A with one stage. Cascading more gain stages, though boosting amplifier gain, makes it difficult to ensure sufficient phase margin. Thus usually one stage amplifier is utilized.

3.2.2.3 Active-feedback TIA

Since the role of R_F is to provide feedback current to input, or a transconductance function, it is natural to consider directly using a G_m device (transistor). Indeed, this topology is known as active-feedback TIA, a variation of shunt-feedback TIA described earlier, shown in Fig. 3.6.

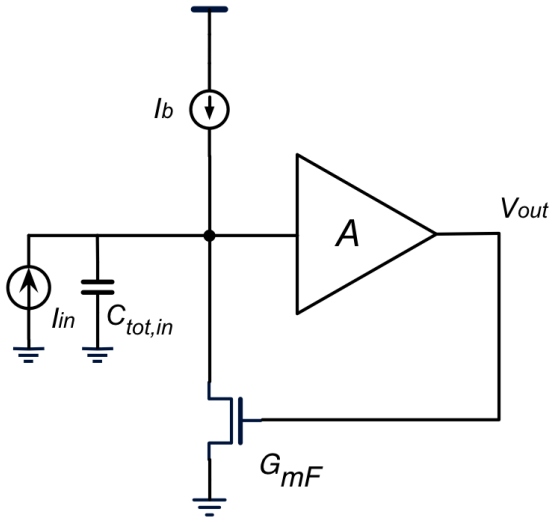


Fig. 3.6 Active-feedback TIA

The transimpedance thus equals $1/G_{mF}$ as long as A is large, while the amplifier should provide non-inverting output to maintain negative feedback. In this topology, the feedback transistor doesn't affect the open-loop output resistance, but adds capacitance to both input and output, which affect both bandwidth and noise. Furthermore, the transistor is less linear than resistor feedback [5].

3.2.3 Common gate TIA

Common gate stage is naturally a TIA if the input current goes to the source of the cascode device, shown in Fig. 3.7.

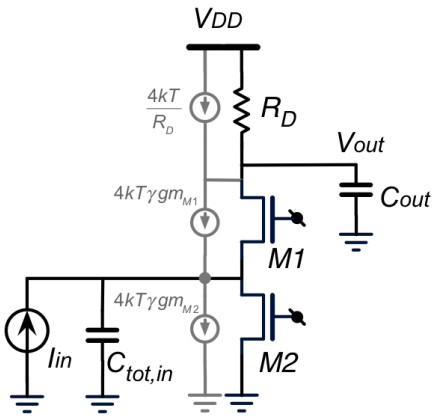


Fig. 3.7 Common gate TIA

The common gate stage is known for its low input impedance, which roughly equals $1/g_{m,M1}$. If output pole doesn't limit the overall bandwidth, the -3-dB bandwidth can be easily obtained at the input:

$$BW_{-3dB} \approx \frac{1}{2\pi R_{in} C_{tot,in}} = \frac{g_{m,M1}}{2\pi C_{tot,in}} \tag{3.23}$$

Since all signal current flows into the load resistor R_D, the transimpedance is

$$R_T = R_D \tag{3.24}$$

The thermal noise current from bias M2 and load resistor R_D are directly referred to the input, we have

$$\overline{I_{n,in}^2} = \overline{I_{n,M2}^2} + \overline{I_{n,R_D}^2} \tag{3.25}$$

$$= 4kT\gamma g_{m,M2} + \frac{4kT}{R_D} \tag{3.26}$$

Eq. 3.26 gives two indications: 1) noise of bias transistor and load resistor are referred to the input by unity gain; 2) noise of bias transistor and load resistor trade with each other, because for a given bias current, the overdrive voltage for bias transistor tends to be maximized for low g_m (noise), whereas this results smaller load resistor and increase its noise, especially in advanced CMOS technology where supply voltage is around 1V. In this sense, common gate TIA is not a good candidate for low-noise when large bias current is needed to ensure speed, where two problems are exacerbated more. Again we are facing the situation of noise-bandwidth trade-off.

On the other hand, high-frequency operation ability and less stability issue as a result of open-loop structure render the common gate TIA a more appealing approach in high-speed region even at the expense of noise. The ability to tolerate larger input capacitance also enables it for short-haul application where larger photodiode capacitance must be tolerated and sensitivity requirement becomes less an issue.

In practice, to alleviate the impact of output pole on bandwidth, the common gate TIA is followed by a shunt-feedback TIA, as shown in Fig. 3.8.

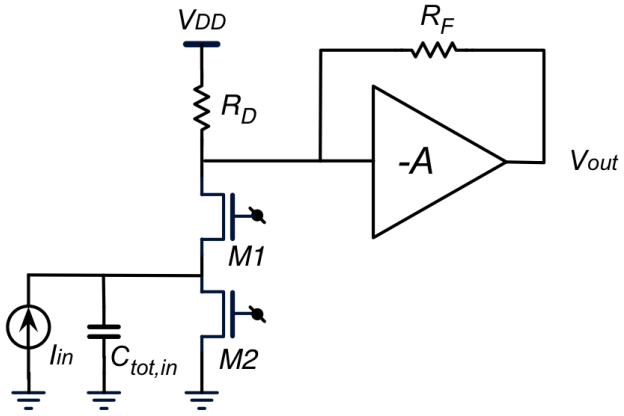


Fig. 3.8 Shunt-feedback TIA with common gate input stage

In this case, the common gate stage acts as a current buffer and the overall transimpedance gain roughly equals R_F, if R_D >> R_F/(A+1). In terms of noise, except the noise from common gate stage, other noise from shunt-feedback stage that we have analyzed before will all exhibit in the inputs and increase overall input-referred noise current.

A modification based on common gate TIA, so-called Regulated Cascode (RGC) [19], is shown in Fig. 3.9.

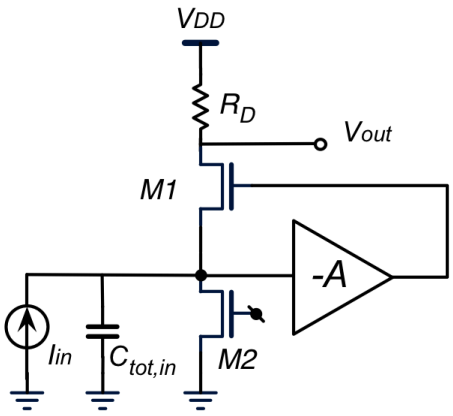


Fig. 3.9 Regulated Cascode TIA

The idea of this topology is by exploring shunt-series feedback, the input impedance is lowered by a factor of A+1, and we have

$$R_{in} \approx \frac{1}{g_{m,M1}(A+1)} \tag{3.27}$$

and thus even larger input capacitance can be tolerated. It should be noted that the original goal of this topology is to boost the output impedance by a factor of A+1 seen into the cascode, a popular gain-boost technique.

In the original design, a shunt-feedback TIA is following the RGC stage, similar to Fig. 3.8. Due to extra noise sources added, RGC is even nosier than common gate TIA. But superior high-frequency operation capability enables it utilized more in 10 Gb/s and higher speed.

To summarize this section, the increased bandwidth requirement poses serious difficulties on low-noise TIA design, due to the critical bandwidth-noise trade-off in virtually every topology. New techniques for low-noise TIA are highly desired and we will dedicate chapter 4 to address this issue.

3.3 Limiting amplifier

The role of the limiting amplifier (LA) is to provide extra gain to amplify the TIA output signal from tens of mV to several hundred mV. This number is in the range of 20~40 dB depending on TIA and CDR. Since noise from LA is referred to the input by transimpedance gain, it is less critical in terms of noise compared with TIA. More rigorously, the input-referred noise current considering both TIA and LA is given by

$$\overline{I_{n,in}^2} = \overline{I_{n,TIA}^2} + \frac{\overline{V_{n,LA}^2}}{|Z_T(f)|^2} \tag{3.28}$$

where $\overline{I_{n,TIA}^2}$ is the input-referred noise current of TIA and $\overline{V_{n,LA}^2}$ is the input-referred noise voltage of LA.

Although the name limiting indicates possible operation under large signal, when input signal is close to sensitivity, the limiting amplifier (or at least first several stages) still works in linear region, which dictates the need for sufficient

bandwidth. Not to impact overall receiver bandwidth set by TIA, the LA bandwidth needs to be set to around the data rate.

3.3.1 Gain-bandwidth product extension

For limiting amplifier operating at several tens of Gb/s, the overall gain-bandwidth product (*GBW*) far exceeds the capability of one stage amplifier, which roughly equals transit frequency (f_T) of the technology. For example, if the limiting amplifier needs to provide 25 GHz -3-dB bandwidth and 30 dB gain, the consequential *GBW* is 790 GHz, much higher than any of the current CMOS technology can do.

On the other hand, multi-stage amplifier is able to solve this because the overall gain from cascading grows faster than the bandwidth shrink and thus we are able to boost the *GBW*. To quantify this effect, we examine $GBW_{\text{tot}}/GBW_{\text{s}}$, the *GBW* extension ratio where GBW_{tot} is the total *GBW* and GBW_{s} is the stage *GBW*. From [5], if cascaded amplifier stage has first-order response, we have

$$\frac{GBW_{tot}}{GBW_s} = A_{tot}^{1-1/n} \cdot \sqrt{2^{1/n} - 1} \tag{3.29}$$

where A_{tot} is the total gain and n is the number of cascaded stages. For cascaded second-order Butterworth stages, it becomes:

$$\frac{GBW_{tot}}{GBW_s} = A_{tot}^{1-1/n} \cdot \sqrt[4]{2^{1/n} - 1} \tag{3.30}$$

Assume A_{tot} is 30 dB and both expressions are plotted in Fig. 3.10. From the plot, to maximize the *GBW* extension ratio, n is 7 for first-order cascading and 14 for second-order Butterworth cascading. In practice, the stage response is more like second-order, e.g. Butterworth response, and the optimum is 14-stages. However, 14-stage is seldom adopted because: 1) power consumption is excessive; 2) input-referred noise becomes larger since gain of first few stages goes lower; 3) as n goes larger than 6, the *GBW* extension becomes marginal. Based on these consideration, n of 4~6 is usually adopted. For example, if we choose n = 5, the *GBW* extension ratio for Butterworth response is still around 10, but power consumption is essentially less than halved. Using the practice we have done earlier, which needs GBW_{tot} of 790 GHz, GBW_{s} is 79 GHz and stage gain is 6 dB, which is OK for modern 65-nm CMOS technology with f_T of around 200 GHz.

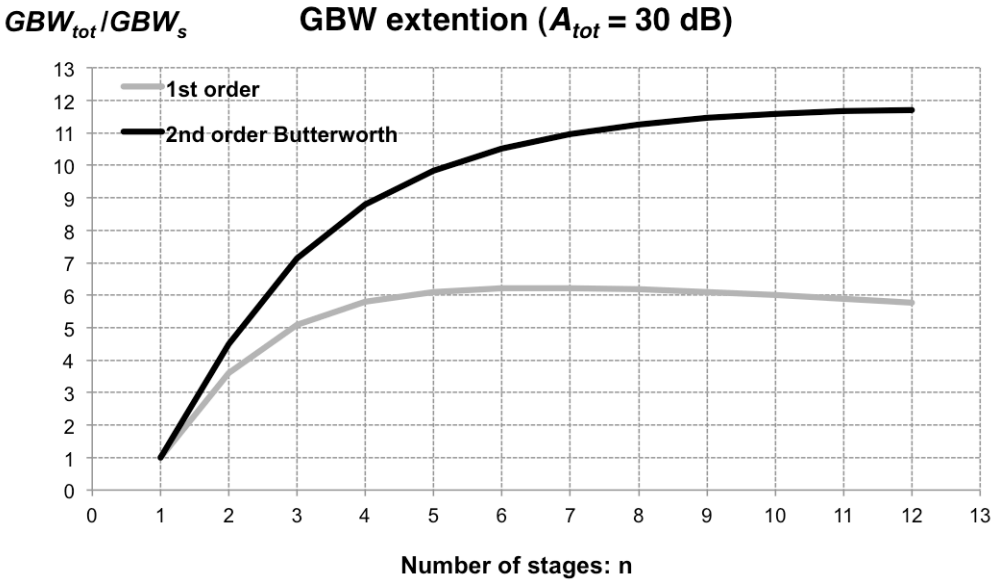


Fig. 3.10 *GBW* extension from single stage to multi-stage (A_{tot} = 30dB)

3.3.2 Bandwidth extension

Review of modern CMOS. In last section, we have arrived at the conclusion from the architecture level that a technology with f_T larger than 79 GHz is enough. To gain more insights at circuit level, let's review the *GBW* product assumption we have used for a single transistor, RC loaded amplifier, which dictates that:

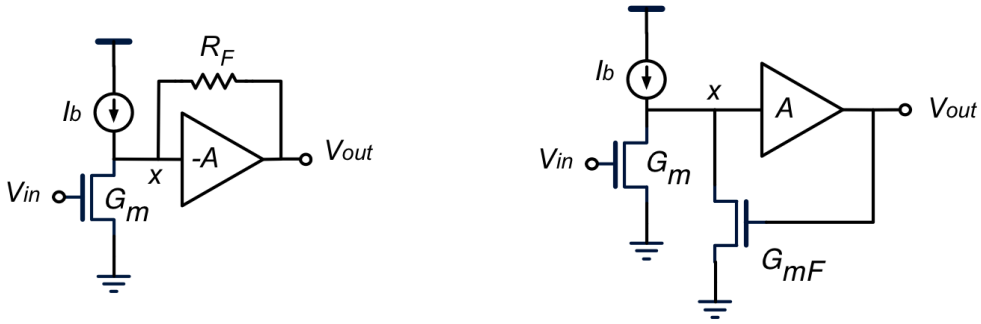
$$A * BW \approx f_T \tag{3.31}$$

where A is the voltage gain, BW is the -3-dB bandwidth and f_T is the transit frequency. However, the modern advanced CMOS technology suffers from several drawbacks that make achievable *GBW* much lower than f_T.

- Miller capacitance. In advanced CMOS, C_{gd}/C_{gs} is around 0.5, and the resulting miller capacitance could be even larger than original gate capacitance, lowering the *GBW* significantly.
- Parasitic capacitances. As the technology miniaturization proceeds, while the intrinsic capacitance scales down, the parasitic capacitances from wiring become comparatively larger, which considerably lowers *GBW*. Furthermore, C_{ds} and C_{dg} that are neglected in f_T calculation, also reduce *GBW* somewhat.
- Low power supply. The technology f_T is a bias dependent parameter, which peaks only at very high bias current density. The voltage headroom from low supply voltage usually refrains the transistor from biased at peak f_T meaning we could only use downgraded transistor.
- Lack of good buffer. Source follower, which could lower the capacitance loading for following amplifier is difficult to be adopted in low power supply.

Given all the above factors, the actual *GBW* is significantly scaled down from peak f_T by a factor of $2 \sim 4$. Therefore, we still need bandwidth extension techniques at circuit level. For the rest of this section, we briefly review some popular bandwidth extension techniques.

Shunt-feedback. Shunt-feedback transimpedance stage is not only a good TIA, but also can be also used as a wideband TIA load. The well-known Cherry-Hooper [20] amplifier and active-feedback [21] both employ this characteristic.



(a) (b)

Fig. 3.11 Shunt-feedback TIA as load: (a) Cherry-Hooper amplifier, (b) Active-feedback.

In both cases, if the feedback factor R_F and $1/G_{mF}$ equal the original load, the overall gain doesn't change. By introducing TIA load, the amplifier becomes second-order from first-order RC load. Assume a Butterworth response and use the result obtained earlier; it can be easily shown that

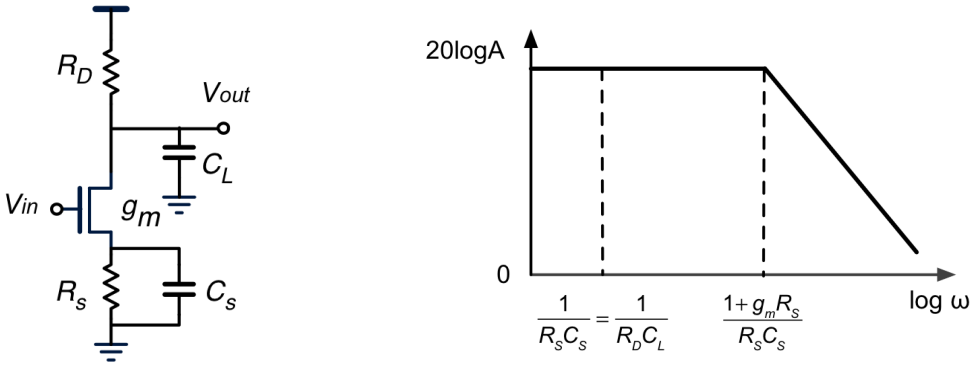
$$BW_{SF} \approx \sqrt{\frac{f_T}{BW}} \cdot BW \tag{3.32}$$

where BW is the -3-dB bandwidth with simple RC load and BW_{SF} is that for TIA load, assuming the capacitance at node X doesn't change. In practice, extra parasitics, limited amplifier gain and other effects stated in last section will make the bandwidth extension smaller than Eq. 3.32 would predict.

It should be noted that the core amplifier has different output polarity in the two cases to maintain negative feedback. Furthermore, in active-feedback amplifier, the gain is set by G_m/G_{mF}, making it more robust to PVT variation than resistor loaded amplifier.

Capacitive degeneration. Another way to speed up the amplifier is by introducing a zero, known as the capacitive degeneration. The half-circuit from differential is plotted in Fig. 3.12 in conjunction with the frequency response. It can be shown that, by the introduction of R_S and C_S, the frequency response becomes [7]

$$A(s) = \frac{g_m R_D}{1 + g_m R_S} \frac{(R_S C_S s + 1)}{(R_S C_S s + 1 + g_m R_S)(R_D C_L s + 1)} \tag{3.33}$$



(a) (b)

Fig. 3.12 Capacitive degeneration [7]: (a) half-circuit, (b) frequency response

If the zero at $1/(R_SC_S)$ is placed to cancel the output pole at $1/(R_DC_L)$, the overall bandwidth can be boosted by a factor of $1+g_mR_S$. On the other hand, the source degeneration also lowers the gain by the same factor and the overall GBW doesn't change.

Another advantage from this approach is that the input capacitance is lowered by Miller effect. Since the voltage gain from gate to source is $g_mR_S/\ (1+g_mR_S)$, C_{gs} which acts as a Miller capacitance referred to the input becomes

$$C_{gs,in} = \frac{C_{gs}}{1 + g_mR_S} \tag{3.34}$$

and thus the input capacitance is lowered to speed up its previous stage. This input capacitance reduction effect can also be utilized to build the f_T-doubler [7], shown in Fig. 3.13.

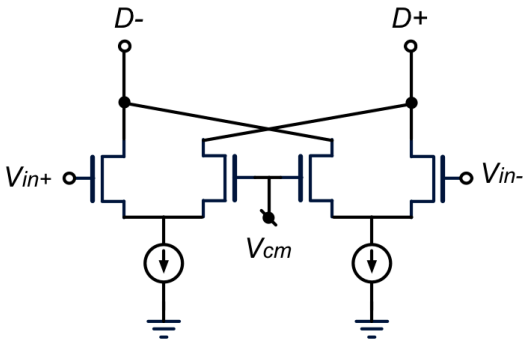
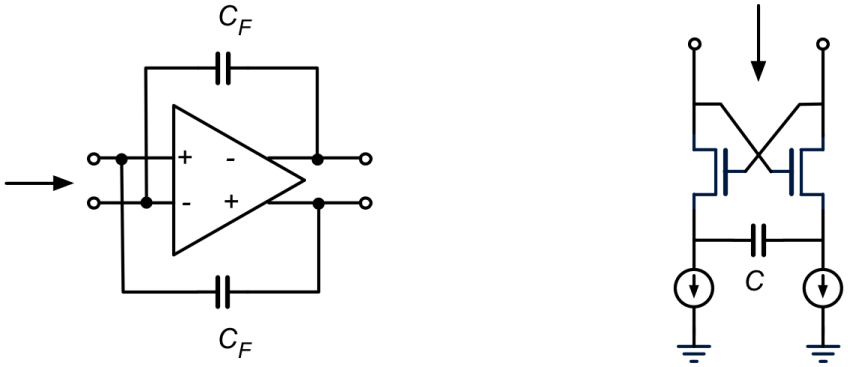


Fig. 3.13 f_T-doubler

If the four transistors are identical, the voltage gain from either input to source node is 1/2. Based on Miller theory, the capacitance seen into the either input is around $C_{gs}/2$. Though the transconductance from the input differential pair is also halved, the common gate transistor compensates the other half and the overall transconductance seen from each output equals that of a pure differential pair. In such a way, the f_T from the combo is roughly doubled.

In reality, capacitance from current tail and other parasitics raise the input capacitance and lower the actual f_T. Other downsides include twice the power consumption and more capacitance to output.

Negative capacitance. Based on Miller theory, if the voltage gain is positive and larger than unity, we are able to create a negative capacitance to input and cancel part of the input capacitance [21], shown in Fig. 3.14. However, as frequency goes higher, the amplifier gain drops and the outputs and inputs are more mismatched in phase, making this technique less effective.



(a) (b)

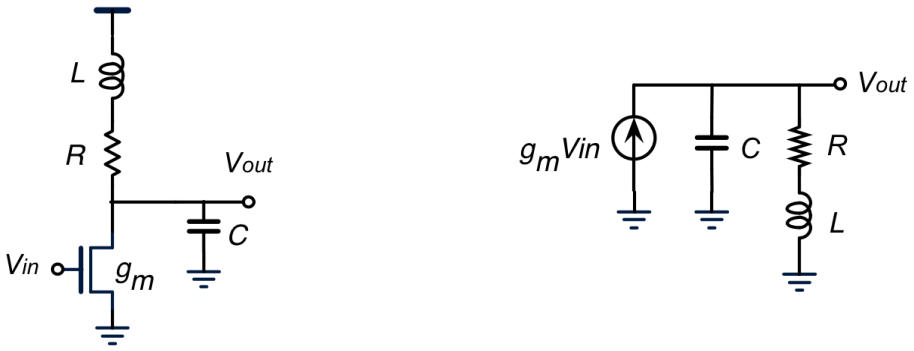
Fig. 3.14 Negative capacitance by (a) Miller effect and (b) NIC

Another way to create negative capacitance is by employing negative impedance converter (NIC) [21], comprised by a cross-coupled pair with a capacitance shunting their sources, as shown in Fig. 3.14. In this circuit, the NIC swaps the terminal voltage across the capacitor while keeping the same charging current. Thus looking from the drain, it is like charging a negative capacitor.

In both techniques, the effective negative capacitance should be controlled under certain range so that the overall input capacitance doesn't become negative and causes the amplifier unstable.

Inductive peaking. The monolithic inductor has opened up new possibilities for wideband circuit design. Intuitively, inductor is able to tune out capacitance to boost bandwidth. Among numerous ways shunt peaking is a popular one. The basic circuit and small-signal equivalent are shown in Fig. 3.15.

The topology of shunt peaking is essentially the simple RC load in addition of an inductor. The role of the inductor is to resonate with the capacitance at load at high frequency, thus boosting overall high-frequency impedance, or in another way, the bandwidth. In time domain, if a step current comes in, the inductor initially blocks the current from going into the resistor and force it charging the capacitor. Thus the output voltage alters faster than pure resistor load. More rigorously, the small signal voltage gain transfer function is given by [7]



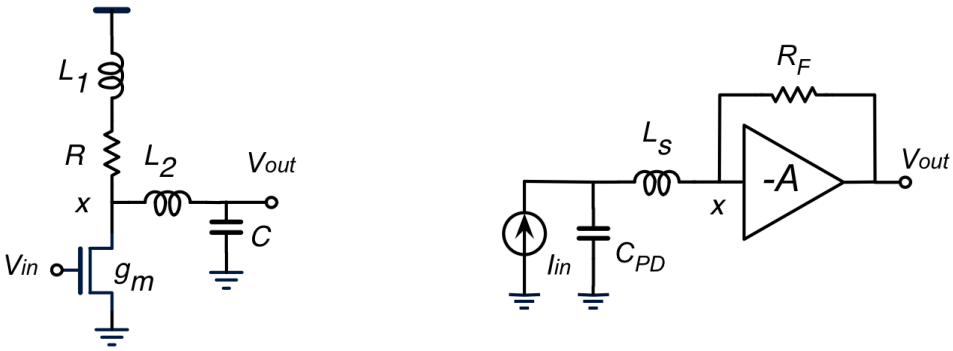
(a) (b)

Fig. 3.15 Shunt peaking: (a) circuit, (b) small-signal equivalent

$$\frac{V_{out}}{V_{in}} = -g_m R \cdot \frac{s + 2\varsigma\omega_n}{s^2 + 2\varsigma\omega_n s + \omega_n^2} \cdot \frac{\omega_n}{2\varsigma} \tag{3.35}$$

where $\varsigma = R/2\sqrt{C/L}$ and $\omega_n = 1/\sqrt{LC}$. Thus, the simple RC is transformed to a two-pole and one-zero second-order system and the bandwidth boost comes from the zero and the complex poles. Assuming 5% overshoot, a bandwidth extension factor of 78% can be obtained showing the power of this technique. In practice, inductor practices limits this number to around 50% [7]. Note that the inductor in shunt peaking can also be realized by active device but suffers from noise and large voltage drop [5], making it difficult to be utilized in scaled technology with low supply voltage.

Another way to use the inductor is series peaking and can be employed by both TIA and LA, as shown in Fig. 3.16.



(a) (b)

Fig. 3.16 use of series peaking: (a) shunt series peaking [22], (b) series peaking in TIA [7]

The series inductor transforms the one-port load network to the two-port interstage network providing more bandwidth extension capability [39]. However, with more pole(s) added the frequency response is also degraded, with possible ISI in time domain [7].

Based on shunt peaking and series peaking, numerous peaking methods have been invented [21], [23], extending bandwidth considerably by a factor of 2~4. In the extreme case, distributed amplification can boost bandwidth to infinity in

the ideal case. Practical limits from lossy transmission line, finite output impedance of transistor, propagation mismatch and Miller effect lower the bandwidth extension to a finite but still considerable number. Certainly, distributed amplifier gains bandwidth from large power consumption and area.

Finally, it should be noted that these techniques are employed either standalone or jointly to obtain greater bandwidth extension capability. These techniques can be equally utilized by TIA and Output buffer.

After introducing several kinds of bandwidth extension techniques, what will the limiting amplifier possibly look like? Fig. 3.17 provides a design perspective.

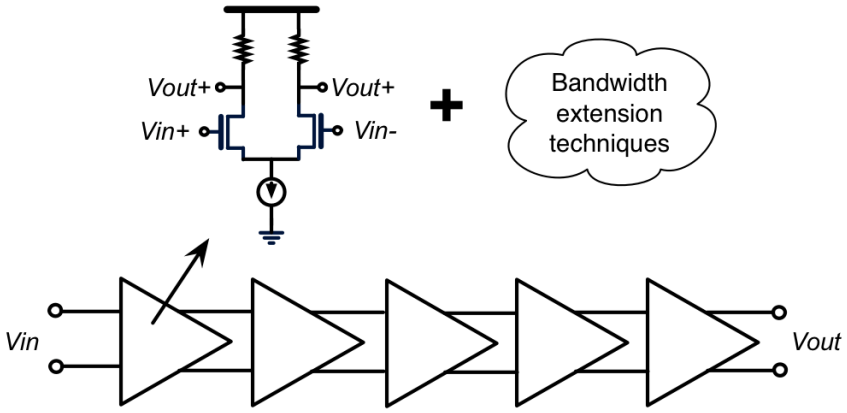


Fig. 3.17 Limiting amplifier: a design perspective

- Differential signaling. Almost all modern CMOS LAs are differential signaled, due to its superior common mode supply/ground noise rejection capability.
- Core LA stage is basically a resistor (or PMOS operating in linear region) loaded differential pair, which minimizes parasitics and provides high-speed.
- Several bandwidth extension techniques co-work to gain satisfied high-frequency operation capability.

3.4 Output buffer

For Output buffer, gain and noise are less concerned but bandwidth needs to be assured for the same reason as LA. Since Output buffer drive 50-Ω (100-Ω differential) off-chip load, output matching and large current driving capability are critical. Therefore, a tapered topology [7] is often necessary, as shown in Fig. 3.18.

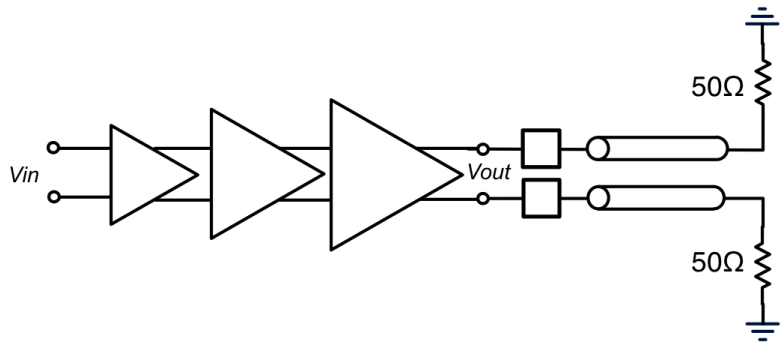


Fig. 3.18 Tapered Output buffer (coupling capacitor not shown)

The basic stage is usually the resistor loaded differential pair like LA. The first stage of Output buffer is also the load for LA, thus a relative small transistor width should be used to not overload LA. As signal proceeds, the transistor width gradually increases. For the last stage, to have good matching, 50-Ω resistor load is usually mandatory. To achieve an output swing of 1 $V_{diff,pp}$, the bias current should be around 20 mA because half of the current charges off-chip load, which doubles the current consumption. Therefore, the last stage uses large transistor width and bias current, which dominates the power consumption of Output buffer.

A disadvantage of using tapered topology is that each stage should drive larger transistor them itself. From the analysis in LA section, this means a downgrade of *GBW*. For example, if each stage drives 2x larger following stage, *GBW* is scaled down by a factor of two. However, since Output buffer doesn't require voltage gain, this is not a problem. Furthermore, the numerous bandwidth extension techniques explained for LA generally apply also to Output buffer, rendering bandwidth not a problem.

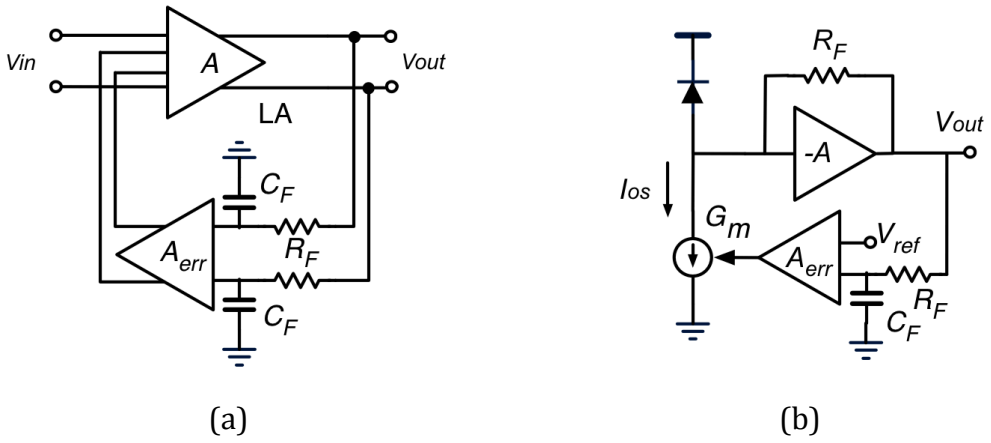
The presence of output pad makes matching difficult since it contributes several tens of fF capacitance. Consider also the large device used for the last stage, the total capacitance is even larger. Furthermore, other package parasitics that could also lower the output impedance at high frequency. Hence, the load resistor can be made larger than 50 Ω, in the range of 75~100 Ω [7], to provide better matching for the whole band. In the mean time, power consumption is lowered since a lower bias current is required now for the same voltage swing.

3.5 Other design issues

3.5.1 Offset correction

There are two kinds of offset in optical receiver: 1) transistor offset. This is because optical receiver has high gain while advanced CMOS has large 3σ offset voltage. 2) Photo-current induced offset. It is introduced by the DC component of photocurrent. When TIA operates near overload limit, a large DC current flows though the feedback resistor and completely changes the bias of TIA even when the TIA is single-ended. For example, a 1mA DC current goes through a 500 Ω feedback resistor generates a 500 mV IR drop, the g_m device is thus driven into

deep-triode region. As a result the feedback is destroyed and heavily non-linear distortion will appear.



(a) (b)

Fig. 3.19 Offset correction in: (a) LA, (b) TIA

Fig. 3.19 shows the principle of offset correction for both LA and TIA where the idea behind is similar: close the loop in low-frequency with a large feedback factor, thus the input offset voltage/current is reduced by a factor of low-frequency loop gain.

In Fig. 3.19 (a), suppose LA's core amplifier has an input offset voltage V_{OS} and the error amplifier has an input off V_{OSE}, the overall input offset V_{OS}' after correction becomes [5]

$$V_{os}' \approx \sqrt{\left(\frac{V_{os}}{AA_{err}}\right)^2 + \left(\frac{V_{OSE}}{A}\right)^2} \tag{3.36}$$

where A is forward amplifier gain and A_{err} is the gain of error amplifier. V_{OS} and V_{OSE} are assumed statistically independent. Since V_{OSE} is suppressed less than V_{OS}, it is important to keep it low, which can be done by using large transistor dimension. Some times, the error amplifier can be omitted from the loop (V_{OSE} = 0 and A_{err} =1), the topology becomes a low-frequency buffer and the input offset equals output offset.

Since the feedback loop only works at low-frequency, the closed-loop gain goes from $1/A_{err}$ at low-frequency to A at high frequency, by a factor of $A \cdot A_{err}$. Thus a zero and pole is introduced to the transfer, as shown in Fig. 3.20.

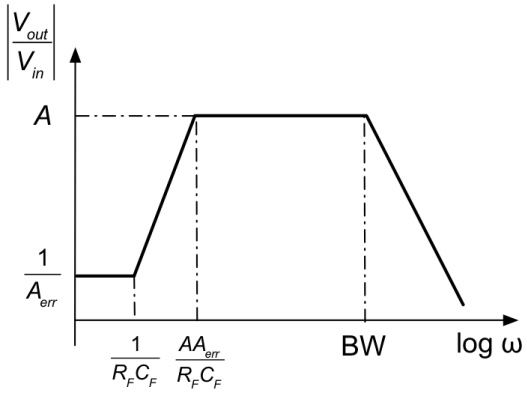


Fig. 3.20 Frequency response with offset correction

The low-frequency cutoff $A*Aerr\ /(R_FC_F)$ should be kept low to avoid baseline wander. Since many communication standards mandate very small low-frequency cutoff to a few kHz, the value of passive component becomes excessively large, requiring possibly off-chip capacitor. Another way to get large capacitor is by employing Miller effect: C_F is connected between the input and output of error amplifier to realize large input capacitance.

The offset correction technique for LA can equally apply to TIA shown in Fig. 3.19 (b) and the feedback becomes current. The common mode reference voltage (V_{ref}) can be obtained either from a dummy TIA output without input signal, or directly taken from the TIA input node if there is no dummy TIA available.

In both cases, stability criteria should be met. Suppose the error amplifier has one pole, to have around 60° phase margin, pole frequency of error amplifier should be set larger than $2A*Aerr/(R_FC_F)$.

3.5.2 Gain control

If the optical receiver needs a large dynamic input range, we should take care of the situation for input signal not only as low as sensitivity, but also close to overload limit, on the range of 1~2 mA_{pp}. At such a large input signal, both the TIA and LA will work heavily non-linear where ISI, distortion and jitter significantly close the output eye. Accordingly, either TIA or LA may need certain gain control mechanism and we explain both situations here.

This gain mechanism shunt-feedback TIA is shown in Fig. 3.21. Since the transimpedance gain is determined by the feedback resistor, gain control can be realized by alter the feedback resistance. However, this will raise the open-loop input pole frequency and lead to instability. According to Eq. 3.18, amplifier gain needs to change to the same extent to maintain the same Q factor.

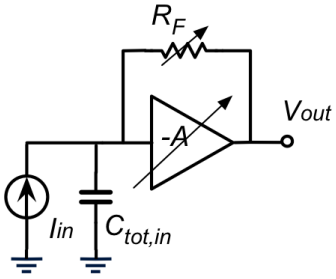
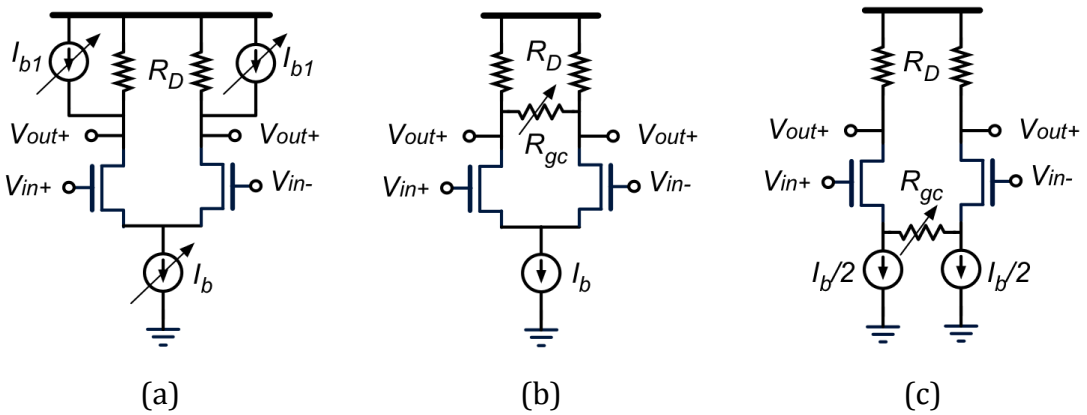


Fig. 3.21 TIA gain control

A shunt NMOS transistor can be added to R_F to realize variable resistance. The gain control of core amplifier is similar to the situation of gain control of LA, which we will describe in a moment.

The gain control can be done either by external tuning or automatically, known as the automatic gain control (AGC) and in this case, a detection mechanism is needed. Since the DC component of photocurrent is proportional to the peak-to-peak input swing assuming a high distinction ratio of the input signal, the offset correction method for TIA described before may be used to generate the gain control signal [7].

LA stage with gain control makes it a variable gain amplifier (VGA) and most popular ways are shown in Fig. 3.22, including [5]: 1) vary g_m; 2) vary load resistance; 3) series feedback.



(a) (b) (c)

Fig. 3.22 VGA gain stage: 1) vary g_m; 2) vary load resistance; 3) series feedback.

- Vary g_m. It is realized by varying bias current I_b. To maintain a constant common-mode output voltage, $2I_{b1}$ needs to be altered by the same amount. The shortcoming here is that when bias current is decreased, the linear input range is also decreased which opposes large signal operation.
- Vary load resistance. A variable resistor R_{gc} is placed shunting the two outputs so that it only impacts the differential gain without impacting the output common mode voltage. The variable resistor can be implemented by NMOS in linear region. The problem with the topology is that the bandwidth increases as gain decreases, making more noise captured.
- Series feedback. R_{gc}. The bias current is spitted into two parts and the series feedback resistor R_{gc} is connected to the source of the differential pair to realize source degeneration. The disadvantage of this technique is that the noise from bias current is no longer common-mode and contribute to the total noise.

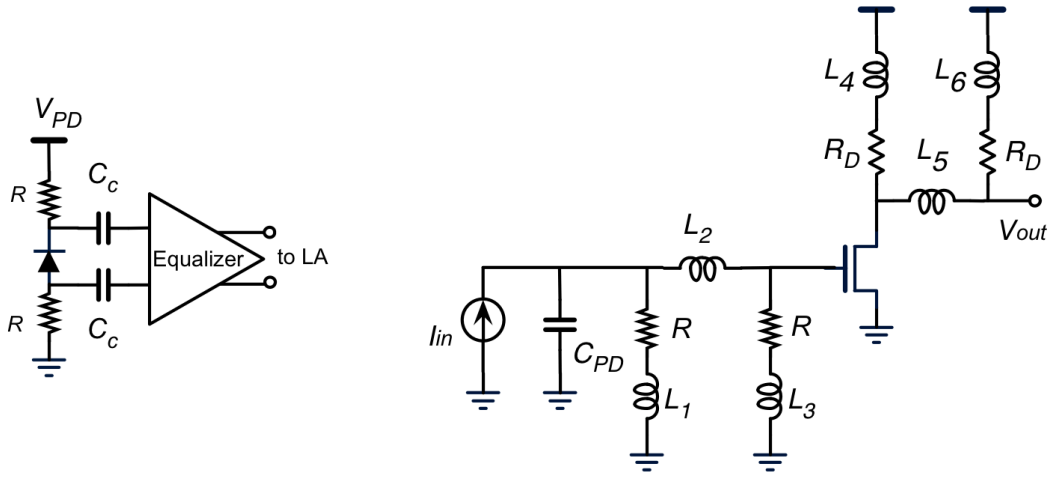
Since there is no correlation between the DC offset and input signal for LA, a peak detection mechanism is required, which can be found in [5], [24], [25].

3.6 State-of-the-art

We briefly review some recent works on low-noise high-speed (≥ 10 Gb/s) CMOS optical receiver design with a particular focus on TIA, which plays the decisive role on noise performance.

3.6.1 Resistor TIA

As has been stated before, resistor TIA suffers from severe noise-bandwidth trade-off in high-speed. With the advancement of silicon photonics, Ge-photodiode presents itself extremely small capacitance (~ 10 fF), more than 10x smaller than the discrete III-V photodiode whose capacitance is around 150 fF for 10Gb/s application. In this sense, the resistor TIA is adopted in a 10 Gb/s silicon photonics optical receiver [16], shown in Fig. 3.23 (a).



(a) (b)

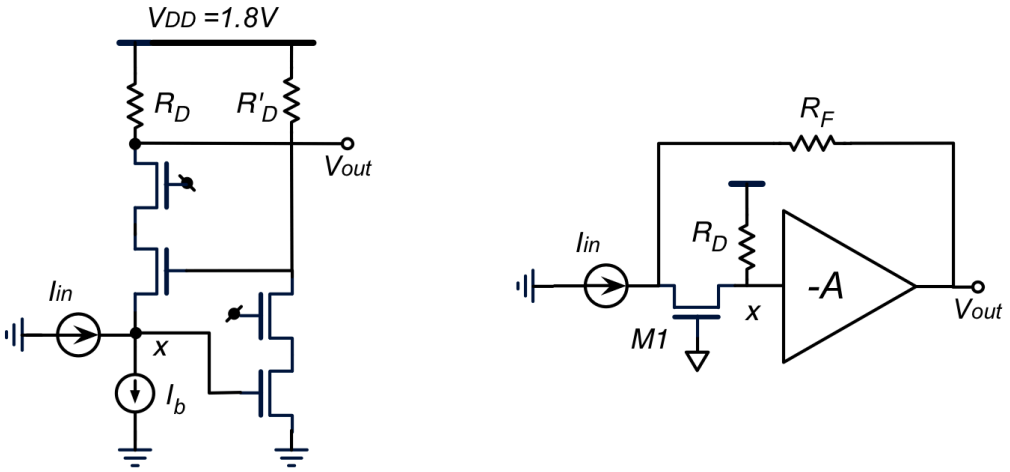
Fig. 3.23 Resistor TIA reported on (a) [16], (b) [23]

The design employs a differential sensing structure, which gives 3 dB advantages on sensitivity. AC coupling is necessary to separate bias for photodiode and following stages, at the price of bottom-plate parasitic capacitance from the coupling capacitor C_C. An equalizer following the resistor TIA enables larger resistor value for better noise performance. The receiver is fabricated in a 0.13-μm SOI CMOS, showing 6 μA_{pp} input sensitivity, which translates to an input-referred noise current spectral density of 5.1 pA/$\sqrt{Hz}$ assuming a 7 GHz -3-dB bandwidth.

In [23], a π-inductor peaking network for bandwidth improvement is used for both resistor TIA and wideband gain stage (x4) to enable 40 Gb/s operation in 0.18-μm CMOS, shown in Fig. 3.23 (b), where only one stage of the gain stage is shown. The design is optimized for 50 Ω input match, resulting a large average input-referred noise current spectral density of 55.7 pA/$\sqrt{Hz}$, more than 3x of pure 50 Ω resistor (18 pA/$\sqrt{Hz}$). This shows the importance of a large transimpedance necessary to suppress noise from following stages.

3.6.2 Common gate TIA

Common gate TIA has become popular from 10 Gb/s to 40 Gb/s due to its good high-frequency operation capability. In [27], the RCG topology is chosen for a 25 Gb/s TIA fabricated in 65-nm CMOS, shown in Fig. 3.24 (a). Both the common gate stage and common source stage utilize cascode to reduce miller capacitance. Consequently, a higher V_{DD} (1.8V) is necessary to provide enough voltage headroom. In [28], a topology modified from RGC is reported, where a common gate transistor is inserted between node X and the gate of feedback amplifier to increase the gain of feedback amplifier in a low V_{DD} (1V).



(a) (b)

Fig. 3.24 Common gate TIA reported on: (a) [27] (b) [25]

In Fig. 3.24 (b), a modified common gate TIA is reported in 40 Gb/s optical receiver fabricated in 90-nm CMOS [25] and only the small signal circuit is shown here. The main difference from conventional "common gate + shunt-feedback" is that the feedback point moves from x to the input. In such a way, the input resistance $1/g_{m,M1}$ is lowered by roughly a factor of $1+AR_D/R_F$, allowing a smaller bias current and thus lower noise. A Reversed Triple-Resonance Network (RTRN) is utilized to boost bandwidth in x (not shown in the Figure). Compared with RGC, since the feedback amplifier to lower the input impedance in the common gate stage is eliminated, it gives better noise performance. The average input-referred noise current spectral density is 22 pA/$\sqrt{Hz}$ assuming different sensing.

3.6.3 Shunt-feedback TIA

Shunt-feedback is becoming more and more difficult to use in higher than 10 Gb/s, unless a very low photodiode capacitance (e.g. Ge-photodiode) or very fast technology is available.

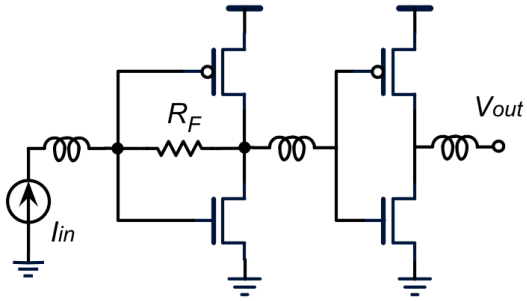


Fig. 3.25 Shunt-feedback TIA reported on [29]

In [29], a 40 Gb/s TIA based on shunt-feedback is realized in 40-nm SOI CMOS whose peak f_T is close to 400 GHz. Push-pull structure is adopted to boost amplifier gain to 4.5 (13 dB). Due to the superior transistor performance, open-loop amplifier bandwidth is still sufficient not to hurt the closed-loop response. A π-network is utilized to boost bandwidth with minimum group-delay variation.

The receiver shows average input-referred noise current spectral density of 20.5 pA/$\sqrt{Hz}$.

In [30], a fully-integrated silicon photonic transceiver is demonstrated in 25 Gb/s and fabricated in a 0.13-μm SOI CMOS. The TIA employs active-feedback topology (cf. Fig. 3.6) and gives an average input-referred noise current spectral density of 40 pA/$\sqrt{Hz}$.

3.7 Conclusion

In this chapter, we have reviewed the main building blocks of CMOS optical receiver circuits, including TIA, LA and Output buffer. Design considerations with respect to the signal integrity issues motioned in previous chapter like bandwidth, gain, noise, jitter and stability are exemplified in circuit design. Furthermore, we understand virtually very TIA topology faces critical bandwidth and noise trade-off, rendering low-noise especially difficult to obtain in high-speed regime. In addition, system level consideration as offset and gain control are discussed. Finally, state-of-the-art TIA design are explored.

Chapter 4

Low noise design techniques

In previous chapter, we have seen that virtually every TIA topology suffers from the critical noise-bandwidth trade-off that low-noise TIA becomes increasing difficult to realize in high-speed regime. In this chapter, we first give an in-depth analysis of the TIA noise composition and the existing noise optimization approach, then we will develop several new low-noise design techniques, which together enables low-noise front-end for high-speed optical receiver.

4.1 TIA Noise analysis

As has been shown in last chapter, shunt-feedback TIA is most favorable in terms of noise but shows limitation as speed goes higher. We therefore focus on this topology as the base for low-noise design. A typical shunt-feedback TIA schematic is shown in Fig. 4.1, where a common source amplifier followed by a source follower form the core amplifier.

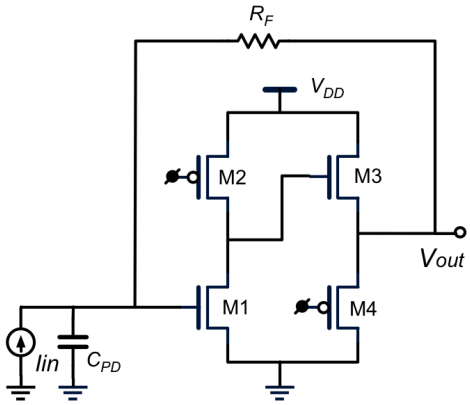
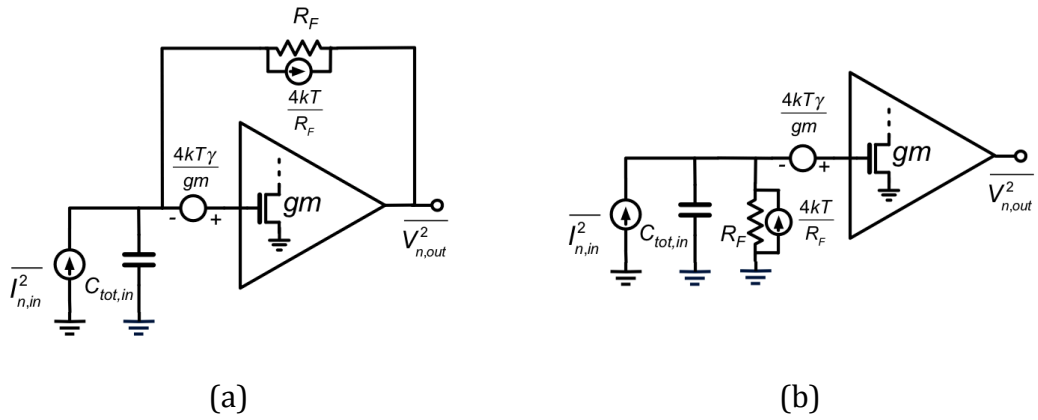


Fig. 4.1 A typical shunt-feedback TIA

Recall that Eq. 3.15 has given the expression of low-frequency input-referred noise current for shunt-feedback TIA. Here, a complete picture of input-referred noise current spectrum is desired since the input impedance (formed by the total input capacitance) decreases at high frequency.



(a) (b)

Fig. 4.2 Noise from shunt-feedback TIA: (a) in closed-loop; (b) in open-loop

The main noise sources of the shunt-feedback TIA from Fig. 4.1 are shown in Fig. 4.2 (a), assuming that the thermal noise of the g_m device dominates in the amplifier. Since feedback doesn't change input-referred noise [31], it is more convenient to obtain the expression of noise from the open-loop configuration shown in Fig. 4.2 (b): R_F noise current direct contributes to input-referred noise current; amplifier input-referred noise voltage refers back to input by a low-pass network formed by $C_{tot,in}$ and R_F, and thus becomes a high-pass input-referred noise current. This can be also verified by more complicated closed-loop approach [32], [33]: evaluate the output noise of each noise source and divide them by the square of the magnitude of closed-loop transimpedance transfer function. In both cases we have

$$\overline{I^2_{n,R_F}(f)} = \frac{4kT}{R_F} \tag{4.1}$$

$$\overline{I^2_{n,amp}(f)} = \frac{4kT\gamma}{g_m} \Big/ \left|\frac{R_F}{1+sR_FC_{tot,in}}\right|^2 \tag{4.2}$$

$$= \frac{4kT\gamma}{g_m R_F^2}\left|1+sR_FC_{tot,in}\right|^2 \tag{4.3}$$

where $\overline{I^2_{n,R_F}(f)}$ and $\overline{I^2_{n,amp}(f)}$ are the input-referred noise current of R_F and amplifier, respectively. As expected, R_F noise is white and amplifier noise has a DC component and a second-order term. Using Eq. 3.16, the output noise voltage of R_F and amplifier are given by

$$\overline{V^2_{n,R_F}(f)} = \frac{4kT}{R_F}|H(s)|^2 \tag{4.4}$$

$$\approx \frac{4kTR_F}{\left|\frac{s^2}{\omega_n^2}+\frac{s}{\omega_n Q}+1\right|^2} \tag{4.5}$$

$$\overline{V^2_{n,amp}(f)} = \frac{4kT\gamma}{g_m R_F^2}\left|1+sR_FC_{tot,in}\right|^2|H(s)|^2 \tag{4.6}$$

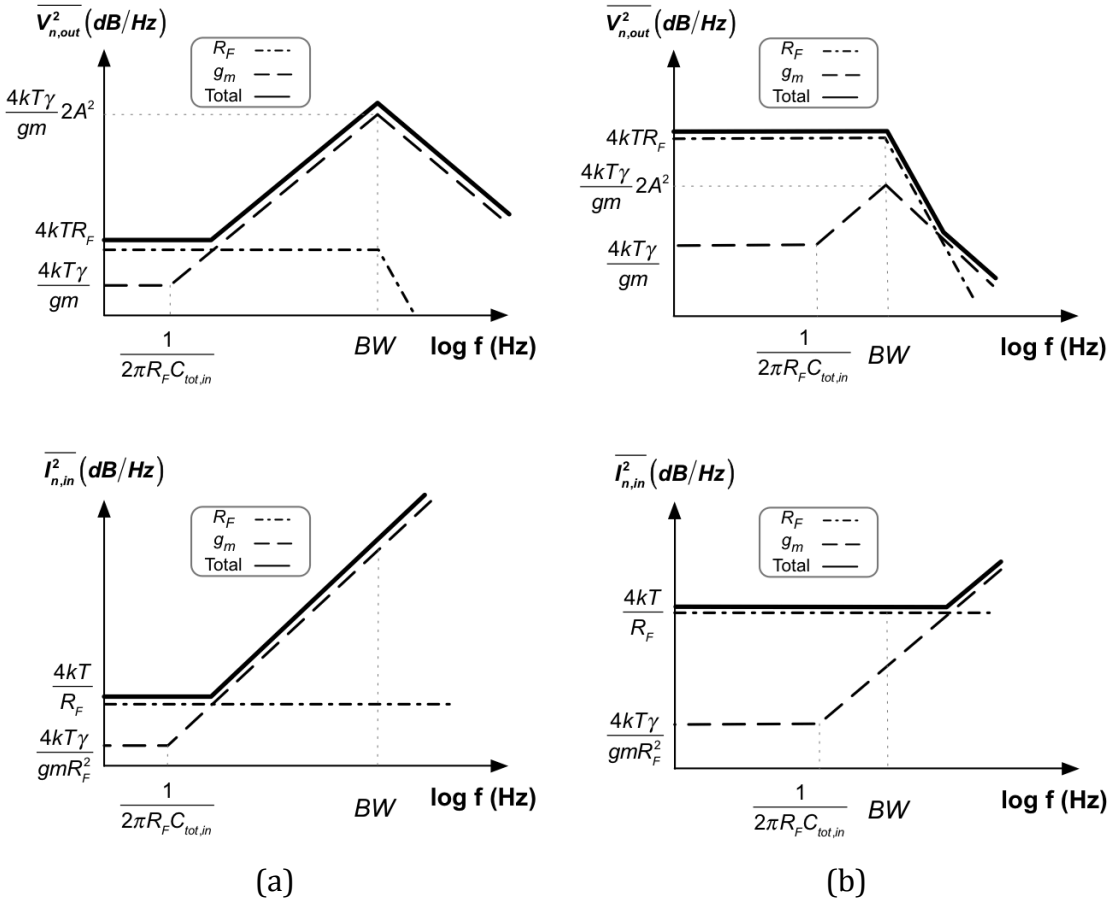
$$\approx \frac{4kT\gamma}{g_m} \frac{\left|1 + sR_F C_{tot,in}\right|^2}{\left|\frac{s^2}{\omega_n^2} + \frac{s}{\omega_n Q} + 1\right|^2} \tag{4.7}$$

where all the parameters are defined the same as in Section 3.2.2, and assume $A >> 1$, thus $A/(A+1) \approx 1$. With the expression of both input-referred and output noise, we can plot the noise spectra in Fig. 4.3. Assume the TIA has a Butterworth response, Eq. 3.20 and 3.21 are re-written as

$$\omega_n = \omega_{-3dB} \approx \frac{\sqrt{2}A}{R_F C_{tot,in}} \tag{4.8}$$

Depending on the application, there are two different scenarios of noise profiles for shunt-feedback TIA, as shown in (a) and (b) of Fig. 4.3.

a) "Low-speed" TIA. This means the TIA -3-dB bandwidth is relatively small compared with the technology f_T. In this case, there is sufficient amplifier voltage gain to allow a relatively large R_F whose noise contribution becomes small. On the other hand, the input-referred amplifier noise exceeds that of R_F in-band, making its contribution significant. Therefore in this scenario, main effort of noise optimization is usually on the amplifier.
b) "High-speed" TIA. In this scenario, the TIA -3-dB bandwidth is close to technology f_T and hence available amplifier voltage gain A is fairly small. This on one hand leads to small R_F (larger R_F noise), and on the other hand means a relatively smaller ramp-up of input-referred amplifier noise in-band. Hence in this scenario, more attention should be paid on R_F noise.



(a) (b)

Fig. 4.3 Input-referred and output noise spectrum in Shunt-feedback TIA in two scenarios: (a) "low-speed" TIA (b) "high-speed" TIA

From Fig. 4.3 and the noise expression we have developed so far, it is clear that noise optimization in the input is easier than that exercised in the output. Furthermore, input-referred noise directly relates to sensitivity (c.f. Eq. 3.4). Our following analysis thus is carried out on the optimization of input-referred noise current, trying to find the minimum integrated input-referred noise current.

4.2 Conventional noise optimization

Conventionally, noise optimization of shunt-feedback TIA is dedicated to amplifier noise (assuming dominated by g_m noise) [5], [33], [34] based on the assumptions that R_F is large enough that its noise is negligible, which assumes scenario (a) of Fig. 4.3.

Expand input-referred noise of amplifier in Eq. 4.3, we have

$$\overline{I_{n,amp}^2(f)} = \overline{I_{n,g_m}^2(f)} = \frac{4kT\gamma}{g_m R_F^2}\left[1 + \left(2\pi R_F C_{tot,in} f\right)^2\right] \tag{4.9}$$

$$= \frac{4kT\gamma}{g_m R_F^2} + 4kT\gamma \frac{\left(2\pi C_{tot,in}\right)^2}{g_m} f^2 \tag{4.10}$$

Since R_F is large enough, the DC component of g_m noise (first term) is small enough to be neglected, we have

$$\overline{I_{n,g_m}^2(f)} \approx 4kT\gamma \frac{\left(2\pi C_{tot,in}\right)^2}{g_m} f^2 \tag{4.11}$$

which becomes a second-order term only, where

$$C_{tot,in} = C_{PD} + C_{in} \tag{4.12}$$

and C_{PD} is the photodiode capacitance (including capacitance from input pad), while C_{in} is the input capacitance from TIA circuit, which equals $C_{gg} = C_{gs} + C_{gd}$, the gate capacitance of the input MOS device of the amplifier. For now, we neglect Miller effect for simplicity. Since

$$f_T = \frac{g_m}{2\pi(C_{gs} + C_{gd})} = \frac{g_m}{2\pi C_{gg}} \tag{4.13}$$

and take it into Eq. 4.11, we have

$$\overline{I_{n,g_m}^2(f)} = 4kT\gamma \frac{2\pi}{f_T} \frac{\left(C_{PD} + C_{gg}\right)^2}{C_{gg}} f^2 \tag{4.14}$$

Observing Eq. 4.14, the minimum integrated input-referred noise can be found by minimizing input-referred noise spectrum. By taking derivative, it reaches minimum when

$$C_{PD} = C_{gg} \tag{4.15}$$

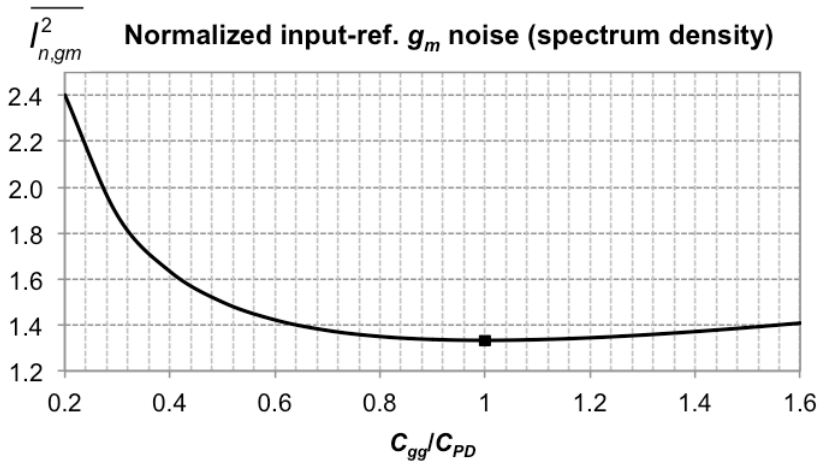


Fig. 4.4 g_m noise optimization in conventional approach

This is also clearly shown in Fig. 4.4, which plots normalized input-referred g_m noise item: $(C_{PD}+C_{gg})^2/C_{gg}$. Thus, to minimize TIA input-referred noise, we should size the MOS g_m device that its input capacitance equals that from photodiode. This "noise matching" technique [5], [33], [34] has been used for quite a long time even in recent work [30].

The other implication from Eq. 4.14 is that the MOS g_m device should be biased as fast as possible to reach the peak f_T because this can maximize the g_m.

4.3 Improved noise optimization

4.3.1 Optimum sizing

In "high-speed" TIA scenario, R_F noise is not negligible and may even dominate, as shown in Fig. 4.3 (b). Hence the noise optimization should consider both R_F and amplifier, we have

$$\overline{I_{n,in}^2(f)} = \frac{4kT}{R_F} + 4kT\gamma\frac{2\pi}{f_T}\frac{\left(C_{PD}+C_{gg}\right)^2}{C_{gg}}f^2 \tag{4.16}$$

The f^2 term is independent of R_F but the sizing of C_{gg} will impact the maximum R_F because larger input capacitance results a smaller R_F for a constant bandwidth. Since R_F and amplifier have different shapes of input-referred noise spectrum, the noise optimization is conducted on integrated input-referred noise.

Because the TIA is followed by several amplifier stages that sharpens out-band roll-off, the noise bandwidth BW_n roughly equals the -3-dB bandwidth[1]. Integrating Eq. 4.16 from 0 to -3-dB bandwidth (BW), we have

$$\overline{I_{n,in,int}^2} = \int_0^{BW}\left[\frac{4kT}{R_F} + 4kT\gamma\frac{2\pi}{f_T}\frac{\left(C_{PD}+C_{gg}\right)^2}{C_{gg}}f^2\right]df \tag{4.17}$$

Define capacitance ratio

$$\alpha = \frac{C_{gg}}{C_{PD}} \tag{4.18}$$

we have

$$\overline{I_{n,in,int}^2} = \frac{4kT}{R_F}BW + \frac{4kT\gamma 2\pi C_{PD}}{3f_T}\frac{(1+\alpha)^2}{\alpha}BW^3 \tag{4.19}$$

Replace R_F by the transimpedance limit (Eq. 3.22), and assume

$$GBW = f_T \cdot \eta \tag{4.20}$$

where η is the ratio between amplifier's actual GBW and f_T. Eq. 4.17 becomes

$$\overline{I_{n,in,int}^2} = \frac{4kT2\pi C_{PD}(1+\alpha)}{f_T\cdot\eta}BW^3 + \frac{4kT\gamma 2\pi C_{PD}}{f_T}\frac{(1+\alpha)^2}{3\alpha}BW^3 \tag{4.21}$$

$$= \frac{4kT2\pi C_{PD}}{f_T}BW^3\left[\frac{1+\alpha}{\eta} + \frac{(1+\alpha)^2}{3\alpha}\right] \tag{4.22}$$

where it is assumed that $\gamma \approx 1$, which holds for advanced CMOS technology. To solve the minimum, make derivation and we get

[1] Assume second-order Butterworth response and an ideal brick-wall filter follows, $BW_n \approx 0.9*BW$ for both DC and f^2 noise term. In reality, following amplifier stages together make a sharp filter but attenuate less than ideal brick-wall filter, $BW_n \approx BW$ is a reasonable approximation.

$$\alpha_{opt} = \sqrt{\frac{\eta}{\eta + 3}} \tag{4.23}$$

Since $\eta < 1$ and may be as low as $1/3$ due to parasitics from load device (passive or active), loading capacitance from next stage, cascode device, wire capacitance, and buffer loss, we obtain that

$$\alpha_{opt} \approx \frac{1}{3} \sim \frac{1}{2} \tag{4.24}$$

The rationale behind this optimization is to trade some amplifier noise for smaller R_F noise: smaller device size C_{gg} on one hand reduces g_m and leads to larger amplifier noise, but on the other hand enables larger R_F and hence smaller R_F noise. Eventually they reach equilibrium at α_{opt}. This optimization exercised on Eq. 4.22 is plotted in Fig. 4.5, normalized to the common coefficient $4kT2\pi C_{PD}/f_T BW^3$. Assuming η = 0.4, α_{opt} in this case is 0.34. The trade-off of g_m noise and R_F noise is clearly shown in the figure too.

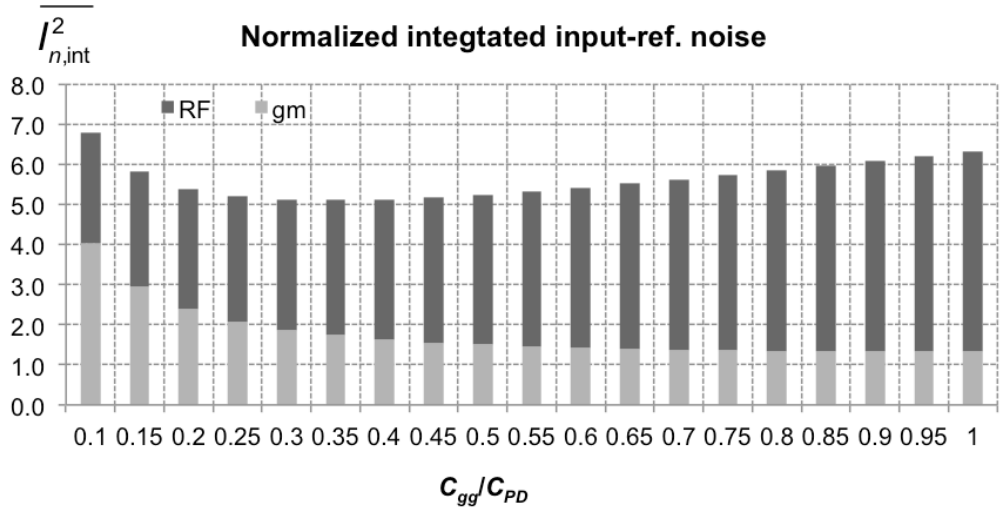


Fig. 4.5 Noise co-optimization on both R_F and amplifier in improved approach

The outcome from this optimization is as follows:

1) Conventional "noise matching" ($\alpha = 1$) approach doesn't achieve minimum input-referred noise, which is clearly shown in Fig. 4.5 due to the lack of consideration of R_F noise. The improved optimization enables minimum input-referred noise since both noise contributors of R_F and g_m are taken into account.
2) Conventional noise matching not only leads to smaller transimpedance gain because of more input capacitance, but also suppresses the noise from latter stages (LA) less due to smaller transimpedance gain.
3) Conventional noise matching means more power dissipated due to larger device used.
4) The improved noise optimization approach, though derived from the "high-speed" scenario, applies to both "low-speed" and "high-speed" TIA scenario, since it takes account of both noise contributions from R_F and g_m.

4.3.2 Optimum sizing: other effects

Impact of Miller effect. In previous optimization Miller effect is neglected, which fits for two cases: 1) a cascode device to suppress Miller effect; 2) C_{gd} is quite small compared with C_{gs}, which is basically true in old technology. However, there may not always be the cascode device to avoid Miller effect and the C_{gd}/C_{gs} ratio is quite high, close to 1/2 in advanced technology, since the gate drain overlap area becomes relatively larger compared to the highly scaled channel length. Thus, it is reasonable to consider Miller effect for certain cases. Notice that Miller effect doesn't change the capacitance that forms the zero [33] and Eq. 4.22 becomes

$$\overline{I_{n,in,int}^2} = \frac{4kT2\pi C_{PD}}{f_T} BW^3 \left[\frac{[1+\alpha(1+A/3)]}{\eta} + \frac{(1+\alpha)^2}{3\alpha}\right] \tag{4.25}$$

and the α_{opt} becomes

$$\alpha_{opt} = \sqrt{\frac{\eta}{\eta + 3(1+A/3)}} \tag{4.26}$$

where A is the amplifier open-loop gain. In advanced technology A is around 2~6. Assume η is 0.4, α_{opt} is around $0.21 \sim 0.27$, smaller than the case without Miller effect. Intuitively, this is because the actual R_F is smaller due to Miller effect, thus α_{opt} is shifted smaller in favor of a larger R_F.

Impact of other noise sources. Until now it is assumed that noise only comes from R_F and g_m device of amplifier. In the real case, other noise sources may contribute considerable noise. For example, noise from load device (M2 in Fig. 4.1) may be quite large because active load, e.g. PMOS current load, is preferred over resistor due to the limited voltage headroom in advanced technology, which increases amplifier noise. Furthermore, the source follower (M3 and M4 in Fig. 4.1) also boosts the noise somewhat. Since noise from both the active load and source follower scales with the g_m noise, a noise multiplication factor λ is introduced to take account of other noise sources, and Eq. 4.22 becomes

$$\overline{I_{n,in,int}^2} = \frac{4kT2\pi C_{PD}}{f_T} BW^3 \left[\frac{1+\alpha}{\eta} + \frac{\lambda(1+\alpha)^2}{3\alpha}\right] \tag{4.27}$$

add the α_{opt} becomes

$$\alpha_{opt} = \sqrt{\frac{\lambda\eta}{\lambda\eta + 3}} \tag{4.28}$$

Assume η is 0.4 and λ is between $1 \sim 2$, α_{opt} becomes $0.34 \sim 0.46$. Intuitively, extra noise makes g_m noise looked larger and α_{opt} (0.34) is correspondingly larger to offset this effect.

Impact of wire parasitics. Parasitics from interconnection may be quite large in advanced CMOS since the wire capacitance doesn't scale with transistor. In this sense, considering transistor capacitance only may lead to too optimistic result.

As a matter of fact, a parameter which stands for wire parasitics is introduced and Eq. 4.22 becomes

$$\overline{I_{n,in,int}^2} = \frac{4kT2\pi C_{PD}}{f_T} BW^3 \left[\frac{1+\alpha p}{\eta} + \frac{(1+\alpha p)^2}{3\alpha}\right] \tag{4.29}$$

where p expresses the capacitance increase from wire parasitics in the input. Thus we have

$$\alpha_{opt} = \sqrt{\frac{\eta}{\eta p^2 + 3p}} \tag{4.30}$$

Still assume η is between $1/3 \sim 1$ and p is 1.3 (30% more input capacitance from TIA, a reasonable assumption in advanced CMOS), we have α_{opt} is $0.27 \sim 0.42$. The result can be interpreted as follows: the wire parasitics impact both g_m and R_F noise, but due to the dominant role of R_F noise, α_{opt} is smaller for larger R_F.

Combine these effects. Finally, all the aforementioned effects are taken account together for a unified expression and Eq. 4.22 becomes

$$\overline{I_{n,in,int}^2} = \frac{4kT2\pi C_{PD}}{f_T} BW^3 \left[\frac{1+p\alpha(1+A/3)}{\eta} + \frac{\lambda(1+p\alpha)^2}{3\alpha}\right] \tag{4.31}$$

where the meaning of the all parameters introduced earlier are listed below

C_{PD}	Photodiode capacitance, including that from bonding pad
f_T	Transistor transit frequency (bias dependent)
BW	TIA -3-dB bandwidth
α	C_{gg}/C_{PD}, where $C_{gg} = C_{gs} + C_{gd}$
A	Amplifier open-loop gain
η	GBW/f_T: amplifier GBW efficiency
λ	Noise multiplication factor due to extra noise sources
p	Input capacitance multiplication factor due to wire capacitance

Thus, α_{opt} is given by

$$\alpha_{opt} = \sqrt{\frac{\lambda\eta}{\lambda\eta p^2 + 3p(1+A/3)}} \tag{4.32}$$

Design example. To verify the developed optimization approach, a circuit example is done in 65-nm CMOS technology and C_{PD} is assumed to be 30 fF. The TIA schematic employs that in Fig. 4. 1. The BW is set to 17 GHz (≈ 25G * 0.7) and the bias in g_m-stage is set to around 180 μA/μm to maximize its transconductance, which corresponds to f_T of around 200 GHz. By varying the input capacitance (α), input-referred noise is plotted in Fig. 4.6.

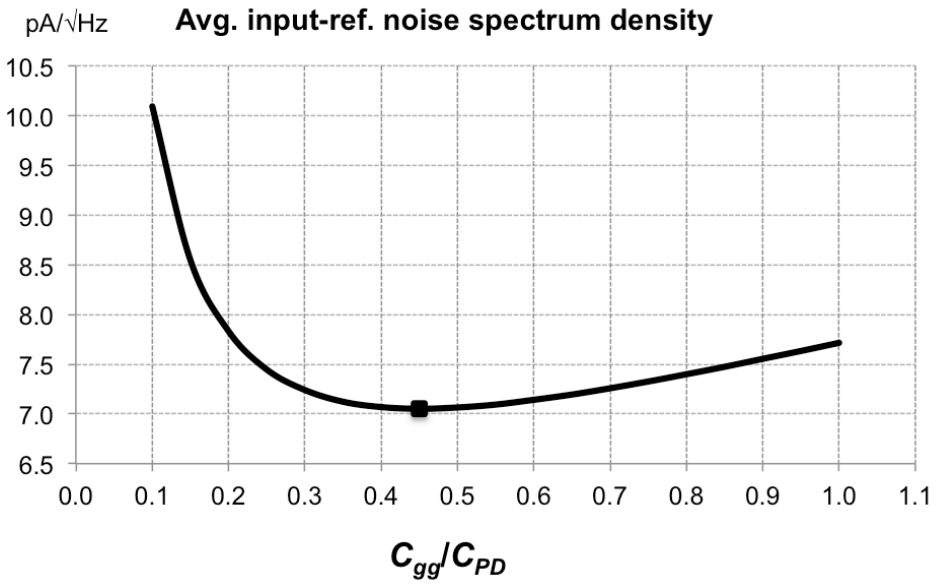


Fig. 4.6 TIA noise optimization example: transistor sizing

From Fig. 4.6, it is clear that α_{opt} = 1 in the conventional approach is sub-optimal. On the other hand, by using Eq. 4.32, α_{opt} in this case is around 0.3~0.4, much closer to the simulation results.

Table 4.1 Comparison of noise optimization on transistor sizing

Item	Conventional	Improved	Ratio (impr./conv.)
α_{opt}	1	0.35	0.35x
Noise (pA/$\sqrt{Hz}$)	7.7	7.1	0.92x
R_T (Ohm)	242	389	1.61x
Power (mW)	8.2	2.9	0.35x
R_T*BW/power (Ω*GHz/mW)	502	2280	4.5x

Quantitative performance comparison in conventional and improved noise optimization is given in Table 4.1. It can be seen that while noise is reduced moderately, R_T and power consumptions are improved considerably due to much smaller transistor size used (smaller input capacitance, smaller bias current). As a result, overall FOM (R_T*BW/power) gains major improvement. It should be mentioned that α_{opt} is set to 0.35 rather than the absolute minimum point (α_{opt} = 0.45) to have more advantage in R_T and power consumptions with negligible noise increase.

4.3.3 Optimum biasing

Observe Eq. 4.22 we may draw the conclusion that f_T should be maximized because not only transistor g_m is maximized but also larger R_F can be achieved due to a larger GBW ($\propto f_T$). Consequently, the g_m transistor should be biased with large current density maximize g_m. While in first-order this is correct, some issues from advanced CMOS technology may change this scenario somewhat: 1) low supply voltage; 2) low intrinsic gain in large bias situation.

Suppose bias current is I_D and $g_m \propto \sqrt{I_D}$, assuming transistor doesn't enter velocity saturated region. Due to the low supply voltage, load device (resistor or active load) is allocated very limited voltage headroom (V_{RD}). Thus $R_D \propto 1/I_D$ holds for either resistor or active load with R_D being load resistance. As a result, the open-loop amplifier gain $A = g_m * R_D$ has the following characteristic [31]

$$A \propto 1/\sqrt{I_D} \tag{4.33}$$

which drops as bias current goes higher. As a result, R_F will become increasingly deviate from its theoretical maximum, the transimpedance limit. Furthermore, the intrinsic gain of transistor drops as bias current is set large, which also limits the amplifier gain.

Therefore, optimum bias is lowered from the maximum to enable sufficient amplifier gain (A) and a large enough R_F. This of course lowers g_m and raises the g_m noise. As a result, equilibrium can be reached again based on the trade-off between R_F and g_m noise. Use the same circuit from Fig. 4.1, fix α at 0.35 and vary the bias current (f_T), a noise plot can be obtained shown in Fig. 4.7.

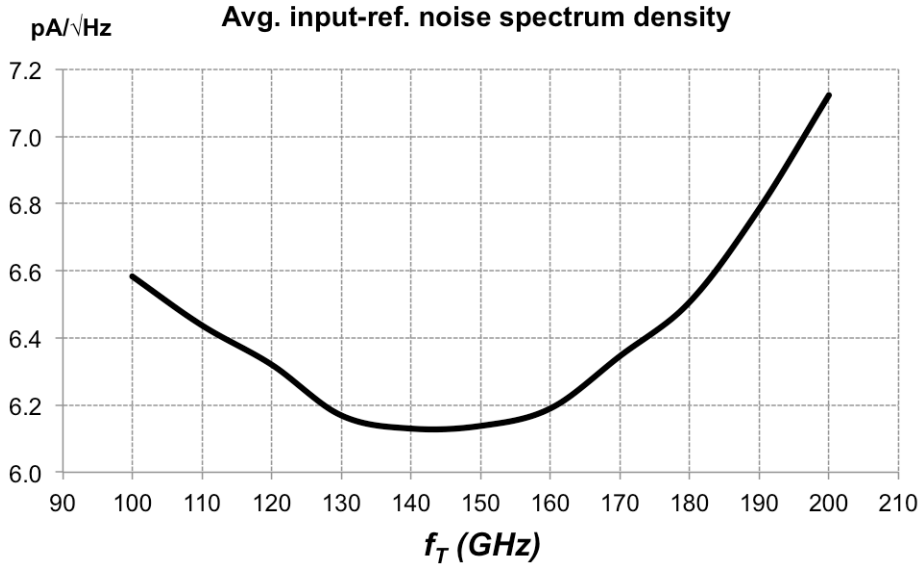


Fig. 4.7 TIA noise optimization example: transistor biasing

It can be seen from Fig. 4.7 that optimum noise is achieved at f_T of around 140~150 GHz (~ 80 μA/μm) instead of peak f_T of 200 GHz (~ 180 μA/μm). Performance comparison is exercised and shown in Table 4.2, with the same conclusion as before: moderate noise reduction, more advantage on R_T and power consumptions, which translates to a major FOM improvement. In general, a bias current density of around 100 μA/μm is found necessary for optimum noise.

Table 4.2 Comparison of noise optimization on transistor biasing

Item	Conventional	Improved	Ratio (impr./conv.)
f_T (GHz)	200	140	0.7x
Noise (pA/$\sqrt{Hz}$)	7.1	6.1	0.85x
R_T (Ohm)	389	648	1.63x
Power (mW)	2.9	1.6	0.62x
R_T*BW/power (Ω*GHz/mW)	2280	6885	3x

By contrast, for a BJT feedback-TIA, $g_m = I_C/V_T$ where I_C is the collector current and V_T is the thermal voltage; thus $g_m \propto I_C$. This suggests that the voltage gain is independent of bias point, which enables bias at its technology peak f_T and still achieves high gain.

4.3.4 Co-optimum: sizing and biasing

Combining result we have obtained from so far transistor sizing (α_{opt}) and biasing (f_T) reported in Table 4.1 and 4.2, determines the overall effect of the optimization, plotted in 3-D in Fig. 4.8.

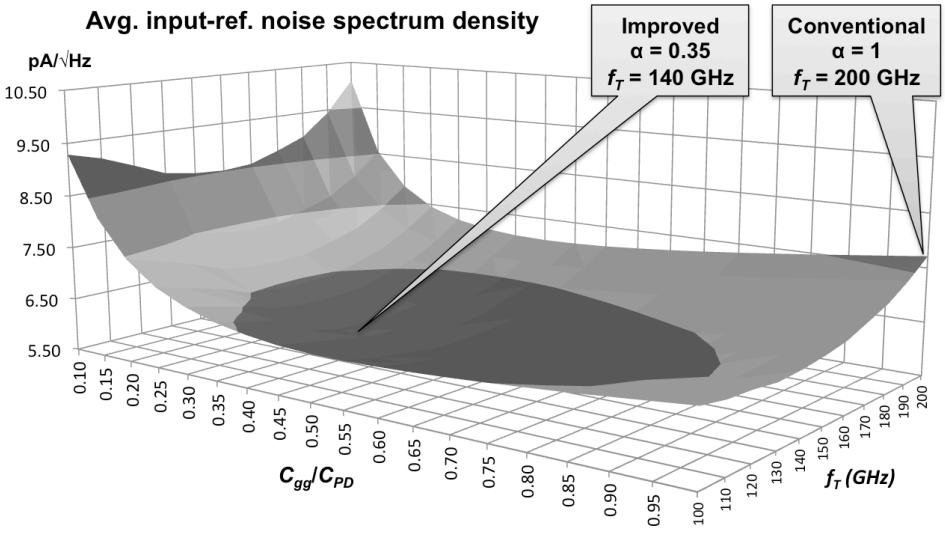


Fig. 4.8 TIA noise co-optimization example: transistor sizing + biasing

A summary is given in Table 4.3. Overall, average input-referred noise spectrum density is reduced by 21%. FOM is boosted by 13.7x resulting from 2.7x R_T and 5x lower power consumption.

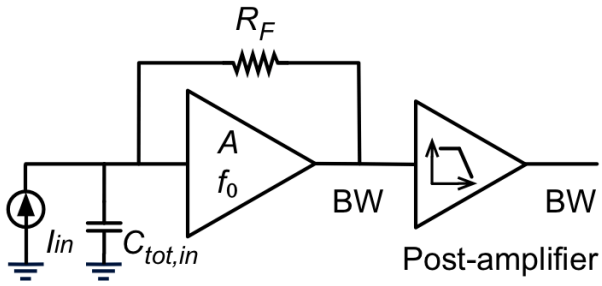
Table 4.3 Noise optimization on transistor sizing and biasing

Item	Conventional	Improved	Ratio (impr./conv.)
α_{opt}	1	0.35	0.35x
f_T(GHz)	200	140	0.7x
Noise (pA/$\sqrt{Hz}$)	7.7	6.1	0.79x
R_T (Ohm)	242	648	2.68x
Power (mW)	8.2	1.6	0.2x
R_T*BW/power (Ω*GHz/mW)	502	6885	13.7x

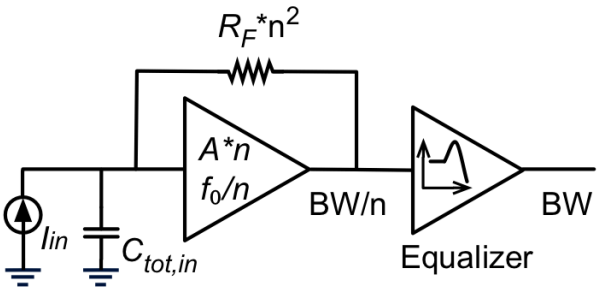
4.4 Low-noise two-stage front-end

The improved noise optimization techniques, though showing considerable advantage compared with the conventional approach in FOM, could reduce the noise just to a moderate degree, mainly because it is based on the trade-off between R_F and g_m noise. Here we propose a more effective noise reduction technique that does not rely on the trade-off, but targeting directly on R_F noise and other low-frequency components.

Shunt-feedback TIA as a low-noise topology in low-speed is the starting point of this technique and the topology is shown in Fig. 4.9.



(a)



(b)

Fig. 4.9 Front-end topologies: (a) Conventional approach; (b) Two-stage front-end

In proposed two-stage front-end, the TIA bandwidth is purposefully scaled down by a factor of *n*, and the following stage becomes an equalizer from amplifier to recover bandwidth to the target value. In such a way the overall

front-end bandwidth remains unchanged compared with conventional approach. Since the TIA bandwidth is n times smaller, according to transimpedance limit (c.f. Eq. 3.22), this enables R_F to be n^2 higher, which means n^2 times R_F noise reduction. On the other hand, the amplifier within the loop should be modified accordingly: n times larger open-loop gain and n times smaller open-loop bandwidth (GBW doesn't change), which is achieved by increasing its output impedance by n times. The modification on amplifier on one hand accommodates the new closed-loop bandwidth requirement (c.f. Eq. 3.11), and on the other hand does not change the phase margin of the loop since open-loop bandwidth and closed-loop bandwidth vary the same factor.

To gain more insights on the overall impact on noise, the input-referred noise current spectrum in conventional approach is given by

$$\overline{I_{n,in,conv}^2(f)} = \overline{I_{n,R_F}^2(f)} + \overline{I_{n,amp}^2(f)} + \overline{I_{n,eq}^2(f)} \tag{4.34}$$

$$= \frac{4kT}{R_F} + \frac{4kT\gamma}{g_m R_F^2}\left|1 + sR_F C_{tot,in}\right|^2 + \frac{\frac{4kT\gamma}{g_{m,eq}}}{|Z_T(s)|^2} \tag{4.35}$$

where noise contribution from equalizer $\overline{I_{n,eq}^2(f)}$ is considered and $g_{m,eq}$ stands for the equalizer transconductance. In this case, the post-amplifier is named equalizer to unify latter analysis and also to distinguish itself from the amplifier within the feedback loop. Since $Z_T(s)$ is usually designed to have Butterworth shape for maximal flat response, we have

$$|Z_T(s)|^2 \approx \frac{R_F^2}{1 + \left(\frac{f}{BW}\right)^4} \tag{4.36}$$

where $A/(A+1)$ is assumed to be 1. Thus the input-referred noise spectrum becomes

$$\begin{aligned}\overline{I_{n,in,conv}^2(f)} = \frac{4kT}{R_F} + \frac{4kT\gamma}{g_m R_F^2} + 4kT\gamma\frac{\left(2\pi R_F C_{tot,in}\right)^2}{g_m R_F^2} f^2 \\ + \frac{4kT\gamma}{g_{m,eq} R_F^2} + \frac{4kT\gamma}{g_{m,eq} R_F^2}\left(\frac{f}{BW}\right)^4\end{aligned} \tag{4.37}$$

Hence g_m noise has a DC term and a f^2 term while EQ (equalizer) noise has a DC term and a f^4 term. On the other hand, by applying the proposed technique, the input-referred noise spectrum density becomes

$$\begin{aligned}\overline{I_{n,in,prop}^2(f)} = \frac{4kT}{R_F n^2} + \frac{4kT\gamma}{g_m R_F^2 n^4} + 4kT\gamma\frac{\left(2\pi R_F n^2 C_{tot,in}\right)^2}{g_m R_F^2 n^4} f^2 \\ + \frac{4kT\gamma}{g_{m,eq} R_F^2 n^4} + \frac{4kT\gamma}{g_{m,eq} R_F^2 n^4}\left(\frac{f}{BW/n}\right)^4\end{aligned} \tag{4.38}$$

$$= \frac{4kT}{R_F n^2} + \frac{4kT\gamma}{g_m R_F^2 n^4} + 4kT\gamma \frac{\left(2\pi R_F C_{tot,in}\right)^2}{g_m R_F^2} f^2 + \frac{4kT\gamma}{g_{m,eq} R_F^2 n^4} + \frac{4kT\gamma}{g_{m,eq} R_F^2} \left(\frac{f}{BW}\right)^4 \tag{4.39}$$

Compare Eq. 4.39 with Eq. 4.37, we have the following observations:

- R_F noise is reduced by a factor of n^2.
- DC noise of g_m and equalizer are reduced by a factor of n^4.
- High-frequency noise of g_m and equalizer remain unchanged.

Thus, as long as we could equalize the bandwidth, this approach reduces all low-frequency noise without trading off high-frequency noise, compared to what we have done in the noise optimization section. If the scaling factor n is sufficiently large, all the low-frequency noise essentially go to zero. Thus this is a much more aggressive approach compared with previous noise optimization.

In the real case, an ideal equalizer doesn't exit and n can't be arbitrarily large. Another limitation comes from the fact that the amplifier gain needs to scale up n times (by boost output impedance) at the same time, which might be difficult to realize since low supply voltages limits the number of cascaded devices. However, when n only equals 2, R_F noise can be reduced by 4 times and DC components of g_m and EQ noise are reduced by 16 times, showing its impressive noise reduction capability.

The theoretical input-referred noise spectrum of conventional approach and proposed two-stage front-end approach (n = 2) are plotted in Fig. 4.10 for comparison. The unchanged high-frequency noises can also be verified from the zero point in both g_m and equalizer noise. For g_m noise, zero frequency is reduced 4x while DC magnitude is reduced 16x. Since the roll-up is second order, the old and the new zero locate in the same roll-up function. Similarly, for equalizer noise, zero frequency is halved and DC magnitude is 16x smaller, the old and the new zero are in the same fourth-order roll-up function.

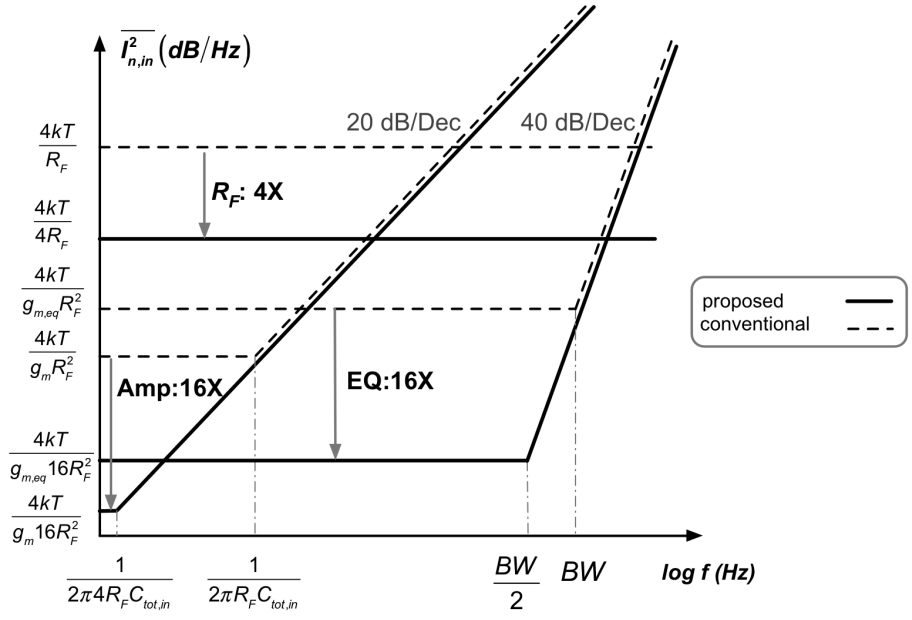


Fig. 4.10 Input-referred noise spectrum: conventional approach and two-stage front-end

It should be noticed that the noise reduction on equalizer is also desirable. This is because $g_{m,eq}$ is usually much smaller than g_m due to smaller input device size on equalizer not to load TIA to maintain sufficient amplifier open-loop bandwidth for stability. Thus the noise from equalizer though has only DC content in-band, might be several times higher DC noise of g_m device. The proposed approach hence is also advantageous on this point.

Noise optimization on low-noise two-stage front-end. The two-stage front-end approach discussed so far assumes a fixed transistor size. In fact, the developed noise optimization approach can be applied also to the two-stage front-end and Eq. 4.22 becomes

$$I_{n,in,int}^2 = \frac{4kT2\pi C_{PD}}{f_T} BW^3 \left[\frac{(1+\alpha)}{\eta \cdot n^2} + \frac{(1+\alpha)^2}{3\alpha}\right] \tag{4.40}$$

and the optimum is

$$\alpha_{opt} = \sqrt{\frac{\eta}{\eta + 3/n^2}} \tag{4.41}$$

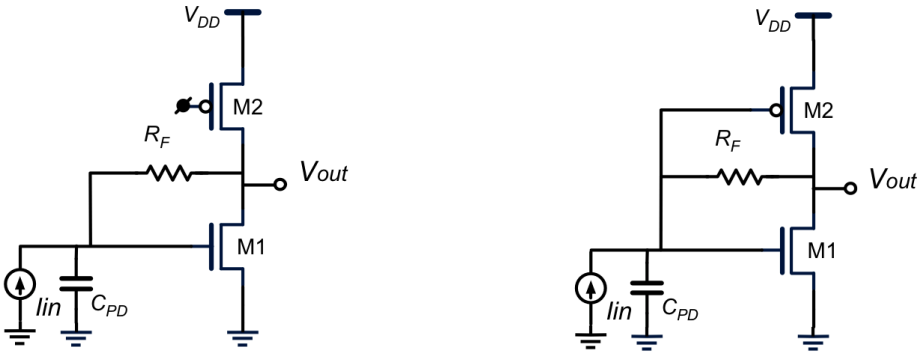
Assume $\eta = 0.4$, and n varies from 2~4, α_{opt} is between 0.59 ~ 0.83. As a result, α_{opt} in two-stage front-end is larger than conventional TIA. If n goes very large, α_{opt} -> 1. The reason for this is that as n is sufficiently large, R_F noise becomes less and the optimization is like exercising on g_m noise alone.

Examples of TIA design employing low-noise two-stage front-end will be given in next chapter.

4.5 G_m-reuse technique

4.5.1 Principle

Previous analyses assume an NMOS input TIA (NMOS-TIA). In this section we investigate the case of CMOS input TIA (CMOS-TIA). This g_m-reuse technique has been exploited in LNA design [35] previously; we will apply the same concept to TIA design here.



(a) (b)

Fig. 4.11 Typical shunt-feedback TIA: (a) NMOS-type; (b) CMOS-type

Fig. 4.11 shows the typical shunt-feedback TIA: NMOS-TIA and CMOS-TIA. In NMOS-TIA, the loading device could be resistor, but it is limited by the low supply voltage and provides low gain; the PMOS load, which consumes less voltage headroom is preferred although nosier.

The g_m-reuse technique derived from LNA design [35] is to use CMOS-TIA, in favor of the extra transconductance and gain contributed from $g_{m,p}$, to get both larger total g_m and R_F. Furthermore, since the load device becomes a transconductor, not only g_m is larger, but also its input-referred noise contribution as load is removed. However, since the addition of PMOS also contributes capacitance at the input, the exact noise reduction for g_m and R_F should be carefully analyzed.

For CMOS-TIA, assume $Wp/Wn = k$, then $g_{m,p}/g_{m,n} \approx 0.5*k$. Thus the amplifier gain becomes $1+0.5k$ times larger from NMOS-TIA to CMOS-TIA. Assuming a conventional noise optimization scenario (noise matching) $C_{PD} = C_{tia}$ (NMOS only) and PMOS dimension twice of NMOS, the total input capacitances are $2C_{PD}$ and $C_{PD}*(2+k)$ respectively for NMOS-TIA and CMOS-TIA. Hence, total input capacitance also becomes $1+0.5k$ times larger from NMOS-TIA to CMOS-TIA. As a result, PMOS device contributes the same amount of amplifier gain and input capacitance, meaning that R_F doesn't change.

Now let's compare the noise of amplifier. The normalized input-referred amplifier noise spectrum for NMOS-TIA is given by ($C_{in,tot}^2/g_m$, normalized to $4k\mathrm{T}\gamma*f^2$ and the same hereafter, c.f. Eq. 4.11)

$$\overline{I_{n,amp,nmos,conv}^2} = \frac{(2C_{PD})^2}{g_{m,n}}\left(1+\frac{k}{2}\right) \tag{4.42}$$

where the factor $k/2$ stands for the noise contributed from PMOS load. The normalized input-referred amplifier noise spectrum for CMOS-TIA is

$$\overline{I_{n,amp,cmos,conv}^2} = \frac{[C_{PD}(2+k)]^2}{g_{m,n}\left(1+\frac{k}{2}\right)} = \frac{(2C_{PD})^2}{g_{m,n}}\left(1+\frac{k}{2}\right) \tag{4.43}$$

Compare Eq. 4.43 with Eq. 4.42, we find the input-referred noise equals with each other in the two situation. The conclusion we get is that CMOS-TIA doesn't offer advantage to NMOS-TIA under conventional noise matching scenario.

On the other hand, when CMOS-TIA topology is applied to the improved noise optimization scenario, the normalized input-referred amplifier noise spectrum for NMOS-TIA and CMOS-TIA are

$$\overline{I_{n,amp,nmos,impr}^2} = \frac{(C_{PD}+\alpha C_{PD})^2}{g_{m,n}\alpha}\left(1+\frac{k}{2}\right) \tag{4.44}$$

$$\overline{I_{n,amp,cmos,impr}^2} = \frac{[C_{PD}(1+\alpha+\alpha k)]^2}{g_{m,n}\alpha\left(1+\frac{k}{2}\right)} \tag{4.45}$$

where α is defined the same as before and the result is plotted in Fig. 4.12.

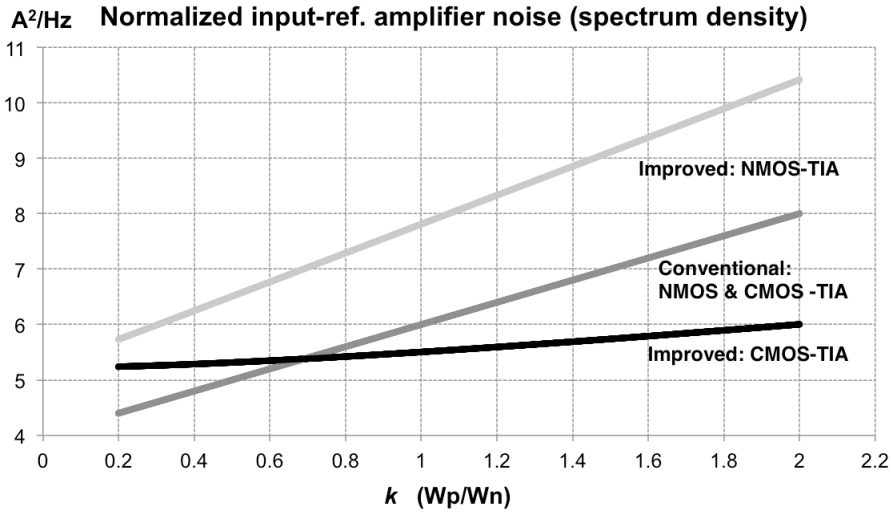


Fig. 4.12 Normalized input-referred amplifier noise spectrum vs. P/N ratio in conventional and improved noise optimization scenarios

Here, α is set to 0.35. The normalized input-referred amplifier noises are plotted in various scenarios, where for conventional approach, NMOS-TIA and CMOS-TIA have the same noise profile as we have already discussed earlier. NMOS-TIA in improved noise optimization has larger noise due to that it trades with smaller R_F noise to achieve overall smaller noise, as we have explained in Sec. 4.3. CMOS-TIA in improved noise optimization not only reduces noise from its NMOS counterpart, but also achieves smaller noise (when k is large enough) compared with CMOS-TIA in conventional approach where g_m noise is exclusively optimized. Meanwhile k should not be kept small since NMOS needs a decent bias current, so the advantage in CMOS-TIA from improved noise optimization scenario to conventional scenario holds.

Furthermore, we examine CMOS-TIA in improved noise optimization scenario on R_F noise, which is proportional to $1/R_F$ or $C_{in,tot}/A$. The normalized input-referred R_F noise spectra are

$$\overline{I^2_{n,R_F,nmos,impr}} = 1 + \alpha \tag{4.46}$$

$$\overline{I^2_{n,R_F,cmos,impr}} = \frac{1 + \alpha + \alpha k}{1 + k/2} \tag{4.47}$$

while in conventional optimization the normalized input-referred R_F noise spectrum is just 2 for both NMOS-TIA and CMOS-TIA. Hence, we plot all of them in Fig. 4.13.

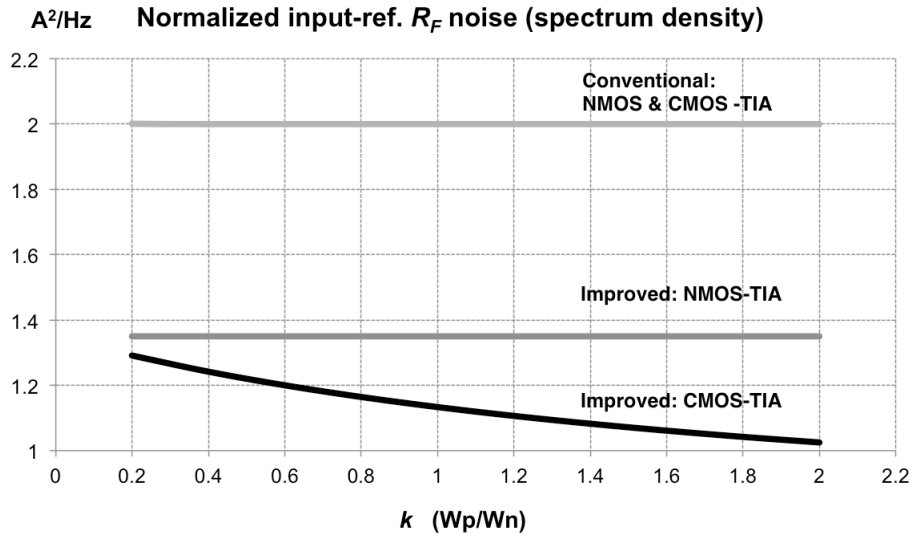


Fig. 4.13 Normalized input-referred R_F noise vs. P/N ratio in conventional and improved noise optimization scenario

As a result, CMOS-TIA on improved noise optimization scenario reduces R_F noise. The rationale behind is that it adds more amplifier gain than capacitance thus enabling a larger R_F.

To summarize, the noise reduction from CMOS-TIA to NMOS-TIA in improved noise optimization scenario is able to decrease both $C_{in,tot}/A$ and $C_{in,tot}^2/g_m$, thus reducing noise for both R_F and amplifier. In contrast, CMOS-TIA doesn't provide these advantages in conventional noise optimization scenario.

4.5.2 P/N ratio

From Fig. 4.12 and 4.13, as k (P/N ratio) goes larger, input-referred amplifier noise increases and input-referred R_F noise decrease, meaning that an optimum k will give the minimum total input-referred noise. Like what we have done in Section 4.3, co-optimization on integrated input-referred noise rather than input-referred noise spectrum density is necessary. Similar to the analysis in Section 4.3, the integrated input-referred noise in CMOS-TIA is

$$\overline{I_{n,in,int}^2} = \frac{4kT2\pi C_{PD}}{f_T} BW^3 \left[\frac{1+\alpha+\alpha k}{\eta(1+k/2)} + \frac{(1+\alpha+\alpha k)^2}{3\alpha(1+k/2)}\right] \tag{4.48}$$

Previous section has shown that g_m-reuse technique works when α is on improved noise optimization scenario. Applying corresponding parameters (α = 0.35, η = 0.4, defined the same as before), the minimum is reached when k = 2.2.

This g_m-reuse technique can also be applied to low-noise two-stage front-end topology we have proposed. The integrated input-referred noise in becomes

$$\overline{I_{n,in,int}^2} = \frac{4kT2\pi C_{PD}}{f_T} BW^3 \left[\frac{1+\alpha+\alpha k}{n^2 * \eta(1+k/2)} + \frac{(1+\alpha+\alpha k)^2}{3\alpha(1+k/2)}\right] \tag{4.49}$$

where n is defined the same as before and the minimum is reached when k = 0.63. The reason for smaller k is that the relative weight of R_F noise becomes less, thus k goes to the direction in favor for smaller g_m noise, as shown in Fig. 4.12.

Examples of g_m-reuse technique will be given in next chapter on both one stage TIA and two-stage front-end.

4.6 Conclusion

In this chapter, starting from an in-depth noise analysis of shunt-feedback TIA, we have built an improved noise optimization approach that shows advantage on noise and considerable improvement on gain and power, compared with the conventional noise optimization approach. Furthermore, we have proposed a low-noise two-stage front-end, which can theoretically eliminate all low-frequency noise contributors thus gives significant noise reduction. Finally, the g_m-reuse technique has been explored for TIA design, showing advantage under our improved noise optimization scenario rather than conventional noise optimization scenario. These low-noise design techniques can work together for low-noise optical receiver front-end design, and will be exemplified in the following chapters.

Chapter 5

A 25 Gb/s optical receiver for discrete photodiode

In this chapter, we present a low-noise 25 Gb/s optical receiver for the IEEE 100GBASE-LR4 standard in 65-nm CMOS technology, intended to interface with a discrete III-V commercial photodiode. Various low-noise design techniques presented in previous chapter have been utilized. A prototype was fabricated and measured, showing excellent low-noise performance and the state-of-the-art FOM.

5.1 Introduction

The IEEE 802.3ba 40/100G Ethernet [2] was issued in June 2010 to solve the urgent bandwidth problem from the boom of Internet traffic, data center, storage network and super computing in recent years, shown in Table 5.1.

Table 5.1 40G/100G Ethernet overview

Physical layer		40 GBASE	100 GBASE
Electrical	1m Backplane	KR4	
	7m Copper cable	CR4	CR10
Optical	100 m OM3 MMF	SR4	SR10
	150 m OM4 MMF		
	10 km over SMF	LR4	**LR4**
	40 km over SMF	N/A	ER4

It can be seen that the majority of 40G/100G Ethernet are implemented on optical communications, where MMF targets short reach and SMF targets medium reach. This works in specific focuses on the 100 GBASE-LR4 standards, where four lanes of 25 Gb/s data forming an aggregate 100 Gb/s data stream, transmit at least 10km in SMF. The corresponding receiver architecture is shown in Fig. 5.1.

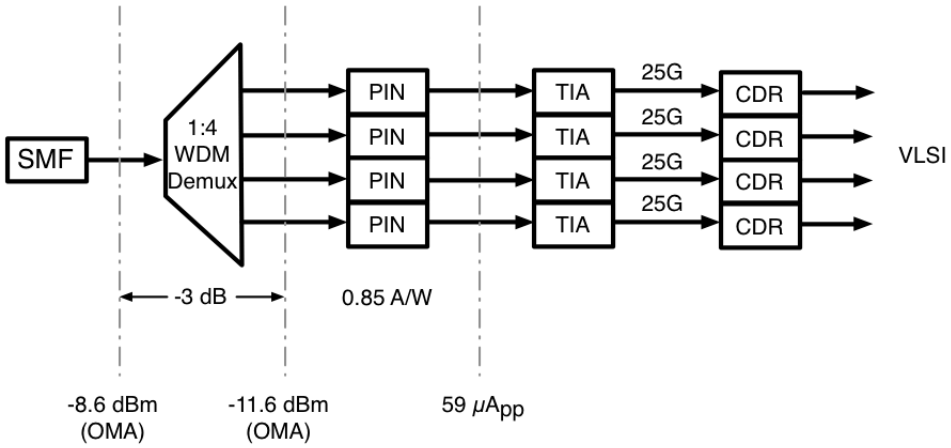


Fig. 5.1 100 GBASE-LR4 receiver architecture

The 100 Gb/s high-speed optical pulse from SMF is demuxed into four 25 Gb/s optical signals and fed to four PIN photodiodes. Then TIAs (receiver) receive the small photocurrent and generate voltage swing of several hundred mV to drive CDR and eventually the data go to VLSI logic.

The standard indicates an minimum input sensitivity of -8.6 dBm OMA. Assume the coupler and DEMUX have total loss of 3 dB, input signal to the PIN photodiode becomes -11.6 dBm. Assume responsivity of PIN photodiode is 0.85 A/W, input current peak-to-peak seen by the TIA is 59 μA_{pp}. To achieve BER < 10^{-12}, the required SNR > 16.9 dB. This corresponds to an input-referred noise current below 4.2 μA_{rms}. The standard also specifies minimum output amplitude of 300 mV_{pp} at sensitivity, leading to a minimum required transimpedance gain of 74 dBΩ. Given the data rate of 25 Gb/s, the bandwidth specification is set to around 17 GHz (≈ 25G * 0.7), in order to give little ISI.

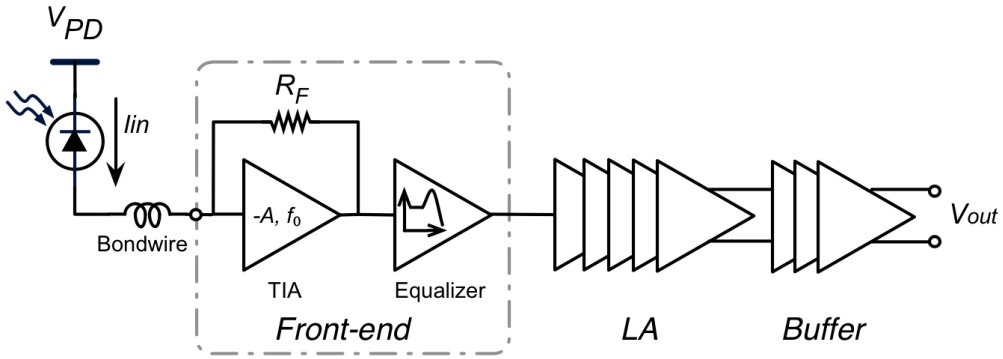


Fig. 5.2 25 Gb/s Receiver architecture

The proposed receiver architecture is given in Fig. 5.2. The external III-V photodiode is intended to be wire-bonded to CMOS receiver chip. Instead of using conventional full bandwidth TIA, this design adopts the two-stage front-end architecture for low-noise performance, explained in previous chapter. A five-stage LA provides more than 20 dB gain for sufficient signal amplification. A three-stage Output buffer offers off-chip 50-Ω driven capability for test purpose. The photodiode needs a higher bias V_{PD}, e.g. 2.5~3V depending the photodiode characteristics, while the receiver circuits are realized in 65-nm digital CMOS

with standard VDD (1V) and a second VDD (1.8V) for performance boost, which will be explained later.

5.2 Two-stage low-noise front-end

Applying the proposed low-noise two-stage front-end architecture, the corresponding schematic is shown in Fig. 5.3. The photodiode is represented by its AC equivalent circuit: a current source in parallel with the photodiode capacitance C_{PD}. Since TIA receives single-end photocurrent, the pseudo-differential architecture is utilized to gain better common-mode noise suppression where the dummy-mirror TIA with floating input is added. This on the other hand, doubles the integrated input-referred noise (3 dB sensitivity penalty) and the power consumption. Since the front-end usually works in small-signal mode while some of the following blocks operate in limiting mode, the power and ground rail tend to be quite noisy and this configuration is often necessary.

A DC current sink realized by an NMOS transistor (drawn in dashed line since it works in DC) is placed to draw the DC component of the photocurrent, so that the DC operating point of TIA is fixed to the optimum. The control voltage of the current sink is set externally.

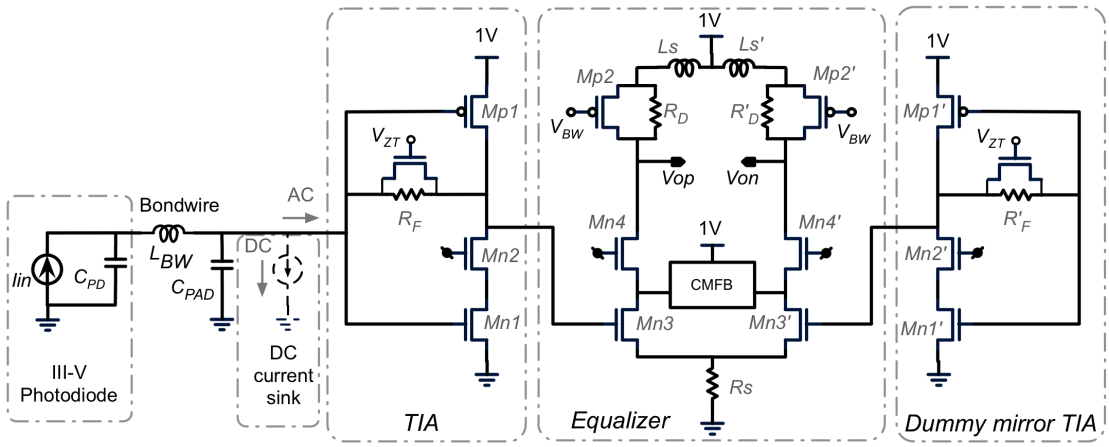


Fig. 5.3 Schematic of the proposed two-stage front-end

The first stage of the front-end is a low-bandwidth TIA. Due to the scaling of the bandwidth of this stage in the two-stage front-end approach, R_F is boosted considerably. The CMOS inverter amplifier is adopted by exploring the g_m-reuse technique explained in previous chapter. An NMOS cascode transistor is inserted to lower the input capacitance due to Miller effect. Extra PMOS cascode transistor is not used in view of the limited supply voltage (1V) and to maintain sufficient open-loop amplifier bandwidth for stability reason. The feedback resistance R_F is an unsilicided poly-resistor in parallel to a variable resistance realized by NMOS transistor. In such a way, overall feedback resistance can be set smaller when large current input feeds in.

The performance of CMOS-TIA is compared to an NMOS-TIA with the same input NMOS dimension and bias. The PMOS dimension is sized 0.8 times of NMOS for minimum noise performance based on the analysis of Section 4.5.2. This also justifies the NMOS cascode rather than the PMOS cascode. This bias

current density is set to 100 μA/μm. The corresponding parameters are compared in Table 5.2.

Table 5.2 CMOS-TIA vs. NMOS-TIA

Item	CMOS / NMOS
Total input cap ($C_{in,tot}$)	1.46X
$g_{m,tot}$	1.48X
Amplifier gain (A)	1.48X
R_F ($A/C_{in,tot}$)	1X
C_{PD} + C_{gg}	1.16X
Input-ref. g_m noise (C_{PD} + C_{gg})$^2/g_{m,tot}$	0.92X
Input-ref. amp noise	0.62X

In Section 4.5, the analysis shows CMOS-TIA has larger R_F than NMOS-TIA without considering Miller effect. From Table 5.2, CMOS-TIA has the same R_F as NMOS-TIA since the Miller capacitance contributed by PMOS has offset this advantage. On the other hand, CMOS-TIA presents only 62% of amplifier noise than NMOS-TIA showing the advantage we have explained in Section 4.5.

The equalizer stage employs a pseudo-differential topology for the same reason as TIA. Since the equalizer is DC coupled to the TIA, which doesn't offer enough common-mode voltage to accommodate a current source below the differential pair, a small resistor is used instead. A cascode NMOS transconductor increases the output impedance of the input transistor pair, and lowers the input capacitance seen by the TIA stage. Shunt peaking is the adopted equalization method, realized by means of low-Q differential inductor, where the amount of peaking can be tuned by the variable resistance formed by shunt pair of resistor and PMOS. Low-Q of the inductor enables the optimization of its parasitic capacitance to minimum and a high self-resonance frequency for safe operation. The width of the metal of the spiral inductor is thus chosen to the minimum constrained by the metal's electro-migration requirement. That kind of the variable resistance connecting the two differential outputs shown in Fig. 3.22 (b) has the advantage of not impacting DC common mode voltage, but cannot be employed here because the equalizer doesn't operate in a fully differential fashion. Hence, equalizer tuning will alter the DC common mode voltage, and a common-mode feedback (CMFB) circuit is inserted to maintain the output common-mode voltage constant. The CMFB circuit realized by a PMOS current load is connected to the drain of the transconductor instead of the output, so that the output capacitance is lowered. Overall, the equalizer can vary its peaking from 0 dB to more than 10 dB, in order to add flexibility accommodating variations in parasitic capacitance and bondwire inductance.

Simulation (pre-layout, single-ended circuit) is performed to decide the first stage bandwidth and other corresponding design parameters like R_F and the amount of equalization. A reference front-end circuit which adopts the conventional bandwidth method (each stages maintains full -3-dB bandwidth: TIA +amplifier, c.f. Fig. 4.9) is used for comparison. The total external input capacitance in this case is 160 fF, where 80 fF comes from the III-V photodiode

and the input pad contributes another 80 fF. The size of the input NMOS transistor is set to have around 1/4 of the external total capacitance (160 fF), rather than the calculated optimum of 0.45 times in order to halve the power consumption since the dummy mirror TIA doubles the power, with less 10% of noise penalty. The reference front-end circuit also employs CMOS-TIA for fair comparison. The P/N ratio for proposed and reference front-end are 0.8 and 2, based on the analysis in Section 4.5.2.

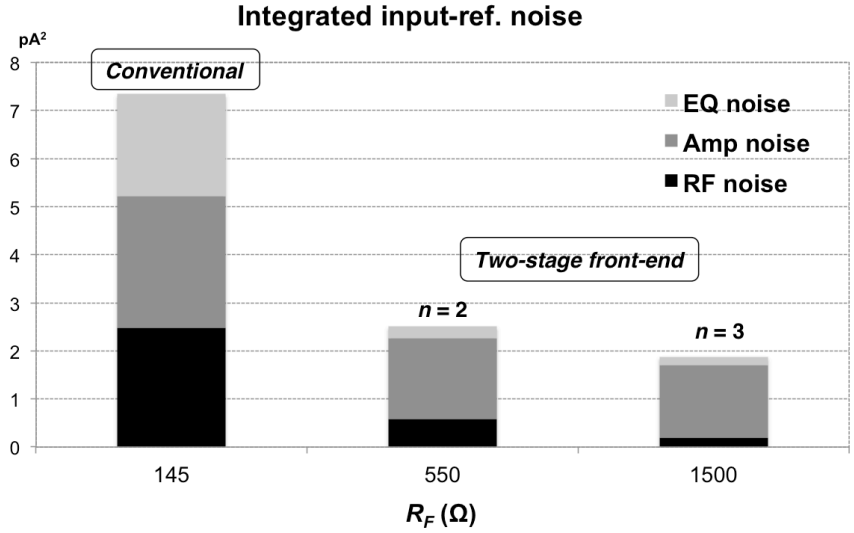


Fig. 5.4 Noise reduction from conventional to two-stage front-end (simulated)

From Fig. 5.4, noise reduction from conventional to the two-stage front-end is evident when R_F grows, due to the dominant role of R_F and the DC component of other noise sources (e.g. equalizer). As the bandwidth scaling factor n increases further from 2 to 3, noise reduction is marginal due to the fact that the dominant noise is from AC part of g_m noise, which doesn't change in the scaling. On the other hand, for the case n = 3, equalizer is tuned almost to its maximum peaking which leads to a frequency response not as flat as the case n = 2 that may potentially degrade the output eye and leaves little margin of equalization tuning to overcome PVT variation. Therefore, n = 2 is chosen as a compromise for low-noise, flat frequency response and tunability.

The input-referred noise power spectra from each noise contributor in conventional and proposed approach (n = 2) are plotted in Fig. 5.5 for comparison, which verifies the theoretical analysis in Fig. 4.10. Fig. 5.6 gives the overall input-referred noise spectrum density. Overall, the proposed two-stage front-end shows 1/4 integrated input referred noise power, compared with Fig. 5.4 of 1/3 times of noise reduction. The difference comes from other noise sources. Furthermore, the two-stage front-end also shows 8 dB more transimpedance gain because: 1) for the TIA stage the conventional approach has only 1/4 of the transimpedance gain; 2) for the second stage the conventional approach is not able to achieve 4x DC gain due to the low supply limit, as we have explained in Section 4.3.3.

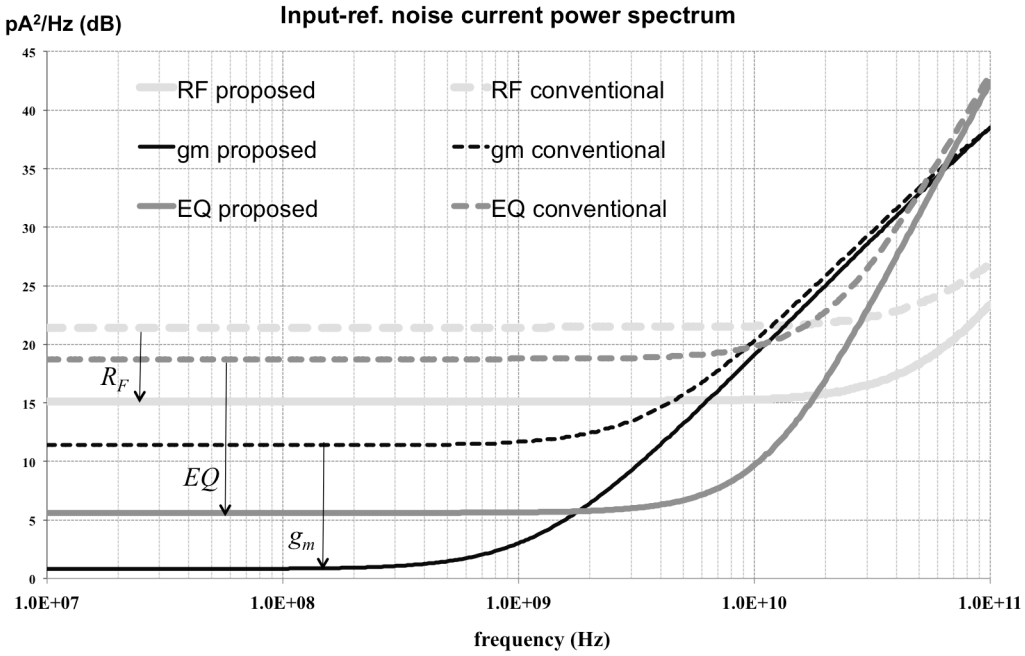


Fig. 5.5 Main noise contributors from conventional to two-stage front-end

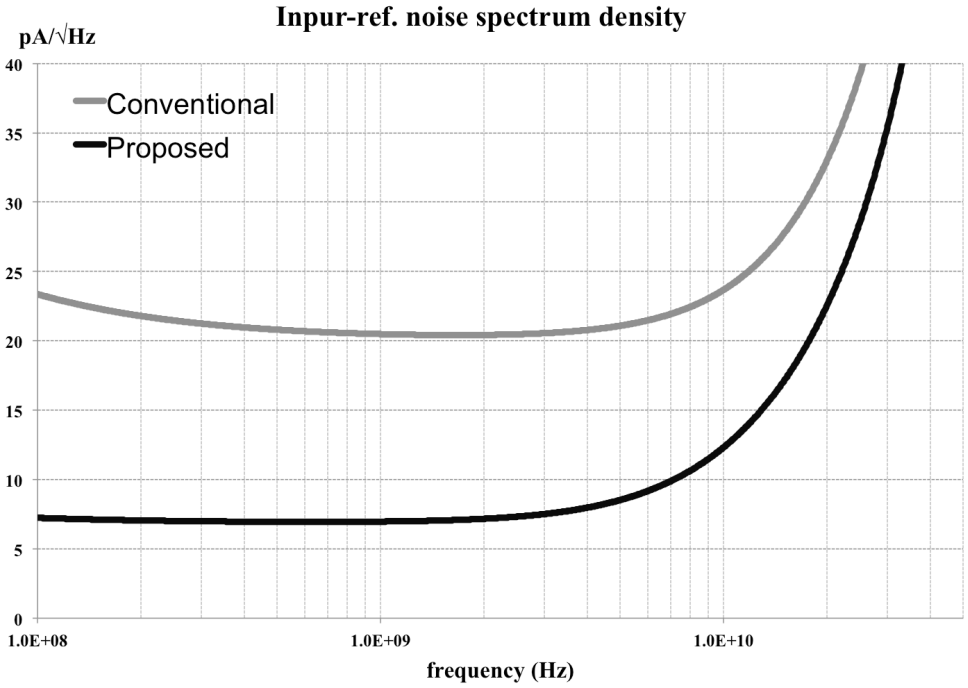


Fig. 5.6 Input-referred noise spectrum density from conventional and proposed two-stage front-end

Since in this design bondwire is the intended connection between photodiode and front-end, series peaking is exploited to get extra bandwidth enhancement. Moreover, series peaking helps suppress high-frequency g_m noise since it cancels some of the input capacitance. The peaking frequency is roughly

$$f_{peak} \approx \frac{1}{2\pi\sqrt{C_{PD}L_{BW}}} \tag{5.1}$$

where L_{BW} is the bondwire inductance. In this design, L_{BW} is supposed to be around 0.5 nH and the corresponding peaking frequency is around 24 GHz. The peaking frequency is placed out-of-band but close to -3-dB bandwidth to gain bandwidth extension with minimum phase distortion, which otherwise may potentially degrade output eye.

The frequency response of the two-stage front-end is given in Fig. 5.7, where it is assumed C_{PD} = 80 fF and L_{BW} = 0.5 nH, and simulated at standard condition (TT, 1V, 27°C) on post-layout circuits. The TIA stage has 8.1 GHz bandwidth, roughly half of the target (17 GHz). The equalizer boosts high frequency and peaks at around 15 GHz while the input series peaking peaks at around 24 GHz. Overall, the two-stage front-end provides 57.2 dBΩ transimpedance gain and 22.4 GHz bandwidth. In-band group delay variation is less than 16 ps to ensure low deterministic jitter.

The simulated input-referred RMS noise current of the two-stage front-end is 2.26 μA_{rms}, which translates to 15.1 pA/$\sqrt{Hz}$ of average input-referred noise spectrum density. The corresponding input sensitivity current for BER < 10^{-12} is thus 32 μA_{pp}, meeting the specification with enough margin.

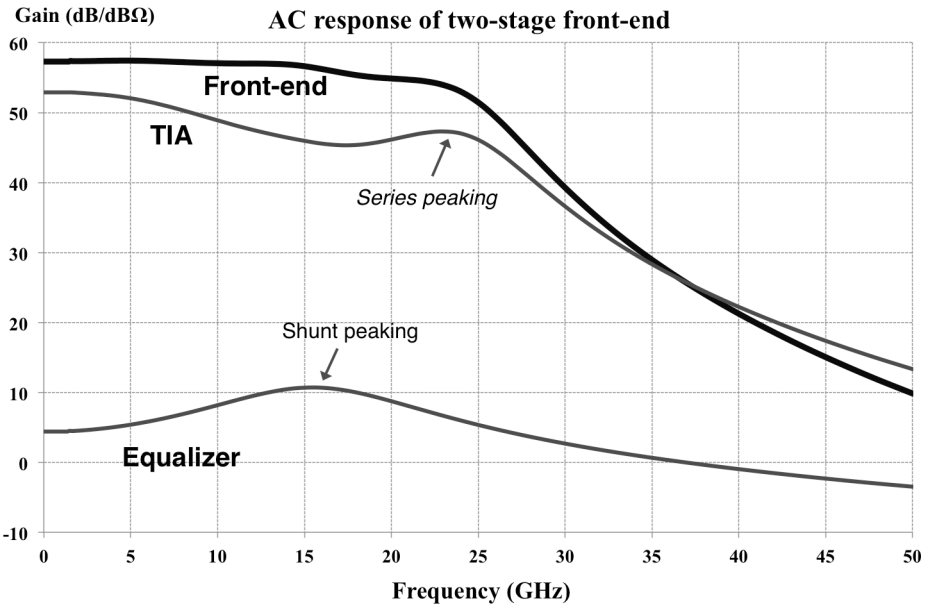


Fig. 5.7 Frequency response of two-stage front-end

To evaluate the quality of transmission, the output eye diagram at the output of two-stage front-end is shown in Fig. 5.8, where 2^7-1 PRBS is used as the source signal for fast characterization. The rise and fall time (10%- 90%) of input signal is 8.3 ps. At sensitivity, the output eye is wide-open and clear with minor signal distortion. The peak-to-peak data-dependent jitter in this case is less than 1.7 ps.

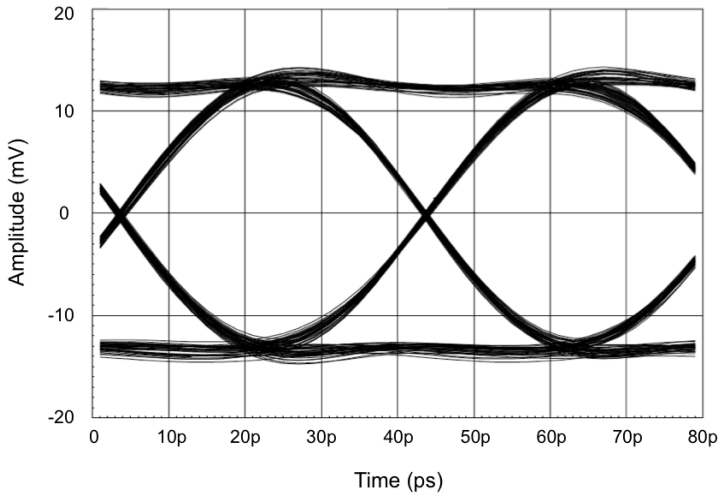


Fig. 5.8 Simulated output eye of front-end at sensitivity

The overload capability of the front-end is also concerned. A large current, e.g. 2 mA_{pp} will drive the g_m device of TIA into deep linear region. Further more, since common-mode output of equalizer is placed around 800 mV, such a large input current will generate highly asymmetrical voltage swing at equalizer output because this equalizer is not fully differential circuit. Consequently, large distortion and jitter will appear. As described earlier, R_F can be configured much smaller for gain control at large input current. However, this will make the feedback loop unstable unless the amplifier gain is reduced by the same amount. In this design, instead of making the amplifier gain variable, which will inevitably lead more parasitics and noise, we exploit the tunability of equalizer. When R_F is set small, a large peaking will be generated in the TIA stage. The equalizer is then tuned to a low bandwidth amplifier, which suppresses the high-frequency peaking from the TIA stage and the two-stage front-end is still able to achieve overall flat frequency response. The frequency response for this gain control mechanism is shown in Fig. 5.9.

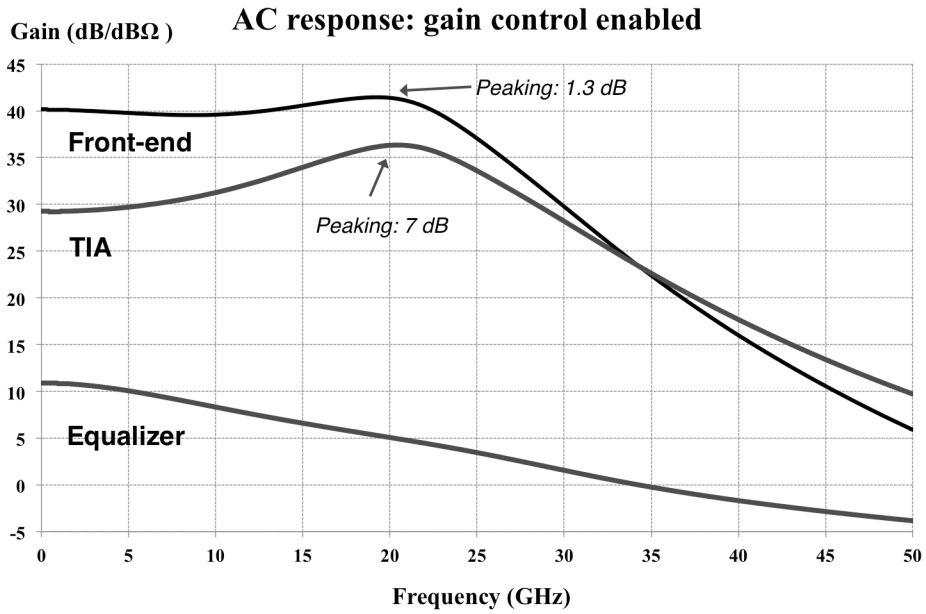
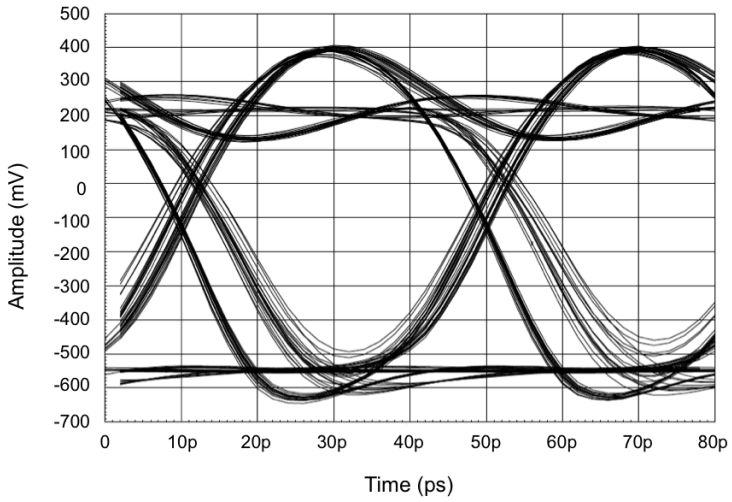


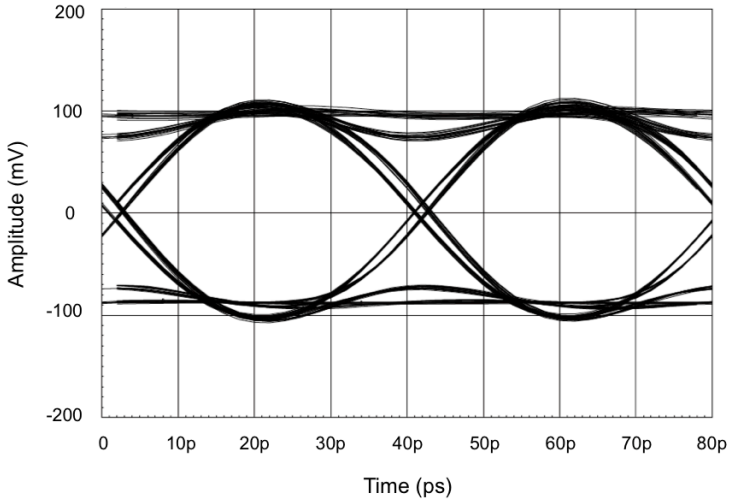
Fig. 5.9 Simulated frequency response for gain control configuration

From Fig. 5.9, as R_F is tuned to minimum (~ 30 Ω), a 7 dB peaking appears in the TIA stage. The equalizer hence is tuned to a low-bandwidth (10 GHz in this case) amplifier to suppress the peaking generate overall flat response. As a result, the two-stage front-end has a maximum gain peaking of only 1.3 dB.

To illustrate this effect in time domain, the output eye with 2 mA_{pp} input current is shown in Fig. 5.10 for two situations: one without gain control and the other with gain control. It is clear that, the output eye in former case is heavily distorted; the eye in the gain control configuration is clear and wide-open, which verifies the effectiveness of this technique.



(a)



(b)

Fig. 5.10 Simulated output eye at front-end with 2 mA_{pp} input current: (a) no gain control; (b) gain control enabled

5.3 Limiting amplifier

From system level simulation, the LA (limiting amplifier) needs DC gain of 22 dB and bandwidth of 22.5 GHz to guarantee overall receiver gain and bandwidth, which translates to total gain-bandwidth product (GBW_{tot}) of 283 GHz for LA. Applying Eq. 3.30 for A_{tot} = 22 dB assuming a second-order Butterworth response, the *GBW* extension ratio is plotted in Fig. 5.11, and the ratio peaks when n = 10. From power and noise point of view, less stages are desirable and it is found that beyond n > 5 the *GBW* extension is marginal. Thus, the LA is set with 5 gain stages, where the corresponding *GBW* extension ratio is 4.71. Hence, *GBW* of each LA stage (GBW_s) is 283/4.71 = 60 GHz. Since the gain of each stage is 22/5 = 4.4 dB (1.66X), bandwidth of each is stage is 60/1.66 = 36 GHz.

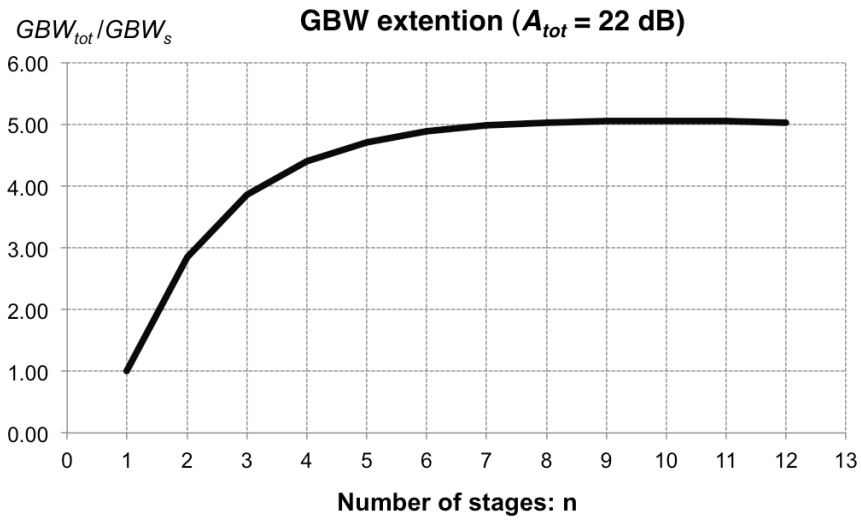


Fig. 5.11 *GBW* extension (A_{tot} = 22 dB) for LA

Active-feedback [21] is adopted to provide further bandwidth and gain control capability. Instead applying this technique to each of the gain stage, which will essentially increase power consumption and parasitics, an inter-stage active-feedback is utilized for the last three stages for optimum gain peaking. The architecture of proposed LA is shown in Fig. 5.12. The more active-feedback (G_{mf}), the larger total bandwidth and the less overall gain. However, the amount of active-feedback should be limited to a certain degree to give rise to small gain peaking. Since active-feedback is essentially shunt peaking, based on the analysis on Section 3.2.2, the bandwidth of open-loop amplifier should be kept large enough to maintain stability (small gain peaking). Based on this, the last two gain stages (Gm4 and Gm5) are half sized to enhance the overall bandwidth of the three-stage cascades within the active-feedback loop.

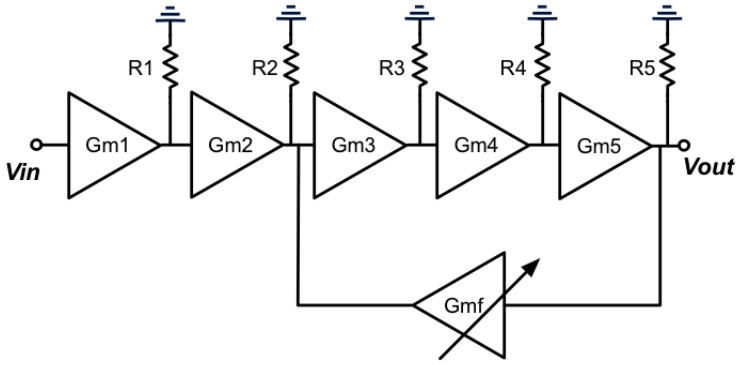


Fig. 5.12 LA architecture (single-ended view)

Having analyzed the LA from architecture level, we move on to the gain stage design. For wideband amplifier, resistor-loaded common-source topology is usually the choice and the voltage gain is given by

$$A = g_m \cdot R_D \tag{5.2}$$

$$= \left(\frac{g_m}{I_D}\right) \cdot I_D \cdot R_D$$

$$= \left(\frac{g_m}{I_D}\right) \cdot V_{R_D} \tag{5.3}$$

where V_{RD} is the voltage drop on load resistor. Since the g_m transistor is biased with high current density for large f_T, the corresponding g_m/I_D is as low as 5 (at current density of 200 μA/μm). To achieve a gain of 1.66, V_{RD} is thus 330 mV, which may be somewhat large since gain stages are DC coupled. Thus, an extra pair of current load that injects current to g_m transistor is adopted to relax the voltage limitation issue. The corresponding LA gain stage schematic is depicted in Fig. 5.13.

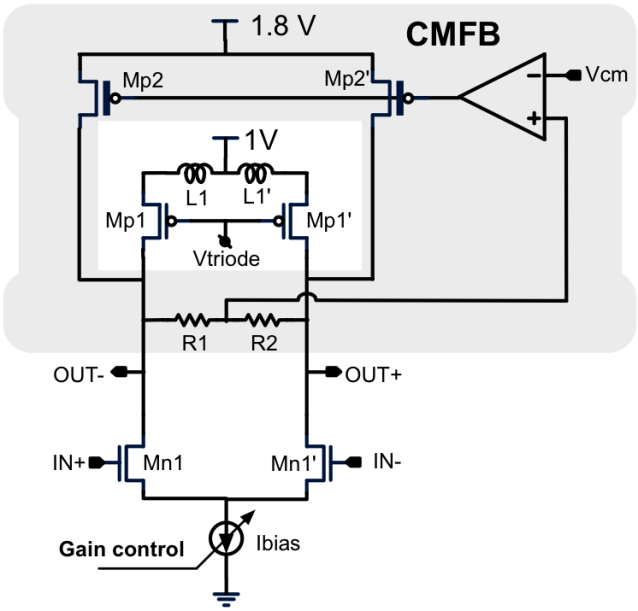


Fig. 5.13 LA gain stage

The LA gain stage employs fully-differential signaling. Shunt peaking is utilized to extend bandwidth and change the gain stage response to second-order, which when cascaded shows slower bandwidth drop, or a larger *GBW* extension ratio as we analyzed before (c.f. Fig. 3.10). Load resistor is realized by PMOS operating in linear region and the resistance can be tuned through V_{triode}. A CMFB loop is added for two purposes: 1) regulating the output common-mode voltage to resist PVT variation; 2) injecting current to g_m device as we explained before. The current load is realized by thick-oxide PMOS (2.5V device) and powered by a higher VDD (1.8V) as a more ideal current source (higher output impedance) that doesn't impact the gain of LA stage.

The bias current can be tuned smaller for LA gain control, which is favorable to avoid excessive jitter when input signal is large [5]. In this design, the bias current of each gain stage can be programmed to half, decreasing the total LA gain by roughly 4x. The CMFB loop works to maintain the output common-mode voltage stable when gain control is exercised (different bias current).

Offset correction is realized at the first stage of LA, where a pair of thick-oxide PMOS the same as that in CMFB but with a smaller dimension is externally controlled to compensate the offset voltage.

Overall, LA achieves 23.5 dB gain and 21.5 GHz bandwidth in nominal condition. Thanks to the active-feedback and tunable resistance from the linear PMOS, through different PVT conditions (FF/TT/SS, 5% VDD, 0~80°C), the variation of gain is 3.9 dB (21.6 ~ 25.5) while bandwidth is larger than 17.8 GHz (17.8 ~ 24.6 GHz), ensuring a stable performance of LA.

5.4 Output buffer

The Output buffer needs to drive 50-Ω off-chip (AC coupling) load, which requires both output impedance matching and large current drive capability. As a result, the output stage of the Output buffer uses large transistor size and bias current. On the other hand, not to overload LA, transistor size in input stage of Output buffer should be kept small. Therefore, a tapered topology is necessary as already depicted in Fig. 3.18. In this design, three gain stages are adopted considering both signal fan-out and power consumption. The gain stage schematic is shown in Fig. 5.14.

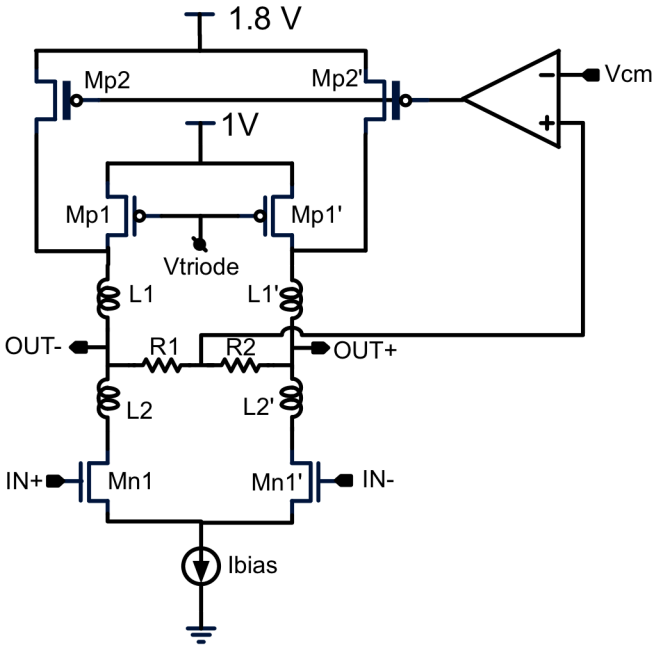


Fig. 5.14 Output buffer gain stage

The topology is similar to that of LA, where the CMFB serves the same purpose. An extra pair of inductor L2/L2' is added to match the inter-stage transmission line impedance and compensates high-frequency loss. The output stage of the buffer employs 50-Ω poly-resistor instead of linear PMOS for accurate output impedance matching.

Not to impact the overall receiver bandwidth, the bandwidth requirement for Output buffer is more challenging. In this design, from system simulation it is around 50 GHz. To cope with this huge bandwidth need, stagger-tuning technique is utilized and the corresponding frequency response is shown in Fig. 5.15. By properly setting the peaking inductor L1, the first stage and second stage peak at 24 GHz and 36 GHz respectively, together with the third stage synthesizing an overall bandwidth more than 50 GHz and out-of-band gain

peaking of 2 dB. Though PVT variation (FF/TT/SS, 10% VDD, 0~80°C), the variation of gain is 3 dB (1 ~ 4) while bandwidth in the worst case is larger 38 GHz.

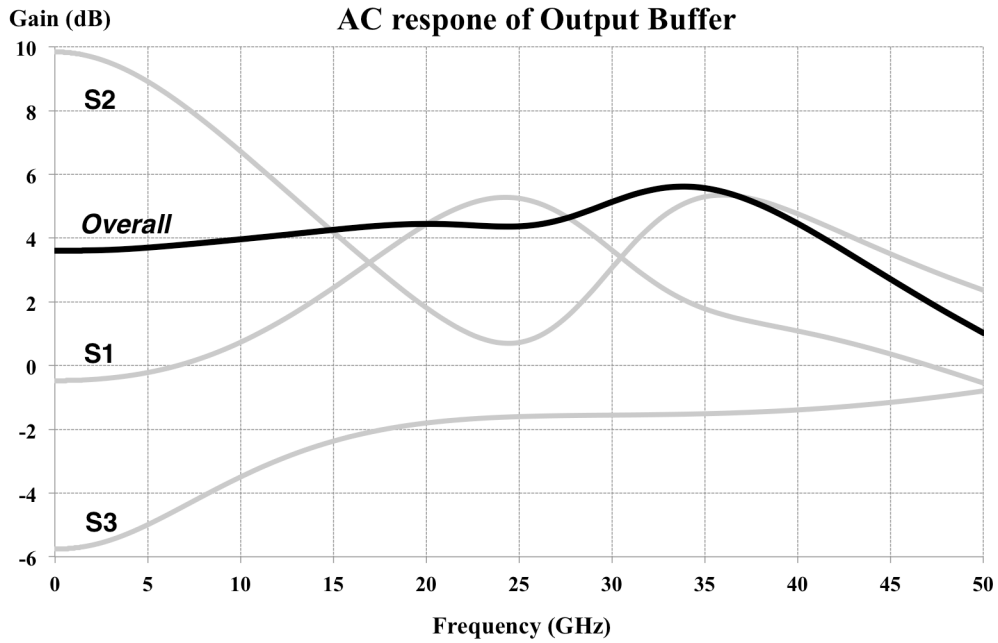


Fig. 5.15 Stagger-tuning on Output buffer

5.5 Receiver performance

After introducing the building blocks, the whole receiver performance has also been examined. Assuming the same external interface as before (C_{PD} = 80 fF, L_{BW} = 0.5 nH), in nominal case (TT, 1V, 27 °C), the receiver gain and bandwidth is summarized in Table 5.3 and Fig. 5.16.

Table 5.3 Receiver performance summary

Block	Front-end	LA	Buffer	**RX**
Gain (dB/dBΩ)	57.2	23.5	3.6	**84.3**
BW (GHz)	22.4	21.5	> 50	**19.0**
Power (mW)	13.3	25.6	49.3	**88.2**

The input-referred RMS noise current of receiver is 2.29 μA_{rms}, which translates to 16.6 pA/$\sqrt{Hz}$ average input-referred noise spectrum density, or a input sensitivity of 32 μA_{pp}.

The receiver output eye at sensitivity is plotted in Fig. 5.17 and the output eye with large input current (2 mA_{pp}) is also examined with and without gain control, shown in Fig. 5.18. The gain control in this case is applied to both front-end and LA to achieve the best output eye. As a result, the receiver achieves an input dynamic range of 36 dB (32 μA_{pp} ~ 2 mA_{pp}).

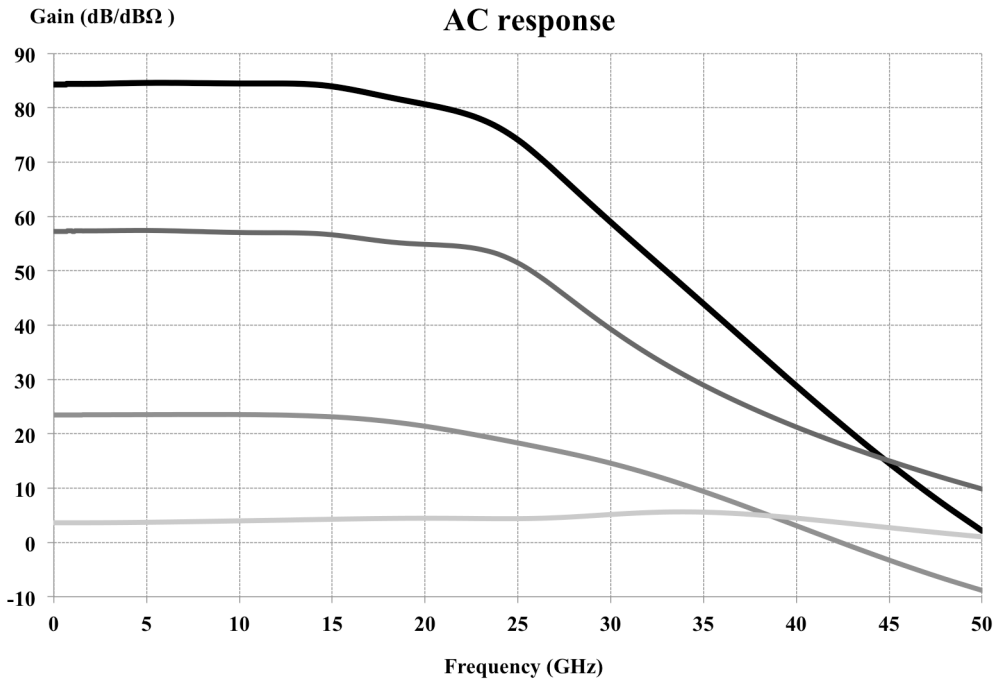


Fig. 5.16 Receiver AC response

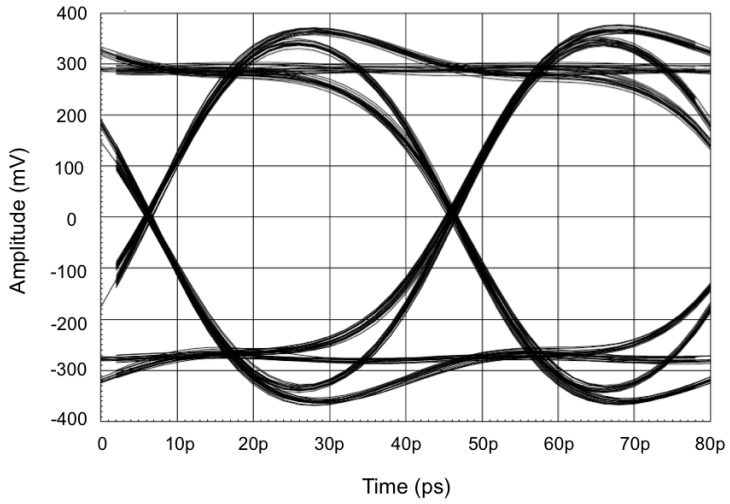
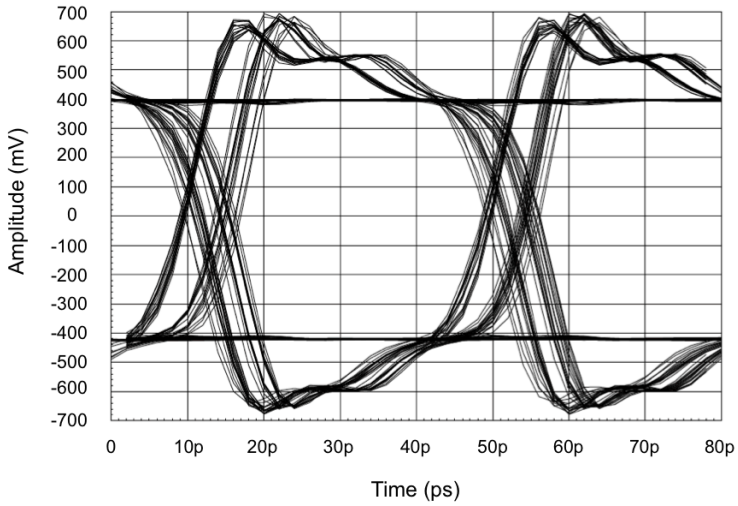
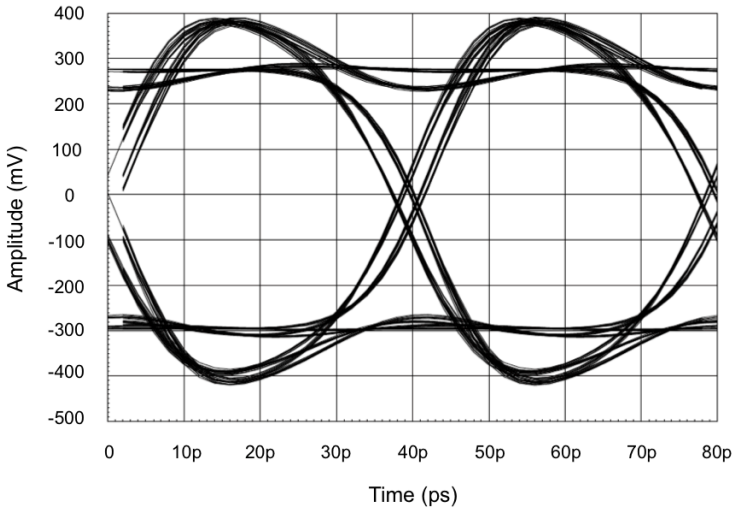


Fig. 5.17 Receiver output eye at sensitivity



(a)



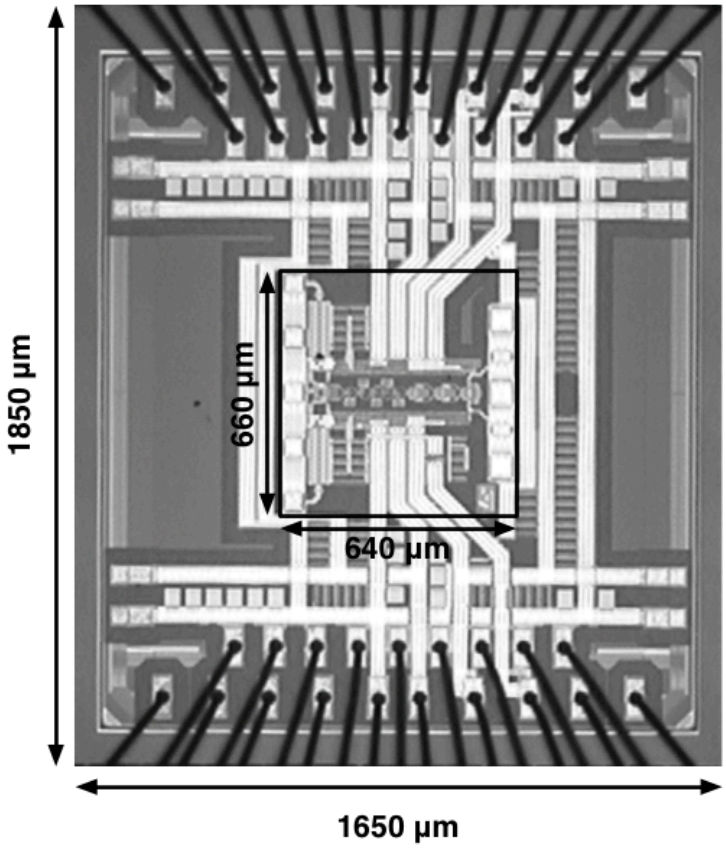
(b)

Fig. 5.18 Receiver output eye with 2 mA_{pp} input current: (a) no gain control; (b) gain control enabled

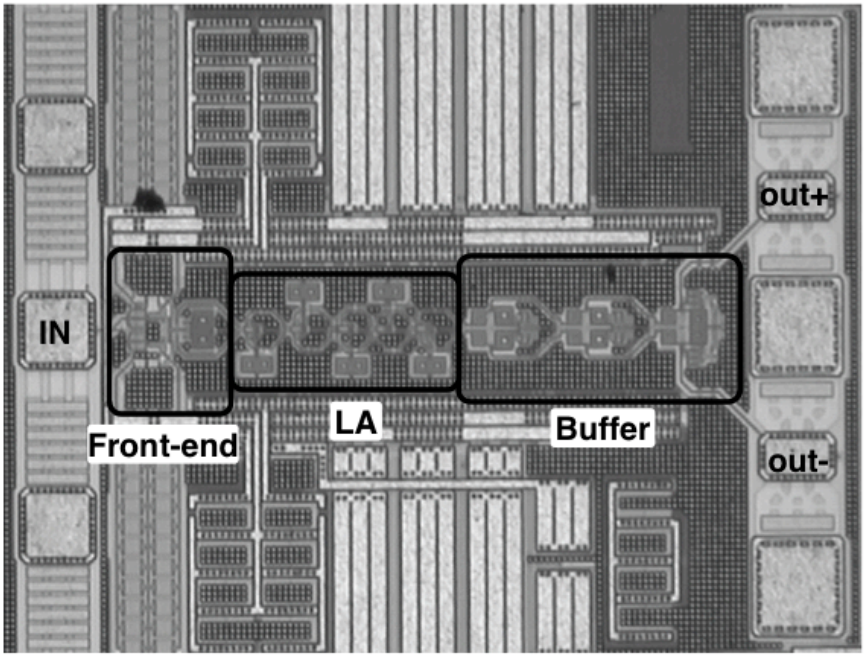
5.6 Experimental results

The receiver chip has been fabricated in STMicroelectronics 65-nm Bulk CMOS technology and the photomicrograph is shown in Fig. 5.19. The core occupies 0.64 mm x 0.66 mm (0.42 mm^2). The chip is pad-limited due to many test pins are added. The input and output RF pads are placed out of the main pad ring for optimized RF performance. The input pads are set in a B-G-S-G-B layout, where the inner G-S-G is used for probe test and the outer B-S-B is used to wire-bond the commercial planar photodiode with large pitch. In such a way, the chip can be measured by either electrical probing or wire-bonded with photodiode.

Additional supply filters are added along the GND and VDD distribution lines to improve immunity to external disturbances.



(a)



(b)

Fig. 5.19 Chip microphotograph: (a) full-size; (b) zoom-in

The receiver draws 26.5 mA from 1V and 36.9 mA from 1.8V, resulting a total DC power consumption of 93 mW. The characterization has been performed without the photodiode which enables a direct comparison of simulation and measurement. G-S-G input pads featuring 80 fF capacitance are used to for electrical probe testing. S-parameter measurements have been performed using a Vector Network Analyzer (VNA) with 50 GHz bandwidth, directly accessing the I/O pads via RF probes. Three-ports characterization is applied to the receiver: one for the input and two for the outputs based on the single-ended in, differential output receiver I/O characteristics. Fig. 5.20 shows the resulting differential transimpedance gain, derived from measured S-parameters as [36]

$$Z_T = \frac{Z_0(S_{21} - S_{31})}{1 - S_{11}} \tag{5.4}$$

where S21 and S31 have the same amplitude but are out-of-phase.

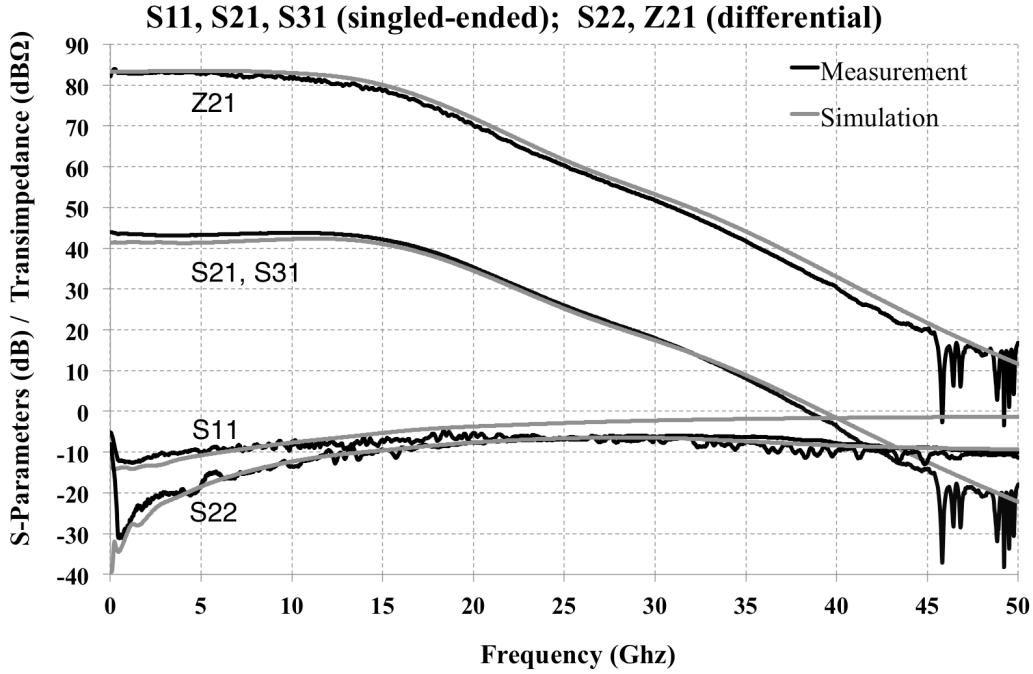


Fig. 5.20 S-parameters and Z-parameter: measurement and simulation

A differential transimpedance gain of 83 dBΩ over a -3-dB bandwidth of 13.6 GHz (pure electrical) has been achieved for the receiver. Comparisons between simulations and measurements show good agreement with each other. Furthermore, a 10.6 GHz ~ 18.2 GHz bandwidth tuning and 87.1dB ~ 78 dB gain tuning are achieved by tuning the bandwidth control signal at the equalizer stage in the front-end, shown in Fig. 5.21.

Measured and simulated input-referred noise current spectral density is reported in Fig. 5.22. The measurement has been performed by detecting the output noise voltage spectral density through spectrum analyzer with receiver input floated. The outcome is then divided by the measured transimpedance gain shown in Fig. 5.20. The in-band value is larger than resulting from Fig. 5.6 due to the contribution from the dummy mirror TIA, introduced in order to improve supply rejection, key in wire-line applications. The out-of-band noise grows more slowly than resulting from Fig. 5.6 due to the absence of the photodiode

determining a lower input capacitance. The input-referred RMS noise current has been computed as the measured output RMS noise, divided by the measured in-band transimpedance gain, leading to 2.44 μA_{rms}, meeting the required sensitivity with margin. Dividing the RMS input referred noise by $\sqrt{BW}$ (c.f. Eq. 3.5) determines an average input-referred noise current spectral density of 20.9 pA/$\sqrt{Hz}$.

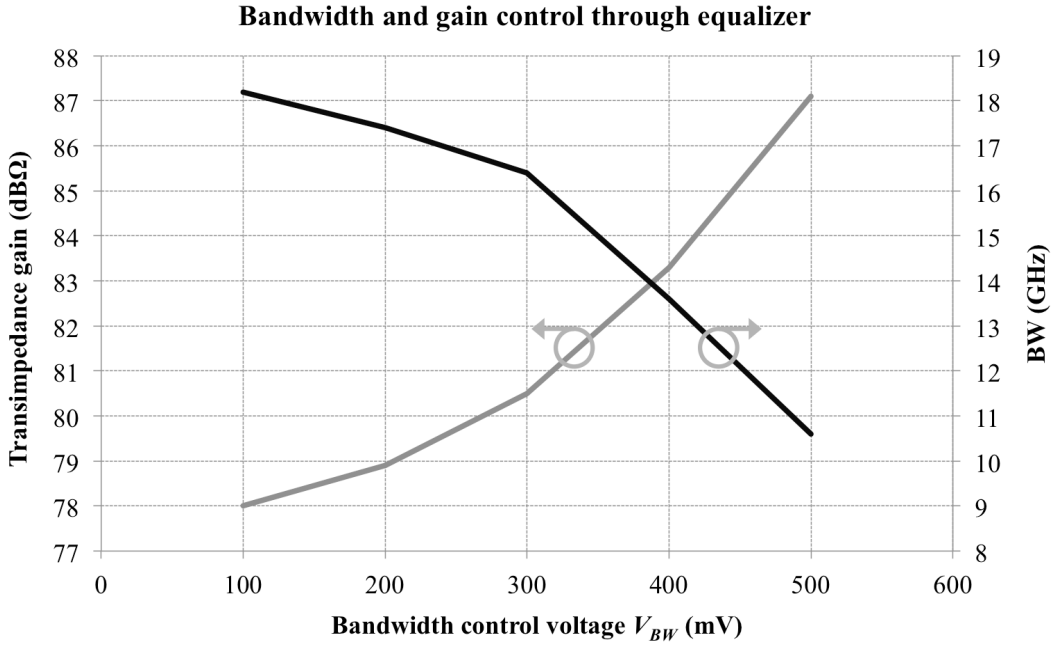


Fig. 5.21 Bandwidth and gain control though equalizer in front-end

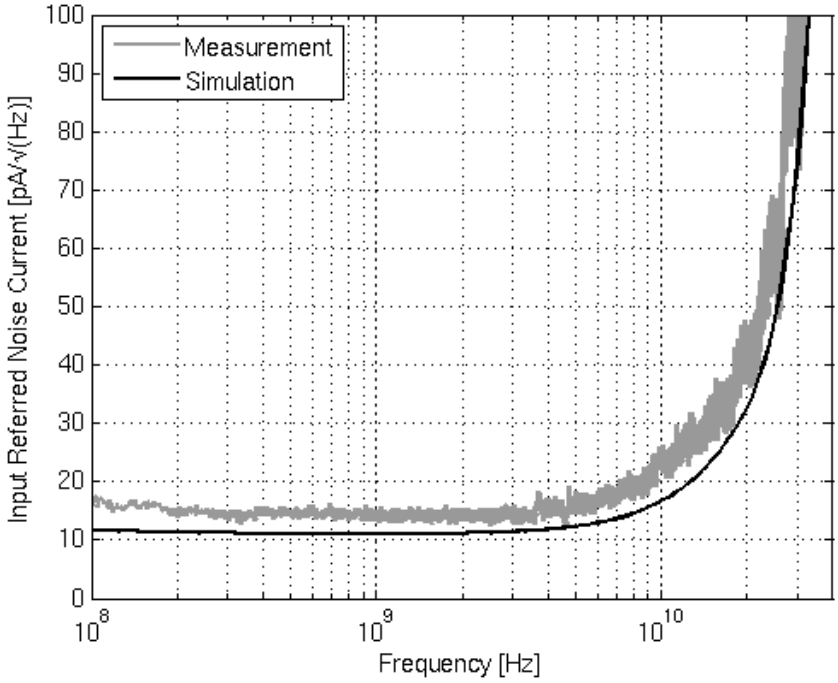


Fig. 5.22 Input-referred noise spectrum density

Fig. 5.23 shows the differential eye diagram at receiver output, measured at data rate of 25 Gb/s using 2^{31}-1 PRBS source, with input voltage of around 14

mV_{pp}. The eye diagram is affected by the input signal from the external pattern generator, in terms of both rise/fall time and jitter. Although the receiver is not optimized for to 50-Ω input source, the output eye is still clear and wide-open, demonstrating good data transmission capability at 25 Gb/s.

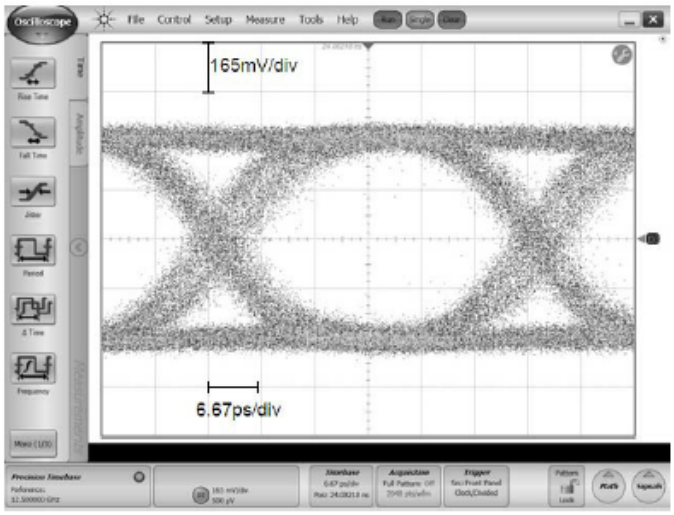


Fig. 5.23 Electrical output eye diagram at receiver output, with 14 mV_{pp} input voltage

Finally, Table 5.4 summarizes the performance of the receiver compared to the state-of-the-art and shows this work achieves the best FOM. Since the dummy mirror TIA doubles the integrated input-referred noise, the noise is slightly larger than [29], which is realized in a SOI technology twice faster.

Table 5.4 Receiver performance summary

Reference	This Work	CICC10 [27]	JSSC12 [30]	JSSC08 [25]	JSSC12 [29]
CMOS Tech. (nm)	**65**	65	130 (SOI)	90	45 (SOI)
Bit Rate (Gb/s)	**25**	25	25	40	40
Bandwidth (GHz)	**13.6**	22.8*	25	22	30
Transimpedance (dBΩ)	**83**	69.8*	67	66	55
Supply (V)	**1 / 1.8**	1 / 1.8	1.2	1.2	1
Power (mW)	**93**	74	48	75	9
Noise ($pA/\sqrt{Hz}$)	**20.9*** **	-	40	22**	20.5
FOM (GHz·Ω/mW)	**2066**	952	1166	585	1874
*: simulated, **: differential, ***: including dummy mirror TIA					

5.7 Conclusion

In this chapter, we have presented a 25 Gb/s 65-nm receiver intended to interface an external commercial photodiode for 100GBASE-LR4 optical communications. Various low-noise approaches proposed have been used in this design, demonstrating the effectiveness of our low-noise design techniques. The receiver has achieved excellent low-noise performance with state-of-the-art FOM.

Chapter 6

A 25 Gb/s optical receiver for silicon photonics

In this chapter, we present a 25 Gb/s optical receiver tailored to silicon photonics application where the optical part highlights low-parasitic capacitance Ge-on-Si (germanium on silicon) waveguide photodiode. The proposed low-noise design techniques in conjunction with the favorable properties of silicon photonics generate the lowest noise for optical receiver with more than 10 Gb/s data rate. The hybrid integration approach is able to provide high performance, high degree of integration, better flexibility and low cost.

6.1 Introduction

In recent years, silicon photonics has evolved from R&D level to maturation pushed by the exponential growth of bandwidth demand and the never-ending inquest for low cost, low power and higher degree of integration, which has been pursued by the semiconductor industry for more than 40 years. After Luxtera corp. demonstrated the first silicon photonics transceiver in 2006, more and more industry players like Intel, IBM, Oracle, STM and etc. have joined this field, stimulated by its promising blueprint. A prediction from STM [37] on silicon photonics evolvement for the next decade is shown in Table 6.1.

Table 6.1 Silicon photonics for next 10 years [37]

Speed increase	80x
Power reduction	20x
Cost reduction	100x
Silicon area reduction	5000x

In this work, we target the optical receiver design intended to interface silicon photonics. Rather than integrating optics and electronics on the same silicon chip, a hybrid approach is utilized, shown in Fig. 6.1.

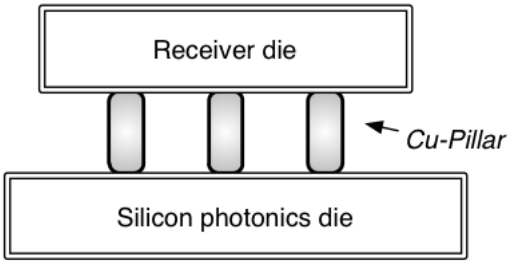


Fig. 6.1 Integration concept of silicon photonics die and receiver die

The separation of silicon photonics die and receiver die enables independent optimization of performance for each part. The silicon photonics die is fabricated in a tailored SOI CMOS technology, where additional semiconductor compound is added for optical performance. The receiver die on the other hand employs a standard Bulk CMOS technology to benefit from the continuous CMOS scaling. Thanks to the advancement of 3D connection technology like Cu-pillar, low parasitic and high performance chip interconnect is obtained while maintaining high degree of integration level. In such a hybrid way, performance, cost, integration, and power are all achieved simultaneously with flexibility.

From receiver circuits point of view, the most critical aspect is the interfacing parasitics from silicon photonics and the interconnection (Cu-pillar). A Ge-on-Si waveguide photodiode on silicon photonics technology is shown in Fig. 6.2. Since the light is coupled horizontally, high quantum efficiency can be achieved with very low parasitic capacitance in contradict with the traditional vertical illuminated photodiode. As a result, parasitic capacitance is as low as 10 ~ 15 fF, 5 ~ 10 times smaller than that of conventional photodiode. The capacitance from Cu-pillar thus needs some attention, since it is on the range of 20 ~ 30 fF, even larger than that from photodiode. Overall, the receiver circuit sees a total external capacitance of around 45 fF.

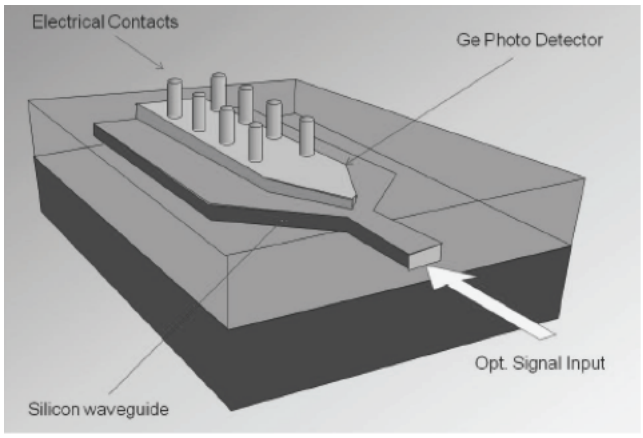


Fig. 6.2 Ge-on-Si waveguide photodiode [37]

6.2 Noise scaling

It is instrumental to examine the noise relationship with external input capacitance, in view of its critical role in our previous analysis. Recall Eq. 4.22

and Eq. 4.40, we conclude that for both conventional approach and proposed two-stage front-end, the integrated input-referred noise

$$\overline{I_{n,in,int}^2} \propto C_{PD} \tag{6.1}$$

where C_{PD} includes the contribution from Cu-pillar and the bandwidth scaling factor in the two-stage front-end doesn't change. Accordingly, the input-referred RMS noise and the average input-referred noise spectrum density

$$I_{n,in}^{rms}, I_{n,in}^{avg} \propto \sqrt{C_{PD}} \tag{6.2}$$

In our previous design, the external input capacitance is 160 fF and the average input-referred noise spectrum density is 20.9 pA/$\sqrt{Hz}$. Apply this noise scaling scenario, for Ge-on-Si photodiode with a total external capacitance of 45 fF, the average input-referred noise spectrum density is expected to be

$$\frac{20.9}{\sqrt{\frac{160}{45}}} = 11.1\ pA/\sqrt{Hz}$$

If we are able to remove the dummy mirror TIA, the noise will become 3 dB smaller, or 7.8 pA/$\sqrt{Hz}$. This analysis gives first-order estimation of the noise performance we may achieve.

6.3 Circuit design

The receiver architecture is shown in Fig. 6.3, similar to the previous realization. Ge-PD and Cu-pillar together gives 45fF capacitance. The two-stage low-noise two-stage front-end topology has been employed for low noise profile and high sensitivity. Smaller total input capacitance results larger transimpedance gain from front-end stage and the LA is made of four stages because less gain is needed. Output buffer has been reused to secure the overall bandwidth.

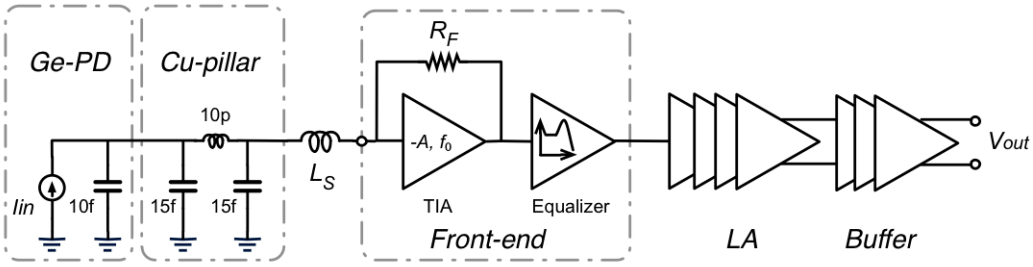


Fig. 6.3 Receiver architecture

A spiral inductor L_S is realized on-chip for series peaking, the same as the role of bondwire, while it is more accurately defined and modeled. The series peaking inductor is optimized based on the trade-off: low parasitic capacitance is desired for low noise due to the important role of input capacitance and large enough self-resonance frequency; however low parasitic capacitance indicates less metal width of inductor and higher series resistance, which lowers the Q of series peaking and contributes more noise.

To fully realize the low noise potential from silicon photonics receiver as we have calculated earlier, the front-end adopts single-ended topology. This on one

hand drops the 3-dB noise overhead and power consumption from the dummy TIA, on the other hand requires a better AC ground. Accordingly, several hundred picofarads of on-chip capacitors are added to for better supply coupling.

The schematic of front-end circuit is shown in Fig. 6.4 where the low-noise two-stage front-end architecture is employed. The bandwidth of TIA stage is designed around half of full bandwidth as a compromise of noise reduction, flat frequency response and equalizer tuning margin. Input device C_{gg} is sized close to its theoretical optimum, half of C_{PD} in this case. CMOS-TIA is adopted for amplifier noise reduction (g_m-reuse), as explained in Section 4.5 and 5.2. An equalizer stage using inductive shunt peaking is utilized with sufficient tuning range (peaking from 0~10 dB). Both TIA and equalizer can be tuned independently by V_{ZT} and V_{BW} for optimum performance and large input signal configuration: gain peaking in the TIA stage due to small R_F can be suppressed by equalizer stage (configured as a low-bandwidth amplifier) for an overall flat frequency response, as explained in Chapter 5. A passive RC low-pass filter draws the low-frequency component of equalizer output to feed the negative input of the first stage of LA.

To ease the on-board tuning for chip test, two analog loops are added for offset correction and CMFB, which will be explained in detail.

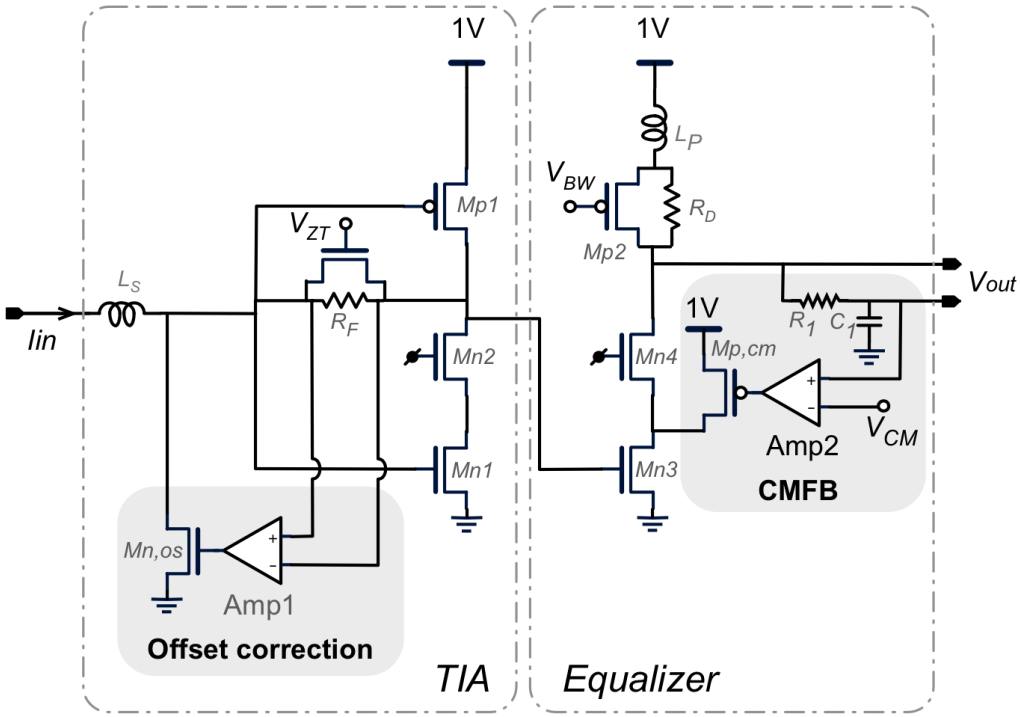
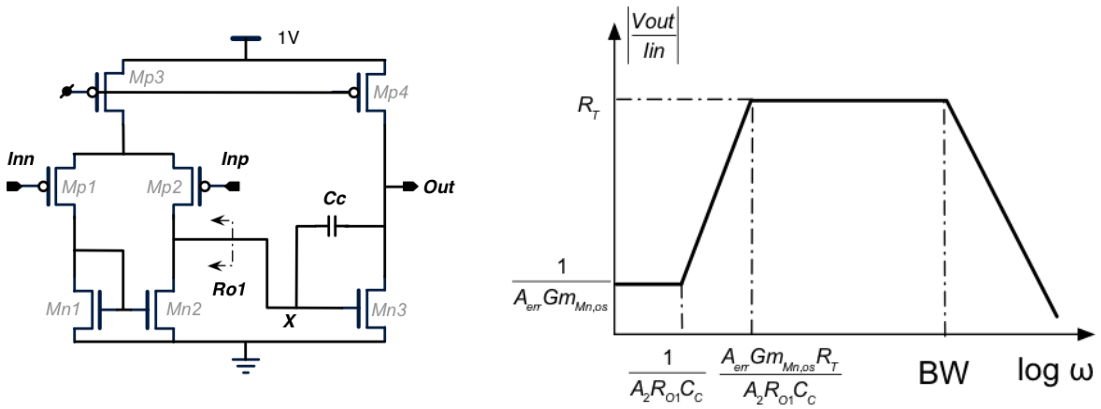


Fig. 6.4 Schematic of two-stage front-end

Offset correction. The input offset current is sensed by the voltage drop across R_F as shown in Fig. 6.4. The concept of offset correction has been introduced in Section 3.5.1 where external capacitor is often needed to make the low-frequency cutoff sufficiently low to avoid baseline wander. In this design, we explore Miller effect to avoid external capacitor. The schematic of error amplifier (Amp1) is shown in Fig. 6.5. (a).



(a) (b)

Fig. 6.5 (a) Amp1; (b) Frequency response of receiver due to offset correction

Based on the common-mode voltage of TIA input (output), a P-type Miller amplifier is adopted. The main pole of the loop ($1/A_2R_{O1}C_C$) is realized at point X due to Miller effect, where A_2 is the gain of second stage of error amplifier and R_{O1} is the output resistance of first stage of error amplifier. Since the capacitive load of second stage is a small, the RHP (right half plane) zero is located at high frequency and the zero compensation resistor is not needed. The low-frequency cutoff is given by

$$\omega_{LF} = \omega_x \cdot LG \tag{6.3}$$

$$\omega_{LF} = \frac{1}{A_2R_{O1}C_C} \cdot A_{err}Gm_{Mn,os}R_T \tag{6.4}$$

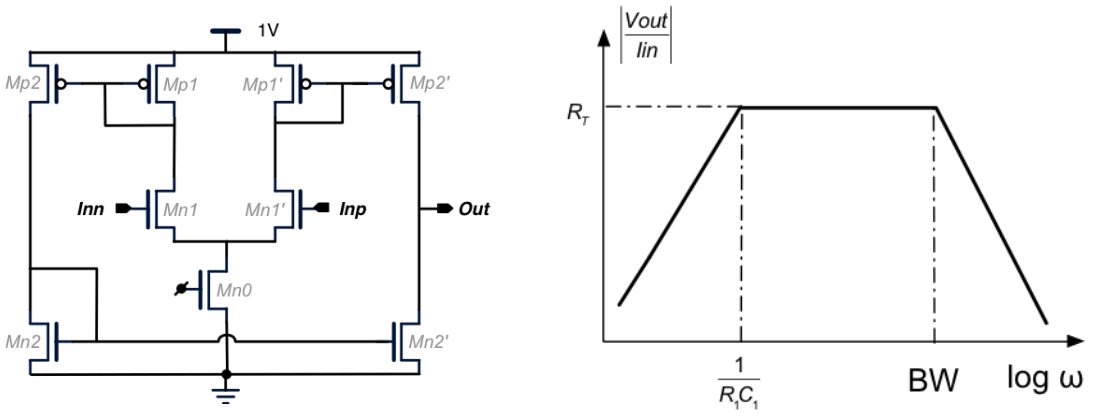
where A_{err} is the gain of error amplifier, $Gm_{Mn,os}$ is the transconductance of the current sink transistor (*Mn,os* in Fig. 6.4) and *LG* is loop gain. The output offset is

$$V_{OS,out} = \frac{Iin}{A_{err}Gm_{Mn,os}} \tag{6.5}$$

and the corresponding frequency response of the front-end is also plotted Fig. 6.5 (b).

In this design, a 1 pF MOM (Metal-Oxide-Metal) capacitor is used for Cc, generating main pole at around 10 KHz for Amp1. Correspondingly, the low-frequency cutoff is *LG* times higher, at around 2 MHz. As input current goes larger, $Gm_{Mn,os}$ also is higher which raises the loop gain and low-frequency cutoff whereas decreases phase margin. However, since the forward amplifier gain (R_T) is lowered for large signal operation, the loop gain keeps roughly unchanged and the phase margin can be secured. Within the input dynamic range, the phase margin is larger than 80°.

CMFB. The equalizer tuning changes the load resistance and thus the output common mode, which requires CMFB, as shown in Fig. 6.4. Since the main pole of the loop is set by the passive RC low-pass filter, the error amplifier (Amp2) is designed to have only one gain stage, as shown in Fig. 6.6 (a).



(a) (b)

Fig. 6.6 (a) Amp2; (b) Frequency response of receiver due to low-pass RC filter

Amp2 employs a typical N-type transconductance amplifier topology since common mode voltage is around 800 mV. For stability, the main pole of Amp2 (output pole) should be set at least LG times higher than $1/(R_1C_1)$. In this design, the phase margin in CMFB loop is around 120°.

Due to the existence of $1/(R_1C_1)$, another high-pass is formed in the receiver transfer function, as shown in Fig. 6.6 (b). Thus R_1C_1 should be set large enough that it doesn't change the low-frequency cutoff generated by offset correction.

6.4 Results

The receiver as a whole has been characterized assuming the aforementioned photodiode and Cu-pillar parasitics. The receiver AC response at sensitivity is shown in Fig. 6.7 and the AC performance is summarized in Table 6.2.

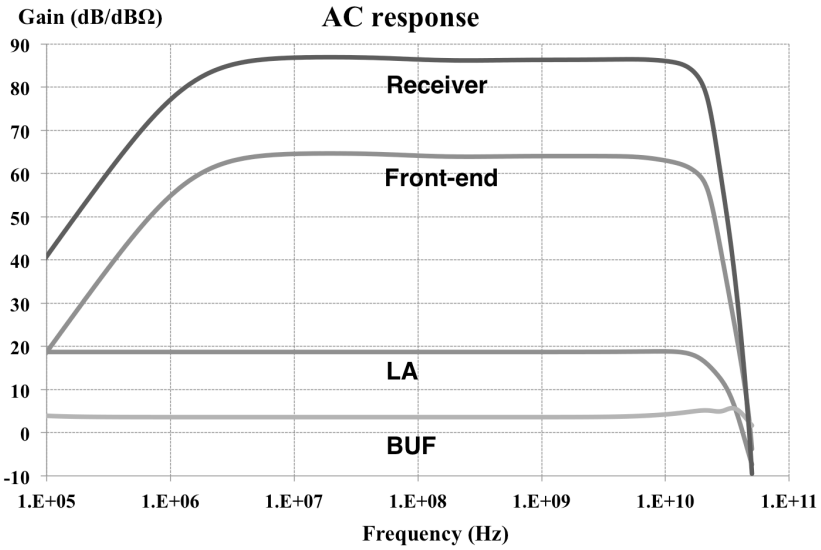


Fig. 6.7 Receiver AC response at sensitivity

Table 6.2 Receiver performance summary

Block	Front-end	LA	Buffer	**RX**
Gain (dB/dBΩ)	64	18.7	3.6	**86.2**
BW (GHz)	16.4	22.1	> 50	**17.4**
Power (mW)	5.2	19.5	49.4	**74.1**

The input-referred RMS noise current of receiver is 1.06 μA_{rms}, which translates to 8.1 pA/$\sqrt{Hz}$ average input-referred noise spectrum density, or a input sensitivity of 15 μA_{pp}. This result matches closely to the predicted value from noise scaling (7.8 pA/$\sqrt{Hz}$). The input-referred noise spectrum density is also plotted in Fig. 6.8. The high-frequency noise suppression from series peaking can be clearly seen from the plot.

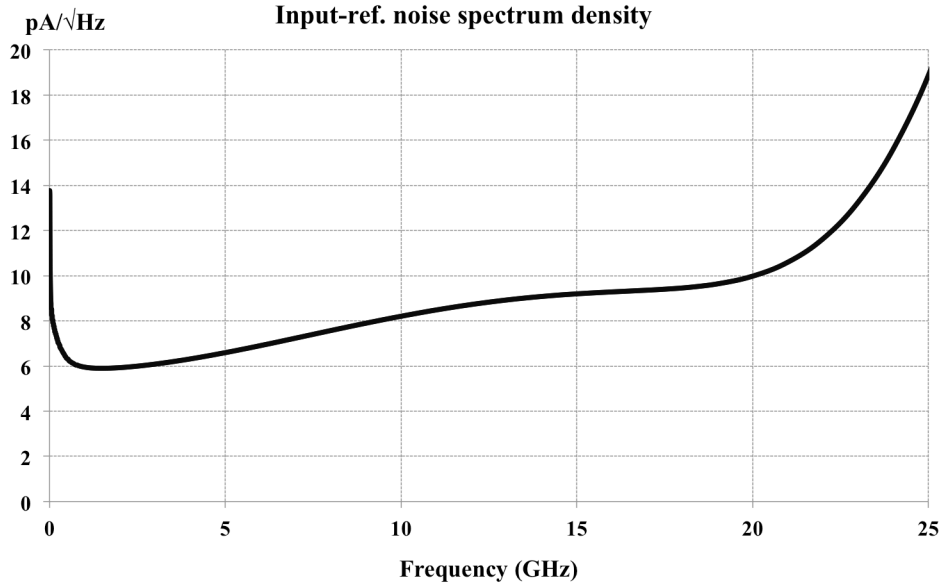
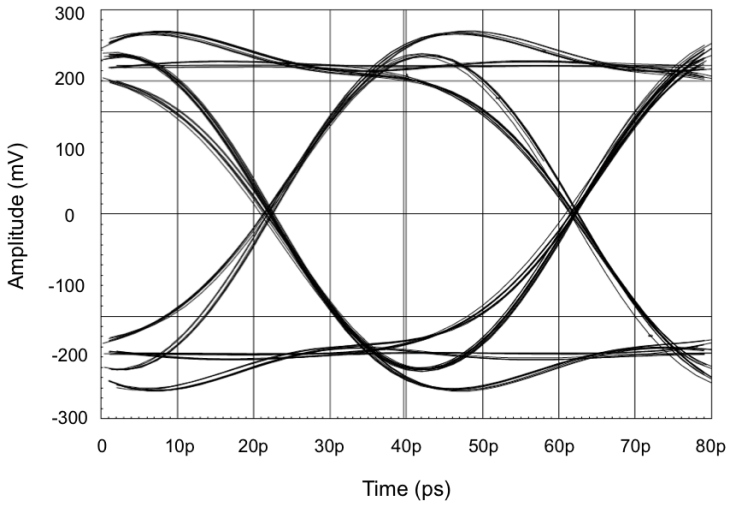
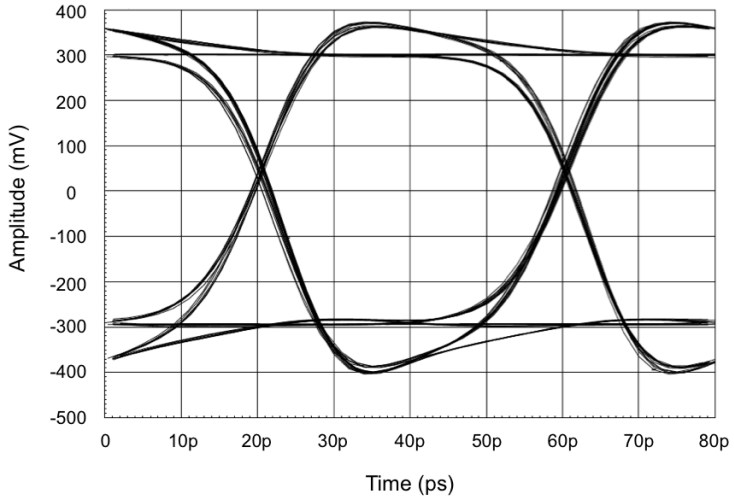


Fig. 6.8 Input-referred noise spectrum density

The receiver achieves an input dynamic range of 34 dB (15 μA_{pp} ~ 750 μA_{pp}). The receiver output eye at sensitivity (15 μA_{pp}) and overload limit (750 μA_{pp}, gain control enabled) simulated with using 2^7-1 PRBS source is plotted in Fig. 6.9, where input signal rise/fall (10%- 90%) time is 8.3 ps. The deterministic jitter for the two cases are 1.8 ps and 2.9 ps respectively.



(a)



(b)

Fig. 6.9 Receiver output eye at: (a) sensitivity (15 μA_{pp}); (b) overload limit (750 μA_{pp}) with gain control

Finally, Table 6.3 summarizes the performance of the receiver compared to the state-of-the-art. The lowest input-referred noise is achieved together with the best FOM, demonstrating the efficacy of the proposed low-noise design technique and the advantage provided by silicon photonics.

Furthermore, compared with [30] where silicon photonics and receiver circuits are integrated on the same chip, this work achieves 5x lower noise and 4x better FOM, which reveals the advantages of our hybrid integration approach. Very soon, faster receiver chip can be integrated with existing silicon photonics chip and more advantages can be expected. With the advancement of wafer level bonding, the hybrid approach can offer the best performance and flexibility with low cost and high degree of integration.

Table 6.3 Receiver performance summary

Reference	This work*	Previous Work****	CICC10 [27]	JSSC12 [30]	JSSC08 [25]	JSSC12 [29]
CMOS Tech. (nm)	**65**	65	65	130 (SOI)	90	45 (SOI)
Bit Rate (Gb/s)	**25**	25	25	25	40	40
Bandwidth (GHz)	**17.4**	13.6	22.8*	25	22	30
Transimpedance (dBΩ)	**86.2**	83	69.8*	67	66	55
Supply (V)	**1 / 1.8**	1 / 1.8	1 / 1.8	1.2	1.2	1
Power (mW)	**74.1**	93	74	48	75	9
Noise (pA/$\sqrt{Hz}$)	**8.1**	20.9***	-	40	22**	20.5
FOM (GHz·Ω/mW)	**4794**	2066	952	1166	585	1874
*: simulated, **: differential, ***: including dummy mirror TIA, ****: work in chapter 5						

6.5 Conclusion

In this chapter, we have applied our proposed low-noise design techniques to a silicon photonics optical receiver: the lowest input-referred noise and the best FOM have been achieved compared with the state-of-the-art. This not only showcases the capability of our low-noise design techniques, but also proves the advantage of silicon photonics and our hybrid integration approach.

Chapter 7

General conclusion

In the late 1990s, optical communication has gone through major development on backbone telecommunication network, accompanied by the transformation from voice communication to data communication. The recent boom of Internet, supercomputing, storage network and data center have again boosted the growth of communication infrastructure, featuring massive utilization of optical communication from core network to lower communication hierarchy/node, e.g. rack-to-rack, backplane and even board level, in view of the near limit of electrical communication. This work has addressed low-noise circuit design techniques for CMOS high-speed optical receiver, one of the most critical issues on optical communication, in several aspects:

- By careful and solid analysis of shunt-shunt feedback front-end transimpedance amplifier working at high-speed regime, new optimum design methodology in terms of device size and bias has been determined. Compared with conventional design methodology, noise, transimpedance gain and power consumption have shown considerable improvement.
- A novel low-noise two-stage optical front-end topology has been proposed, that theoretically eliminates all major low-frequency noise contributors, which have become dominant in high-speed optical receiver circuits. This technique enables high-speed receiver circuits benefit the superior noise performance as a low-speed receiver but still maintain the bandwidth needed to operate at high-speed.
- G_m-reuse technique for LNA has been explored under the context of transimpedance amplifier design. It is found that this technique is more favorable in proposed device optimization scenario rather than the conventional noise optimization scenario.

The aforementioned low-noise design techniques can work together to give superior overall noise reduction, proved by two design examples.

- A 65-nm 25 Gb/s CMOS optical receiver for discrete photodiode tailored to 100GBASE-LR4 has been proposed employing low-noise two-stage front-end topology, providing 4x noise power reduction compared to a traditional design approach. A fabricated receiver prototype demonstrates a transimpedance gain of 83 dBΩ, an input-referred equivalent RMS noise current of 2.44 μA and an electrical analog bandwidth tunable between 10.6 GHz and 18.2 GHz. The power consumption is 93 mW with a FOM of 2066 GHz·Ω/mW.
- A 65-nm 25 Gb/s CMOS optical receiver for silicon photonics has been designed employing various proposed low-noise design techniques. The Ge-on-Si waveguide photodiode offered by silicon photonics technology exhibits low capacitance, favorable for low-noise optical receiver design. A hybrid integration approach has been adopted, which enables

independent optimization of both receiver circuits and photonics, where integration, performance, cost and power are achieved simultaneously. The receiver has accomplished a transimpedance gain of 86.2 dBΩ, an input-referred equivalent RMS noise current of 1.06 μA and an electrical analog bandwidth of 17.4 GHz. The power consumption is 74.1 mW with a FOM of 4794 GHz·Ω/mW. This work demonstrates the lowest input-referred noise and best FOM for high-speed (> 10 Gb/s) optical receiver compared with the state-of-the-art, which proves the efficacy of our low-noise design techniques, and the advantage of silicon photonics hybrid integrated with receiver circuits.

The fast growth of optical communication will continue, given the constant quest of bandwidth in the information age. In the future, optical communication will expand to a more extensive and deeper level, e.g. chip-to-chip and even intro-chip, thanks to the advancement of silicon photonics and CMOS scaling on future technology node, which together will offer more advantages on integration, speed, cost and power. Moore's law in terms of miniaturization and scaling has upgraded to a broader sense, from pure electronics to hybrid on-chip systems with the integration of optics, MEMS, Bio-sensors, etc.; and from 2D to 3D with the progress of new integration and interconnect technology. For the circuit designers, pursue of new circuit techniques that enable low noise, high gain, large bandwidth and low power will continue to cope with the rapid development of optical communication and other hybrid systems made on-chip.

Bibliography

[1] "IP/MPLS Networks: Optimize Video Transport for Broadcasters", white paper, Cisco corporation, 2011.

[2] "IEEE 802.3ba". [Online], available: http://www.ieee802.org/3/ba/index.html

[3] "Fiber-optic communication", [Online], available: http://en.wikipedia.org/wiki/ Fiber-optic_communication

[4] "An overview: the next generation of Ethernet", [Online], available: http://www.ieee802.org/3/hssg/public/nov07/ HSSG_Tutorial_1107.zip

[5] E. Sackinger, Broadband Circuits for Optical Fiber Communication, John Wiley & Sons, 2005.

[6] "Implementation Agreement for Integrated Dual Polarization Intradyne Coherent Receivers", white paper, Optical Internetworking Forum (OIF), 2010.

[7] B. Razavi, Design of Integrated Circuits for Optical Communications, McGraw-Hill, 2003.

[8] "NRZ Bandwidth - HF Cutoff vs. SNR", application note, Maxim Integrated.

[9] M. Azadeh, Fiber Optics Engineering, Springer, 2009.

[10] V. Lal et al. "Terabit photonic integrated circuits in InP: 10-channel coherent PM-QPSK transmitter and receiver PICs operating at 100 Gb/s per wavelength", in Photonics Society Summer Topical Meeting Series, Jul. 2011, pp. 117-118.

[11] A. Huang, C. Gunn, G.-L. Li, Y. Liang, S. Mirsaidi, A. Narasimha, and T. Pinguet, "A 10Gb/s photonic modulator and WDM MUX/DEMUX integrated with electronics in 0.13μm SOI CMOS" in IEEE ISSCC Dig. Tech. Papers, 2006, pp. 922–929.

[12] "The 50Gbps Si photonics link", white paper, Intel corp., 2010.

[13] Y. A. Vlasov, "Silicon integrated nanophotonics: road from scientific explorations to practical applications", in Conference on Lasers and Electro-Optics (CLEO), 2012.

[14] G. Masini, J. Witzens, S. Sahni, B. Analui, C. Gunn, and G. Capellini, "Germanium photodetectors enable scalable silicon photonics", [Online], available:http://spie.org/x25789.xml.

[15] L. Fulbert, “State of the art in silicon photonics and technology roadmap”, Silicon photonics workshop - European Silicon Photonics Cluster, Munich, 2011.

[16] D. Kucharski, D. Guckenberger, G. Masini, S. Abdalla, Je. Witzens, and S. Sahni, “10Gb/s 15mW optical receiver with integrated germanium photo-detector and hybrid Inductor peaking in 0.13μm SOI CMOS technology”, in IEEE ISSCC Dig. Tech. Papers, 2010, pp. 360–361.

[17] Y. A. Vlasov, “Silicon CMOS-integrated nano-photonics for computer and data communications beyond 100G”, IEEE Communications Magazine, vol. 50, pp. 67-72, Feb. 2012.

[18] S. Mohan, M. Hershenson, S. Boyd, and Thomas H. Lee, “Bandwidth extension in CMOS with optimized On-Chip inductors”, IEEE J. Solid-State Circuits, vol. 35, no. 3, pp. 346-355, Mar. 2000.

[19] S. M. Park and H.-J. Yoo, “1.25 Gb/s regulated cascode CMOS transimpedance amplifier for gigabit Ethernet applications,” IEEE J. Solid-State Circuits, vol. 39, no. 1, pp. 112–121, Jan. 2004.

[20] E. M. Cherry and D. E. Hooper. “The design of wide-band transistor feedback amplifiers”, Proceedings of IEE, vol. 110, no. 2, Feb. 1963.

[21] S. Galal and B. Razavi, “10-Gb/s limiting amplifier and laser/modulator driver in 0.18 μm CMOS technology,” IEEE J. Solid-State Circuits, vol. 38, no. 12, pp. 2138–2146, Dec. 2003.

[22] T. H. Lee, The Design of CMOS Radio-Frequency Integrated Circuits, Second Edition, Cambridge University Press, 2003.

[23] J.-D. Jin and S. H. Hsu, “A 40-Gb/s transimpedance amplifier in 0.18-μm CMOS technology”, IEEE J. Solid-State Circuits, vol. 43, no. 6, pp. 1449–1457, Jun. 2008.

[24] W.-Z. Chen, Y.-L. Cheng, and D.-S. Lin, “A 1.8-V 10-Gb/s fully integrated CMOS optical receiver analog front-end”, IEEE J. Solid-State Circuits, vol. 40, no. 6, pp. 1388–1396, Jun. 2005.

[25] C.-F. Liao and S.-I. Liu “40Gb/s transimpedance-AGC amplifier and CDR circuit for broadband data receivers in 90nm CMOS,” IEEE J. Solid-State Circuits, vol. 43, no. 3, pp. 642–655, Mar. 2008.

[26] B. Analui, D. Guckenberger, D. Kucharski, and A. Narasimha, “A fully integrated 20-Gb/s optoelectronic transceiver implemented in a standard 0.13-μm CMOS SOI technology”, IEEE J. Solid-State Circuits, vol. 41, no. 12, pp. 2945-2955, Dec. 2006.

[27] T. Takemoto, F. Yuki, H. Yamashita, T. Ban, M. Kono, Y. Lee, T. Saito, S. Tsuji, and S. Nishimura, “A 25Gb/s x 4-channel 74mW/ch transimpedance amplifier in 65nm CMOS,” in Proc. IEEE Custom Integrated Circuits Conf. (CICC), 2010, pp. 1–4.

[28] C. Kromer, G. Sialm, T. Morf, M. Schmatz, F. Ellinger, D. Erni, and H. Jäckel, "A low-power 20-GHz 52-dBΩ transimpedance amplifier in 80-nm CMOS", IEEE J. Solid-State Circuits, vol. 39, no. 6, pp. 885-894, Jun. 2004

[29] J. Kim and J. F. Buckwalter, "A 40-Gb/s optical transceiver front-End in 45 nm SOI CMOS", IEEE J. Solid-State Circuits, vol. 47, no. 3, pp. 615-626, Mar. 2012.

[30] J. F. Buckwalter, X. Zheng, G. Li, K. Raj, and A Krishnamoorthy, "A monolithic 25-Gb/s transceiver with photonic ring modulators and Ge detectors in a 130-nm CMOS SOI process", IEEE J. Solid-State Circuits, vol. 47, no. 6, pp. 1309-1322, Jun. 2012.

[31] B. Razavi, Design of Analog CMOS Integrated Circuits, McGraw-Hill, 2001.

[32] D. Johns and K. Martin, Analog Integrated Circuits Design, John Wiley & Sons, 1997.

[33] A. Abidi, "Gigahertz transresistance amplifiers in fine Line NMOS", IEEE J. Solid-State Circuits, vol. 19, no. 6, pp. 986-994, Dec. 1984.

[34] R. G. Smith and S. D. Personick, "Receiver design for optical fiber communication systems," in Semiconductor Devices for Optical Communication. Berlin, Germany: Springer Verlag, 1980.

[35] F. Gatta, E. Sacchi, F. Svelto, P. Vilmercati, and R. Castello, "A 2-dB noise figure 900-MHz differential CMOS LNA", IEEE J. Solid-State Circuits, vol. 36, no. 10, pp. 1444-1452, Oct. 2001.

[36] J.-Y. Dupuy, F. Jorge, M. Riet, A. Konczykowska, and J. Godin, "InP DHBT transimpedance amplifiers with automatic offset compensation for 100-Gbit/s optical communications," in Proc. Microwave Integrated Circuits Conf. (EuMIC), 2010, pp. 341-344.

[37] M. Zuffada, "The industrialization of the silicon photonics: technology road map and applications", in Proc. Eur. Solid-State Circuit Conf. (ESSCIRC), 2012, pp. 7–13.

[38] D. V. Thourhout, "Silicon Photonics", Tutorial, Optical Fiber Conference (OFC), Mar. 2010.

[39] H. W. Bode. Network Analysis and Feedback Amplifier Design. D. Van Nostrand Company, New York, 1945.

[40] B. Zhu, T.F. Taunay, M. Fishteyn, X. Liu, S. Chandrasekhar, M. F. Yan, J. M. Fini, E. M. Monberg, and F. V. Dimarcello, "112-Tb/s Space-division multiplexed DWDM transmission with 14-b/s/Hz aggregate spectral efficiency over a 76.8-km seven-core fiber", Optics Express, vol. 19, no. 17, pp. 16665-16671, 2011.